



UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA

University of Modena and Reggio Emilia
Department of Engineering “Enzo Ferrari”

XXXVIII cycle of the Doctorate School in
Information and Communication Technologies (ICT)

From Human Visual Attention to Computational Models:

Scanpath Prediction and Real-World Applications

Ph.D. Dissertation

GIUSEPPE CARTELLA

Advisor: Prof. Rita Cucchiara
PhD Programme Coordinator: Prof. Luigi Rovati

Advisor:

Prof. Rita Cucchiara

University of Modena and Reggio Emilia

PhD Programme Coordinator:

Prof. Luigi Rovati

University of Modena and Reggio Emilia

Review Committee:

Prof. Andreas Bulling

University of Stuttgart

Prof. Raffaella Lanzarotti

University of Milan

“ITSERR - Italian Strengthening of the ESFRI RI RESILIENCE” - cod. Progetto IR0000014 – CUP B53C22001770006 - PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) MISSIONE 4, COMPONENTE 2, INVESTIMENTO 3.1 “Fondo per la realizzazione di un sistema integrato di infrastrutture di ricerca e innovazione” finanziato dall’Unione europea – NextGenerationEU - rif. Avviso MUR 3264/2021 - Decreto di concessione n.0000111 del 20/06/2022.



Finanziato
dall'Unione europea
NextGenerationEU



Ministero
dell'Università
e della Ricerca



Italiadomani
PIANO NAZIONALE
DI RIPRESA E RESILIENZA



UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA

The work described in this thesis has been carried out within the International Doctorate in Information and Communication Technologies (ICT), at the AImageLab research laboratory of the University of Modena and Reggio Emilia.

This dissertation was typeset by the author using $\LaTeX 2_{\epsilon}$, originally developed by Leslie Lamport and based on Donald Knuth's $\TeX$. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface.

Copyright © 2026 by Giuseppe Cartella

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission from the author.

To my family

From Human Visual Attention to Computational Models:

Scanpath Prediction and Real-World Applications

ABSTRACT

Human visual attention is a fundamental cognitive process that enables individuals to efficiently interpret complex visual environments. By selectively prioritizing salient and task-relevant regions, humans extract critical information while suppressing irrelevant stimuli. This selective mechanism supports perception, decision-making, and interaction with a dynamic world. Foundational work in neuroscience and cognitive science has established key principles of spatial attention, feature-based selection, and the interaction between top-down and bottom-up control. In parallel, advances in Artificial Intelligence and Computer Vision have inspired autonomous systems that emulate human prioritization of visual information. Unlike traditional saliency models that predict static regions of interest, scanpath models capture both spatial and temporal aspects of visual exploration, including the sequence, timing, and trajectory of gaze shifts. This thesis introduces probabilistic and generative frameworks for scanpath prediction, contributing to a deeper understanding and modeling of temporal dynamics and inter-subject variability.

We propose TPP-Gaze, a mathematically principled framework, inspired by the theory of Neural Temporal Point Processes, which conceptualizes scanpaths as sequences of stochastic events occurring irregularly over time. This formulation addresses a key limitation of existing approaches: while many accurately model *where* to look, they often lack a principled representation of *when* gaze shifts occur. By explicitly modeling temporal evolution, TPP-Gaze provides a

more complete description of visual attention dynamics. Human attention is inherently variable across individuals, influenced by cognitive biases, preferences, and context. To capture such variability, we introduce ScanDiff, a stochastic diffusion-based model capable of generating diverse and realistic gaze trajectories. Unlike methods that produce a single average scanpath, ScanDiff reflects the natural heterogeneity of visual behavior across observers.

Beyond methodological contributions, this research provides a comprehensive overview of current models and datasets, highlighting trends, challenges, and opportunities for integrating human attentional mechanisms into artificial systems. The practical relevance of the proposed models is demonstrated through eye-tracking experiments and data collection across multiple domains. These include studies on gaze behavior in authentic versus manipulated images, the integration of perceptual signals into deepfake detection frameworks, and the analysis of attention in neurodevelopmental conditions. Additionally, the thesis introduces new resources, including a large-scale multimodal egocentric dataset capturing gaze, motion, and environmental context in culturally diverse settings. These data reveal distinctive attentional patterns and cognitive mechanisms, informing the design of human-centered computational models for applications such as media authenticity assessment, behavioral analysis, and multimodal attention modeling.

In conclusion, this thesis advances the study of human visual attention by *(i)* developing computational models that capture the spatio-temporal structure of scanpaths, *(ii)* providing empirical evidence into gaze dynamics under varied conditions, and *(iii)* exploring real-world applications of attention modeling. By bridging cognitive science and artificial intelligence, this work deepens theoretical understanding while enhancing the computational modeling of human-like visual behavior.

Dall' Attenzione Visiva Umana ai Modelli Computazionali: Predizione dello Scanpath e Applicazioni in Scenari Reali

SOMMARIO

L'attenzione visiva umana è un processo cognitivo che consente agli individui di interpretare in modo efficiente ambienti visivi complessi. Attraverso la prioritizzazione di regioni salienti e rilevanti per il problema, gli esseri umani estraggono informazioni critiche sopprimendo al contempo gli stimoli irrilevanti. Questo meccanismo di selezione supporta la percezione, il processo decisionale e l'interazione con un mondo dinamico. Studi nelle neuroscienze hanno stabilito principi chiave dell'attenzione spaziale, della selezione basata sugli attributi visivi e dell'interazione tra controllo top-down e bottom-up. I progressi nell'Intelligenza Artificiale e nella Computer Vision hanno ispirato lo sviluppo di sistemi autonomi in grado di emulare la prioritizzazione umana delle informazioni visive.

A differenza dei modelli di salienza tradizionali, che predicono regioni di interesse statiche, i modelli di scanpath catturano sia gli aspetti spaziali che quelli temporali dell'esplorazione visiva, inclusi la sequenza, la tempistica e la traiettoria dello sguardo. Questa tesi introduce framework probabilistici e generativi per la generazione degli scanpath, contribuendo a una comprensione più approfondita delle dinamiche temporali e della variabilità tra individui.

Proponiamo TPP-Gaze, un framework ispirato alla teoria dei Neural Temporal Point Processes, che concettualizza gli scanpath come sequenze di eventi stocastici che accadono in modo irregolare nel tempo. Tale formulazione affronta un limite chiave degli approcci esistenti: mentre molti modelli rappresentano accuratamente *dove* guardare, spesso mancano di una rappresentazione rigorosa di *quando* avvengono gli spostamenti dello sguardo. Modellando esplicitamente

l'evoluzione temporale, TPP-Gaze fornisce una descrizione più completa delle dinamiche dell'attenzione visiva. L'attenzione umana è intrinsecamente variabile tra individui, influenzata da bias cognitivi, preferenze e contesto. Per catturare tale variabilità, introduciamo ScanDiff, un modello stocastico basato su diffusione, capace di generare traiettorie di sguardo diverse e realistiche. A differenza dei metodi che producono un singolo scanpath medio, ScanDiff riflette la naturale eterogeneità del comportamento visivo.

Oltre ai contributi metodologici, questa ricerca fornisce una panoramica completa dei modelli e dei dataset attuali, evidenziando trend, sfide e opportunità per integrare i meccanismi attentivi umani nei sistemi artificiali. La rilevanza pratica dei modelli proposti è dimostrata attraverso esperimenti di eye-tracking e raccolta dati in diversi domini. Questi includono studi sul comportamento visivo in immagini autentiche rispetto a immagini manipolate, l'integrazione di segnali percettivi per il rilevamento dei deepfake e l'analisi dell'attenzione in condizioni neuroevolutive.

Inoltre, la tesi introduce un dataset egocentrico multimodale che cattura sguardo, movimento e contesto ambientale in ambiti culturalmente diversificati. Questi dati rivelano pattern attentivi distintivi, contribuendo alla progettazione di modelli centrati sull'uomo per applicazioni quali la valutazione dell'autenticità dei media, l'analisi comportamentale e la modellazione multimodale dell'attenzione.

In conclusione, questa tesi avanza lo studio dell'attenzione visiva umana attraverso (i) lo sviluppo di modelli computazionali che catturano la struttura spazio-temporale degli scanpath, (ii) la definizione di evidenze empiriche sulle dinamiche dello sguardo in diverse condizioni e (iii) l'esplorazione di applicazioni reali della modellazione dell'attenzione. Collegando scienze cognitive e intelligenza artificiale, questo lavoro approfondisce la comprensione teorica e migliora la modellazione computazionale del comportamento visivo umano.

Contents

ABSTRACT	I
SOMMARIO	III
1 INTRODUCTION	1
1.1 Research Objectives & Contributions	4
1.2 Thesis Organization	9
2 HUMAN VISUAL ATTENTION MODELING: FOUNDATION AND ADVANCES	11
2.1 Saliency Models	12
2.2 Scanpath Models	15
2.3 Benchmarks & Evaluation Metrics for Human Attention Models	19
2.3.1 Benchmarks for Saliency and Scanpath Prediction . . .	19
2.3.2 Evaluation of Saliency Prediction Models	23
2.3.3 Evaluation of Scanpath Prediction Models	26
3 PREDICTING HUMAN GAZE BEHAVIOR: FROM TEMPORAL DYNAMICS TO GENERATION DIVERSITY	31
3.1 Modeling Gaze Dynamics in Space and Time with Neural Temporal Point Processes	33
3.1.1 Temporal Point Process Models for Human Gaze Dynamics	35
3.1.2 The TPP-Gaze Model	39
3.1.3 Model Inference	44
3.1.4 Experimental Evaluation	44
3.1.5 Scanpath Prediction	46
3.1.6 Applications	55
3.1.7 Extension to Visual Search Tasks	55
3.1.8 Discussion and Limitations	59
3.2 Modeling Human Gaze Behavior with Diffusion Models for Unified Scanpath Prediction	60
3.2.1 Diffusion Models for Trajectory Generation	62

3.2.2	ScanDiff: A Diffusion-based Model for Unified Scan-path Prediction	67
3.2.3	Training and Inference	71
3.2.4	Experimental Results	73
3.2.5	Scanpath Variability Analysis	83
3.2.6	Discussion and Limitations	85
4	INTEGRATION OF HUMAN ATTENTION IN AI MODELS: A REVIEW	87
4.1	Image and Video Processing	90
4.1.1	Visual Recognition	90
4.1.2	Graphic Design	91
4.1.3	Image Enhancement and Manipulation	91
4.1.4	Image Quality Assessment (IQA)	92
4.2	Vision-and-Language Applications	94
4.2.1	Automatic Captioning	94
4.2.2	Visual Question Answering (VQA)	95
4.3	Language Modelling	96
4.3.1	Machine Reading Comprehension	96
4.3.2	Natural Language Understanding	97
4.4	Domain-Specific Applications	98
4.4.1	Robotics	98
4.4.2	Autonomous Driving	99
4.4.3	Medicine	100
4.5	From Gaze-Guided AI to Gaze-Guided Understanding	102
5	GAZE-GUIDED LEARNING FOR VISION AND BEHAVIORAL UNDERSTANDING	103
5.1	Gaze Patterns as Cues for DeepFake Detection	105
5.1.1	Background on Deepfake Detection and The Role of Human Gaze	106
5.1.2	Human Visual Perception of Manipulated Images	107
5.1.3	Visual Saliency to Detect Religious Image Manipulation	116
5.2	Unraveling Neurodivergent Gaze Behavior through Visual Attention Causal Graphs	131
5.2.1	Background on Gaze Modeling for Neurodevelopmental Disorders	133
5.2.2	Proposed Method	134
5.2.3	Aggregated Item-Level Attention Allocation Traces	136

5.2.4	Visual Attention Causal Graphs (VA-CG)	136
5.2.5	ASD Detection via VA-CGs	139
5.2.6	Experimental Analysis	140
5.3	Sanctuarial-Gaze: A Multimodal Egocentric Dataset for Human Attention Analysis in Religious Sites	143
5.3.1	An Overview on Egocentric Datasets	144
5.3.2	Sanctuarial-Gaze Dataset	147
5.3.3	Dataset Analysis	153
5.3.4	Analysis of the Exploration Patterns	160
6	CONCLUSIONS	163
6.1	Summary of Contributions	163
6.2	Open Challenges and Future Directions	166
6.3	Closing Remarks	169
	LIST OF PUBLICATIONS	171
	BIBLIOGRAPHY	173
	ACKNOWLEDGEMENTS	215
A	PH.D. ACTIVITIES	219
B	SUPPLEMENTARY MATERIAL FOR CHAPTER 3	223
C	SUPPLEMENTARY MATERIAL FOR CHAPTER 5	247

Everything evolves, don't cling to anything. Be a river, not a rock.

1

Introduction

HUMAN visual attention constitutes one of the fundamental mechanisms through which humans perceive and interact with the real world. The human visual system enables individuals to navigate complex environments, prioritize relevant information, and make rapid decisions under uncertainty. Vision is not merely a passive recording of the external stimuli, but rather an active process of information sampling and interpretation. What we look at, and in what order, reflects our goals, expectations, and experiences. Attention plays an essential role in human cognition and survival. Due to the overwhelming amount of incoming sensory information at any given time, evolution has molded the ability of selective attention. The totality of inputs cannot be fully processed; thus, the brain transforms sensory data into structured and meaningful representations, allowing for the most effective action to survive in a changing environment. Eye movements provide a tangible trace of this process, offering measurable insights into the dynamics of attention.

The study of visual attention has long been central to cognitive science, neuroscience and psychology. Behavioral studies, neurophysiological investigations, and computational neuroscience research have established theoretical frame-

works describing how perceptual prioritization operates in the brain. These frameworks distinguish between *bottom-up* and *top-down* processes, *overt* and *covert* shifts of attention, and *spatial* versus *feature-based* selection mechanisms, providing foundational conceptual tools for understanding attentional selection. Within computational neuroscience, attention has been modeled as a set of modulatory mechanisms influencing sensory processing and decision-making, including seminal frameworks such as the biased competition theory of Desimone and Duncan [113] and the normalization model of attention proposed by Reynolds and Heeger [295]. Together, these perspectives have contributed to a more comprehensive understanding of the neural and cognitive processes.

However, early theories of attention were largely descriptive. Although they provided comprehensive explanations of attentional phenomena, they did not yield predictive mechanisms capable of generalizing across diverse contexts. Moreover, many empirical findings were derived from controlled laboratory settings using simplified stimuli and constrained tasks. While these approaches enabled precise hypothesis testing, they often failed to capture the complexity and semantic richness of natural visual behavior. As a consequence, there emerged a practical need to complement descriptive theories with more quantitative and predictive frameworks, motivating the development of computational models of visual attention. Computational modeling serves several essential purposes. It translates theoretical concepts into explicit mathematical formulations and algorithmic procedures. Furthermore, it enables simulation and prediction, allowing researchers to test how well a model reproduces empirical findings such as fixation patterns, reaction times, or neural responses. Over the past few decades, extensive research has been conducted, resulting in the proposal of several models aimed at predicting attention allocation towards visual and, more recently, multimodal stimuli. These models have garnered increasing interest in the computer vision and pattern recognition community since the seminal publication of the Itti et al. [177] model, which translated principles of early visual processing into a practical algorithm for predicting bottom-up attentional allocation, paving the way to potential applications in various fields ranging from image and video analysis, robotics, natural language processing, and autonomous driving. Early

computational approaches relied on hand-crafted features and explicitly designed components grounded in neurobiological discoveries. While interpretable and theoretically motivated, these models were limited in their ability to scale to complex, real-world stimuli.

The advent of Deep Learning, together with advances in eye-tracking technology and the availability of larger gaze datasets, facilitated the transition toward data-driven methodologies. Deep neural networks have demonstrated remarkable improvements in saliency and scanpath prediction by learning high-dimensional mappings between visual input and fixation distributions. In this regard, it has become evident that computational models that mirror biological processes can enhance agents' ability to maximize their information intake over time with respect to an internal or external goal.

In robotics and active perception [8, 23], attention mechanisms have been studied as a means of improving perceptual efficiency and guiding action in real-time environments. In these settings, attentional selection is closely tied to the problem of resource allocation under computational and sensory constraints, enabling agents to focus processing on task-relevant regions of the environment. In addition, as noted in [404], the capacity of an AI agent to perceive and interpret human gaze contributes to an efficient and seamless human-machine interaction. By deciphering the nuances of gaze patterns, an AI agent gains the ability to engage with users on a deeper level, anticipating their needs and preferences, reacting intuitively to user inputs. In human-computer interaction (HCI), attentional processes have also been extensively investigated both as a modality for gaze-based interaction and as a valuable source of information for understanding user behavior. Insights derived from eye-tracking studies have informed user modeling and the design of interfaces that better align with human perceptual strategies, ultimately improving usability, safety, and the overall user experience [178, 179].

Understanding and modeling human visual attention is also important not only for theoretical reasons but also for its broad societal and technological impact. In clinical psychology, studying attentional mechanisms provides a deeper understanding of disorders such as attention-deficit/hyperactivity disorder (ADHD) or autism spectrum conditions.

A substantial body of eye-tracking research has documented atypical visual attention patterns, including reduced fixation on eyes and socially relevant stimuli, and altered scanpath behavior in naturalistic environments [100, 259, 357]. Developmental studies further indicate that these differences emerge early in life, particularly in joint attention and gaze-following behaviors, which are considered foundational mechanisms for social cognitive development [34, 342]. These findings suggest that atypical attentional allocation in autism may reflect both altered perceptual salience and differences in top-down attentional control.

In education, eye-tracking methodologies have been increasingly employed to investigate learning strategies, expertise acquisition, reading comprehension, and classroom interactions, providing objective measures of learners' cognitive processes and supporting the design of more effective instructional environments [181]. In marketing and consumer research, the analysis of visual attention has offered valuable insights into how individuals explore advertisements, product displays, and retail environments, contributing to the understanding of consumer decision-making and informing the design of more effective communication and product presentation strategies [365]. Collectively, these applications underscore the broad relevance of visual attention research across disciplines and highlight its potential to address both scientific and practical challenges.

1.1 RESEARCH OBJECTIVES & CONTRIBUTIONS

Despite these advances, several challenges remain, and a fundamental question arises: *how closely do these systems reflect the mechanisms that guide human perception?*

CURRENT LIMITATIONS IN ATTENTION MODELING. Most computational models conceptualize attention as a static mapping from an image to a saliency map. In contrast, human gaze behavior unfolds over time: the decision of where to look next is inherently sequential, shaped by prior fixations. The limitations of saliency models have motivated the development of scanpath models, which aim to predict gaze trajectories in both space and time, thereby capturing the temporal dynamics of visual exploration. However, existing scanpath prediction so-

lutions often lack a theoretically grounded formulation of temporal dynamics. Instead, they typically rely on heavily engineered architectures that do not explicitly disentangle spatial selection from fixation duration. As a result, these models provide a limited understanding of the distinct mechanisms underlying where and how long we look.

Another major challenge concerns the variability of gaze behavior. Most computational models are trained to predict aggregate scanpaths averaged across observers. Yet visual attention is shaped by individual differences in cognitive strategies, prior experiences, task interpretation, and perceptual biases. Treating variability as noise obscures potentially meaningful structure in gaze behavior. Failing to explicitly model such variability limits the ability to capture the nuanced ways in which individuals explore visual scenes. A principled framework for gaze prediction should therefore account for both shared attentional patterns and systematic individual differences. Current evaluation protocols predominantly reward performance on averaged behaviors and offer little support for assessing how well models capture variability across observers.

While gaze prediction has matured into a well-established task with clear practical relevance, many application domains remain underexplored in leveraging gaze as a supervisory or guiding signal. Moving beyond the development of attention models, this thesis investigates gaze as a functional component within real-world AI systems, examining how human visual behavior can actively inform and enhance intelligent decision-making processes. The research spans emerging technical challenges such as deepfake detection, as well as visually and culturally significant environments, including religious architecture. In parallel, it explores the causal characteristics of gaze behavior in neurodevelopmental disorders. In these contexts, incorporating human gaze signals into the training process may improve interpretability, robustness, and alignment with human perceptual priorities. These considerations are particularly salient in the era of modern AI. As artificial agents are increasingly deployed in socially embedded, safety-critical, and interactive environments, there is growing recognition that systems must also be interpretable, human-aligned, and capable of operating in ways that resonate with human perceptual and behavioral patterns.

THESIS CONTRIBUTIONS. In light of the challenges outlined above, the central objective of this thesis is to develop a unified computational framework that models human visual attention as a temporally structured, goal-dependent, and informative process, while leveraging gaze signals to improve the interpretability and performance of artificial intelligence systems.

More specifically, this thesis has the following objectives and makes the following contributions:

- **Formalizing gaze as a sequential and adaptive information sampling strategy.** We introduce a computational approach that explicitly models the sequential dependencies between fixations, capturing the dynamic evolution of attention over time. Inspired by the theory of Neural Temporal Point Processes, the proposed framework enables the generation of realistic scanpaths that better reflect natural viewing behavior.
- **Accounting for inter-individual variability and rethinking evaluation protocols.** The thesis proposes a diffusion-based strategy that treats variability as structured information rather than noise. This contributes to a more nuanced and cognitively grounded representation of attentional processes. We critically examine current evaluation protocols based on population-averaged scanpaths and propose alternative perspectives that better assess individual-level behavior.
- **Leveraging gaze as a supervisory and alignment signal for AI systems.** The thesis sheds light on the integration of human gaze behavior as a supervisory signal in domains where its potential remains largely underexplored. In particular, within the context of deepfake detection, we demonstrate that human observers exhibit systematic and distinguishable attentional patterns when viewing authentic versus manipulated images. By incorporating such gaze-derived signals into deepfake detection pipelines, we show that it is possible to guide model attention toward perceptually informative regions, thereby improving detection performance. Another contribution concerns neurodevelopmental disorders. While it is widely

recognized that gaze is an important biomarker for the early diagnosis of autism, the causal nature of gaze is still underexplored in the field. Our work aims to understand whether the causal structure of attention allocation allows to differentiate subjects with autism disorders from typically developing controls, enabling a more interpretable approach.

- **Collecting new eye-tracking data in ecologically valid scenarios.** This work introduces the first multimodal eye-tracking dataset collected in the distinctive context of religious environments, capturing gaze behavior as individuals freely explore different sanctuaries. This resource reflects attention allocation in a culturally, architecturally, and semantically rich setting, where visual exploration is driven by symbolic meaning, prior beliefs, and personal engagement. The analysis of the collected dataset reveals systematic differences in the objects and regions that attract the attention of religious and non-religious participants, highlighting the role of individual background and belief systems in guiding visual exploration.

A graphical overview of the thesis is reported in Fig. 1.1. It maps the main research areas, models, datasets, and how each chapter or publication connects to these components.

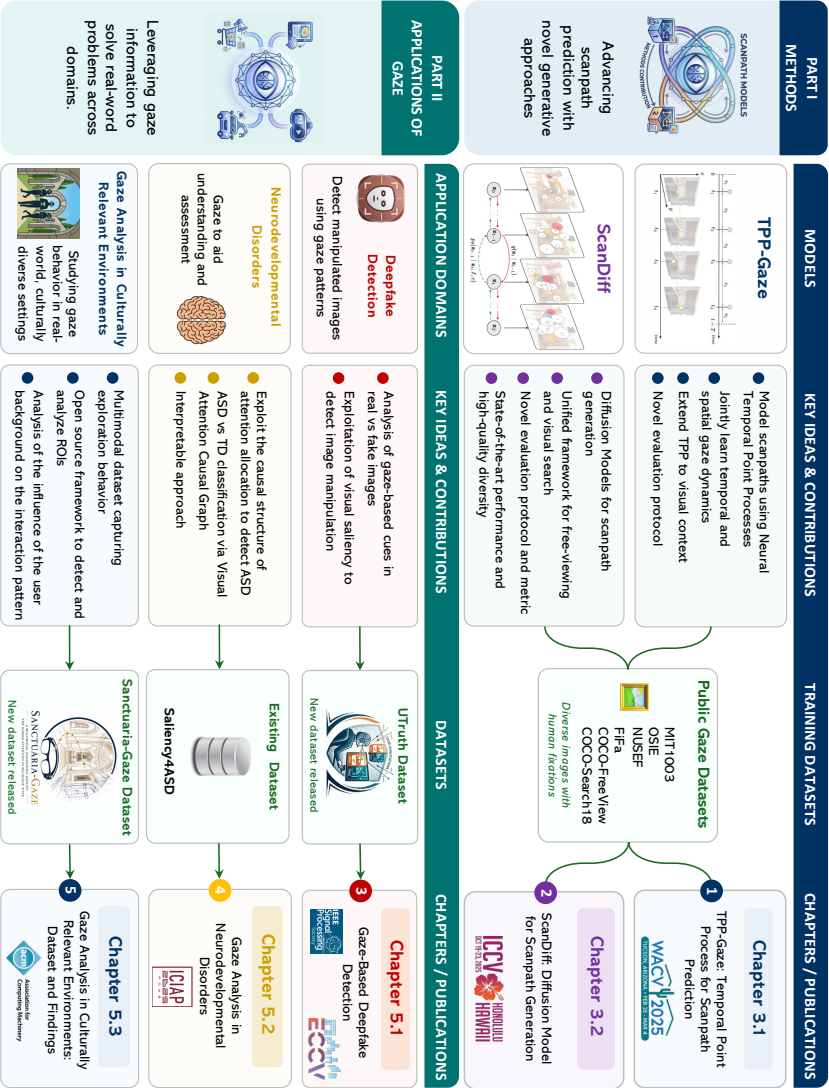


Figure 1.1: Overview of the PhD Thesis. This thesis advances the state-of-the-art in gaze modeling and demonstrates the power of gaze information in addressing challenges across several domains, including technology, health, and society.

1.2 THESIS ORGANIZATION

The remainder of the thesis is organized as follows:

- **Chapter 2** discusses theoretical and methodological foundations of human visual attention from both cognitive and computational perspectives. It reviews classical saliency models and scanpath prediction approaches, tracing their evolution from handcrafted formulations to modern deep learning-based methods. This chapter also summarizes the principal benchmarks and evaluation metrics commonly adopted in the field.
- **Chapter 3** identifies the key limitations of current scanpath prediction models and motivates the need for more principled frameworks capable of explicitly modeling the temporal dynamics and stochastic nature of gaze behavior. The TPP-Gaze and ScanDiff models are presented, along with detailed analyses and experimental validation.
- **Chapter 4** broadens the perspective by reviewing how human attention mechanisms can be integrated into AI systems across diverse domains, including image and video processing, vision-and-language, language modeling, robotics, autonomous driving, and medical applications.
- **Chapter 5** presents the application-oriented contributions of this thesis, focusing on gaze-assisted deepfake detection and behavioral analysis in neurodevelopmental contexts. It also introduces Sanctuaria-Gaze, a multi-modal egocentric dataset designed to investigate attention in religious and culturally significant environments.
- **Chapter 6** concludes the thesis by summarizing the main findings and outlining open challenges and future research directions.

2

Human Visual Attention Modeling: Foundation and Advances

VISUAL attention is a fundamental cognitive process that enables humans to efficiently process complex visual scenes by selectively prioritizing relevant information while filtering out irrelevant stimuli [192]. The Feature Integration Theory, proposed by Treisman and Gelade [346], provides a framework for understanding this selective mechanism, describing visual perception as a two-stage process. In the first stage, named *preattentive stage*, the brain processes low-level features (*e.g.* color, shape, size, and movement), automatically and in parallel, without conscious awareness. The second stage, the *focused attention stage*, involves the integration of these features, enabling the perception of coherent and unified objects. Attentional allocation is guided by two complementary mechanisms. *Bottom-up* attention is driven by the inherent salience of visual stimuli. In contrast, *top-down* attention reflects higher-level cognitive factors, including task demands, prior knowledge, expectations, and semantic

This Chapter discusses topics from the following paper: “G. Cartella *et al.*, Trends, Applications, and Challenges in Human Attention Modelling, IJCAI 2024” [60].

context. These mechanisms operate across both preattentive and focused stages, shaping how humans prioritize information in complex scenes.

The study of visual attention in Computer Vision has seen significant progress since the seminal works in [8, 25, 26, 177]. Empirical studies of visual attention typically rely on eye-tracking experiments, which record sequences of fixations and saccades, referred to as *scanpaths*, providing a proxy for overt attention. Depending on the experimental setup, eye movements can be collected under *free-viewing* conditions, where observers explore images without an explicit task, or under *goal-directed* settings, where attention is guided by specific objectives such as visual search, questions, or referring expressions. While free-viewing behavior is dominated by bottom-up factors, goal-directed viewing exhibits a stronger influence of top-down control, leading to different gaze patterns.

Computational modeling of visual attention has given rise to two closely related, yet conceptually distinct, prediction tasks. The first is *saliency prediction*, which aims to estimate a spatial map indicating how likely each image location is to attract human gaze. Such saliency maps summarize fixation behavior across observers and time. The second task is *scanpath prediction*, which seeks to model the sequential nature of visual exploration by predicting ordered sequences of fixations. In contrast to saliency maps, scanpaths preserve temporal dynamics and capture not only *where* people look, but also *in which order* they attend to different regions of a scene. In the following, we review prior work in both directions, covering both classical and deep learning-based approaches.

2.1 SALIENCY MODELS

Visual saliency prediction aims to estimate the spatial distribution of human visual attention in an image or scene. It is commonly represented as a saliency map, where each location indicates the likelihood of being fixated by an observer. Saliency prediction has proven valuable across a wide range of applications, including graphic design [222], visual recognition [282], image enhancement [252], and image quality assessment [414]. The objective of saliency prediction models is to approximate this distribution by computing or learning which

regions are visually or semantically prominent, based on low-level cues (*e.g.* color, contrast, orientation, size, motion) and high-level factors such as objects, faces, text, and scene context. Saliency prediction abstracts away the temporal order of fixations and focuses on estimating the spatial importance of image regions under a given viewing condition.

Early models are largely inspired by findings from cognitive neuroscience and focused on bottom-up, stimulus-driven mechanisms. They exploit image processing techniques based on the identification of hand-crafted features correlated with human attention. Inspired by the work of Koch and Ullman [204], a seminal implementation of a saliency prediction model was provided by Itti et al. [177]. Attention is described as a combination of bottom-up processes, based on conspicuity maps associated with basic features of the visual stimuli such as colour, intensity, and edges. These maps are then combined into a single saliency map. Following this direction, other methods explored the idea of combining low-level features [54, 130, 154], also including center-surround cues [257, 398]. However, the research by Einhäuser et al. [123] challenged the assumption that early visual features directly drive attention. Instead, they showed that observers attend to salient objects, with low-level saliency having only an *indirect* influence on gaze allocation.

In a similar vein, Judd et al. [194] demonstrated that reliance on purely bottom-up computation fails to accurately reflect human viewing behavior. By training an SVM classifier that incorporates higher-level semantic features, their method achieved a substantial performance improvement, highlighting the importance of top-down cues. Together, these studies marked a shift from purely handcrafted, bottom-up pipelines toward more data-driven approaches that integrate low-level information with higher-level, semantically meaningful concepts. As a result, Goferman et al. [144] defined a context-aware saliency model that combines both local and global visual cues. Zhao and Koch [411] further investigated learning-based fusion strategies to adaptively weigh multiple saliency maps, improving robustness across diverse scenes. Borji [44] employed boosting techniques to combine a large set of heterogeneous features, demonstrating the benefits of ensemble learning for saliency prediction. Despite leveraging more ad-

vanced machine learning techniques, these approaches still rely on handcrafted features and manually designed representations. A major shift occurred with the advent of deep learning, which enabled saliency models to learn hierarchical feature representations directly from data and significantly surpass the performance of handcrafted methods. Vig et al. [351] followed an entirely automatic data-driven approach and proposed an *Ensemble of Deep Networks*. Unfortunately, the first deep-learning-based solutions were still limited due to the lack of large-scale datasets suitable for training.

To overcome this limitation, Kümmerer *et al.* applied transfer learning and developed DeepGaze I [215] and DeepGaze II [216] models, built upon AlexNet [208] and VGG-19 [317], respectively. Collecting large-scale datasets of human fixations through eye-tracking devices is known to be very challenging and time-consuming. At the same time, the performance of deep-learning models is highly dependent on the availability of sufficiently large datasets. For this reason, Jiang et al. [187] presented SALICON, a dataset collected by using mouse clicks as a proxy for human visual attention, and that still represents one of the biggest data sources to train saliency models. The introduction of SALICON has represented a significant development in data-driven approaches, leading to the emergence of more modern deep learning-based architectures for the task of saliency prediction, especially based on convolutional neural networks [89, 210, 269, 270, 338].

Among them, Cornia et al. [91] proposed a convolutional long short-term memory that focuses on the most salient regions of the input image to iteratively refine the predicted saliency map. To improve saliency prediction performance, other works integrated additional sources of information such as external knowledge [407], scene context [209], and object dissimilarity [18]. Lou et al. [242] advanced the field by exploring, for the first time, the integration of Transformer-based models within a saliency prediction framework. In particular, a Transformer encoder is built on top of the features extracted by a convolutional network to capture long-range contextual information. Aydemir et al. [19] developed a data augmentation strategy based on latent diffusion to face the limited diversity and quantity of annotated data. Human visual attention is not static;

rather, it evolves over time. Therefore, from a different perspective, based on the temporally evolving patterns characterizing human attention, Fosco et al. [134] and Aydemir et al. [20] defined novel saliency models capable of performing time-specific saliency prediction. Beyond temporal evolution, visual attention exhibits high inter-subject variability, and region prioritization also depends on the observer's preferences. A related trend [72, 375, 376] has tackled this challenge, laying the foundations for the design of custom applications based on users' interests. Saliency prediction has also expanded to cover diverse types of data beyond natural scenes, including commercial advertisements [133, 184, 230] and user interfaces [190, 316]. Research on computational visual attention has made significant progress to integrate image and video saliency within a single unified framework [118]. On a similar line, some works developed unified image-based saliency models that can work across different types of visual stimuli (*e.g.* natural scenes, web pages, e-commercial images) and tasks [162, 223].

2.2 SCANPATH MODELS

Saliency prediction estimates a spatial distribution representing the prominence of information at each location in the visual field. However, by collapsing attention to a saliency map, the temporal component of the attentional deployment is lost. Human eye movements unfold as sequences of *fixations* and *saccades* (*i.e.* shifts from one fixation to another), commonly referred to as *scanpaths*. Two scanpaths that visit analogous locations in an image but in completely different orders generate identical saliency maps. Yet, the exploration order of a visual stimulus is of particular importance since it better reflects the conspicuity of the image elements with respect to the observer, as well as their internal strategy. Scanpath prediction offers a more comprehensive and natural representation of attention, providing deeper insights into the mechanisms of visual exploration. The literature on scanpath prediction exhibits a trend similar to saliency prediction, shifting from model-based to deep-learning approaches. Classical models for simulating gaze behavior emphasize the central importance of the saliency map. Inspired by the work of Itti et al. [177], scanpaths are typically simulated using the

Winner-Take-All (WTA) algorithm[52, 174, 175, 353]. In this framework, the initial fixation is selected as the location with the highest saliency. Subsequent fixations are generated by iteratively inhibiting the previously attended location and then shifting attention to the next highest saliency location. However, this greedy procedure yields deterministic and often unrealistic scanpaths and suffers from several limitations: *i*) accuracy of the generation is dependent on the quality of the saliency map, *ii*) next fixation prediction is not conditioned on the visual information and the history of previous fixations, *iii*) assumption that gaze shifts are equally likely in all directions contradicts empirical observations of human eye movement behavior.

Therefore, many studies have been dedicated to modeling scanpaths, in a more sophisticated manner than the previously described winner-take-all algorithm [163, 234, 332, 339, 361]. Boccignone and Ferraro [40] introduced the constrained Levy algorithm and modeled gaze shifts as a stochastic process to account for the variability of human gaze. Le Meur and Liu [221] showed that combining eye-movement generation with oculomotor biases, such as saccade amplitudes and orientations, improves the quality of the generated scanpath. Wang et al. [361] used the principle of information maximization to generate a scanpath, consisting of three key components: Visual Working Memory (VWM), reference sensory responses, and foveated imaging. Jiang et al. [186] adopted the least-squares policy iteration (LSPI), based on reinforcement learning, to mimic human behavior.

Differently, Zanca et al. [394] conceived a visual attention scanpath as the motion of a unitary mass within gravitational fields generated by salient visual features. Recent approaches exploited the semantics of features learned by deep neural networks to improve modeling. A considerable amount of research has been dedicated to predicting scanpaths under free-viewing conditions, where the observer has no predefined task [15, 80, 214, 260, 283, 331]. PathGAN [15] adopts adversarial training and is composed of a generator and a discriminator trained on image features and gaze sequences. Ngo and Manjunath [260] proposed a model that extracts visual features from a CNN and are given as input to a Long-Short Term Memory (LSTM) model, which acts as a form of VWM

integrating scene information over time. The work by Chen and Sun [80], further extended in [331], explicitly incorporates the Inhibition-of-Return (IOR) mechanism. The model, based on an LSTM network, operates over regions of interest (ROIs), instead of predicting fixations at the pixel level. DeepGaze III represents another influential model designed by Kümmerer et al. [214]. This model entangles a spatial priority network that outputs a priority map, with a scanpath network conditioned on the history of previous fixations.

While most of the research has concentrated on developing models in free-viewing conditions, recent research has shifted focus toward goal-driven attention modeling, where observers actively engage in a specific task. A prominent example is visual search [122, 160, 369, 396], in which an observer is required to locate a specific target within a scene. Early visual search models typically assumed simple targets known to the observer [368] or relied on explicit images of the target as input [122, 396]. Subsequent works incorporated richer sources of information, such as spatial relationships between objects [17] and global scene context [344]. More recently, deep learning-based approaches have emerged [4, 402], enabling flexible modeling of goal-directed gaze behavior.

In this context, the introduction of the COCO-Search18 dataset [78] marked a significant milestone, providing a large-scale benchmark for studying human gaze behavior in visual search scenarios [256, 386, 388, 389, 395]. Yang et al. [386] defined a visual search framework based on inverse reinforcement learning, treating each fixation as a possible source of reward. Mondal et al. [256] introduced GazeFormer, a transformer-based model that encodes the search target using a language model and jointly processes visual and semantic information through an encoder-decoder architecture. The authors proposed the ZeroGaze task, which evaluates the ability to predict gaze behavior for unseen target categories. Another related line of research investigates how humans behave when searching for objects that are absent from the scene, further highlighting the role of high-level goals and expectations in guiding visual attention [389].

Beyond visual search, other works studied human visual attention in the context of Visual Question Answering (VQA), a task where the model is asked to answer a question about an image by jointly reasoning over visual and linguis-

tic information [10, 147]. The study of human attention for the VQA task was pioneered by Das et al. [107]. They compared human attention with machine attention, showing that artificial models do not look at the same regions as humans. Another related work [71] introduced a novel dataset and a computational model that unfolds attention in several steps, emulating the visual reasoning process of humans. Chen et al. [74] proposed an approach based on deep reinforcement learning to generate scanpaths conditioned on the visual question answering task.

In contrast to these works that only analyze attention on the visual modality, Sood et al. [326] took a different direction and collected VQA-MHUG, a multimodal dataset containing attention data both on the images and questions. Another important study in the field of cognitive science is the capacity of humans to use referring expressions to drive attention. While hearing a referring expression, humans incrementally integrate visual information with word-by-word linguistic guidance [195, 334]. Motivated by this evidence, recent works have investigated predicting attention dynamically as observers process visual scenes while hearing an audio trace. In particular, Mondal et al. [255] introduced the task of *incremental object referral*, aiming to predict gaze fixations grounded with each word of the sentence referring to a specific target in the scene. Instead, Liu et al. [237] studied scanpath prediction driven by more general auditory cues describing the overall scene, rather than a specific target.

The growing interest in studying diverse attention tasks has driven the development of unified architectures designed to adapt seamlessly across a variety of scenarios [76, 388]. For instance, Chen et al. [76] leveraged vision-and-language models to jointly predict scanpaths and generate language-based explanations. The majority of the models discussed so far treat attention in a uniform manner, training on aggregated data from multiple observers learning an “average” gaze behavior. However, human attention is inherently variable [297, 379]. Each individual exhibits unique exploration strategies, preferences, and intents. Accurately modeling the uniqueness of each individual is crucial for the development of adaptive attention models in applications like personalized human-computer interaction or assistive technologies, to improve usability. In this context, recent methods tackled this challenge and predicted gaze sequences conditioned on a

specific user embedding [75, 189]. Xue et al. [378] further advanced the field by addressing the problem in a few-shot setting.

Another line of research has focused on scanpath generation targeting 360° images [16, 248, 284, 330]. Assens Reina et al. [16] introduced saliency volumes to capture the temporal evolution of scanpaths. Martin et al. [248] proposed the use of a GAN to generate scanpaths, while Sui et al. [330] modeled the mechanism of VWM by updating the visual state through a Markov chain. Similar works [191, 363] explored the use of diffusion models for scanpath generation, demonstrating promising results, even if limited to 360° images. To the best of our knowledge, this thesis is the first to investigate the potential of diffusion-based architectures for free-viewing and visual search tasks in natural scenes. A detailed description of the proposed method is introduced in Section 3.2. Following our research work, a related approach was introduced in [196], further highlighting the growing interest in diffusion-based models for gaze behavior modeling on natural scenes.

2.3 BENCHMARKS & EVALUATION METRICS FOR HUMAN ATTENTION MODELS

Standardized benchmarks and evaluation metrics have been established to assess progress in computational models of human visual attention. However, evaluating attention models is inherently challenging due to the stochastic and task-dependent nature of human gaze behavior. This section reviews the most widely adopted benchmarks and metrics for saliency and scanpath prediction.

2.3.1 BENCHMARKS FOR SALIENCY AND SCANPATH PREDICTION

Datasets for studying visual attention typically consist of eye-tracking recordings, where raw gaze data from multiple observers is captured while viewing a stimulus (*e.g.* an image or video). Processed fixations can be used to derive spatial distributions, such as fixation density maps, or to analyze sequential patterns of gaze. As a result, the same dataset can support both saliency and scanpath prediction.

Aggregation into spatial maps is commonly performed by convolving each fixation with a Gaussian kernel, yielding a smooth representation that captures the variability and uncertainty of human visual attention. To support fair and reproducible comparison, a number of datasets have been introduced. However, these benchmarks differ substantially in terms of data collection protocol, type of stimuli, viewing conditions, and task constraints. Below, we summarize the most widely adopted resources in the field:

FIXATION IN FACES (FiFa) [69]. The FiFa database was designed to study human attention on faces and social stimuli in natural settings. It contains 200 images displayed for 2 seconds during the experiment. Data analysis clearly reveals that faces are strong attractors of attention.

MIT1003 [194]. It is one of the first datasets that laid the foundation for saliency evaluation. It contains 1, 003 natural images from Flickr Creative Commons and LabelMe and provides fixation data collected from 15 subjects observing stimuli for 3 seconds, under free-viewing conditions.

NUS EYE FIXATION (NUSEF) [289]. The NUSEF dataset features a diverse collection of images, representing a range of semantic concepts such as faces, nudes, and actions, including objects of varying scale, illumination, and orientation. The collection includes images extracted from Flickr, Photo.net, Google Images, and emotion-evoking IAPS [219]. Each free-viewing experiment lasted 5 seconds, and 75 subjects were involved.

MIT300 [193]. A benchmark for model evaluation rather than training. It contains 300 natural indoor and outdoor images collected from Flickr Creative Commons and personal image collections. Gaze data was recorded from 39 users during free viewing using a table-mounted, video-based eye-tracker. Each image was displayed for 3 seconds. This dataset was introduced as a held-out test set with an online evaluation server. Ground-truth fixations and saliency maps are held out, and models are submitted to the MIT Saliency Benchmark [55], ensuring reproducibility and unbiased comparison across methods.

OSIE [374]. The OSIE dataset comprises 700 images with eye-tracking data of 15 viewers. It was explicitly devised to incorporate high-level semantic attributes. Annotations include segmentation masks and semantic attribute labels for all ob-

jects in the images.

CAT2000 [45]. A large-scale free-viewing dataset that was proposed to explicitly analyze model sensitivity to specific visual factors and solve some dataset-specific biases of previous benchmarks, such as center bias. It features 4,000 images (2,000 for training and 2,000 for testing) evenly distributed across 20 distinct categories, including indoor/outdoor environments, sketches, low-resolution images, and cartoon-style images. The data collection process involved 120 viewers (24 users per image) and each image was shown for 5 seconds.

SALICON [187]. Collecting large-scale eye-tracking data presents several practical challenges. Accurate eye-tracking requires expensive and specialized hardware, while general-purpose webcams are still limited in precision and reliability. Furthermore, recruiting a sufficient number of participants and designing a rigorous experimental setup is both time-consuming and labor-intensive. All these factors naturally limit the size and diversity of available datasets. In parallel, modern deep learning models require thousands, or even tens of thousands, of annotated images to achieve generalization and robust performance. To address these limitations, Jiang et al. [187] proposed using mouse movements as a scalable proxy for human visual attention, enabling the collection of larger datasets. They designed an interface where subjects can move the mouse to direct the high-resolution fovea to regions of interest in the image. By aggregating mouse trajectories across multiple subjects, SALICON approximates human fixations and produces continuous saliency maps. The introduction of SALICON has represented a significant development in data-driven approaches, leading to the emergence of several deep learning-based architectures.

COCO-FREEVIEW [79]. It shares the same natural images as COCO-Search18 [386], but gaze data was recorded under the free-viewing setting. Each image was viewed for 5 seconds by 10 users.

VQA-HAT [107]. It is a task-driven dataset designed for Visual Question Answering (VQA). Its data collection protocol is inspired by SALICON and relies on an interactive, mouse-based annotation procedure. Specifically, participants are presented with a blurred image and are instructed to progressively sharpen regions via mouse clicks that they consider important for answering a given ques-

tion. This process results in spatial attention maps that highlight image regions relevant to the VQA task. VQA-HAT is among the first datasets to enable analysis of whether humans and machine attention mechanisms focus on similar regions when answering visual questions.

COCO-SEARCH18 [386]. COCO-Search18 is the largest dataset for visual search. It comprises 6,202 images, selected from the COCO trainval set [232], and more than 300,000 fixations. Each image is paired with a target object category that participants are instructed to search for within the scene. The dataset includes two visual search scenarios: **target-present (TP)**, in which the target object appears in the scene, and **target-absent (TA)**, where the target object is not present. The images are evenly divided between these two settings, with 3,101 images for each condition. In TP images, only a single instance of the target object appears. In TA images, at least two instances of the target’s semantic siblings (as defined by the COCO taxonomy) are included to act as distractors, thereby increasing task difficulty. COCO-Search18 covers 18 object categories (e.g., keyboard, cup, bowl, ...). Each image is viewed by 10 participants. Participants are instructed to search for the specified target and to terminate the trial as soon as they decide whether the target is present or absent.

AiR-D [74]. Gaze data was recorded from users answering a specific question about the image with an eye-tracking device. Such a protocol enables the study of human attention allocation during visual reasoning tasks as well as the comparison with the attention of artificial models. Overall, AiR-D includes 987 images collected from the GQA dataset [170]. Each visual stimulus is presented for 3 seconds to the user after carefully reading the question.

VQA-MHUG [326]. VQA-MHUG is another dataset designed for VQA. Unlike AiR-D, which records gaze exclusively on images, VQA-MHUG provides synchronized eye-tracking data over both the visual and textual modality. The eye-tracking study resulted in 3,990 question-image pairs and involved 49 participants.

REFCOCO-GAZE [255]. This dataset was designed to study the novel task of *incremental object referral*, which links visual attention dynamics with linguistic input. The goal is to predict a scanpath in which each fixation is temporally

grounded to a specific word or phrase in a referring expression. In particular, participants are instructed to search for a target in the image while simultaneously listening to a spoken referring expression describing that object. As the expression unfolds over time, observers incrementally integrate linguistic and visual information, producing scanpaths that reflect the progressive grounding of language in the visual scene. RefCOCO-Gaze consists of 2, 094 image-expression pairs from the RefCOCO dataset [392], and eye-tracking data was collected from 220 participants. The dataset enables the study of how humans incrementally aggregate visual evidence in response to language, and supports the development and evaluation of multimodal scanpath prediction models.

EYEAR-20K [237]. It is an eye-tracking dataset containing 20k gaze points from 8 subjects while simultaneously looking at an image and hearing an audio clip describing the overall scene instead of a specific target.

2.3.2 EVALUATION OF SALIENCY PREDICTION MODELS

Metrics for saliency prediction can be roughly divided into two main categories: *localization-based* and *distribution-based* metrics. The first category evaluates saliency predictions at discrete fixation locations, assessing whether fixated points receive higher saliency values than non-fixated ones. In contrast, distribution-based metrics treat saliency maps as continuous spatial probability distributions and measure how closely the predicted distribution matches the ground-truth fixation density over the entire image.

LOCALIZATION-BASED METRICS

AREA UNDER ROC CURVE (AUC). It measures how well a predicted saliency map assigns higher values to human fixation locations compared to non-fixated locations. AUC treats the saliency map as a binary classifier distinguishing fixated from non-fixated pixels. By varying a threshold over saliency values, a ROC curve is obtained by plotting the true positive rate (TPR) against the false positive rate (FPR). The resulting AUC score corresponds to the probability that a randomly selected fixated pixel is assigned a higher saliency value than a randomly selected

non-fixated pixel. Several variants of AUC have been proposed, differing in how the true and false positives are computed.

AUC-Judd. For a given threshold, the true positive (TP) rate is the fraction of fixated pixels whose saliency values exceed the threshold, while the false positive (FP) rate is the fraction of non-fixated pixels exceeding the same threshold.

AUC-Borji [46]. This metric is a variant of AUC-Judd that reduces bias from the large number of non-fixated pixels. Instead of using all non-fixated pixels, AUC-Borji randomly samples an equal number of non-fixated locations as negatives and computes the ROC curve by comparing saliency values at fixated versus sampled non-fixated pixels.

SHUFFLED AUC (sAUC). It is designed to account for center bias in human fixations, which can artificially inflate standard AUC scores. Instead of comparing fixated locations to randomly sampled non-fixated pixels from the same image, sAUC samples negative pixels from fixation locations of other images in the dataset. This effectively removes the influence of common biases, such as the tendency to fixate near the center, and measures how well the saliency map predicts fixations beyond these general tendencies.

NORMALIZED SCANPATH SALIENCY (NSS). NSS evaluates how well a saliency map predicts human fixations by directly comparing saliency values at fixation locations. The predicted saliency map is first normalized to zero mean and unit standard deviation, and the saliency values at the human fixation points are sampled. NSS is computed as the average of these normalized saliency values at fixation locations, with higher scores indicating better alignment between predicted saliency and human fixations.

DISTRIBUTION-BASED METRICS

SIMILARITY (SIM). It measures how closely the predicted saliency map matches the ground-truth fixation density map as a probability distribution. Both maps are normalized to sum to 1, and SIM is computed as the sum of the minimum values at each pixel location. The resulting score ranges from 0 to 1, where 1 indicates perfect overlap between predicted and ground-truth distributions, 0 otherwise.

A limitation of this metric is that it is sensitive to the choice of the Gaussian sigma used to construct the ground-truth and saliency maps.

PEARSON’S CORRELATION COEFFICIENT (CC). CC interprets saliency and fixation maps as random variables and measures the linear correlation between them. Value range from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no linear relationship. CC captures linear relationships rather than absolute overlap, making it sensitive to how well the predicted map reflects variations in human fixation densities.

KULLBACK-LEIBLER DIVERGENCE (KL). In general terms, KL quantifies the difference between two probability distributions. In the context of saliency evaluation, it measures the information lost when the predicted saliency map approximates the ground-truth distribution. A lower KL score indicates that the predicted saliency closely matches human fixation patterns, whereas a higher score indicates greater discrepancy.

EARTH MOVER’S DISTANCE (EMD). EMD measures the dissimilarity and the spatial misalignment between the predicted saliency map and the ground-truth fixation density map by computing the minimum cost required to transform one distribution into the other. The saliency value of each pixel is treated as a “mass,” and EMD accounts for both the amount of mass to move and the distance it must be moved. A lower EMD indicates that the predicted saliency closely matches the spatial distribution of human fixations.

INFORMATION GAIN (IG) [216]. IG evaluates how much a predicted saliency map improves over a baseline model (*e.g.* center prior) in explaining human fixations. Both the predicted map and the baseline (commonly the center bias) are treated as probability distributions. IG is computed as the average log-likelihood difference between the predicted saliency map and the baseline at human fixation locations. A higher IG indicates that the model captures fixation patterns beyond what is explained by generic biases, such as the tendency to look near the image center.

2.3.3 EVALUATION OF SCANPATH PREDICTION MODELS

Different from the case of saliency prediction, when assessing the similarity between visual attention scanpaths, accounting for the sequential nature of fixations is crucial. Various metrics have been proposed to compare two scanpaths. Some of them behave like a similarity measure, and some like a dissimilarity measure. We divide scanpath prediction metrics into four main categories: *string-based*, *vector-based*, *curve similarity*, and *time-series* metrics. String-based measures work by quantizing the spatial landscape into different bins and using string alignment metrics. Vector-based methods try to geometrically align scanpaths. Curve similarity and time-series metrics may be more discriminative in revealing differences between viewing patterns or computational models.

STRING-BASED METRICS

LEVENSHTEIN DISTANCE (OR STRING-EDIT DISTANCE) [50, 279]. This metric measures the distance between two variable-length sequences by computing the minimum number of operations required to transform one sequence into the other. The allowed operations are: *insertion*, *deletion*, and *substitution*. In the context of scanpath evaluation, the image is first discretized into a grid, and each cell is assigned a unique symbol. Fixations are then mapped to their corresponding cells, producing symbolic sequences for the ground-truth and predicted scanpaths. The Levenshtein distance quantifies the dissimilarity of the sequences in terms of both the order of fixations and sequence length, penalizing missing, extra, or mismatched fixations. However, the Levenshtein distance has several limitations. Fixations that are spatially close may fall into different bins, while multiple fixations within the same bin are represented with the same symbol, thereby ignoring local differences. Moreover, the metric does not account for fixation duration.

SCANMATCH (SM) [97]. ScanMatch is an enhanced version of the Levenshtein distance that incorporates spatial and temporal information as well as semantic context. It is based on the Needleman–Wunsch algorithm [258] and integrates temporal encoding by repeating symbols according to the duration of each fix-

ation. The similarity between two scanpaths is computed using a substitution matrix that assigns smaller penalties to substitutions between close regions and larger penalties for substitutions between distant regions. This allows ScanMatch to better capture local spatial deviations compared to the standard Levenshtein distance. ScanMatch is still sensitive to the discretization of the Areas of Interest (AOIs).

SEQUENCE SCORE (SS) [47]. SS adopts a clustering-based approach to convert fixations into symbolic sequences. A clustering algorithm (*e.g.*, Mean-Shift) groups together the ground-truth fixations for each visual stimulus, and each cluster is assigned a unique symbol. Each ground-truth scanpath is then transformed into a string by mapping each fixation to its nearest cluster. The same mapping is applied to the predicted scanpath. The similarity between the two resulting sequences is computed using the Needleman–Wunsch algorithm, which aligns the sequences and accounts for insertions, deletions, and substitutions, producing a similarity score that reflects both the spatial correspondence and temporal order of fixations.

SEMANTIC SEQUENCE SCORE (SEMSS) [389]. SemSS was introduced by Yang et al. [389] to evaluate goal-directed attention prediction in a more semantically meaningful way. Unlike the standard Sequence Score (SS), which captures only *where* a person looks, SemSS aims to capture *what* a person is looking at. Instead of mapping fixations to cluster IDs, SemSS converts a scanpath into a sequence of object categories, leveraging the segmentation annotations provided by COCO [232]. This metric assesses whether predicted scanpaths attend to the same semantic objects as human observers, providing a more task-relevant evaluation of goal-directed attention models.

VECTOR-BASED METRICS

MULTIMATCH (MM) [114, 180]. is a multi-dimensional similarity metric for comparing scanpaths, designed to capture not only spatial alignment but also temporal, directional, and amplitude-related properties. Unlike string-based metrics, MultiMatch represents scanpaths as sequences of geometric vectors connecting successive fixations. To reduce complexity, scanpaths are first simplified

by merging small saccades based on their amplitude and angle. The two simplified scanpaths are then temporally aligned. The similarity between the aligned sequences is computed along five complementary dimensions using simple vector-based comparisons:

- **Shape:** The vector difference between consecutive saccade pairs.
- **Length:** The difference in amplitude (length) between the endpoints of saccade vectors.
- **Direction:** The difference in angle between saccade vectors.
- **Position:** The distance between two fixations. Note that a saccade vector connects two fixations.
- **Duration:** The difference of fixation durations between aligned fixations.

Each of these metrics ranges from 0 (no similarity) to 1 (perfect match).

CURVE SIMILARITY METRICS

DISCRETE FRECHET DISTANCE [11, 124]. This metric measures the similarity among curves, taking into account the position and ordering of the fixations. $G = (g_1, g_2, \dots, g_m)$ and $P = (p_1, p_2, \dots, p_n)$ denote the ground-truth and predicted scanpaths, where each element corresponds to a 2D fixation coordinate. The Discrete Fréchet Distance (DFD) computes an optimal alignment between these two sequences under the constraint that the order of fixations is preserved in both scanpaths (*i.e.* alignments are monotonic and do not allow reordering). For any such alignment, the distance between the two scanpaths is defined as the maximum Euclidean distance between all paired fixation points along the alignment. The Discrete Fréchet Distance is then the minimum of this maximum distance over all valid monotonic alignments.

TIME-SERIES ANALYSIS

DYNAMIC TIME WARPING (DTW) [31]. Given two temporal sequences of lengths N and M , the DTW formulation constructs an $N \times M$ cost matrix D , where each entry $D(i, j)$ represents the Euclidean distance between the i -th element of the first sequence and the j -th element of the second. The algorithm

then searches for an optimal warping path from $(1, 1)$ to (N, M) that aligns the two sequences while minimizing the cumulative alignment cost.

TIME DELAY EMBEDDING (TDE) [308]. TDE emphasizes temporal dynamics rather than just the spatial proximity of fixations. It represents each fixation not only by its spatial coordinates but also by a window of preceding fixations, effectively embedding temporal context into each point. Formally, given a scanpath $S = (s_1, s_2, \dots, s_N)$ of N fixations, a time-delay embedding with embedding dimension d constructs vectors of the form:

$$\mathbf{v}_i = (s_i, s_{i-1}, \dots, s_{i-d+1}), \quad i = d, \dots, N.$$

Each vector $\mathbf{v}_i$ represents the current fixation together with the previous $d - 1$ fixations, capturing the local temporal trajectory of the scanpath. The embedded sequences can then be compared using standard distance measures (*e.g.*, Euclidean distance) or sequence alignment techniques to evaluate how similar the temporal patterns of two scanpaths are, rather than just the absolute fixation positions.

3

Predicting Human Gaze Behavior: From Temporal Dynamics to Generation Diversity

UNDERSTANDING and predicting human visual attention remains a central problem in Computer Vision [60, 76, 189, 214, 256], with implications for human-computer interaction [190], autonomous driving [267], and cognitive robotics [280]. While recent deep learning approaches have substantially improved the performance of scanpath prediction in several tasks ranging from free-viewing and visual search to visual question answering, two fundamental challenges remain. First, human gaze behavior is naturally a temporal process, and the decision of where to look next strongly depends on the history of previous fixations. Although state-of-the-art methods accurately predict the spatial location of fixations (*i.e. where* to look), they typically over-

This Chapter discusses topics from the following papers: “A. D’Amelio *et al.*, TPP-Gaze: Modelling Gaze Dynamics in Space and Time with Neural Temporal Point Processes, WACV 2025” [102] and “G. Cartella *et al.*, Modeling Human Gaze Behavior with Diffusion Models for Unified Scanpath Prediction, ICCV 2025” [64].

look the timing of gaze shifts. Existing attempts to model the temporal evolution of scanpaths rely on heavily engineered solutions or weak formulations that do not explicitly model the dependency on previously attended locations. The second challenge concerns the intrinsic variability of visual attention. Different observers, or even the same observer across trials, may exhibit diverse yet equally plausible exploration strategies for the same scene. Nevertheless, the majority of existing scanpath models fail to capture such variability, collapsing to an “average” behavior. This flaw can be attributed to the training paradigm, which commonly relies on pointwise regression objectives that encourage mean predictions at the expense of diversity. In light of these limitations, this chapter addresses both challenges through two complementary contributions that progressively move from structured temporal modeling to a stochastic generative modeling of scanpaths.

The first part of the chapter presents TPP-Gaze, a scanpath prediction architecture grounded in the theory of neural temporal point processes. In this formulation, scanpaths are interpreted as sequences of events governed by learned temporal dynamics and contextual cues. This perspective enables a more mathematically principled representation of fixation timing and ordering. The remainder of the chapter addresses the complementary challenge of variability in human gaze behavior. Rather than producing “averaged” scanpaths, this thesis introduces ScanDiff, a diffusion-based framework that treats scanpath prediction as a stochastic generative process, capable of sampling semantically rich gaze trajectories conditioned on a visual stimulus and task. By leveraging multimodal representation learning, the proposed approach unifies free-viewing and task-driven scanpath prediction within a single architecture, while explicitly modeling uncertainty and inter-subject variability.

Together, TPP-Gaze and ScanDiff advance the field of computational attention modeling with the aim to push scanpath prediction beyond deterministic and averaged solutions, toward models that more faithfully capture the temporal dynamics and diversity of human visual attention.

3.1 MODELING GAZE DYNAMICS IN SPACE AND TIME WITH NEURAL TEMPORAL POINT PROCESSES

Gaze is considered a good measure of overt attention and, more generally, a window to the observer’s thoughts, intentions, and emotions [50]. With roots in psychology and neuroscience, computational models of attention [212] have gained traction in the computer vision and pattern recognition fields since the seminal Itti et al. [177] model. As a matter of fact, the vast majority of works in the field have focused on the computational modeling of saliency maps.

Nevertheless, a growing number of models (*i.e.*, scanpath models) are addressing the prediction of a sequence of fixations. These models explicitly unfold the dynamics of overt attention allocation over a stimulus [36, 213].

It is worth remarking, though, that barely predicting the spatial sequence of fixations does not entail proper modeling of the temporal evolution of attention. By and large, most scanpath models predict an ordered sequence of events while neglecting their continuous timestamp information. As a result, these models are able to tell *where* to look and in what order, but fail in answering *when*. In many respects, this is not an innocent flaw. Human actions often rely on visual information. Therefore, it is essential to direct attention to the right place at the right time [337]. Modeling *when* to perform a saccade translates to devising scanpath models able to predict the sequence of both fixation positions and corresponding durations. Albeit recently few approaches have successfully dealt with such a problem via fully engineered approaches [74, 75, 80, 256, 331], only a marginal subset of them has tackled it in a mathematically principled way [39, 101, 337]. This has typically resulted in a weaker generality of the methods, which are tailored to specific contexts or applications. Under such circumstances, one of the goals of this thesis is to introduce a fresh, general, and simple view on the problem of scanpath modeling.

With this aim, we present TPP-Gaze, a novel and principled approach to model scanpath dynamics grounded on the theory of *Neural Temporal Point Processes* (*Neural TPPs*), that jointly learns the temporal dynamics of fixation position and duration, integrating deep learning methodologies with point process

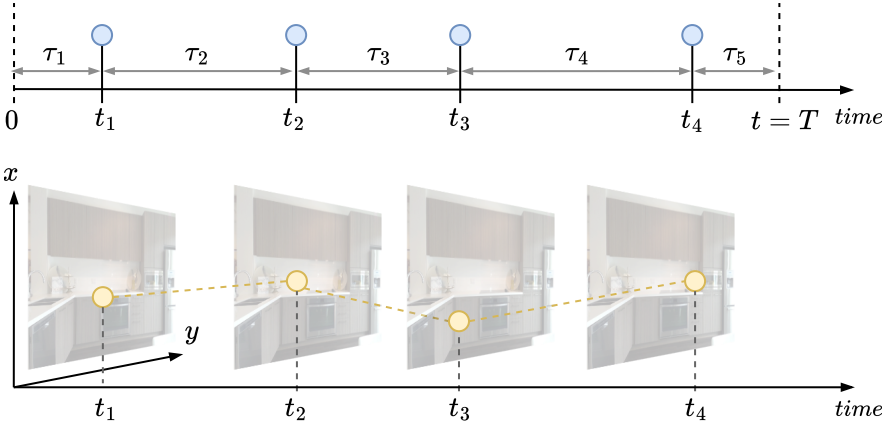


Figure 3.1: Scanpath dynamics as a marked TPP. Time is represented on the horizontal axis, and different scanpath fixations occur at times t_1, t_2, t_3 and t_4 .

theory. Specifically, we consider a scanpath as the realization of a point process in space and time. Remarkably, the modeling assumptions of Neural TPPs align perfectly with the structure of scanpath data. A scanpath consists of a series of events (saccades) occurring at irregular intervals (fixation durations), which is *exactly* what Neural TPPs are designed to model. While the psychological and neuroscience literature has used traditional point processes for eye movement *analysis* [28, 126, 391], these tools are not well-suited for scanpath *prediction* due to their inability to handle stimuli. In other words, traditional TPPs are effective for studying the observer but fall short when addressing Computer Vision tasks related to attention allocation prediction. In contrast, Neural TPP-based models offer the best of both worlds: they combine the robust theoretical framework of TPPs with the flexibility and power of modern neural networks. Nevertheless, in this thesis, we present the first attempt to adopt them for the problem of scanpath modeling.

Our key contributions can be summarized as follows: 1) We propose a novel scanpath model able to jointly learn the temporal dynamics of both fixation position and duration. 2) We extend recent Neural TPP models to deal with visual data (*i.e.*, images) and connect scanpath modeling and prediction to point process theory. To assess our proposal, which can be appreciated at a glance

in Fig. 3.1, we conduct experiments on five publicly available datasets, showing an overall superior performance of the proposed model when compared to state-of-the-art approaches. Source code and trained models are publicly available*.

3.1.1 TEMPORAL POINT PROCESS MODELS FOR HUMAN GAZE DYNAMICS

Real-world event streams, such as earthquake occurrences, trades in a financial market, user engagements in recommendation systems, or web-based activity, are rarely periodic. Instead, they emerge from complex continuous-time processes that yield events at irregular intervals. Modeling these streams requires more than simple time-series analysis based on regularly sampled observations. It demands a framework capable of jointly capturing (i) the timing of the next arrival, (ii) what kind of event it will be (its mark or type), and (iii) how the history of past events influences the system’s future dynamics. Temporal Point Processes (TPPs) provide a principled probabilistic framework in this direction. They describe the evolution of a system through a history-dependent intensity function that represents the instantaneous rate of events over time. This perspective supports descriptive analysis (*e.g.*, detecting inhibition or periodicity) and enables predictive modeling (*e.g.*, forecasting future events and quantifying uncertainty). As a result, TPPs have become a key tool for modeling asynchronous data across the natural and social sciences, finance, and modern online platforms.

BACKGROUND ON TEMPORAL POINT PROCESSES. Temporal Point Processes (TPPs) are probabilistic generative models designed for continuous-time event sequences. TPPs natively model the timing at which events occur, making them particularly suitable for irregular, asynchronous data such as user interactions or system logs.

Formally, a Temporal Point Process is a random process whose realization consists of a sequence of discrete events with strictly increasing arrival times $\mathcal{T} = \{t_1, t_2, \dots, t_N\}$, where $t_i \in \mathbb{R}^+$, $t_1 < t_2 < \dots < t_N$ and N , the number of events, is itself a random variable. The history of all past events up to time

*<https://github.com/phuselab/tppgaze>

t is defined as $\mathcal{H}_t = \{t_n \in \mathcal{T} : t_n < t\}$. In more traditional approaches, the dependency of the next arrival time t on the history $\mathcal{H}_t$ is modeled through the conditional intensity function $\lambda^*(t) = \lambda(t \mid \mathcal{H}_t)$ whose functional form determines the properties of the TPP. Instead of directly modeling the next arrival time t_i , common practice is to represent the TPP as a sequence of strictly positive inter-event times $\tau_i = t_i - t_{i-1} \in \mathbb{R}^+$. Note that the representations in terms of t_i and τ_i are isomorphic. Knowing the conditional intensity function allows to recover the conditional probability of the time τ_i until the next event as:

$$p^*(\tau_i) = p(\tau_i \mid \mathcal{H}_{t_i}) = \lambda^*(t_{i-1} + \tau_i) \exp \left(- \int_0^{\tau_i} \lambda^*(t_{i-1} + s) ds \right). \quad (3.1)$$

For instance, under the assumptions of no dependence on the history and constancy over time (*i.e.*, $\lambda^*(t) = k$, with $k \geq 0$), the homogeneous Poisson process is recovered, with inter-event times distributed according to the exponential distribution. Choosing more complex functional forms for $\lambda^*(t)$ allows to recover many well-known TPPs such as Hawkes or self-correcting processes [155, 173].

MARKED TPPs. The basic mathematical formalism of TPPs naturally handles the dynamics of arrival times of events. However, the distribution of time until the next event might depend on factors other than the history. Event data is often accompanied by some kind of covariate indicating the nature of the specific event being predicted. In the context of TPPs, such covariates are called *marks*. Formally, a marked TPP is a random process whose realization consists of a sequence of discrete events localized in time $\mathcal{S} = \{(t_1, m_1), \dots, (t_N, m_N)\}$ with the timing $t_i \in \mathbb{R}^+$ and the mark $m_i \in \mathcal{M}$. The mark m_i is typically modeled as an integer representing the type of event, but other kinds of marks (*e.g.*, $\mathcal{M} = \mathbb{R}^d$) can eventually be adopted. For a given $\mathcal{S}$, the history of past events at time t is denoted as $\mathcal{H}_t = \{(t_i, m_i) : t_i < t\}$. Specifying a marked-TPP involves the definition of the joint conditional density function of the next event, with inter-event time τ_i and mark m_i , given the history of past events: $p^*(\tau_i, m_i) = p(\tau_i, m_i \mid \mathcal{H}_{t_i})$.

NEURAL TPPs. Designing TPPs typically involves trade-offs along three dimensions: *flexibility* (*i.e.*, can the model approximate any distribution?), *efficiency* (*i.e.*, can the likelihood function be evaluated in closed form?), and *ease of use*

(i.e., is sampling straightforward?). Classical TPPs, such as Poisson, Hawkes, and self-correcting processes, rely on hand-crafted parametric forms for the intensity function. While these models provide analytical tractability and interpretable dynamics, restricting $\lambda^*(t)$ to a specific parameterization limits the general applicability of such models. Indeed, they impose strong assumptions on the underlying event-generating process, thus limiting the ability to model complex, high-dimensional, or nonlinear real-world sequences, as well as long-term dependencies between different events. To overcome these limitations, the most recent solutions have resorted to neural approaches, called *Neural TPPs*. Neural TPPs [251, 314, 403, 417] integrate key concepts from point process literature with deep learning methodologies and implement learnable parametric forms of the intensity function, $\lambda_\theta^*(t)$ [119, 166].

Such an approach provides increased flexibility to capture complex temporal dependencies, incorporate high-dimensional event features, and model nonlinear interactions, all while retaining the probabilistic foundation of classical TPPs.

By assuming a conditional distribution parameterized by the weights of a neural model, $p_\theta^*(\tau_i, m_i)$, inference can be performed by maximizing the joint likelihood of the N observed events in a sequence:

$$\theta^* = \operatorname{argmax}_{\theta} \prod_{i=0}^N p_\theta(\tau_i, m_i \mid \mathcal{H}_t) = \prod_{i=0}^N p_\theta^*(\tau_i, m_i). \quad (3.2)$$

Early Neural TPPs, such as the Neural Hawkes Process [251], used RNNs to model the intensity function. More recently, self-attention mechanisms have been employed for the same purpose [403, 417]. The main challenge of such intensity-based approaches lies in choosing a good parametric form for $\lambda_\theta^*(t)$.

The choice of the parametric form for the intensity function has to take into account the necessity of a closed-form solution of the integral in Eq. (3.1), thus practically restricting the expressiveness of the model. More complex parametric forms would require Monte Carlo approximation of the integral [251], or the sampling is expensive [265]. To mitigate this issue, Shchur et al. [314] proposed to directly learn the parametric conditional distribution $p_\theta^*(\tau)$ of the inter-arrival

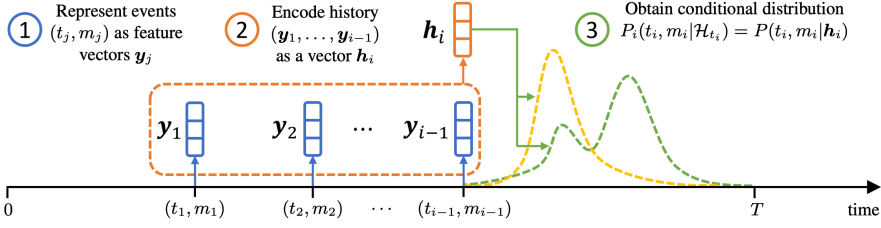


Figure 3.2: Representation of an autoregressive neural TPP model. Figure from [315].

times rather than the conditional intensity function $\lambda_{\theta}^*(t)$, thus recasting learning Neural TPPs as a density estimation problem.

Following Du et al. [119], Neural TPPs can be defined as autoregressive models that sequentially predict the time and mark of the next event (t_i, m_i) . This process can be broadly decomposed into three main stages, as illustrated in Fig. 3.2:

1. **Representing an Event:** Each event (t_i, m_i) is first encoded as a feature vector $\mathbf{y}_i$, obtained by combining temporal and mark-related information. The temporal component can be derived from the inter-event time τ_i (or its logarithm $\log \tau_i$) [119, 265], or from trigonometric functions, inspired by the positional encodings used in Transformer-based models [403, 417]. The mark component can take various forms, including categorical labels, scalar values, or more generally d -dimensional real-valued feature vectors.
2. **Encoding the History of Events:** The core idea is to summarize the event history $\mathcal{H}_{t_i} = \{\mathbf{y}_1, \dots, \mathbf{y}_{i-1}\}$, which is inherently a variable-length sequence, into a fixed-dimensional representation $\mathbf{h}_i$. In practice, this is achieved by encoding the past events $(\mathbf{y}_1, \dots, \mathbf{y}_{i-1})$ into a history embedding $\mathbf{h}_i$. One way to encode the history is through an update function f , implemented as a recurrent encoder such as RNN, LSTM, or GRU. Starting with an initial hidden state $\mathbf{h}_1$, after each event, the hidden state is updated as $\mathbf{h}_{i+1} = f(\mathbf{h}_i, \mathbf{y}_i)$. In this manner, it is possible to compute the history embedding for all N events of the sequence. Another possibility is to compute the embedding $\mathbf{h}_i$ directly from the set of past event representations $(\mathbf{y}_1, \dots, \mathbf{y}_{i-1})$. In this case, the encoding is performed

independently for each event i , for example, by using the self-attention mechanism[403, 417].

3. **Predicting the Arrival Time of the Next Event:** Neural TPPs model the conditional distribution of the next inter-event time τ_i (or equivalently the next arrival time t_i) given the history $\mathcal{H}_{t_i}$, denoted by $P_i^*(\tau_i)$. This distribution can be specified through equivalent representations, such as the probability density, cumulative distribution, survival function, hazard function, or cumulative hazard function. In neural TPPs, a parametric form is chosen for one of these functions, and its parameters are predicted from the history embedding $\mathbf{h}_i$ via a neural network. It is important to ensure that the chosen parametrization defines a valid probability distribution. Negative log-likelihood (NLL) is the default training objective for both neural and classical TPP models. However, other objectives based on adversarial losses [380] and inverse reinforcement learning [224] have been proposed.

Applications of Neural TPPs span a variety of research fields [315], such as healthcare [128], finance [21], social network analysis [347], earthquake forecasting [51], and recommendation systems [211]. They have also been employed for the spatio-temporal analysis of eye movement sequences [391]. In this work, we make a step further and leverage the Neural TPP framework to directly model attention dynamics on visual data.

3.1.2 THE TPP-GAZE MODEL

Modeling scanpaths involves defining a mapping from visual data, $\mathbf{I}$ (raw image data representing either a static picture or a stream of images), to a sequence of time-stamped gaze locations $\mathcal{S} = \{(\mathbf{r}_{F_1}, t_1), (\mathbf{r}_{F_2}, t_2), \dots, (\mathbf{r}_{F_N}, t_N)\}$. Here $\mathbf{r}_{F_n} \in \mathbb{R}^2$ represents the two-dimensional vector of spatial coordinates of the n -th fixation on the stimulus $\mathbf{I}$, while $t_n \in \mathbb{R}^+$ represents its arrival time. Eventually, a perceptual representation of the input stimuli, $\mathcal{Z}$, is computed, with the

aim of locating the relevant objects inside the scene:

$$\mathbf{I} \rightarrow \mathcal{Z} \rightarrow \{(\mathbf{r}_{F_1}, t_1), (\mathbf{r}_{F_2}, t_2), \dots, (\mathbf{r}_{F_N}, t_N)\}. \quad (3.3)$$

3

Here we assume that no specific external task or goal is given to the observer (*i.e.*, free-viewing condition). The dynamics of the attentive process, which unrolls as a sequence of fixation locations with corresponding duration/arrival time, is characterized by an inherent randomness which likely stems from internal stochastic fluctuations affecting sensory and information processing, movement planning, and execution [348], in both fixation locations and corresponding durations. Many scanpath models proposed in the recent literature [15, 16, 214, 330] get rid of fixations' timestamp information by rearranging the sequence $\{(\mathbf{r}_{F_1}, t_1), \dots, (\mathbf{r}_{F_N}, t_N)\}$ as $\{\mathbf{r}_F(1), \dots, \mathbf{r}_F(N)\}$, assuming $(\mathbf{r}_{F_n}, t_n) = \mathbf{r}_F(n)$.

Several approaches [39, 101, 337] have dealt with this problem, relying on specific theoretical frameworks. Tatler et al. [337] modeled saccade timings as an evidence accumulation process with clear neurobiological significance. Similarly, in [101], a Langevin-type SDE race model [42] was adopted to predict fixations and their durations in socially relevant contexts, while in [39], fixation duration was equated to the patch residence time of a forager searching for nourishment. Conversely, the vast majority of recent methods [74, 75, 80, 256, 331] simply model fixation duration by employing specific neural architectural choices that aim at associating each fixation with its corresponding duration.

In a different vein, this work recasts the whole visual attention allocation process in the mathematical framework of point process theory [99]. This emphasizes the central role of visual attention's spatio-temporal dynamics by explicitly representing scanpaths as sequences of discrete events happening at irregular intervals. Specifically, we conceive a scanpath as realization of a random process whose events happen at strictly increasing arrival times $\mathcal{T} = \{t_1, \dots, t_N\}$. Fixation duration can be recovered by resorting to inter-event times $\tau_n = t_n - t_{n-1}$, while their locations can be represented as the two-dimensional continuous mark associated with the n -th event. Under this assumption, Neural Temporal Point Processes represent the natural choice for modeling this kind of data.

Given a stimulus (image) $\mathbf{I}_j$, an ensemble of N_{obs} observers performs a sequence of fixations and saccades (scanpath) on it, thus obtaining a set of sequences $\mathcal{C}_j = \{S^1, \dots, S^{N_{obs}}\}$. Each scanpath S^i is a sequence of pairs (events) $S_n^i = (\mathbf{r}_{F_n}, t_n)$ each composed by a fixation position (marker) $\mathbf{r}_{F_n} \in \mathbb{R}^2$, and a corresponding arrival time $t_n \in \mathbb{R}^+$. At the most general level, we are interested in modeling the stochastic generative process that, given a semantic representation of the image $\mathcal{Z}_j$ and the history of past events $\mathcal{H}_t$, simulates the next fixation position and duration. More formally:

$$S_{n+1}^i \sim p_\theta(\mathbf{r}_{F_{n+1}}, t_{n+1} \mid \mathcal{H}_t, \mathcal{Z}_j), \quad (3.4)$$

where $p_\theta(\cdot)$ represents the parametric joint conditional distribution of a Neural TPP [314].

In the following, we present the architecture of TPP-Gaze, implementing a scanpath model on an image as a Neural TPP. A qualitative overview of the proposed architecture is depicted in Fig. 3.3.

REPRESENTING SCENE SEMANTICS. As outlined in Eq. (3.3), the sequence of events composing a scanpath depends not only on the history of past events, but on a perceptual representation of the input stimulus $\mathbf{I}_j$, encoding scene semantics and relevant object locations. We extract the perceptual representation of the input image via a CNN architecture inspired by [214]. Specifically, the input image is first processed by a pre-trained DenseNet201 CNN [165]. Activation maps from various convolutional layers (as reported in [214]) are extracted, thus obtaining a 2, 048 channels volume, each representing the location of semantic features inside the scene. It is worth noticing that learning to predict fixation locations (*i.e.*, marks) involves a mapping between coordinates in Cartesian space, a task in which standard convolutions have been reported to fail [235]. Following [248, 330], we adopt a CoordConv layer [235] to give convolutions access to their own input coordinates. This results in a 2, 051 channels volume which is fed as input to 3 layers of 1×1 convolutions, followed by a linear layer mapping to $\mathbf{z}_j$ acting as our semantic representation.

REPRESENTING HISTORY. Neural TPPs employ either Recurrent Neural Net-

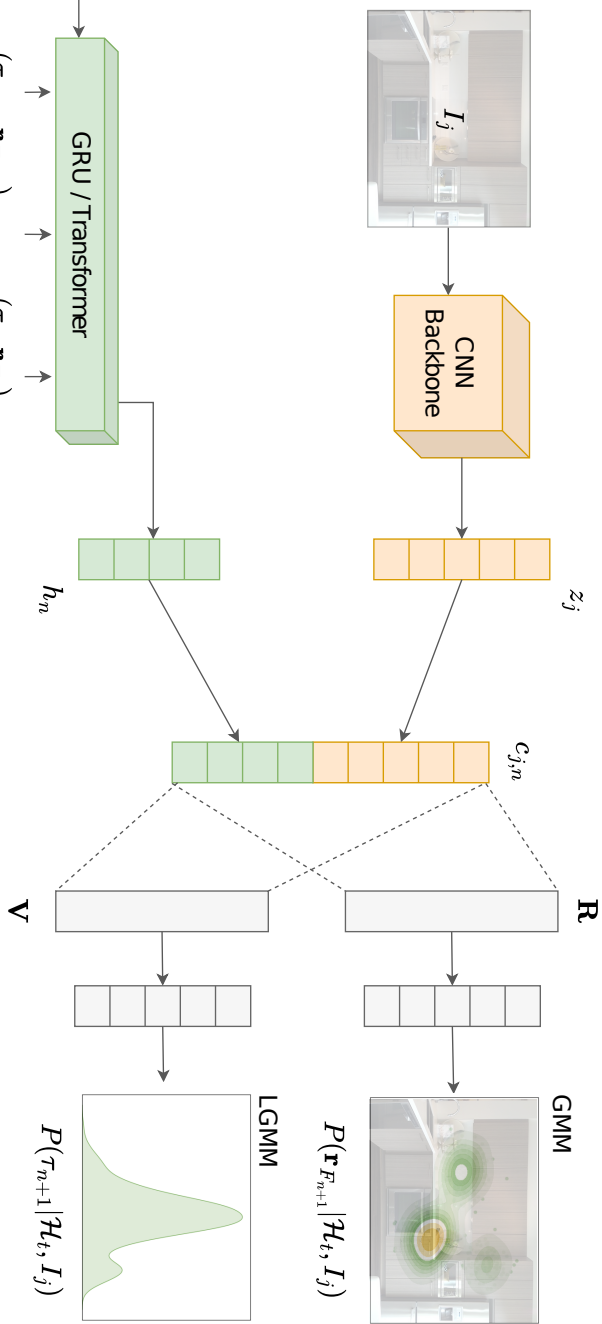


Figure 3.3: Overview of TPP-Gaze model architecture. Given a semantic representation of the image (z_j) and the history of past events (h_n), the next fixation position and duration are simulated.

works (RNNs) and their variants (*e.g.*, LSTM, GRU) [119, 314, 372] or Transformer encoders [403, 417] to model the nonlinear dependency over both the markers and the timings from past events [315]. As shown in Fig. 3.3, the pair $(\mathbf{r}_{F_n}, \tau_n)$ representing the event occurring at the time t_n with fixation position $\mathbf{r}_{F_n}$ and duration $\tau_n = t_n - t_{n-1}$, is fed as the input into either a GRU or a Transformer encoder as described in [417]. The Transformer/GRU state embedding $\mathbf{h}_n$ represents the influence of the history up to the n -th fixation. Hence, it can be employed as a vector space representation of $\mathcal{H}_{t_n}$. Taking into account the semantic representation $\mathbf{z}_j$ and the history embedding $\mathbf{h}_n$, Eq. (3.4) can be rewritten as:

$$\mathcal{S}_{n+1}^i \sim p_\theta(\mathbf{r}_{F_{n+1}}, t_{n+1} \mid \mathbf{h}_n, \mathbf{z}_j). \quad (3.5)$$

FIXATION DURATION GENERATION. We model the conditional dependence of the distribution $p_\theta(\tau_{n+1} \mid \mathbf{h}_n, \mathbf{z}_j)$ on both past events and stimulus by concatenating the history embedding and semantic vectors into a context vector $\mathbf{c}_{j,n} = [\mathbf{h}_n \parallel \mathbf{z}_j]$. In the vein of [314], the latter is employed to learn the parameters of a Log-Gaussian Mixture Model (LGMM) via an affine transform:

$$\mathbf{w} = \text{softmax}(\mathbf{V}_w \mathbf{c}_{j,n}), \quad \mathbf{s} = \exp(\mathbf{V}_s \mathbf{c}_{j,n}), \quad \mathbf{m} = \mathbf{V}_m \mathbf{c}_{j,n} \quad (3.6)$$

where $\mathbf{w} \in \mathbb{R}_+^K$ are the mixture weights, $\mathbf{m} \in \mathbb{R}^K$ are the mixture means, and $\mathbf{s} \in \mathbb{R}_+^K$ are the standard deviations. K represents the number of mixture components. The fixation duration for the n -th event can be generated by sampling from the LGMM defined by:

$$p_\theta^*(\tau_n \mid \mathbf{c}_{j,n}) = p(\tau_n \mid \mathbf{w}, \mathbf{m}, \mathbf{s}) = \sum_{k=1}^K w_k \frac{1}{\tau_n s_k \sqrt{2\pi}} \exp\left(-\frac{(\log \tau_n - m_k)^2}{2s_k^2}\right). \quad (3.7)$$

FIXATION POSITION (MARK) GENERATION. Similarly, given the context vector $\mathbf{c}_{j,n}$, we define the conditional probability of the next mark (fixation position), $p_\theta(\mathbf{r}_{F_{n+1}} \mid \mathbf{h}_n, \mathbf{z}_j)$, as a 2D Gaussian Mixture Model (GMM) whose parameters

are obtained via another affine projection:

$$\omega_g = \text{softmax}(\mathbf{R}_\omega^g \mathbf{c}_{j,n}), \quad \Sigma_g = \text{diag}(\exp(\mathbf{R}_\Sigma^g \mathbf{c}_{j,n})), \quad \mu_g = \mathbf{R}_\mu^g \mathbf{c}_{j,n} \quad (3.8)$$

where $\omega_g \in \mathbb{R}_+^2$ are the mixture weights, $\mu_g \in \mathbb{R}^2$ are the mixture means, and $\Sigma_g \in \mathbb{R}^{2 \times 2}$ are the diagonal covariance matrices of G bi-variate Gaussian distributions. The x and y coordinates of the n -th fixation can be generated by sampling from the GMM defined by:

$$\begin{aligned} p_\theta^*(\mathbf{r}_{F_n} \mid \mathbf{c}_{j,n}) &= p(\mathbf{r}_{F_n} \mid \omega, \mu, \Sigma) \\ &= \sum_{g=1}^G \omega_g \frac{\exp\left(-\frac{1}{2}(\mathbf{r}_{F_n} - \mu_g)^T \Sigma_g^{-1} (\mathbf{r}_{F_n} - \mu_g)\right)}{\sqrt{(2\pi)^2 \mid \Sigma_g \mid}}. \end{aligned} \quad (3.9)$$

3.1.3 MODEL INFERENCE

Consider a set of stimuli $\mathcal{I} = \{\mathbf{I}_1, \dots, \mathbf{I}_j, \dots, \mathbf{I}_J\}$ each gazed by N_{obs} human observers. Each observer produces an ensemble of scanpaths $\mathcal{C}_j = \{S^1, \dots, S^{N_{obs}}\}$ with $S_n^i = (\mathbf{r}_{F_n}^i, \tau_n^i)$ representing an event (*i.e.*, fixation position and duration). Model inference is performed by minimizing a negative log-likelihood loss with respect to the parameters of the semantic network, the GRU/Transformer encoding history of events, and the affine transforms of the LGMM and GMM. Formally, the loss function is defined as follows:

$$\mathcal{L}(\theta) = - \sum_j \sum_i \sum_n [\log p_\theta^*(\tau_n^i \mid \mathbf{c}_{j,n}) + \log p_\theta^*(\mathbf{r}_{F_n}^i \mid \mathbf{c}_{j,n})]. \quad (3.10)$$

3.1.4 EXPERIMENTAL EVALUATION

DATASETS. Regarding the stimuli and eye tracking data, we select five publicly available datasets of human-recorded scanpaths comprising both fixation positions and durations: FiFa [69], NUSEF [289], OSIE [374], MIT1003 [194], and COCO-FreeView [79].

IMPLEMENTATION & TRAINING DETAILS. COCO-FreeView and MIT1003 datasets are used for model training. To this end, 70% of the images from both datasets are used for training, while the remaining 30% is equally partitioned between validation and test sets. We use AdamW as optimizer, with weight decay set to 10^{-1} , and the learning rate set to 10^{-3} . Batch size is equal to 128. We employ early stopping after 20 epochs with no improvement on the validation set. Following previous literature [80, 214], during training and evaluation, we discard the first fixation and remove all scanpaths containing fewer than four fixations.

SCANPATH EVALUATION METRICS. We employ the MultiMatch [114, 180], ScanMatch [97], and Sequence Score [47] evaluation metrics since they explicitly consider fixation duration in the evaluation process. In addition, the String Edit Distance [50] is adopted to further evaluate predicted scanpaths.

In our experiments, for ScanMatch, the longest dimension of the stimuli is divided into 14 bins, while the shortest is split into 8 bins. The temporal bin size is set to 50 ms for scanpath models delivering fixation duration estimates.

EVALUATION PROTOCOL. We compare the scanpaths generated by different models with those recorded from human observers. The objective is to evaluate whether the simulated behaviors exhibit the statistical properties observed in humans during eye-tracking experiments while viewing the same stimulus.

The evaluation protocol unfolds as follows. Suppose there are N_{obs} human observers. For each stimulus, we first compute the evaluation scores for every possible pair of the N_{obs} observers (Real vs. Real). Then, for each model, (i) we generate gaze trajectories from artificial observers and (ii) calculate the evaluation scores for every possible pair of real and artificial scanpaths (Real vs. Simulated).

For a given metric, this procedure yields a target distribution P of similarity scores between observers (Real vs. Real) and a distribution Q of similarity scores for the given model w.r.t. humans (Real vs. Simulated). As reported in [213], MM, SM, and SS average values may deliver inconsistent results: models exhibiting less variability w.r.t. humans, can score systematically better than the ground truth model. This issue can be tackled by considering a good model as the one that minimizes the discrepancy between the target and model-derived score distributions. We quantified such discrepancy using the Kullback-Leibler Divergence

	Dim		GMM		MM (KL-Div) ↓		SM (KL-Div) ↓		SED ↓
CNN	Img	TPP	K	G	Dur	Avg	w/ Dur	w/o Dur	Avg
<i>Image Backbone</i>									
RN	256	256	4	16	0.011	0.037	0.113	0.101	17.575
DN	256	256	4	16	0.012	0.028	0.078	0.060	17.032
<i>Image and TPP Dimensionalities</i>									
DN	128	128	4	16	0.010	0.031	0.094	0.069	16.959
DN	128	256	4	16	0.012	0.030	0.084	0.063	16.887
DN	256	128	4	16	0.009	0.037	0.105	0.082	17.413
DN	256	256	4	16	0.012	0.028	0.078	0.060	17.032
DN	256	512	4	16	0.010	0.031	0.101	0.095	17.462
DN	512	256	4	16	0.008	0.032	0.110	0.093	17.497
DN	512	512	4	16	0.009	0.027	0.104	0.077	17.154
<i>Mixture Components</i>									
DN	256	256	2	16	0.014	0.027	0.092	0.071	16.944
DN	256	256	4	16	0.012	0.028	0.078	0.060	17.032
DN	256	256	2	32	0.009	0.030	0.109	0.098	17.216
DN	256	256	4	32	0.009	0.031	0.103	0.076	17.252

Table 3.1: Ablation study results comparing different model configurations and hyperparameters. We report the results for ResNet50 (RN) and DenseNet201 (DN) visual backbones, various embedding vector dimensions for the image representation and the TPP history, and different numbers of Gaussian mixture components.

(KL-Div): $D_{KL}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log(P(x)/Q(x))$. Conversely, as SED is an evaluation metric not requiring alignment, it is not susceptible to the inconsistency issues associated with MM, SM, and SS. Consequently, its values are directly reported without any further processing.

3.1.5 SCANPATH PREDICTION

ABLATION STUDIES. The TPP-Gaze architecture consists of three main blocks: image encoding (CNN backbone), history encoding (RNN/Transformer), and fixation/inter-time prediction (GMM/LGMM). To break down these components and make the adopted design choices explicit, we perform extensive ablation studies. Specifically, we evaluate two different CNN backbones for image encoding (*i.e.*, a ResNet50 [157] and a DenseNet201 [165]) as well as three embedding vector dimensions for the image semantic representation ($\mathbf{z}_j$) and the history embedding ($\mathbf{h}_n$, TPP dimensionality). Moreover, different numbers of

components for the GMM/LGMM are considered. Table 3.1 reports the results of the ablation studies conducted on the COCO-FreeView dataset. In our experiments, we select the hyperparameters yielding the best trade-off according to the considered evaluation metrics, resulting in a DesNet201 backbone and a dimensionality equal to 256 for all embedding vectors. Moreover, the parameters K and G representing mixture components are respectively set to 4 and 16.

COMPARISON WITH THE STATE OF THE ART. We compare the proposed model with state-of-the-art approaches that reach high performance in recent scanpath benchmarks [212], offer source code availability, and are representative of different architectural solutions. As to the latter criteria, following the taxonomy proposed in [212], scanpath models can be aggregated into the following categories: biologically inspired (*e.g.* Itti-Koch model [177] and G-Eymol [394]); statistically inspired (*e.g.* CLE model [41]); cognitively inspired (*e.g.* IOR-ROI-LSTM [80]); engineered models (*e.g.* DeepGazeIII [214] and Scanpath-VQA [74]); but see [212–214] for an in-depth review. Under such circumstances, we assess the performance of TPP-Gaze against the aforementioned models.

Table 3.2 reports quantitative results on the COCO-FreeView and MIT1003 datasets in terms of all considered metrics, while model performance on OSIE, NUSEF, and FiFa is reported in Table 3.3 in terms of MM and SM and in Table 3.4 in terms of SS and SED. In all experiments, we compare the aforementioned approaches using the pre-trained model weights released by the authors. As DeepGaze models were trained on the entire MIT1003 dataset, the results from DeepGazeIII and CLE (DG) are not included in this comparison. Additionally, to explicitly measure the effect of the proposed architecture and mathematical framework, we retrain and test the two most recent models (DeepGazeIII and Scanpath-VQA) under the same conditions adopted for TPP-Gaze (see Section 3.1.4). Specifically, beyond training on the same data, the large-scale pre-training of DeepGazeIII, as well as the fine-tuning stage of Scanpath-VQA based on reinforcement learning [294], have been inhibited. These results are reported in gray color at the bottom of the tables.

When trained under the same settings and datasets, TPP-Gaze (with either

	COCO-FreeView											MIT1003										
	MM (KL-Div) ↓					SM (KL-Div) ↓					SS (KL-Div) ↓					SED ↓						
	Sh	Len	Dir	Pos	Dur	Avg	w/ Dur	w/o Dur	Dur		Avg	Sh	Len	Dir	Pos	Dur	Avg	w/ Dur	w/o Dur	Dur		Avg
Int-Koch [177]	0.42	0.40	0.21	1.02	-	0.51	-	2.54	-	1.01	14.00	0.91	0.64	0.71	1.53	-	0.95	-	2.27	-	6.30	8.86
CLE (tst) [41, 177]	0.07	0.30	0.35	1.43	-	0.54	-	2.50	-	1.27	14.37	0.10	0.10	0.32	1.04	-	0.39	-	2.16	-	6.25	9.09
CLE (DG) [41, 215]	0.06	0.18	0.23	1.31	-	0.44	-	2.37	-	1.22	14.31	-	-	-	-	-	-	-	-	-	-	
G-Eymol [394]	0.37	0.73	0.93	1.22	1.99	1.05	9.00	6.67	8.75	6.30	14.20	0.68	0.68	0.46	1.54	1.03	0.88	15.90	4.89	3.32	6.96	
IOR-ROI-LSTM [80]	1.15	0.47	0.03	0.19	0.05	0.38	1.54	0.76	0.56	0.64	13.55	0.59	0.27	0.07	0.57	0.05	0.31	0.69	0.45	5.08	8.61	
DeepGazeIII [214]	0.04	0.02	0.03	0.03	-	0.03	-	0.33	-	0.33	13.15	-	-	-	-	-	-	-	-	-	-	
Scapath-VQA [74]	0.05	0.16	0.10	0.06	0.25	0.12	1.07	0.34	0.43	0.28	12.76	0.04	0.05	0.08	0.05	0.14	0.07	0.06	0.05	0.05	0.11	
DeepGazeIII [214]	0.01	0.03	0.05	0.05	-	0.04	-	0.34	-	0.36	13.15	0.05	0.01	0.20	0.05	-	0.08	-	0.19	-	5.06	8.28
Scapath-VQA [74]	0.62	0.41	0.02	0.05	0.03	0.23	0.08	0.03	0.03	0.31	14.34	0.20	0.14	0.15	0.08	0.02	0.12	0.23	0.19	0.14	0.25	
TPP-Gaze (GRU)	0.06	0.02	0.02	0.03	0.01	0.03	0.08	0.06	0.05	0.11	17.03	0.01	0.03	0.09	0.04	0.01	0.04	0.15	0.11	0.12	0.11	
TPP-Gaze (Transformer)	0.05	0.01	0.02	0.03	0.01	0.03	0.10	0.07	0.06	0.12	16.93	0.01	0.02	0.09	0.07	0.02	0.04	0.22	0.16	0.14	0.14	

Table 3.2: Comparison of various models on COCO-FreeView and MIT1003. **Gray color** indicates models trained under the same settings and datasets. Within this group, **bold** values represent the best performance for each metric. Underline values indicate the overall best performance across all models and metrics.

	OSIE										NUSEF										FiFa										
	MM (KL-Div) ↓					SM (KL-Div) ↓					MM (KL-Div) ↓					SM (KL-Div) ↓					MM (KL-Div) ↓					SM (KL-Div) ↓					
	Sh	Len	Dir	Pos	Dur	Avg	Sh	Len	Dir	Pos	Dur	Avg	Sh	Len	Dir	Pos	Dur	Avg	Sh	Len	Dir	Pos	Dur	Avg	Sh	Len	Dir	Pos	Dur	Avg	
Itri-Koch [177]	1.62	0.89	0.45	3.69	-	1.66	-	2.22	-	2.22	-	0.45	-	0.61	-	0.61	-	0.61	-	0.61	-	0.61	-	0.61	-	0.61	-	0.61	-	0.61	-
CLE (ltri) [41, 177]	0.13	0.03	0.20	0.75	-	0.28	-	1.98	-	1.98	-	0.20	-	0.79	-	0.79	-	0.79	-	0.79	-	0.79	-	0.79	-	0.79	-	0.79	-	0.79	-
CLE (DG) [41, 215]	0.17	0.03	0.15	0.60	-	0.24	-	1.43	-	1.43	-	0.15	-	0.50	-	0.50	-	0.50	-	0.50	-	0.50	-	0.50	-	0.50	-	0.50	-	0.50	-
G-Eymol [394]	1.18	1.08	0.25	2.12	1.18	1.16	16.17	7.29	0.38	0.30	0.05	0.29	3.02	0.81	1.76	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	
IOR-ROI-LSTM [80]	1.72	0.73	0.03	0.96	0.03	0.69	0.75	0.76	0.90	0.36	0.12	0.23	0.17	0.36	0.11	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	
DeepGazeIII [214]	0.14	0.08	0.06	0.15	-	0.11	-	0.12	-	0.12	-	0.07	-	0.07	-	0.07	-	0.07	-	0.07	-	0.07	-	0.07	-	0.07	-	0.07	-	0.07	-
Scanpath-VQA [74]	0.07	0.07	0.04	0.04	0.16	0.08	0.03	0.03	0.11	0.04	0.02	0.05	0.08	0.06	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
DeepGazeIII [214]	0.04	0.03	0.09	0.14	-	0.08	-	0.22	0.11	0.07	0.09	0.04	-	0.08	-	0.06	-	0.06	-	0.06	-	0.06	-	0.06	-	0.06	-	0.06	-	0.06	-
Scanpath-VQA [74]	0.49	0.35	0.09	0.20	0.02	0.23	0.40	0.28	0.11	0.07	0.06	0.03	0.16	0.09	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
TPP-Gaze (GRU)	0.03	0.04	0.05	0.12	0.03	0.05	0.20	0.20	0.03	0.02	0.01	0.02	0.10	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
TPP-Gaze (Transformer)	0.02	0.04	0.06	0.14	0.05	0.06	0.25	0.44	0.03	0.01	0.02	0.01	0.13	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04

Table 3.3: Comparison of various models on OSIE, NUSEF, and FiFa datasets. **Gray color** indicates models trained under the same settings and datasets. Within this group, **bold** values represent the best performance for each metric. Underline values indicate the overall best performance across all models and metrics.

	OSIE			NUSEF			FiFa		
	SS (KL-Div) ↓		SED ↓	SS (KL-Div) ↓		SED ↓	SS (KL-Div) ↓		SED ↓
	w/ Dur	w/o Dur	Avg	w/ Dur	w/o Dur	Avg	w/ Dur	w/o Dur	Avg
Itti-Koch [177]	-	3.93	9.07	-	1.89	9.97	-	14.70	8.65
CLE (Itti) [41, 177]	-	3.24	9.29	-	1.40	10.16	-	12.62	8.86
CLE (DG) [41, 215]	-	3.65	9.23	-	1.35	10.07	-	14.38	8.83
G-Eymol [394]	12.28	2.95	8.00	1.99	0.53	<u>8.02</u>	13.17	5.00	<u>6.13</u>
IOR-ROI-LSTM [80]	0.20	2.84	8.82	0.06	1.10	9.69	0.30	12.44	8.27
DeepGazeIII [214]	-	2.51	8.47	-	1.04	9.38	-	12.08	7.97
Scanpath-VQA [74]	<u>0.02</u>	<u>0.09</u>	<u>7.55</u>	<u>0.02</u>	0.10	8.39	<u>0.03</u>	0.44	6.81
DeepGazeIII [214]	-	2.52	8.57	-	1.04	9.42	-	12.25	8.00
Scanpath-VQA [74]	0.29	0.31	9.70	0.06	0.18	10.61	0.35	0.90	9.73
TPP-Gaze (GRU)	0.25	0.30	8.05	<u>0.02</u>	0.03	8.41	0.15	<u>0.24</u>	7.00
TPP-Gaze (Transformer)	0.29	0.35	8.10	<u>0.02</u>	0.04	8.40	0.21	0.31	7.05

Table 3.4: Additional results on OSIE, NUSEF, and FiFa datasets in terms of SS and SED. **Gray color** indicates models trained under the same settings and datasets. Within this group, **bold** values represent the best performance for each metric. Underline values indicate the overall best performance across all models and metrics.

GRU or Transformer-based history encoding) outperforms all the other approaches on most of the adopted metrics. Interestingly, in many cases, the proposed approach offers the best overall performance, even when considering the pre-trained models released by the authors, except for ScanMatch, where Scanpath-VQA, which is directly optimized via reinforcement learning on this metric, understandably proves to be the best. It is also worth noting that, especially for the NUSEF and FiFa datasets, our approach can achieve the best overall results in terms of SS with and without duration. Some qualitative results are shown in Fig. 3.5, where we report sampled scanpaths from five models alongside those from humans. TPP-Gaze can predict fixations that better align with those recorded from human subjects, confirming the advantages of the proposed approach for predicting scanpaths during free-viewing. We refer the reader to Section B.1 of the appendix for additional qualitative results.

Additional analyses in Fig. 3.4 and Fig. 3.6 show empirical distributions summarizing TPP-Gaze’s scanpath statistics compared to those yielded by human observers and other methods. TPP-Gaze features scanpath statistics that better align with human behavior when compared with IOR-ROI-LSTM [80], DeepGazeIII [214] and Scanpath-VQA [74]. Even when tested on MIT1003,

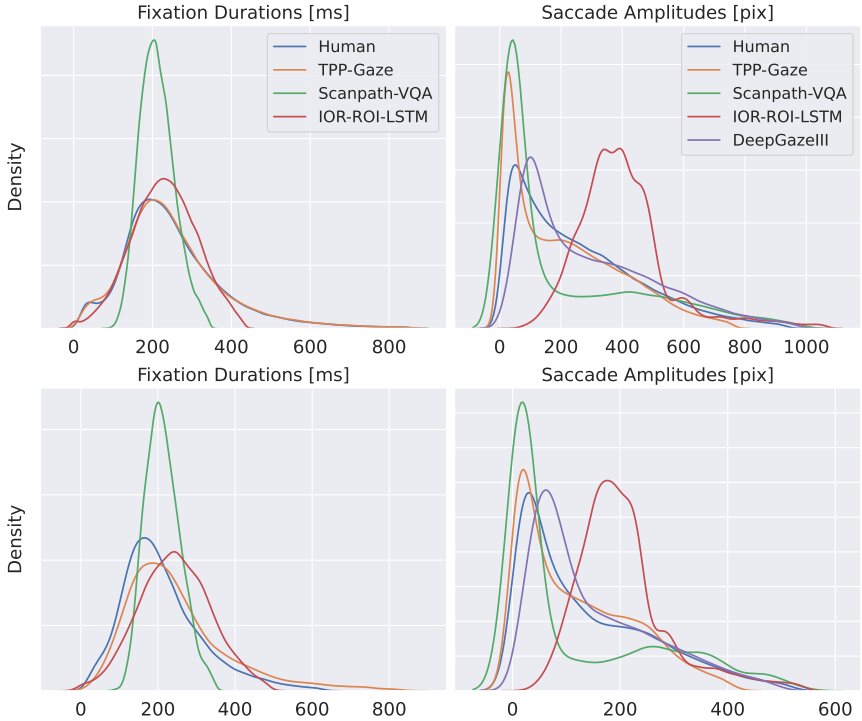


Figure 3.4: Statistical properties exhibited by TPP-Gaze and other methods relative to those of human observers, in terms of empirical fixation durations and saccade amplitudes on the COCO-FreeView (top row) and OSIE (bottom row) datasets.

NUSEF, and FiFa, TPP-Gaze effectively models the long-tail distribution of both fixation durations and saccade amplitudes. In contrast, other methods tend to capture only the average human gaze dynamics. An exception is DeepGazeIII on NUSEF, which achieves comparable results for saccade amplitudes but does not model fixation duration.

Beyond common scanpath statistics, we further evaluate the proposed approach using a return fixations (RF) analysis [400]. RF analysis describes the tendency of observers (either human or simulated) to revisit previously foveated locations. The frequency of RFs and the temporal offset (*i.e.*, the number of intervening fixations before returning to a location) at which they occur, provide a more nuanced description of the cognitive processes underlying attention alloca-

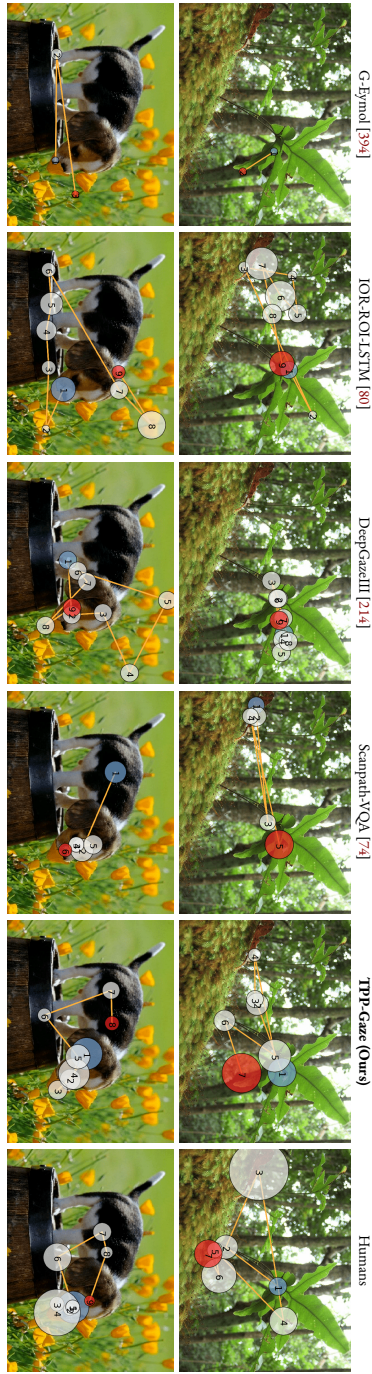


Figure 3.5: Comparison of simulated and human scanpaths. Each circle represents a fixation point, with its diameter proportional to the fixation duration. For methods that do not model fixation duration, circles are shown with a uniform size.

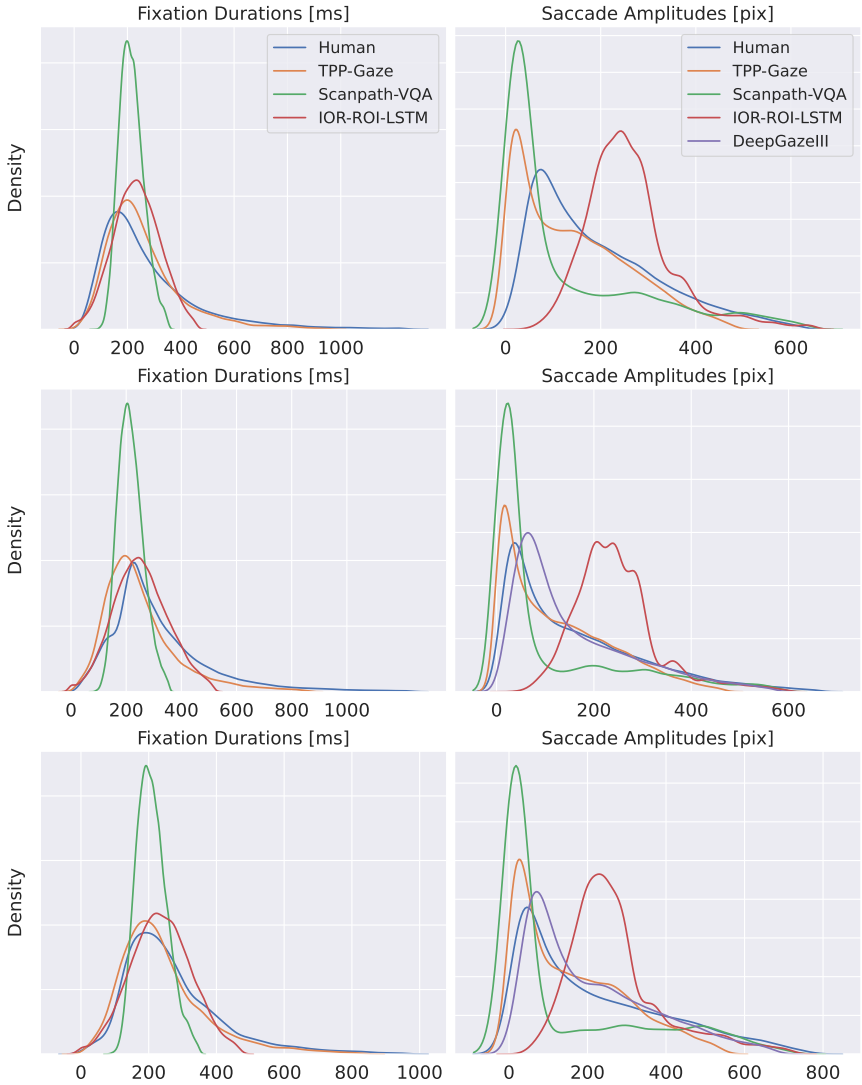


Figure 3.6: Statistical properties exhibited by TPP-Gaze and other methods relative to those of human observers, in terms of empirical fixation durations and saccade amplitudes on MIT1003 (top row), NUSEF (middle row) and FiFa (bottom row) datasets.

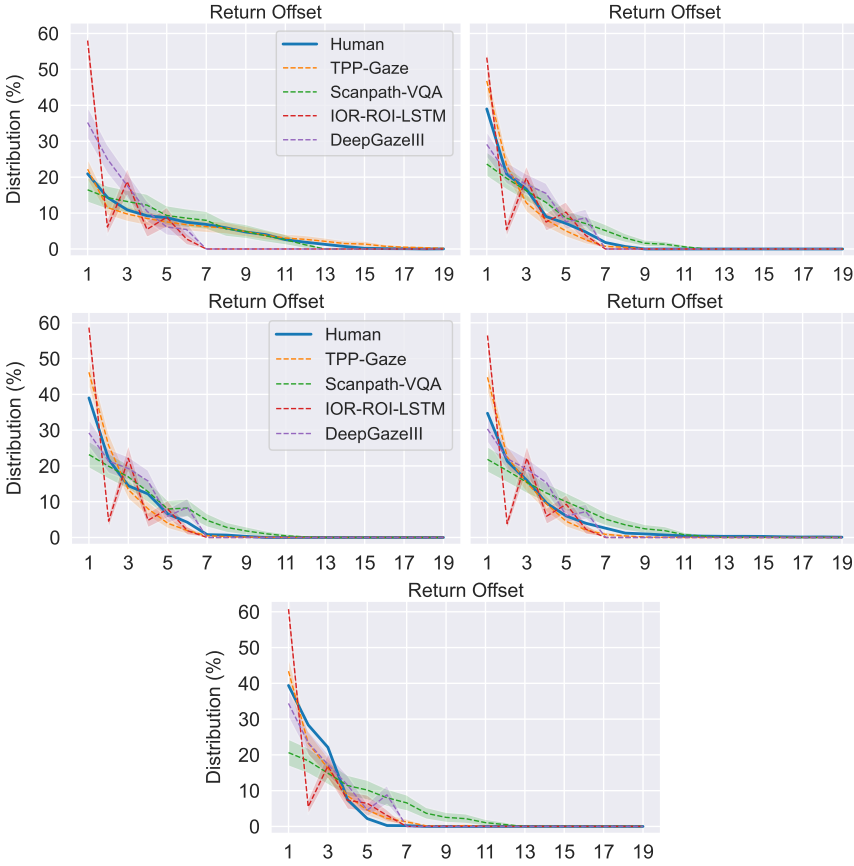


Figure 3.7: Return fixations analysis comparing TPP-Gaze with other methods and human observers. Results are shown on COCO-FreeView (top-left), OSIE (top-right), MIT1003 (middle-left), NUSEF (middle-right) and FiFa (bottom) datasets.

tion [400]. Fig. 3.7 reports the results of this analysis in comparison with existing methods across several datasets. Although TPP-Gaze was not explicitly trained for this objective, it produces the most accurate RF patterns, generally presenting an RF distribution that aligns better with human observers compared to other methods.

3.1.6 APPLICATIONS

SALIENCY PREDICTION. The performance of TPP-Gaze is further evaluated by comparing the saliency maps “backward” generated from fixations with those of human observers across all evaluated scanpath models. The results, presented in Table 3.5, are measured using three commonly adopted saliency metrics [55, 56, 92]: Kullback-Leibler Divergence (KL-Div), Judd’s Area Under the Curve (AUC), and Normalised Scanpath Saliency (NSS). DeepGazeIII and CLE (DG) are reported here only as references for the performance of a saliency prediction model, given their adoption of an extensive pre-training phase designed expressly for saliency generation (DeepGazeIII), or explicit adoption of a saliency model (CLE (DG)). Overall, TPP-Gaze obtains the best or second-best performance across all metrics and datasets. It yields results that are comparable to or surpass those of IOR-ROI-LSTM [80] and Scanpath-VQA [74], which are significantly better than all other approaches. This further demonstrates the effectiveness of our approach in predicting fixation points that better resemble human scanpaths than those predicted by existing methods. Qualitative results are reported in Section B.1 of the appendix.

3.1.7 EXTENSION TO VISUAL SEARCH TASKS

Recently, several works [74, 116, 256, 389, 402] have focused on predicting attention allocation on specific targets (visual search tasks). Although TPP-Gaze was originally devised and evaluated for the free-viewing scenario, we extend the TPP-Gaze architecture to handle the visual search task by forcing the model to learn a task-specific semantic representation of the input image (see Fig. 3.8). Recall that the TPP-Gaze’s semantic representation module consists of a DenseNet201 CNN backbone and a learnable readout network composed of three 1×1 convolutional layers with 8, 16, and 1 channels, respectively. The spatial priority map is then projected to a fixed-dimensional vector, $\mathbf{z}_j$, to obtain the j -th image semantic representation. Specifically, the last layer performing a 1×1 convolution is responsible for learning a non-linear combination of the feature maps from the previous layers.

	COCO-FreeView			MIT1003			OSIE			NUSEF			Fifa		
	KL-Div ↓	AUC ↑	NSS ↑	KL-Div ↓	AUC ↑	NSS ↑	KL-Div ↓	AUC ↑	NSS ↑	KL-Div ↓	AUC ↑	NSS ↑	KL-Div ↓	AUC ↑	NSS ↑
<i>Saliency-based</i>															
CLE (DG) [41, 215]	8.65	0.55	0.09	-	-	-	5.08	0.59	0.28	4.99	0.63	0.38	6.39	0.59	0.25
DeepGazeIII [214]	0.85	0.84	1.75	-	-	-	0.32	0.87	2.01	0.49	0.85	1.89	0.62	0.88	2.52
<i>Saliency-free</i>															
Itti-Koch [177]	8.94	0.56	0.24	5.01	0.64	0.47	3.35	0.65	0.51	4.84	0.63	0.40	5.47	0.64	0.42
CLE (Itti) [41, 177]	7.45	0.54	0.07	4.15	0.61	0.23	3.45	0.61	0.23	3.36	0.63	0.31	4.84	0.60	0.23
G-Fymol [394]	10.98	0.56	0.26	7.64	0.62	0.35	4.58	0.67	0.60	5.09	0.66	0.55	9.04	0.62	0.47
IOR-ROI-LSTM [80]	1.30	0.77	0.99	0.78	0.81	1.40	0.50	0.83	1.46	0.74	0.80	1.32	0.83	0.85	1.72
Scannpath-VQA [74]	3.53	0.77	1.56	2.12	0.82	2.01	1.26	0.84	2.12	2.45	0.80	1.76	1.88	0.86	2.89
TPP-Gaze (GRU)	1.01	0.84	1.65	0.78	0.86	2.06	<u>0.67</u>	0.84	<u>1.72</u>	0.84	0.84	<u>1.71</u>	<u>1.07</u>	0.86	<u>2.06</u>
TPP-Gaze (Transformer)	<u>1.11</u>	0.83	1.54	0.83	0.85	1.93	0.68	0.84	1.68	<u>0.79</u>	0.84	1.70	1.11	0.85	1.91

Table 3.5: Saliency prediction results on COCO-FreeView, MIT1003, OSIE, NUSEF, and Fifa datasets. Models are grouped into saliency-based and saliency-free methods, where the former (*i.e.*, CLE (DG) and DeepGazeIII) incorporate components trained to predict saliency maps. **Bold** values represent the best performance within each metric, while underline values indicate the second-best results.

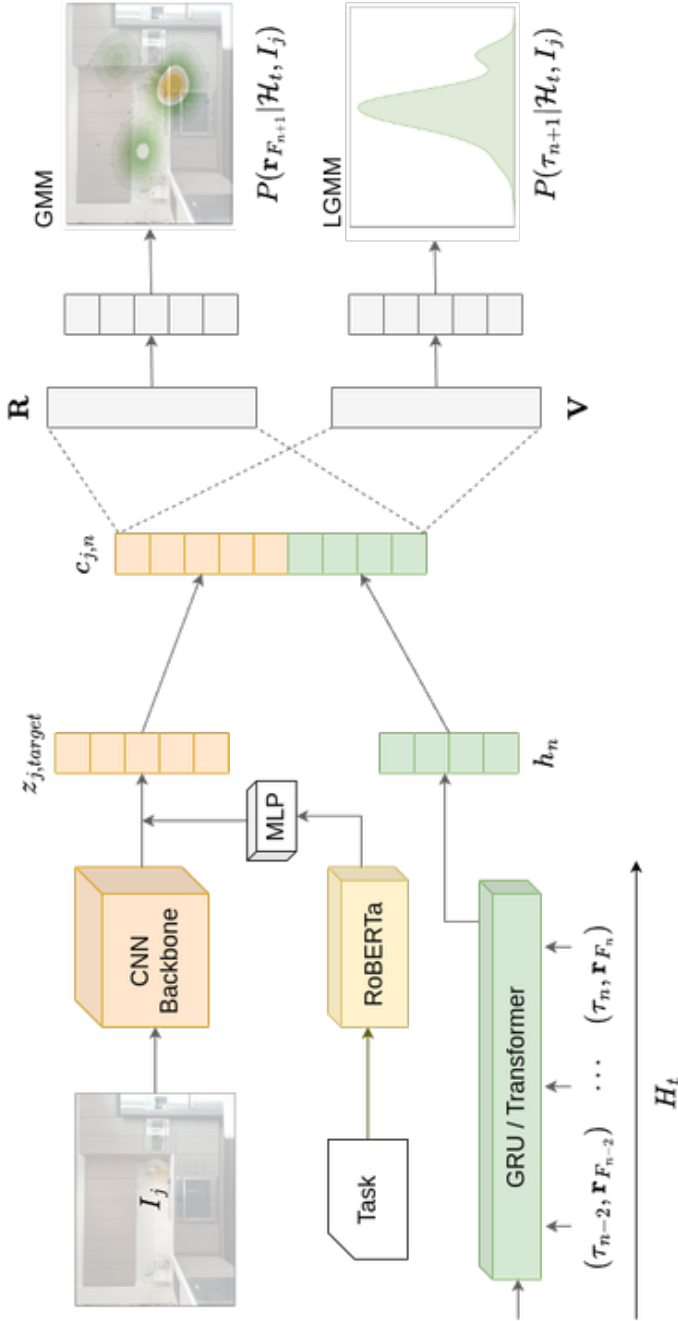


Figure 3.8: Overview of TPP-Gaze model architecture extended to handle the visual search task. A linguistic embedding (RoBERTa) of the search target is employed to learn a task-driven semantic representation (z_j). The latter, together with the history of past events (h_n), is used to simulate the next fixation position and duration.

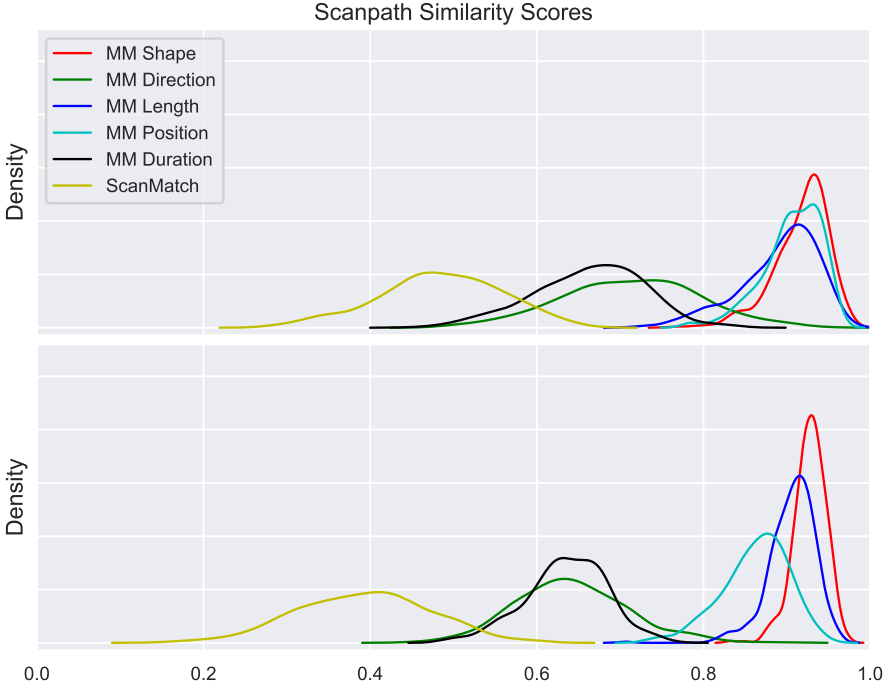


Figure 3.9: Empirical distributions of the adopted metrics quantifying inter-humans (top) and human *vs.* TPP-Gaze (bottom) scanpath similarity for the visual search task on COCO-Search18.

To guide the model toward a specific search objective, we redefine the architecture to enable TPP-Gaze to learn such a combination conditioned on a given text string. To this end, we first obtain a linguistic embedding of the search target using the RoBERTa language model [238]. Let $\mathbf{F}_{target}$ be the embedding vector representing the search objective. The readout network for the visual search model consists of three 1×1 convolutional layers with 16, 64, and 256 channels, respectively. Thus, it is modified to output $M = 256$ feature maps. Let $\mathbf{X} = [\mathbf{x}_0; \dots; \mathbf{x}_M] \in \mathbb{R}^{M \times d}$ represent the matrix of flattened image features.

The task-specific semantic representation for the j -th image, $\mathbf{z}_{j,target}$, is then obtained as follows:

$$w = \text{softplus}(\text{MLP}(\mathbf{F}_{target})), \quad \mathbf{z}_{j,target} = \sum_{i=1}^M w_i \mathbf{x}_i. \quad (3.11)$$



Figure 3.10: Human (left) and simulated (right) scanpaths for the visual search task. Search objective is “Sink”.

Preliminary results show encouraging trends on the COCO-Search18 dataset [78, 386], as illustrated in Fig. 3.9, where visual search patterns produced by TPP-Gaze are compared to human patterns using the MM and SM metrics. A qualitative example is depicted in Fig. 3.10. For further results, see Section B.1.

3.1.8 DISCUSSION AND LIMITATIONS

TPP-Gaze is built upon the intensity-free learning formulation of temporal point processes [314], thereby inheriting several of the associated limitations. Instead of explicitly modeling a conditional intensity function, it adopts an intensity-free formulation where the distribution over future gaze events is conditioned on a latent encoding of the observed scanpath. While this removes the optimization burden associated with intensity-based neural TPPs, it does not eliminate dependence on past observations. Rather, it shifts it into a compressed hidden representation. This introduces a potential information bottleneck, since long-range temporal dependencies and fine-grained behavioral cues must be encoded into a fixed-dimensional state. As a result, subtle aspects of gaze behavior that depend on extended context or higher-order interactions may not be fully preserved. Moreover, the reliance on event-based conditioning means that broader influences on visual attention, such as semantic scene understanding, task structure, and cognitive state, are only indirectly captured unless explicitly incorporated into the model. While the intensity-free approach improves flexibility, it may reduce interpretability compared to explicit intensity-based formulations,

making it harder to relate predicted event dynamics to meaningful psychological constructs. Finally, as with most TPP-based approaches, generalization to out-of-distribution viewing conditions remains limited, particularly in unconstrained environments where gaze behaviour is strongly influenced by context, task semantics, and individual variability not explicitly encoded in the event formulation.

3.2 MODELING HUMAN GAZE BEHAVIOR WITH DIFFUSION MODELS FOR UNIFIED SCANPATH PREDICTION

Visual attention deployment is a dynamic and selective mechanism through which humans efficiently process the vast amount of information in complex visual stimuli. Beyond temporal evolution, a fundamental aspect of visual attention is its inherent variability across humans. Consequently, human gaze behavior is more appropriately described as a distribution over plausible trajectories rather than as a single canonical sequence of fixations. However, much of the existing work on scanpath prediction generates scanpaths that reflect an averaged behavior, failing to capture the rich variability observed in individual visual exploration [220, 297, 379].

As noted in [59], the decision of where to look next at any given moment is neither entirely deterministic nor completely random. Modeling the effects of randomness allows us to efficiently address the influence of complex factors, such as oculomotor biases, traits, and motor response variability, at both the individual and group levels. Indeed, variability and the resulting stochasticity of gaze allocation go beyond merely revealing individual idiosyncrasies, which are significant in clinical and psychological studies. It also enables the observer to remain responsive to new signals and promotes a flexible shift of attention. This flexibility, in turn, facilitates efficient learning and exploration of the environment, an essential capability for autonomous systems in computer vision and robotics [24, 336].

This mismatch between the stochastic nature of human visual exploration and the predominantly deterministic formulation of current models constitutes a fundamental limitation in both modeling and evaluation, motivating the need for new solutions that explicitly represent variability in visual attention. Recent advances in generative modeling, particularly diffusion probabilistic models [115, 161], offer a promising alternative by using stochastic sampling to learn and generate diverse sequential outputs. Early applications to scanpath prediction have shown promise in generating human-like gaze behaviors that capture the inherent variability of individual scanpaths on 360° images [191, 363] and text [43]. When combined with the sequence modeling strengths of Transformer-based architectures [117, 350], these approaches move beyond deterministic predictions to simulate a broader range of plausible scanpaths aligned with cognitive theories of attention [410].

Recent works have begun to address this issue through personalized scanpath prediction, typically by conditioning models on subject-specific embeddings or identity representations [75, 189, 378]. While this direction is an important step toward accounting for inter-subject differences, it primarily focuses on adapting predictions to a known individual. However, this thesis focuses on a related but distinct setting. Rather than assuming a single representative scanpath or conditioning on a specific observer identity, we model group-level visual attention while explicitly accounting for the inter-subject variability within the group.

Based on this observation, we propose *ScanDiff*, a unified architecture that integrates diffusion models with Vision Transformers [117, 266] to generate diverse and realistic gaze scanpaths. Unlike existing approaches, *ScanDiff* explicitly models scanpath variability by leveraging the stochastic nature of diffusion models, enabling the generation of diverse yet plausible gaze trajectories. Furthermore, our method incorporates textual conditioning and a length prediction module, enabling the model to adapt to diverse visual search objectives within a unified framework.

Through extensive experiments on COCO-FreeView [79], MIT1003 [194], and COCO-Search18 [78, 386], we demonstrate that *ScanDiff* sets a new state-of-the-art in scanpath prediction across both free-viewing and task-driven scenar-

ios. Additionally, we present a novel analysis of the variability of predicted scanpaths, highlighting that our approach generates highly diverse eye trajectories and better captures human gaze behaviors than competing methods. This is achieved by leveraging existing scanpath prediction metrics and incorporating a new measure that penalizes excessive similarity among generated scanpaths.

In summary, our key contributions are as follows:

- A novel diffusion-based architecture that models the inherent stochasticity of human gaze, enabling the generation of diverse and realistic gaze trajectories.
- A unified framework that integrates textual conditioning and a length prediction module, allowing the model to adapt to both free-viewing and task-driven scenarios.
- A comprehensive evaluation that includes a novel analysis of scanpath variability, demonstrating that `ScanDiff` outperforms existing methods in capturing the diversity of human gaze behavior, along with achieving state-of-the-art results in traditional scanpath prediction metrics.

Source code and trained models are publicly released[†].

3.2.1 DIFFUSION MODELS FOR TRAJECTORY GENERATION

For many years, Generative Adversarial Networks (GANs) [146] have been the dominant paradigm for high-quality sample generation. GANs rely on an adversarial training scheme between a generator and a discriminator. Despite their empirical success, the min-max training process is notoriously unstable; they are sensitive to hyperparameters, and they suffer from mode collapse, which limits the coverage of the data distribution. Moreover, the lack of an explicit likelihood makes both evaluation and theoretical analysis challenging.

Unlike classical generative models that learn a direct mapping from noise to data, Diffusion Models [161, 320] have emerged as a compelling alternative, addressing several of these limitations. They represent a family of generative methods that formulate the generation process as a gradual transformation over time,

[†]<https://aimagelab.github.io/ScanDiff>

progressively refining coarse structures into fine details. By replacing adversarial training with a well-defined probabilistic framework based on progressive noising and denoising, diffusion models offer more stable training dynamics, while achieving high sample quality across a wide range of domains [53, 254, 262]. Although this comes at the cost of increased computational effort during generation due to the iterative sampling process, the improved robustness, scalability, and coverage of the data distribution have led diffusion models to largely supplant GANs in many applications.

This central idea has been developed through three main perspectives:

- **Variational View** [161, 202, 320]: It frames diffusion as learning a denoising process through a variational objective.
- **Score-Based View** [3, 324]: It learns the score function, the gradient of the log data density, which guides how to gradually remove noise from samples.
- **Flow-Based View** [233, 296]: It models generation as a continuous transformation that transports samples from a simple prior toward the data distribution.

These perspectives provide complementary ways to understand and implement diffusion modeling. In this thesis, we view diffusion models through a variational lens.

VARIATIONAL PERSPECTIVE: DDPM

The variational framework gave rise to the Denoising Diffusion Probabilistic Models (DDPMs) [161, 320], which represent a cornerstone of diffusion modeling. At their core, DDPMs involve two distinct stochastic processes: the *forward process* and the *reverse denoising process*. An illustration of this framework is depicted in Fig. 3.11.

THE FORWARD PROCESS. Given a sample $\mathbf{x}_0$ drawn from a real-world data distribution $\mathbf{x}_0 \sim q(\mathbf{x})$, the forward process is a non-trainable operation, fixed to a Markov chain that gradually corrupts the original data by adding Gaussian noise $\varepsilon \sim \mathcal{N}(0, I)$ for a number of timesteps T , producing a sequence of noisy sam-

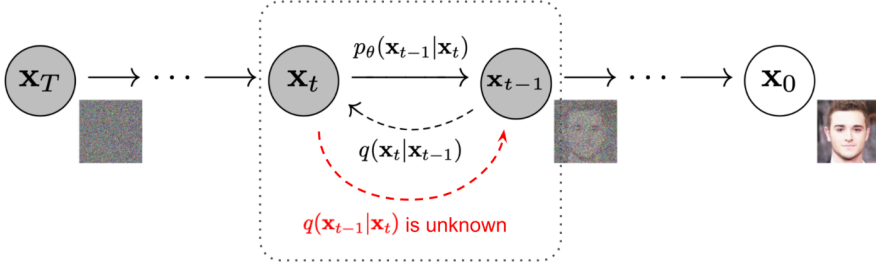


Figure 3.11: Illustration of the forward process and reverse process of the Denoising Diffusion Probabilistic Models (DDPM). Figure from [367].

ples $\mathbf{x}_1, \dots, \mathbf{x}_T$, according to a noise variance schedule $\{\beta_t \in (0, 1)\}_{t=1}^T$. This produces, at each timestep t , a latent variable $\mathbf{x}_t$ with distribution $q(\mathbf{x}_t | \mathbf{x}_{t-1})$, defined as:

$$\begin{aligned} q(\mathbf{x}_t | \mathbf{x}_{t-1}) &= \mathcal{N}\left(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}\right), \\ q(\mathbf{x}_{1:T} | \mathbf{x}_0) &= \prod_{t=1}^T q(\mathbf{x}_t | \mathbf{x}_{t-1}), \end{aligned} \quad (3.12)$$

with $\mathbf{I}$ being the identity matrix.

As a result, the forward process gradually corrupts the distinguishable features of the original sample $\mathbf{x}_0$ while the step t increases. When $T \rightarrow \infty$, $\mathbf{x}_T$ approaches an isotropic Gaussian distribution $p_{prior} := \mathcal{N}(0, \mathbf{I})$.

A useful property of the forward process is that it allows the sampling of $\mathbf{x}_t$ at an arbitrary timestep t given the original data $\mathbf{x}_0$. By defining $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$, we obtain:

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) \mathbf{I}). \quad (3.13)$$

THE REVERSE DENOISING PROCESS. The essence of the reverse process is to reverse the forward process. Starting from pure noise $\mathbf{x}_T \sim p_{prior}$, the final goal is to progressively denoise this randomness until a meaningful data sample $\mathbf{x}_0$ emerges. Unfortunately, directly computing the target distribution $q(\mathbf{x}_{t-1} | \mathbf{x}_t)$

is challenging. Indeed, by Bayes' theorem, we would need to evaluate:

$$q(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \frac{q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) q_{t-1}(\mathbf{x}_{t-1})}{\underbrace{q_t(\mathbf{x}_t)}_{\text{intractable}}}. \quad (3.14)$$

Thus, a closed-form evaluation is not possible because the data distribution p_{data} is unknown. The solution is to learn a model p_θ to approximate these conditional probabilities and run the reverse process. Formally, this process is defined by a Markov chain parameterized by θ :

$$p_\theta(\mathbf{x}_{0:T}) := p_\theta(\mathbf{x}_T) \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t). \quad (3.15)$$

In particular, each transition $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ is modeled as a Gaussian:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}\left(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)\right), \quad (3.16)$$

where μ_θ and Σ_θ represent the predicted mean and variance of the true posterior distribution, respectively. In most implementations, the variance Σ_θ is typically fixed according to a schedule β_t , even if some works proposed to learn it [261]. In typical DDPM implementation, training is usually based on the ε - *prediction* formulation. Specifically, the network p_θ predicts the noise ε that was added in the forward process, and a noisy sample $\mathbf{x}_i \sim q(\mathbf{x}_i \mid \mathbf{x})$ at timestep t is generated by:

$$\mathbf{x}_t = \sqrt{\alpha_t} \mathbf{x}_0 + \sqrt{1 - \alpha_t} \varepsilon_t, \quad \mathbf{x}_0 \sim p_{data}, \quad \varepsilon \sim \mathcal{N}(0, \mathbf{I}). \quad (3.17)$$

Knowing ε , we can analytically express the mean μ_θ of the posterior:

$$\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \alpha_t}} \underbrace{\varepsilon_\theta(\mathbf{x}_t, t)}_{\varepsilon\text{-prediction}} \right). \quad (3.18)$$

Through some simplifications, it is possible to demonstrate that the final train-

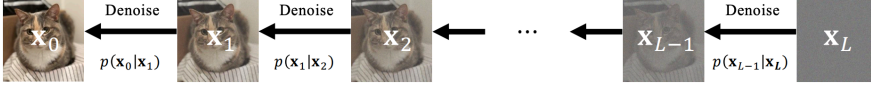


Figure 3.12: Sampling procedure with a ϵ - prediction model $\epsilon_\theta(\mathbf{x}_t, t)$. Figure from [367].

3

ing objective is equivalent to a simple ℓ_2 loss between the true noise and the network prediction:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x}_0, \epsilon, t} \left[\left\| \epsilon - \epsilon_\theta(\mathbf{x}_t, t) \right\|^2 \right]. \quad (3.19)$$

Minimizing this loss encourages the network to approximate the true reverse posterior without requiring access to the unknown $\mathbf{x}_0$ at generation time.

SAMPLING PROCEDURE. After training the ϵ -prediction model $\epsilon_\theta(\mathbf{x}_t, t)$, sampling is performed sequentially as illustrated in Fig. 3.12, using the parameterized transition $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$. Starting from a random sample

$$\mathbf{x}_T \sim p_{\text{prior}} = \mathcal{N}(0, \mathbf{I}),$$

we recursively sample from $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ following the update rule for $t = T, T-1, \dots, 1$:

$$\mathbf{x}_{t-1} \leftarrow \underbrace{\frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) \right)}_{\mu_\theta(\mathbf{x}_t, t)} + \sigma_t \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \mathbf{I}), \quad (3.20)$$

This *denoising* process continues until $\mathbf{x}_0$ is obtained as the final clean generated sample.

GENERATING SEQUENCES WITH DIFFUSION MODELS. Diffusion models have also gained popularity as a non-autoregressive alternative for modeling sequences [385], demonstrating success in generating various types of sequences, including continuous timeseries [87, 96, 231], text [145, 225], and audio [205]. These models have recently been adopted for generating spatio-temporal data, such as GPS trajectories [366, 415, 416] and human motion data [83, 323, 401], including eye movement patterns [43, 191, 363]. In contrast to these methods,

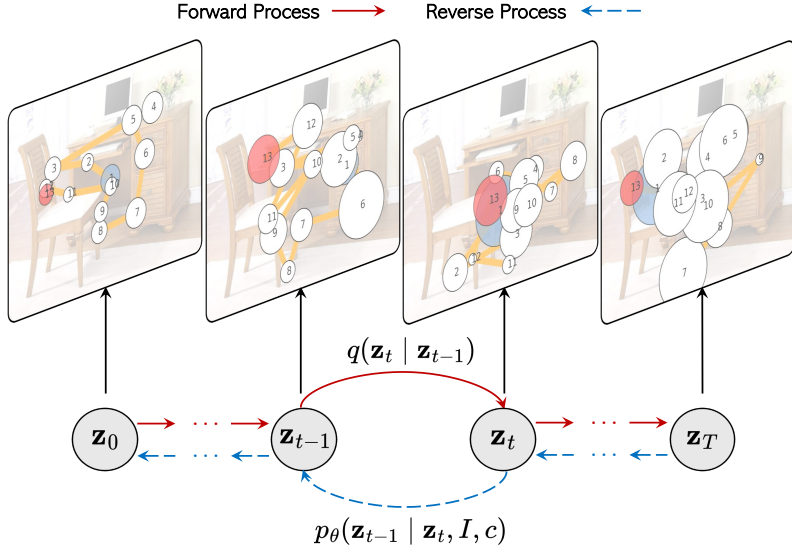


Figure 3.13: The diffusion process of ScanDiff that generates realistic scanpaths through learned transitions conditioned on image I and viewing task c .

we focus on both standard free-viewing and goal-oriented settings, introducing a novel approach able to predict scanpaths of variable lengths on natural images, thereby enabling greater variability and more realistic gaze behavior.

3.2.2 SCANDIFF: A DIFFUSION-BASED MODEL FOR UNIFIED SCANPATH PREDICTION

We propose ScanDiff, a novel scanpath prediction architecture based on diffusion models to generate realistic and diverse gaze patterns (see Fig. 3.13). Its multimodal nature enables the unified prediction of various types of visual attention, seamlessly adapting to different viewing tasks and stimuli.

Given an image or stimulus $I \in \mathbb{R}^{H \times W \times 3}$ and a viewing task c , the final objective is to predict a human-like scanpath represented as an ordered sequence of N fixations $\mathbf{s} = \{f_1, f_2, \dots, f_N\}$. Each fixation f_i consists of a pair (r_i, m_i) , where $r_i = (x_i, y_i) \in \mathbb{R}^2$ is the 2D spatial fixation location, while $m_i \in \mathbb{R}^+$ is the fixation duration. In this work, we propose a non-autoregressive approach to generate scanpath trajectories by learning a diffusion model ϕ_θ . Starting from a

noisy sample drawn from a Gaussian distribution, the learned model iteratively refines it to produce the final scanpath trajectory. An overview of the proposed architecture is depicted in Fig. 3.14.

3

FORWARD PROCESS: SCANPATH EMBEDDING

Let $\mathbf{s} \in \mathbb{R}^{N \times 3}$ be a sequence of N ground-truth fixations. To enable a structured latent space that better captures temporal and spatial dependencies of the scanpath, we learn a linear projection $g_\theta : \mathbb{R}^{N \times 3} \rightarrow \mathbb{R}^{N \times d}$ to map the scanpath $\mathbf{s}$ into an augmented embedding space, thus obtaining the initial uncorrupted latent variable $\mathbf{z}_0 = g_\theta(\mathbf{s}) \in \mathbb{R}^{N \times d}$. During the forward process (see Fig. 3.13), we gradually corrupt the whole embedded sequence $\mathbf{z}_0$ by adding Gaussian noise over T timesteps, following a predefined variance schedule. At each timestep t , the noisy latent representation $\mathbf{z}_t$ is obtained through a Markovian diffusion process, as defined in Eq. (3.12). By the final timestep, the representation $\mathbf{z}_T$ approaches an isotropic Gaussian distribution, effectively removing any trace of the original scanpath structure.

CONDITIONAL DENOISING PROCESS

Scanpath prediction involves the generation of an ordered sequence of N fixations conditioned on a given stimulus I and a viewing task c . Therefore, referring to Eq. 3.15, the conditioned denoising process can be rewritten as:

$$p_\theta(\mathbf{z}_{0:T} \mid I, c) := p(\mathbf{z}_T) \prod_{t=1}^T p_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t, I, c). \quad (3.21)$$

Our model φ_θ is based on an encoder-only Transformer [350], modified to incorporate an additional multi-head cross-attention layer between the self-attention and feed-forward layers [273]. This enables the model to effectively condition on the image I and the viewing task c .

IMAGE ENCODING. We process each stimulus $I \in \mathbb{R}^{H \times W \times 3}$ using a Transformer-based visual backbone $v(\cdot)$ which outputs a dense feature map $v(I) \in \mathbb{R}^{h \times w \times d_v}$. Here, h and w denote the number of patches along the height

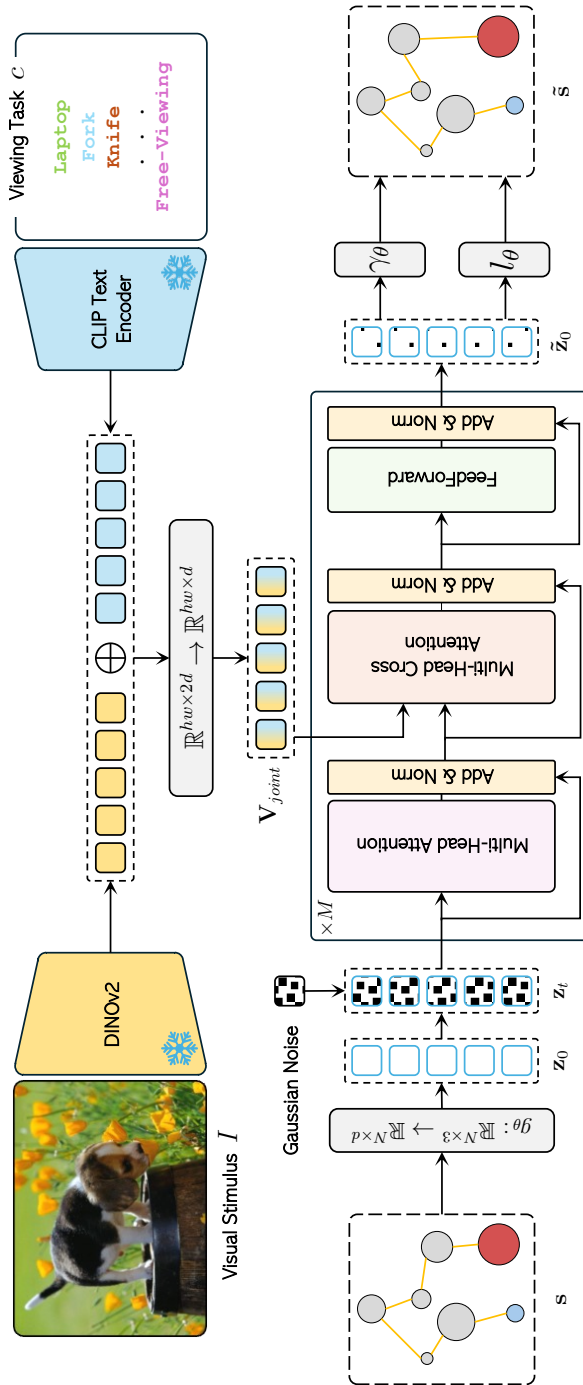


Figure 3.14: Overview of ScanDiff. Given a stimulus I and a viewing task c , a scanpath $\bar{s}$ is generated through a diffusion process.

and width of the image, respectively, while d_v refers to the dimensionality of the visual embedding space.

TASK ENCODING. Our model features a unified architecture that seamlessly adapts to different tasks without requiring any architectural modifications. This flexibility is achieved by using a text encoder to represent the viewing task c . Specifically, for the free-viewing task, we represent c as an empty string. In contrast, for visual search, c corresponds to the textual label of the target object to look for in the image (e.g. “laptop”). To extract task representations, we employ a pre-trained text encoder $\psi(\cdot)$ which maps c to a feature vector $\psi(c) \in \mathbb{R}^{d_t}$ in the textual embedding space, where d_t denotes the dimensionality of the textual features.

MULTIMODAL CONDITIONING. To condition the denoising process on both the image I and task c , we project the visual features $v(I)$ and the textual features $\psi(c)$ in a joint multimodal embedding space. Following previous works [256], we first map the visual and textual features into a common d -dimensional space using two independent linear transformations. The textual features are then repeated hw times and concatenated with the visual features along the channel dimension, resulting in a feature map of size $hw \times 2d$. Finally, this feature map is linearly projected into a d -dimensional feature space to obtain the final multimodal embedding $\mathbf{V}_{joint} \in \mathbb{R}^{hw \times d}$, which effectively combines visual and task semantic information. This multimodal conditioning enables a unified model to adapt to various tasks, from free-viewing to visual search. The resulting multimodal embedding $\mathbf{V}_{joint}$ is then passed through the cross-attention layer of the Transformer. In parallel, $\mathbf{z}_t$ is augmented with a learnable positional encoding and a sinusoidal diffusion timestep embedding. For the sake of simplicity, in what follows, we refer to $\mathbf{z}_t$ as the combination of the noisy scanpath embedding, the positional, and the diffusion timestep embeddings.

Unlike previous approaches [191, 363] that directly concatenate the noisy gaze sequence with the image embedding, we combine $\mathbf{z}_t$ and the visual-semantic features $\mathbf{V}_{joint}$ only in the cross-attention layer. This design choice allows the model to dynamically modulate the interaction between gaze dynamics and visual-semantic information, rather than enforcing a rigid concatenation.

SCANPATH RECONSTRUCTION & LENGTH PREDICTION. The Transformer encoder output is defined as $\tilde{\mathbf{z}}_0 = \varphi_\theta(\mathbf{z}_t, \mathbf{V}_{joint}) \in \mathbb{R}^{N \times d}$. The spatial coordinates and the duration of the fixations are reconstructed starting from the predicted sample $\tilde{\mathbf{z}}_0$. Specifically, a feed-forward network γ_θ with three linear layers followed by a ReLU activation function is adopted to decode $\tilde{\mathbf{z}}_0$ to an approximation of the original scanpath $\tilde{\mathbf{s}} = \{\tilde{f}_1, \tilde{f}_2, \dots, \tilde{f}_N\}, \tilde{f}_i \in \mathbb{R}^3$. The visual response to a given stimulus and the corresponding scanpath length can vary across subjects. Existing works that leverage diffusion models for scanpath generation [191, 363] typically produce fixed-length scanpaths. In contrast and crucially, we take a different approach: we introduce a length prediction module in the model architecture, which allows for greater flexibility. In particular, this module predicts the probability $\tilde{u}_i$ of each token in the reconstructed sample $\tilde{\mathbf{z}}_0$ to be valid through a linear function $l_\theta : \mathbb{R}^{N \times d} \rightarrow \mathbb{R}^N$. The final predicted length is equal to the number of consecutive valid tokens.

3.2.3 TRAINING AND INFERENCE

The training objective $\mathcal{L}$ is defined as the combination of four different components:

$$\mathcal{L} = \mathcal{L}_{VLB} + \mathcal{L}_{rec} + \mathcal{L}_{val} + \mathcal{L}_T. \quad (3.22)$$

The first component $\mathcal{L}_{VLB}$ aims to minimize the difference between the uncorrupted sample $\mathbf{z}_0$ and the model prediction. Formally, it is defined as:

$$\mathcal{L}_{VLB} = \sum_{t=1}^T \text{mid} \mathbf{z}_0 - \varphi_\theta(\mathbf{z}_t, \mathbf{V}_{joint}) \text{mid}^2. \quad (3.23)$$

This simplification can be derived from the variational lower bound [43]. To reduce the noise in the optimization of $\mathcal{L}_{VLB}$ we adopt importance sampling [261].

The second term $\mathcal{L}_{rec}$ measures the scanpath reconstruction error and is defined as the L_1 loss between the ground-truth scanpath $\mathbf{s}$ and the reconstructed

one $\tilde{\mathbf{s}}$:

$$\mathcal{L}_{rec} = \| \mathbf{s} - \tilde{\mathbf{s}} \| = \frac{1}{N} \sum_{i=1}^N (|x_i - \tilde{x}_i| + |y_i - \tilde{y}_i| + |m_i - \tilde{m}_i|). \quad (3.24)$$

Here, $\tilde{x}_i$ and $\tilde{y}_i$ denote the spatial coordinates of the reconstructed fixation $\tilde{f}_i$, and $\tilde{m}_i$ is the relative fixation duration. During training, the ground-truth scanpaths are padded or truncated to a maximum length of L . The $L - N$ padding fixations are masked out during the computation of $\mathcal{L}_{rec}$.

The term $\mathcal{L}_{val}$ represents binary cross-entropy loss for predicting the validity of each reconstructed fixation:

$$\mathcal{L}_{val} = \frac{1}{L} \sum_{i=0}^L \text{BCE}(u_i, \tilde{u}_i). \quad (3.25)$$

At the final diffusion step, the mean prediction should converge to zero under the assumption that the noise prior follows a standard Gaussian distribution. To stabilize the diffusion process, we define the loss $\mathcal{L}_T = \| \mu(\mathbf{z}_T) \|^2$, which penalizes any residual bias in the mean prediction at the final timestep, ensuring it ends in a clean isotropic Gaussian. This also regularizes training by enforcing theoretical constraints otherwise weakened by finite-step approximations.

SAMPLING A SCANPATH. At inference time, our model generates a scanpath in response to a visual stimulus I and a viewing task c (e.g. free-viewing or visual-search). We sample $\mathbf{z}_T \sim \mathcal{N}(0, \mathcal{I}) \in \mathbb{R}^{N \times d}$ and the model φ_θ iteratively denoises $\mathbf{z}_T$ to $\mathbf{z}_0$. At each sampling step t , the multimodal features $\mathbf{V}_{joint}$ are fused with $\mathbf{z}_t$ in the cross-attention layer to condition the reverse process. After denoising $\mathbf{z}_T$ into $\mathbf{z}_0$, this is fed through two independent inverse embedding layers to obtain the predicted scanpath and its relative length N . To account for the variability in the observational patterns across different subjects, we generate multiple scanpath trajectories by sampling distinct noisy samples $\mathbf{z}_T$.

3.2.4 EXPERIMENTAL RESULTS

EXPERIMENTAL SETUP

DATASETS. We evaluate our model on both the free-viewing and visual search tasks. For free-viewing, we adopt COCO-FreeView [79] and MIT1003 [194]. To train our model, we combine images from both datasets, using 70% for training, 15% for validation, and 15% for testing. To assess the performance on the visual search task, we employ COCO-Search18 [78, 386]. In this setting, we employ the training, validation, and test sets used in previous works [76, 256].

EVALUATION METRICS. The similarity between generated and human scanpaths is measured through the MultiMatch (MM) [114, 180], Scan-Match (SM) [97], Sequence Score (SS) [386], and Semantic Sequence Score (SemSS) [389] metrics. In particular, we adopt the same evaluation protocol proposed in TPP-Gaze [102], where the distribution of human vs. generated metrics is compared against the distribution of the human consistency metrics, using the Kullback-Leibler divergence. Beyond similarity, capturing the diversity of generated scanpaths is crucial to prevent the model from collapsing into a deterministic solution, thereby preserving the natural variability observed in human eye movements. To this end, we adopt two additional metrics: the Individual Scanpath Recall [388], which we rename as Recall Sequence Score (RSS), and a newly introduced metric that favors the diversity of generated scanpaths, termed as Diversity-aware Sequence Score (DSS).

In particular, the RSS measures the extent to which the generated scanpaths cover the variability of human scanpaths for a given stimulus. For a human scanpath in the dataset, it is considered covered if its SS with at least one generated scanpath surpasses a predefined threshold. The RSS is then computed as the ratio of covered human scanpaths to the total number of human scanpaths. This metric evaluates whether the model can replicate the range of individual behaviors observed in humans.

The novel DSS metric we propose extends the standard sequence similarity measures by incorporating a term that penalizes excessive similarity among the generated scanpaths when humans do not reflect such behavior. Given a set of

generated scanpaths $\mathbf{s}_g$ and corresponding human scanpaths $\mathbf{s}_h$ for a specific visual stimulus, DSS is computed as

$$\text{DSS}(\mathbf{s}_g, \mathbf{s}_h) = \frac{\text{SS}(\mathbf{s}_g, \mathbf{s}_h)}{1 + |\text{SS}(\mathbf{s}_g, \mathbf{s}_g) - \text{SS}(\mathbf{s}_h, \mathbf{s}_h)|} \quad (3.26)$$

where SS is the average sequence score calculated over the possible combinations of different scanpaths. The denominator penalizes models that produce overly uniform predictions, encouraging outputs that not only match human behavior but also reflect its natural variability.

IMPLEMENTATION & TRAINING DETAILS. We employ the DINOv2 ViT-B/14 model [266] as the pre-trained visual backbone, considering its rich semantic understanding of the visual scene. In particular, we use the DINOv2 variant with registers [106]. To align with its training resolution, we resize all images to a resolution of 518×518 , resulting in a feature map $v(I)$ of 37×37 patches, with an embedding dimension of $d_v = 768$. For textual encoding, we adopt the pre-trained CLIP ViT-B/32 model [285], which projects the viewing task c into a feature space of dimension $d_t = 512$. The modified architecture of the Transformer encoder consists of $M = 6$ layers, each with 8 attention heads and a hidden dimension $d = 512$. The model is trained with the AdamW optimizer, a batch size of 128, a learning rate set to 1×10^{-4} , a weight decay of 1×10^{-2} , and a number of diffusion steps $T = 1000$. In addition, a square-root noise schedule is adopted. Following previous works [74, 76], the maximum scanpath length is set to 16. During training, the spatial coordinates of each fixation are scaled in the range $[0, 1]$, and fixation durations are retained in seconds.

COMPARISON WITH THE STATE OF THE ART

We evaluate ScanDiff by comparing it with existing scanpath prediction models in both free-viewing and visual search tasks. Our evaluation includes a diverse range of approaches and architectures, covering both traditional model-based methods (*e.g.* Itti-Koch [177], CLE [41], and G-Eymol [394]) and deep learning-based models (*e.g.* PathGAN [15], IOR-ROI-LSTM [80], DeepGazeIII [214],

ChenLSTM [74], Gazeformer [256], HAT [388], ChenLSTM-ISP [75], GazeXplain [76], TPP-Gaze [102]). ChenLSTM-ISP [75] is originally designed to generate user-specific scanpaths. In contrast, our model accounts for population diversity. To adapt ChenLSTM-ISP to our setting, we make it predict a single scanpath conditioned on an image, a task, and a user identifier. By simulating this process across different user identifiers, we obtain a population of scanpaths representing diverse subjects. We retrain TPP-Gaze on the new splits, as training data used in the original model may include samples from the new test set.

FREE-VIEWING RESULTS. Table 3.6 presents a comprehensive evaluation across the considered free-viewing datasets. For a fair comparison, the most recent models were re-trained using identical settings and datasets to ScanDiff. These results are reported in gray color at the bottom of the table.

For the COCO-FreeView dataset, ScanDiff demonstrates competitive performance across multiple metrics. Our approach achieves the best results in the MM-direction metric among models trained with identical settings. This indicates a superior ability of our model to predict saccade directions that match human scanpaths. Furthermore, our model shows strong performance in the MM-position metric and achieves the best overall SM and SS metrics when considering fixation duration, demonstrating effective modeling of temporal dynamics. The MIT1003 dataset results further validate the effectiveness of ScanDiff. Our model achieves the highest MM average score among all competitors, demonstrating its overall strong performance. The superior results in SM and SS with duration are confirmed also on this dataset, outperforming all competing methods. This strong performance on the duration-aware metrics highlights the ability of our model to effectively represent the temporal aspects of visual attention. The unified diffusion-based architecture allows capturing both the spatial patterns of eye movements and their temporal characteristics, enabling more realistic scanpath generation.

ZERO-SHOT EVALUATION. As a complement to Table 3.6, we report in Table 3.8 the results obtained on the OSIE dataset [374], which is used exclusively for zero-shot evaluation and not included in the training process. This analysis allows us to assess the generalization capability of ScanDiff on unseen sam-

	COCO-FreeView										MIT1003																										
	MM↓					SM↓					SS↓					MM↓					SM↓					SS↓											
	Sh	Len	Dir	Pos	Dur	Avg	w/Dur	w/o Dur	Dur	w/Dur	w/o Dur	Sh	Len	Dir	Pos	Dur	Avg	w/Dur	w/o Dur	Dur	w/Dur	w/o Dur	Sh	Len	Dir	Pos	Dur	Avg	w/Dur	w/o Dur	Dur	w/Dur	w/o Dur				
Int-Koch [177]	0.504	0.507	0.237	1.325	-	0.643	-	4.317	-	1.624	-	1.040	0.702	0.353	2.900	-	1.249	-	3.233	-	6.639	-	0.504	0.507	0.237	1.325	-	0.643	-	4.317	-	1.624	-	6.639	-		
CLE (Int) [41, 177]	0.052	0.317	0.427	1.966	-	0.691	-	3.576	-	1.747	-	0.061	0.124	0.414	1.515	-	0.529	-	3.454	-	5.397	-	0.052	0.317	0.427	1.966	-	0.691	-	3.576	-	1.747	-	5.397	-		
CLE (DG) [41, 215]	0.037	0.180	0.323	1.823	-	0.591	-	3.657	-	1.750	-	0.099	0.038	0.458	1.066	-	0.415	-	2.566	-	6.234	-	0.037	0.180	0.323	1.823	-	0.591	-	3.657	-	1.750	-	6.234	-		
PathGAN [15]	0.070	0.406	1.009	0.073	0.031	0.318	1.210	1.383	0.718	1.012	0.063	0.234	1.603	0.514	0.165	0.516	2.255	1.069	1.087	1.190	0.070	0.406	1.009	0.073	0.031	0.318	1.210	1.383	0.718	1.012	0.063	0.234	1.603	0.514	1.087	1.190	
G-Eyemol [394]	0.583	0.741	1.296	0.550	0.676	0.769	9.350	8.990	8.622	4.127	0.870	0.523	0.444	0.431	0.187	0.491	9.942	2.513	9.799	3.771	0.583	0.741	1.296	0.550	0.676	0.769	9.350	8.990	8.622	4.127	0.870	0.523	0.444	0.431	0.187	0.491	
IOR-ROI-LSTM [80]	1.107	0.442	0.013	0.444	0.028	0.407	1.540	1.520	0.546	1.005	0.677	0.446	0.021	1.099	0.051	0.459	0.985	0.875	0.437	5.302	1.107	0.442	0.013	0.444	0.028	0.407	1.540	1.520	0.546	1.005	0.677	0.446	0.021	1.099	0.051	0.459	0.985
DeepGazeIII [214]	0.037	0.016	0.019	0.028	-	0.025	-	0.368	-	0.393	-	-	-	-	-	-	-	-	-	-	0.037	0.016	0.019	0.028	-	0.025	-	0.368	-	0.393	-	-	-	-	-		
ChenLSTM [74]	0.034	0.128	0.105	0.045	0.189	0.100	0.574	0.373	0.344	0.442	0.028	0.073	0.149	0.107	0.110	0.094	0.168	0.161	0.192	0.316	0.034	0.128	0.105	0.045	0.189	0.100	0.574	0.373	0.344	0.442	0.028	0.073	0.149	0.107	0.110	0.094	
HAT [388]	1.099	0.434	0.042	0.444	-	0.505	-	1.025	-	0.331	1.196	0.522	0.381	2.386	-	1.121	-	2.112	-	1.305	-	1.099	0.434	0.042	0.444	-	0.505	-	1.025	-	1.025	-	1.305	-	1.305	-	
ChenLST-ISP [75]	0.038	0.173	0.166	0.077	0.188	0.128	0.683	0.576	0.377	0.579	0.034	0.124	0.175	0.114	0.095	0.108	0.264	0.214	0.267	0.240	0.038	0.173	0.166	0.077	0.188	0.128	0.683	0.576	0.377	0.579	0.034	0.124	0.175	0.114	0.095	0.108	
GazeXplain [76]	0.151	0.195	0.874	0.164	0.382	0.353	3.915	3.423	2.278	5.616	0.018	0.065	0.079	0.058	0.188	0.082	0.035	0.094	0.072	1.419	0.151	0.195	0.874	0.164	0.382	0.353	3.915	3.423	2.278	5.616	0.018	0.065	0.079	0.058	0.188	0.082	
DeepGazeIII [214]	0.009	0.017	0.059	0.038	-	0.031	-	0.348	-	0.417	-	0.025	0.020	0.210	0.074	-	0.082	-	0.210	-	3.878	0.009	0.017	0.059	0.038	-	0.031	-	0.348	-	0.417	-	3.878	-			
ChenLSTM [74]	0.715	0.411	0.056	0.129	0.092	0.280	0.116	0.110	0.022	0.093	0.251	0.153	0.181	0.136	0.059	0.156	0.373	0.251	0.284	0.236	0.715	0.411	0.056	0.129	0.092	0.280	0.116	0.110	0.022	0.093	0.251	0.153	0.181	0.136	0.059	0.156	
GazeXplain [76]	0.346	0.226	0.032	0.033	0.068	0.141	0.049	0.038	0.017	0.007	0.060	0.046	0.065	0.025	0.044	0.048	0.158	0.051	0.128	0.043	0.346	0.226	0.032	0.033	0.068	0.141	0.049	0.038	0.017	0.007	0.060	0.046	0.065	0.025	0.044	0.048	
TTP-Gaze [102]	0.063	0.017	0.061	0.038	0.010	0.038	0.125	0.226	0.033	0.130	0.039	0.036	0.139	0.068	0.027	0.062	0.244	0.257	0.144	0.280	0.063	0.017	0.061	0.038	0.010	0.038	0.125	0.226	0.033	0.130	0.039	0.036	0.139	0.068	0.027	0.062	
Scandiff (Ours)	0.131	0.048	0.021	0.037	0.151	0.078	0.015	0.027	0.013	0.038	0.050	0.015	0.042	0.019	0.072	0.040	0.041	0.065	0.026	0.047	0.131	0.048	0.021	0.037	0.151	0.078	0.015	0.027	0.013	0.038	0.050	0.015	0.042	0.019	0.072	0.040	

Table 3.6: Performance comparison of different models on the COCO-FreeView [79] and MIT1003 [194] datasets. Models trained using identical settings and datasets to Scandiff are highlighted in **gray**. Among these, the highest performance for each metric is marked in **bold**. Underlined values denote the top overall performance across all models and metrics.

	Target-Present						Target-Absent					
	MM ↓		SM ↓		SS ↓		MM ↓		SM ↓		SS ↓	
	Avg	w/ Dur	w/o Dur	w/ Dur	w/o Dur	w/ Dur	Avg	w/ Dur	w/o Dur	w/ Dur	w/o Dur	w/ Dur
PathGAN [15]	0.513	1.891	2.451	0.808	0.939	0.313	0.468	0.125	0.792	0.357	0.498	0.944
ChenLSTM [74]	0.197	0.011	0.236	0.040	0.262	0.084	0.162	0.075	0.010	0.012	0.036	0.775
Gazeformer [256]	0.281	0.027	0.340	0.119	0.268	0.131	0.208	0.089	0.061	0.075	0.102	0.245
HAT [388]	0.118	-	0.263	-	0.148	-	3.837	0.052	-	0.063	-	0.097
ChenLSTM-ISP [75]	0.174	0.013	0.306	0.015	0.257	0.043	0.097	0.082	0.028	0.146	0.063	0.727
GazeXplain [76]	0.166	0.023	0.237	0.070	0.232	0.140	0.188	0.046	0.062	0.048	0.038	0.191
TPP-Gaze [102]	0.524	1.618	3.218	0.579	1.590	0.554	1.147	0.098	0.511	0.529	0.242	0.325
Gazeformer [256]	0.251	0.045	0.508	0.048	0.349	0.095	0.262	0.526	1.184	1.688	0.319	0.520
GazeXplain [76]	0.167	0.010	0.238	0.050	0.217	0.092	0.224	0.037	0.030	0.026	0.028	0.146
TPP-Gaze [102]	0.507	2.317	3.995	0.893	2.381	0.736	1.605	0.135	0.775	0.887	0.427	0.537
ScanDiff (Ours)	0.048	0.037	0.079	0.019	0.074	0.072	0.078	0.020	0.005	0.008	0.008	0.031
												0.007
												0.024

Table 3.7: Performance comparison of different models on the COCO-Search18 dataset [78, 386] for both target-present and target-absent settings. Models trained using identical settings and training splits to ScanDiff are highlighted in **gray**. Among these, the highest performance for each metric is marked in **bold**. Underlined values denote the top overall performance across all models and metrics.

	OSIE									
	MM ↓						SM ↓		SS ↓	
	Sh	Len	Dir	Pos	Dur	Avg	w/ Dur	w/o Dur	w/ Dur	w/o Dur
Itti-Koch [177]	1.886	1.293	0.482	2.910	-	1.643	-	4.467	-	4.086
CLE (Itti) [41, 177]	0.070	0.049	0.294	1.007	-	0.355	-	2.560	-	3.185
CLE (DG) [41, 215]	0.105	0.025	0.215	0.906	-	0.313	-	2.021	-	3.766
PathGAN [15]	0.070	0.108	0.575	2.148	0.199	0.620	3.504	3.155	2.061	1.960
G-Eymol [394]	1.531	0.782	0.238	2.159	0.324	1.007	14.068	7.125	9.468	3.341
DeepGazeIII [214]	0.058	0.025	0.143	0.200	-	0.107	-	0.333	-	2.465
ChenLSTM [74]	0.723	0.477	0.122	0.420	0.026	0.353	0.781	0.638	0.402	0.350
HAT [388]	2.793	1.248	0.207	2.236	-	1.621	-	3.371	-	1.548
TPP-Gaze [102]	0.070	0.070	0.085	0.288	0.067	0.116	1.058	0.648	0.779	0.365
ScanDiff (Ours)	0.036	0.024	0.040	0.150	0.035	0.057	0.219	0.305	0.253	0.226

Table 3.8: Performance comparison of different models on the OSIE [374] dataset for zero-shot prediction. Models with the highest performance for each metric is marked in **bold**.

ples. Our model achieves the best overall results when compared with models that don’t use OSIE as part of the training set. This highlights the ability of the model to generate plausible and diverse scanpaths without requiring dataset-specific fine-tuning, demonstrating its robustness also in out-of-distribution scenarios.

VISUAL SEARCH RESULTS. The results summarized in Table 3.7 demonstrate the superior performance of our proposed model on the COCO-Search18 dataset across multiple evaluation metrics and search scenarios. We evaluate performance in both target-present and target-absent conditions, which represent fundamentally different search behaviors in human visual attention. In the target-present scenario, ScanDiff achieves state-of-the-art performance across all metrics. Our model exhibits a significant improvement in MM distributions, with an average KL divergence of 0.048, which represents a 71.3% reduction compared to the second best model (*i.e.*, GazeXplain at 0.167). For SemSS, ScanDiff attains KL divergence values of 0.072 and 0.078 with and without duration information, respectively, demonstrating consistent performance across temporal aspects of gaze behavior. The target-absent condition presents a particularly challenging scenario, as human attention patterns become more exploratory when the target object cannot be found. Even in this setting, ScanDiff outperforms all baseline methods by substantial margins.

	Target-Present					Target-Absent				
	MM ↓					MM ↓				
	Sh	Len	Dir	Pos	Dur	Sh	Len	Dir	Pos	Dur
PathGAN [15]	0.594	0.365	0.937	0.333	0.336	0.030	0.103	0.167	0.153	0.172
ChenLSTM [74]	0.253	0.276	0.337	0.054	0.066	0.075	0.044	0.111	0.052	0.092
Gazeformer [256]	0.581	0.301	0.316	0.056	0.150	0.067	0.047	0.054	0.044	0.233
HAT [388]	0.161	0.089	0.115	0.108	-	0.108	0.040	0.024	0.034	-
ChenLSTM-ISP [75]	0.272	0.232	0.302	0.019	0.044	0.114	0.087	0.132	0.016	0.060
GazeXplain [76]	0.238	0.255	0.245	0.038	0.052	0.019	0.016	0.015	0.051	0.128
TPP-Gaze [102]	0.676	0.250	0.845	0.825	<u>0.025</u>	0.051	0.017	0.117	0.287	<u>0.018</u>
Gazeformer [256]	0.436	0.326	0.313	0.031	0.147	0.081	0.418	0.269	1.613	0.250
GazeXplain [76]	0.244	0.224	0.295	0.030	0.045	0.009	0.023	0.024	0.035	0.095
TPP-Gaze [102]	0.445	0.254	0.581	1.216	0.037	0.023	0.032	0.052	0.547	0.023
ScanDiff (Ours)	0.077	0.048	0.067	0.015	0.033	0.008	0.010	0.007	0.008	0.067

Table 3.9: Additional results on COCO-Search18 dataset [78] for both target-present and target-absent settings. Models trained using identical settings and training splits to ScanDiff are highlighted in **gray**. Among these, the highest performance for each metric is marked in **bold**. Underlined values denote the top overall performance across all models and metrics.

It is worth noting that while some competing methods like GazeXplain perform reasonably well in specific metrics, they lack the consistent performance across all evaluation dimensions that ScanDiff demonstrates. This consistency across metrics and conditions indicates that our model better captures the underlying mechanisms of human visual search behavior in both goal-directed (target-present) and more exploratory (target-absent) scenarios.

Table 3.9 presents a detailed breakdown of ScanDiff’s performance on the COCO-Search18 dataset across individual MM metrics for both target-present and target-absent scenarios. Extending the results reported in Table 3.7, ScanDiff consistently outperforms existing methods in modeling spatial characteristics of scanpaths with substantial margins. In the target-present condition, our approach achieves KL-divergence reductions of 68–78% for shape, 79–81% for length, 77–88% for direction, and 50–99% for position compared to models trained under identical settings (highlighted in gray). Similar improvements are observed in the target-absent condition. While TPP-Gaze shows slightly better performance in duration modeling for target-absent cases, ScanDiff maintains competitive performance in duration metrics for target-present scenarios

	Backbones		COCO-FreeView			COCO-Search18			
	Textual	Visual	MM ↓	SM ↓	SS ↓	MM ↓	SM ↓	SS ↓	SemSS ↓
<i>Effect of Textual Backbone</i>									
	RoBERTa	DINOv2	0.110	0.181	0.111	0.076	0.143	0.072	0.132
ScanDiff	CLIP	DINOv2	0.078	0.015	0.013	0.048	0.037	0.019	0.072
<i>Effect of Visual Backbone</i>									
	CLIP	RN50	0.049	0.019	0.029	0.070	0.052	0.031	0.084
	CLIP	CLIP	0.058	0.199	0.112	0.090	0.079	0.033	0.069
ScanDiff	CLIP	DINOv2	0.078	0.015	0.013	0.048	0.037	0.019	0.072

Table 3.10: Performance comparison of different textual and visual backbones on COCO-FreeView [387] and COCO-Search18 [78] (TP) datasets. Best results are highlighted in **bold**.

(*i.e.*, 0.033). These results further validate the effectiveness of our model in capturing the complex dynamics of task-driven visual behavior across different search conditions.

QUALITATIVE RESULTS. Fig. 3.15 shows some qualitative results comparing ScanDiff with other competitors in both free-viewing and visual search settings. These results confirm the effectiveness of our model, also from a qualitative standpoint, highlighting its ability to generate human-like eye movement trajectories across diverse scenarios. Additional qualitatIVES are reported in Section B.2 of the appendix.

ABLATION STUDIES

To provide insights into the design choices of our model, we conduct a series of ablation studies examining the impact of different components on scanpath prediction performance. Tables 3.10 and 3.11 summarize these results across free-viewing and visual search tasks.

EFFECT OF TEXTUAL AND VISUAL BACKBONES. We first investigate the influence of textual and visual backbones on model performance. As shown in Table 3.10, replacing the CLIP text encoder with RoBERTa [238] leads to a degradation in performance across all metrics on both datasets. This highlights the importance of vision-language pre-training for scanpath prediction, as CLIP’s joint embedding space better captures the semantic relationships between textual

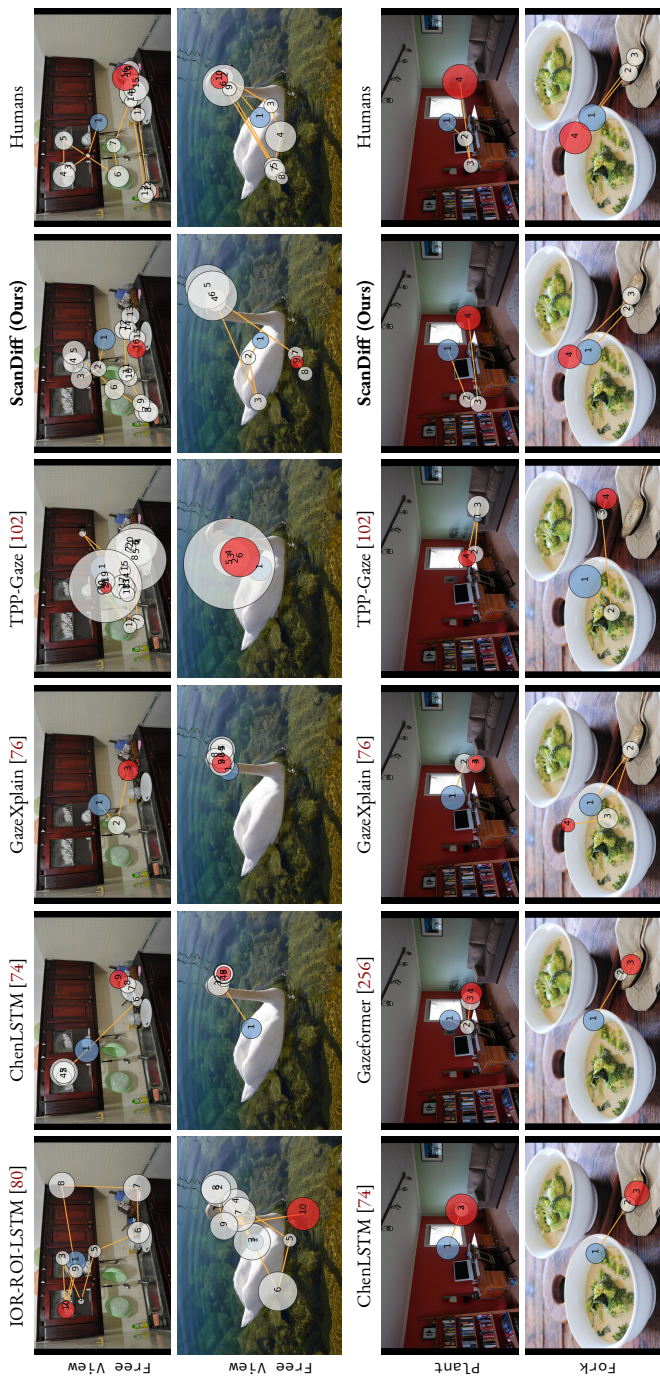


Figure 3.15: Comparison of simulated and human scanpaths across different datasets for both free-viewing and visual search tasks. From top to bottom: results on COCO-FreeView [387], MIT1003 [194], COCO-Search18 TP [78], and COCO-Search18 TA [389] datasets.

	$\mathcal{L}_T$	T	COCO-FreeView			COCO-Search18			
			MM ↓	SM ↓	SS ↓	MM ↓	SM ↓	SS ↓	SemSS ↓
<i>Effect of $\mathcal{L}_T$ Loss Function</i>									
	✗	1000	0.088	0.011	0.009	0.058	0.040	0.018	0.076
ScanDiff	✓	1000	0.078	0.015	0.013	0.048	0.037	0.019	0.072
<i>Varying Diffusion Timesteps</i>									
	✓	200	0.043	0.072	0.064	0.069	0.040	0.023	0.085
	✓	500	0.049	0.145	0.085	0.046	0.050	0.030	0.098
	✓	1500	0.045	0.131	0.122	0.057	0.086	0.061	0.130
ScanDiff	✓	1000	0.078	0.015	0.013	0.048	0.037	0.019	0.072

Table 3.11: Ablation study on the effect of alignment loss ($\mathcal{L}_T$) and diffusion timesteps (T) on COCO-FreeView [387] and COCO-Search18 [78] (TP) datasets. Best results are highlighted in bold.

queries and visual features that guide human attention.

For the visual backbone comparison, we test our model with ResNet-50-FPN [156], CLIP visual encoder (always using the ViT-B version), and DINOv2. The results indicate that while ResNet-50 achieves the best MM average score on COCO-FreeView, DINOv2 consistently outperforms other visual backbones across most metrics, particularly on the more challenging COCO-Search18 dataset. Interestingly, the CLIP visual encoder performs best on the SemSS metric, suggesting its strength in capturing semantic relationships between fixations and image regions, likely due to its vision-language pre-training.

EFFECT OF $\mathcal{L}_T$ LOSS FUNCTION. We evaluate the contribution of the diffusion prior alignment loss $\mathcal{L}_T$. As shown in Table 3.11, including $\mathcal{L}_T$ improves overall performance, but the benefits are more pronounced on the COCO-Search18 dataset, where it improves MM, SM and SemSS. This suggests that the convergence loss is particularly valuable for modeling the sequential nature of fixations in goal-directed visual search tasks, where the temporal order of fixations is more structured compared to free-viewing scenarios.

VALIDATING DIFFUSION TIMESTEPS. Finally, we investigate the impact of diffusion timesteps (T) on model performance by varying T from 200 to 1500. Using a number of timesteps $T = 1000$ achieves the best overall balance across metrics and datasets. While smaller timesteps (e.g., $T = 200$) can achieve better

	Len	COCO-FreeView			COCO-Search18 (TP)			
		MM ↓	SM ↓	SS ↓	MM ↓	SM ↓	SS ↓	SemSS ↓
w/o cross-attention	16	0.108	0.173	0.111	0.439	1.839	0.862	0.962
	32	0.052	0.204	0.113	0.084	0.069	0.030	0.056
	24	0.058	0.046	0.029	0.075	0.084	0.044	0.092
ScanDiff	16	0.078	0.015	0.013	0.048	0.037	0.019	0.072

Table 3.12: Ablation study on the effect of cross-attention compared to input concatenation and on the maximum scanpath length.

MM scores on COCO-FreeView, they underperform on the other spatial metrics. Similarly, $T = 500$ achieves a slightly better MM score on COCO-Search18, but at the cost of poorer performance on other metrics. Increasing T beyond 1000 (*i.e.*, $T = 1500$) leads to deteriorated performance across most metrics, likely due to overfitting to noise patterns. These results highlight the critical role of properly calibrating the diffusion process: sufficient timesteps are needed to learn complex distributions, but excessive noise can degrade the ability of the model to capture meaningful patterns in scanpath data.

EFFECT OF CROSS-ATTENTION & MAXIMUM SCANPATH LENGTH. Previous diffusion-based approaches [191, 363] perform the conditioning by directly concatenating the noisy gaze sequence with the image embedding. On a different line, we condition the denoising process through the cross-attention layer. Results in Table 3.12 demonstrate that the proposed approach allows for a better semantic alignment between the scanpath and the multimodal features compared to the rigid concatenation of the input. Additionally, in Table 3.12, we also report an ablation study on the maximum scanpath length. We set this value to 16, matching the highest median across all datasets used in our experiments and the value used in [74, 76]. This hyperparameter serves as an upper bound, though the model can predict variable lengths.

3.2.5 SCANPATH VARIABILITY ANALYSIS

Human visual exploration is inherently variable. Individuals perceive the same stimulus in different ways depending on factors such as attention, context, and

	COCO-FV		MIT1003		TP		TA	
	DSS ↑	RSS ↑	DSS ↑	RSS ↑	DSS ↑	RSS ↑	DSS ↑	RSS ↑
IOR-ROI-LSTM [80]	0.185	0.428	0.264	0.579	-	-	-	-
ChenLSTM [74]	0.174	0.420	0.257	0.534	0.386	0.635	0.247	0.591
Gazeformer [256]	-	-	-	-	0.377	0.578	0.206	0.417
HAT [388]	-	0.645	-	0.615	-	0.861	-	0.748
ChenLSTM-ISP [75]	0.190	0.501	0.264	0.619	0.423	0.735	0.268	0.670
GazeXplain [76]	0.099	0.032	0.302	0.674	0.406	0.689	0.283	0.716
TPP-Gaze [102]	0.271	0.732	0.313	0.758	0.284	0.516	0.221	0.663
ScanDiff (Ours)	0.277	0.736	0.354	0.815	0.425	0.747	0.312	0.800

Table 3.13: Analysis of scanpath variability on free-viewing (COCO-FreeView [387] and MIT1003 [194]) and task oriented datasets (COCO-Search18 target-present [78] and target-absent [389]).

cognitive processes [35, 159]. Capturing such variability is essential for developing models that accurately reflect the diverse range of human traits. However, existing scanpath prediction models tend to align closely with the statistical mean of human gaze behavior. This limitation mainly stems from how the task is formulated. Most approaches treat scanpath prediction as a supervised regression or sequence prediction problem, minimizing pointwise distances between predicted and ground-truth fixations.

In practice, this is often implemented with simple regression losses such as the ℓ_1 loss or mean squared error (MSE) loss, applied to fixation coordinates, thus encouraging predictions that approximate the mean of the data distribution. While this approach may improve performance on traditional evaluation metrics, it does not reflect the natural variability in human visual attention. Commonly used metrics such as MM, SM, and SS tend to reward predictions that closely match an aggregated ground truth, thus favoring models that generate a single representative scanpath [212]. This is clear in several works [74, 76, 256] where scanpath models surpass human consistency. Indeed, the average similarity between ground-truth scanpaths can be smaller than the average similarity between generated scanpaths if they well reflect an average behavior.

Building upon these considerations, we present a first attempt to quantitatively assess the ability of a model to generate diverse, yet human-like, gaze tra-

jectories. Specifically, for this study, we adopt the DSS and RSS metrics introduced in Section 3.2.4. Table 3.13 reports the results on free-viewing and visual search. ScanDiff achieves the best overall performance on all settings and datasets, highlighting its effectiveness in predicting accurate eye movement trajectories well aligned with the human scanpath variability. Interestingly, the performance gap between ScanDiff and state-of-the-art methods becomes even more evident in the visual search task and is further supported by qualitative results in the appendix (see Section B.2). Goal-oriented scanpaths tend to be more deterministic [66, 396], particularly in the target-present setting, and are generally shorter than those in free-viewing scenarios. Nevertheless, our model effectively captures even the more subtle variability present in human gaze behavior.

3.2.6 DISCUSSION AND LIMITATIONS

While ScanDiff demonstrate state-of-the-art performance in human attention modeling, it still faces some limitations. First, as a diffusion-based generative model, it incurs higher computational complexity and inference time compared to more deterministic scanpath prediction approaches, which may limit its suitability for real-time applications. This limitation can be attributed to the use of a standard DDPM sampling strategy, where iterative denoising steps are required to generate each scanpath. Note that this aspect is not the primary focus of the work, which instead emphasizes the modeling of scanpath variability rather than computational efficiency. Nevertheless, reliance on a basic DDPM sampling strategy may be suboptimal, and more advanced diffusion acceleration techniques such as improved samplers [321], fewer-step generation schemes [243, 304], or distillation-based approaches [322] could significantly reduce inference cost without compromising generative quality. Second, while ScanDiff captures variability at the population level, it does not explicitly account for individual differences, as it is not conditioned on user identity. This hinders its ability to reproduce consistent subject-specific viewing strategies and its applicability to personalized applications. However, this limitation could be addressed through relatively straightforward extensions such as user embeddings, subject-conditioned

diffusion mechanisms, or other personalization mechanisms that incorporate observer identity and gaze data. Third, the model operates on discrete scanpath representations without explicitly modeling continuous gaze trajectories, thereby omitting fine-grained oculomotor dynamics such as saccadic motion and velocity profiles. Recent work [408] suggests that jointly modeling discrete scanpaths and continuous trajectories within a unified diffusion process can improve both realism and performance, indicating a natural direction for extending the ScanDiff framework toward more physiologically realistic gaze modeling. Finally, the model is sensitive to the quality and diversity of the training data. Since it learns the underlying distribution of human scanpaths from annotated eye-tracking datasets, any biases or noise in the training data are directly reflected in the learned generative process. This can lead to reduced generalization and an overemphasis on dominant viewing strategies, while underrepresenting rare but meaningful behavioral patterns.

Despite these limitations, ScanDiff represents an important step forward in the modeling of human visual attention, contributing to a broader shift in the field toward probabilistic and diversity-aware approaches. This focus on variability is increasingly becoming a central aspect of recent research in scanpath prediction and visual attention modeling, where accounting for inter- and intra-observer variability is recognized as essential for more realistic and generalizable models.

4

Integration of Human Attention in AI Models: A Review

HUMAN attention modeling has proven, in recent years, to be particularly valuable not only for understanding the cognitive processes underlying visual exploration, but also for providing support to artificial intelligence models designed to solve complex problems across multiple areas, bridging cognitive science and advanced computational approaches.

Research in this area has shown a growing interest in leveraging human attention signals, derived either from eye-tracking data or computational attention models, to address challenges in both established and emerging tasks. An intriguing development is the integration of human attention principles beyond visual stimuli. This expansion encompasses tasks ranging from vision-only to those involving text, where attention modeling incorporates text gaze patterns.

The goal of this chapter is to provide a structured overview of recent efforts to integrate human attention mechanisms into contemporary deep learning ar-

This Chapter discusses topics from the following paper: “G. Cartella *et al.*, Trends, Applications, and Challenges in Human Attention Modelling, IJCAI 2024” [60].

Models	Input				Human Attention		
	I	V	A	T	Type	Real	Task
Wang et al. [362]	✓				Fixation points	✓	Visual Recognition
Min and Corso [253]	✓				Fixation points	✓	Visual Recognition
Qiao et al. [281]	✓	✓			Saliency maps	✓	Visual Recognition
Bylinskii et al. [57]	✓				Saliency maps	✓	Graphic Design
Fosco et al. [133]	✓				Saliency maps	✓	Graphic Design
Leiva et al. [222]	✓				Saliency maps	✓	Graphic Design
Jiang et al. [185]	✓				Saliency maps	✓	Image Manipulation
Aberman et al. [1]	✓				Saliency maps		Image Manipulation
Miangoleh et al. [252]	✓				Saliency maps		Image Manipulation
Yang et al. [383]	✓				Saliency maps		IQA
Zhu et al. [414]	✓				Saliency maps		IQA
Sui et al. [330]	✓				Scanpaths		IQA
Sugano and Bulling [329]	✓		✓		Fixation points	✓	Captioning
Yu et al. [393]		✓	✓		Saliency maps		Captioning
Cornia et al. [90]	✓		✓		Saliency maps		Captioning
Qiao et al. [282]	✓		✓		Saliency maps		VQA
Sood et al. [325]	✓		✓		Saliency maps		VQA
Ilaslan et al. [171]		✓	✓		Saliency maps	✓	VQA
Berzak et al. [32]			✓		Fixation points	✓	Reading Comprehension
Reich et al. [292]			✓		Scanpaths	✓	Reading Comprehension
Skerath et al. [318]			✓		Fixation points	✓	Reading Comprehension
Sood et al. [327]			✓		Saliency maps		Language Understanding
Khurana et al. [200]			✓		Scanpaths		Language Understanding
Deng et al. [111]			✓		Scanpaths		Language Understanding

Table 4.1: A summary of representative works incorporating human attention across various tasks. We outline the adopted modalities (*I*: Image, *V*: Video, *A*: Audio, *T*: Text), how human attention is processed (via *saliency maps*, full *scanpaths*, or raw *fixation points*), and whether the model employs *real* human attention data from eye-trackers or predictions from computational models.

chitectures for image, text, and multimodal data. The proposed taxonomy organizes existing works according to the application type, distinguishing among image and video processing, vision-and-language tasks, and language modeling. Additionally, we also emphasize applications in specific domains that present extensive literature, such as robotics, autonomous driving, and medicine.

Fig. 4.1 depicts some representative examples of how human attention can be integrated as a supervisory signal across different tasks, while Table 4.1 summa-

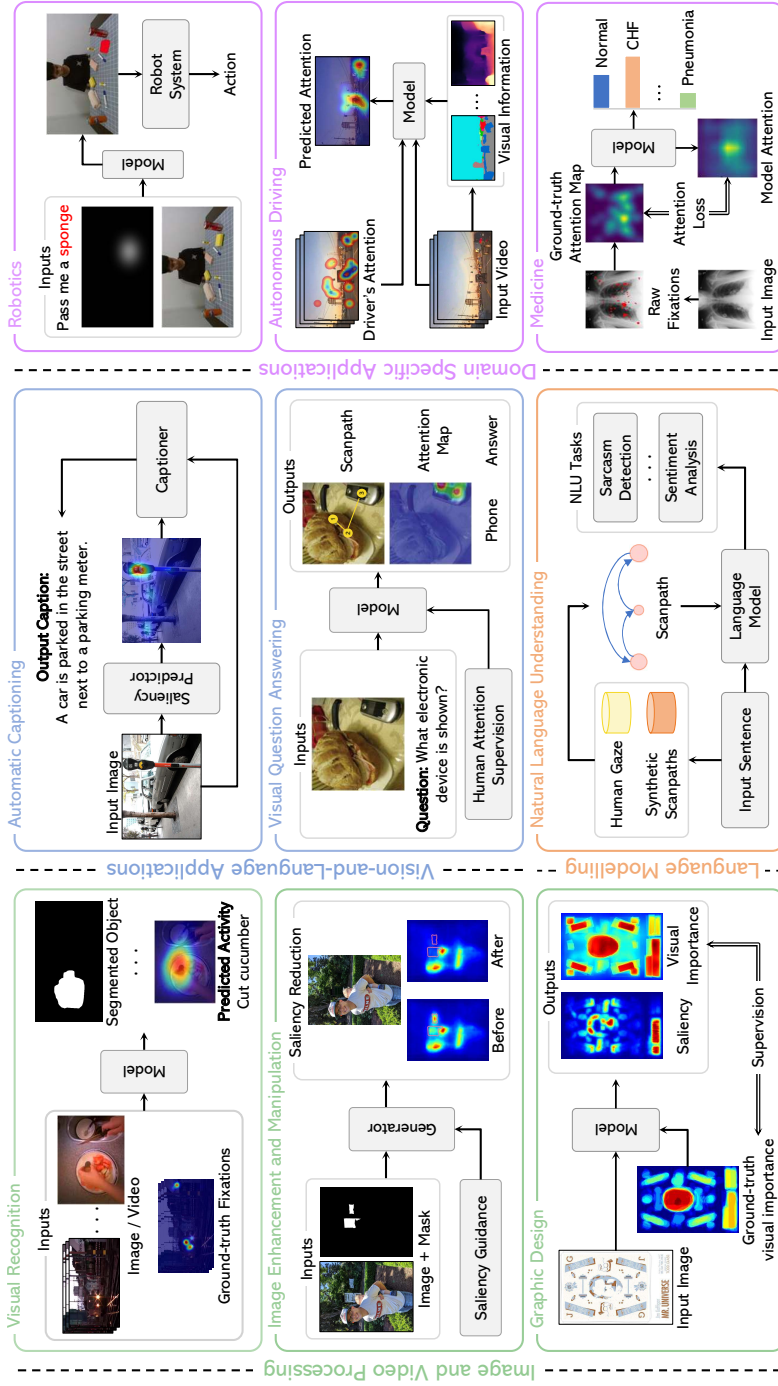


Figure 4.1: An overview of sample architectures integrating human visual attention with different input and output modalities. Human visual attention has been employed to solve tasks in diverse domains spanning from image and video processing, automatic captioning, visual question answering, and language understanding, as well as robotics, autonomous driving, and medicine.

rizes the proposed categorization. Specifically, we examine how gaze information, such as scanpaths or saliency maps, can be incorporated to enhance model performance in various settings. For instance, in image and video processing, such integration leads to better object recognition, scene understanding, and overall visual comprehension. In the field of vision-and-language applications (*e.g.* automatic captioning or visual question answering), incorporating gaze data enhances the contextual awareness of the models, allowing for a more refined interpretation of the relationship between visual and textual information. Similarly, the injection of gaze signals into language modeling contributes to more contextually informed predictions, improving the model's proficiency in reading comprehension and language understanding.

To the best of our knowledge, the perspective of the interplay between deep learning-based applications and visual attention is lacking in most recent reviews of the field. For an updated and curated overview of ongoing research in this area, we refer the reader to our dedicated repository*. The following sections explore approaches benefiting from human attention adoption, offering insights into its broad applicability.

4.1 IMAGE AND VIDEO PROCESSING

4.1.1 VISUAL RECOGNITION

Human attention modeling has been successfully adopted in a variety of computer vision tasks such as object detection [131], segmentation [362], and activity recognition [253]. The recorded human gaze data assists AI models in detecting salient objects in videos, segmenting videos without supervision, and recognizing activities in egocentric video sequences. In [281], human visual attention is employed in a multi-task learning framework for audio–visual saliency prediction and sound source localization, proving that attention is a valuable guide to pinpoint the origin of sounds in videos. Other notable examples relate to the concept of foveated processing, *i.e.* prioritizing computations around a focal point

*<https://github.com/aimagelab/awesome-human-visual-attention>

as in the human visual system. Foveated processing has been recently introduced as a novel structural component of modern deep learning architectures, delivering significant inference speed-ups compared to classic convolutional networks and competitive results on different image and video recognition tasks [343]. Interestingly enough, the concept of foveated processing has been applied to other sub-fields of computer vision, such as volume rendering [29] and image super-resolution [355].

4.1.2 GRAPHIC DESIGN

One compelling application of human attention modeling is related to the automatic evaluation of design choices of websites, data visualizations and infographics, mobile UIs, or advertisements and posters [57, 184, 222]. Visual attention allocation in graphic designs can be interpreted as a proxy for the perceived relative importance of design elements. Being able to effectively and automatically perform such an assessment paves the way to a spectrum of downstream applications, including automatic image retargeting and thumbnailing [57] or interactive design [133]. In the context of Mobile User Interfaces, Schoop et al. [310] proposed a model to predict the perceived tappability of mobile UI elements. Unlike natural images, graphic designs try to direct attention using a mixture of captivating text fonts and animations. Chakraborty et al. [70] defined a two-stage approach to model the spatio-temporal attention on graphics design documents during free-viewing. Recent approaches, while keeping the neural architectural structure of saliency models, are trained to predict *importance maps*. Adopted datasets and ground-truth importance maps are collected either via BubbleView mouse clicks [57] or by manual labeling of important regions [133].

4.1.3 IMAGE ENHANCEMENT AND MANIPULATION

Visual attention allocation models, particularly saliency models, have shown effectiveness in various tasks aimed at manipulating images to enhance the attractiveness of specific objects therein. Such tasks, sometimes referred to as attention retargeting or re-attentionizing, aim at automatically enhancing (or weak-

ening) some portions of the original stimulus, thus reducing visual distraction and highlighting the desired object. In [141], this is achieved by training a convolutional neural network to transform images so as to satisfy a given fixation distribution eventually predicted by a saliency model. Target fixation distribution is obtained by either increasing or decreasing the saliency of particular objects or globally scaling the saliency map. Similarly, in [185], the authors proposed a saliency-guided image translation model, named SalG-GAN, designed for image-to-image translation based on user-specified saliency maps. SalG-GAN employs a disentangled representation framework to encourage diverse translations for the same saliency and introduces a saliency-based attention module to enhance the key components.

More recently, Aberman et al. [1] proposed a method to reduce distractions in images by utilizing a saliency model. By inputting an image and a mask specifying the editing region, the approach employs backpropagation to parameterize a differentiable editing operator, diminishing saliency in the edited area. Various operators are defined, encompassing a recoloring operator for seamless integration of distractors, a warping operator to gradually eliminate objects, and a generative adversarial network operator to substitute regions with less salient alternatives using a semantic prior. Similarly, in [252], a realistic saliency-guided image enhancement method was proposed to maintain high manipulation realism while attenuating distractors and amplifying objects of interest. To this end, a realism loss for saliency-guided image enhancement was designed for training the system to optimize the saliency of a region while penalizing for deviations from realism.

4.1.4 IMAGE QUALITY ASSESSMENT (IQA)

IQA involves using computational models to measure image quality in alignment with subjective human evaluations. As significant distortions can impact visual attention allocation on a given stimulus, saliency and scanpath models have been adopted to increase the effectiveness of perceptual IQA methods.

In [82], a CNN-based no-reference image quality metric was proposed. Initially, the image degradation type is identified via a dedicated CNN, and the cor-

responding saliency map is computed. The latter is employed to choose patches based on their perceptual relevance. A different CNN model is subsequently employed to assess the quality of each patch, and the overall quality is determined by averaging the obtained scores. Similarly, Jia and Zhang [183] used a saliency map to assign an importance value to a set of patches extracted from the original image. A CNN model is then applied to the subset of most salient patches to predict a local quality score; the final quality score is obtained by averaging.

Yang et al. [383] proposed SGDNet, an end-to-end deep neural network for no-reference IQA. It optimizes visual saliency and image quality prediction by employing a multi-task learning framework. The authors highlight how this approach can leverage the correlation between saliency information and image quality, utilizing saliency maps and quality scores during training. Handling saliency prediction as a sub-task enhances feature fusion, thus producing more perceptually consistent outcomes. Zhu et al. [414] introduced TranSLA, a saliency-guided Transformer network for no-reference IQA. TranSLA jointly learns to predict mean opinion scores on a given image and the corresponding saliency map. The latter is used as the query value of the multi-head self-attention mechanism of a Transformer architecture, while the key-value pair is provided by the input image feature representation from a ResNet-based convolutional network. Eventually, this forces the model to focus on the most salient regions, employing a pre-trained saliency model to provide ground-truth saliency maps.

Recent studies have demonstrated that viewing behaviors (*i.e.* scanpaths) are useful for perceptual IQA on 360° images and videos, where eventual distortions are localized in space and time. In [330], a scanpath prediction model on 360° images is employed to sample a sequence of rectilinear projections of viewpoints (*i.e.* frames) along different generated scanpaths. Quality scores are then computed on each frame. Video-level quality scores are obtained by temporally pooling the frame-level quality scores. The proposed approach shows better performance compared to naively applying IQA to equirectangular projections.

4.2 VISION-AND-LANGUAGE APPLICATIONS

4.2.1 AUTOMATIC CAPTIONING

Within the domain of vision-and-language, automatic captioning has gained significant attention as an effective way of generating natural language descriptions for both images and videos. Captioning models are designed not only to identify all pertinent objects and entities within a scene but also to generate coherent and grammatically correct sentences describing them. While advancements in this field are promising, some works have explored the possibility of integrating human gaze information to enhance the generation process of textual descriptions.

The first attempt in this direction was presented by Sugano and Bulling [329], in which fixation points are used as an additional input to guide the caption generation of a recurrent-based image captioning model. In particular, they integrated gaze information, represented in the form of fixation histograms, at each time step of the recurrent neural network, where fixated and non-fixated regions are combined to produce the output sentence. This approach is limited by the need to have both eye fixation and caption annotations for the same visual input. As a solution, Cornia et al. [90] shifted to the use of saliency maps predicted by a pre-trained visual saliency model instead of employing raw fixation points, thus being able to apply the proposed approach to any input image and potentially extend training to any publicly available captioning dataset. Moreover, given that humans tend to pay attention not only to the salient regions but also to the surrounding context when describing an image, they also proposed to let the model learn how to weigh the contribution of both salient and contextual visual features during the generation of the caption.

On a similar line, other works [393] employed predicted saliency maps in conjunction with additive attention mechanisms to effectively capture all semantic cues typically described in a textual sentence. While previous methods take images as input, Yu et al. [393] proposed to use saliency information directly predicted inside the model for describing video sequences. He et al. [158], instead, extracted fixated locations from predicted saliency maps via the winner-take-all

approach and used them to obtain a set of visual features, each corresponding to the spatial image region centered on a fixated location, that is given as input to the language model. Additionally, they introduced a novel dataset for the task composed of 4,000 images with corresponding eye movements and verbal descriptions recorded synchronously over images. Similar annotations were employed in the model proposed by Takmaz et al. [333], in which a saliency map is extracted for each word of the caption, taking advantage of the alignment between eye fixations and recorded audio traces with image descriptions. This sequence of saliency maps is used as an additional input for the captioning model, using the saliency map corresponding to the word to be predicted at each time step of the generation. A related research area is personalized image captioning [85, 240, 373], where the aim is to predict a caption given a user context. However, prior approaches overlook human attention. On this line, Xue et al. [377] proposed for the first time to incorporate personalized human attention as an input signal for image captioning.

4.2.2 VISUAL QUESTION ANSWERING (VQA)

In the context of vision-and-language research, there has been a growing emphasis on integrating human attention to enhance the performance of VQA models, which aim to effectively answer questions based on a given image or video sequence. In this context, Das et al. [107] conducted a preliminary investigation and introduced the VQA-HAT dataset, comprising image-question pairs and corresponding attention maps collected through a mouse-based procedure. Their evaluation of attention maps from state-of-the-art VQA models against human attention revealed a discrepancy between machine-generated and human attention, highlighting the need for enhanced attention mechanisms. A subsequent work [282] employed the VQA-HAT dataset to address the scarcity of gaze-based data in VQA. Specifically, this study proposed a network capable of generating human-like attention maps, which are employed to enhance attention-based VQA models with human-like attention supervision, resulting in significant improvements. Chen et al. [71] continued in this direction by leveraging

human attention to promote reasoning behavior in a VQA model and collected an eye-tracking dataset to enhance the understanding of human attention during the visual reasoning process. In a different vein, Sood et al. [326] presented the VQA-MHUG dataset, focusing on multimodal human gaze on both images and questions in VQA. Their analysis revealed a significant correlation between human attention on text and VQA performance across various models, underscoring the importance of text attention in understanding visual scenes. Subsequent research [325] adopted this dataset and proposed a VQA model that integrates human-like attention on both image and text during training, showing competitive performance against previous methods. While all the aforementioned approaches were designed to work on images, Ilaslan et al. [171] shifted towards video question answering and introduced a task-oriented VQA scenario for collaborative tasks. Their contribution involved the creation of a dataset with human gaze data collected for first-person view videos and the introduction of a novel approach for task-specific question answering, emphasizing the importance of gaze information in understanding user intent during collaborative tasks.

4.3 LANGUAGE MODELLING

4.3.1 MACHINE READING COMPREHENSION

Advancements in machine reading comprehension, particularly in predicting native language from gaze patterns, have been a focal point in recent natural language processing research. Berzak et al. [32] introduced an innovative methodology for studying cross-linguistic influence by analyzing eye movement patterns in second-language reading. The proposed framework demonstrated the predictive power of gaze fixations, revealing a systematic influence of native language properties on reading in a second language. Building on this, Skerath et al. [318] conducted a study to validate the identification of native language from gaze records. By utilizing a new corpus with a larger set of native languages, they reinforced the robustness of the approach, showcasing comparable classification performance with reduced data. Reich et al. [292], instead, designed the first

LSTM-based model capable of inferring reading comprehension, text difficulty, and the reader's native language from scanpaths in reading. Their model, which incorporates lexical, semantic, and linguistic features, treats eye movements as an independent variable, highlighting the potential to infer reader properties comprehensively.

4.3.2 NATURAL LANGUAGE UNDERSTANDING

The integration of human attention into natural language understanding tasks has recently emerged as a promising avenue for improving the performance of various models. In this context, the work by Sood et al. [327] stood out for adopting human gaze information into neural attention mechanisms for natural language processing tasks. Faced with data scarcity challenges, the authors proposed a novel hybrid text saliency model, combining a cognitive model of reading behavior with human gaze supervision. The proposed model was further integrated into an attention layer, demonstrating substantial performance gains in paraphrase generation and sentence compression tasks.

Khurana et al. [200] contributed to this field with the introduction of a model capable of generating scanpaths from existing eye-tracking corpora. The generated scanpaths proved beneficial for various tasks, such as sentiment classification, paraphrase detection, and sarcasm detection. Following this direction, Deng et al. [111] leveraged synthetic gaze data and developed a model that combined scanpath generation with a BERT-based language model, eliminating the need for human gaze data. The proposed approach outperforms the underlying language model in multiple tasks, especially in low-resource settings, showcasing the potential of synthetic eye-gaze data.

On a different line, Ren and Xiong [293] tackled the problem of shortcut learning in large language models and proposed a framework that incorporates human attention to encourage LLM-based task-specific models to focus on relevant features. This human attention-guided approach effectively identified shortcut tokens and improved the robustness of LLM-based models, even on out-of-distribution samples.

4.4 DOMAIN-SPECIFIC APPLICATIONS

4.4.1 ROBOTICS

Human-robot interaction has witnessed a paradigm shift towards the integration of socially intelligent robots into various facets of human life. When it comes to patients with motor disabilities, assistive robotics makes a fundamental contribution to daily tasks, making the surrounding environment more inclusive. In this context, the human gaze represents a fundamental nonverbal communication channel used to convey our intentions, goals, or preferences, while offering a natural and non-invasive human-machine interface. Robot assistance can be enabled by individuals directing their gaze towards a specific target they wish to grasp or manipulate. As emphasized in [280], such an approach allows humans to convey their object-referring intentions without ambiguity.

In the specific domain of collaborative robotics, effective interaction and mutual understanding between humans and robots are essential for the successful accomplishment of a given task. In early teleoperation applications, robots were controlled through joystick input signals. However, this control mode poses inherent complexities and demands substantial human expertise. Furthermore, joystick inputs constrain the capacity to anticipate user intentions by conveying only incremental instructions, thus limiting the system's awareness of the ultimate goal. In response to these limitations, researchers have embarked on the development of shared control systems able to accommodate eye-gaze input modalities, thereby diverging from the exclusive reliance on joystick signals [12–14].

Another fundamental area of robotics is learning from demonstrations. It refers to the process of a robot acquiring knowledge through the imitation of an expert. Saran et al. [306] characterized the gaze patterns of human instructors during task demonstrations and observed a tendency among users to disregard task-irrelevant objects. Their work proved that integrating gaze information with inverse reinforcement learning enhances the robot's ability to learn reward structures. Huang et al. [168] attempted to model robot gaze as a feedback signal intended to communicate with humans and enhance the efficacy of human

instructions and the overall quality of the learning process.

Beyond human-robot interaction, gaze has also been integrated into robotics computational models not directly involving humans. An interesting example is aerial robotics, which has recently benefited from the modeling of human visual attention [105]. The problem has been extended to multiple discontinuous images, and recent studies exploited both spatial [84] and temporal information [167] for saliency prediction.

4.4.2 AUTONOMOUS DRIVING

Developing intelligent agents able to perceive the environment and autonomously perform actions is a primary objective of AI research. In this context, the design of autonomous vehicles is one of the most striking examples. The human visual attention system is unique in this respect. Human beings have the innate ability to rapidly discern the most relevant stimuli, filter out unnecessary information, and, while driving, promptly recognize and preempt potential risks. Transferring these inherent human abilities onto autonomous vehicles constitutes a non-trivial challenge that has attracted the interest of the research community.

Palazzi et al. [268] pioneered a significant endeavor in the field by introducing one of the earliest attempts, in the deep learning era, to model driver's attention. With this aim, they presented the first publicly available dataset, named DR(eye)VE, which features in-car recordings under different traffic and weather conditions, as well as varied landscapes. Accurately modeling the behavior of autonomous agents in critical circumstances, such as collisions and accidents, represents a major concern for the safety and security of passengers. Nonetheless, collisions are quite rare, and the in-car protocol delineated in [268] makes the modeling of such events nearly impossible due to the scarcity of data.

As a consequence, Xia et al. [371] and Bae et al. [22] presented an in-lab data gathering protocol to simulate accident-prone scenarios, involving collisions, near-collisions, and braking events. These efforts resulted in the creation of the BDD-A and EyeCar datasets, respectively. The model proposed in [371] demon-

strated generalization to real-world scenarios, even when trained on in-lab data only. Baee et al. [22] presented a model based on an inverse reinforcement learning schema that predicts an attention policy treating each fixation point as a potential source of reward. An additional distinguishing feature of the human visual system is peripheral vision. It denotes the capacity to focus our attention on one object while concurrently attending to another. Pal et al. [267] took into consideration this particular capability and introduced a gaze detection model augmented by the semantics of the scene, comprising the most relevant objects to the driving task, as well as distances and speed. Classical driver's attention prediction models typically predict saliency, neglecting long-term cumulative information, such as a driver acquiring experience over time. To this aim, Chen et al. [77] integrated a feedback loop mechanism within a CNN-Transformer-based architecture. The feedback loop module gets the output from the decoder and iteratively generates incremental knowledge.

4.4.3 MEDICINE

In recent years, the advancement of sophisticated deep-learning techniques has contributed to the field of medical image analysis, consequently leading to the success of computer-aided diagnosis techniques. Conventional methodologies usually employ convolutional or Transformer-based architectures to learn the textural pattern of diseases and the spatial relationships among image patches.

Medical images exhibit intricate anatomical structures. This poses challenges for a deep network, uniquely based on its inner representations, to precisely capture diagnostically relevant regions. Moreover, conventional models lack the required domain knowledge to correctly diagnose a disease in a more interpretable manner, while avoiding spurious correlations [245]. In the medical field, practitioners often rely on visual information, and the outcome of a diagnosis is closely linked to the observation and interpretation of a medical image. Here, human gaze emerges as a natural way to capture visual attention during the diagnosis process and motivates recent research towards the implementation of human attention-augmented architectures for the healthcare domain. Advancements

in this direction have been facilitated by the introduction of medical datasets that incorporate eye-gaze data obtained from human experts during the diagnostic process. Karargyris et al. [198] released a dataset of medical images equipped with raw gaze annotations provided by radiologists, complemented by bounding boxes marking the discriminative areas. The authors highlighted the significance of acquiring fine-grained data extending beyond basic labels representing disease classes. Indeed, more detailed annotations lead to better localization capabilities, thereby enhancing interpretability, an essential requirement for high-risk applications.

Building upon this contribution, several models trained on gaze medical data have been developed. From an architectural point of view, two principal research directions can be identified. The first entails the conversion of raw gaze data into a visual attention map, employing it as a source of supervision to enhance the alignment of model attention with human attention. Bhattacharya et al. [33] proposed a teacher-student framework in conjunction with a visual attention loss, guiding the student attention towards regions identified by the attention map generated by the teacher, that was trained on the ground-truth visual search patterns. Concurrently, Wang et al. [358] defined an attention-consistency module, suppressing inter-observer variability. The second line of research, instead, is based on a two-stream architecture. The input image and the related attention map are typically processed separately before being concatenated to predict the final output. In particular, Ma et al. [245] proposed a gaze-guided Vision Transformer that receives image patches as input, appropriately masked to exclude irrelevant regions for the diagnosis.

In contrast to the previously outlined research directions, Wang et al. [354] proposed for the first time to directly embed raw gaze data into a graph neural network for X-ray classification. This architecture avoids the adoption of visual attention maps as commonly defined in more classical models.

4.5 FROM GAZE-GUIDED AI TO GAZE-GUIDED UNDERSTANDING

The literature reviewed in this chapter demonstrates that gaze information can support AI systems across a wide range of tasks, including image and video processing, vision-language learning, language understanding, medicine, robotics and autonomous driving. Despite the diversity of applications, a common principle emerges: gaze provides access to information about human attentional priorities that is often unavailable from the primary input modality alone (*e.g.* image, video, or text). As a consequence, gaze can serve not only as a supervisory signal for improving machine performance, but also as a proxy for the perceptual, cognitive, and behavioral processes underlying human decision making. The taxonomy presented in this chapter can be naturally extended to three complementary research directions. First, gaze can guide machine perception by highlighting regions that humans consider informative for a visual task. Second, gaze can be analyzed directly to characterize cognitive and behavioral processes, particularly in clinical and developmental settings. Third, gaze can be used to study human interaction with complex visual and cultural environments, providing insights that extend beyond conventional computer vision benchmarks. These directions motivate the contributions presented in the following chapter.

5

Gaze-Guided Learning for Vision and Behavioral Understanding

HUMAN gaze and eye movements provide a rich, real-time window into visual attention, offering insights that span perceptual, cognitive, and behavioral processes. Over the past decade, the integration of gaze information into computational models has proven valuable not only for advancing our understanding of visual attention but also for informing downstream applications that leverage human perceptual strategies.

In this chapter, we explore a series of contributions that harness gaze and eye tracking data to address diverse and societally relevant tasks, ranging from media integrity and clinical assessment to cultural heritage and environmental attention

This Chapter discusses topics from the following papers: “G. Cartella *et al.*, Unveiling the Truth: Exploring Human Gaze Patterns in Fake Images, IEEE SPL 2024” [61], “G. Cartella *et al.*, Pixels of Faith: Exploiting Visual Saliency to Detect Religious Image Manipulation, ECCV Workshops 2024” [62], “G. Cartella *et al.*, Unravelling Neurodivergent Gaze Behaviour through Visual Attention Causal Graphs, ICIAP 2025” [65], and “G. Cartella *et al.*, Sanctuaria-Gaze: A Multimodal Egocentric Dataset for Human Attention Analysis in Religious Sites, ACM JOCCH 2025” [63].

analysis. The first contribution concerns the use of gaze patterns to support media authenticity and manipulation detection, a visual analysis task that extends the image and video processing category reviewed in the previous chapter.

As generative models increasingly produce highly realistic synthetic imagery, traditional detection methods face growing challenges. Section 5.1 presents our contribution on fake image perception. Our research provides empirical evidence of distinctive viewing behaviors that differentiate real from altered content, offering a human-centric perspective on deepfake detection that complements algorithmic approaches (see Section 5.1.2). Following this research line, in Section 5.1.3, we further investigate how human visual saliency can be incorporated into machine learning frameworks to improve detection of religious image manipulation, aligning model attention with perceptually relevant image regions and emphasizing the preservation of culturally significant content.

Beyond media forensics, the second contribution focuses on neurodevelopmental disorders and relates to the medical and cognitive applications of gaze, with implications in healthcare and behavioral science. In this case, gaze is no longer used to improve machine perception but is itself the primary object of analysis, serving as a behavioral biomarker capable of revealing underlying cognitive mechanisms. In Section 5.2, we present a study on neurodivergent visual attention. In particular, we model the causal mechanisms underlying gaze allocation in individuals with and without developmental conditions. This work highlights the potential of gaze-based modeling to support objective assessment in neurodevelopmental research.

Finally, understanding how individuals allocate attention in real-world environments requires large, ecologically valid datasets. To this end, the last section of this chapter introduces Sanctuaria-Gaze, a multimodal egocentric dataset collected from visits to architecturally and culturally significant religious sites. Captured with wearable devices that synchronize video, gaze coordinates, head motion, and spatial data, this dataset enables detailed analysis of attention allocation in dynamic, real-life contexts, and it opens new avenues for interdisciplinary research in human behavior and environmental interaction. Together, these contributions span three distinct yet interconnected roles of gaze: *perceptual guid-*

ance, cognitive characterization, and environmental interaction. They provide a practical validation of the broader framework introduced in the previous chapter while illustrating the versatility of gaze-guided learning across more specialized domains.

5.1 GAZE PATTERNS AS CUES FOR DEEPPFAKE DETECTION

5

Creating high-quality and realistic content is now possible thanks to impressive advances in generative artificial intelligence. A description in natural language of your desired output is all you need to obtain breathtaking results.

In recent years, the spread of generative models has significantly transformed various aspects of everyday activities. Generative models are adopted at a steady pace for realistic content creation [199, 290, 300, 303], including images [161, 298, 300], text [58], and audio [206] and to enhance the perception of visual quality [151, 390]. Such accessibility fosters innovation and increases production in fields such as design, entertainment, and education.

Such advancements are the direct consequence of the development of new deep learning architectures based on diffusion models [115, 161, 261, 298, 320, 321], which have marked a new era in the field of generative modeling. In addition, the training of increasingly large deep networks can be empowered by the availability of a massive volume of synthetic data. However, the widespread adoption of generative models also introduces significant challenges regarding the propagation of misinformation and malicious content. The ability of these systems to synthesize highly convincing yet entirely fabricated images, videos, and audio clips has raised concerns about their potential misuse. Synthetic media can be exploited to manipulate public opinion, interfere with political processes, and distort narratives on social media platforms. These risks highlight the need for models capable of distinguishing genuine content from synthetic data.

5.1.1 BACKGROUND ON DEEPPAKE DETECTION AND THE ROLE OF HUMAN GAZE

The field of deepfake detection has seen significant advancements in recent years, primarily driven by the continuous evolution of generative techniques [93, 182, 263, 359, 370]. Early efforts in this area focused on detecting GAN-generated content [299, 356, 384]. As image generation methods have evolved to include diffusion models [278, 298], detection techniques expanded accordingly. Most of the computer vision literature focuses on the recognition of fake images and videos, with a particular emphasis on face manipulation [229, 299, 409]. However, recent trends have emerged that extend the recognition to natural images (*i.e.* landscapes, urban scenes, etc.) [9, 27, 129, 241]. This adaptation marked a significant shift, demonstrating the need for versatile detection methods capable of handling various generation techniques.

Typically, state-of-the-art techniques ground the detection of fake samples on the analysis of the generative models' feature space. Various studies [94, 136] demonstrated that images stemming from different generative models present discernible fingerprints left behind by the model during the generation process. Another approach [364] focused on the discrepancies between input images and their reconstructions produced by pre-trained diffusion models. Wang et al. [359] showed that by adopting a proper post-processing and data augmentation pipeline, it is possible to generalize across different GAN-based models. However, the introduction of diffusion models shed light on a major challenge, highlighting how generalization across different families of generative models is still an open research problem [94, 263]. To increase the robustness of deepfake detection methods to unseen generators, recent strategies have tackled the deepfake detection problem by leveraging CLIP [285] to extract visual features for the task [9, 86, 95, 263, 313]. These approaches emphasize the visual patterns identified by the CLIP backbone rather than its semantic text-image alignment properties. Only a few attempts have been made to include human visual attention in the deepfake detection task. Gupta et al. [152] collected FakeET, an eye-tracking dataset collected from 40 participants to investigate human visual perception of

deepfake videos. Their findings reveal that attention patterns in fake videos tend to be more concentrated, whereas real videos elicit a more distributed viewing behavior. In the real case, observers allocate attention not only to the facial region but also to surrounding contextual areas. Boyd et al. [48] proposed a training strategy to address the task by incorporating human visual annotations into a loss function. The acquisition of the training set required subjects to manually annotate regions in the presented images, while providing an answer on which image is either the synthetic or real one, in a two-alternative forced choice manner.

Different from previous research, we tackle deepfake detection by relying on training data directly collected from an eye-tracking experiment and present a novel solution to detect partially manipulated religious images, a domain not previously explored in the literature.

5.1.2 HUMAN VISUAL PERCEPTION OF MANIPULATED IMAGES

From a semantic perspective, as assisted in the field of face manipulation, modifications can be carried out starting from real images and altering only portions of them in order to obtain a result that preserves the original context, making fake detection an increasingly challenging task. Drawing inspiration from the visual attention literature [135, 271, 275], we bring humans into the loop, aiming to exploit their semantic knowledge and generalization skills acquired through evolution and further developed in a lifelong learning process. This study investigates the possibility of including human visual attention as a source of supervision in deepfake detection frameworks. In particular, we seek an answer to the following research question:

Does an underlying attentive pattern exist governing the human visual perception when looking at partially manipulated images compared to genuine ones?

We posit the existence of such a disparity and, to validate this hypothesis, we begin by collecting authentic samples from various existing datasets. For each image under investigation, we produce three distinct altered variants by implementing as many editing methodologies based on state-of-the-art diffusion models [53, 298]. To assess human visual perception of authentic samples versus their

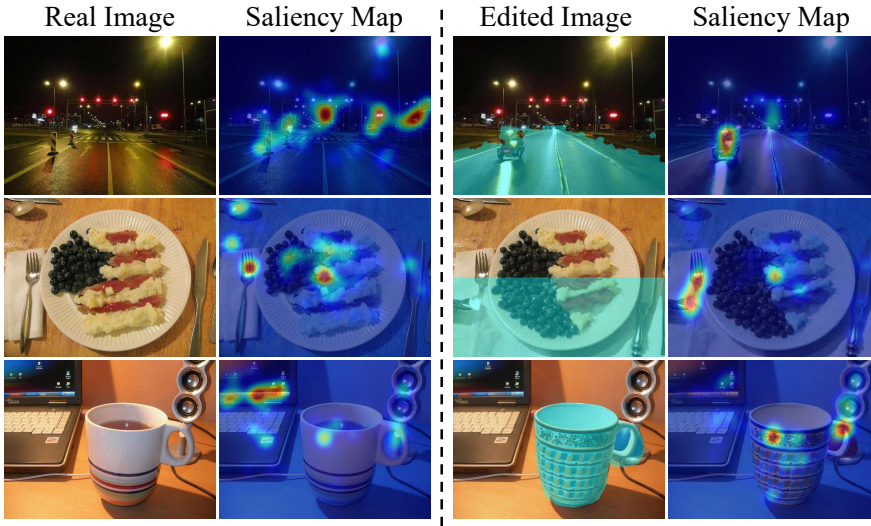


Figure 5.1: Overview of the human gaze patterns when observing real and altered images. Interestingly, humans tend to focus on more circumscribed areas when looking at counterfeit samples. Light-blue masks of edited images represent inpainted regions.

counterfeit equivalents, an eye-tracking experiment is further devised. During this experiment, a sequence of images is presented to the participants, and their ocular movements are recorded as they attempt to distinguish real from fake samples. The collected dataset, including gaze recordings, is publicly released*. A statistical analysis of the collected fixations reveals significant disparities in the viewing patterns that occur when looking at genuine and fake images. In particular, when analyzing the entropy distribution of the acquired saliency maps, we find that fake images elicit high fixation concentrations in specific regions, resulting in lower entropy values when compared to their original counterpart (Fig. 5.1). As a result, we draw the conclusion that humans tend to direct their attention to more delimited areas when perceiving manipulated images. We believe our preliminary results open up further research towards the integration of human gaze information within automatic fake detection pipelines.

*<https://github.com/aimagelab/unveiling-the-truth>

PROPOSED APPROACH

Given a set of real and fake images, our goal is to conduct a statistical analysis to investigate the existence of an underlying pattern governing the visual perception of partially manipulated images, enabling further studies on how the human gaze could improve existing fake detection models. With this aim, and due to the lack of an existing dataset in the literature, we collect images from different sources encompassing a variety of scenes, ranging from indoor to outdoor environments.

DATASET COLLECTION. We define a set of stimuli that covers scenes and environments of varying degrees of complexity. In our dataset, three distinct categories can be identified: *indoor*, *outdoor urban*, and *outdoor natural*. The first two categories feature a plethora of intricate details, a wide set of objects, and a rich diversity of color contrasts. On the opposite, the outdoor natural category is characterized by a lower amount of salient objects, scarce occurrences of high-frequency details, and a prevalence of more uniform color palettes. Such diversity strongly influences how people perceive images [247, 264], thus guaranteeing appropriate data heterogeneity.

Images are extracted from three publicly available datasets, namely COCO [232], ADE20K [413], and LHQ [319]. Since text, faces, and animals are known to be very salient [67, 68, 194], we filter out all the images including these three classes to avoid any possible bias, following recent literature [387]. In the filtering process, we also discard low-resolution images for all datasets, keeping those with a minimum size of 640×480 (or 480×640).

IMAGE EDITING TECHNIQUES. In contrast to prior literature on deepfake detection, which has been focusing on recognizing entirely generated images [9, 94, 263], we concentrate on manipulating images in a subtle manner, preserving the realism, semantics, and context of the original image, thus making the deepfake detection task even more challenging. To this end, three different types of intervention are implemented (see Fig. 5.2).

- **SEMANTIC-AGNOSTIC (SA).** Given a real source image I , we desire to create its fake counterpart $\tilde{I}$. This task can be formally classified as an inpainting problem where the inpainting region is defined as a binary mask M , constructed by

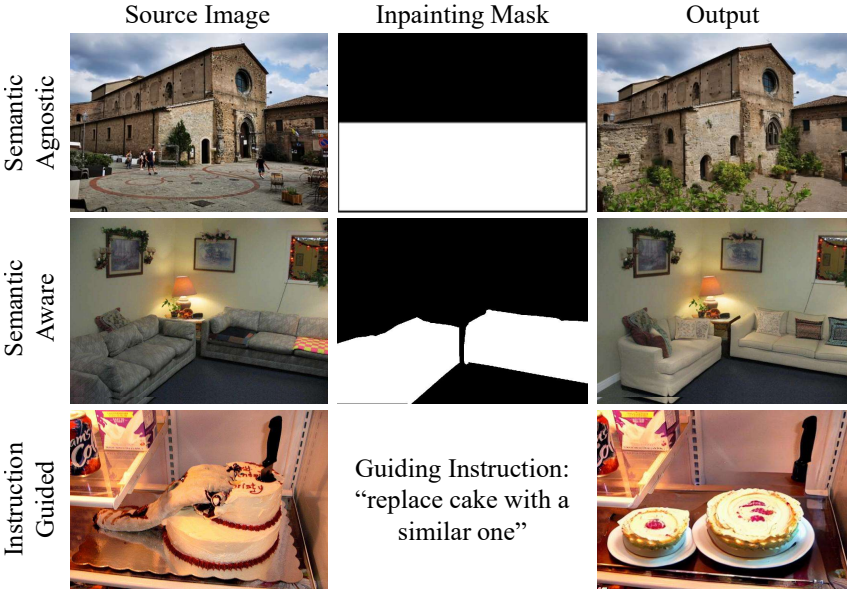


Figure 5.2: Image editing examples where the white masks represent the inpainting regions.

randomly masking half of the image I . Inspired by [2], for each sample, we randomly inpaint one among the following parts: bottom, upper, left, right, upper left diagonal, upper right diagonal, bottom left diagonal, bottom right diagonal, or random patches (until we cover at least 50% of the image). As an inpainting model, we adopt the Stable Diffusion v2.0 inpainting pipeline[†] [298] which we refer to as f_{θ} . To enhance the final output and avoid the generation of undesired objects, we make use of negative prompts N . In our experiments, we qualitatively find that having no guiding caption c as input leads to more realistic outputs. Formally, the complete inpainting process can be defined as $\tilde{I} = f_{\theta}(I, M, c = \emptyset, N)$.

- **SEMANTIC-AWARE (SW).** Although semantic-agnostic manipulation leads to realistic results, there is no control over the generated output. As a different image editing technique, we propose semantic-aware inpainting, where objects present in the scene are substituted with others of the same class (*e.g.* replace

[†]<https://huggingface.co/stabilityai/stable-diffusion-2-inpainting>

a bed with another bed). This choice is driven by the intent to preserve the context and semantics of the scene, and ensure that the generated object well fits the given inpainting mask. With respect to the previous manipulation category, edits affect smaller areas and are more challenging to discern. To define the inpainting region, we start from the image segmentation maps of the images. COCO and ADE20K already provide a segmentation mask for each object O_i and the corresponding textual class label y , while for LHQ, we construct the needed information by making use of the RAM-Grounded-SAM model[‡] [203, 406]. We designate as an inpainting region of the real image I , the binary mask M_{O^*} corresponding to a randomly chosen object O^* among those with an occupancy between 10% and 40% of the total image area. The generation process is performed in the same manner as the *semantic-agnostic*, but to guarantee context preservation, the guiding caption c for the Stable Diffusion model is the textual class label y of the selected object. Negative prompts N are maintained the same. Overall, we obtain $\tilde{I} = f_{\theta}(I, M_{O^*}, c = y, N)$.

- **INSTRUCTION-GUIDED (IG).** Recently, the field of generative artificial intelligence has witnessed the emergence of innovative generative methods, with instruction-guided image editing standing out as a prominent research direction [53, 405]. In our context, we adopt the InstructPix2Pix model[§] [53], able to follow a given editing instruction in natural language to produce the manipulated version of the given sample. In our problem, we feed InstructPix2Pix, referred to as g_{θ} , with the source input I , and the following editing instruction c : `replace {y} with a similar one`. Specifically, y represents the textual class label of an object selected by following the same procedure of the *semantic-aware* editing technique. In our implementation, we use a fine-tuned version of InstructPix2Pix [399], which has been shown to produce better images according to human evaluation. The final synthesized image is defined as $\tilde{I} = g_{\theta}(I, c(y))$.

[‡]<https://github.com/IDEA-Research/Grounded-Segment-Anything>

[§]<https://huggingface.co/vinesmsuic/magicbrush-jul7>

EYE-TRACKING EXPERIMENT. Acquiring eye-tracking data is crucial to understanding the patterns governing the way humans look at real and fake stimuli. With this aim, we set up an eye-tracking experiment involving 20 participants. Each person sits in front of a screen with a resolution of 1920×1080 pixels, equipped with a screen-based eye tracker, at a distance of 68 cm. The screen size is $54 \text{ cm} \times 30.3 \text{ cm}$. To accommodate the inherent variability in the perceptual process of different people, we ensure that the image is viewed by five different observers. Each stimulus is shown for 5 seconds, and the user is instructed to carefully observe the presented image while assessing its authenticity. Following this time-lapse, the stimulus is replaced with a rating screen where the user is required to evaluate the realism by selecting a numerical value on a 5-point Likert scale. The lower end denotes a greater level of confidence in the image's artificial nature, whereas the upper range suggests a strong belief in the image's authenticity. Before presenting the next stimulus, a gray screen featuring a small black cross at the center is exhibited for one second, aimed at engaging the user's attention. To ensure high-quality acquisition, we calibrate the eye-tracker at the beginning of each user session.

Overall, the experiment is based on a total of 400 stimuli, comprising 100 unique genuine samples, uniformly distributed between the three datasets and their respective edited counterparts. For each observer, we randomly select 100 images and, to avoid any bias, we ensure that at most one instance of the same image is shown (*i.e.* if an observer is presented a semantic-agnostic edited image, then no other versions of the same image appear to the same participant).

EXPERIMENTAL ANALYSIS

We conduct an in-depth statistical analysis of the recorded fixations to assess the evidence of a distinguishable pattern in the gaze behavior of individuals when viewing genuine versus counterfeit images. Overall, the analysis is performed on 2,000 recorded human scanpaths.

HUMAN ANNOTATIONS. In the first part of the investigation, we examine how users' ratings are distributed across the images. Fig. 5.3a depicts the histogram of the users' realism perception. In general, the observers are able to distinguish

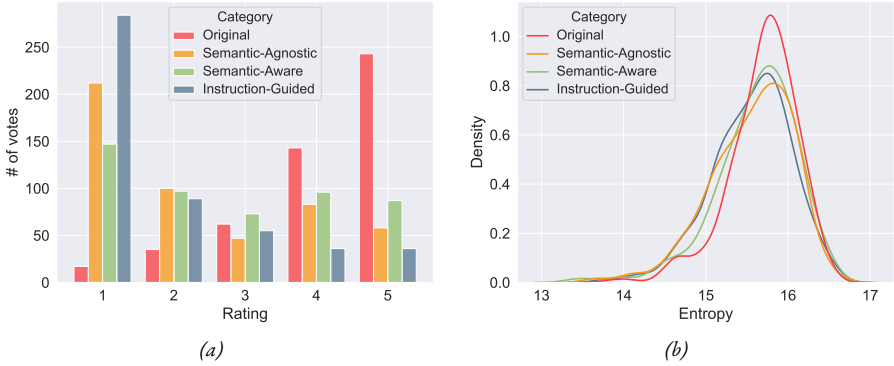


Figure 5.3: Statistical analysis. (a) Histogram of the ratings of realism given by the users in the eye-tracking experiment. (b) Kernel density estimation of the saliency maps' entropy across viewers.

the real nature of most of the images. A considerable number of fake images are given a low rating (*i.e.* 1 or 2), and the majority of genuine samples are correctly identified. Nonetheless, some real images are classified as altered or possibly altered (*i.e.* ratings 1, 2, or 3). Such results can be ascribed to the bias introduced by the requested task or the nature of the image itself (*e.g.* a badly taken picture or out-of-context objects). On the same line, occurrences of improperly categorized edited images (*i.e.* ratings 4 or 5) are observable, meaning that the adopted generative models, in several cases, synthesize highly realistic content. As a final consideration, we point out that the instruction-guided category is the easiest to detect for humans, while the opposite holds for semantic-aware editing, leading to the most realistic outputs. The explanation behind such a result is that the generation through InstructPix2Pix is much more challenging. Indeed, reference inpainting masks are absent, and our constructed textual prompt does not heavily constrain the generation process. As a consequence, such methodology is more prone to the creation of artifacts or out-of-context details. On the contrary, semantic-aware editing represents a more constrained type of intervention, with limited inpainting areas and a textual caption that guides the generation of a specific object that surely fits the context by construction.

SALIENCY ENTROPY. Another important analysis is the one regarding the eye-gaze pattern, enabling the study of the perceptual response of individuals to stim-

uli of different natures. As a first step, we adopt the velocity-threshold fixation identification algorithm proposed in [305] to distinguish between fixations and saccades. The algorithm computes point-to-point velocities and classifies each raw point as fixation or saccade based on a simple velocity threshold.

To evaluate the consistency of human fixations over an image, we measure the entropy of the saliency maps across all the observers. Given a sample image and its recorded fixations, a ground truth saliency map is obtained by convolving a fovea-sized Gaussian kernel over all the fixations locations.

Empirical distributions of the saliency maps' entropy for each of the four considered categories are reported in Fig. 5.3b. In a qualitative assessment, a significant distinction is noticeable between the distributions of real and fake data. Simultaneously, there is evident similarity among the outcomes of the three image editing techniques employed. Particularly, there is a higher population of edited images in correspondence of lower entropy values, while saliency maps of genuine samples exhibit higher entropy.

Our findings bring to light an interesting outcome. While looking at altered images, humans tend to explore less when compared to genuine samples. We attribute this behavior to the inclination of individuals to focus more on specific details when encountering unfamiliar content within the image, in order to enhance their comprehension of the surrounding context. If we consider the proposed task, observers tend to rapidly shift their gaze from one location to another if they do not perceive anything unfamiliar, thus leading to higher degrees of disorders of the saliency maps. Sample saliency maps for both original and corresponding edited images are shown in Fig. 5.1.

STATISTICAL ANALYSIS. Given the real and fake entropy distributions, we further evaluate the results from a quantitative perspective. Specifically, we conduct the two-sample Kolmogorov-Smirnov (K-S) test in order to reveal a possible statistical difference between what is observed with the original images versus all the possible edits. Indeed, this is a non-parametric statistical test that evaluates whether the two sets of data come from the same population. A larger statistic value indicates greater dissimilarity between the distributions. The K-S test is sensitive to differences in the tails of the distributions, making it a good choice to

Category	K-S Test		C-M Test		MWU Test		
	Statistic	<i>p</i>	Statistic	<i>p</i>	Statistic	<i>p</i>	
O	SA	0.154	<.001	2.068	<.001	144409	<.001
	SW	0.106	.007	0.759	.009	136013	.016
	IG	0.166	<.001	2.807	<.001	148184	<.001
SA	SW	0.074	.129	0.387	.078	116543	.064
	IG	0.054	.460	0.127	.468	127714	.552
SW	IG	0.086	.050	0.690	.013	136541	.011

Table 5.1: Statistical tests' results to evaluate the difference between entropy distributions across categories. K-S, C-M and MWU refer to Kolmogorov-Smirnov, Cramér-von Mises and Mann Whitney U tests, respectively. Bolded results refer to significant *p*-values ($\alpha = 0.05$) and higher statistics.

detect even minor discrepancies across the entire range. We corroborate the outcomes of the K-S test by employing the two-sample Cramér-von Mises (C-M) test. This test, akin to the K-S test, quantifies the divergence between the cumulative distribution functions (CDFs) of the two datasets. However, it does not solely concentrate on the maximum difference but instead considers the entirety of the CDF. It computes a test statistic that measures the overall discordance between the distributions, assigning more weight to disparities in the middle of the distributions, rendering it suitable for detecting variances in the central portion of the data.

Finally, evidence of the validity of our findings is obtained through the Mann-Whitney U (MWU) test, to assess the null hypothesis that two samples have the same central tendency. In other words, this test is well-suited for comparing two independent samples when our aim is to establish whether one group typically exhibits greater values than the other. Table 5.1 presents all the statistical results, including both the test statistic and its associated *p*-value. In our analysis, if a *p*-value obtained from a test falls below the significance value $\alpha = 0.05$, it indicates that the observed differences between the two datasets are statistically significant. More specifically, there is strong evidence to reject the null hypothesis and conclude that the two sets of data are different in a meaningful way. On the other hand, if the *p*-value is greater than α , it suggests that the observed differences are

not statistically significant, and we do not have sufficient evidence to reject the null hypothesis. Comparing the entropy distribution of original images against semantic-agnostic, semantic-aware, and instruction-guided editing, the p -value is below the threshold in all cases and for all tests. As a consequence, we reject the null hypothesis that there is no significant difference between the distributions, highlighting that, in terms of entropy, there exists a distinguishable pattern in the perception of real and fake stimuli. By considering the difference between semantic-aware, semantic-agnostic, and instruction-guided editing classes, the null hypothesis is confirmed, meaning that these distributions can be assimilated into the same population of counterfeit images. The only exception stands between the semantic-aware and instruction-guided classes, where for both C-M and MWU tests a statistical difference exists. We argue that the primary reason lies in the output generation quality. As previously discussed, instruction-guided editing is the most easily detectable category as fake, while the opposite holds for the semantic-aware class. However, we are mainly interested in distinguishing between the real and all editing classes. In this case, the statistical difference between original and semantic-aware or between original and instruction-guided still holds.

5.1.3 VISUAL SALIENCY TO DETECT RELIGIOUS IMAGE MANIPULATION

In the context of religious studies, the problem of fake content generation is particularly relevant. Indeed, religious images have been and still are of great significance to billions of people worldwide. The history of faith communities that do not apply aniconic prohibitions shows that religious imagery and iconography have played a fundamental role in the dissemination and transmission of faith, devotion, and knowledge about sacred texts, but also had polemic and propaganda purposes in intra- and inter-religious dynamics. In medieval and modern times, images have been modified mainly with the aim of reproducing sacred images to adapt them to different geographical and cultural contexts or to ameliorate the subjects. In the cases when images have been used in overtly propagandist ways,

supporting one identity, culture, religious, or political party over another, manipulation of the elements obviously opted for pejorative or ridiculing elements.

More recently, the adoption of images in a digital format and the spread of AI changed the way images are manipulated and their effects: deepfakes make it much more difficult to distinguish real images from those created by AI, and they are spread at unprecedented speed, with chances to “go viral” on a global scale in a short time. Such a change has also affected Christian imagery: numerous devotional images produced by AI algorithms have caused controversy as they depict physical defects. These examples have alarmed believers and various Christian Churches around the world because the images, while apparently suitable for private devotion, exhibit subtle traits of blasphemy. Moving from such specific cases, it has been demonstrated that deepfakes, created ad hoc, generate conflicts with high social costs and have an impact at social, religious, and political levels that cannot be ignored. This is particularly relevant when deepfakes are improperly used for targeting specific groups or individuals based on their race, gender, or religion [30]. This form of discrimination conveys prejudices that heighten existing social tensions and further reinforce harmful stereotypes.

To counterbalance these trends and work on accurate historical and cultural knowledge, interdisciplinary teams of scholars can develop control systems based on robust methods for automatic fake recognition. However, the discrepancy between human and algorithmic deepfake detection capabilities presents an intriguing challenge in this field. As revealed in [207], state-of-the-art algorithms often struggle to detect deepfake videos that human observers find obviously fake. This counterintuitive finding suggests that current models may not be capturing the same visual cues that humans instinctively recognize as indicators of manipulation. Several factors could contribute to this phenomenon: (i) humans may rely on subtle contextual or semantic cues that AI models are not attuned to; (ii) algorithms might overfit to specific artifacts rather than learning generalizable features; and (iii) there could be a mismatch between fake images used to train algorithms and those shown to human participants. Moreover, human perception can be influenced by expectations, context, and cognitive biases, while algorithms apply their detection criteria more consistently [246]. This discrepancy reveals

gaps in our understanding of both human and machine perception of deepfakes, highlighting the need for further research to bridge this gap.

The principle of incorporating human supervision can be particularly beneficial in tasks where subtle visual cues are crucial, such as in the detection of deepfakes. Following this direction, recent studies have started the integration of human visual attention into deepfake detection models. Korshunov and Marcel [207] highlighted the discrepancies between human and algorithmic detection capabilities, suggesting that models could benefit from mimicking human perceptual strategies.

Our study, presented in the previous section, has demonstrated that humans are particularly susceptible to the alteration of images. There exists a difference in the human observational pattern when observing at real or fake images.

The application of human attention and saliency models in detecting partially manipulated religious images is still underexplored. In fact, religious images hold significant cultural and emotional value, making the detection of alterations critical to preserving the integrity of the cultural heritage. Following the research direction of our previous study, in this section, we present a first attempt to exploit the potential of human visual attention for the automatic detection of partially manipulated images, with a specific focus on the field of religious studies. By analyzing where humans focus their attention when identifying deepfakes, we create human-derived saliency maps that highlight the most perceptually relevant regions for deepfake detection. In this context, saliency maps can highlight areas that might have been manipulated, as these regions often exhibit subtle inconsistencies that draw attention. By incorporating this visual attention information into visual backbones, we aim to guide the feature extraction process toward areas that are more likely to show signs of manipulation. This human-in-the-loop approach is shown to enhance the performance of deepfake detection systems, especially on images that are visually obvious to human observers but challenging for current algorithms.

To sum up, the main contributions of this research work are as follows:

- We exploit human attention and visual saliency for the task of deepfake detection on partially manipulated images.

- We introduce a new testing dataset specifically designed to address the distinctive challenges posed by religious images and, more broadly, partially modified ones (Fig. 5.5).
- Our experiments show how the integration of visual attention into the training process of a model for the deepfake detection task contributes positively to its performance. Moreover, we show that even in the presence of a small amount of data recorded by eye-trackers, they are of value and beneficial compared to the exclusive use of generated data from saliency models.

PROPOSED APPROACH

The main goal of our study is to assess, from an objective standpoint, the potential benefits of incorporating visual attention information into the design of models for the automatic identification of manipulated images.

Starting with the basic formulation, let $\mathcal{X} = \{x_1, \dots, x_N\}$ be a set of N input images and $\mathcal{Y} = \{y_1, \dots, y_N\}$ their corresponding labels, where $y_i \in \{0, 1\}$ indicates whether the i -th image is genuine (0) or manipulated (1). The model employs a visual backbone to extract features, denoted by the function $f_\theta(\cdot)$, where θ represents the learnable parameters. The feature vector extracted from the backbone for image x_i is given by $z_i = f_\theta(x_i)$. A classifier $g_\phi(\cdot)$ with parameters ϕ is then applied to these features, producing a probability distribution over the two classes: $p_i = g_\phi(z_i)$. The model is therefore trained to minimize the binary cross-entropy loss:

$$\mathcal{L}_{BCE} = -\frac{1}{N} \sum_{i=1}^N [y_i \log p_i + (1 - y_i) \log(1 - p_i)], \quad (5.1)$$

where p_i represents the predicted probability of the i -th image being manipulated. The parameters θ of the visual backbone and ϕ of the classification head are optimized jointly using stochastic gradient descent to minimize the binary cross-entropy loss over the training dataset.

Inspired by [48], we incorporate human generalization capabilities with in-

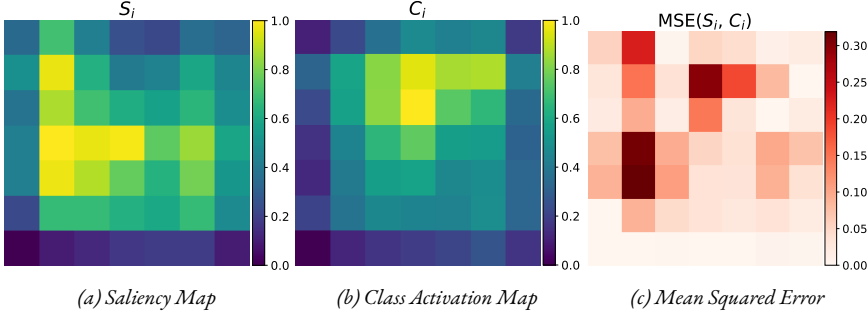


Figure 5.4: Visualization of mean squared error between saliency map (S_i) and Class Activation Map (C_i) for a manipulated image x_i .

sights from visual attention models within the aforementioned framework. In particular, given an image x_i , we aim to focus on specific regions that are likely to be informative for the classification task. This information can be represented as an empirical fixation density S_i^b , obtained by convolving the fixation locations of all the observers on x_i by an isotropic bidimensional Gaussian function, or via an ideal saliency model, defined as $S_i^m = \mathcal{M}(x_i)$, where $\mathcal{M}(\cdot)$ is a function that maps an input image to a saliency map based on low-level visual features and higher-order statistics. To leverage this information, we compute Class Activation Maps (CAMs) [412] for each image, denoted as $C_i = h_\psi(z_i)$, where $h_\psi(\cdot)$ is the CAM generation function with parameters ψ . The model is therefore trained to minimize a combined loss function:

$$\mathcal{L} = \lambda \mathcal{L}_{BCE} + (1 - \lambda) \mathcal{L}_{Sal}, \quad (5.2)$$

where $\mathcal{L}_{Sal}$ is the saliency loss that measures the discrepancy between the saliency map and the CAM. Formally, this is defined as follows:

$$\mathcal{L}_{Sal} = \frac{1}{N} \sum_{i=1}^N D(S_i, C_i), \quad (5.3)$$

where $D(\cdot, \cdot)$ is a distance function (*e.g.* mean squared error or KL divergence) between $S_i \in \{S_i^b, S_i^m\}$ and the CAM C_i related to the image x_i , while $\lambda \in [0, 1]$ is

an hyperparameter that balances the two loss components. To provide a graphical representation of the distance function, we report in Fig. 5.4 the mean squared error obtained from a generic manipulated image x_i . Here S_i refers to the saliency map while C_i is the 7×7 pixels feature map obtained from one of the last layers of a CNN. Minimizing Eq. (5.2) encourages the model to learn features that effectively distinguish between real and fake images (through the classification loss $\mathcal{L}_{BCE}$) while focusing on image regions that are likely to be informative for the classification task, guided by the saliency map and the CAM similarity term.



Figure 5.5: Qualitative examples of our proposed dataset comprising original religious images and corresponding manipulated versions using segmentation masks to preserve religious figures. From top to bottom and from left to right: Marian apparition of Banneux (Belgium), Marian apparition of Fatima (Portugal), Marian apparition of Guadalupe (Mexico), Marian apparition of Lourdes (France).

TESTING DATASET

DATASET CONSTRUCTION. To evaluate the effectiveness of the proposed approach, we introduce a novel testing dataset specifically designed to address the distinctive challenges posed by religious images and, more broadly, partially modified ones. Our proposed dataset is composed of images representing the iconography attested within Marian devotion and the phenomena of Marian apparitions.

tions over the centuries, which could be used in the study of devotional phenomena and the perception of religious images in the contemporary age. In particular, we collect 50 images of Marian apparitions related to 12 different religious representations. Examples of the collected images are depicted in Fig. 5.5a. Our objective is to define a manipulation process that exclusively focuses on the context of the religious representation. Indeed, the partial modification of the original image, coupled with the use of a conservative mask, precludes the introduction of awkward elements or distortions on the sacred figure, thus making the detection task particularly challenging also from a semantic standpoint.

In detail, given an image x_i , we extract the segmentation mask s_i (see Fig. 5.5b) pertaining to the individuals depicted in the scene through the SAM model [203]. In the editing phase, we employ the inpainting pipeline of the Stable Diffusion XL model (SDXL [278])[‡] to generate manipulated images. This process takes as input the original image x_i , a segmentation mask representing the inpainting regions (*i.e.* in our case the inverse of s_i since we want to preserve the religious figures), and a textual prompt t_i which guides the regeneration of the image context while keeping the main subjects unchanged. Specifically, t_i is randomly sampled from a template of possible prompts that we manually devise for the scope. Examples of t_i are: *“Peaceful orchard with fruit-laden trees and buzzing bees”*, *“Tranquil garden with blooming flowers and greenery”*. Each of the 50 original images is linked to six distinct contextual modifications, resulting in a total of 350 examples (Fig. 5.5c).

DATASET ANALYSIS. Considering the manipulation pipeline previously described, we aim to understand the variation in the focus of attention between a real image and its altered counterpart. We demonstrate that the difference in the attentive pattern across the images of our constructed dataset is of considerable importance. By utilizing TempSAL [20], a state-of-the-art saliency predictor, we generate saliency maps for both real and altered images to simulate human visual attention. The qualitative analysis presented in Fig. 5.6 provides evidence of a discernible attentive pattern across the two image categories. Compared to the real samples, the generation of the surrounding context through a diffusion-based

[‡]<https://huggingface.co/diffusers/stable-diffusion-xl-1.0-inpainting-0.1>

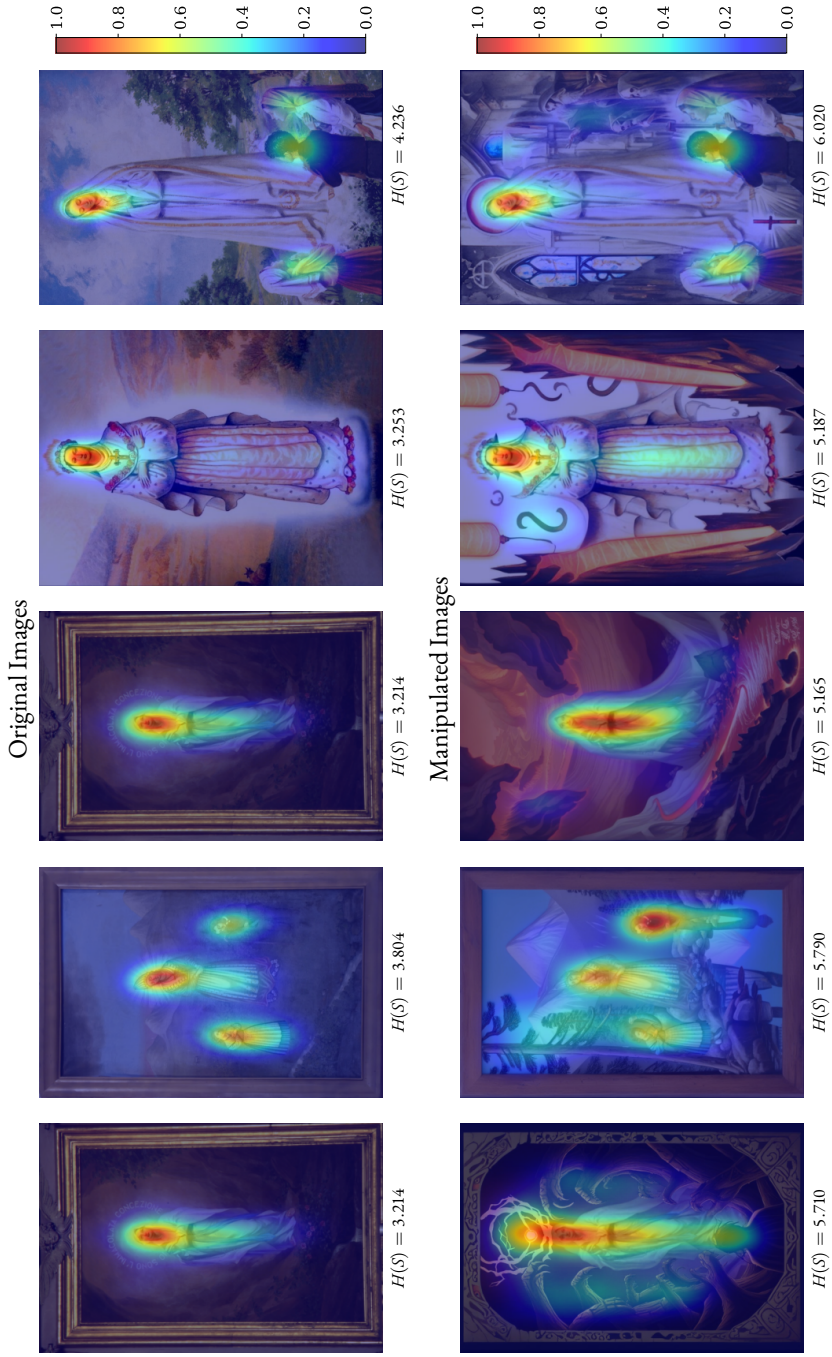


Figure 5.6: Examples of the computed saliency maps before (top) and after (bottom) the manipulation. H represents the entropy while S is the saliency map predicted by [20].

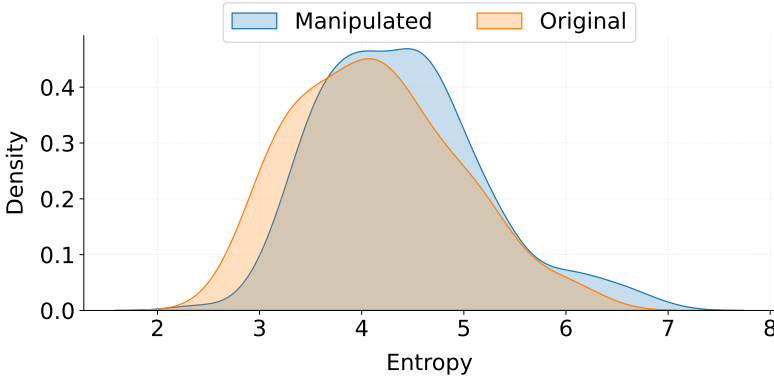


Figure 5.7: Comparison of saliency map entropy distributions between original and manipulated images.

model causes a significant shift towards the inpainted regions.

To quantify this observed difference in attention distribution, we employ entropy as a measure of dispersion for our saliency maps. Entropy, in this context, provides an indication of how focused or dispersed the attention of the model is across the image. A lower entropy value suggests that the attention of the model is concentrated on specific regions, while a higher entropy indicates a more uniform distribution of attention across the image. Given a saliency map S of size $m \times n$, we first normalize it to form a probability distribution:

$$P(i, j) = \frac{S(i, j)}{\sum_{x=1}^m \sum_{y=1}^n S(x, y)}, \quad (5.4)$$

where $P(i, j)$ represents the normalized saliency value at position (i, j) . The entropy H of the saliency map is then calculated as:

$$H = - \sum_{i=1}^m \sum_{j=1}^n P(i, j) \log_2 P(i, j). \quad (5.5)$$

This formulation yields entropy values in bits. A saliency map with perfectly uniform attention would have a maximum entropy of $\log_2(mn)$, while a map with all attention focused on a single pixel would have a minimum entropy of 0.

As a consequence of the attention shift towards inpainted regions in synthetic images, the entropy of the corresponding saliency maps presents higher values, as indicated in Fig. 5.7, where the empirical distributions of the saliency entropy values are compared to the real case. This higher entropy in synthetic images suggests a more dispersed attention pattern, likely due to the model detecting inconsistencies or artifacts introduced by the inpainting process across a wider area of the image. In contrast, the lower entropy observed in real images indicates more focused attention, possibly on natural salient features of the original, unaltered image, such as the faces of the people featured. This quantitative entropy analysis corroborates our qualitative observations, providing a numerical basis for the difference in attention patterns between real and synthetic images. It emphasizes the potential of entropy as a discriminative feature in distinguishing between original and manipulated images, offering insights into the spatial characteristics of model attention.

EXPERIMENTAL SETTING

DATASETS. As previously described, the training procedure entails the calculation of a saliency loss, which employs images and associated saliency maps representative of the region of attention. Consequently, a dataset comprising both types of data, collected in the context of partially modified images, is essential for effective learning. The Unveiling the Truth (UTruth) dataset presented in [61] represents the only collection of both real and manipulated images paired with visual attention data obtained through eye-tracking. The dataset consists of 400 images: 100 authentic and 300 partially modified using diffusion models. The modifications were carefully executed to preserve the realism, semantics, and context of the original images. During an eye-tracking experiment, each image was viewed by five distinct subjects, resulting in a saliency map that combines all the recorded fixations for every image in the dataset. To achieve a balanced ratio between real and manipulated images, we supplement this initial set with 200 authentic images and their corresponding saliency maps from CAT2000 [45], a widely-used saliency dataset that features images from different categories.

However, the limited number of available images resulting from the oner-

ous process of data acquisition using an eye-tracker led us to integrate into the training dataset 2, 000 images collected in D³ [27], a comprehensive dataset designed for large-scale deepfake detection. D³ includes 9.2 million generated images created using four state-of-the-art diffusion model generators and their original counterpart. In particular, we randomly selected 1, 000 original samples and 1, 000 corresponding generated versions among all the images with at least 400 pixels on each side. Given that images from D³ do not include human fixations, we equip them with computational saliency maps obtained from TempSAL [20], a recent saliency prediction model that can well simulate human attention.

ARCHITECTURE & TRAINING DETAILS. Our approach utilizes a ResNet-50 architecture [157] as the visual backbone, pre-trained on ImageNet-1k [110]. We modify the final fully connected layer to output two classes, corresponding to “real” and “manipulated” images. The model is trained using stochastic gradient descent with a momentum of 0.9 and weight decay of 1×10^{-6} . We employ a step learning rate scheduler, initially setting the learning rate to 0.005 and reducing it by a factor of 0.1 every 12 epochs. The training process spans 50 epochs with a batch size of 32. As described in Eq. (5.2), our loss function is a weighted combination of two components: binary cross-entropy loss for classification $\mathcal{L}_{BCE}$ and mean squared error as saliency loss $\mathcal{L}_{Sal}$ prediction. The best combination of the two loss components is obtained with $\lambda = 0.9$.

To visualize the focus of the model, we implement Class Activation Mapping (CAM) [412] by extracting features from the final convolutional layer of ResNet-50. The CAM is computed by weighting these features with the corresponding class weights from the final fully connected layer. This allows us to generate heatmaps highlighting the regions most influential in the model decision-making process. All input images are resized to 224×224 .

EVALUATION METRICS. Our evaluation methodology is tailored to address the significant imbalance in our testing dataset, which comprises 50 original images and 400 fake images. This 1:8 ratio necessitates a careful selection of evaluation metrics that are robust to class imbalance.

We prioritize Average Precision (AP) as our primary metric. AP provides a single-value summary of the precision-recall curve, offering a concise yet com-

	AP	AUC-PR	Precision	Recall	AUC-ROC
Grag2021 [148]	87.81	87.69	100.00	0.25	51.68
Wang2020 [359] (Blur+JPEG 0.5)	90.44	90.38	100.00	0.25	56.78
Wang2020 [359] (Blur+JPEG 0.1)	92.28	92.25	100.00	0.50	61.66
DIRE [364] (LSUN/ADM)	85.83	85.65	75.00	3.00	43.14
DIRE [364] (LSUN/PNDM)	87.79	87.94	89.04	95.50	47.15
Ours w/o $\mathcal{L}_{Sal}$	92.97	92.95	92.79	48.25	62.13
Ours w/ $\mathcal{L}_{Sal}$	95.35	95.34	94.76	63.25	72.93

Table 5.2: Accuracy results on our testing dataset comparing our solution with state-of-the-art deepfake detection methods. Precision and recall are not in bold as they are not individually significant in the case of an unbalanced dataset. Ours w/o $\mathcal{L}_{Sal}$ means our approach with $\lambda = 1$, while Ours w/ $\mathcal{L}_{Sal}$ means our approach with $\lambda = 0.9$ which represents our best configuration.

prehensive measure of model performance. It is particularly informative in our imbalanced scenario, as it emphasizes the model ability to detect the minority class (*i.e.* original images) without being influenced by the majority class (*i.e.* manipulated images). Complementing AP, we utilize the Area Under the Precision-Recall Curve (AUC-PR) that offers a more detailed view of the precision-recall trade-off across different thresholds. Given the prevalence of manipulated images in our dataset, Precision becomes a critical metric. It quantifies the proportion of correct fake detections among all images classified as fake, thereby measuring our model ability to avoid false alarms. This is particularly important given the potential real-world implications of misclassifying original images as fake. Complementing Precision, we also consider Recall (or sensitivity). Recall indicates the proportion of actual fake images successfully identified by our model. With the large number of fake images in our dataset, a high recall ensures that we detect a significant portion of the manipulated content.

Lastly, we include the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) for completeness. While AUC-ROC is a popular metric that measures the model discriminative ability across various classification thresholds, we interpret it cautiously in our context. AUC-ROC can potentially overestimate performance on imbalanced datasets like ours, as it is less sensitive to class imbalance compared to precision-recall based metrics.

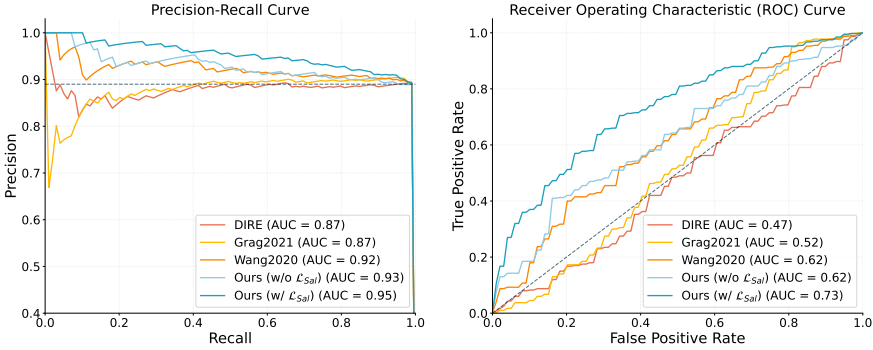


Figure 5.8: Performance comparison of image manipulation detection methods on our testing dataset. We report the Precision-Recall curve (left) and ROC curve (right). The dashed line indicates the performance of a random classifier.

COMPARISON WITH STANDARD DEEPFAKE DETECTORS

We evaluate our model in detecting real and fake images on our proposed testing dataset. In particular, we compare our model against state-of-the-art approaches, namely the model proposed by Gragnaniello *et al.* (Grag2021) [148], the one introduced by Wang *et al.* (Wang2020) [359], and the DIRE architecture [364]. The results for deepfake detection are reported in Table 5.2 and plotted in Fig. 5.8. Our fake detection approach surpasses state-of-the-art detectors across all key metrics, including AP, AUC-PR, and AUC-ROC. Although the considered approaches achieve higher absolute precision or recall values, it should be noted that in the context of an imbalanced dataset, precision and recall individually may not provide a comprehensive evaluation. In addition to these results, in Table 5.2 we also report a comparison of our model with and without the saliency loss (cf. Eq. (5.2)). It is clear that augmenting training with saliency maps representing the human focus of attention leads to better performance. Saliency maps represent a guide for the detector that, during training, learns to adjust its internal attention representation towards regions that are the most discriminative for the real-fake classification, thus mimicking human behavior.

ABLATION STUDIES

ANALYZING THE IMPACT OF THE TRAINING SET. In the previous section, we demonstrated the efficacy of our method in detecting manipulated images depicting religious representations. The integration of saliency within the fake detection pipeline proved to be fundamental. In this section, we conduct an ablation study to shed light on the contributions of human attention in contrast to the attention maps generated by a predictive model. In Table 5.3, we compare different training strategies considering three combinations of training sets. The first row refers to a setting where we employ 300 real and 300 fake images from the UTruth dataset presented in [61] and CAT2000 *et al.* [45], along with their corresponding human saliency maps. In the second case, we train our model on 1, 000 real and 1, 000 fake images derived from D³ [27]. Although training on D³ achieves superior performance, the approach based entirely on human attention remains competitive. We attribute this performance gap to the different sizes of the two training sets. Indeed, focusing on the latter training setting in the table, where the dataset includes a combination of real and generated saliency maps, the performance gain is maximized. Such results prove that the integration of human attention is beneficial for the detection of partially manipulated images.

GENERALIZATION ACROSS DIFFERENT BACKBONES. We conduct extensive experiments across various visual backbones to evaluate the robustness of the proposed method. In detail, we report in Table 5.4 the results for four backbones, each trained with ($\lambda = 0.9$) and without ($\lambda = 1$) the incorporation of saliency maps. In addition to reporting the performance of ResNet-50, we consider ResNet-18 [157], DenseNet-201 [165], and a Vision Transformer-based backbone [117] like Swin-Tiny [239]. ResNet-50 leads to the best results according to all evaluation metrics. Most importantly, the adoption of the saliency loss leads to better detection performance across all the key metrics, regardless of the backbone employed, thus further demonstrating the effectiveness of human attention for the deepfake detection task.

ANALYZING THE IMPACT OF THE λ PARAMETER. Fig. 5.9 illustrates the impact of varying the lambda parameter from Eq. (5.2) on the accuracy of detect-

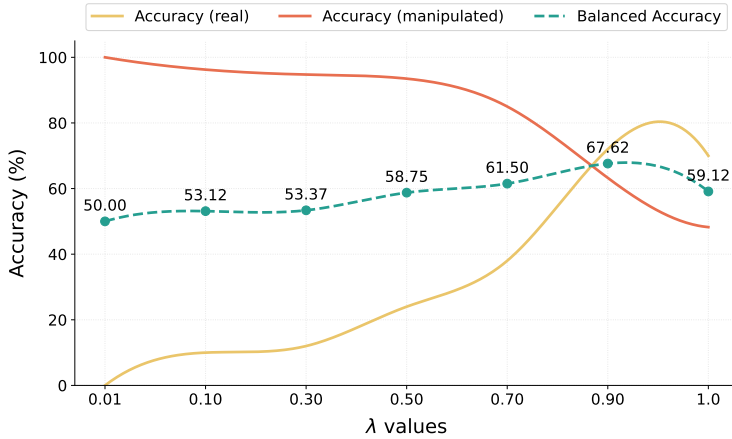


Figure 5.9: Impact of the lambda parameter on image classification accuracy. The optimal balance between real and manipulated image detection is achieved at $\lambda = 0.9$, demonstrating the benefit of combining classification and saliency losses.

Human Saliency		Computed Saliency					
UTruth [61]	CAT2000 [45]	D ³ [27]	AP	AUC-PR	Precision	Recall	AUC-ROC
✓	✓	-	89.97	89.93	89.72	96.00	56.06
-	-	✓	93.57	93.55	94.40	29.50	64.31
✓	✓	✓	95.35	95.34	94.76	63.25	72.93

Table 5.3: Performance comparison of different training set strategies.

ing real and manipulated images, as well as the balanced accuracy between the two. Here, we consider the best configuration of our approach obtained with a ResNet-50 backbone. As lambda increases from 0.01 to 1.0 (representing less weight given to the saliency loss), we observe that the accuracy for real images steadily increases, reaching a peak around $\lambda = 0.9$. The accuracy for manipulated images sharply decreases, especially for λ values above 0.7. The balanced accuracy shows a gradual improvement up to $\lambda = 0.9$, after which it declines with $\lambda = 1.0$ which corresponds to a training without the saliency loss.

The performance in detecting manipulated images drops significantly as lambda increases, indicating that reducing the influence of the saliency loss negatively impacts the model ability to identify manipulated content. This indicates the importance of combining both class loss and saliency loss for optimal perfor-

	$\mathcal{L}_{Sal}$	AP	AUC-PR	Precision	Recall	AUC-ROC
ResNet-18	✗	91.29	91.26	92.00	34.50	57.67
	✓	91.75	91.72	90.32	56.00	58.00
ResNet-50	✗	92.97	92.95	92.79	48.25	62.13
	✓	95.35	95.34	94.76	63.25	72.93
DenseNet-201	✗	92.35	92.33	91.60	57.25	61.96
	✓	95.17	95.16	98.00	24.50	71.73
Swin-Tiny	✗	90.54	90.50	89.81	35.25	54.86
	✓	92.40	92.38	94.26	28.75	57.94

Table 5.4: Performance comparison of various visual backbone architectures.

mance. The balanced accuracy peaks at $\lambda = 0.9$, suggesting this is the optimal value for balancing the detection of both real and manipulated images. Beyond this point, corresponding to a global loss entirely based on the classification loss, the overall performance degrades.

5.2 UNRAVELING NEURODIVERGENT GAZE BEHAVIOR THROUGH VISUAL ATTENTION CAUSAL GRAPHS

Decades of research on the nature of eye movements and visual attention have shaped the interplay between the determinism and randomness of where our eyes alight in the visual scene. This interplay arises from bottom-up cues (salient features due to stark contrasts of luminance, color, and motion, e.g. [176]) that harmonize with the deliberate direction of top-down influences [88, 113, 176]. The demands of our current pursuits, our accumulated knowledge, and the expectation of what might lie ahead form attention. Attention goes beyond simple reaction and becomes a sophisticated reflection of our inner world [113, 311, 397]. Under the assumption of perception as hypothesis testing, eye movements can be considered as experiments, in which data are gathered to test (counterfactual) hypotheses or beliefs about how those data are caused [137, 142].

Specifically, the idea that attention allocation follows a causal structure suggests that top-down attention control may be represented as a form of high-level causal reasoning that guides the observer’s attentive dynamics towards specific

items. Eventually, the sequence of fixations chosen to sample the scene may be influenced by underlying cognitive factors such as semantic relatedness [61], task relevance, social meaning [39], and individual cognitive characteristics [309]. Indeed, humans possess sophisticated mechanisms that enable reasoning about cause and effect in both abstract and formal terms [272]. Recent research has demonstrated how these mechanisms, particularly counterfactual reasoning, can be revealed by an observer's gaze behavior [142]. This causal influence likely reflects a dynamic interaction between perceptual inputs and the observer's internal state, suggesting that visual behavior can serve as an indirect window into higher-level cognitive functioning.

This perspective may have profound implications for developmental and clinical research. Autism Spectrum Disorder (ASD) is a complex disorder characterized by dysfunctions of between- and within-networks functional connectivity involving the somatomotor, visual, salience, and default mode networks (DMN) [360]. The DMN critically modulates the theory of mind (ToM), social interaction, and emotions. Individuals with ASD exhibit atypical patterns of visual attention that are often characterized by diminished attention to socially salient stimuli, such as faces or eye regions [274, 307], though variations might occur as a function of the social context in which stimuli are embedded. Traditional metrics of gaze behavior, such as total fixation duration or number of fixations on regions of interest, even when complemented by sophisticated statistical analyses (e.g., [172]), provide valuable descriptive information but fail to capture the underlying causal structure of attentional deployment (see Section 5.2.1). Instead, modeling gaze as a phenomenon governed by causal relationships potentially allows for a deeper understanding of how attention is orchestrated moment-to-moment, and how this differs between groups of individuals.

The goal of this thesis is to verify whether the causal structure governing the dynamics of attention allocation on natural scenes (images) has the potential to differentiate subjects with Autism Spectrum Disorder (ASD) from typically developing (TD) controls. Although it is largely agreed that eye tracking offers measurable variables that are promising in the study of neurodevelopmental disorders [138, 349] and could serve as biomarkers for early diagnosis, monitoring, and fur-

ther care interventions [138], the very causal nature of gaze unfolding is substantially unexplored in this field.

More specifically, by means of a state-of-the-art causal discovery method from time series data [301], we estimate the directed acyclic graphs (DAGs) describing the causal mechanisms of attention allocation from eye-tracking data (Section 5.2.2). Crucially, these causal graphs capture not only the statistical patterns of gaze transitions on specific regions of interest but also infer the directional causal influence that fixating on a given item has on another, shedding light on the hierarchical structure of attentional control. We hypothesize that individuals with ASD exhibit distinct causal structures of gaze deployment, reflecting alterations in how cognitive factors modulate attention. Section 5.2.6 reports and discusses results achieved on a publicly available dataset.

To the best of our knowledge, this is the first study to investigate the diagnostic potential of causal modeling of gaze behavior to discover the cognitive phenotypes of autism spectrum disorder.

5.2.1 BACKGROUND ON GAZE MODELING FOR NEURODEVELOPMENTAL DISORDERS

Modeling gaze behavior has become an important area of study in computer vision and cognitive science [60], particularly for understanding attention mechanisms and supporting diagnostic applications in neurodevelopmental disorders.

Gaze behavior has proven especially valuable in the context of ASD detection and ASD/ADHD comorbidity [172]. Some studies have developed frameworks that utilize multi-modal behavioral data, including eye-tracking, to distinguish ASD from TD [73, 188, 236]. Others have proposed models capable of simulating the dynamic gaze behavior (scanpaths) of neurodivergent individuals [75]. Additionally, multi-level saliency models have been introduced to capture distinct attentional patterns in ASD, such as a central fixation bias and reduced attention to faces and socially relevant cues [357]. The Saliency4ASD Grand Challenge further advanced this field by providing datasets and tools to support the development of visual attention models tailored to ASD characteristics [153]. From

a biometric perspective, eye movements have been modeled using stochastic processes rooted in foraging theory, yielding personalized behavioral signatures for identification [121]. Other studies have analyzed gaze in social contexts, linking it to personality traits using probabilistic models [98]. Similarly, sequence models trained on eye movement data have shown success in ADHD detection, emphasizing the importance of dynamic gaze features and task design [112]. Recent solutions have started to employ high-level graph representations to model human attention by means of eye-tracking data. One recent approach is the Attention Graph [382], which encodes saliency and semantic-driven scanpaths in a graph-based structure to capture common gaze behaviors across individuals. In a different vein, GazeGraph introduced a flexible and generalizable framework that combines spatial-temporal gaze graphs with deep learning and meta-learning for rapid adaptation and robust context recognition [218].

While such representations enhance both interpretability and utility in tasks such as ASD screening and age classification [382], their modeling approach is limited to quantifying statistical associations, without accounting for higher-level causal drivers of attention allocation. Recent research has shown that counterfactual simulation during gaze tasks plays a significant role in causal reasoning [142], supporting the development of causal frameworks for attention modeling. This work takes a first step in that direction by proposing Visual Attention Causal Graphs, aiming to integrate attention modeling and causal reasoning for a deeper understanding of gaze behavior.

5.2.2 PROPOSED METHOD

An overview of our approach is summarized in Fig. 5.10. A causal graph is estimated, which describes the mechanisms governing the dynamics of attention allocation towards semantically relevant *items* within a scene. The goal is to validate the assumption that such graphs, referred to as Visual Attention Causal Graphs (VA-CG), effectively capture the observer’s top-down attention allocation process and provide insights into their behavioral cues. Specifically, we aim to quantify their ability to assess high-level cognitive functioning, with a focus

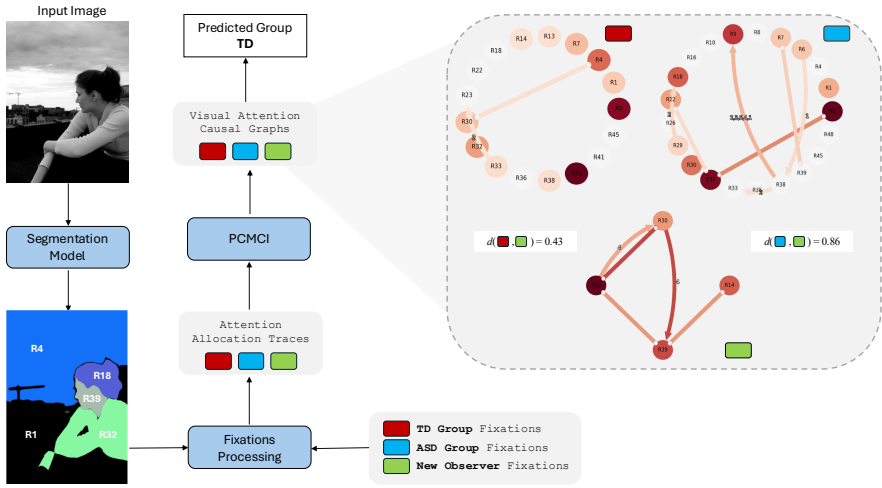


Figure 5.10: The proposed method at a glance. Semantically relevant items are automatically extracted from the input image using a segmentation model. Visual Attention Causal Graphs (VA-CGs) are then generated as templates for the ASD (blue) and TD (red) observer groups. A VA-CG is also constructed for the new observer (green). Finally, the new observer is classified as either ASD or TD by minimizing the Jaccard distance between their VA-CG and the template VA-CGs.

on quantitative screening for autism spectrum disorder (ASD). The approach is divided into three phases:

1. *Construction of Aggregated Item-Level Attention Allocation Traces:* The visual attention behavior of a group of subjects is represented as a multivariate time-series that describes the aggregated group-level attention toward each semantically relevant item inside the scene.
2. *Estimation of the Visual Attention Causal Graph (VA-CG):* The causal mechanism of attention allocation dynamics is reformulated as a multivariate time-series causal discovery problem. The state-of-the-art PCMCi causal discovery method [301] is applied to estimate the associated causal graph.
3. *ASD vs. TD Classification via the Obtained VA-CG Representations:* The final phase involves using the VA-CG representations to classify subjects as

either exhibiting traits of autism spectrum disorder (ASD) or as typically developing (TD).

5.2.3 AGGREGATED ITEM-LEVEL ATTENTION ALLOCATION TRACES

The first step of the proposed approach implies the construction of the multivariate time series that describes the aggregated group-level attention toward semantically relevant items. Consider a stimulus j (image), observed by a group of S subjects with their eye movements tracked in real-time.

A segmentation approach is adopted to define semantically relevant items within an image, where each item corresponds to a distinct region or object [244]. For each time instant, we quantify the visual attention allocated to each item by counting the number of fixations directed towards it by the subjects.

This procedure generates a multivariate time series, where each time series corresponds to an individual item, describing the aggregated attention allocation across multiple subjects over time. The resulting data, shown in Fig. 5.11, captures the group-level temporal dynamics of visual attention at the item level, providing insight into how attention is distributed across different regions of the image as it is processed by the viewers.

5.2.4 VISUAL ATTENTION CAUSAL GRAPHS (VA-CG)

To estimate the causal graph from the obtained aggregated item-level attention allocation traces, we employ the PCMCI causal discovery method [301]. PCMCI (Peter–Clark Momentary Conditional Independence) is a method designed to uncover causal relationships within multivariate time series data. By analyzing the temporal dependencies among variables, PCMCI infers causal structures that can be compactly represented using Directed Acyclic Graphs (DAGs). As with any causal discovery framework, PCMCI relies on a set of foundational assumptions. These include the causal Markov condition and the faithfulness assumption [328], as well as specific conditions tailored to time series: causal stationarity (the assumption that causal relationships remain constant over time), the absence

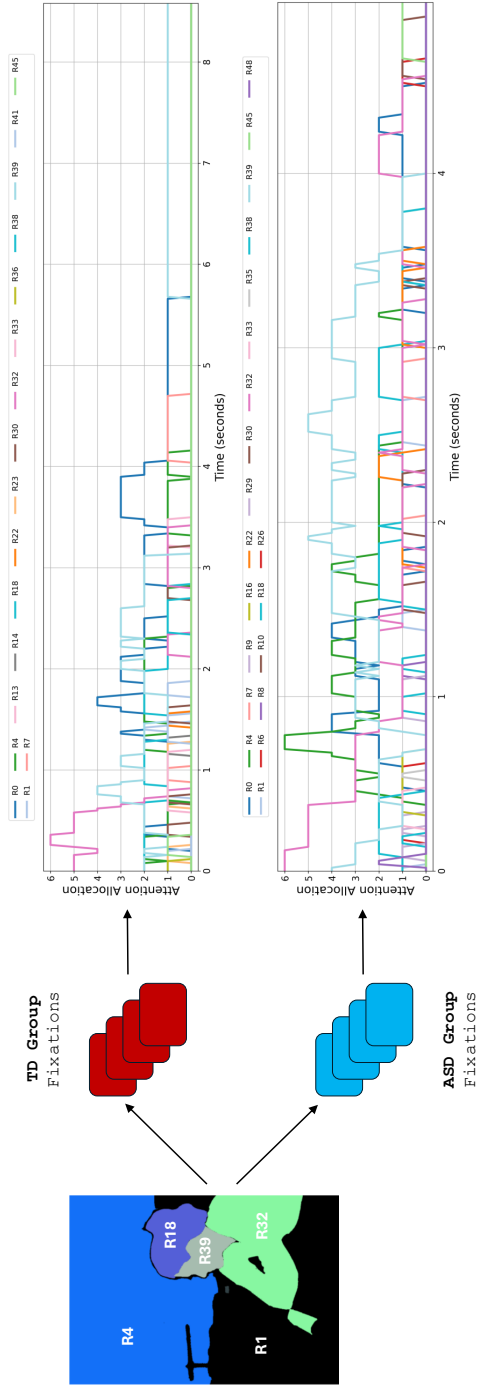


Figure 5.11: Illustration of the construction of a multivariate time-series representing group-level visual attention allocation on a scene. The input image is segmented into semantically meaningful regions (e.g., R1, R4, R18), and fixation data from the TD (top) and ASD (bottom) observer groups are aggregated over time. The resulting time-series plots show the temporal allocation of visual attention to each region, capturing group-level dynamics in visual attention distribution.

of instantaneous (contemporaneous) causal links, and causal sufficiency - meaning that all causally relevant variables are observed and included in the model [301]. In this study, we apply PCMCI to the aggregated semantic-level attention allocation traces described above. Specifically, for each time series $x_i(t)$, corresponding to the amount of attention allocated to a given item at time t , PCMCI employs Conditional Independence (CI) testing to detect causal dependencies. Causal links between traces $x_i(t)$ and $x_j(t)$ are determined by evaluating their conditional independence ($\perp$) with respect to a set of potentially mediating variables $\mathbf{z}(t)$:

$$x_i(t) \perp x_j(t) \mid \mathbf{z}(t),$$

where $\mathbf{z}(t)$ encompasses other variables that may influence the interaction between $x_i(t)$ and $x_j(t)$. The application of PCMCI requires the specification of several key hyperparameters:

- *Conditional Independence Test*: The choice of CI test is critical, with available options including Partial Correlation (ParCorr), its robust variant (r-ParCorr), as well as discrete-variable tests such as the G-squared test. Selection depends on the nature of the data, its distributional properties, and sample size.
- *Time Lag Range* (τ_{min}, τ_{max}): Defines the minimum and maximum temporal delays over which causal relationships are assessed. Larger lag windows facilitate the discovery of long-range dependencies, but at the cost of increased computational complexity.
- *Significance Threshold* (α): Specifies the alpha level for the CI tests, serving to filter out statistically insignificant or spurious causal links.

By analyzing directed, time-lagged dependencies among the attention traces, PCMCI constructs a causal graph that reflects the structure of attention flow across the items in a scene. We refer to such graphs as Visual Attention Causal Graphs (VA-CG). Time-unrolled versions of VA-CGs, estimated on the same image for both TD and ASD groups, are shown in Fig. 5.12.

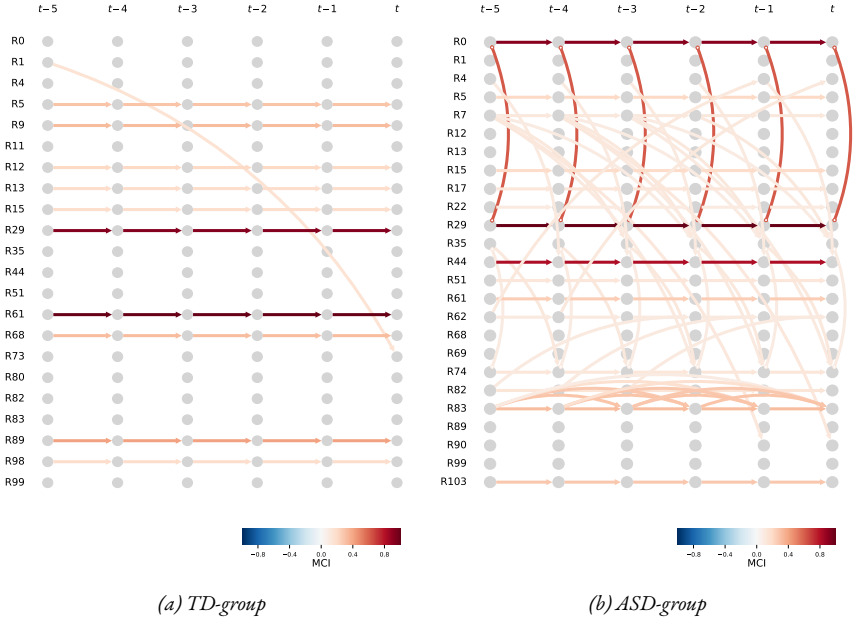


Figure 5.12: Time unrolled VA-CGs estimated on the same sample image from the TD (a) and ASD (b) groups, providing a causal representation of visual attention dynamics across semantically segmented regions. Each node represents a region of interest (e.g., R0–R103) at successive time steps ($t-5$ to t). Directed edges indicate causal transitions of attention between regions, with edge color denoting the causal strength as estimated by the Momentary Conditional Independence (MCI) test. The TD-group displays more consistent, linear patterns of attention allocation, while the ASD-group exhibits more complex and diverse transitions, indicating group-level differences in visual attention behavior.

5.2.5 ASD DETECTION VIA VA-CGS

For each image j in the dataset, available subjects are used to build the Aggregated Item-Level Attention Allocation Traces for both the ASD and TD groups. These traces are then used to construct the VA-CGs for the ASD (G_j^{ASD}) and TD (G_j^{TD}) groups. These graphs serve as templates for classifying a new subject's (s) scanpath on the j -th image by computing its visual attention causal graph G_j^s using the procedure described in Section 5.2.2.

The assignment of the s -th subject's eye tracking data to one of the groups is determined by calculating the Jaccard distance between the ASD and TD VA-CGs and the new subject's causal graph:

$$\text{group}_j^* = \arg \min_{g \in \{\text{ASD}, \text{TD}\}} d(G_j^g, G_j^s).$$

Specifically, given a threshold $\beta \in (0, \alpha)$, that sets the minimum causal strength, the set of links (edges) considered in the graph G_j with N nodes is defined as:

$$\mathcal{E}_j = \{(m, n, \tau) \mid P_{m,n,\tau} < \beta \wedge |T_{m,n,\tau}| > 0\}$$

where $T \in \mathbb{R}^{N \times N \times (\tau_{\max} + 1)}$ and $P \in [0, 1]^{N \times N \times (\tau_{\max} + 1)}$ are the matrices of test statistic values and the matrices of p-values for the graph G_j , respectively. The distance between the two graphs, based on the Jaccard distance, is:

$$d(G_j^g, G_j^s) = 1 - \frac{|\mathcal{E}_j^g \cap \mathcal{E}_j^s|}{|\mathcal{E}_j^g \cup \mathcal{E}_j^s|}.$$

Following previous work [73, 382], more robust subject-level classification can be obtained by a majority voting scheme based on the predictions across multiple stimuli.

5.2.6 EXPERIMENTAL ANALYSIS

DATASET

We adopt the Saliency4ASD dataset [153], which offers a collection of 300 images accompanied by eye-tracking records from children diagnosed with ASD and TD children. This dataset is intended to facilitate research on visual attention modeling as it relates to ASD diagnosis. In the original study, the authors gathered eye movement data from a cohort comprising 14 children with ASD and 14 children with TD, all aged between 5 and 12 years, with a mean age of approximately 8 years. The dataset does not include subject identifiers distinguishing individuals within each group. Consequently, in alignment with previous work [73], fixation sequences indexed by i within the same group are assumed to originate from the same individual.

Methods	Accuracy	Precision	Recall	AUC
Chen et al. (Independent) [73]	0.89	0.86	0.93	0.92
Chen et al. (Full) [73]	0.93	0.93	0.93	0.98
AttentionGraph+S _{scan} [382]	0.93	-	-	-
AttentionGraph+S _{scan} ' [382]	0.86	-	-	-
Ours	0.93	0.93	0.93	0.93

Table 5.5: Comparison of the proposed method with state-of-the-art competing approaches on the Saliency4ASD dataset.

EXPERIMENTAL SETUP

Each image from the Saliency4ASD dataset is automatically processed to extract relevant items within the scene. The SAM 2.1 Large model [291] is employed for accurate segmentation of areas of interest, which are considered as regions of gaze attraction (items). Aggregate item-level attention allocation traces (see Section 5.2.3) are generated by integrating the amount of attention (gaze points) that fall within the segmented regions. These traces are then fed into the Visual Attention Causal Discovery method outlined in Section 5.2.2. Items that do not attract attention are excluded from the computation of the Visual Attention Causal Graph. Consistent with previous work [73, 382], the classification task is performed using a leave-one-subject-out strategy on 100 randomly selected images. For the computation of the Visual Attention Causal Graph using the PCMCi method, we set the significance threshold to $\alpha = 0.01$ and define a time lag range of $\tau_{min} = 0$ and $\tau_{max} = 5$. Given the univariate and discrete/categorical nature of the variables in the Aggregated Item-Level Attention Allocation Traces, the G-squared Conditional Independence test is adopted [301].

RESULTS

We compare the proposed method against [73] and [382] and evaluate it in terms of accuracy, precision, recall, and AUC. Results are reported in Table 5.5.

Among these methods, our approach demonstrates strong and well-rounded performance, achieving an accuracy, precision, recall, and AUC of 0.93. This places it among the top-performing models, on par with Chen et al. (Full) [73]

and AttentionGraph with S_{scan} in terms of accuracy. In contrast, Chen et al. (Independent) [73] shows slightly lower performance, with an accuracy of 0.89 and an AUC of 0.92, while AttentionGraph with S'_{scan} underperforms with an accuracy of 0.86.

It is worth noting that, unlike the purely predictive and, to some extent, black-box approach proposed in [73], the solution presented here offers a high degree of interpretability. This is achieved through the use of causal discovery techniques, which enable the method to characterize the underlying causal mechanisms governing the dynamics of attention allocation. By constructing Visual Attention Causal Graphs (VA-CGs), the proposed method not only predicts group membership accurately but also provides insights into how and why visual attention shifts over time within and across observer groups.

In this respect, the proposed solution shares some conceptual similarities with the AttentionGraph approach introduced in [382]; however, their method remains at a purely correlational level of explanation. In contrast, the adoption of VA-CGs allows the proposed approach to operate at a higher level on the ladder of causation [272], enabling the inspection of attention allocation mechanisms in neurodivergent subjects with greater explanatory depth.

As illustrated in Fig. 5.12, the time-unrolled VA-CGs offer an interpretable causal representation of visual attention dynamics across semantically segmented regions. The figure captures the temporal evolution of attention across regions of interest, with directed edges signifying causal transitions between them, and edge color reflecting causal strength. Notably, the TD-group exhibits more consistent, linear attention patterns, whereas the ASD-group demonstrates more complex and varied transitions, highlighting group-level differences in visual attention behavior. This makes the approach particularly valuable in domains where understanding the decision-making process is as important as the predictions themselves, such as in cognitive and clinical research. Overall, the proposed approach stands out for its balanced and complete performance, combining high accuracy with strong precision, recall, and AUC, while offering substantial explanatory power.

5.3 SANCTUARIA-GAZE: A MULTIMODAL EGOCENTRIC DATASET FOR HUMAN ATTENTION ANALYSIS IN RELIGIOUS SITES

Sacred spaces have long served as repositories of cultural heritage, architectural innovation, and religious expression. These environments present particular opportunities to study human attention and interaction patterns, as visitors engage with a range of elements, including architectural features and religious artifacts. Understanding how individuals explore and interact with these spaces is crucial for fields spanning architecture [6], religious studies [201], and cognitive science [197]. Recent advancements in wearable technology, particularly eye-tracking devices [127], have enabled researchers to capture and analyze human attention patterns with greater precision [60, 61].

Eye-tracking has been increasingly applied in architectural research to study human interaction with built environments [108, 125, 302, 352], providing insights into cognitive and behavioral responses to architectural designs. However, most existing studies rely on laboratory settings with static images or 3D renderings, which fail to capture the dynamic nature of real-world exploration [5, 37, 38, 49, 120]. Furthermore, while sacred spaces have been extensively studied from historical and architectural perspectives, there is a limited body of empirical research on how visitors actually experience and interact with these environments. The introduction of eye-tracking technology to in-situ architectural studies presents both opportunities and challenges.

This is particularly pertinent in the context of sacred architecture, where visual engagement is shaped not only by spatial configuration but also by symbolic, iconographic, and ritual elements that elicit complex forms of attention and interpretation. Eye-tracking studies in sanctuaries offer insights into how visitors engage with these complex spaces, highlighting patterns of attention around features like altars, relics, and Marian imagery. Such analyses offer new perspectives on devotional behavior, the effectiveness of liturgical space design, and the evolving modes of interaction with sacred imagery. Moreover, this approach supports

interdisciplinary inquiry, integrating insights from semiotics, anthropology, and art history to better understand how sacred spaces guide visual experience and structure spiritual engagement.

While these tools can provide rich data about visitor behavior, the complexity of real-world environments makes it difficult to automatically identify and track objects of interest, especially those that are specific to the sacred environment. Conventional methods need laborious manual annotation, limiting the scalability and reproducibility of such studies [250]. Recent developments [109, 345] aim to streamline this process by automating object detection and tracking in dynamic scenes, thereby enhancing the efficiency of eye-tracking data analysis.

To address these challenges, we present Sanctuaria-Gaze, a novel dataset and analysis framework focused on visitor interactions with sacred spaces. Our work makes three primary contributions: (i) a multimodal dataset capturing natural exploration behavior in four significant sanctuaries, (ii) an open-source framework for detecting and analyzing Areas of Interest (AOIs) in dynamic environments, and (iii) empirical insights into how visitors’ backgrounds influence their interaction patterns with sacred spaces.

We refer the reader to Fig. 5.13 for qualitative examples of the recordings and their corresponding annotations. Additional samples are reported in the appendix (Section C.1). Collected data and source code are publicly available^{||}.

5.3.1 AN OVERVIEW ON EGOCENTRIC DATASETS

Egocentric vision represents a distinctive paradigm for understanding human behavior, utilizing wearable cameras to capture first-person perspectives [277]. Over the last decade, the development of egocentric datasets enriched with multimodal data, including gaze and audio, showcased the potential of first-person views for addressing a wide range of challenges, from action recognition and anticipation to gaze understanding and prediction [39, 139, 140, 169, 217, 226, 227, 249, 340]. Among these, the Activity of Daily Living (ADL) dataset [276] features recordings of individuals performing unscripted, everyday actions in real-

^{||}<https://github.com/aimagelab/Sanctuaria-Gaze>

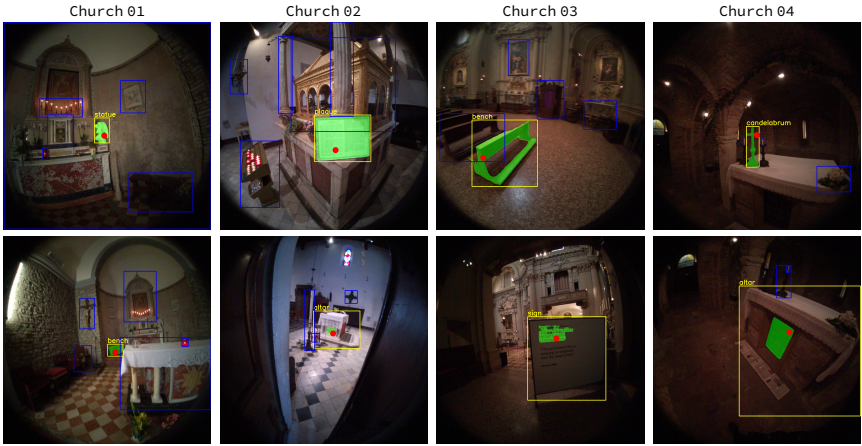


Figure 5.13: Examples of recorded videos in the four churches with corresponding annotations. The figures highlight key annotations such as gaze points, segmentation mask, object bounding boxes and label of the observed object.

world environments. Fathi et al. [132] advanced the field by integrating gaze data with egocentric videos to study human behavior, leading to the creation of the GTEA Gaze+ dataset and its subsequent extension [228]. The EPIC-KITCHENS dataset [103, 104] marked a significant milestone by offering large-scale, diverse egocentric video data focused on kitchen activities.

The Ego4D dataset [149] stands as one of the most comprehensive resources to date. It encompasses over 3,000 hours of recordings of several daily activities across hundreds of environments. Some subsets of this dataset also provide gaze data collected using a diverse range of wearable devices.

Ragusa et al. [287, 288] introduced MECCANO, a dataset of egocentric videos in industrial-like scenarios for studying human interactions with small objects and tools. Similarly, Sener et al. [312] collected Assembly101, a multi-view dataset for procedural activity analysis. In recent years, datasets focusing on cultural contexts have also been introduced, emphasizing visitor behavior in such environments. For example, the EGO-CH dataset [286] provides a comprehensive resource for studying visitor behavior in cultural sites, focusing on activities such as object attention and exploration patterns. Mazzamuto et al. [250] extended this dataset, combining gaze signals with weak supervision to detect at-

Dataset	Setting	Signal	Frames	Sequences	Object Classes	Participants
EGO4D [149]	Multi Domain	RGB, Audio, 3D Environments, Stereo, Gaze, IMU, Multi-View	418M	9650	449	931
EPIC-KITCHENS-100 [104]	Kitchens	RGB	20M	700	300	37
EGTEA Gaze+ [341]	Kitchens	RGB, Gaze	2.4M	86	-	32
ADL [276]	Daily Activities	RGB	1M	20	42	20
THU-READ [335]	Daily Activities	RGB, Depth	343,626	1920	-	8
MECCANO Multimodal [287]	Industrial	RGB, Depth, Gaze	299,376	20	20	20
Assembly101 [312]	Industrial	RGB, Multi-View, 3D Hand-Pose	20M	362	90	53
EGO-CH [286]	Cultural Sites	RGB	867,957	120	226	70
EGO-CH-Gaze [250]	Cultural Sites	RGB, Gaze	220,300	14	15	7
Sanctuaria-Gaze (Ours)	Cultural Sites	RGB, Gaze, Depth, IMU	241,355	40	30	20

Table 5.6: Comparison of Sanctuaria-Gaze with other egocentric datasets.

Subject ID	Group 01										Group 02									
	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20
Belief	NC	C	NC	NC	NC	NC	C	C	C	C	NC	C	C	NC	NC	C	NC	C	C	NC
Visited	-	-	02	-	-	-	01-02	-	-	01	04	03-04	-	03	-	-	-	03-04	03-04	04
Vision	N	N	C	C	N	N	C	C	N	C	C	C	C	N	C	N	C	C	C	N

Table 5.7: Information about participants. *Belief* indicates whether the participant identifies as Catholic (C) or non-Catholic (NC). *Visited* specifies the ID of the church visited in the past by the participant, if applicable. *Vision* refers to normal (N) or corrected-to-normal vision (C).

tended objects in cultural environments. While these datasets have driven significant progress in egocentric vision, a focus on religious environments remains largely unexplored. Our work addresses this gap by introducing a new egocentric dataset of visits to religious sites, equipped with detailed gaze data.

In Table 5.6, we report a comparison of Sanctuaria-Gaze with other egocentric datasets. To the best of our knowledge, this is the largest dataset within this domain, encompassing different data signals. It offers unique insights into human behavior and visual attention in sacred spaces, thereby bridging egocentric vision research with the study of cultural and religious heritage.

5.3.2 SANCTUARIA-GAZE DATASET

We introduce Sanctuaria-Gaze, a novel dataset of egocentric videos, equipped with eye-tracking data, captured from visitors exploring four religious sites in Northern Italy while wearing wearable devices. Along with the dataset, we also release an open-source framework designed for the automatic extraction and analysis of AOIs, representing locations that captured the user’s attention. This framework not only enables the identification of AOIs but also provides tools for visualizing attention patterns, assessing attention shifts over time, and conducting comparative analyses across different individuals and sites.

DATASET ACQUISITION

RELIGIOUS SITES. Our dataset was collected in four religious sites in Northern Italy: *Sanctuary of Bismantova*, located in Reggio Emilia (Church 01), *Sanctuary of San Pellegrino in Alpe*, located in Lucca (Church 02), *Sanctuary of Santa Maria della Vita* (Church 03) and *Sanctuary of Santi Vitale e Agricola*

(Church 04), both located in Bologna. All these buildings differ in various aspects, including their size, the number of rooms, and their architectural style. Notably, Church 01 and Church 02 are mountain hermitages, while Church 03 and Church 04 exhibit rich historical and artistic features.

PARTICIPANTS. The Sanctuaria-Gaze dataset was collected from 20 unique participants with normal or corrected-to-normal vision, over two separate days. They were divided into two groups of 10 people each. The first group visited Church 01 and Church 02, while the second explored Church 03 and Church 04. The participants were selected by guaranteeing diversity in terms of age, religious belief, and gender. Their ages range from 18 to 65 years. Among all the visitors, two religious groups can be identified: Catholics (C) and non-Catholics (NC), a distinction that formed the basis for the analysis presented in the following sections. It is worth noting that some participants may have previously visited one or more of the sanctuaries involved in the data collection process. Table 5.7 reports detailed information about the participants. The study adhered to ethical guidelines for the protection of human subjects, in compliance with the legal requirements of the country, and obtained formal approval from the Ethical Committee of the participating research institutions.

WEARABLE DEVICES. Participants were required to wear a pair of Project Aria Glasses [127], essential for the recording. Each video was captured at a resolution of 1408×1408 pixels with a frame rate of 15 fps. To facilitate the study of human visual attention and analyze participants' areas of interest, gaze data were collected from the eye-tracking cameras integrated into the glasses, operating at a rate of 30 fps. SLAM cameras were enabled at a frequency of 15 fps, while IMU sensors worked at a frequency of 1000 Hz and 800 Hz, respectively. Data from a magnetometer operating at 10 Hz and a barometer operating at 50 Hz were also recorded to capture additional environmental and spatial data.

EXPERIMENTAL PROTOCOL. Each participant was instructed to freely explore the church, simulating a typical visit, while wearing Project Aria Glasses. Participants were given complete autonomy in deciding their movements and activities within the church. Before each experiment, the wearable device was calibrated according to the standard Project Aria calibration procedure. The duration of each

experiment was limited to a maximum of ten minutes, with no restrictions on the minimum duration. To ensure an undisturbed experience, each participant conducted the exploration individually, thus avoiding potential distractions deriving from the simultaneous presence of other people.

COLLECTED DATA

We collected 40 egocentric videos with an average duration of 6.7 minutes, for a total of 4.47 hours of recordings. The recorded gaze data revealed different viewing patterns across the four churches, as shown in Fig. 5.14. The gaze heatmaps illustrate different patterns of visual attention, with Church 01 and Church 03 showing a centralized focus, while the others demonstrate rightward-shifted distributions. This suggests that architectural elements or points of interest in the latter churches may have drawn participants' attention toward the right side of their visual field. The fixation duration distributions follow a consistent pattern across all churches, with a predominance of shorter fixations and a long tail of longer durations, typical of natural viewing behavior.

However, Church 04 shows the highest frequency of shorter fixations, indicating either a greater number of visual elements prompting rapid visual exploration, or alternatively, a more intuitive spatial organization requiring less prolonged examination. These findings suggest that architectural richness may influence not only where people look, but also how they temporally structure their visual exploration of sacred spaces.

PROPOSED FRAMEWORK FOR AUTOMATIC AOIs ANNOTATION

In addition to the dataset, we propose a methodological framework for the automatic analysis of the collected data, enabling the study of human visual attention within religious spaces. Given an egocentric video V with N frames of size $H \times W$ and a synchronized stream of gaze coordinates $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$, where each $\mathbf{x}_i \in \mathbb{R}^2$ represents a 2D gaze point (x, y) in the frame, the objective is to automatically annotate the objects fixated by the observer. These annotations include a segmentation mask $M \in \{0, 1\}^{H \times W}$, representing the pixel region of

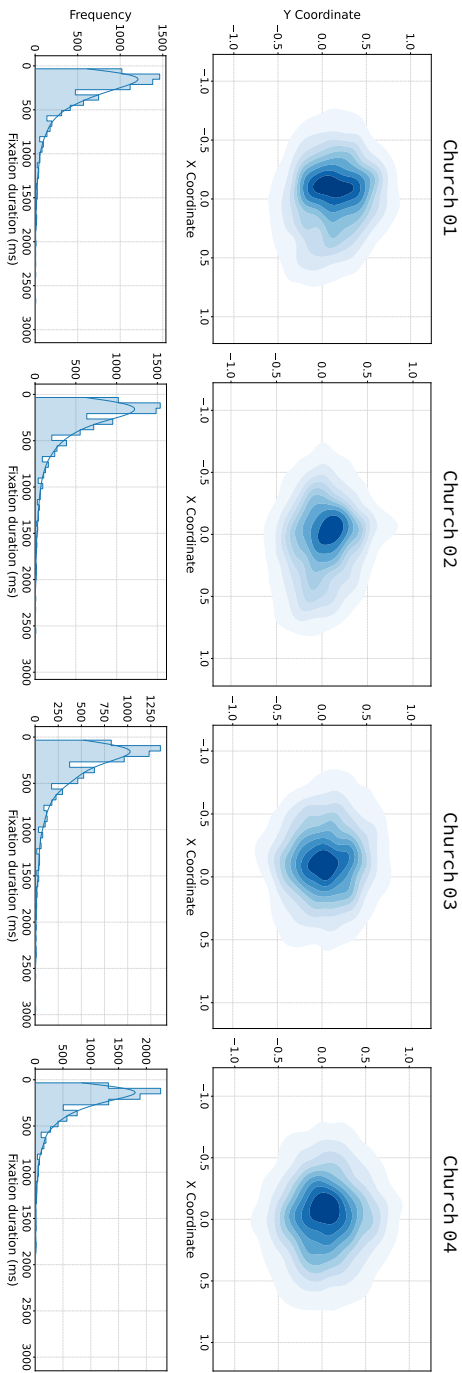


Figure 5.14: Gaze heatmap (top) and distribution of the fixation durations (bottom) for the four churches. The top row shows the spatial distribution of participant gaze patterns, with darker blue indicating higher gaze concentration. The bottom row displays histograms of fixation durations (in milliseconds), indicating the frequency and distribution of gaze fixation lengths.

the fixated object, a bounding box $B \in \mathbb{R}^4$ and a category label $C \in \mathcal{L}$, where $\mathcal{L}$ is the set of possible object labels. An overview of the proposed annotation pipeline is depicted in Fig. 5.15 and summarized in the following steps.

FIXATION DETECTION. A fixation detection algorithm $f : \mathcal{X} \rightarrow \mathcal{E}$ is applied to classify each gaze point as either a fixation or a saccade. Here, $\mathcal{E} = \{e_1, e_2, \dots, e_N\}$, $e_i \in \{0, 1\}$, where $e_i = 1$ if $\mathbf{x}_i$ is classified as a fixation, and $e_i = 0$ otherwise.

FRAME FILTERING. To reduce the computational burden, frames corresponding to non-fixation points ($e_i = 0$) are discarded. For consecutive fixation points, only the first associated frame is retained. The resulting subset of frames is denoted by $\mathcal{F} = \{f_j \mid j \in \mathcal{I}\}$, $\mathcal{I} = \{i \mid e_i = 1, e_{i-1} = 0 \text{ or } i = 1\}$, where $\mathcal{I}$ is the index set of retained frames.

SEGMENTATION MASK EXTRACTION. For each retained frame f_j , a segmentation model $g : (f_j, \mathbf{x}_j) \rightarrow M_j$ is employed to extract the segmentation mask $M_j \in \{0, 1\}^{H \times W}$ corresponding to the attended object, where $\mathbf{x}_j$ is the fixation point.

OBJECT DETECTION AND LABELING. In parallel with the segmentation model, an open-vocabulary object detector $h : (f_j, \mathcal{L}) \rightarrow \{(B_j^k, C_j^k) \mid k = 1, \dots, K\}$ is applied to detect K candidate objects in frame f_j , where each candidate object is associated with a bounding box $B_j^k \in \mathbb{R}^4$, and a label $C_j^k \in \mathcal{L}$.

OBJECT FILTERING. The final object annotation is selected by finding the detected bounding box B_j^k with the greatest overlap with the segmentation mask M_j , provided the overlap (IoU) exceeds a threshold τ . This can be formalized as:

$$k^* = \arg \max_{k \in \mathcal{K}} \text{IoU}(B_j^k, M_j), \quad \mathcal{K} = \{k \mid \text{IoU}(B_j^k, M_j) \geq \tau\},$$

$$\text{where } \text{IoU}(B_j^k, M_j) = \frac{|B_j^k \cap M_j|}{|B_j^k \cup M_j|}$$

If no bounding box satisfies the threshold τ , the frame is discarded, otherwise the object annotation is $(B_j^{k^*}, C_j^{k^*})$.

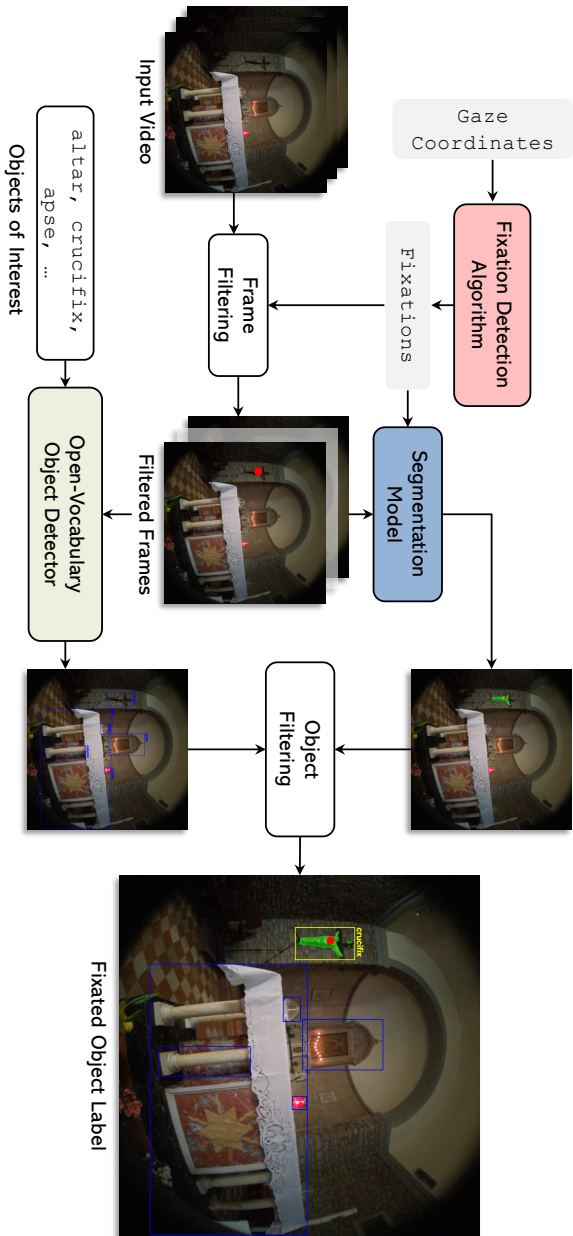


Figure 5.15: Pipeline of the proposed annotation framework. Given an egocentric video and a stream of gaze coordinates, frames containing a fixation point are selected and passed through a segmentation model to generate a mask for the attended object. Simultaneously, an open-vocabulary object detector identifies bounding boxes for all objects of interest within the scene. The bounding box with the highest overlap with the segmentation mask is selected, and its associated category label is assigned as the final object annotation.

5.3.3 DATASET ANALYSIS

The proposed framework is implemented and released as an open-source tool. In particular, we evaluate the efficacy of our pipeline on the Sanctuaria-Gaze dataset, focusing on two primary areas: the behaviors exhibited by individual participants and the group-level dynamics, with particular attention on the most frequently fixated object categories.

DATA PROCESSING

Each recorded session is initially processed using the Project Aria Machine Perception Services (MPS) [127]. Specifically, the eye gaze service extracts the gaze directions (yaw values) of the left and right eyes, as well as the corresponding depth of intersection, based on images captured by the eye-tracking cameras. These data are then projected into camera space to compute 2D gaze coordinates.

Subsequently, we apply the I-DT fixation detection algorithm [305], using a gaze-dispersion threshold equal to 0.05 in normalized pixel distance within the frame and a time threshold of 100 milliseconds. This process results in a total of 36,990 fixations, which are used to guide the filtering of an equivalent number of frames extracted from the RGB recordings. These frames, combined with the corresponding gaze coordinates, serve as input to the segmentation model. We adopt *Segment Anything Model* (SAM 2.1 Base Plus) [291] for object segmentation. This model represents the current state-of-the-art in segmentation and is specifically designed to support point-based prompts, making it a natural fit for our use case where each gaze fixation corresponds to a single 2D coordinate. This alignment allows for an accurate extraction of object masks centered on the user’s visual attention. In addition, given a frame f_j from a video V , and a textual vocabulary $\mathcal{L}$ containing the object names to detect, we use an open-vocabulary detector (*i.e.*, YOLO-World [81]) to localize all instances present in the scene and included in $\mathcal{L}$. The choice of YOLO-World over conventional closed-vocabulary detectors is motivated by the specificity and semantic richness of the object categories typically found in religious heritage sites. Standard object detectors are unlikely to recognize many of the relevant classes (*e.g.*, *crucifix*, *altar*, *relic*)

without retraining. In contrast, YOLO-World allows flexible detection via textual prompts, making it well-suited to handle the domain-specific vocabulary required for our study.

The vocabulary is composed of six macro-categories for a total of 30 objects of interest grouped as follows:

- **Architectural Features:** apse, arch, column, nave.
- **Contextual Objects:** bag, camera, lighting fixture, phone, sign.
- **Cultural and Historical Features:** book, painting, plaque, relic, tapestry.
- **Furnishings:** bench, candelabrum, chair, lamp, lectern, offering box.
- **Religious Items:** altar, candle, crucifix, holy water font, pulpit, statue, tabernacle.
- **Temporal or Event-Based Objects:** flower, musical instrument, seasonal decoration.

The set of objects composing the vocabulary $\mathcal{L}$ was chosen in collaboration with experts in religious studies, considering the content of the four sanctuaries where the experiments were carried out. The list of bounding boxes extracted during this phase is filtered using the segmentation mask, with $\tau = 0.40$. This process results in a total of 15,798 classified observed objects, forming the baseline for further analysis at both individual and group levels.

INDIVIDUAL ANALYSIS

An analysis of individual visitor behavior in the four churches reveals distinct patterns of object observation that appear to correlate with the distinctive characteristics of each church. The data, shown in Fig. 5.16, demonstrate how architectural relevance and religious importance influence visitors' attention patterns. In Church 03 and Church 04, distinguished by their architectural significance, observations of architectural elements (*e.g.*, apse, arch, column, and nave) account for an average of $43\% \pm 12\%$ of visitor's attention, a proportion that is significantly higher than in the other churches ($t(38) = -7.73$, $p < 0.001$). This architectural focus suggests that visitors were naturally drawn to the design features of the buildings. Conversely, the mountain hermitages of Church 01

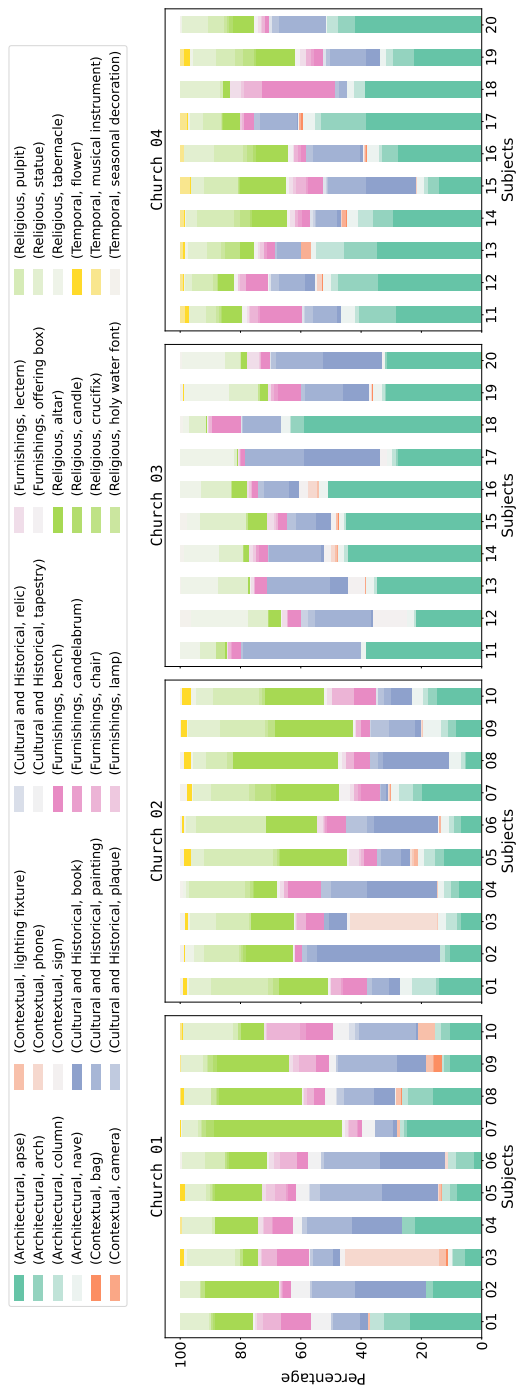


Figure 5.16: Distribution of object categories across 20 different subjects in four churches. Each plot represents the percentage composition of different objects in different churches. Each macro-category is further subdivided into specific object categories.

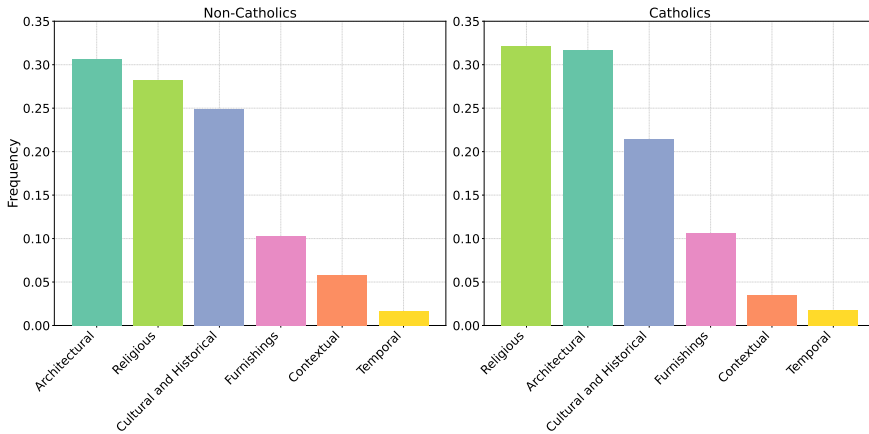


Figure 5.17: Comparison of descriptive category frequencies between Catholic and non-Catholic participants, aggregated over the four churches. Left and right plots show frequency distributions across six object macro-categories.

and Church 02 exhibit the highest proportion of religious object observations compared to the other churches ($t(38) = 4.30, p < 0.001$), particularly towards the altar, averaging $16\% \pm 14\%$ of total observations, possibly reflecting their more contemplative, spiritually-focused settings. These patterns suggest that while visitors develop personal exploration strategies, their attention is significantly influenced by the distinct characteristics and primary points of interest of each church. The mountainous environments of Church 01 and Church 02 appear to encourage a heightened focus on religious elements, while the architectural richness of Church 03 and Church 04 naturally draws attention to structural features, demonstrating how the physical and spiritual attributes of sacred spaces can guide visitor engagement.

GROUP ANALYSIS

For the purpose of the analysis and based on participants' self-reported religious affiliation in the pre-visit questionnaire, subjects were clustered into two groups for each visit: Catholics and non-Catholics, with five participants for each group. The non-Catholic group included both individuals from other religious backgrounds and those reporting no religious affiliation.

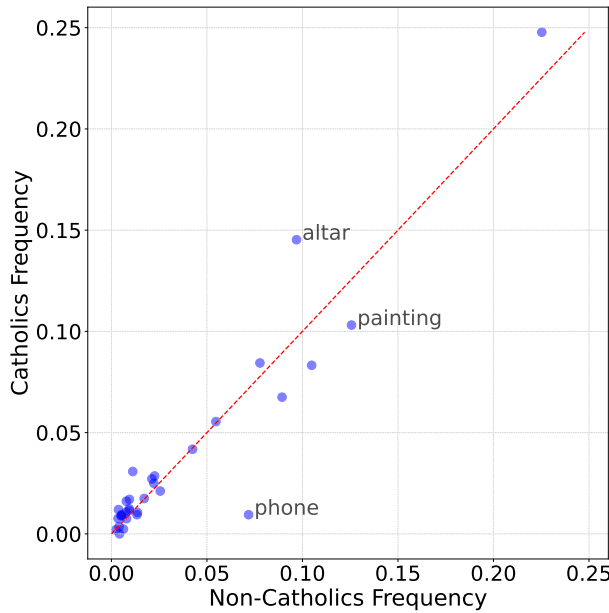


Figure 5.18: Correlation between Catholic and non-Catholic group frequencies of object categories, with the diagonal line representing equal frequencies between groups. Object labels are shown for the top-3 differences among the two groups.

While this binary classification simplifies a more complex spectrum of religious beliefs, it provides a meaningful starting point for investigating the relationship between religious identity and visitor behavior in sacred spaces. The choice of this categorization is grounded in the context of the sanctuaries involved, which are all sites of Catholic religious significance. Catholic participants may approach these spaces with culturally ingrained expectations and emotional resonance, whereas non-Catholic visitors might engage from a more observational perspective. While acknowledging the heterogeneity within each group (*e.g.*, varying levels of religiosity or spiritual orientation), this binary distinction enables an exploratory analysis of how religious affiliation may influence egocentric visual attention patterns. Further studies could incorporate more fine-grained categorizations (*e.g.*, practicing vs. non-practicing, interfaith backgrounds), but such an analysis is beyond the scope of this study.

Analysis of visitor behavior stratified by religious background reveals notable

similarities and distinct differences in observation patterns between Catholic and non-Catholic participants. At the macro-category level (Fig. 5.17), both groups show comparable attention to *Architectural* elements with an average of 31% of observations. The most significant discrepancy emerges in the *Religious* category, where Catholic visitors exhibit significantly higher levels of attention (33% of observations) compared to non-Catholic visitors (28%). This finding suggests that religious literacy and personal connection to Catholic symbolism may influence viewing patterns. A more detailed analysis at the object-specific level, as depicted in Fig. 5.18, reveals several salient differentiating elements. The altar emerges as one of the most discriminating objects, along with paintings and phones.

To quantify the magnitude of such differences, we compute Cohen's d effect sizes (see Fig. 5.19) and relative percentage changes for each object category. The strongest effect sizes are observed in the usage of *phones* ($d = -1.26$, 86.8% less frequent among Catholics) and attention to *altars* ($d = 0.98$, 49.9% more frequent among Catholics). The negative effect size for phone usage suggests that non-Catholic visitors were substantially more likely to use their phones during the visit, possibly for documentation purposes. Conversely, the large positive effect size for altar observation indicates Catholics' greater engagement with this central liturgical element. Moderate effects are found for *paintings* ($d = -0.46$), *apses* ($d = 0.45$), *pulpits* ($d = -0.44$), and *books* ($d = -0.44$).

Interestingly, while some religious elements show strong differences (altars), others display more modest effects (*crucifixes*: $d = 0.15$, *statues*: $d = 0.13$), suggesting that the influence of religious background varies across different types of sacred objects. Several categories show relevant percentage changes despite smaller effect sizes. For instance, attention to *relics* shows a 223.5% increase among Catholics ($d = 0.17$), and *signs* show a 176% increase ($d = 0.40$). Seasonal decorations show the maximum possible negative change (−100%), though with a relatively small effect size ($d = -0.09$), indicating complete avoidance by one group but high variability within groups. These attention patterns suggest that while religious background strongly influences certain aspects of church exploration, such as documentation behavior and attention to central liturgical elements, its effect on observation of other church features is more nu-

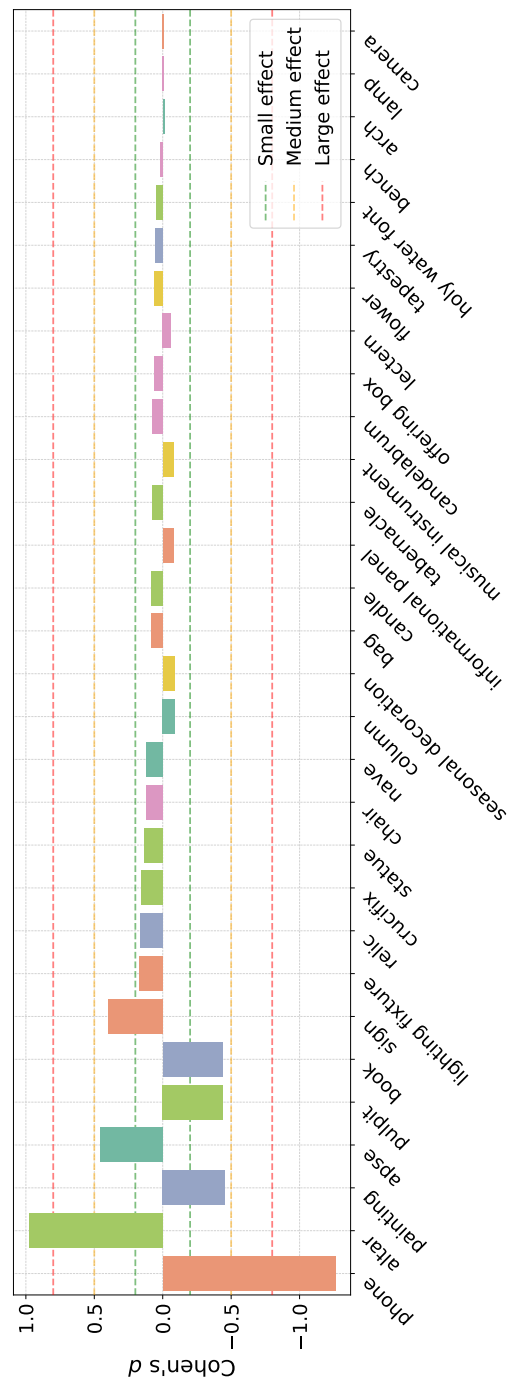


Figure 5.19: Cohen's d effect sizes for different objects, showing differences between Catholic and non-Catholic groups. Horizontal dashed lines indicate conventional thresholds for small (green), medium (orange), and large (red) effect sizes. Bars are colored by macro-category, with positive values indicating higher frequency in Catholic group and negative values indicating higher frequency in non-Catholic group.

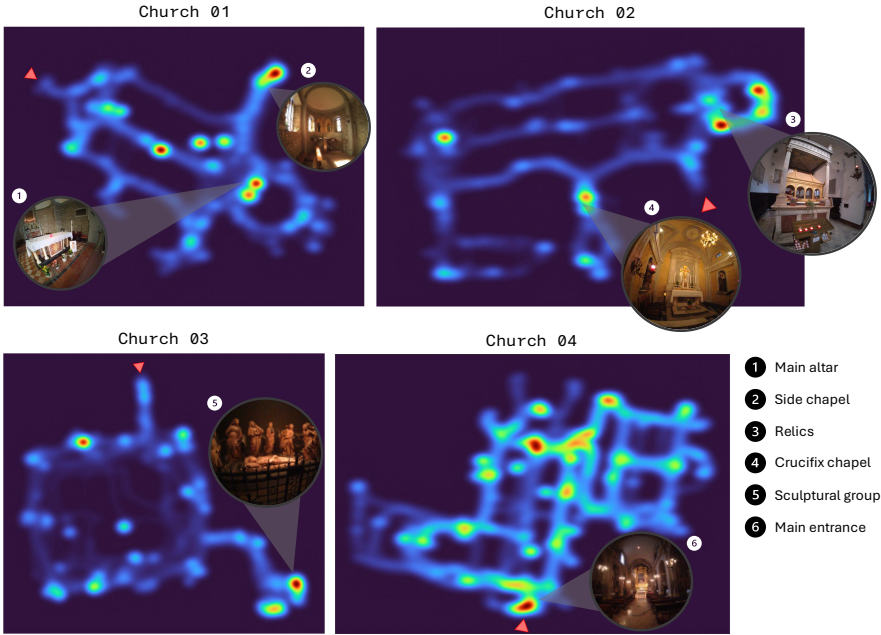


Figure 5.20: Participants movement trajectories throughout each church space, where warmer colors (yellow-red) indicate longer dwelling times and cooler colors (blue) represent transit areas. The red triangles mark the entrance locations.

anced. The discrepancy between effect sizes and percentage changes for certain objects, such as relics, suggests high individual variability in observation patterns within both groups.

5.3.4 ANALYSIS OF THE EXPLORATION PATTERNS

The presence of the SLAM cameras integrated into the Project Aria Glasses allows us to capture the visitor movement trajectories inside the churches. Fig. 5.20 illustrates the regions where the visitors spent their time during the exploration of the sanctuary. The resulting positional heatmaps, which aggregate the movement data of ten participants per church, reveal varying degrees of spatial coverage, with some areas receiving more attention across multiple participants. These areas often correspond to key architectural or religious features, as supported by the compositional analysis presented in the previous sections.

For instance, Church 01 shows concentrated movement around the main altar, a side chapel illuminated by natural light, and the central nave offering a comprehensive view of the interior. Similarly, in Church 02, movement was primarily directed toward three key areas: the relics of the saints, an alternative entrance from which the entire building could be viewed, and the chapel with the crucifix. In contrast, Church 03, characterized by its circular plan and architectural richness, prompted visitors to move across various chapels, with particular density around the sculptural group placed at the bottom-right corner of the map. Church 04 also elicited broader spatial engagement, with movement distributed across areas housing the many artworks present.

These findings, combined with the ones presented in the sections above, suggest that spatial experience in sacred spaces is modulated not only by the architectural affordances but also by the distribution and nature of the visual and devotional stimuli, which together influence both movement and attention during exploration. Simpler and more centralized layouts tend to focus movement around a few key points, while more complex and ornamented environments encourage broader and more varied trajectories.

LIMITATIONS

While the proposed framework for AOIs detection enables large-scale, gaze-driven object annotation in complex environments such as religious heritage sites, it also has some limitations. First, the accuracy of gaze-to-object association depends on the precision of the eye-tracking system and the fixation detection algorithm. In particular, the adopted I-DT algorithm [305], though widely used in traditional screen-based setups, may not be ideally suited for egocentric vision, where natural head and body movements introduce noise and require compensation mechanisms that the algorithm does not account for [7]. Second, although the use of SAM 2.1 and YOLO-World allows for flexible and domain-adaptable annotation, the segmentation quality may degrade when objects are occluded, poorly illuminated, or not visually salient. Additionally, the open-vocabulary detection performance of YOLO-World is influenced by the specificity and clarity of the textual prompts, which can introduce variability across instances. Finally,

the object vocabulary used in this study is domain-specific and manually curated with expert input. Adapting the framework to other domains would require the definition of new object categories and potentially further prompt engineering or detector fine-tuning.

Despite these limitations, the modularity of our approach facilitates future extensions and targeted improvements, such as integrating more robust fixation detection tailored to egocentric data or enhancing detection with context-aware language models.

6

Conclusions

This thesis investigated the role of human visual attention from both computational and application-driven perspectives, examining how gaze modeling can advance our understanding of visual behavior and enhance intelligent systems. Key limitations of existing computational attention models were addressed, particularly with respect to temporal dynamics and behavioral variability. By framing gaze as a sequential and inherently stochastic process, the proposed approaches move beyond static and deterministic formulations, enabling richer representations of attention mechanisms and more principled evaluation strategies. The proposed research also demonstrated how human gaze can serve as a powerful supervisory signal across diverse domains. From deepfake detection to neurodevelopmental behavioral analysis and attention modeling in ecologically valid cultural environments.

6.1 SUMMARY OF CONTRIBUTIONS

This thesis is organized around two complementary research directions, encompassing both the design and development of new computational models and the application of gaze-informed methodologies to societally relevant domains.

In Chapter 3, we focused on computational models of human visual attention, treating gaze as a dynamic cognitive process to be modeled. The lack of architectures that explicitly model the temporal evolution of visual attention and its variability across observers motivated the introduction of two novel frameworks for scanpath prediction, presented in detail in Chapter 4. First, TPP-Gaze formulates gaze behavior within the mathematical framework of Neural Temporal Point Processes, enabling explicit modeling of both fixation locations and their durations. By treating scanpaths as sequences of temporally dependent events, the proposed approach captures the spatio-temporal evolution of attention in a principled manner, achieving state-of-the-art performance in scanpath similarity and fixation duration prediction while exhibiting human-like gaze dynamics.

Building on the need to account for variability in visual exploration, ScanDiff reframes scanpath prediction as a stochastic generative problem through diffusion-based modeling. By explicitly representing the probabilistic nature of gaze allocation, this approach moves beyond deterministic formulations and produces diverse yet plausible scanpaths that better capture the variability inherent in human visual exploration. Modeling such diversity is beneficial not only for predictive accuracy but also essential for downstream applications requiring realistic simulation of human visual behavior, such as human-computer interaction, autonomous systems, and cognitive robotics.

Complementing the methodological advances, the second part of the thesis further advances the field, demonstrating how gaze can serve as a learning signal for specific applications. Chapter 4 offers a comprehensive overview of the current state of gaze-assisted AI, highlighting both its theoretical foundations and practical implications. In Chapter 5, we present novel contributions in the field of deepfake detection, neurodevelopmental research, and the collection of new multimodal data resources for studying attention in culturally significant environments. In the context of media forensics, the work investigates how human observers visually perceive manipulated images, revealing systematic differences in attention allocation between authentic and forged content. Our findings reveal that when humans examine counterfeit images, their attention tends to be directed toward more confined regions, in contrast to genuine images, where the

observed visual pattern is more evenly distributed across the presented stimuli. Such results were supported through statistical tests conducted on the entropy distributions of the saliency maps, thereby confirming our initial hypothesis. Building on these findings, we demonstrated the efficacy of integrating human visual attention into the automatic detection pipeline of manipulated images. Our experiments strongly support the hypothesis that human-derived saliency maps can significantly enhance the performance of deepfake detection models, particularly for images that are visually obvious to human observers but challenging for current algorithms.

The introduction of our novel testing dataset, specifically designed for religious visual data, has provided valuable insights into the unique challenges posed by partially modified images in this domain. Our research demonstrates that even a small amount of eye-tracking data can yield substantial improvements compared to relying solely on generated data from saliency models. This finding underscores the value of human-in-the-loop approaches in AI development and suggests that expanding this dataset would be highly beneficial for further advancements in the field. Future work should prioritize the collection of more extensive eye-tracking data across diverse scenarios and cultural contexts to enhance the robustness and generalizability of these models. As generative AI technologies continue to evolve, the approach presented in this thesis offers a promising direction for future research, suggesting that further exploration of human-AI collaboration could lead to more robust, context-aware, and ethically aligned deepfake detection systems.

Beyond media analysis, the thesis explores gaze as a behavioral biomarker in neurodevelopmental research. We proposed a framework based on Visual Attention Causal Graphs (VA-CGs), which captures how attention dynamically shifts between semantically relevant items in a visual scene. Experiments demonstrate the potential of VA-CGs in terms of classification performance compared to existing methods. Crucially, beyond mere classification, the proposed method offers improved interpretability. By explicitly modeling causal relationships, our framework provides insights into the distinct attentional strategies employed by different groups. This capability may be essential for advancing our understanding

of the cognitive processes that differentiate ASD from TD individuals. Finally, the thesis contributes a large-scale multimodal egocentric dataset collected in religious environments, providing an ecologically valid resource for studying attention allocation in complex real-world settings. The dataset is complemented by an open-source framework designed for the automatic annotation of fixated objects and the analysis of areas of interest (AOIs) in the spatial environment. Our analyses reveal how movement and exploration patterns vary across churches, influenced both by the architectural configuration of each site and by the participants' cultural and religious backgrounds. Such findings suggest that sacred spaces elicit differentiated forms of spatial and visual engagement, shaped by the historical, symbolic, and aesthetic affordances of the environment. By combining gaze, video, and spatial trajectories, this dataset supports interdisciplinary research across computer vision, human–computer interaction, and cultural studies. Sanctuaria-Gaze not only serves as a benchmark dataset for visual and behavioral analysis in religious settings, but also as a methodological bridge between computational techniques and the interpretive work of the humanities. Future work may expand this research to include larger participant groups and specific ritual contexts, such as pilgrimages or feast days, allowing for comparative studies across time, space, and beliefs.

6.2 OPEN CHALLENGES AND FUTURE DIRECTIONS

Throughout this thesis, we have shown that understanding human visual attention constitutes a multifaceted challenge, spanning principled computational modeling and real-world deployment. Observable manifestations of attention, such as gaze dynamics and eye movement patterns, provide rich and structured signals that can enhance both the performance and interpretability of AI systems. However, incorporating this information into AI models still represents a major challenge.

DATA SCARCITY. One possible reason for this is the scarcity and cost of collecting human gaze data. When it comes to training machine learning models that rely on human attention, the shortage of data poses a significant challenge. One

possible solution is to try to reduce the cost of eye-tracking hardware and make data collection less dependent on human instructors. This might be fostered by the increased interest and proliferation of wearable devices, such as smart glasses, virtual and augmented reality headsets, which would allow large amounts of data to be collected in an easy way and, most importantly, in an ecological setting. However, this is likely to raise a new challenge concerning ethical and privacy-aware data collection and sharing.

PRIVACY ISSUES. Currently, there is a lack of research into the privacy issues raised by eye tracking on a broad scale [150]. To better understand the privacy risks inherent in naturalistic real-world situations, it is essential to focus on eye-tracking datasets that document extensive gaze data during typical daily activities. A notable example of such datasets is Ego4D [149]. It contains 3,670 hours of daily-life activity videos together with eye-tracking data. This approach is unique compared to most research datasets. The analysis involves examining the real-world implications of eye tracking and exploring the dynamics of privacy-sensitive situations that may occur during everyday activities. The focus is on actual situations rather than artificial scenarios. By carefully examining these instances and assessing the level of privacy risks from the user's perspective, it would be possible to gain valuable insights into the immediate concerns related to eye tracking. This can establish a proactive understanding of potential long-term risks, enabling the development of strategies and safeguards to anticipate, mitigate, and prevent privacy-related challenges in the evolving landscape of eye-tracking technologies.

SYNTHETIC DATA. Another way to address both the aforementioned problem and that of data scarcity is to use synthetic eye movements to supplement the existing data sources. Researchers have developed models that integrate synthetic scanpath generation with a scanpath-augmented language model, eliminating the need for human gaze data [111, 200]. This approach has shown promising results for natural language processing tasks, while minimal exploration has been undertaken in computer vision.

PERSONALIZATION. The personalization of gaze-informed models raises several methodological and practical challenges. First, learning subject-specific rep-

resentations requires sufficient individual-level data, which exacerbates the already present issue of data scarcity. Second, personalization mechanisms must generalize to unseen users, requiring meta-learning or few-shot strategies capable of rapidly adapting to new behavioral patterns. While recent works [75, 189, 378] have started exploring possible solutions in this direction, such approaches risk overfitting to unique traits and neglecting shared attention patterns, underscoring the need to balance deterministic and stochastic components of gaze behavior. Third, evaluation protocols should also move beyond aggregate metrics that favor averaged patterns and instead assess how well models capture meaningful variability across individuals. From an application perspective, personalization is particularly relevant in domains such as assistive technologies, adaptive user interfaces, and immersive environments, where user-specific gaze modeling can enhance usability. However, balancing personalization with privacy preservation and fairness remains a critical concern, as gaze patterns can encode sensitive behavioral and cognitive information. Overall, advancing personalization in gaze-guided learning requires a shift toward uncertainty-aware frameworks that ensure ethical and responsible deployment.

WEARABLE DEVICES. Back to the spread of head-mounted displays like augmented and virtual reality headsets, by adapting the capabilities of vision-and-language models to stimuli obtained in real-time from the surrounding world, it would be possible to enhance the user experience and improve the applicability and effectiveness of AI agents in everyday scenarios [381]. This integration enables a focus on user preferences, providing precise and relevant responses by understanding intent, context, and individual interests. This approach is particularly beneficial for visually impaired individuals who rely on their gaze to communicate their intent.

Eventually, other related research areas present challenges that still need to be addressed, such as multimodal gaze following, reliable real-time gaze prediction, and gaze modeling for dynamic environments [143, 164]. Efforts spent in these research areas have the potential to significantly advance human gaze modeling and enable the development of more effective applications.

6.3 CLOSING REMARKS

In conclusion, this thesis advances the state-of-the-art in scanpath prediction while offering a principled framework to capture temporal structure and behavioral variability. Beyond methodological contributions, the work has demonstrated the value of gaze-informed learning across societally relevant domains. Together, these contributions highlight the importance of integrating human-centered signals into AI, fostering systems that are not only more accurate and robust but also more interpretable and aligned with real-world human behavior.

List of Publications

- [1] Baldrati, A., Morelli, D., Cartella, G., Cornia, M., Bertini, M., and Cucchiara, R. (2023). Multimodal garment designer: Human-centric latent diffusion models for fashion image editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [2] Cartella, G., Baldrati, A., Morelli, D., Cornia, M., Bertini, M., and Cucchiara, R. (2023). Openfashionclip: Vision-and-language contrastive learning with open-source fashion data. In *Proceedings of the International Conference on Image Analysis and Processing*.
- [3] Cartella, G., Cornia, M., Cuculo, V., D’Amelio, A., Zanca, D., Boccignone, G., and Cucchiara, R. (2024a). Trends, applications, and challenges in human attention modelling. In *Proceedings of the International Joint Conferences on Artificial Intelligence*.
- [4] Cartella, G., Cuculo, V., Cornia, M., and Cucchiara, R. (2024b). Unveiling the Truth: Exploring Human Gaze Patterns in Fake Images. *IEEE Signal Processing Letters*.
- [5] Cartella, G., Cuculo, V., Cornia, M., Papasidero, M., Ruozzi, F., and Cucchiara, R. (2024c). Pixels of Faith: Exploiting Visual Saliency to Detect Religious Image Manipulation. In *Proceedings of the European Conference on Computer Vision Workshops*.
- [6] Cartella, G., Cuculo, V., Cornia, M., Papasidero, M., Ruozzi, F., and Cucchiara, R. (2025a). Sanctuaria-Gaze: A Multimodal Egocentric Dataset for Human Attention Analysis in Religious Sites. *ACM Journal on Computing and Cultural Heritage*.
- [7] Cartella, G., Cuculo, V., D’Amelio, A., Cornia, M., Boccignone, G., and Cucchiara, R. (2025b). Modeling human gaze behavior with diffusion models for unified scanpath prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [8] Cartella, G., Cuculo, V., D’Amelio, A., Cucchiara, R., and Boccignone, G. (2025c). Unravelling neurodivergent gaze behaviour through visual attention

causal graphs. In *Proceedings of the International Conference on Image Analysis and Processing*.

- [9] D’Amelio, A., Cartella, G., Cuculo, V., Lucchi, M., Cornia, M., Cucchiara, R., and Boccignone, G. (2025). Tpp-gaze: Modelling gaze dynamics in space and time with neural temporal point processes. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.
- [10] Morelli, D., Baldrati, A., Cartella, G., Cornia, M., Bertini, M., and Cucchiara, R. (2023). Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. In *Proceedings of the ACM International Conference on Multimedia*.

Bibliography

- [1] Aberman, K., He, J., Gandelsman, Y., et al. (2022). Deep saliency prior for reducing visual distraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [2] Aboutaleb, H., Mao, D., Xu, C., and Wong, A. (2023). Deepfakeart challenge: A benchmark dataset for generative ai art forgery and data poisoning detection. *arXiv preprint arXiv:2306.01272*.
- [3] Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. (1985). A learning algorithm for boltzmann machines. *Cognitive Science*.
- [4] Adeli, H. and Zelinsky, G. (2018). Deep-bcn: Deep networks meet biased competition to create a brain-inspired model of attention control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.
- [5] Adhanom, I. B., MacNeilage, P., and Folmer, E. (2023). Eye tracking in virtual reality: A broad review of applications and challenges. *Virtual Reality*.
- [6] Alavi, H. S., Churchill, E. F., Wiberg, M., Lalanne, D., Dalsgaard, P., Fatah gen Schieck, A., and Rogers, Y. (2019). Introduction to human-building interaction (hbi): Interfacing hci with architecture and urban design. *ACM Transactions on Computer-Human Interaction*.
- [7] Alinaghi, N. and Giannopoulos, I. (2022). Consider the head movements! saccade computation in mobile eye-tracking. In *Proceedings of the Symposium on Eye Tracking Research & Applications*.
- [8] Aloimonos, J., Weiss, I., and Bandyopadhyay, A. (1987). Active vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [9] Amoroso, R., Morelli, D., Cornia, M., Baraldi, L., Del Bimbo, A., and Cucchiara, R. (2024). Parents and children: Distinguishing multimodal deep-fakes from natural images. *ACM Transactions on Multimedia Computing, Communications, and Applications*.

- [10] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [11] Aronov, B., Har-Peled, S., Knauer, C., Wang, Y., and Wenk, C. (2006). Fréchet distance for curves, revisited. In *European Symposium on Algorithms*.
- [12] Aronson, R. M. and Admoni, H. (2022). Gaze complements control input for goal prediction during assisted teleoperation. In *Robotics Science and Systems*.
- [13] Aronson, R. M., Almutlak, N., and Admoni, H. (2021). Inferring goals with gaze during teleoperated manipulation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*.
- [14] Aronson, R. M., Santini, T., Kübler, T. C., Kasneci, E., Srinivasa, S., and Admoni, H. (2018). Eye-hand behavior in human-robot shared manipulation. In *ACM/IEEE International Conference on Human-Robot Interaction*.
- [15] Assens, M., Giro-i Nieto, X., McGuinness, K., and O'Connor, N. E. (2018). Pathgan: Visual scanpath prediction with generative adversarial networks. In *Proceedings of the European Conference on Computer Vision Workshops*.
- [16] Assens Reina, M., Giro-i Nieto, X., McGuinness, K., and O'Connor, N. E. (2017). Saltinet: Scan-path prediction on 360 degree images using saliency volumes. In *Proceedings of the International Conference on Computer Vision Workshops*.
- [17] Aydemir, A., Sjöö, K., Folkesson, J., Pronobis, A., and Jensfelt, P. (2011). Search in the real world: Active visual object search based on spatial relations. In *Proceedings of the IEEE International Conference on Robotics and Automation*.
- [18] Aydemir, B., Bhattacharjee, D., Zhang, T., Kim, S., Salzmann, M., and Süssstrunk, S. (2022). Modeling object dissimilarity for deep saliency prediction. *Transactions on Machine Learning Research*.
- [19] Aydemir, B., Bhattacharjee, D., Zhang, T., Salzmann, M., and Süssstrunk, S. (2024). Data augmentation via latent diffusion for saliency prediction. In *Proceedings of the European Conference on Computer Vision*.

- [20] Aydemir, B., Hoffstetter, L., Zhang, T., Salzmann, M., and Süsstrunk, S. (2023). Tempsal - uncovering temporal information for deep saliency prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [21] Bacry, E., Mastromatteo, I., and Muzy, J.-F. (2015). Hawkes processes in finance. *Market Microstructure and Liquidity*.
- [22] Bae, S., Pakdamanian, E., Kim, I., Feng, L., Ordonez, V., and Barnes, L. (2021). Medirl: Predicting the visual attention of drivers via maximum entropy deep inverse reinforcement learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [23] Bajcsy, R. (1988). Active perception. *Proceedings of the IEEE*.
- [24] Bajcsy, R., Aloimonos, Y., and Tsotsos, J. K. (2018). Revisiting active perception. *Autonomous Robots*.
- [25] Bajcsy, R. and Campos, M. (1992). Active and exploratory perception. *CVGIP: Image Understanding*.
- [26] Ballard, D. H. (1991). Animate vision. *Artificial Intelligence*.
- [27] Baraldi, L., Cocchi, F., Cornia, M., Baraldi, L., Nicolosi, A., and Cucchiara, R. (2024). Contrasting deepfakes diffusion via contrastive learning and global-local similarities. In *Proceedings of the European Conference on Computer Vision*.
- [28] Barthelmé, S., Trukenbrod, H., Engbert, R., and Wichmann, F. (2013). Modeling fixation locations using spatial point processes. *Journal of Vision*.
- [29] Bauer, D., Wu, Q., and Ma, K.-L. (2022). Fovolnet: Fast volume rendering using foveated deep neural networks. *IEEE Transactions on Visualization and Computer Graphics*.
- [30] Berg, A. and Valaskivi, K. (2023). Representational silence and racial biases in commercial image recognition services in the context of religion. In *Handbook of Critical Studies of Artificial Intelligence*.
- [31] Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

- [32] Berzak, Y., Nakamura, C., Flynn, S., and Katz, B. (2017). Predicting native language from gaze. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.
- [33] Bhattacharya, M., Jain, S., and Prasanna, P. (2022). Radiotransformer: A cascaded global-focal transformer for visual attention-guided disease classification. In *Proceedings of the European Conference on Computer Vision*.
- [34] Billeci, L., Narzisi, A., Campatelli, G., Crifaci, G., Calderoni, S., Gagliano, A., Calzone, C., Colombi, C., Pioggia, G., and Muratori, F. (2016). Disentangling the initiation from the response in joint attention: an eye-tracking study in toddlers with autism spectrum disorders. *Translational psychiatry*.
- [35] Bisley, J. W. (2011). The neural basis of visual attention. *The Journal of physiology*.
- [36] Boccignone, G., Cuculo, V., and D’Amelio, A. (2019a). Problems with saliency maps. In *Proceedings of the International Conference on Image Analysis and Processing*.
- [37] Boccignone, G., Cuculo, V., and D’Amelio, A. (2019b). Problems with saliency maps. In *Proceedings of the International Conference on Image Analysis and Processing*.
- [38] Boccignone, G., Cuculo, V., and D’Amelio, A. (2020a). How to look next? a data-driven approach for scanpath prediction. In *Proceedings of the International Symposium on Formal Methods*.
- [39] Boccignone, G., Cuculo, V., D’Amelio, A., Grossi, G., and Lanzarotti, R. (2020b). On gaze deployment to audio-visual cues of social interactions. *IEEE Access*, 8:161630–161654.
- [40] Boccignone, G. and Ferraro, M. (2004a). Modelling gaze shift as a constrained random walk. *Physica A: Statistical Mechanics and its Applications*.
- [41] Boccignone, G. and Ferraro, M. (2004b). Modelling gaze shift as a constrained random walk. *Physica A: Statistical Mechanics and its Applications*.
- [42] Bogacz, R., Brown, E., Moehlis, J., Holmes, P., and Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*.

- [43] Bolliger, L. S., Reich, D. R., Haller, P., Jakobi, D. N., Prasse, P., and Jäger, L. A. (2023). Scandl: A diffusion model for generating synthetic scanpaths on texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- [44] Borji, A. (2012). Boosting bottom-up and top-down visual features for saliency estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [45] Borji, A. and Itti, L. (2015). Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581*.
- [46] Borji, A., Sihite, D. N., and Itti, L. (2012). Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*.
- [47] Borji, A., Tavakoli, H. R., Sihite, D. N., and Itti, L. (2013). Analysis of scores, datasets, and models in visual saliency prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [48] Boyd, A., Tinsley, P., Bowyer, K. W., and Czajka, A. (2023). Cyborg: Blending human saliency into the loss improves deep learning-based synthetic face detection. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.
- [49] Bozkir, E., Özdel, S., Wang, M., David-John, B., Gao, H., Butler, K., Jain, E., and Kasneci, E. (2023). Eye-tracked virtual reality: A comprehensive survey on methods and privacy challenges. *arXiv preprint arXiv:2305.14080*.
- [50] Brandt, S. A. and Stark, L. W. (1997). Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of Cognitive Neuroscience*.
- [51] Bray, A. and Schoenberg, F. P. (2013). Assessment of point process models for earthquake forecasting. *Statistical Science*.
- [52] Brockmann, D. and Geisel, T. (2000). The ecology of gaze shifts. *Neurocomputing*.
- [53] Brooks, T., Holynski, A., and Efros, A. A. (2023). Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

- [54] Bruce, N. and Tsotsos, J. (2005). Saliency based on information maximization. In *Advances in Neural Information Processing Systems*.
- [55] Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., and Torralba, A. (2012). Mit saliency benchmark. <http://saliency.mit.edu/>.
- [56] Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., and Durand, F. (2019). What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [57] Bylinskii, Z., Kim, N. W., O'Donovan, P., Alsheikh, S., Madan, S., Pfister, H., Durand, F., Russell, B., and Hertzmann, A. (2017). Learning visual importance for graphic designs and data visualizations. In *Proceedings of the ACM Symposium on User Interface Software and Technology*.
- [58] Caffagni, D., Cocchi, F., Barsellotti, L., Moratelli, N., Sarto, S., Baraldi, L., Baraldi, L., Cornia, M., and Cucchiara, R. (2024). The revolution of multimodal large language models: A survey. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, Findings*.
- [59] Canosa, R. L. (2009). Real-world vision: Selective perception and task. *ACM Transactions on Applied Perception*.
- [60] Cartella, G., Cornia, M., Cuculo, V., D'Amelio, A., Zanca, D., Boccignone, G., and Cucchiara, R. (2024a). Trends, applications, and challenges in human attention modelling. In *Proceedings of the International Joint Conferences on Artificial Intelligence*.
- [61] Cartella, G., Cuculo, V., Cornia, M., and Cucchiara, R. (2024b). Unveiling the truth: Exploring human gaze patterns in fake images. *IEEE Signal Processing Letters*.
- [62] Cartella, G., Cuculo, V., Cornia, M., Papasidero, M., Ruozzi, F., and Cucchiara, R. (2024c). Pixels of faith: Exploiting visual saliency to detect religious image manipulation. In *Proceedings of the European Conference on Computer Vision Workshops*.
- [63] Cartella, G., Cuculo, V., Cornia, M., Papasidero, M., Ruozzi, F., and Cucchiara, R. (2025a). Sanctuaria-gaze: A multimodal egocentric dataset for human attention analysis in religious sites. *Journal on Computing and Cultural Heritage*.

- [64] Cartella, G., Cuculo, V., D'Amelio, A., Cornia, M., Boccignone, G., and Cucchiara, R. (2025b). Modeling human gaze behavior with diffusion models for unified scanpath prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [65] Cartella, G., Cuculo, V., D'Amelio, A., Cucchiara, R., and Boccignone, G. (2025c). Unravelling neurodivergent gaze behaviour through visual attention causal graphs. In *Proceedings of the International Conference on Image Analysis and Processing*.
- [66] Castelhana, M. S., Mack, M. L., and Henderson, J. M. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision*.
- [67] Cerf, M., Frady, E. P., and Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision*.
- [68] Cerf, M., Harel, J., Einhäuser, W., and Koch, C. (2007). Predicting human gaze using low-level saliency combined with face detection. In *Advances in Neural Information Processing Systems*.
- [69] Cerf, M., Harel, J., Einhäuser, W., and Koch, C. (2008). Predicting human gaze using low-level saliency combined with face detection. In *Advances in Neural Information Processing Systems*.
- [70] Chakraborty, S., Wei, Z., Kelton, C., Ahn, S., Balasubramanian, A., Zelinsky, G. J., and Samaras, D. (2022). Predicting visual attention in graphic design documents. *IEEE Transactions on Multimedia*.
- [71] Chen, S., Jiang, M., Yang, J., and Zhao, Q. (2020). Air: Attention with reasoning capability. In *Proceedings of the European Conference on Computer Vision*.
- [72] Chen, S., Valliappan, N., Shen, S., Ye, X., Kohlhoff, K., and He, J. (2023a). Learning from unique perspectives: User-aware saliency modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [73] Chen, S. and Zhao, Q. (2019). Attention-based autism spectrum disorder screening with privileged modality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

- [74] Chen, X., Jiang, M., and Zhao, Q. (2021a). Predicting human scanpaths in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [75] Chen, X., Jiang, M., and Zhao, Q. (2024a). Beyond average: Individualized visual scanpath prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [76] Chen, X., Jiang, M., and Zhao, Q. (2024b). Gazexplain: Learning to predict natural language explanations of visual scanpaths. In *Proceedings of the European Conference on Computer Vision*.
- [77] Chen, Y., Nan, Z., and Xiang, T. (2023b). Fblnet: Feedback loop network for driver attention prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [78] Chen, Y., Yang, Z., Ahn, S., Samaras, D., Hoai, M., and Zelinsky, G. (2021b). Coco-search18 fixation dataset for predicting goal-directed attention control. *Scientific Reports*.
- [79] Chen, Y., Yang, Z., Chakraborty, S., Mondal, S., Ahn, S., Samaras, D., Hoai, M., and Zelinsky, G. (2022). Characterizing target-absent human attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.
- [80] Chen, Z. and Sun, W. (2018). Scanpath prediction for visual attention using ior-roI lstm. In *Proceedings of the International Joint Conferences on Artificial Intelligence*.
- [81] Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., and Shan, Y. (2024). Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [82] Chetouani, A. (2018). Convolutional neural network and saliency selection for blind image quality assessment. In *Proceedings of the IEEE International Conference on Image Processing*.
- [83] Chu, C., Zhang, H., Wang, P., and Lu, F. (2024). Simulating human mobility with a trajectory generation framework based on diffusion model. *International Journal of Geographical Information Science*.

- [84] Chu, T.-T., Chen, P.-H., Huang, P.-J., and Chen, K.-W. (2021). Collaborative learning of multiple-discontinuous-image saliency prediction for drone exploration. In *Proceedings of the IEEE International Conference on Robotics and Automation*.
- [85] Chunseong Park, C., Kim, B., and Kim, G. (2017). Attend to you: Personalized image captioning with context sequence memory networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [86] Cocchi, F., Baraldi, L., Poppi, S., Cornia, M., Baraldi, L., and Cucchiara, R. (2023). Unveiling the impact of image transformations on deepfake detection: An experimental analysis. In *Proceedings of the International Conference on Image Analysis and Processing*.
- [87] Coletta, A., Gopalakrishnan, S., Borrajo, D., and Vyetrenko, S. (2023). On the constrained time-series generation problem. In *Advances in Neural Information Processing Systems*.
- [88] Corbetta, M. and Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3(3):201–215.
- [89] Cornia, M., Baraldi, L., Serra, G., and Cucchiara, R. (2016). A deep multi-level network for saliency prediction. In *Proceedings of the International Conference on Pattern Recognition*.
- [90] Cornia, M., Baraldi, L., Serra, G., and Cucchiara, R. (2018a). Paying more attention to saliency: Image captioning with saliency and context attention. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 14(2):1–21.
- [91] Cornia, M., Baraldi, L., Serra, G., and Cucchiara, R. (2018b). Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing*, 27(10):5142–5154.
- [92] Cornia, M., Baraldi, L., Serra, G., and Cucchiara, R. (2018c). Sam: Pushing the limits of saliency prediction models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.
- [93] Corvi, R., Cozzolino, D., Poggi, G., Nagano, K., and Verdoliva, L. (2023a). Intriguing properties of synthetic images: From generative adversarial net-

- works to diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.
- [94] Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., and Verdoliva, L. (2023b). On the detection of synthetic images generated by diffusion models. In *International Conference on Acoustics, Speech, and Signal Processing*.
 - [95] Cozzolino, D., Poggi, G., Corvi, R., Nießner, M., and Verdoliva, L. (2024). Raising the bar of ai-generated image detection with clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.
 - [96] Crabbé, J., Huynh, N., Stanczuk, J., and Van Der Schaar, M. (2024). Time series diffusion in the frequency domain. In *Proceedings of the International Conference on Machine Learning*.
 - [97] Cristino, F., Mathôt, S., Theeuwes, J., and Gilchrist, I. D. (2010). Scanmatch: A novel method for comparing fixation sequences. *Behavior Research Methods*.
 - [98] Cuculo, V., D’Amelio, A., Lanzarotti, R., and Boccignone, G. (2018). Personality gaze patterns unveiled via automatic relevance determination. In *Federation of International Conferences on Software Technologies: Applications and Foundations*.
 - [99] Daley, D. J. and Vere-Jones, D. (2007). *An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure*. Springer Science & Business Media.
 - [100] Dalton, K. M., Nacewicz, B. M., Johnstone, T., Schaefer, H. S., Gernsbacher, M. A., Goldsmith, H. H., Alexander, A. L., and Davidson, R. J. (2005). Gaze fixation and the neural circuitry of face processing in autism. *Nature neuroscience*.
 - [101] D’Amelio, A. and Boccignone, G. (2021). Gazing at social interactions between foraging and decision theory. *Frontiers in Neurorobotics*.
 - [102] D’Amelio, A., Cartella, G., Cuculo, V., Lucchi, M., Cornia, M., Cucchiara, R., and Boccignone, G. (2025). Tpp-gaze: Modelling gaze dynamics in space and time with neural temporal point processes. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.

- [103] Damen, D., Doughty, H., Farinella, G. M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al. (2018). Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision*.
- [104] Damen, D., Doughty, H., Farinella, G. M., Furnari, A., Kazakos, E., Ma, J., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al. (2022). Rescaling egocentric vision: Collection pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*.
- [105] Dang, T., Papachristos, C., and Alexis, K. (2018). Visual saliency-aware receding horizon autonomous exploration with application to aerial robotics. In *Proceedings of the IEEE International Conference on Robotics and Automation*.
- [106] Darcet, T., Oquab, M., Mairal, J., and Bojanowski, P. (2024). Vision transformers need registers. In *Proceedings of the International Conference on Learning Representations*.
- [107] Das, A., Agrawal, H., Zitnick, L., Parikh, D., and Batra, D. (2016). Human attention in visual question answering: Do humans and deep networks look at the same regions? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- [108] De la Fuente Suárez, L. A. (2020). Subjective experience and visual attention to a historic building: A real-world eye-tracking study. *Frontiers of Architectural Research*.
- [109] Deane, O., Toth, E., and Yeo, S.-H. (2023). Deep-saga: A deep-learning-based system for automatic gaze annotation from eye-tracking data. *Behavior Research Methods*.
- [110] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [111] Deng, S., Prasse, P., Reich, D., Scheffer, T., and Jäger, L. (2023). Pre-trained language models augmented with synthetic scanpaths for natural language understanding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

- [112] Deng, S., Prasse, P., Reich, D. R., Dziemian, S., Stegenwallner-Schütz, M., Krakowczyk, D., Makowski, S., Langer, N., Scheffer, T., and Jäger, L. A. (2022). Detection of adhd based on eye movements during natural viewing. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*.
- [113] Desimone, R. and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual review of neuroscience*.
- [114] Dewhurst, R., Nyström, M., Jarodzka, H., Foulsham, T., Johansson, R., and Holmqvist, K. (2012). It depends on how you look at it: Scanpath comparison in multiple dimensions with multimatch, a vector-based approach. *Behavior Research Methods*.
- [115] Dhariwal, P. and Nichol, A. (2021). Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*.
- [116] Ding, Z., Ren, X., David, E., Vo, M., Kreiman, G., and Zhang, M. (2022). Efficient zero-shot visual search via target and context-aware transformer. *arXiv preprint arXiv:2211.13470*.
- [117] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*.
- [118] Droste, R., Jiao, J., and Noble, J. A. (2020). Unified image and video saliency modeling. In *Proceedings of the European Conference on Computer Vision*.
- [119] Du, N., Dai, H., Trivedi, R., Upadhyay, U., Gomez-Rodriguez, M., and Song, L. (2016). Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.
- [120] D’Amelio, A., Cartella, G., Cuculo, V., Lucchi, M., Cornia, M., Cucchiara, R., and Boccignone, G. (2025). Tpp-gaze: Modelling gaze dynamics in space and time with neural temporal point processes. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.
- [121] D’Amelio, A., Patania, S., Bursic, S., Cuculo, V., and Boccignone, G. (2023). Using gaze for behavioural biometrics. *Sensors*.

- [122] Ehinger, K. A., Hidalgo-Sotelo, B., Torralba, A., and Oliva, A. (2009). Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition*.
- [123] Einhäuser, W., Spain, M., and Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*.
- [124] Eiter, T., Mannila, H., et al. (1994). Computing discrete fréchet distance. *Technical Report*.
- [125] Emery, K. J., Zannoli, M., Warren, J., Xiao, L., and Talathi, S. S. (2021). Openneeds: A dataset of gaze, head, hand, and scene signals during exploration in open-ended vr environments. In *Proceedings of the Symposium on Eye Tracking Research & Applications*.
- [126] Engbert, R., Trukenbrod, H. A., Barthelmé, S., and Wichmann, F. A. (2015). Spatial statistics and attentional dynamics in scene viewing. *Journal of Vision*.
- [127] Engel, J., Somasundaram, K., Goesele, M., Sun, A., Gamino, A., Turner, A., Talattof, A., Yuan, A., Souti, B., Meredith, B., et al. (2023). Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*.
- [128] Enguehard, J., Busbridge, D., Bozson, A., Woodcock, C., and Hammerla, N. (2020). Neural temporal point processes for modelling electronic health records. In *Machine Learning for Health*.
- [129] Epstein, D. C., Jain, I., Wang, O., and Zhang, R. (2023). Online detection of ai-generated images. In *Proceedings of the International Conference on Computer Vision Workshops*.
- [130] Erdem, E. and Erdem, A. (2013). Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of Vision*.
- [131] Fan, D.-P., Wang, W., Cheng, M.-M., and Shen, J. (2019). Shifting more attention to video salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [132] Fathi, A., Li, Y., and Rehg, J. M. (2012). Learning to recognize daily actions using gaze. In *Proceedings of the European Conference on Computer Vision*.

- [133] Fosco, C., Casser, V., Bedi, A. K., O'Donovan, P., Hertzmann, A., and Bylinskii, Z. (2020a). Predicting visual importance across graphic design types. In *Proceedings of the ACM Symposium on User Interface Software and Technology*.
- [134] Fosco, C., Newman, A., Sukhum, P., Zhang, Y. B., Zhao, N., Oliva, A., and Bylinskii, Z. (2020b). How much time do you have? modeling multi-duration saliency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [135] Foulsham, T. and Kingstone, A. (2010). Asymmetries in the direction of saccades during perception of scenes and fractals: Effects of image type and image features. *Vision Research*.
- [136] Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., and Holz, T. (2020). Leveraging frequency analysis for deep fake image recognition. In *Proceedings of the International Conference on Machine Learning*.
- [137] Friston, K., Adams, R. A., Perrinet, L., and Breakspear, M. (2012). Perceptions as hypotheses: Saccades as experiments. *Frontiers in psychology*.
- [138] Frye, R. E., Vassall, S., Kaur, G., Lewis, C., Karim, M., and Rossignol, D. (2019). Emerging biomarkers in autism spectrum disorder: a systematic review. *Annals of translational medicine*.
- [139] Fuhl, W. and Kasneci, E. (2021). A multimodal eye movement dataset and a multimodal eye movement segmentation analysis. In *Proceedings of the Symposium on Eye Tracking Research & Applications*.
- [140] Furnari, A. and Farinella, G. M. (2019). What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [141] Gatys, L. A., Kümmerer, M., Wallis, T. S., and Bethge, M. (2017). Guiding human gaze with convolutional neural networks. *arXiv preprint arXiv:1712.06492*.
- [142] Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., and Tenenbaum, J. B. (2017). Eye-tracking causality. *Psychological science*.
- [143] Ghosh, S., Dhall, A., Li, Z., Mittal, A., Kumar, G., Choudhury, D., and Thies, J. (2023). Automatic gaze analysis: A survey of deep learning based approaches. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

- [144] Goferman, S., Zelnik-Manor, L., and Tal, A. (2011). Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [145] Gong, S., Li, M., Feng, J., Wu, Z., and Kong, L. (2022). Diffuseq: Sequence to sequence text generation with diffusion models. In *Proceedings of the International Conference on Learning Representations*.
- [146] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*.
- [147] Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. (2017). Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [148] Gragnaniello, D., Cozzolino, D., Marra, F., Poggi, G., and Verdoliva, L. (2021). Are gan generated images easy to detect? a critical analysis of the state-of-the-art. In *ICME*.
- [149] Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al. (2022). Ego4d: Around the world in 3000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [150] Gressel, C., Overdorf, R., Hagenstedt, I., Karaboga, M., Lurtz, H., and Sun, J. (2023). Privacy-aware eye tracking: Challenges and future directions. *IEEE Pervasive Computing*.
- [151] Guo, L., Wang, C., Yang, W., Huang, S., Wang, Y., Pfister, H., and Wen, B. (2023). Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [152] Gupta, P., Chugh, K., Dhall, A., and Subramanian, R. (2020). The eyes know it: Fakeet - an eye-tracking database to understand deepfake perception. In *Proceedings of the International Conference on Multimodal Interaction*.
- [153] Gutiérrez, J., Che, Z., Zhai, G., and Le Callet, P. (2021). Saliency4asd: Challenge, dataset and tools for visual attention modeling for autism spectrum disorder. *Signal Processing: Image Communication*.

- [154] Harel, J., Koch, C., and Perona, P. (2006). Graph-based visual saliency. *Advances in Neural Information Processing Systems*.
- [155] Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*.
- [156] He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [157] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [158] He, S., Tavakoli, H. R., Borji, A., and Pugeault, N. (2019). Human attention in image captioning: Dataset and analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [159] Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in cognitive sciences*.
- [160] Henderson, J. M., Brockmole, J. R., Castelano, M. S., and Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. In *Eye Movements*.
- [161] Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*.
- [162] Hosseini, A., Kazerooni, A., Akhavan, S., Brudno, M., and Taati, B. (2025). Sum: Saliency unification through mamba for visual attention modeling. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.
- [163] Hou, X. and Zhang, L. (2008). Dynamic visual attention: Searching for coding length increments. In *Advances in Neural Information Processing Systems*.
- [164] Hou, Y., Zhang, Z., Horanyi, N., Moon, J., Cheng, Y., and Chang, H. J. (2024). Multi-modal gaze following in conversational scenarios. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.
- [165] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

- [166] Huang, H., Wang, H., and Mak, B. (2019a). Recurrent poisson process unit for speech recognition. In *Proceedings of the Conference on Artificial Intelligence*.
- [167] Huang, P.-J., Lu, C.-A., and Chen, K.-W. (2022). Temporally-aggregating multiple-discontinuous-image saliency prediction with transformer-based attention. In *Proceedings of the IEEE International Conference on Robotics and Automation*.
- [168] Huang, S. H., Huang, I., Pandya, R., and Dragan, A. D. (2019b). Non-verbal robot feedback for human teachers. In *Proceedings of the Conference on Robot Learning*.
- [169] Huang, Y., Cai, M., Li, Z., and Sato, Y. (2018). Predicting gaze in egocentric video by learning task-dependent attention transition. In *Proceedings of the European Conference on Computer Vision*.
- [170] Hudson, D. A. and Manning, C. D. (2019). Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [171] Ilaslan, M., Song, C., Chen, J., Gao, D., Lei, W., et al. (2023). Gazevqa: A video question answering dataset for multiview eye-gaze task-oriented collaborations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- [172] Ioannou, C., Seernani, D., Stefanou, M. E., Riedel, A., Tebartz van Elst, L., Smyrnis, N., Fleischhaker, C., Biscaldi-Schaefer, M., Boccignone, G., and Klein, C. (2020). Comorbidity matters: Social visual attention in a comparative study of autism spectrum disorder, attention-deficit/hyperactivity disorder and their comorbidity. *Frontiers in psychiatry*.
- [173] Isham, V. and Westcott, M. (1979). A self-correcting point process. *Stochastic Processes and Their Applications*.
- [174] Itti, L. and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*.
- [175] Itti, L. and Koch, C. (2001a). Computational modelling of visual attention. *Nature Reviews - Neuroscience*.

- [176] Itti, L. and Koch, C. (2001b). Computational modelling of visual attention. *Nature Reviews - Neuroscience*.
- [177] Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [178] Jacob, R. J. (1991). The use of eye movements in human-computer interaction techniques: what you look at is what you get. *ACM Transactions on Information Systems (TOIS)*.
- [179] Jacob, R. J. and Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In *The mind's eye*.
- [180] Jarodzka, H., Holmqvist, K., and Nyström, M. (2010). A vector-based, multidimensional scanpath similarity measure. In *Proceedings of the Symposium on Eye Tracking Research & Applications*.
- [181] Jarodzka, H., Skuballa, I., and Gruber, H. (2021). Eye-tracking in educational practice: Investigating visual perception underlying teaching and learning in the classroom. *Educational psychology review*.
- [182] Jeong, Y., Kim, D., Ro, Y., and Choi, J. (2022). Frepgan: Robust deep-fake detection using frequency-level perturbations. In *Proceedings of the Conference on Artificial Intelligence*.
- [183] Jia, S. and Zhang, Y. (2018). Saliency-based deep convolutional neural network for no-reference image quality assessment. *Multimedia Tools and Applications*.
- [184] Jiang, L., Li, Y., Li, S., Xu, M., Lei, S., Guo, Y., and Huang, B. (2022). Does text attract attention on e-commerce images: A novel saliency prediction dataset and method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [185] Jiang, L., Xu, M., Wang, X., and Sigal, L. (2021). Saliency-guided image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [186] Jiang, M., Boix, X., Roig, G., Xu, J., Van Gool, L., and Zhao, Q. (2016). Learning to predict sequences of human visual fixations. *IEEE Transactions on Neural Networks and Learning Systems*.

- [187] Jiang, M., Huang, S., Duan, J., and Zhao, Q. (2015). Salicon: Saliency in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [188] Jiang, M. and Zhao, Q. (2017). Learning visual attention to identify people with autism spectrum disorder. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [189] Jiang, Y., Guo, Z., Rezazadegan Tavakoli, H., Leiva, L. A., and Oulasvirta, A. (2024). Eyeformer: Predicting personalized scanpaths with transformer-guided reinforcement learning. In *Proceedings of the ACM Symposium on User Interface Software and Technology*.
- [190] Jiang, Y., Leiva, L. A., Rezazadegan Tavakoli, H., RB Houssel, P., Kylmä, J., and Oulasvirta, A. (2023). Ueyes: Understanding visual saliency across user interface types. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.
- [191] Jiao, C., Wang, Y., Zhang, G., Bâce, M., Hu, Z., and Bulling, A. (2025). Diffgaze: A diffusion model for modelling fine-grained human gaze behaviour on 360° images. *ACM Transactions on Interactive Intelligent Systems*.
- [192] Jonides, J., Irwin, D. E., and Yantis, S. (1982). Integrating visual information from successive fixations. *Science*.
- [193] Judd, T., Durand, F., and Torralba, A. (2012). A benchmark of computational models of saliency to predict human fixations. *MIT Technical Report*.
- [194] Judd, T., Ehinger, K., Durand, F., and Torralba, A. (2009). Learning to predict where humans look. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [195] Kamide, Y., Altmann, G. T. M., and Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye-movements. *Journal of Memory and Language*.
- [196] Kara, O., Nisar, H., and Rehg, J. M. (2025). Diffeye: Diffusion-based continuous eye-tracking data generation conditioned on natural images. In *Advances in Neural Information Processing Systems*.
- [197] Karakas, T. and Yildiz, D. (2020). Exploring the influence of the built environment on human experience through a neuroscience approach: A systematic review. *Frontiers of Architectural Research*.

- [198] Karargyris, A., Kashyap, S., Lourentzou, I., Wu, J., Sharma, A., Tong, M., Abedin, S., Beymer, D., Mukherjee, V., Krupinski, E. A., and Moradi, M. (2021). Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for ai development. *Scientific Data*.
- [199] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [200] Khurana, V., Singla, Y. K., et al. (2023). Synthesizing human gaze feedback for improved nlp performance. In *Proceedings of the European Chapter of the Association for Computational Linguistics*.
- [201] Kilde, J. H. (2014). Approaching religious space: An overview of theories, methods, and challenges in religious studies. *Religion and Theology*.
- [202] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- [203] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [204] Koch, C. and Ullman, S. (1987). Shifts in selective visual attention: Towards the underlying neural circuitry. *Matters of Intelligence: Conceptual Structures in Cognitive Neuroscience*.
- [205] Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. (2021a). Dif-fwave: A versatile diffusion model for audio synthesis. In *Proceedings of the International Conference on Learning Representations*.
- [206] Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. (2021b). Dif-fwave: A versatile diffusion model for audio synthesis. In *Proceedings of the International Conference on Learning Representations*.
- [207] Korshunov, P. and Marcel, S. (2021). Subjective and objective evaluation of deepfake videos. In *International Conference on Acoustics, Speech, and Signal Processing*.
- [208] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*.

- [209] Kroner, A., Senden, M., Driessens, K., and Goebel, R. (2020). Contextual encoder-decoder network for visual saliency prediction. *Neural Networks*.
- [210] Kruthiventi, S. S., Ayush, K., and Babu, R. V. (2017). Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*.
- [211] Kumar, S., Zhang, X., and Leskovec, J. (2019). Predicting dynamic embedding trajectory in temporal interaction networks. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.
- [212] Kümmerer, M. and Bethge, M. (2021). State-of-the-art in human scan-path prediction. *arXiv preprint arXiv:2102.12239*.
- [213] Kümmerer, M. and Bethge, M. (2023). Predicting visual fixations. *Annual Review of Vision Science*.
- [214] Kümmerer, M., Bethge, M., and Wallis, T. S. A. (2022). Deepgaze iii: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision*.
- [215] Kümmerer, M., Theis, L., and Bethge, M. (2014). Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *Proceedings of the International Conference on Learning Representations Workshops*.
- [216] Kümmerer, M., Wallis, T. S., and Bethge, M. (2015). Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences*.
- [217] Lai, B., Liu, M., Ryan, F., and Rehg, J. M. (2024). In the eye of transformer: Global–local correlation for egocentric gaze estimation and beyond. *International Journal of Computer Vision*.
- [218] Lan, G., Heit, B., Scargill, T., and Gorlatova, M. (2020). Gazegraph: Graph-based few-shot cognitive context sensing from human visual behavior. In *Proceedings of the Conference on Embedded Networked Sensor Systems*.
- [219] Lang, P. J., Bradley, M. M., Cuthbert, B. N., et al. (2005). *International affective picture system (IAPS): Affective ratings of pictures and instruction manual*. NIMH, Center for the Study of Emotion & Attention Gainesville, FL.

- [220] Le Bras, T., Allibe, B., and Doré-Mazars, K. (2024). The way we look at an image or a webpage can reveal personality traits. *Scientific Reports*.
- [221] Le Meur, O. and Liu, Z. (2015). Saccadic model of eye movements for free-viewing condition. *Vision Research*.
- [222] Leiva, L. A., Xue, Y., Bansal, A., Tavakoli, H. R., et al. (2020). Understanding visual saliency in mobile user interfaces. In *ACM MobileHCI*.
- [223] Li, P., He, J., Li, G., Bhargava, R., Shen, S., Valliappan, N., Liang, Y., Gu, H., Ramachandran, V., Li, Y., et al. (2024). Uniar: A unified model for predicting human attention and responses on visual content. In *Advances in Neural Information Processing Systems*.
- [224] Li, S., Xiao, S., Zhu, S., Du, N., Xie, Y., and Song, L. (2018a). Learning temporal point processes via reinforcement learning. *Advances in Neural Information Processing Systems*.
- [225] Li, X., Thickstun, J., Gulrajani, I., Liang, P. S., and Hashimoto, T. B. (2022). Diffusion-lm improves controllable text generation. In *Advances in Neural Information Processing Systems*.
- [226] Li, Y., Fathi, A., and Rehg, J. M. (2013). Learning to predict gaze in ego-centric video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [227] Li, Y., Liu, M., and Rehg, J. M. (2018b). In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European Conference on Computer Vision*.
- [228] Li, Y., Liu, M., and Rehg, J. M. (2021). In the eye of the beholder: Gaze and actions in first person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [229] Li, Y. and Lyu, S. (2018). Exposing deepfake videos by detecting face warping artifacts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.
- [230] Liang, S., Liu, R., and Qian, J. (2021). Fixation prediction for advertising images: Dataset and benchmark. *Journal of Visual Communication and Image Representation*.

- [231] Lim, H., Kim, M., Park, S., and Park, N. (2023). Regular time-series generation using sgm. *arXiv preprint arXiv:2301.08518*.
- [232] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*.
- [233] Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. (2022). Flow matching for generative modeling. In *Proceedings of the International Conference on Learning Representations*.
- [234] Liu, H., Xu, D., Huang, Q., Li, W., Xu, M., and Lin, S. (2013). Semantically-based human scanpath estimation with hmms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [235] Liu, R., Lehman, J., Molino, P., Petroski Such, F., Frank, E., Sergeev, A., and Yosinski, J. (2018). An intriguing failing of convolutional neural networks and the coordconv solution. In *Advances in Neural Information Processing Systems*.
- [236] Liu, W., Li, M., and Yi, L. (2016). Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework. *Autism Research*.
- [237] Liu, X., Cheng, X., Sun, Y., Wu, X., Song, R., Sun, H., and Zhang, D. (2025). Eyear: Learning audio synchronized human gaze trajectory based on physics-informed dynamics. In *Proceedings of the Conference on Artificial Intelligence*.
- [238] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- [239] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [240] Long, C., Yang, X., and Xu, C. (2020). Cross-domain personalized image captioning. *Multimedia Tools and Applications*.

- [241] Lorenz, P., Durall, R. L., and Keuper, J. (2023). Detecting images generated by deep diffusion models using their local intrinsic dimensionality. In *Proceedings of the International Conference on Computer Vision Workshops*.
- [242] Lou, J., Lin, H., Marshall, D., et al. (2022). Transalnet: Towards perceptually relevant visual saliency prediction. *Neurocomputing*.
- [243] Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. (2022). Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems*.
- [244] Lumetti, L., Pipoli, V., Bolelli, F., Ficarra, E., and Grana, C. (2024). Enhancing patch-based learning for the segmentation of the mandibular canal. *IEEE Access*.
- [245] Ma, C., Zhao, L., Chen, Y., Wang, S., Guo, L., Zhang, T., Shen, D., Jiang, X., and Liu, T. (2023). Eye-gaze-guided vision transformer for rectifying shortcut learning. *IEEE Transactions on Medical Imaging*.
- [246] Maiano, L., Benova, A., Papa, L., Stockner, M., Marchetti, M., Convertino, G., et al. (2024). Human versus machine: A comparative analysis in detecting artificial intelligence-generated images. *IEEE Security & Privacy*.
- [247] Marr, D. (1976). Analyzing natural images: A computational theory of texture vision. In *Cold Spring Harbor Symposia on Quantitative Biology*.
- [248] Martin, D., Serrano, A., Bergman, A. W., Wetzstein, G., and Masia, B. (2022). Scangan360: A generative model of realistic scanpaths for 360° images. *IEEE Transactions on Visualization and Computer Graphics*.
- [249] Mazzamuto, M., Furnari, A., and Farinella, G. M. (2025). Gazing into missteps: Leveraging eye-gaze for unsupervised mistake detection in egocentric videos of skilled human activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [250] Mazzamuto, M., Ragusa, F., Furnari, A., and Farinella, G. M. (2024). Learning to detect attended objects in cultural sites with gaze signals and weak object supervision. *Journal on Computing and Cultural Heritage*.
- [251] Mei, H. and Eisner, J. M. (2017). The neural hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems*.

- [252] Miangleh, S. M. H., Bylinskii, Z., Kee, E., Shechtman, E., and Aksoy, Y. (2023). Realistic saliency guided image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [253] Min, K. and Corso, J. J. (2021). Integrating human gaze into attention for egocentric activity recognition. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.
- [254] Mokady, R., Hertz, A., Aberman, K., Pritch, Y., and Cohen-Or, D. (2023). Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [255] Mondal, S., Ahn, S., Yang, Z., Balasubramanian, N., Samaras, D., Zelinsky, G., and Hoai, M. (2024). Look hear: Gaze prediction for speech-directed human attention. In *Proceedings of the European Conference on Computer Vision*.
- [256] Mondal, S., Yang, Z., Ahn, S., Samaras, D., Zelinsky, G., and Hoai, M. (2023). Gazeformer: Scalable, effective and fast prediction of goal-directed human attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [257] Murray, N., Vanrell, M., Otazu, X., and Parraga, C. A. (2011). Saliency estimation using a non-parametric low-level vision model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [258] Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*.
- [259] Neumann, D., Spezio, M. L., Piven, J., and Adolphs, R. (2006). Looking you in the mouth: abnormal gaze in autism resulting from impaired top-down modulation of visual attention. *Social cognitive and affective neuroscience*.
- [260] Ngo, T. and Manjunath, B. (2017). Saccade gaze prediction using a recurrent neural network. In *Proceedings of the IEEE International Conference on Image Processing*.
- [261] Nichol, A. Q. and Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In *Proceedings of the International Conference on Machine Learning*.

- [262] Nichol, A. Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. (2022). Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the International Conference on Machine Learning*.
- [263] Ojha, U., Li, Y., and Lee, Y. J. (2023). Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [264] Oliva, A. and Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*.
- [265] Omi, T., Aihara, K., et al. (2019). Fully neural network based model for general temporal point processes. In *nips*.
- [266] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. (2024). Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*.
- [267] Pal, A., Mondal, S., and Christensen, H. I. (2020). Looking at the right stuff: Guided semantic-gaze for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [268] Palazzi, A., Abati, D., Solera, F., Calderara, S., and Cucchiara, R. (2018). Predicting the driver’s focus of attention: The dr(eye)ve project. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [269] Pan, J., Ferrer, C. C., McGuinness, K., O’Connor, N. E., Torres, J., Sayrol, E., and Giro-i Nieto, X. (2017). Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*.
- [270] Pan, J., Sayrol, E., Giro-i Nieto, X., McGuinness, K., and O’Connor, N. E. (2016). Shallow and deep convolutional networks for saliency prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [271] Parkhurst, D., Law, K., and Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*.
- [272] Pearl, J. (2009). *Causality*. Cambridge university press.

- [273] Peebles, W. and Xie, S. (2023). Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [274] Pelphrey, K. A., Sasson, N. J., Reznick, J. S., Paul, G., Goldman, B. D., and Piven, J. (2002). Visual scanning of faces in autism. *Journal of autism and developmental disorders*.
- [275] Peters, R. J., Iyer, A., Itti, L., and Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research*.
- [276] Pirsiavash, H. and Ramanan, D. (2012). Detecting activities of daily living in first-person camera views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [277] Plizzari, C., Goletto, G., Furnari, A., Bansal, S., Ragusa, F., Farinella, G. M., Damen, D., and Tommasi, T. (2024). An outlook into the future of egocentric vision. *International Journal of Computer Vision*.
- [278] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. (2023). Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.
- [279] Privitera, C. M. and Stark, L. W. (2002). Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [280] Qian, K., Zhang, Z., Song, W., and Liao, J. (2023). Gvgnet: Gaze-directed visual grounding for learning under-specified object referring intention. *IEEE Robotics and Automation Letters*.
- [281] Qiao, M., Liu, Y., Xu, M., Deng, X., Li, B., Hu, W., and Borji, A. (2024). Joint learning of audio–visual saliency prediction and sound source localization on multi-face videos. *International Journal of Computer Vision*.
- [282] Qiao, T., Dong, J., and Xu, D. (2018). Exploring human-like attention supervision in visual question answering. In *Proceedings of the Conference on Artificial Intelligence*.
- [283] Qiu, M., Guo, Y., Zhang, M., Zhang, J., Lan, T., and Liu, Z. (2023). Simulating human visual system based on vision transformer. In *Proceedings of the ACM Symposium on Spatial User Interaction*.

- [284] Quan, R., Lai, Y., Qiu, M., and Liang, D. (2024). Pathformer3d: A 3d scanpath transformer for 360° images. In *Proceedings of the European Conference on Computer Vision*.
- [285] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*.
- [286] Ragusa, F., Furnari, A., Battiato, S., Signorello, G., and Farinella, G. M. (2020). Ego-ch: Dataset and fundamental tasks for visitors behavioral understanding using egocentric vision. *Pattern Recognition Letters*.
- [287] Ragusa, F., Furnari, A., and Farinella, G. M. (2023). Meccano: A multimodal egocentric dataset for humans behavior understanding in the industrial-like domain. *Computer Vision and Image Understanding*.
- [288] Ragusa, F., Furnari, A., Livatino, S., and Farinella, G. M. (2021). The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.
- [289] Ramanathan, S., Katti, H., Sebe, N., Kankanhalli, M., and Chua, T.-S. (2010). An eye fixation database for saliency detection in images. In *Proceedings of the European Conference on Computer Vision*.
- [290] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. (2021). Zero-shot text-to-image generation. In *Proceedings of the International Conference on Machine Learning*.
- [291] Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al. (2025). Sam 2: Segment anything in images and videos. In *Proceedings of the International Conference on Learning Representations*.
- [292] Reich, D. R., Prasse, P., Tschirner, C., Haller, P., Goldhammer, F., and Jäger, L. A. (2022). Inferring native and non-native human reading comprehension and subjective text difficulty from scanpaths in reading. In *Proceedings of the Symposium on Eye Tracking Research & Applications*.
- [293] Ren, Y. and Xiong, D. (2023). Huaslim: Human attention motivated shortcut learning identification and mitigation for large language models. In

- Proceedings of the Annual Meeting of the Association for Computational Linguistics, Findings.*
- [294] Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., and Goel, V. (2017). Self-critical sequence training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
 - [295] Reynolds, J. H. and Heeger, D. J. (2009). The normalization model of attention. *Neuron*.
 - [296] Rezende, D. and Mohamed, S. (2015). Variational inference with normalizing flows. In *Proceedings of the International Conference on Machine Learning*.
 - [297] Risko, E. F., Anderson, N. C., Lanthier, S., and Kingstone, A. (2012). Curious eyes: Individual differences in personality predict eye movement behavior in scene-viewing. *Cognition*.
 - [298] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
 - [299] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
 - [300] Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. (2023). Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
 - [301] Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*.
 - [302] Rusnak, M., Szmigiel, M., Geniusz, M., Koszewicz, Z., and Magdziak-Tokłowicz, M. (2024). Exploring the impact of cultural context on eye-tracking studies of architectural monuments in selected european cities: Sustainable heritage management. *Journal of Cultural Heritage*.

- [303] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*.
- [304] Salimans, T. and Ho, J. (2022). Progressive distillation for fast sampling of diffusion models. In *Proceedings of the International Conference on Learning Representations*.
- [305] Salvucci, D. D. and Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the Symposium on Eye Tracking Research & Applications*.
- [306] Saran, A., Short, E. S., Thomaz, A., and Niekum, S. (2019). Understanding teacher gaze patterns for robot learning. In *Proceedings of the Conference on Robot Learning*.
- [307] Sasson, N. J., Turner-Brown, L. M., Holtzclaw, T. N., Lam, K. S., and Bodfish, J. W. (2008). Children with autism demonstrate circumscribed attention during passive viewing of complex social and nonsocial picture arrays. *Autism Research*.
- [308] Sauer, T., Yorke, J. A., and Casdagli, M. (1991). Embedology. *Journal of Statistical Physics*.
- [309] Schiatti, L., Gori, M., Schrimpf, M., Cappagli, G., Morelli, F., Signorini, S., Katz, B., and Barbu, A. (2023). Modeling visual impairments with artificial neural networks: A review. In *Proceedings of the International Conference on Computer Vision Workshops*.
- [310] Schoop, E., Zhou, X., Li, G., Chen, Z., Hartmann, B., and Li, Y. (2022). Predicting and explaining mobile ui tappability with vision modeling and saliency analysis. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.
- [311] Schütt, H. H., Rothkegel, L. O., Trukenbrod, H. A., Reich, S., Wichmann, F. A., and Engbert, R. (2017). Likelihood-based parameter estimation and comparison of dynamical cognitive models. *Psychological review*.
- [312] Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhanian, D., Wang, R., and Yao, A. (2022). Assembly101: A large-scale multi-view video dataset for

- understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [313] Sha, Z., Li, Z., Yu, N., and Zhang, Y. (2023). De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *ACM CCS*.
 - [314] Shchur, O., Biloš, M., and Günnemann, S. (2020). Intensity-free learning of temporal point processes. In *Proceedings of the International Conference on Learning Representations*.
 - [315] Shchur, O., Türkmen, A. C., Januschowski, T., and Günnemann, S. (2021). Neural temporal point processes: A review. In *Proceedings of the International Joint Conferences on Artificial Intelligence*.
 - [316] Shen, C. and Zhao, Q. (2014). Webpage saliency. In *Proceedings of the European Conference on Computer Vision*.
 - [317] Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*.
 - [318] Skerath, L., Toborek, P., Zielińska, A., et al. (2023). Native language prediction from gaze: a reproducibility study. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics Workshops*.
 - [319] Skorokhodov, I., Sotnikov, G., and Elhoseiny, M. (2021). Aligning latent and image spaces to connect the unconnectable. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
 - [320] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the International Conference on Machine Learning*.
 - [321] Song, J., Meng, C., and Ermon, S. (2020). Denoising diffusion implicit models. In *Proceedings of the International Conference on Learning Representations*.
 - [322] Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. (2023). Consistency models. In *Proceedings of the International Conference on Machine Learning*.

- [323] Song, Y., Ding, J., Yuan, J., Liao, Q., and Li, Y. (2024). Controllable human trajectory generation using profile-guided latent diffusion. *ACM Transactions on Knowledge Discovery from Data*.
- [324] Song, Y. and Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*.
- [325] Sood, E., Kögel, F., Müller, P., et al. (2023). Multimodal integration of human-like attention in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.
- [326] Sood, E., Kögel, F., Strohm, F., Dhar, P., and Bulling, A. (2021). Vqamhug: A gaze dataset to study multimodal neural attention in visual question answering. In *Proceedings of the Conference on Computational Natural Language Learning*.
- [327] Sood, E., Tannert, S., Müller, P., and Bulling, A. (2020). Improving natural language processing tasks with human gaze-guided neural attention. In *Advances in Neural Information Processing Systems*.
- [328] Spirtes, P., Glymour, C., and Scheines, R. (2001). *Causation, Prediction, and Search*. MIT press.
- [329] Sugano, Y. and Bulling, A. (2016). Seeing with humans: Gaze-assisted neural image captioning. *arXiv preprint arXiv:1608.05203*.
- [330] Sui, X., Fang, Y., Zhu, H., Wang, S., and Wang, Z. (2023). Scandmm: A deep markov model of scanpath prediction for 360° images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [331] Sun, W., Chen, Z., and Wu, F. (2019). Visual scanpath prediction using ior-roi recurrent mixture density network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [332] Sun, X., Yao, H., Ji, R., and Liu, X.-M. (2014). Toward statistical modeling of saccadic eye-movement and visual saliency. *IEEE Transactions on Image Processing*.

- [333] Takmaz, E., Pezzelle, S., Beinborn, L., and Fernández, R. (2020). Generating image descriptions via sequential cross-modal alignment guided by human gaze. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- [334] Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1996). Using eye movements to study spoken language comprehension: Evidence for visually mediated incremental interpretation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.
- [335] Tang, Y., Tian, Y., Lu, J., Feng, J., and Zhou, J. (2017). Action recognition in rgb-d egocentric videos. *Proceedings of the IEEE International Conference on Image Processing*.
- [336] Tang, Y., Zhao, C., Wang, J., Zhang, C., Sun, Q., Zheng, W. X., Du, W., Qian, F., and Kurths, J. (2022). Perception and navigation in autonomous systems in the era of learning: A survey. *IEEE Transactions on Neural Networks And Learning Systems*.
- [337] Tatler, B. W., Brockmole, J. R., and Carpenter, R. H. (2017). Latest: A model of saccadic decisions in space and time. *Psychological Review*.
- [338] Tavakoli, H. R., Borji, A., Laaksonen, J., and Rahtu, E. (2017). Exploiting inter-image similarity and ensemble of extreme learners for fixation prediction using deep features. *Neurocomputing*.
- [339] Tavakoli, H. R., Rahtu, E., and Heikkilä, J. (2013). Stochastic bottom-up fixation prediction and saccade generation. *Image and Vision Computing*.
- [340] Tavakoli, H. R., Rahtu, E., Kannala, J., and Borji, A. (2019). Digging deeper into egocentric gaze prediction. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.
- [341] Tech, G. (2018). Extended gtea gaze+.
- [342] Thorup, E., Nyström, P., Gredebäck, G., Bölte, S., Falck-Ytter, T., and Team, E. (2016). Altered gaze following during live interaction in infants at risk for autism: an eye tracking study. *Molecular autism*.
- [343] Tiezzi, M., Marullo, S., Betti, A., Meloni, E., Faggi, L., Gori, M., and Melacci, S. (2022). Foveated neural computation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*.

- [344] Torralba, A., Oliva, A., Castelano, M. S., and Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*.
- [345] Trajkovska, K., Kljun, M., and Pucihar, K. Č. (2024). Gaze2aoi: Open source deep-learning based system for automatic area of interest annotation with eye tracking data. *arXiv preprint arXiv:2411.13346*.
- [346] Treisman, A. M. and Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*.
- [347] Trivedi, R., Farajtabar, M., Biswal, P., and Zha, H. (2019). Dyrep: Learning representations over dynamic graphs. In *Proceedings of the International Conference on Learning Representations*.
- [348] van Beers, R. (2007). The sources of variability in saccadic eye movements. *The Journal of Neuroscience*.
- [349] Van Rijn, S., Urbanus, E., and Swaab, H. (2019). Eyetracking measures of social attention in young children: How gaze patterns translate to real-life social behaviors. *Social Development*.
- [350] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*.
- [351] Vig, E., Dorr, M., and Cox, D. (2014). Large-scale optimization of hierarchical features for saliency prediction in natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [352] Vinnikov, M., Motahari, K., Hamilton, L. I., and Altin, B. O. (2021). Understanding urban devotion through the eyes of an observer. In *Proceedings of the Symposium on Eye Tracking Research & Applications*.
- [353] Walther, D. and Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks*.
- [354] Wang, B., Pan, H., Aboah, A., Zhang, Z., Keles, E., Torigian, D., Turkbey, B., Krupinski, E., Udupa, J., and Bagci, U. (2024a). Gazegnn: A gaze-guided graph neural network for chest x-ray classification. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.

- [355] Wang, L., Hajiesmaili, M., and Sitaraman, R. K. (2021). Focas: Practical video super resolution using foveated rendering. In *Proceedings of the ACM International Conference on Multimedia*.
- [356] Wang, R., Juefei-Xu, F., Ma, L., Xie, X., Huang, Y., Wang, J., and Liu, Y. (2020a). Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces. In *Proceedings of the International Joint Conferences on Artificial Intelligence*.
- [357] Wang, S., Jiang, M., Duchesne, X. M., Laugeson, E. A., Kennedy, D. P., Adolphs, R., and Zhao, Q. (2015). Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking. *Neuron*.
- [358] Wang, S., Ouyang, X., Liu, T., Wang, Q., and Shen, D. (2022). Follow my eye: Using gaze to supervise computer-aided diagnosis. *IEEE Transactions on Medical Imaging*.
- [359] Wang, S.-Y., Wang, O., Zhang, R., Owens, A., and Efros, A. A. (2020b). Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [360] Wang, T., Xue, Y., Mohamed, Z. A., and Jia, F. (2025). Developmental functional brain network abnormalities in autism spectrum disorder comorbid with attention deficit hyperactivity disorder. *European Journal of Pediatrics*.
- [361] Wang, W., Chen, C., Wang, Y., Jiang, T., Fang, F., and Yao, Y. (2011). Simulating human saccadic scanpaths on natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [362] Wang, W., Song, H., Zhao, S., Shen, J., Zhao, S., Hoi, S. C., and Ling, H. (2019). Learning unsupervised video object segmentation through visual attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [363] Wang, Y., Zhang, F.-L., and Dodgson, N. A. (2024b). Scantd: 360° scan-path prediction based on time-series diffusion. In *Proceedings of the ACM International Conference on Multimedia*.
- [364] Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., and Li, H. (2023). Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

- [365] Wedel, M. and Pieters, R. (2008). Eye tracking for visual marketing. *Foundations and Trends® in Marketing*.
- [366] Wei, T., Lin, Y., Guo, S., Lin, Y., Huang, Y., Xiang, C., Bai, Y., and Wan, H. (2024). Diff-rntraj: A structure-aware diffusion model for road network-constrained trajectory generation. *IEEE Transactions on Knowledge and Data Engineering*.
- [367] Weng, L. (2021). What are diffusion models? <https://lilianweng.github.io/posts/2021-07-11-diffusion-models/>.
- [368] Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*.
- [369] Wolfe, J. M. and Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behaviour*.
- [370] Wu, H., Zhou, J., and Zhang, S. (2023). Generalizable synthetic image detection via language-guided contrastive learning. *arXiv preprint arXiv:2305.13800*.
- [371] Xia, Y., Zhang, D., Kim, J., Nakayama, K., Zipser, K., and Whitney, D. (2018). Predicting driver attention in critical situations. In *Proceedings of the Asian Conference on Computer Vision*.
- [372] Xiao, S., Yan, J., Yang, X., Zha, H., and Chu, S. (2017). Modeling the intensity function of point process via recurrent neural networks. In *Proceedings of the Conference on Artificial Intelligence*.
- [373] Xiong, K., Jiang, L., Dang, X., Wang, G., Ye, W., and Qin, Z. (2020). Towards personalized aesthetic image caption. In *Proceedings of the International Joint Conference on Neural Networks*.
- [374] Xu, J., Jiang, M., Wang, S., Kankanhalli, M. S., and Zhao, Q. (2014). Predicting human gaze beyond pixels. *Journal of Vision*.
- [375] Xu, Y., Gao, S., Wu, J., Li, N., and Yu, J. (2018). Personalized saliency and its prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [376] Xu, Y., Li, N., Wu, J., Yu, J., Gao, S., et al. (2017). Beyond universal saliency: Personalized saliency prediction with multi-task cnn. In *Proceedings of the International Joint Conferences on Artificial Intelligence*.

- [377] Xue, R., Le, H., Xu, J., Mondal, S., Leite, A., Zelinsky, G., Hoai, M., and Samaras, D. (2025a). Personalized image descriptions from attention sequences. *arXiv preprint arXiv:2512.06662*.
- [378] Xue, R., Xu, J., Mondal, S., Le, H., Zelinsky, G., Hoai, M., and Samaras, D. (2025b). Few-shot personalized scanpath prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [379] Yamashita, J., Takimoto, Y., Oishi, H., and Kumada, T. (2024). How do personality traits modulate real-world gaze behavior? generated gaze data shows situation-dependent modulations. *Frontiers In Psychology*.
- [380] Yan, J., Liu, X., Shi, L., Li, C., and Zha, H. (2018). Improving maximum likelihood estimation of temporal point process via discriminative and adversarial learning. In *Proceedings of the International Joint Conferences on Artificial Intelligence*.
- [381] Yan, K., Wang, Z., Ji, L., Wang, Y., Duan, N., and Ma, S. (2024). Voila-a: Aligning vision-language models with user’s gaze attention. In *Advances in Neural Information Processing Systems*.
- [382] Yang, K.-F. and Li, Y.-J. (2025). Visual attention graph. In *Behavior Research Methods*.
- [383] Yang, S., Jiang, Q., Lin, W., and Wang, Y. (2019a). Sgdnnet: An end-to-end saliency-guided deep neural network for no-reference image quality assessment. In *Proceedings of the ACM International Conference on Multimedia*.
- [384] Yang, X., Li, Y., and Lyu, S. (2019b). Exposing deep fakes using inconsistent head poses. In *International Conference on Acoustics, Speech, and Signal Processing*.
- [385] Yang, Y., Jin, M., Wen, H., Zhang, C., Liang, Y., Ma, L., Wang, Y., Liu, C., Yang, B., Xu, Z., et al. (2024a). A survey on diffusion models for time series and spatio-temporal data. *ACM Computing Surveys*.
- [386] Yang, Z., Huang, L., Chen, Y., Wei, Z., Ahn, S., Zelinsky, G., Samaras, D., and Hoai, M. (2020a). Predicting goal-directed human attention using inverse reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

- [387] Yang, Z., Huang, L., Chen, Y., Wei, Z., Ahn, S., Zelinsky, G., Samaras, D., and Hoai, M. (2020b). Predicting goal-directed human attention using inverse reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [388] Yang, Z., Mondal, S., Ahn, S., Xue, R., Zelinsky, G., Hoai, M., and Samaras, D. (2024b). Unifying top-down and bottom-up scanpath prediction using transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [389] Yang, Z., Mondal, S., Ahn, S., Zelinsky, G., Hoai, M., and Samaras, D. (2022). Target-absent human attention. In *Proceedings of the European Conference on Computer Vision*.
- [390] Yi, X., Xu, H., Zhang, H., Tang, L., and Ma, J. (2023). Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [391] Ylitalo, A.-K. (2017). *Statistical inference for eye movement sequences using spatial and spatio-temporal point processes*. PhD thesis, University of Jyväskylä.
- [392] Yu, L., Poirson, P., Yang, S., Berg, A. C., and Berg, T. L. (2016). Modeling context in referring expressions. In *Proceedings of the European Conference on Computer Vision*.
- [393] Yu, Y., Choi, J., Kim, Y., Yoo, K., Lee, S.-H., and Kim, G. (2017). Supervising neural attention models for video captioning by human gaze data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [394] Zanca, D., Melacci, S., and Gori, M. (2019). Gravitational laws of focus of attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [395] Zelinsky, G., Yang, Z., Huang, L., Chen, Y., Ahn, S., Wei, Z., Adeli, H., Samaras, D., and Hoai, M. (2019). Benchmarking gaze prediction for categorical visual search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.
- [396] Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review*.

- [397] Zelinsky, G. J. and Bisley, J. W. (2015). The what, where, and why of priority maps and their interactions with visual working memory. *Annals of the New York Academy of Sciences*.
- [398] Zhang, J. and Sclaroff, S. (2013). Saliency detection: A boolean map approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [399] Zhang, K., Mo, L., Chen, W., Sun, H., and Su, Y. (2023). Magicbrush: A manually annotated dataset for instruction-guided image editing. In *Advances in Neural Information Processing Systems*.
- [400] Zhang, M., Armendariz, M., Xiao, W., Rose, O., Bendtz, K., Livingstone, M., Ponce, C., and Kreiman, G. (2022). Look twice: A generalist computational model predicts return fixations across tasks and species. *PLoS Computational Biology*.
- [401] Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., and Liu, Z. (2024a). Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [402] Zhang, M., Feng, J., Ma, K. T., Lim, J. H., Zhao, Q., and Kreiman, G. (2018). Finding any waldo with zero-shot invariant and efficient visual search. *Nature Communications*.
- [403] Zhang, Q., Lipani, A., Kirnap, O., and Yilmaz, E. (2020a). Self-attentive hawkes process. In *Proceedings of the International Conference on Machine Learning*.
- [404] Zhang, R., Saran, A., Liu, B., Zhu, Y., Guo, S., Niekum, S., Ballard, D., and Hayhoe, M. (2020b). Human gaze assisted artificial intelligence: A review. In *Proceedings of the International Joint Conferences on Artificial Intelligence*.
- [405] Zhang, S., Yang, X., Feng, Y., Qin, C., Chen, C.-C., Yu, N., Chen, Z., Wang, H., Savarese, S., Ermon, S., et al. (2024b). Hive: Harnessing human feedback for instructional visual editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [406] Zhang, Y., Huang, X., Ma, J., Li, Z., Luo, Z., Xie, Y., Qin, Y., Luo, T., Li, Y., Liu, S., et al. (2024c). Recognize anything: A strong image tagging model.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

- [407] Zhang, Y., Jiang, M., and Zhao, Q. (2021). Saliency prediction with external knowledge. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.
- [408] Zhao, B. N., Kara, O., Kim, J., and Rehg, J. M. (2026). St-diffeye: Diffusion-based continuous gaze generation via joint scanpath-trajectory modeling. *arXiv preprint arXiv:2606.15486*.
- [409] Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., and Yu, N. (2021). Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [410] Zhao, M., Xu, D., and Gao, T. (2024). From cognition to computation: A comparative review of human attention and transformer architectures. *arXiv preprint arXiv:2407.01548*.
- [411] Zhao, Q. and Koch, C. (2011). Learning a saliency map using fixated locations in natural scenes. *Journal of Vision*.
- [412] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [413] Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., and Torralba, A. (2017). Scene parsing through ade20k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [414] Zhu, M., Hou, G., Chen, X., Xie, J., Lu, H., and Che, J. (2021). Saliency-guided transformer network combined with local embedding for no-reference image quality assessment. In *Proceedings of the International Conference on Computer Vision Workshops*.
- [415] Zhu, Y., Ye, Y., Zhang, S., Zhao, X., and Yu, J. (2023). Difftraj: Generating gps trajectory with diffusion probabilistic model. In *Advances in Neural Information Processing Systems*.
- [416] Zhu, Y., Yu, J. J., Zhao, X., Liu, Q., Ye, Y., Chen, W., Zhang, Z., Wei, X., and Liang, Y. (2024). Controltraj: Controllable trajectory generation with topology-constrained diffusion model. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

- [417] Zuo, S., Jiang, H., Li, Z., Zhao, T., and Zha, H. (2020). Transformer hawkes process. In *Proceedings of the International Conference on Machine Learning*.

Acknowledgments

More than three years have passed since I joined the AImageLab family, and looking back, I could never have imagined becoming the person I am today. Along the way, I have collected experiences, memories, and lessons that I will cherish for the rest of my life. When I began this journey, I did not truly know what it meant to do research. Today, I am proud to have contributed, in my own small way, to the advancement of knowledge and to the pursuit of scientific progress. As this fundamental chapter of my life comes to an end, I would like to thank everyone who has been part of this incredible journey.

First and foremost, I would like to thank my supervisor, Professor Rita Cucchiara, who gave me the opportunity to enter the world of research and to experience opportunities that profoundly shaped both my personal and professional growth. A special acknowledgment goes to my co-advisors, Vittorio and Marcella. Together, we were the “Gaze Team,” and it is largely thanks to you that I have become the person and researcher I am today. We shared countless memorable moments, and I will always carry your teachings with me. Thank you for supporting me in every possible way, for standing by me during the most difficult times, and for trying to understand me even during my moments of discouragement. Vittorio, I hope you will always remember me as “your first PhD student”.

This experience would not have been the same without my fellow PhD students. Thank you for all the lunches we shared, the coffee breaks, and the endless conversations about our concerns, ambitions, and plans for the future. Each of you taught me something new and showed me a different way of looking at the world.

A very special mention goes to my desk mates, or rather, to the “Kindergarten” of AImageLab. It is true: we were probably the loudest and most chaotic group in the lab. Thank you, Nic, Luca, and Dinuz. Actually, thank you, you clowns. How much I will miss our unique way of showing affection for one another. I still remember the first day we all met in that nearly empty meeting room. Even then, there was already a feeling that something special was beginning. We were complete strangers from different cities, yet somehow fate brought us together. Over the course of three years, we went from being strangers to becoming true friends. Thank you for always being there, for sharing joys, fears, doubts, and hopes. And thank you for all the teasing and jokes. I will always remember you as the best fullback in Villafranca, the Tuscan who eats 5&5, and the complainer whose endless grievances I could only silence with a suppressed pistol.

I would also like to thank all the colleagues and people I met during my internship in Stuttgart, my first experience in Germany. You welcomed me warmly into your laboratory and made me feel like a valued member of the group from the very first day. It is an experience I will always treasure.

Beyond this, I am also deeply grateful to all the researchers I met around the world during summer schools, conferences, and seminars. Whether through a brief conversation, a collaboration, or a simple piece of advice, each of you contributed in some way to the path that brought me here.

While colleagues, professors, and researchers had a direct impact on my doctoral experience, there are also those whose support came quietly from behind the scenes, yet left an even deeper mark on this journey.

My family, to whom I dedicate this thesis and these years of doctoral study. You have always been my first supporters.

To my mother and father, the pillars of my life. Thank you for listening to every complaint, for giving me advice whenever I needed it, and for sharing every decision with me, even those that ultimately turned out to be mistakes. After all, Mom, we both know that I have always done things my own way. Thank you for never judging me for my mistakes and for always helping me turn them into opportunities for growth.

To my brother. Thank you for all the times you showed interest in what I was

working on in the lab and for all the times you asked whether pursuing a PhD was really worth it. As you know, my answer will always be the same: yes. Today, thanks in part to you, I will continue to pursue research.

Finally, I cannot forget my friends. Thank you for all the times you asked me, “So, what exactly do you research?” Now, at last, you can read this thesis and find out in detail. Thank you for all the days we spent together, which helped me navigate moments of anxiety, worry, and uncertainty. Thank you for understanding when I had to say, “Guys, I can’t make it today. I need to finish writing a paper.” Thank you for always supporting my ambition and my desire to aim high. And above all, thank you for listening to my endless concerns about career choices after the PhD. For a while, we thought our paths would diverge and that we would see each other much less often, perhaps for a very long time. Things turned out differently, however, and I could not be happier knowing that you will have to put up with me for quite a while longer.

I do not see this as the closing of a chapter, but rather as a new beginning: a precious collection of knowledge, relationships, and experiences upon which I can build the foundations of my future. I will always be grateful to those who stood by me throughout this journey and to all those whom fate chose to place along my path.

With gratitude,

Giuseppe



Ph.D. Activities

THIS final section presents a list of the main activities carried out by the candidate during the Ph.D. program in Information and Communication Technologies (ICT).

A.1 EXCHANGE PERIODS

14 July - 30 November 2025: Visiting Student Researcher at the University of Stuttgart in the “Collaborative Artificial Intelligence (CAI)” research laboratory, supervised by Professor Andreas Bulling.

A.2 TEACHING ACTIVITIES

2023: Project Tutor for “Computer Vision and Cognitive Systems” course held by Prof. Rita Cucchiara

A.3 HONORS AND AWARDS

2025: Outstanding Reviewer - International Conference on Computer Vision - ICCV 2025.

2024: Outstanding Reviewer - European Conference on Computer Vision - ECCV 2024.

2024: Outstanding Reviewer - ACM International Conference on Multimedia - ACMMM 2024.

A.4 CONFERENCE ATTENDANCES

19th October - 23rd October 2025: International Conference on Computer Vision (ICCV), 2025, *Honolulu, United States of America*.

29th September - 4th October 2024: European Conference on Computer Vision (ECCV), 2024, *Milan, Italy*.

29th October - 3rd November 2023: ACM International Conference on Multimedia (ACMMM), 2023, *Ottawa, Canada*.

A.5 SEMINARS AND WORKSHOPS ATTENDANCES

June 2023: Attendance at “Academic English Workshop II” seminar by Prof. Silvia Cavalieri.

May 2023: Attendance at “Academic English Workshop I” seminar by Prof. Silvia Cavalieri.

December 2022: 3D Computer Vision for Animals, Speaker: Silvia Zuffi.

December 2022: From Handcrafted to End-to-End Learning, and Back: a Journey far Multi-Object Tracking, Speaker: Laura Leal-Taixé.

November 2022: Graph Signal Processing for Machine Learning: Challenges and Use-cases, Speaker: Laura Toni.

A.6 SUMMER SCHOOLS

7th - 13rd July 2024: International Computer Vision Summer School - ICVSS 2024, Scicli, Italy.

18th -22nd September 2023: ELLIS Summer School on Large-Scale AI for Research and Industry - 2023, Modena, Italy.

4th - 8th September 2023: International Summer School on Machine Vision - VISMAL 2023, Padova, Italy.

A.7 PARTICIPATION IN RESEARCH PROJECTS

2023 - 2025: PNRR project Italian Strengthening of Esfri RI Resilience (IT-SERR) funded by the European Union - NextGenerationEU.

A.8 REVIEWING SERVICE

CONFERENCES.

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, 2026.

ACM Symposium on Eye Tracking Research & Applications (ETRA), 2026.

IEEE/CVF International Conference on Computer Vision (ICCV), 2025.

European Conference on Computer Vision (ECCV), 2024.

ACM International Conference on Multimedia (ACMMM), 2023, 2024.

International Conference on Image Analysis and Processing (ICIAP), 2023.

JOURNALS.

Elsevier Pattern Recognition, 2024.

AI Communications, 2023.

B

Supplementary Material for Chapter 3

B.1 MODELLING GAZE DYNAMICS IN SPACE AND TIME WITH NEURAL TEMPORAL POINT PROCESSES

Additional qualitative results are depicted from Fig. B.1 to Fig. B.5 on COCO-FreeView, MIT1003, OSIE, NUSEF, and FiFa datasets, respectively. Each fixation is represented by a circle, with its diameter proportional to the fixation duration. For methods that do not model fixation duration, circles are shown with a uniform size. The first fixation of each scanpath is omitted. The qualitative results support the quantitative results, highlighting the accuracy of TPP-Gaze in predicting human-like scanpaths. Other methods, instead, either overfit on a few salient objects, especially people and faces in the case of Scanpath-VQA, or predict scanpath trajectories containing fixations on unlikely locations (see the bottom sample in Fig. B.14 or the top sample in Fig. B.4).

Considering the saliency prediction task, given a sample image, we construct the aggregated saliency map by convolving a Gaussian kernel over all the locations of predicted fixations [194]. To support the quantitative analysis, we present the saliency prediction of our model against some competitors from Fig. B.6

to Fig. B.10 on COCO-FreeView, MIT1003, OSIE, NUSEF, and FiFa datasets, respectively. Note that we include DeepGazeIII in the comparison for reference, even though its results are not directly comparable. Indeed, DeepGazeIII was specifically trained on a large scale dataset [187] to predict saliency maps along with scanpaths. Nevertheless, TPP-Gaze outperforms DeepGazeIII and the other models in many cases, demonstrating better alignment with humans.

Finally, in Fig. B.11 we show additional qualitative results on sample images from COCO-Search18 for the visual search task. As can be observed, TPP-Gaze can effectively simulate human-like goal-directed visual attention patterns for various target objects. The model demonstrates its ability to adapt, with a simple architectural variation, from a free-viewing setting to a task-specific visual search scenario. The results illustrate how TPP-Gaze generates plausible attention trajectories that focus on regions likely to contain the target object, mimicking the efficient search strategies employed by humans when looking for specific items in complex scenes.

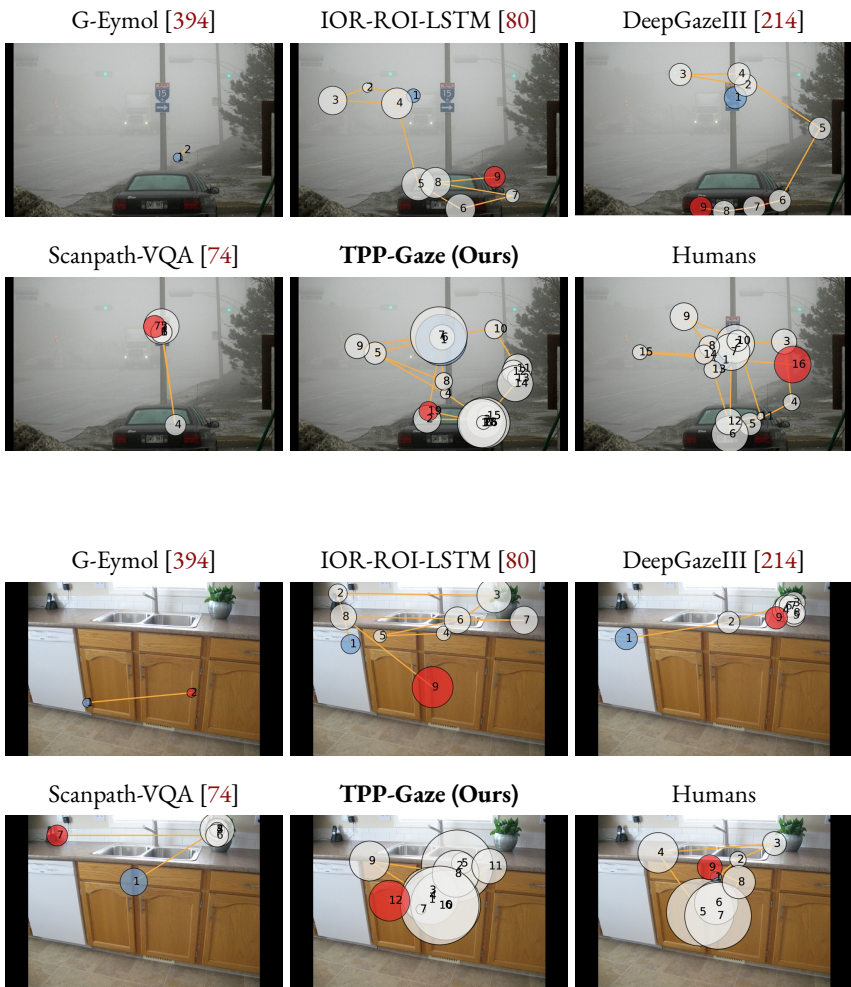


Figure B.1: Qualitative comparison of simulated and human scanpaths on the COCO-FreeView dataset.

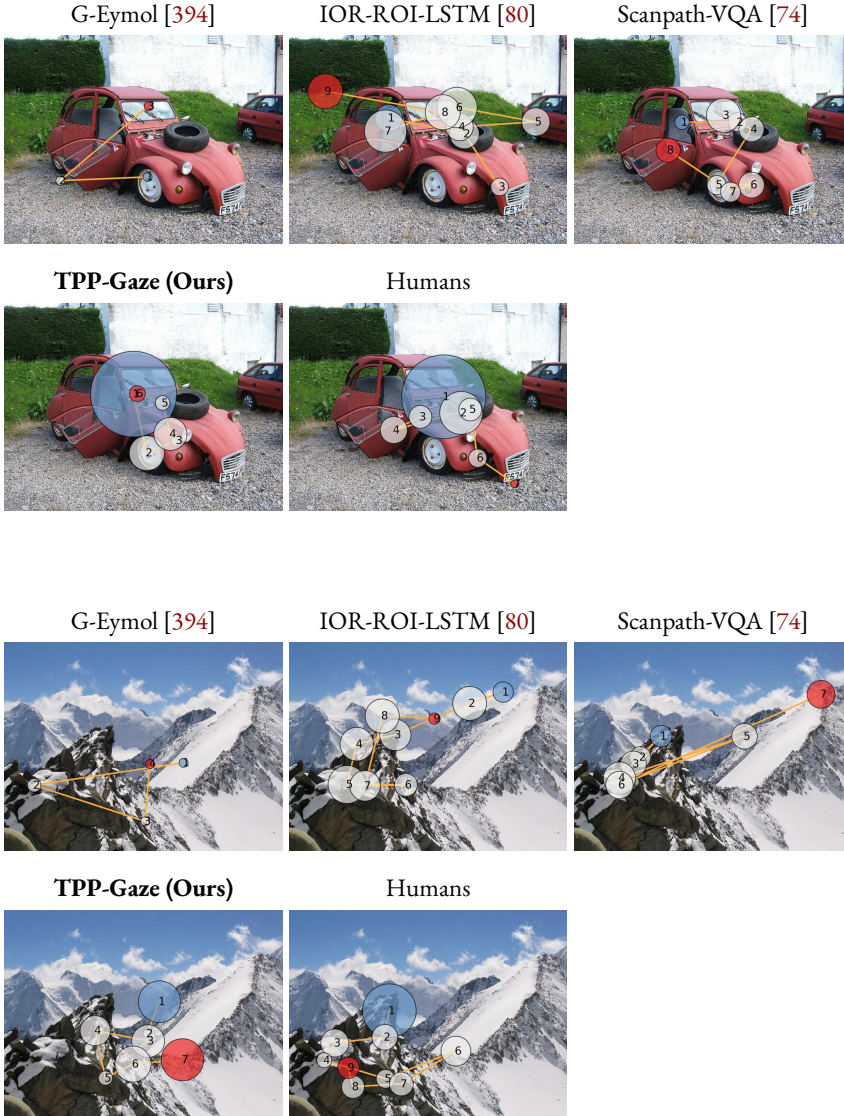


Figure B.2: Qualitative comparison of simulated and human scanpaths on the MIT1003 dataset. We omit DeepGazeIII for consistency with the experimental settings described in the thesis.

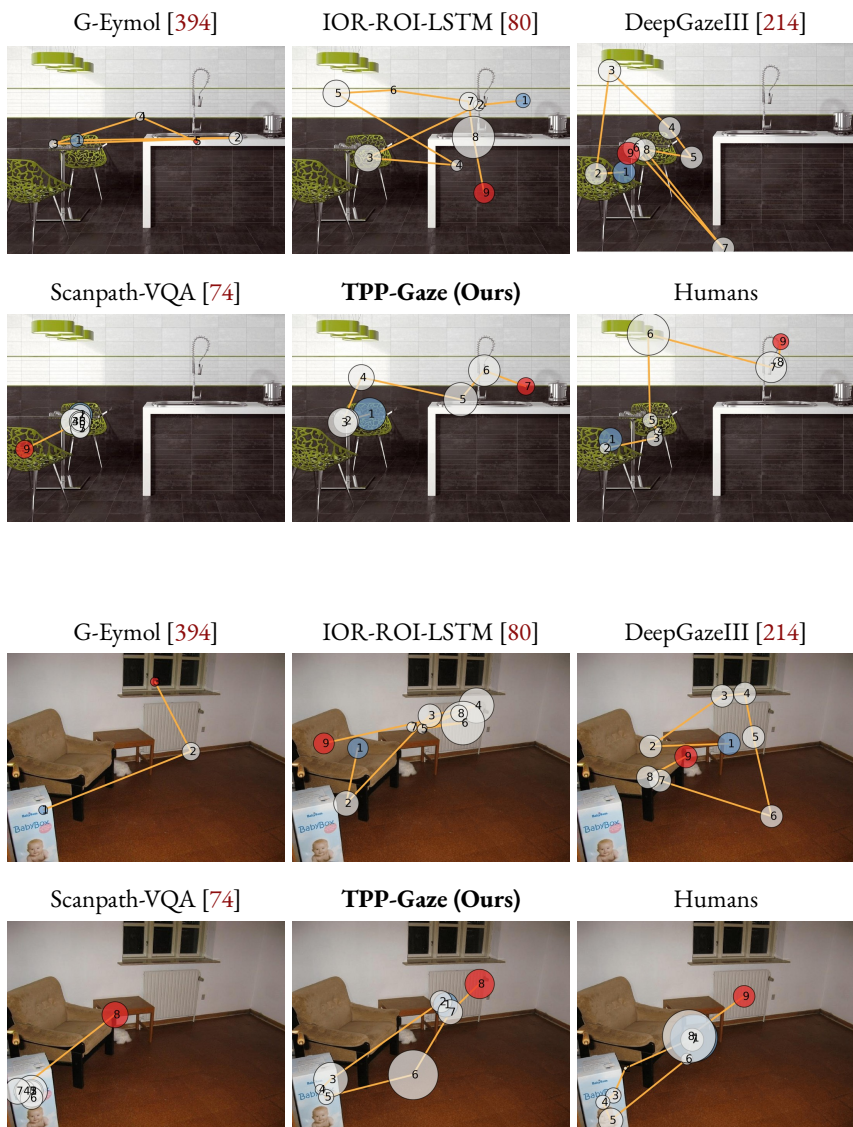


Figure B.3: Qualitative comparison of simulated and human scanpaths on the OSIE dataset.

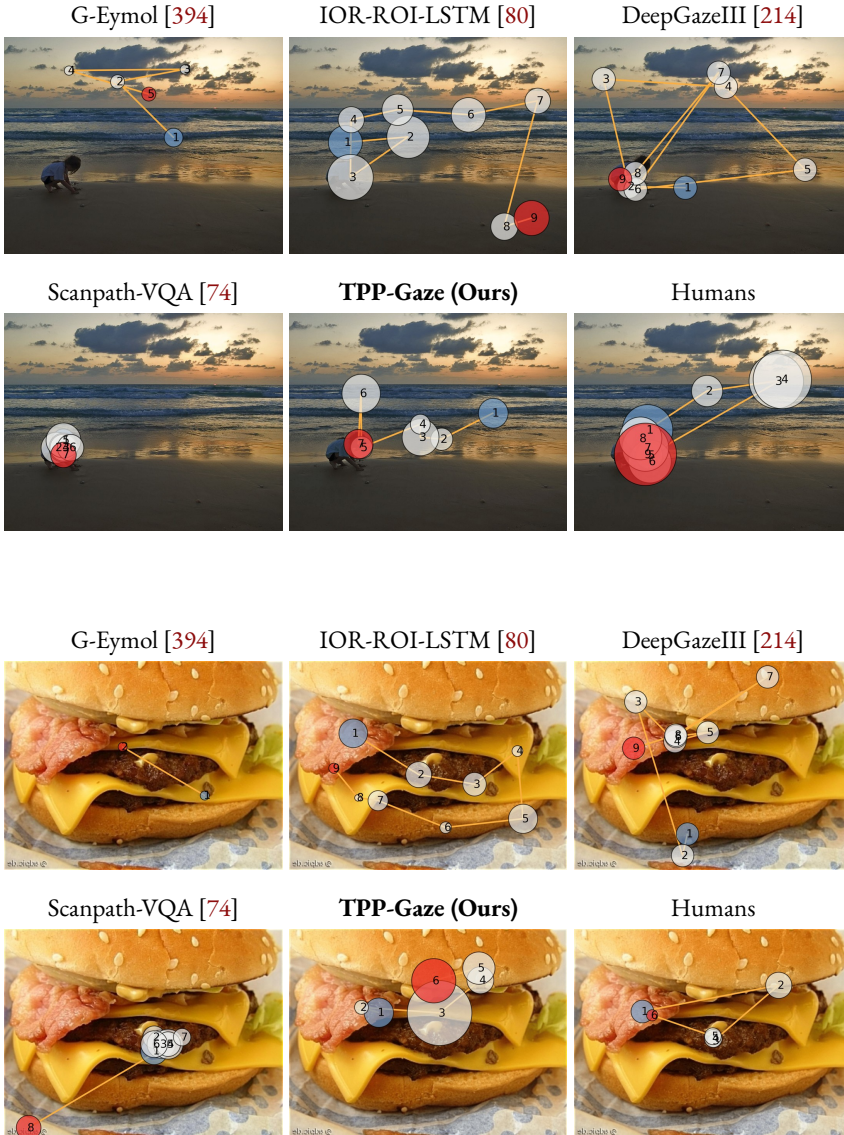


Figure B.4: Qualitative comparison of simulated and human scanpaths on the NUSeF dataset.



Figure B.5: Qualitative comparison of simulated and human scanpaths on the FiFa dataset.



Figure B.6: Saliency maps of sample images from COCO-FreeView dataset computed from the fixations generated by the considered scanpath models. For completeness, we include DeepGazeIII, but note that its training procedure also involves saliency prediction.



Figure B.7: Saliency maps of sample images from MIT1003 dataset computed from the fixations generated by the considered scanpath models. We omit DeepGazeIII for consistency with the experimental settings described in the thesis.

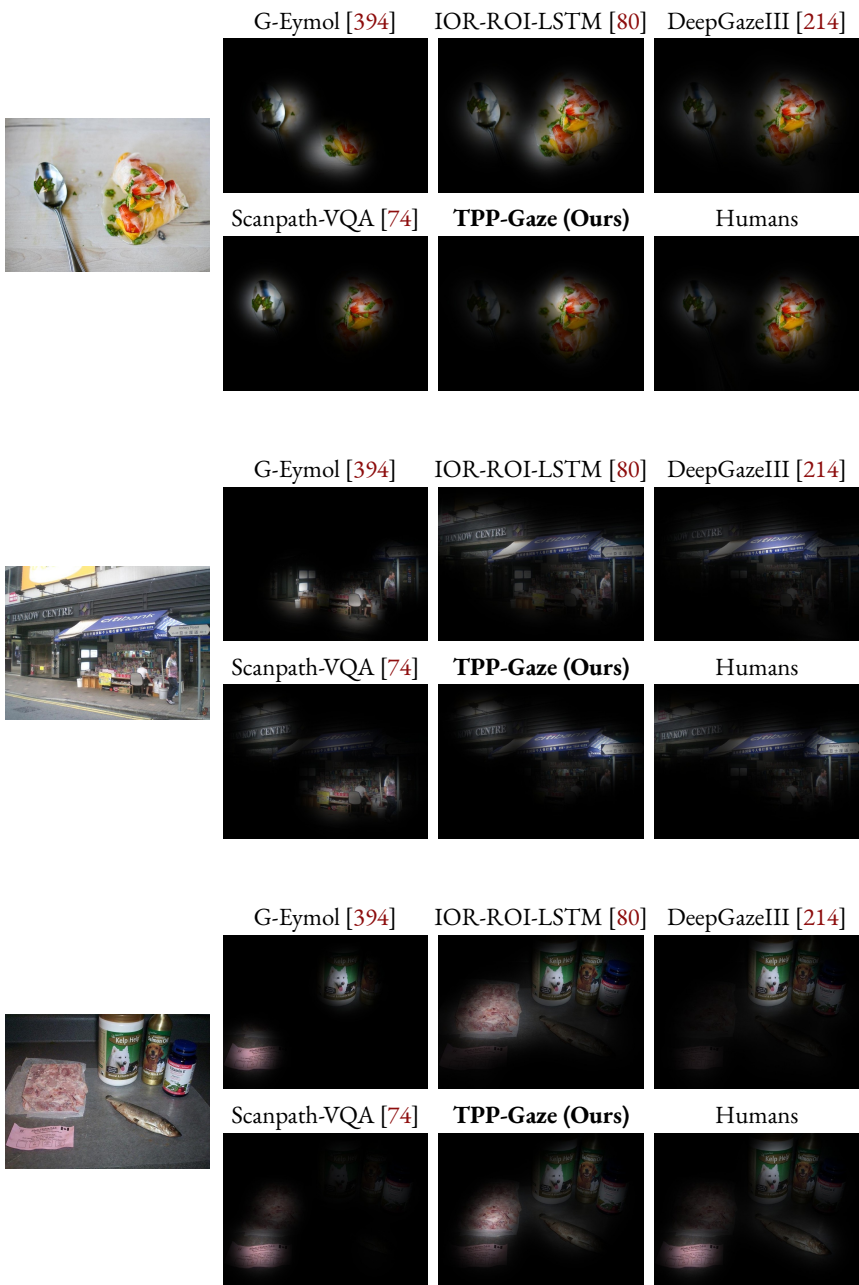


Figure B.8: Saliency maps of sample images from OSIE dataset computed from the fixations generated by the considered scanpath models. For completeness, we include DeepGazeIII, but note that its training procedure also involves saliency prediction.

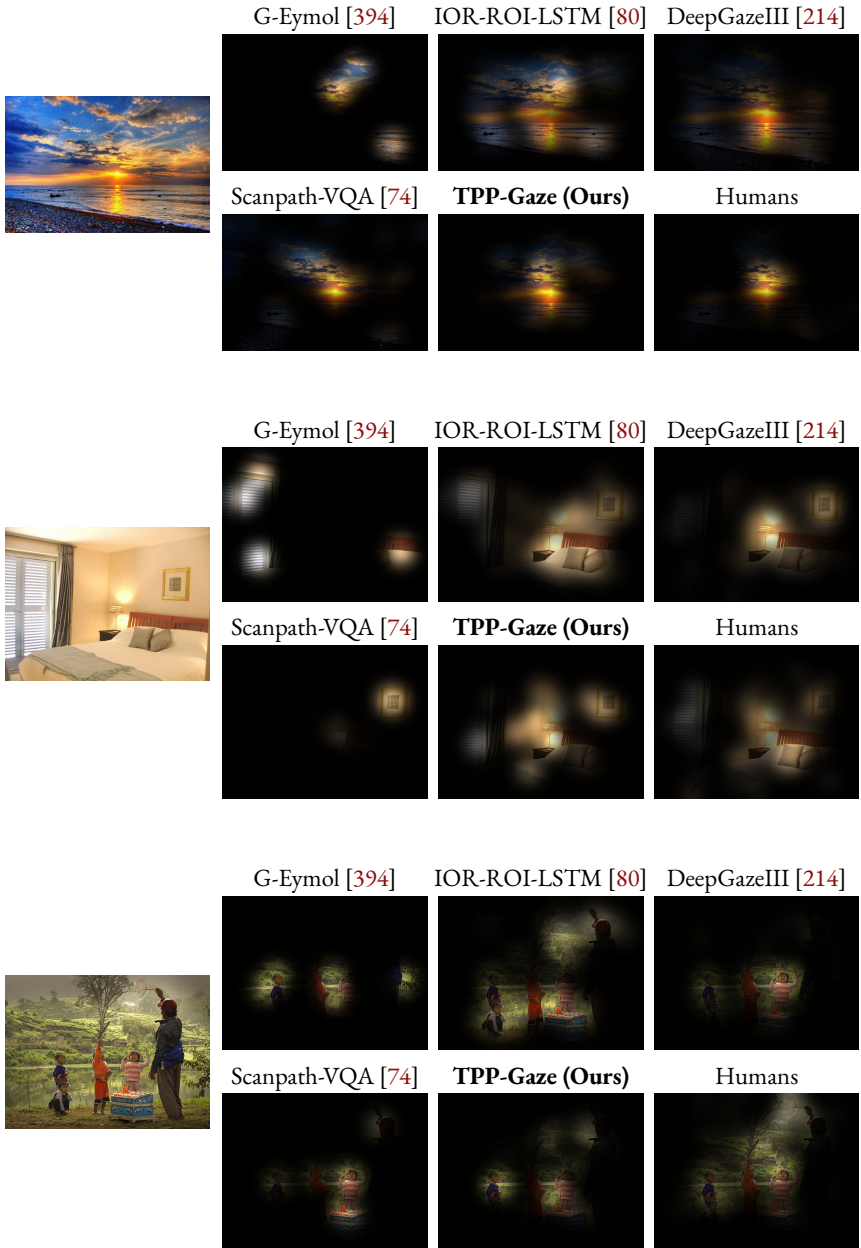


Figure B.9: Saliency maps of sample images from NUSEF dataset computed from the fixations generated by the considered scanpath models. For completeness, we include DeepGazeIII, but note that its training procedure also involves saliency prediction.

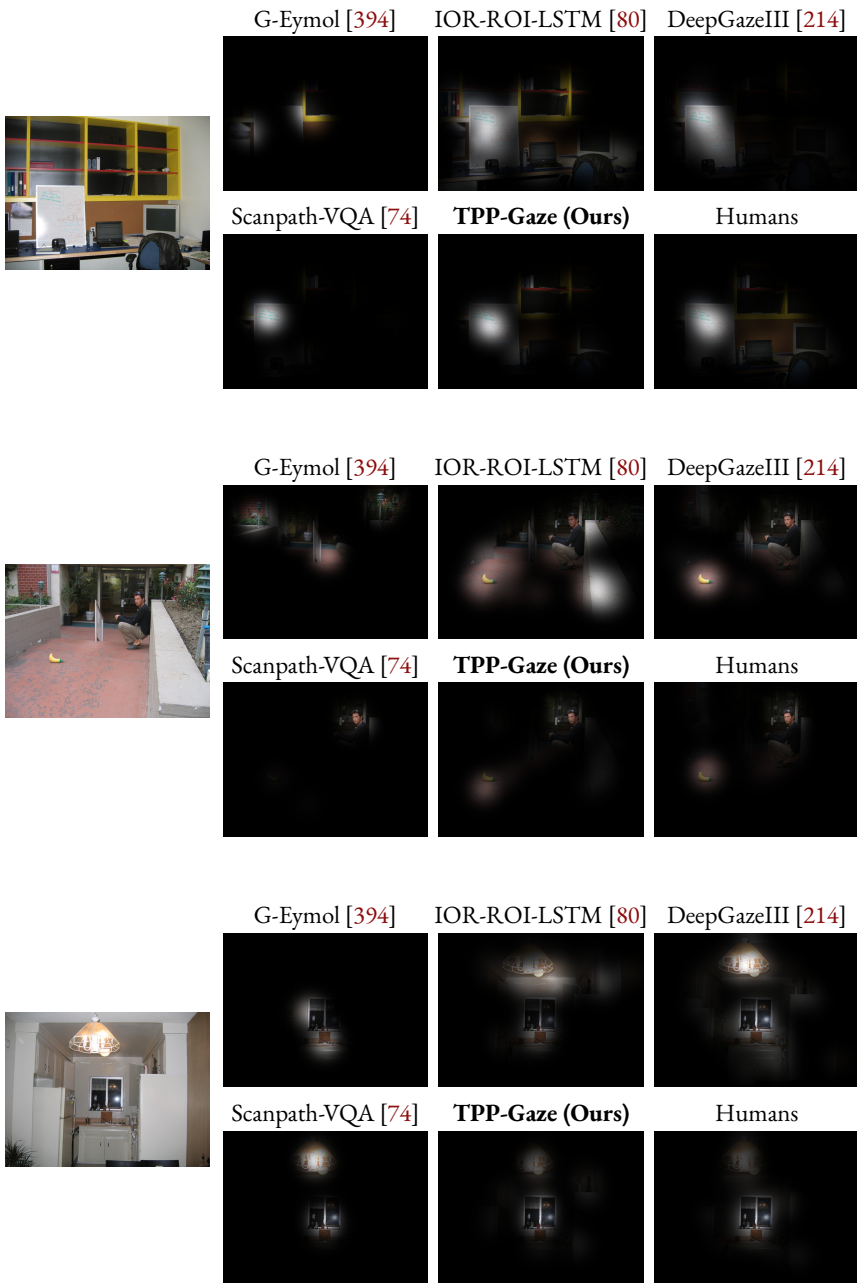


Figure B.10: Saliency maps of sample images from FiFa dataset computed from the fixations generated by the considered scanpath models. For completeness, we include DeepGazeIII, but note that its training procedure also involves saliency prediction.

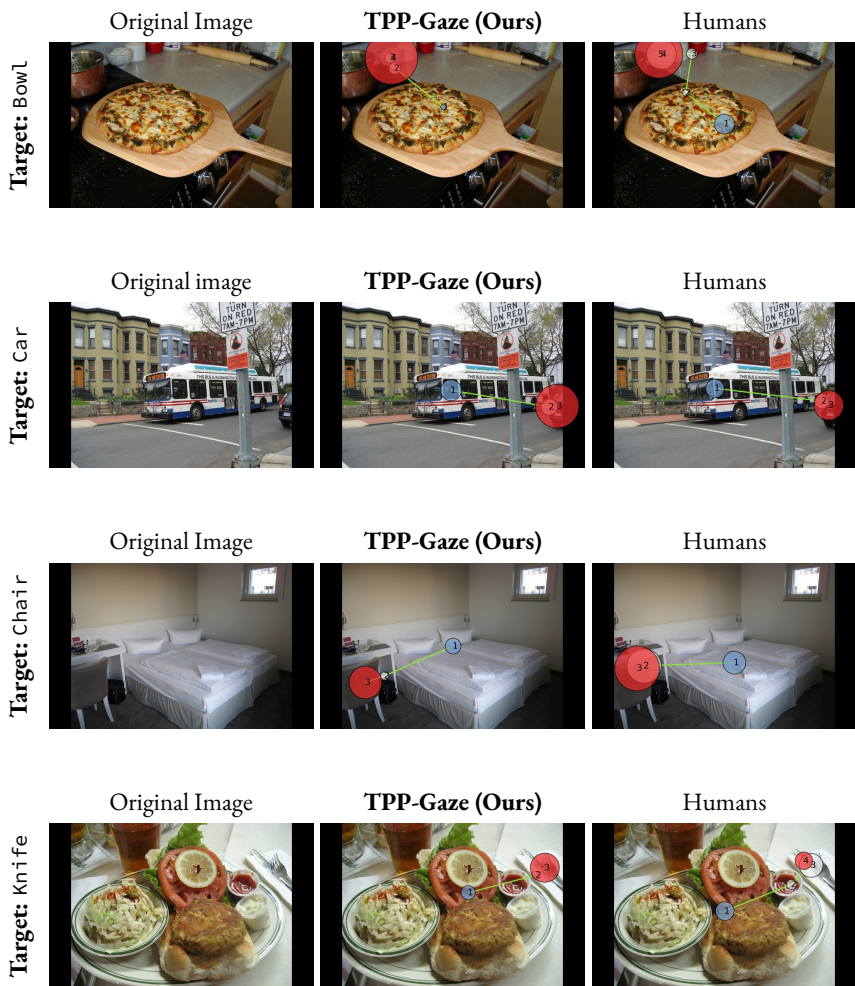


Figure B.11: Qualitative comparison of simulated and human scanpaths on the COCO-Search18 dataset for the visual search task.

B.2 MODELING HUMAN GAZE BEHAVIOR WITH DIFFUSION MODELS FOR UNIFIED SCANPATH PREDICTION

Additional qualitative results are depicted from Fig. B.12 to Fig. B.15 on COCO-FreeView [387], MIT1003 [194], and OSIE [374] for free-viewing and COCO-Search18 [78] for the visual search task, respectively. The qualitative results highlight the accuracy of ScanDiff in predicting human-like scanpaths. Other methods, however, demonstrate limitations by either focusing excessively on specific elements or producing shorter scanpaths than the ones exhibited by humans.

The qualitative results further support the findings of our scanpath variability analysis, highlighting the ability of ScanDiff to generate diverse yet human-like scanpaths across different viewing tasks. As shown in the comparison evaluation from Fig. B.17 to Fig. B.20 existing methods often produce scanpaths that are either overly deterministic – failing to capture the natural variability of human gaze behavior – or overly stochastic, resulting in implausible trajectories. This is particularly true for Gazeformer that, as shown in Fig. B.19 and Fig. B.20, produces scanpaths that are identical to each other over the simulations, essentially generating the same fixation pattern repeatedly regardless of the inherent variability present in human visual attention processes. While Gazeformer achieves reasonable performance on task-oriented datasets as shown in the quantitative results, its deterministic nature fundamentally limits its ability to model the stochastic aspects of human gaze behavior that our approach successfully captures.

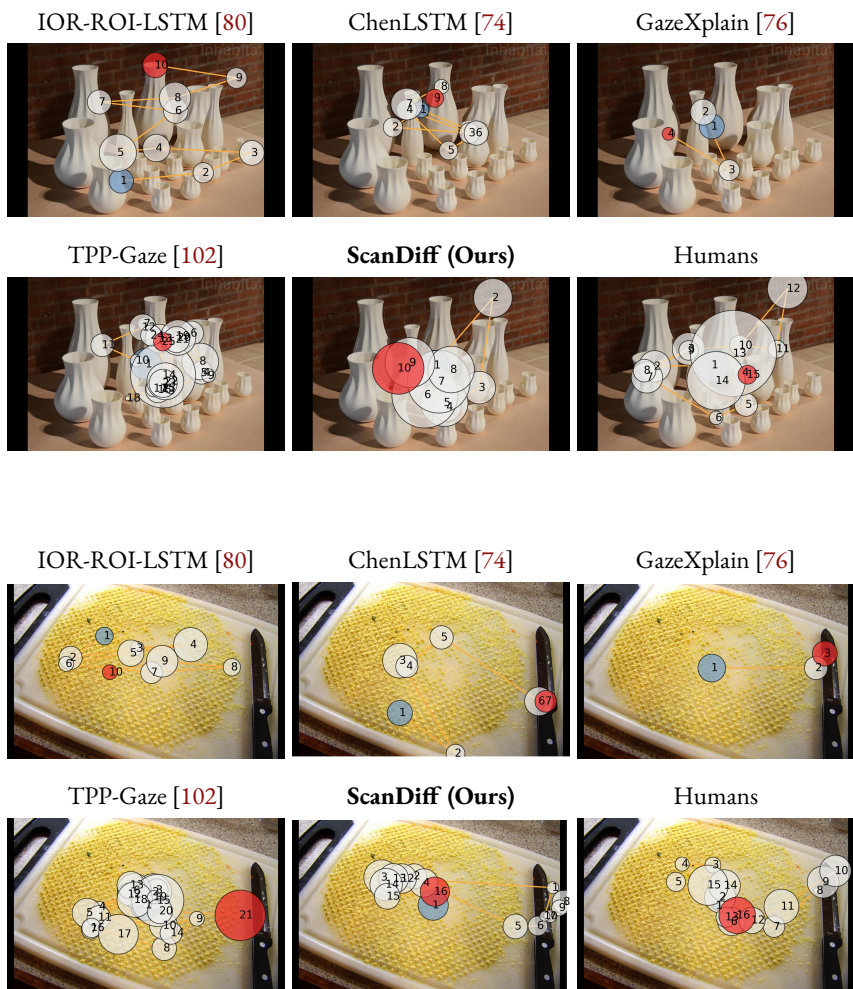


Figure B.12: Qualitative comparison of simulated and human scanpaths on the COCO-FreeView dataset.

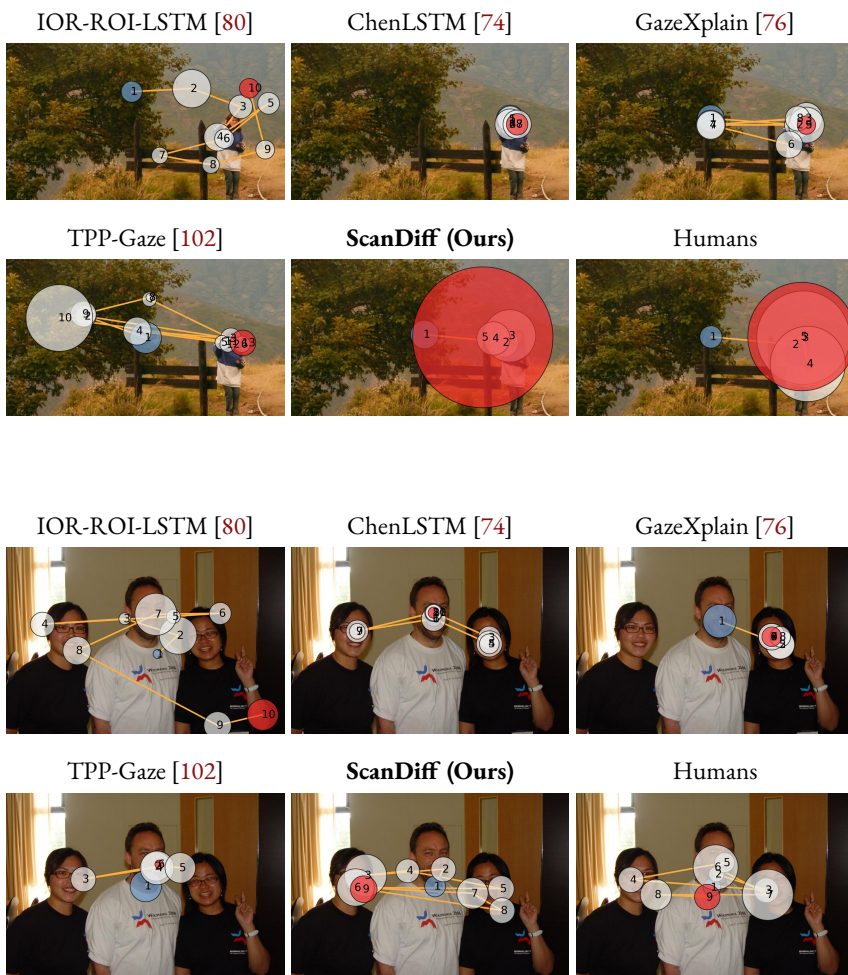


Figure B.13: Qualitative comparison of simulated and human scanpaths on the MIT1003 dataset.

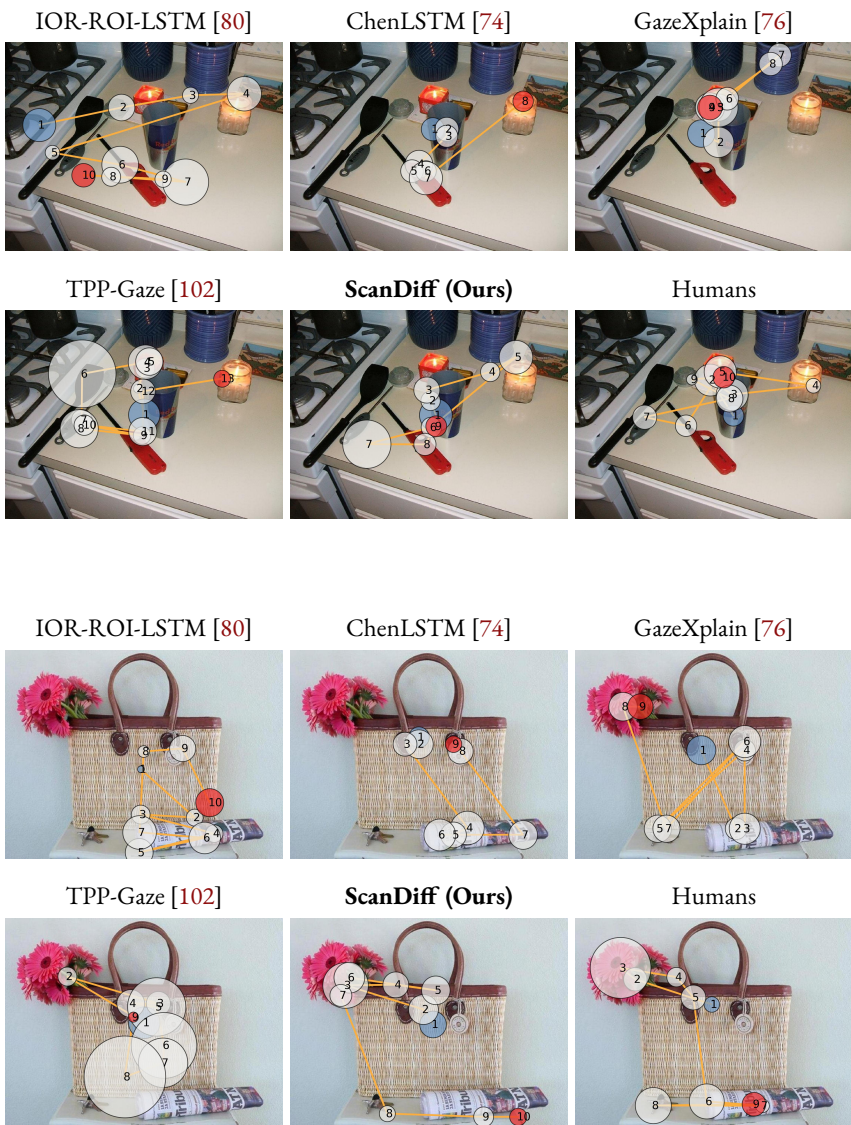


Figure B.14: Qualitative comparison of simulated and human scanpaths on the OSIE dataset.

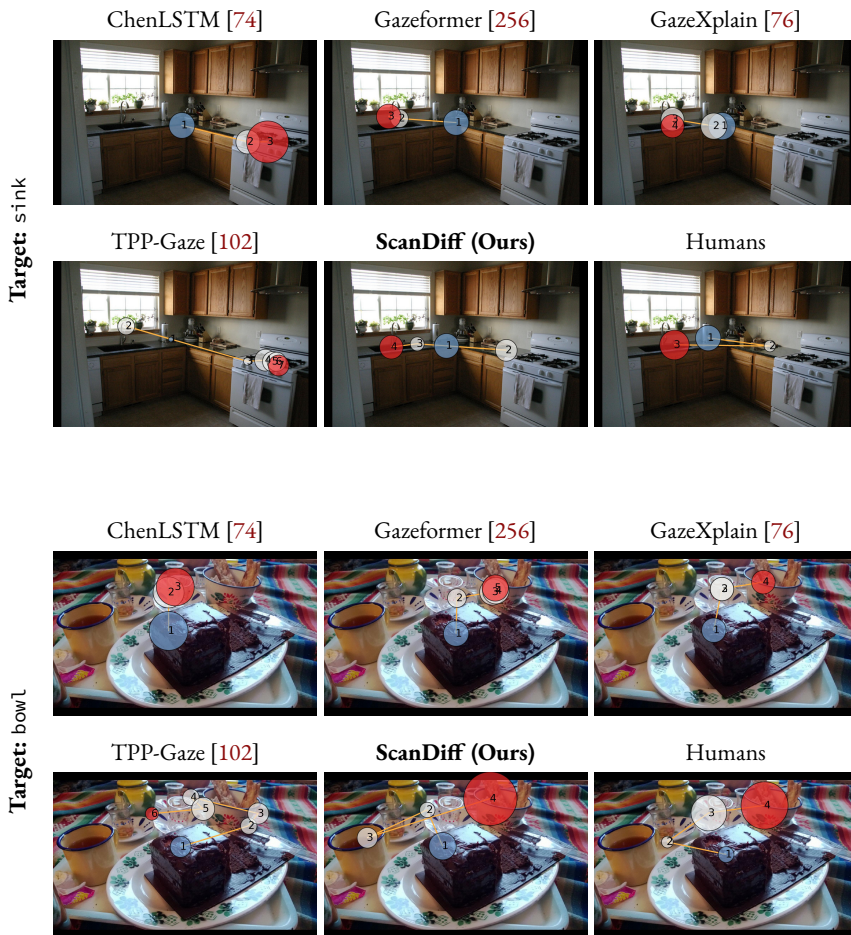


Figure B.15: Qualitative comparison of simulated and human scanpaths on the COCO-Search18 (TP) dataset for the visual search task.

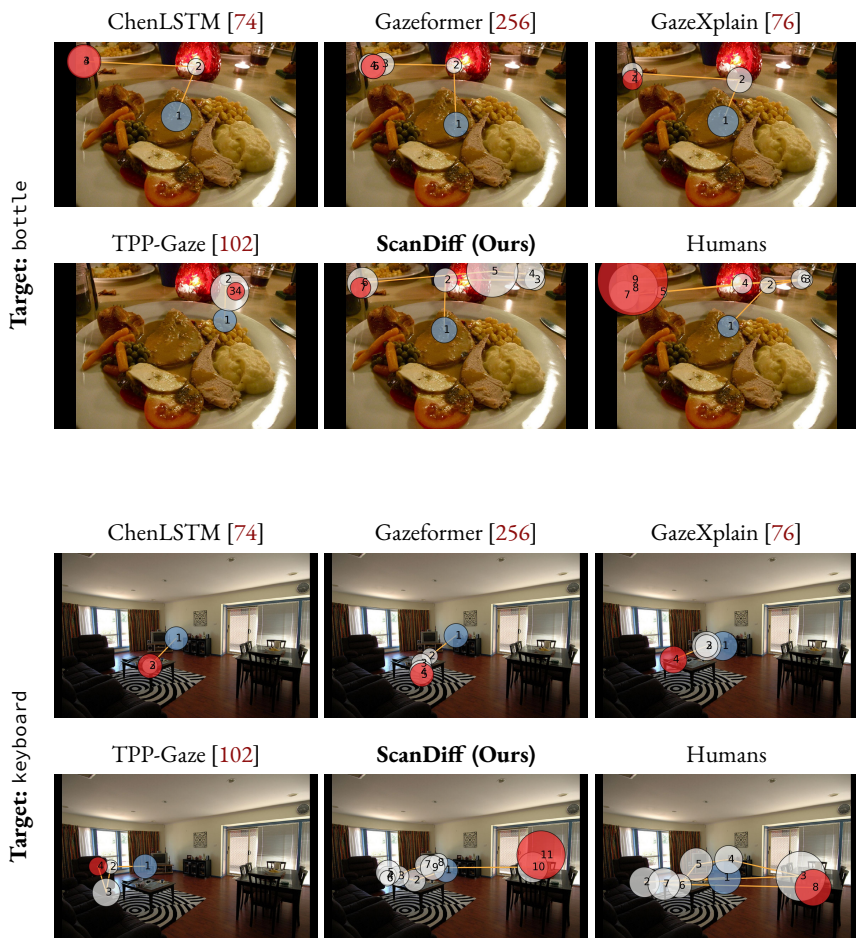


Figure B.16: Qualitative comparison of simulated and human scanpaths on the COCO-Search18 (TA) dataset for the visual search task.

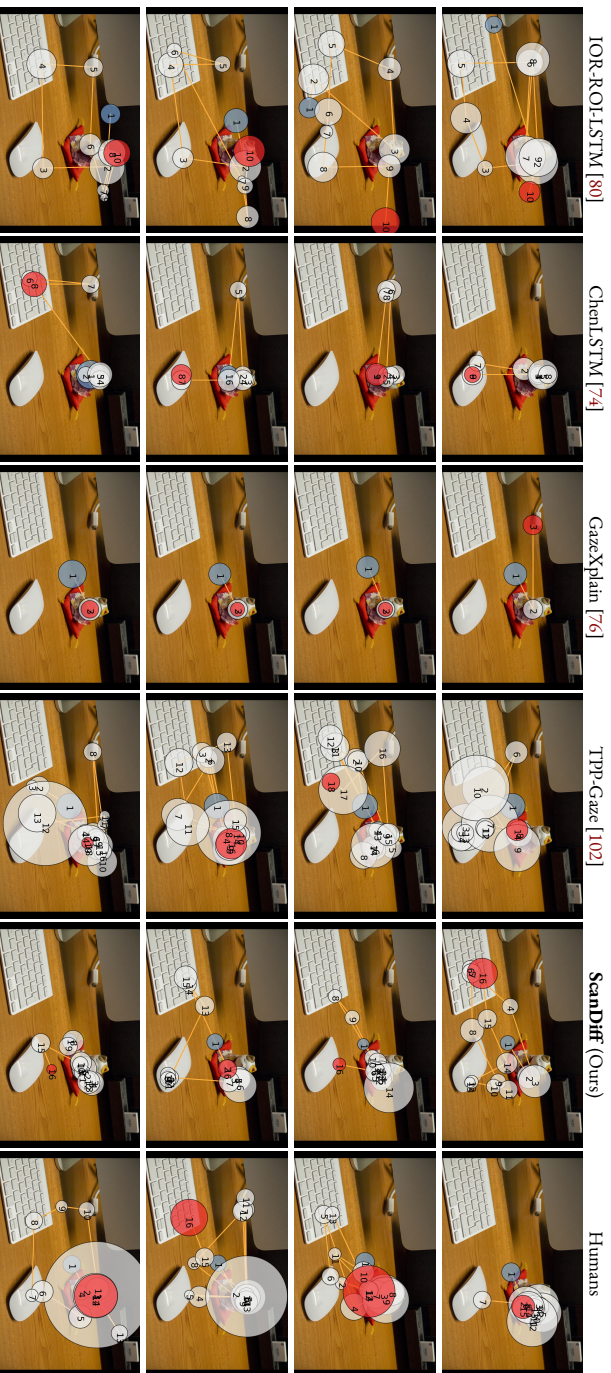


Figure B.17: Qualitative comparison of the variability in simulated and human scanpaths on the COCOFreeView dataset. Each row corresponds to a different simulation or a different human observer.



Figure B.18: Qualitative comparison of the variability in simulated and human scanpaths on the MIT1003 dataset. Each row corresponds to a different simulation or a different human observer.

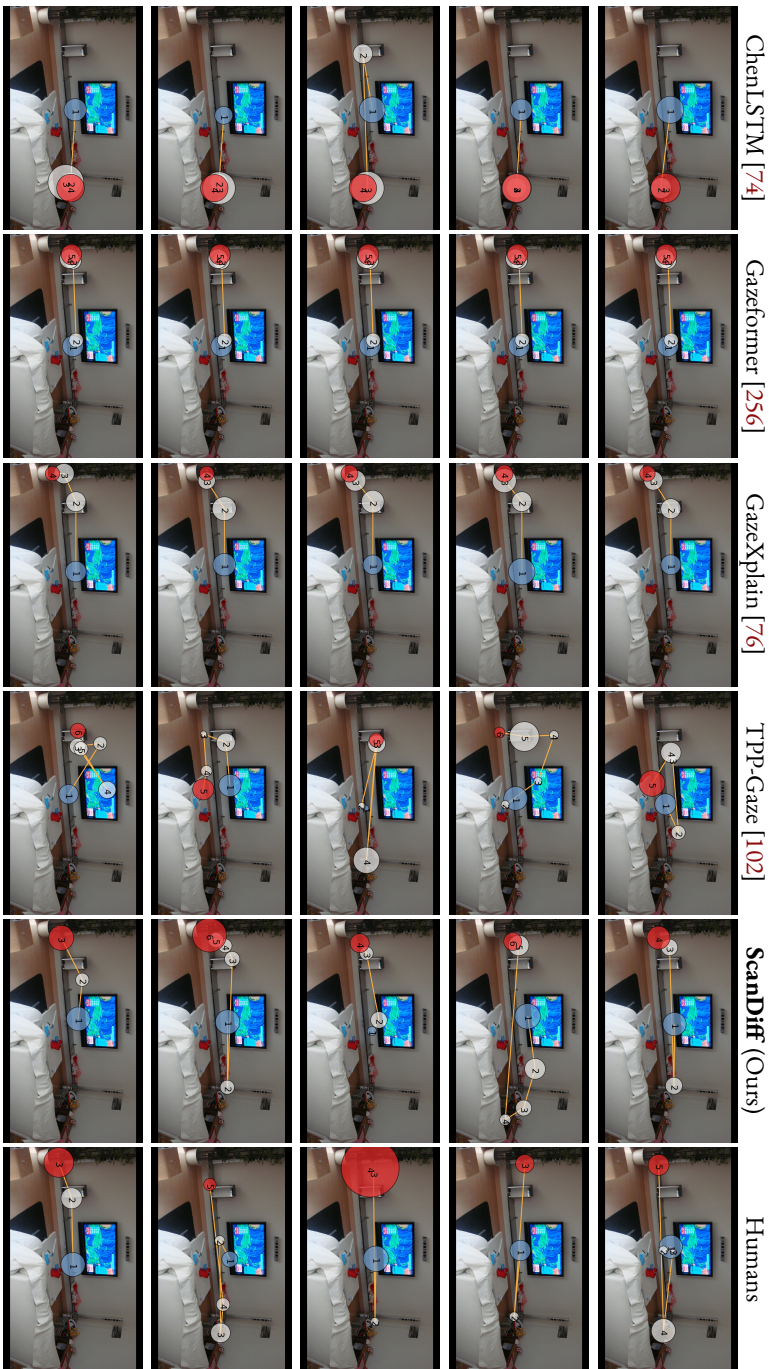


Figure B.19: Qualitative comparison of the variability of simulated and human scapaths on the COCO-Search18 (TP) dataset for the viewing task: potted plant. Each row corresponds to a different simulation or a different human observer.

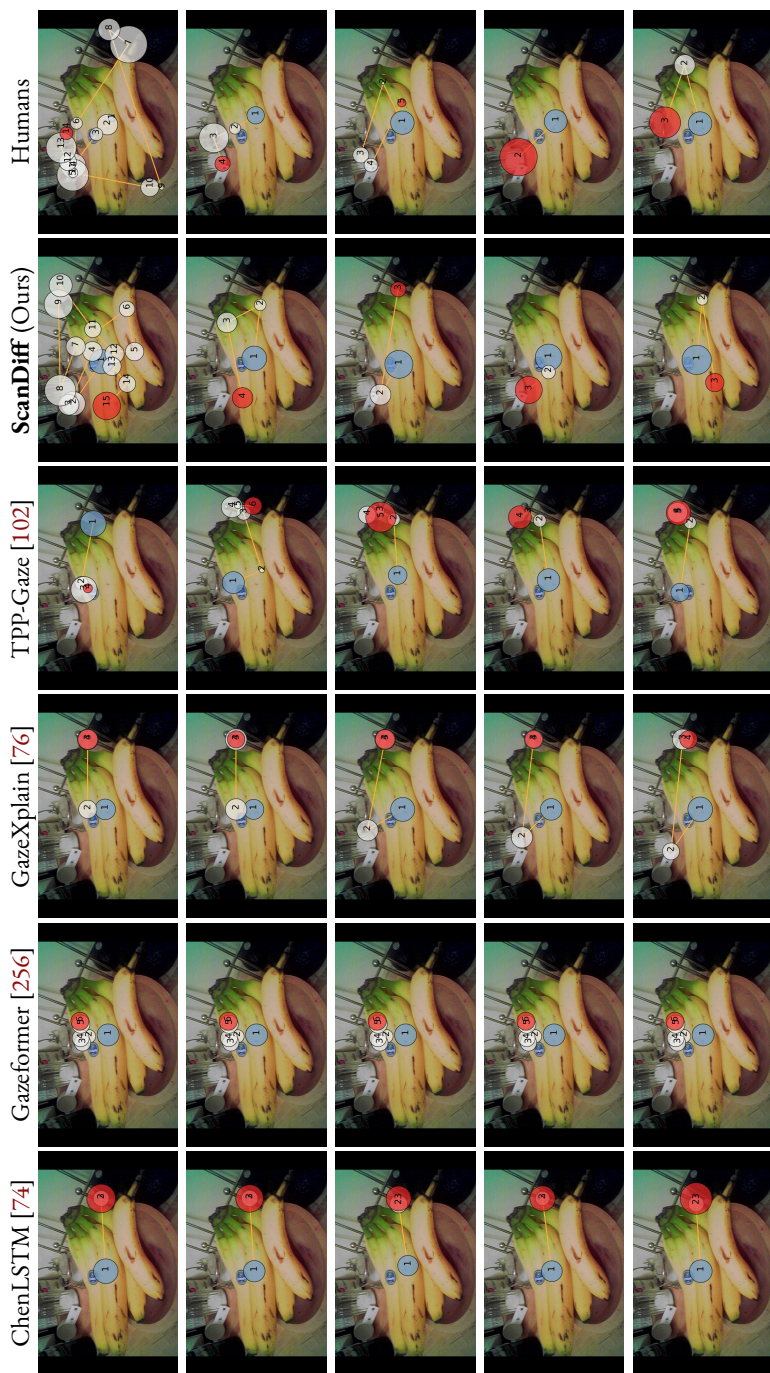
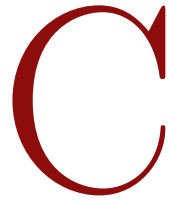


Figure B.20: Qualitative comparison of the variability of simulated and human scanpaths on the COCO-Search18 (TA) dataset for the viewing task: fork. Each row corresponds to a different simulation or a different human observer.



Supplementary Material for Chapter 5

C.1 SANCTUARIA-GAZE: A MULTIMODAL EGOCENTRIC DATASET FOR HUMAN ATTENTION ANALYSIS IN RELIGIOUS SITES

Additional examples from the RGB videos collected in the four sanctuaries are presented in Fig. C.1, while in Fig. C.2, we show further qualitative samples of the proposed annotation framework.

Church 01



Church 02



Church 03



Church 04



Figure C.1: Qualitative examples from the egocentric videos captured in the four churches.

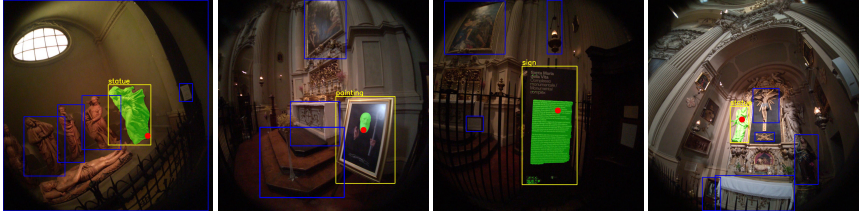
Church 01



Church 02



Church 03



Church 04



Figure C.2: Qualitative examples of the final object annotations obtained through our automatic annotation framework.

