UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA

University of Modena and Reggio Emilia
Department of Engineering "Enzo Ferrari"

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

Scaling Vision–Language Understanding:

From Image Captioning to Knowledge-Grounded Multimodal Large Language Models

Ph.D. Dissertation
in Computer Engineering and Science

Nicholas Moratelli

Advisor: Prof. Rita Cucchiara
Director of the School: Prof. Luigi Rovati

Advisor:

Prof. Rita Cucchiara — University of Modena and Reggio Emilia

Director of the School:

Prof. Luigi Rovati — University of Modena and Reggio Emilia

Review Committee:

Leonardo Filipe Rodrigues Ribeiro, Ph.D. — Amazon AGI
Prof. Paolo Rota — University of Trento

PhD thesis funded by the European Union – NextGenerationEU, Mission 4, Component 2 "From Research to Business" – Investment 3.3 "Introduction of innovative PhDs that respond to the innovation needs of companies and promote the hiring of researchers by companies".

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_ε, 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.

"The more I learn, the more I realize how much I don't know."

— Albert Einstein

to those who taught me how much there is still to learn

Scaling Vision–Language Understanding: From Image Captioning to Knowledge-Grounded Multimodal Large Language Models

Abstract

The multimodal integration of different modalities, most prominently text and images, has become a central pursuit in Artificial Intelligence. Enabling machines to perceive visual content and express or reason about it through language represents a crucial step toward more general and interpretable intelligence. However, despite remarkable recent progress, achieving consistent alignment between visual understanding, language generation, and factual reasoning remains a fundamental challenge. This thesis investigates this continuum, tracing the evolution of multimodal understanding from small-scale, task-specific models for image captioning to large-scale Multimodal Large Language Models (MLLMs) capable of knowledge-grounded reasoning. The first part of the thesis focuses on image captioning, a foundational vision–language task in which a model learns to generate natural language descriptions of visual input. Traditional captioning approaches often rely on reinforcement learning with hand-crafted evaluation metrics that can restrict fluency and semantic depth. To overcome these limitations, we introduce a self-trained reward modeling framework that learns to assess image–caption alignment without predefined rewards, producing captions that are more fluent, informative, and semantically consistent. In a complementary line of work, we develop a direct optimization strategy that unifies reward estimation and caption generation within a single training objective. By leveraging learnable, perceptually aligned evaluation signals, this approach provides a stable and flexible optimization process that bridges the gap between metric-driven and semantically grounded captioning. The second part broadens this perspective to the large-scale integration of vision and language. A comprehensive survey of MLLMs maps the evolving landscape of architectures, modality-alignment techniques, pretraining datasets, and evaluation benchmarks, provid-

ing a conceptual framework for understanding how these models process and reason over visual inputs. The third part explores Knowledge-Based Visual Question Answering (KB-VQA) and the integration of external knowledge into multimodal reasoning. While MLLMs demonstrate strong generalization, they often lack access to factual information beyond their training data. To address this limitation, we develop two complementary Retrieval-Augmented Generation (RAG) frameworks. Wiki-LLaVA integrates large-scale encyclopedic retrieval into multimodal reasoning, enabling access to external knowledge during inference, while ReflectiVA extends this paradigm with a dual reflection mechanism that allows the model to decide when retrieval is necessary and to validate retrieved evidence before answer generation. These approaches substantially improve factual accuracy and interpretability on knowledge-intensive VQA benchmarks while maintaining performance on visual-only tasks. Beyond knowledge access, this thesis addresses limitations of purely parametric multimodal models in compositional understanding by introducing a causal, dependency-aware modeling framework that injects explicit structural inductive biases into vision–language representations, enabling more faithful reasoning over complex linguistic relations. Overall, this thesis presents a coherent progression from captioning models to large-scale multimodal reasoning systems, contributing new methods for visual–linguistic alignment, a comprehensive synthesis of the MLLM landscape, retrieval-augmented architectures for knowledge grounding, and causal compositional modeling, moving multimodal AI beyond perception toward reliable, knowledge-grounded, and structurally informed vision–language understanding.

Scalare la comprensione visivo–linguistica: dall'Image Captioning ai Modelli Multimodali di Grandi Dimensioni basati sull'integrazione di conoscenza esterna

Sommario

L'integrazione multimodale di diverse modalità, in particolare testo e immagini, è diventata un obiettivo centrale dell'Intelligenza Artificiale. Consentire alle macchine di percepire contenuti visivi ed esprimerli o ragionarvi attraverso il linguaggio rappresenta un passo fondamentale verso forme di intelligenza più generali e interpretabili. Nonostante i recenti progressi, ottenere un allineamento coerente tra comprensione visiva, generazione linguistica e ragionamento fattuale rimane una sfida aperta. Questa tesi si colloca in questo continuum, tracciando l'evoluzione della comprensione multimodale dai modelli di image captioning ai Multimodal Large Language Models (MLLM) capaci di ragionamento fondato sulla conoscenza. La prima parte è dedicata all'image captioning, un compito fondativo visione–linguaggio in cui un modello genera descrizioni in linguaggio naturale a partire da ingressi visivi. Gli approcci tradizionali fanno spesso uso dell'apprendimento per rinforzo con metriche progettate manualmente, che possono limitare la fluidità e la ricchezza semantica delle descrizioni generate. Per superare tali limiti, proponiamo un framework di reward modeling auto-addestrato in grado di valutare l'allineamento immagine–didascalia senza ricompense predefinite, producendo descrizioni più fluide, informative e coerenti. In parallelo, sviluppiamo una strategia di ottimizzazione diretta che unifica la stima della ricompensa e la generazione della didascalia in un unico obiettivo di addestramento, offrendo un processo stabile e flessibile basato su segnali valutativi apprendibili e percettivamente allineati. La seconda parte estende l'analisi all'integrazione su larga scala di visione e linguaggio. Un'analisi sistematica dei MLLM esamina architetture, tecniche di allineamento tra modalità, dataset di pre-addestramento e benchmark di valutazione, fornendo un quadro concettuale per comprendere come questi modelli elaborano e ragionano sugli input visivi. La terza parte esplora il Knowledge-Based Visual Question Answer-

ing (KB-VQA) e il ruolo della conoscenza esterna nel ragionamento multimodale. Sebbene i MLLM mostrino una forte capacità di generalizzazione, essi spesso non dispongono di informazioni fattuali oltre quelle apprese in fase di addestramento. Per colmare questo limite, proponiamo due framework di Retrieval-Augmented Generation (RAG). Wiki-LLaVA integra il recupero di conoscenza enciclopedica nel ragionamento multimodale, mentre ReflectiVA introduce un meccanismo di riflessione duale che consente al modello di decidere quando il recupero sia necessario e di validare le evidenze prima della generazione della risposta. Questi approcci migliorano l'accuratezza fattuale e l'interpretabilità su benchmark di VQA che richiedono conoscenza esterna, mantenendo le prestazioni sui compiti puramente visivi. Infine, la tesi affronta i limiti dei modelli multimodali puramente parametrici nella comprensione composizionale, introducendo un framework causale e sensibile alle dipendenze che incorpora bias induttivi strutturali nelle rappresentazioni visione–linguaggio, consentendo un ragionamento più fedele su relazioni linguistiche complesse. Nel complesso, la tesi presenta un percorso coerente che va dai modelli di captioning ai sistemi di ragionamento multimodale su larga scala, contribuendo con nuovi metodi per l'allineamento visivo–linguistico, architetture di recupero per il grounding della conoscenza e modelli causali composizionali, spingendo l'Intelligenza Artificiale multimodale oltre la percezione verso una comprensione visione–linguaggio affidabile e fondata sulla conoscenza.

Contents

1 Introduction

Humans naturally integrate perception, language, and prior knowledge to make sense of the world. When observing a scene, we recognise objects, understand their relationships, recall relevant facts, and communicate this understanding through structured and meaningful language. This process is immediate and intuitive, yet notably complex. How can artificial systems approach this ability, combining vision, language, and reasoning in a grounded and reliable manner?

Recent advances in deep learning have led to remarkable progress in both computer vision and natural language processing. However, integrating these modalities into unified systems that can reliably perceive, reason, and communicate remains an open challenge. Vision–language models have demonstrated strong performance across a wide range of tasks, including image captioning, visual question answering, and multimodal dialogue. Yet, their success often relies on large-scale datasets and parametric memorization, rather than on principled mechanisms for grounding, reasoning, and knowledge integration.

For many years, Computer Vision and Natural Language Processing evolved as largely independent research areas. Vision systems focused on recognizing ob-

jects, detecting regions, and extracting visual features, while language models concentrated on syntax, semantics, and text generation. The advent of deep learning radically transformed both fields, enabling end-to-end learning and rich data-driven representations. As a result, the boundary between vision and language progressively blurred, giving rise to vision–language models capable of jointly processing images and text.

Image captioning represented one of the first successful attempts to bridge visual perception and language generation. By learning to describe images in natural language, captioning models provided a controlled setting to study multimodal alignment. Over time, captioning architectures evolved from recurrent models conditioned on convolutional features to Transformer-based systems optimized with reinforcement learning objectives. These models demonstrated how visual content could be translated into fluent text, but also revealed important limitations: captions often remained generic, underspecified, or overly dependent on dataset-specific biases.

Beyond captioning, many multimodal tasks require more than describing what is visible. Answering questions about images often demands access to external knowledge that is not explicitly present in the scene, such as identifying specific entities, recalling encyclopedic facts, or reasoning about attributes and events. For example, a model may correctly recognize objects in an image, yet hallucinate unsupported facts or fail to answer questions that require background knowledge beyond visual evidence. Such failure modes highlight the limitations of purely parametric multimodal models and motivate the integration of retrieval mechanisms into vision–language systems.

The emergence of Large Language Models (LLMs) further reshaped the multimodal landscape. When coupled with visual encoders, these models—commonly referred to as Multimodal Large Language Models (MLLMs)—exhibit strong generalization across tasks such as visual question answering, multimodal dialogue, and instruction following. Despite their impressive capabilities, MLLMs still suffer from critical shortcomings: they may hallucinate facts, struggle with long-tail knowledge, and often fail to respect the compositional structure of language. For instance, models may confuse subjects and objects in

complex descriptions or generate plausible but unsupported explanations. These observations suggest that scaling alone is insufficient, and that principled approaches to reward design, knowledge integration, and linguistic structure are required.

The challenges outlined above motivate the central research questions addressed in this thesis.

1.1 Core Challenges in Vision–Language Modeling

The overarching objective of this thesis is to improve the grounding, expressiveness, and reasoning capabilities of vision–language models. In particular, the work addresses four core challenges that repeatedly emerge across multimodal tasks.

Improving the descriptiveness of image captions. Standard captioning models are typically optimized using cross-entropy pretraining followed by reinforcement learning objectives such as CIDEr optimization through SCST [236]. While effective at improving benchmark scores, these objectives often encourage safe and generic descriptions that lack semantic richness. This thesis proposes and evaluates learned reward models that better align caption generation with human preferences and promote more informative descriptions.

Understanding the design principles of Multimodal LLMs. The rapid proliferation of MLLMs has introduced a wide range of architectural choices, including different visual encoders, projection layers, and training pipelines. However, the impact of these design decisions is not always well understood. This thesis contributes a systematic analysis of multimodal architectures, with a particular focus on the role of visual backbones in shaping multimodal reasoning and alignment.

Integrating external knowledge through multimodal retrieval. Although parametric multimodal models encode vast amounts of information in their weights, they struggle with long-tail, factual, or evolving

knowledge. Retrieval-Augmented Generation (RAG) offers a promising alternative by enabling models to access external multimodal documents at inference time. This thesis explores how retrieval mechanisms can be structured, when they should be triggered, and how retrieved evidence can be effectively integrated into multimodal reasoning.

Enforcing compositional structure in vision–language models. Many existing vision–language models treat language as a flat sequence and compress it into a single embedding, ignoring syntactic and semantic dependencies between words. This often leads to failures in compositional understanding, such as confusing subjects and objects or misinterpreting relational statements. The final part of this thesis addresses this limitation by introducing causal and syntactic structure into multimodal models, enabling more faithful reasoning over linguistic relations.

Together, these challenges define a research agenda that moves beyond surface-level multimodal alignment toward models that are descriptive, knowledgeable, and grounded in explicit structural representations.

This thesis argues that progress in multimodal intelligence requires moving beyond purely data- and scale-driven solutions. Instead, it advocates for models that explicitly integrate learned reward signals, external knowledge access, and structural inductive biases reflecting the causal and compositional nature of language.

1.2 Thesis Structure and Research Trajectory

This dissertation is organised into four main technical chapters, each addressing a key aspect of multimodal vision–language modeling. The overall structure reflects the evolution of the research carried out during the Ph.D., progressing from task-specific image captioning models to general-purpose multimodal systems capable of reasoning, knowledge integration, and compositional understanding.

Image Captioning. Chapter 3 is dedicated to image captioning and presents two complementary studies that investigate how captioning models can be trained to generate more fluent, descriptive, and human-aligned captions. These

works examine alternatives to hand-crafted evaluation metrics by introducing learned reward models and direct optimisation strategies, highlighting the limitations of traditional captioning objectives.

Multimodal Large Language Models. Chapter 4 shifts the focus to multimodal large language models and presents two contributions analyzing the design principles and empirical behavior of large-scale multimodal architectures. In particular, this chapter investigates how different visual backbones influence multimodal reasoning and performance.

Multimodal Retrieval-Augmented Generation. Chapter 5 addresses the challenge of grounding multimodal models in external knowledge through retrieval-augmented generation. The works presented in this chapter focus on knowledge-intensive visual question answering and analyze how multimodal retrieval pipelines can be designed, evaluated, and integrated into generative models.

Compositional Vision–Language Modeling. Chapter 6 focuses on compositionality in vision–language models and presents a contribution that introduces explicit causal and dependency-based structure into multimodal representations, improving compositional reasoning and generalization.

In summary, this thesis presents a coherent research trajectory that starts from image captioning, progresses through large-scale multimodal modeling, incorporates external knowledge via retrieval, and culminates in structured compositional reasoning. Chapter 7 concludes the dissertation by summarizing the contributions and outlining promising directions for future research in multimodal artificial intelligence.

1

2

Literature Review

This chapter reviews the literature most relevant to the research directions and contributions presented in this thesis. The review is organized to reflect the conceptual trajectory of the dissertation, progressing from early vision–language models for image captioning to large-scale multimodal systems, retrieval-augmented reasoning, and compositional vision–language modeling.

We begin in Section 2.1 by surveying the evolution of image captioning, from early template-based and encoder–decoder approaches to attention-based, Transformer-based, and large-scale pre-trained captioning systems. Image captioning is discussed both as a foundational vision–language task and as a testbed for studying multimodal alignment, training objectives, and semantic expressiveness.

Section 2.2 then reviews Multimodal Large Language Models (MLLMs), focusing on architectural design choices, visual encoders, instruction tuning strategies, and the role of large-scale pre-training in shaping multimodal behavior. Retrieval-augmented multimodal models are discussed separately in Subsection 2.2.2, highlighting how external knowledge integration addresses limi-

tations of purely parametric systems in knowledge-intensive tasks such as visual question answering.

Finally, Section 2.3 examines compositional vision–language modeling, reviewing approaches that explicitly target syntactic, semantic, and causal structure in multimodal representations. These include data-centric strategies based on structured supervision as well as model-based approaches grounded in causal graphical models and syntactic representations.

Rather than providing an exhaustive survey, this chapter emphasizes representative approaches and recurring design patterns that inform the methodological choices explored in later chapters. Throughout the review, particular attention is given to the limitations of existing methods, which directly motivate the contributions developed in this dissertation.

2.1 Image Captioning

Image captioning aims to automatically describe visual content through natural language, bridging the gap between computer vision and natural language understanding. Early approaches relied on predefined templates populated by object detectors that identified entities and attributes within the image [255, 314]. The advent of deep learning revolutionized this paradigm with the introduction of encoder–decoder architectures, where a convolutional neural network (CNN) encodes the image and a recurrent neural network (RNN) decodes the resulting representation into a natural language sequence [280, 133, 236]. This architecture, originally inspired by neural machine translation, became the standard for several years and was extensively explored in many vision–language tasks [218, 217, 32].

A key advancement came with the introduction of attention mechanisms, which enabled the model to focus selectively on salient visual regions when generating each word [306, 188, 22]. Later developments enriched these mechanisms by incorporating spatial and semantic graphs to explicitly model object relationships and context [315, 312]. With the success of Transformers in NLP, attention-based architectures naturally evolved into fully Transformer-based cap-

tioners [117, 68, 67, 166, 29]. These models capture long-range dependencies and better align visual and textual modalities, often extending the paradigm with external memory modules and richer region-level representations [68, 29, 166].

More recently, the field has witnessed a paradigm shift through large-scale cross-modal pre-training. Models such as OSCAR [163], VinVL [337], OFA [286], GIT [285], and similar architectures [114] have been pre-trained on millions of image–text pairs collected from the web, learning general-purpose multimodal representations that can be fine-tuned for captioning. In parallel, the CLIP model [225] introduced a powerful way to align vision and language embeddings, inspiring new captioning pipelines that exploit CLIP-based guidance [248, 62, 324, 74, 9]. These approaches use CLIP either to encode visual inputs or as a reward function to optimize semantic alignment, thereby producing richer and more descriptive captions.

2.1.1 Training Objectives and Supervision

From a training perspective, early models were trained using cross-entropy loss [280, 133, 306]. However, optimizing sequence-level metrics soon led to reinforcement learning (RL) strategies, where captioners are viewed as agents that maximize rewards derived from quality metrics such as BLEU [209], ROUGE [169], CIDEr [279], and SPICE [20]. Several refinements have been proposed along this line, including contrastive learning [69], discriminability-based rewards [191], and unified embedding control [235]. Nevertheless, reliance on reference captions often limits the semantic diversity of generated text. To address this issue, recent approaches propose CLIP-based supervision that removes the dependency on ground-truth captions and instead leverages semantic similarity in the embedding space [62, 324, 74, 9]. This trend highlights a shift from syntactic imitation to semantic alignment as the main objective for caption quality.

2.1.2 LLM-Based Image Captioning

The success of LLMs has inspired a new wave of research integrating vision and language through shared architectures. Several studies have explored ways to endow pre-trained LLMs with visual understanding capabilities [351, 238, 140, 139, 89]. For instance, ZeroCap [265] aligns text generated by GPT-2 [226] with visual inputs using CLIP as external feedback, performing optimization at inference time without explicit retraining. Other approaches, such as ClipCap and SmallCap [197, 230], instead introduce small cross-attention modules that learn to project visual embeddings into the language model's latent space. These lightweight strategies demonstrate that the expressive power of LLMs can be transferred to multimodal tasks with minimal architectural changes.

At a larger scale, the field has converged toward MLLMs that unify text and image processing within the same generative framework [159, 158, 49]. In these architectures, image captioning plays a dual role: it is both a pre-training task for aligning visual and textual modalities, and a downstream benchmark for evaluating descriptive and reasoning capabilities. Thanks to their underlying language model, these systems produce captions with richer contextualization, style, and factual grounding [1, 78, 162].

A further development involves instruction tuning and human alignment. Modern LLM-based captioners are often refined through instruction datasets [290, 291, 121, 61] or reinforcement learning from human feedback (RLHF) [357, 208, 271, 261]. These techniques aim to control model behavior and improve the faithfulness and safety of generated content [179, 177]. They also address biases and hallucinations that may arise when coupling visual information with text generation.

Overall, the evolution from early neural captioners to LLM-based multimodal systems represents a substantial change in both scale and philosophy. The task of image captioning, once narrowly defined as describing an image, is now embedded within a broader framework of multimodal reasoning and language understanding. In this thesis, we build upon these advances by exploring strategies to enhance the semantic quality and descriptiveness of captions without depend-

ing on pre-trained LLMs, focusing instead on robust multimodal alignment and efficient training objectives.

2.2 Multimodal Large Language Models

While image captioning has traditionally served as a focused benchmark for vision–language modeling, recent advances in large language models have enabled a shift toward general-purpose multimodal systems.

LLMs have profoundly reshaped artificial intelligence research, demonstrating exceptional reasoning, comprehension, and generative capabilities. Prominent examples such as GPT-4 [15] and Gemini [23] have showcased the power of large-scale pre-training combined with alignment techniques like instruction tuning [208, 262] and RLHF [258]. Open-source initiatives including Flan-T5 [65], LLaMA [271, 96], Vicuna [61], and Alpaca [262] have greatly accelerated research progress, enabling fine-tuning and adaptation in diverse domains. More recent efforts, such as the Gemma [264], Qwen [27], Phi [13, 14], and SmolLM [19] families, have further demonstrated that model efficiency and high-quality training data can often rival sheer model scale, marking an important shift in LLM development philosophy.

2.2.1 General Multimodal LLMs

Building upon the success of language models, Multimodal Large Language Models extend LLMs to process and reason jointly over visual and textual modalities. Early explorations in this direction, such as VisualGPT [48] and Frozen [275], leveraged pre-trained language models to enhance vision–language capabilities for tasks like image captioning and visual question answering. Subsequent works introduced more structured multimodal integration through architectures like Flamingo [18] and BLIP-2 [158], which inject image features via trainable cross-attention layers or Q-Former modules that map visual embeddings into the language model's latent space. This architectural trend was later extended by FROMAGe [140], Kosmos-1 [118], MiniGPT-4 [352], and Instruct-

BLIP [70], which further refined the interplay between modalities and enhanced instruction-following skills.

A major breakthrough came with the LLaVA family [179, 177, 178], which introduced visual instruction tuning — aligning vision and language representations using datasets curated with GPT-4. This approach quickly became a standard recipe for developing MLLMs, and it has since influenced a wide range of recent multimodal systems. Alongside fusion strategies, the quality of the visual encoder plays a critical role in overall multimodal performance. Most architectures employ Vision Transformers (ViTs) trained with CLIP-style contrastive learning [225], where CLIP ViT-L is a common backbone. Improvements have been proposed through variants such as SigLIP [330, 273], which refines the contrastive objective, and DINO-based encoders [39, 206], which rely on large-scale self-supervised pre-training. Recent work has also explored multi-backbone fusion [174, 90, 186, 269, 270], where multiple encoders are combined to enrich visual feature diversity. Other approaches, such as PaLI [56, 53], scale up the visual backbone to billions of parameters, while alternative studies [249] demonstrate that smaller, multi-scale ViT-based models can achieve comparable results with greater efficiency. Despite the impressive progress, challenges remain regarding fair evaluation and the reproducibility of results, as many systems rely on curated or proprietary datasets [146, 73, 27, 184, 147].

2.2.2 Retrieval-Augmented Multimodal LLMs

While MLLMs effectively integrate visual understanding and linguistic reasoning, their performance remains bounded by the knowledge encoded during pre-training. To mitigate this limitation, recent research has incorporated retrieval-augmented generation (RAG), a paradigm where models dynamically access external knowledge sources at inference time. Retrieval-augmented techniques were first explored in language modeling [100, 33, 123, 129, 282], expanding the input context of frozen LLMs with relevant passages retrieved from large text corpora or the web [201]. These methods improve factual grounding and interpretability while reducing the need for massive parameter growth.

Recently, retrieval augmentation has been extended to multimodal domains [29, 240, 115, 219], enabling models to retrieve not only text but also image–text pairs relevant to a query. For instance, REVEAL [115] encodes world knowledge into a large-scale multimodal memory, while SmallCap [230] applies retrieval for image captioning tasks. In the context of knowledge-based visual question answering (VQA), benchmarks such as OK-VQA [193, 246], Encyclopedic-VQA [194], and InfoSeek [57] have emphasized the need for external knowledge to answer complex visual–textual queries. Recent works [293, 225, 303, 259, 148] use contrastive encoders to retrieve entity-linked passages or images, while others combine retrieval with multimodal reasoning [221, 310]. Specifically, RoRA-VLM [221] introduces a visual token refinement mechanism to filter irrelevant visual information, and EchoSight [310] adopts a Q-Former-based re-ranking module to reorder retrieved text before feeding it to the MLLM. Following this direction, this thesis presents Wiki-LLaVA [3] as a retrieval-augmented multimodal framework grounded in hierarchical access to large-scale external knowledge, and subsequently advances this paradigm by investigating retrieval-awareness and relevance modeling mechanisms for multimodal generation (see Chapter 5).

In summary, the development of MLLMs represents a natural extension of language models toward multimodal reasoning. While general MLLMs focus on aligning vision and language representations through large-scale instruction tuning and architectural design, retrieval-augmented approaches extend these capabilities by integrating external knowledge dynamically. Together, these directions move toward more general-purpose, knowledge-aware multimodal systems capable of bridging perception, reasoning, and grounded generation.

2.3 Compositional Vision–Language Modeling

2.3.1 Compositional Methods

Despite the impressive capabilities of large-scale multimodal models, many systems still struggle with compositional understanding, motivating methods that

aim to enhance the reasoning capabilities of Vision–Language Models (VLMs) through explicitly structured supervision. A common strategy consists of generating *hard negatives*, where specific words in the ground-truth caption are replaced or swapped to produce challenging training examples [327, 336, 119, 35, 198, 81, 254, 203, 320, 104]. Such samples are typically produced through rule-based approaches or with the help of Large Language Models (LLMs). Other works generate dense captions using synthetic or real video data [40, 253], or by combining multiple modules such as segmentation networks (SAM [137]) and captioners (BLIP-2 [158]), as in DAC [80]. SAM has also been coupled with Stable Diffusion [237] to generate hard negative *images* [239], while other approaches [153, 66, 143] use Diffusion Models (DMs) [106] as alternative VLMs, exploiting their noise prediction error as an estimate of image–caption similarity. Finally, Stable Diffusion has been used as a regularizer to fine-tune CLIP [30].

Beyond DM-based models, most compositional approaches build upon fine-tuning or adapting CLIP. For example, [336] employ a hinge loss with a curriculum-based adaptive margin, while DAC [80] uses a Multiple Instance Learning objective. Curriculum learning is also adopted in [254], and [349] iteratively retrain CLIP using sparse codebook representations of image features. [203] propose a local hard-negative loss based on dense alignment between image patch embeddings and textual token embeddings. In VerbCLIP [292], a dependency parser is used to extract (subject, verb, object) triplets from captions, representing subjects and objects as vectors and verbs as transformation matrices applied to them. Other approaches embed CLIP within larger multimodal frameworks; for instance, CoVLM [157] combines a detection network with an LLM that communicates via special tokens. A few methods adopt *generative pre-trained* VLMs, showing clear advantages over encoder-only models, likely due to the richer compositionality induced by next-token prediction pre-training. For example, [104] introduce *Adaptive Scene Graph Tokens* to adapt both CLIP and BLIP-2 [158] to predict scene graph structures, while [281] employ LLaVA [179] with classifier-free guidance to compare predictions on original and masked images. BLIP is also used by [173] to mitigate linguistic bias during pre-training, and Cap/CapPa [274] introduce generative architectures trained via a hybrid

autoregressive and parallel prediction objective. Inspired by these findings, this thesis proposes a semi-parallel decoding strategy guided by a causal graph model (CGM), demonstrating its applicability to both encoder-based and generative VLMs (see Chapter 6).

2.3.2 Causal Graphical Models

Causal Graphical Models (CGMs) are widely employed in causal learning to represent dependencies among variables [241, 214]. Each node denotes a variable, and directed edges encode causal relations, defining the parent–child structure of the system. The joint distribution of all variables can thus be expressed through the *disentangled factorization* [241, 214], obtained as the product of conditional distributions of each variable given its parents. This formulation is assumed to be *causally sufficient*, which reduces the number of dependencies to learn and, consequently, the amount of data required for training [241]. To the best of our knowledge, the only work applying CGMs to vision–language compositional reasoning is ComCLIP [128], where *Independent Causal Mechanisms* [210, 94, 241] model subject–object–action relations. However, ComCLIP differs substantially from our approach, as each mechanism there is computed directly from CLIP similarities between words and sub-images, without explicit causal modeling.

2.3.3 Syntactic Trees

In *Dependency Grammar*, dependency relations express syntactic and semantic connections between words, where one word (the "head") governs the grammatical behavior of its dependents [202]. These relations can be represented as a *Dependency Tree* (DT), extracted automatically by parsers such as [108, 341, 82]. DTs have been integrated into language models to enhance syntactic awareness, for example by predicting whether future tokens are heads or dependents [313], or by modulating attention maps with learned dependency structures [72]. Dependency-based distances have also been employed as auxiliary losses [83] or to modulate attention [109].

In the vision–language domain, syntactic structures are leveraged to improve multimodal alignment. For instance, [256] use DTs to transform textual questions into templates for CLIP fine-tuning, and [161] generate modified sentences by swapping words to define pretext tasks. Constituency trees, which group words by grammatical categories, have also been used to generate phrase-level hard negatives [320] or to extend contrastive losses by maximizing intra-phrase similarity [338]. In contrast to these works, we interpret the syntactic and semantic dependencies extracted from DTs as causal relations that guide the construction of our CGM, integrating syntactic structure directly into the compositional reasoning process.

3

Image Captioning with Learned Rewards

The task of automatically describing images in natural language has historically relied on a two–stage training protocol. Captioning networks are first trained with teacher forcing to minimize cross-entropy loss and then fine-tuned with Self-Critical Sequence Training (SCST) to maximize a hand-crafted metric such as CIDEr. While this pipeline produces high scores on traditional benchmarks, it often encourages bland "average" captions that lack specificity and semantic richness. More recently, attempts to replace hand-crafted rewards with modern multimodal metrics derived from contrastive vision–language models (such as CLIP-Score) have revealed additional challenges. In particular, optimizing these metrics via SCST often leads to overly long or repetitive captions, grammatical degradation, and unstable training dynamics, including reward collapse or uninformative generations. The central challenge is therefore to design training strategies that allow image captioners to produce

This chapter is related to the publications "**N. Moratelli** *et al.*, Fluent and Accurate Image Captioning with a Self-Trained Reward Model, ICPR 2024" [10] and "**N. Moratelli** *et al.*, Revisiting Image Captioning Training Paradigm via Direct CLIP-based Optimization, BMVC 2024" [9]. See the list of Publications on page 149 for more details.

fluent, detailed, and semantically accurate descriptions that align with human preferences, while enabling optimization of modern reward functions without sacrificing stability or linguistic quality. These observations motivate a central question that underlies this chapter: how to design training strategies for image captioning that improve semantic richness and alignment with human preferences without sacrificing fluency or optimization stability.

Scope and role of this chapter. This chapter establishes the conceptual and methodological foundations of this thesis by studying multimodal alignment in its simplest controlled setting: single-image caption generation. Image captioning provides a controlled testbed in which vision–language alignment, reward design, and optimization stability can be analyzed in isolation, without the confounding factors introduced by long contexts, multi-turn interaction, or open-ended instruction following. The analysis developed here introduces two recurring failure modes that will appear throughout the thesis. The first is *reward mismatch*, which arises when proxy objectives fail to capture human-aligned notions of correctness and fluency. The second is *optimization mismatch*, which emerges from the interaction between expressive evaluators and unstable optimization procedures. These concepts will be revisited and expanded in later chapters when scaling to Multimodal Large Language Models.

Research question. This chapter addresses the following research question. *How can reward signals for image captioning be learned and optimized to improve semantic alignment with visual content while preserving linguistic fluency and training stability?* More specifically, the chapter investigates how proxy rewards such as CLIP-based similarity interact with policy gradient optimization, whether reward learning can mitigate instability and degeneration effects, and whether alternative optimization paradigms can avoid reward hacking by construction.

3.1 Approach Overview

While both approaches are grounded in the idea of learning reward signals from multimodal representations, they differ fundamentally in how these signals are

incorporated into the training process: the first retains reinforcement learning while stabilizing it through self-trained rewards, whereas the second questions the necessity of reinforcement learning altogether by embedding reward optimization directly into the captioner.

Self-Cap: Self-Trained Reward Model. The first method introduces a learnable reward model to replace fixed, hand-crafted metrics. A contrastive image–text discriminator is fine-tuned using self-generated negative captions, creating a self-supervised environment in which the model learns to prefer fluent and descriptive sentences over its own erroneous outputs. This discriminator is then used as the reward in the SCST loop, guiding the captioner to produce captions that better reflect human judgement without the adverse effects observed when using naïve CLIP-based rewards. Empirical results show that this self-trained reward leads to captions that are both more linguistically fluent and semantically rich, with faster convergence.

DiCO: Direct Preference Optimization. Building on the observation that optimizing modern metrics via SCST can be unstable, the second method proposes a direct optimization paradigm that eliminates reinforcement learning. DiCO jointly learns and optimizes a reward model distilled from a high-correlation captioning evaluator by solving a weighted classification problem inside the captioner itself. This formulation prevents reward hacking and divergence, maintains fluency, and achieves stable training even when maximizing CLIP-Score or PAC-Score. The resulting captions align more closely with human preferences while retaining competitive performance on traditional metrics.

Together, these contributions advance image captioning by demonstrating how adaptive and directly optimized reward models can yield captions that are both accurate and human-like. The remainder of the chapter provides a detailed technical description and empirical analysis of each method.

Both methods can be interpreted as instances of *learning from preferences* induced by a multimodal evaluator. Self-Cap keeps the classical SCST pipeline but improves the reward so that it penalizes the most common degeneration pat-

terns encountered under CLIP-based optimization (verbosity, repetition, syntactic drift). DiCO instead keeps the evaluator fixed and changes the learning rule, turning scalar reward maximization into likelihood-based learning from relative preferences among candidate captions.

3.2 Self-Cap: Learned Reward for Captioning

3.2.1 From Hand-Crafted to Learnable Rewards

Traditional image captioning systems rely on fixed, hand-crafted reward functions to guide model optimization during reinforcement learning. Metrics such as CIDEr, BLEU, or METEOR were originally designed for evaluation rather than optimization, and encode only partial notions of caption quality, often emphasizing n-gram overlap with reference sentences. While effective at improving benchmark scores, these rewards implicitly bias captioning models toward safe and generic descriptions, limiting semantic richness and expressiveness.

Recent advances in vision–language representation learning have challenged this paradigm. Large-scale contrastive models such as CLIP provide a shared embedding space for images and text that exhibits a substantially higher correlation with human judgment than traditional reference-based metrics. This observation has motivated the use of CLIP-derived scores as alternative reward signals for image captioning, with the promise of producing captions that are more descriptive, semantically aligned, and human-like.

However, directly replacing hand-crafted rewards with CLIP-based metrics exposes a fundamental limitation of existing training pipelines. When optimized through standard reinforcement learning strategies such as Self-Critical Sequence Training, these modern rewards tend to induce unstable learning dynamics, leading to reward hacking, excessive verbosity, and grammatical degradation. This reveals a critical gap: while learnable, multimodal rewards are conceptually superior, they cannot be naively integrated into existing optimization frameworks.

The first contribution presented in this chapter addresses this gap by asking a foundational question: can reward functions for image captioning be learned in a

way that preserves the benefits of modern vision–language models while avoiding the instability and linguistic collapse observed in prior approaches? The method introduced in the following section provides an affirmative answer by learning a reward model that is explicitly shaped by the captioner's own failure modes, enabling stable optimization and improved alignment with human notions of caption quality.

The rest of Sec. 3.2 follows the technical development of the corresponding publication. To improve readability in a thesis setting, we emphasize (i) the source of instability when optimizing CLIP-derived rewards under SCST, (ii) the design choice of learning a discriminator from *self-generated* hard negatives, and (iii) the empirical evidence that the resulting reward improves both semantic descriptiveness and linguistic well-formedness.

3.2.2 SCST Limitations with CLIP Rewards

Despite substantial progress in image captioning, optimizing modern multimodal reward functions within policy-gradient frameworks remains challenging.

In particular, when SCST [236] is combined with CLIP-based evaluators [225], the optimization process often becomes unstable. While CLIP-based metrics such as CLIP-Score [105] and PAC-Score [240] show strong correlation with human judgment, their use as rewards frequently leads to degenerate behaviors, including excessively long captions, word repetitions, grammatical inconsistencies, and violations of proper word order. These failure modes reveal a fundamental mismatch between sequence-level policy-gradient optimization and embedding-based semantic evaluators. More specifically, CLIP-based evaluators operate on a global image–text embedding space and primarily reward semantic compatibility between the generated caption and the visual content. As a consequence, they are relatively insensitive to local linguistic well-formedness: repeating salient concepts, appending semantically related modifiers, or producing loosely structured phrases may preserve or even increase the image–text similarity score. When such a score is used inside SCST, the policy is therefore encouraged to exploit directions in the reward landscape that improve semantic matching

but do not reliably correspond to fluent language. This mismatch is amplified by sequence-level policy-gradient updates, which optimize the final scalar reward without explicitly constraining syntax, word order, or brevity, and can thus lead to verbosity, repetition, and grammatical drift.

To address these issues, we propose a novel approach based on SCST, wherein the image captioning model learns to generate captions by iteratively refining its output through a self-evaluation mechanism. An overview of the proposed training pipeline is illustrated in Fig. 3.1. First, we conduct a fine-tuning process for a caption discriminator using a self-supervised methodology inspired by CLIP. Specifically, alongside the usual positive image-caption pairs, we introduce a set of negative texts generated by the captioning model fine-tuned with the original CLIP-S and PAC-S as reward. The overall goal is to create a self-supervised environment that improves the correlation with human judgment, preserves syntactic accuracy, and allows the model to learn from its errors. As a second step, we integrate this discriminator as the reward used to fine-tune a captioning model, further enhancing its ability to generate high-quality and semantically richer captions.

We assess the effectiveness of the proposed approach by conducting several experiments on the COCO dataset [171], thereby showcasing its robust performance across a range of different backbones. To enhance the comprehensiveness of our analysis and validate the zero-shot capability of our approach, we expand our investigations to include out-of-domain experiments conducted on additional datasets like CC3M [247], nocaps [16], and VizWiz [99], providing insights into its potential applicability in various real-world scenarios.

3.2.3 Method Preliminaries

In this section, we recap the definition of the training protocol typically used in image captioning, of Contrastive Language-Image Pre-training [225], and of learnable image captioning metrics. Also, we introduce the terminology employed in the rest of the section.

Captioning training protocol. Image captioning models are usually

trained with a two-stage training approach. The network f_θ is first pre-trained by encoding an image I_i, described through a sequence of $R = (v_1, v_2, ..., v_R)$ visual features, with a time-wise cross-entropy loss in relation to ground-truth sentences $s_{ij} = (w_1, w_2, ..., w_T)$. In the second stage, the network undergoes fine-tuning through a RL strategy aimed at maximizing the CIDEr score [279] on the training dataset. During the first stage, the model is trained from scratch through a conditioning mechanism, wherein caption generation depends not only on visual features R but also on all previous ground-truth tokens up to time step $t-1$, where w_t is a token belonging to a pre-defined vocabulary. During this phase, f_θ is optimized using a cross-entropy loss (XE) as follows:

$$L_{\mathrm{XE}}(\theta) = -\sum_{t=1}^{T} \log\Big(P(w_t|w_{1:t-1}, R)\Big). \tag{3.1}$$

The network then operates in an autoregressive manner, generating one token per time step. The model f_θ outputs a discrete probability distribution, where the token w_t is chosen as the one with the highest probability, determined by preceding tokens. This selection involves passing the final network embeddings through an MLP followed by a softmax function. In the second training stage, at each time step t tokens are sampled from the probability distribution generated by the model at time step $t-1$. Once the entire caption is generated, the CIDEr score is computed as reward to guide a policy-gradient RL update step [236].

Contrastive Language-Image Pre-Training (CLIP). CLIP [225] represents a state-of-the-art model for the computation of similarities between images and texts. In this context, the computation of matrix similarities and the training of the network through contrastive learning assume a critical role, as it serves as a fundamental step in learning the intrinsic relationships between textual and visual elements, denoted as T and V respectively. The effectiveness of the contrastive method is particularly evident when applied to large-scale datasets. Here, the matrix T is defined as comprising N_t textual instances, each characterized by a D-dimensional embedding. Likewise, the visual representation matrix V has a size of $N_v \times D$. To calculate the similarity matrix S, the cosine similarity func-

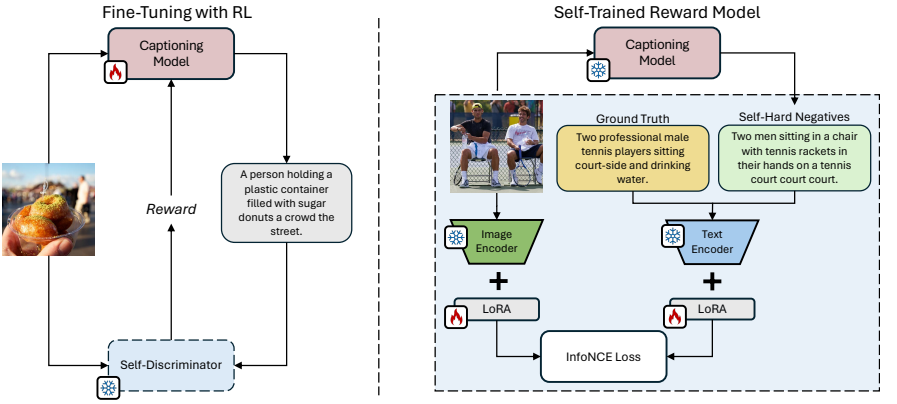


Figure 3.1: Overview of our approach. On the left, the training strategy of the captioner model is shown. The model acts as an agent providing rewards from a discriminator obtained with textual negatives directly derived from the model itself (right).

tion is adopted. For each textual instance T_i and visual instance V_j, the similarity score S_{ij} is computed as follows: $S_{ij} = \text{sim}(T_i, V_j)$, where $\text{sim}(\cdot)$ represents the cosine similarity. This leads to a matrix S, with dimensions $N_t \times N_v$, where each element S_{ij} represents the similarity score between the i-th textual instance and the j-th visual instance.

LEARNABLE CAPTIONING METRICS FROM HUMAN FEEDBACK. A recent yet underexplored research direction involves leveraging a model trained with language-image pre-training as an image captioning metric, given its robust alignment capabilities between visual and textual domains. Following [105], the evaluation score of a caption s'_i can be computed with a cosine similarity $\text{sim}(I_i, s'_i)$ between the visual embedding of the input image and the generated caption. In particular, in [105] a score proportional to the ReLU of the predicted similarity is employed. Additionally, to confine the score within the range of $[0, 1]$ for convenience, the final result is scaled by a multiplicative factor denoted as w:

$$\text{Score}(I_i, s'_i) = w \cdot \text{ReLU}(\text{sim}(I_i, s'_i)). \tag{3.2}$$

One of the most commonly used learnable scores is CLIP-S [197], where the underlying architecture was pre-trained on 400M noisy (image, text) pairs sourced from the internet. Despite demonstrating better alignment with human

judgment compared to traditional captioning metrics (*e.g.* BLEU, METEOR, CIDEr), which rely on reference captions, the use of noisy data during training leads to significant performance degradation when this score is used to directly optimize a captioning model, resulting in disparities between the score and the overall quality of captions. To mitigate this, a recent approach termed PAC-S [240] involves fine-tuning the model on cleaned data, thereby enhancing correlation with human evaluations. Specifically, PAC-S score is trained using a similarity matrix constructed from human-curated captions and machine-generated ones. Nevertheless, although these two metrics appear to yield improved correlation with humans, they tend to favor longer texts that are semantically rich yet grammatically flawed over shorter yet grammatically correct captions.

3.2.4 Self-Trained Reward Model

The SCST approach has proven to be effective in increasing the quality of description with respect to a single XE training stage. However, it also tends to bias the model towards the "average" caption that reflects the most general mode contained in the training set [51]. This comes with some critical disadvantages, including reduced descriptiveness, semantic richness, and discriminative power of the generated captions. What is more, one could argue that employing the CIDEr metric as a reward is an obsolete choice, as it achieves a low correlation with human judgments in comparison with recent alternatives.

Following this intuition, in this section we propose a novel training scheme which is based on a self-supervised reward. In our approach, the classical CIDEr reward is replaced by a learnable language-image discriminator $\mathcal{D}_r$, which takes the form of a language-image model. Following the REINFORCE algorithm, the expected gradient of the reward function can be computed as

$$\nabla_\theta L_{\text{SCST}}(I_i, s_i', \theta) = (\mathcal{D}_r(I_i, s_i') - b)\nabla_\theta \log f_\theta(s_i'), \tag{3.3}$$

where the expected gradient has been approximated using a single Monte-Carlo sample, and b is a baseline employed to reduce the variance of the gradient estimate, which is usually computed as a function of the rewards computed inside

a mini-batch. A classical choice when generating multiple descriptions for the same image through beam search is that of computing b as the average reward of all descriptions generated for I_i, so that $b = \sum_j \mathcal{D}_r(I_i, s'_{ij})/n$.

There are three conceptual advantages in replacing a hand-crafted captioning metric with a learnable discriminator. First, unlike standard text-only metrics, $\mathcal{D}_r$ explicitly conditions on the input image I_i and can therefore evaluate image–text alignment directly. Second, being learnable rather than manually designed, $\mathcal{D}_r$ can be adapted to capture qualitative aspects of captions beyond simple n-gram overlap, including semantic relevance and linguistic well-formedness. Third, once trained, $\mathcal{D}_r$ can be used as a reward signal without requiring ground-truth captions during the fine-tuning stage. The reward depends only on the input image and the generated caption, which makes the approach applicable even in domains where multiple human-written reference captions are unavailable or expensive to collect.

In this regard, a straightforward choice for $\mathcal{D}_r$ would be that of employing a pre-trained CLIP-based model, which also has a large semantic coverage, as explored in [62]. However, when employing learnable rewards, we observed a significant decrease of performance on reference-based metrics, which nonetheless serve as crucial benchmarks for assessing caption quality. Moreover, it is well known that CLIP-based architectures, if not properly fine-tuned, tend to focus heavily on the semantics of the caption, strongly neglecting its grammatical aspect, which is one of the most important aspects of image captioning. From a pragmatic perspective, several works have analyzed the embedding space of CLIP and consistently find that it excels in aligning object categories with images using a bag-of-words approach. This results in robustness against word swapping, rather than mere repetition of identical concepts. Therefore, we introduce a novel fine-tuning methodology grounded in self-supervised learning, which comprises two distinct stages: (i) refinement of CLIP through fine-tuning conditioned on self hard-negatives sourced from the model itself post fine-tuning with CLIP-S and PAC-S; (ii) fine-tuning of the pre-trained model employing our self-discriminator as a reward model.

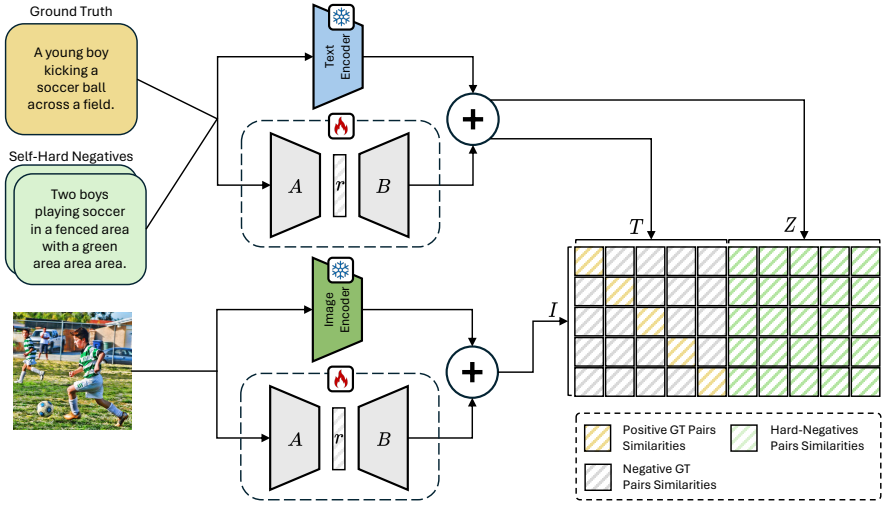


Figure 3.2: Overview of our self-discriminator approach, in which both CLIP encoders are fine-tuned with low-rank adaptation (LoRA) using additional textual negatives.

3.2.5 Fine-Tuning the Self-Discriminator

As mentioned above, the first stage involves refining the CLIP-based discriminator $\mathcal{D}_r$ through generation-aware mining of hard-negatives. In our setting, the term "hard negatives" refers to captions generated by models that were already fine-tuned using CLIP-based rewards. These captions typically maintain strong semantic alignment with the image while exhibiting the degeneration patterns induced by CLIP optimization, such as repetition, verbosity, or syntactic degradation. Because they often receive high scores from the original evaluator despite being linguistically flawed, they constitute particularly informative counterexamples for discriminator training. In practice, for each training image we collect captions produced by two captioners optimized respectively with CLIP-S and PAC-S rewards, and use them as negative textual samples during discriminator fine-tuning. Initially, we employ captioner models trained with CLIP-based rewards to generate these negative instances, which are then exploited to fine-tune CLIP. This process aims to condition CLIP against enforcing alignment styles particularly unsuitable for image captioning. Specifically, through fine-tuning, the goal is to modify the noisy embedding space of CLIP based on the errors

obtained from the captioning model. When CLIP is employed in SCST, it results in a meager grammatical reward, despite its strong semantic robustness. For this purpose, we have generated two distinct types of negatives for each sample (*i.e.* $Z_i = \{Z_i^1, Z_i^2\}$) derived from the fine-tuned captioner using SCST with rewards based on CLIP-S and PAC-S in their reference-based versions, respectively. This choice allows the model to learn not only to better align the embedding space but also to provide self-supervised reward and thus learn from its own mistakes.

To fine-tune the CLIP-based discriminator $\mathcal{D}_r$, we propose a simple modification to the CLIP objective (see Figure 3.2). In particular, given a batch of N images $\mathcal{I} = \{I_1, ..., I_N\}$ and N captions $\mathcal{T} = \{T_1, ..., T_N\}$, we concatenate the textual negatives in such a way as to obtain $\bar{\mathcal{T}} = \{T_1, ..., T_N, Z_1^1, Z_1^2, ..., Z_N^1, Z_N^2\}$. Next, we compute the similarity matrix $S \in \mathbb{R}^{N \times 3N}$. Here, the row-wise and column-wise cross-entropy losses are computed as in CLIP, with the difference that we do not compute the loss for the negative captions column-wise (as there is no matching image for a negative caption). To reduce the number of trainable parameters and save memory, we employ low-rank adaptation (LoRA) [111] during the fine-tuning phase of our CLIP-based discriminator, on all layers of both visual and textual encoders.

3.2.6 Fine-Tuning the Captioner

Once the fine-tuning of the discriminator is completed, it is employed as a reward signal to fine-tune the captioner through SCST. Our fine-tuned discriminator $\mathcal{D}_r$ is capable of providing feedback not only on semantics but it is also sensitive to grammar and syntax. Finally, the reward perceived by our agent is conditioned not only on the generated text but also on the input image and implicitly on the errors that our model would have generated without any correction and modification of the embedding space.

3.2.7 Experimental Setup

We train our model on the COCO dataset [171] which contains around 120k images each associated to five different captions, using the splits defined in [133] where 5,000 images are used for validation, another 5,000 for testing, and the remainder for training. We then evaluate the effectiveness of our solution on the COCO test set and on the validation set of different image captioning datasets, namely nocaps [16], VizWiz [99], and CC3M [247].

To evaluate our results, we employ both standard captioning metrics, such as BLEU [209], METEOR [28], ROUGE [169], CIDEr [279], and SPICE [21], and more recent learning-based scores like CLIP-Score [105] and PAC-Score [240] in their reference-free and reference-based versions. In addition, we employ a novel measure to evaluate the grammatical correctness of the generated captions. Specifically, we define Rep-n with $n = 1, 2, 3, 4$ as the average number of n-grams which are repeated in the generated captions.

3.2.8 Implementation Details

CLIP fine-tuning. Regarding the fine-tuning of CLIP, we use ViT-B/32 as backbone for encoding both images and textual sentences, leveraging the original OpenAI implementation*. As positive examples, we exploit image-caption pairs from the COCO dataset. We use AdamW [185] as optimizer with a learning rate set to $1 \cdot 10^{-4}$ and a batch size of 256. Additionally, to reduce the number of trainable parameters and make fine-tuning more efficient, we employ LoRA [111] with a rank equal to 8.

Architecture. As our captioning model, we employ a standard encoder-decoder Transformer with 3 layers in both encoder and decoder, a hidden size of 512, and 8 attention heads. To encode input images, we use different CLIP-based backbones, such as RN50, ViT-B/32, and ViT-L/14. To implement our model, we employ the Hugging Face library [296].

Training details. We first pre-train the model with the classical cross-entropy loss for sentence generation. Next, we optimize our model using different re-

*https://github.com/openai/CLIP

Backbone	Reward	Supervised ↑							Unsupervised ↑		Grammar ↓			
		B-4	M	R	C	S	RefCLIP-S	RefPAC-S	CLIP-S	PAC-S	Rep-1	Rep-2	Rep-3	Rep-4
RN50	-	32.8	28.1	55.0	109.8	20.3	0.796	0.853	0.743	0.817	1.516	0.108	0.022	0.009
	CIDEr	39.7	29.2	58.3	126.8	21.2	0.797	0.855	0.739	0.817	1.384	0.05	0.008	0.005
	CLIP-S	14.3	24.7	34.9	3.1	21.2	0.765	0.830	0.804	0.837	11.762	5.168	2.809	1.518
	PAC-S	18.5	26.5	42.2	32.2	21.7	0.785	0.849	0.799	**0.860**	5.453	1.588	0.645	0.288
	CLIP-S [62]	6.3	19.7	29.5	11.2	12.3	0.786	0.823	**0.843**	0.837	5.619	1.541	0.466	0.151
	CLIP-S+Gr [62]	16.9	25.9	45.6	71.2	19.6	**0.792**	0.849	0.779	0.839	**1.536**	**0.097**	**0.015**	**0.003**
	Self-Cap	**20.8**	**26.8**	**48.2**	**72.0**	**21.8**	**0.792**	**0.851**	0.780	0.844	2.706	0.495	0.153	0.049
ViT-B/32	-	33.1	28.2	55.4	112.4	20.5	0.804	0.861	0.755	0.830	1.468	0.091	0.017	0.005
	CIDEr	39.4	29.5	58.3	129.0	22.2	0.809	0.866	0.757	0.833	1.360	0.055	0.006	0.001
	CLIP-S	11.4	23.1	31.2	1.1	18.5	0.778	0.830	**0.851**	0.846	11.166	3.566	1.232	0.395
	PAC-S	20.3	27.1	44.1	40.7	22.4	0.796	0.858	0.810	**0.870**	5.078	1.443	0.584	0.260
	Self-Cap	**23.6**	**27.3**	**49.3**	**81.4**	**22.9**	**0.808**	**0.862**	0.800	0.861	**2.626**	**0.483**	**0.156**	**0.063**
ViT-L/14	-	37.3	30.4	58 1	126.6	23.3	0.811	0.868	0.758	0.831	1.402	0.062	0.007	0.002
	CIDEr	43.6	30.8	61.0	143.3	23.2	0.809	0.866	0.750	0.826	0.239	0.498	0.616	0.349
	CLIP-S	10.2	23.0	30.3	1.1	15.3	0.793	0.827	**0.865**	0.834	8.788	2.113	0.716	0.248
	PAC-S	22.3	28.4	46.2	51.1	24.6	0.801	0.861	0.805	**0.862**	4.612	1.199	0.479	0.206
	Self-Cap	**22.6**	**28.4**	**50.2**	**82.7**	**24.7**	**0.809**	**0.864**	0.787	0.853	**2.216**	**0.376**	**0.118**	**0.039**

Table 3.1: Comparison between different reward signals in terms of supervised, unsupervised, and grammar-based metrics. Results are reported on the COCO test set.

wards based on unsupervised and supervised metrics (*i.e.* our Self-Cap strategy, both CLIP-Score [105] and PAC-Score [240], and the CIDEr score). During cross-entropy pre-training, we train our network with the Adam optimizer [136], a batch size of 1,024, and for up to 20,000 steps. During this phase, we linearly warmup for 1,000 steps, then keep a constant learning rate of $2.5 \cdot 10^{-4}$ until 10,000 steps, then sub-linearly decrease until 15,000 steps to 10^{-5} and keep the value constant until the end of the training. For the second stage, we further optimize our model with $1 \cdot 10^{-6}$ as learning rate using a batch size of 32. During caption generation, we employ a beam size equal to 5.

3.2.9 Results

Results on COCO test set. We start by comparing our solution against other CLIP-based rewards (*i.e.* CLIP-S and PAC-S) using different visual backbones to encode input images. Results are reported in Table 3.1 in terms of supervised, unsupervised, and grammar-based metrics. For completeness, we also include the results of the model trained after cross-entropy loss and using a standard CIDEr score as reward. In all experiments, we employ the same Transformer-based architecture with three layers in both the encoder and decoder. Regarding

a comparison with previous works, it is important to note that the only work within the same settings is proposed by Cho *et al*. [62] which however only adopts CLIP RN50 backbone as visual encoder. Specifically, two variants both optimized using CLIP-S are proposed, where the former only employs CLIP-S as reward while the latter combines CLIP-S with a grammar-based reward.

From the results, we can notice that adopting a reward relying on CLIP-based models significantly alters the performance of the model, leading to word repetitions and a lack of logical or grammatical structure within the caption. Indeed, within a few steps, the model appears to hack the metric by finding alternative ways to boost the semantics and consequently the value of the metric itself (*i.e*. CLIP-S or PAC-S), completely disregarding the syntactic structure of the caption. In particular, considering the results of our proposal (*i.e*. Self-Cap) with ViT-B/32 as visual backbone, it can be seen that our reward strategy can significantly improve the results on standard supervised metrics (*e.g*. 81.4 CIDEr points compared to 40.7 and 1.1 achieved with PAC-S and CLIP-S rewards respectively). This demonstrates the effectiveness of Self-Cap in better preserving the coherence of the predicted caption with the image and the ability to generate "human-like" and thus structurally correct captions. As expected, directly optimizing a specific metric leads to the best results on that metric, as showed by the results of the models trained with CLIP-S or PAC-S as reward. Nonetheless, this is not confirmed on the reference-based versions of CLIP-S and PAC-S for which Self-Cap achieves the best performance according to all employed backbones, further confirming a better correlation with human-written captions.

To further clarify the problems associated with unsupervised metrics when used as rewards, we also report the average number of repeated n-grams for each caption (*i.e*. Rep-n with $n = 1, 2, 3, 4$). Notably, Self-Cap significantly reduces the number of repetitions within the generated sentences, decreasing the 1-gram repetitions from 11.166 and 5.078 respectively using CLIP-S and PAC-S to 2.626, always when employing visual features from ViT-B/32. These results are confirmed also considering a larger number of n-grams and across all considered visual backbones, further demonstrating the effectiveness of our training strategy in reducing the grammatical incorrectness of captions generated by captioners

Backbone	Strategy	Unsupervised		Recall			
		CLIP-S	PAC-S	R@1	R@5	R@10	MRR
RN50	XE	0.743	0.817	21.2	44.2	57.6	31.2
	SCST (CIDEr)	0.739	0.817	19.8	43.4	55.7	29.8
	Self-Cap	**0.780**	**0.844**	**37.7**	**67.3**	**78.6**	**50.3**
ViT-B/32	XE	0.755	0.830	24.8	50.8	62.8	35.7
	SCST (CIDEr)	0.757	0.833	25.7	51.7	64.4	36.7
	Self-Cap	**0.800**	**0.861**	**47.1**	**74.6**	**84.9**	**58.9**
ViT-L/14	XE	0.758	0.831	27.7	52.6	64.2	38.5
	SCST (CIDEr)	0.750	0.826	23.9	49.8	61.6	34.9
	Self-Cap	**0.787**	**0.853**	**44.7**	**71.8**	**82.6**	**56.5**

Table 3.2: Descriptiveness analysis of generated captions in terms of unsupervised scores and retrieval-based metrics. Results are reported on the COCO test set.

optimized using standard CLIP-based rewards.

When instead comparing our model with the one proposed in [62] using RN50 visual features, we can notice that the model optimized only with CLIP-S version yields a high value of CLIP-S, while totally degrading the reference-free metrics (*i.e.* 11.2 CIDEr points with respect to 72.0 of Self-Cap) and producing numerous repetitions (*i.e.* 5.619 and 1.541 of Rep-1 and Rep-2 compared to 2.706 and 0.495 of our approach). The scenario is different when considering the second variant, which is optimized with a combination of CLIP-S and a grammar-based reward. Specifically, while Self-Cap still achieves higher results in terms of all supervised metrics, it presents slightly higher values of repetitions. Nevertheless, it is noteworthy that Self-Cap does not exploit any explicit grammatical reward, as it is learned directly within the embedding space of the discriminator itself during the refinement process.

Analysis on the descriptiveness of generated captions. To effectively compare the captions generated by Self-Cap with those generated by a captioning model trained with a standard training paradigm (*i.e.* cross-entropy loss followed by SCST with CIDEr reward), we complement the results shown in Table 3.1 with retrieval-based metrics reported in Table 3.2. Retrieval-based metrics are generally used to measure the discriminative degree of the generated captions, which is usually a viable strategy to estimate their descriptiveness and semantic richness.

In particular, following recent works [141, 43], we measure the quality of gen-

Negatives			Supervised							Unsupervised	
Manual	CLIP-S	PAC-S	B-4	M	R	C	S	RefCLIP-S	RefPAC-S	CLIP-S	PAC-S
✓			19.7	27.4	44.0	41.2	**22.3**	0.799	0.856	**0.812**	**0.865**
	✓		21.6	**27.5**	46.2	57.3	22.3	0.801	0.858	0.808	0.865
		✓	23.1	27.4	48.5	78.9	21.9	0.805	0.861	0.803	0.864
✓		✓	21.3	27.1	47.5	70.0	21.8	0.807	0.862	0.798	0.861
✓	✓	✓	21.0	27.3	46.0	60.4	21.7	**0.808**	**0.862**	0.802	0.862
	✓	✓	**23.6**	27.3	**49.3**	**81.4**	21.9	**0.808**	**0.862**	0.800	0.861

Table 3.3: Ablation study on COCO test set, using different negative textual sentences and CLIP ViT-B/32 as image encoder.

erated captions in distinguishing images in a dataset and compute the percentage of the times the image corresponding to each generated caption is retrieved among the first k retrieved items. This is done by ranking the images in terms of CLIP similarity between visual and textual embeddings, using the CLIP ViT-B/32 model, and computing recall at K with $k = 1, 5, 10$. We also compute the mean reciprocal rank (MRR) for each generated caption: higher MRR scores indicate that captions are more discriminative and therefore usually more detailed. Notably, Self-Cap can significantly increase the results obtained with a standard training paradigm (*i.e.* 24.8 and 25.7 achieved by XE and SCST (CIDEr) in terms of R@1 vs. 47.1 achieved by Self-Cap with ViT-B/32), highlighting a higher degree of descriptiveness in generated captions.

Ablation study on negative examples. As mentioned in Sec. 3.2.3, to compute the reward during the RL-based optimization, we employ a CLIP-based discriminator fine-tuned using a combination of self-generated negative samples obtained by two different captioners, one trained with CLIP-S reward and the other trained with PAC-S reward. In Table 3.3, we evaluate the effectiveness of the chosen negative samples. In particular, we consider negative samples generated by a single captioning model (*i.e.* either trained with CLIP-S or PAC-S) and manually-constructed negative samples, or a combination of them. When generating manual negatives, we consider the failure cases typically produced by a captioner fine-tuned with CLIP-based rewards: (i) premature termination of captions (*e.g.* "a man playing with a cat in"); (ii) redundancy of the final term (*e.g.* "a man with an umbrella in the background background background"); and (iii) duplication of concepts within captions (*e.g.* "a cat in the garden and a cat in

		nocaps						VizWiz						CC3M					
Backbone	**Reward**	B-4	R	C	S	CLIP-S	PAC-S	B-4	R	C	S	CLIP-S	PAC-S	B-4	R	C	S	CLIP-S	PAC-S
RN50	CLIP-S	3.7	23.2	4.6	12.9	0.738	0.799	8.70	29.8	6.7	8.8	0.667	0.78	1.0	13.9	4.3	6.5	0.678	0.78
	PAC-S	4.0	25.3	20.9	**14.1**	**0.741**	**0.850**	9.22	31.6	13.0	10.3	**0.688**	**0.816**	0.8	12.4	5.8	6.5	**0.699**	**0.814**
	Self-Cap	**4.9**	**27.1**	**30.4**	13.9	0.737	0.844	**10.1**	**35.4**	**19.7**	8.1	0.667	0.795	**1.2**	**14.9**	**15.9**	**7.7**	0.686	0.798
ViT-B/32	CLIP-S	4.0	27.1	9.8	13.2	**0.754**	0.810	5.5	23.8	1.3	8.5	**0.737**	0.814	0.8	11.4	0.6	6.0	**0.718**	0.784
	PAC-S	5.2	28.5	35.7	**16.2**	0.750	**0.854**	11.0	34.3	20.1	**9.8**	0.715	**0.837**	1.2	14.1	9.8	7.6	0.698	**0.809**
	Self-Cap	**6.2**	**29.8**	**46.3**	16.0	0.751	**0.854**	**13.0**	**37.8**	**27.0**	9.1	0.702	0.828	**1.3**	**15.2**	**19.4**	**8.5**	0.688	0.803
ViT-L/14	CLIP-S	5.2	28.9	10.2	17.3	**0.750**	0.819	4.1	21.8	1.2	7.0	**0.766**	0.775	0.6	10.2	0.6	4.4	**0.747**	0.765
	PAC-S	5.7	30.0	44.8	**18.1**	0.746	**0.850**	11.2	36.0	26.8	**12.2**	0.701	**0.820**	1.4	15.1	13.2	8.6	0.701	**0.811**
	Self-Cap	**6.9**	**31.3**	**62.8**	**18.1**	0.742	0.839	**11.4**	**37.4**	**28.5**	10.2	0.690	0.809	**1.6**	**16.7**	**21.9**	**9.6**	0.696	0.809

Table 3.4: Out-of-domain performance analysis on nocaps, VizWiz, and CC3M validation sets in terms of supervised and unsupervised metrics.

the garden"). We therefore manually corrupt COCO captions either manually repeating or removing one or more random words, performing a random swap of two words, or substituting one word with a randomly selected word from the entire vocabulary of the COCO dataset.

As it can be seen, the best results are obtained using a combination of negative samples deriving from the combination of CLIP-S and PAC-S, which achieves significantly higher CIDEr values compared to the manually created negatives (*i.e.* 81.4 vs. 41.2) and all other alternatives. Overall, the use of manual negatives does not prove effective also when used in combination with other considered negative samples, leading to performance degradation on all supervised metrics.

Out-of-domain evaluation. To assess the out-of-domain capabilities of our model, we evaluated Self-Cap on three distinct datasets, namely nocaps [16], CC3M [247], and VizWiz [99]. While nocaps is specifically tailored for the novel object captioning task encompassing object classes absent in COCO, CC3M and VizWiz respectively comprises images sourced from the web and captured by visually impaired people. Except for captions from CC3M which are automatically generated, all other datasets are composed of manually-curated textual sentences. Table 3.4 shows the results obtained using three different visual backbones, comparing our approach with models fine-tuned using CLIP-S and PAC-S rewards. Also in this setting, Self-Cap achieves significantly higher results in terms of standard evaluation metrics, demonstrating the effectiveness and generalization capabilities of our approach even in out-of-domain scenarios.

PAC-S: A group of people sitting at a dinner table with a wine glass in the background of a boat setting of wine in the background of a restaurant.
Self-Cap (Ours): A group of men and women sitting around a dinner table at restaurant.

PAC-S: The big ben clock tower towering over the city of London at night at night time with cars driving past it at night.
Self-Cap (Ours): The big ben clock tower towering over the city of London at night.

PAC-S: A woman in a yellow raincoat checking her cell phone against a grey wall with yellow raincoat in the background.
Self-Cap (Ours): A woman in a yellow jacket looking at her cell phone against a brick wall.

PAC-S: Two boys playing soccer in a fenced area with a green soccer ball in the background of a home area setting.
Self-Cap (Ours): A young boy kicking a soccer ball in a field with other players.

PAC-S: A young man sitting on a couch holding a wii remote control in his hand while playing video games in the living room area area.
Self-Cap (Ours): A young man laying in a black leather couch holding a wii remote.

PAC-S: A herd of sheep grazing on a lush green field with a baby sheep grazing in the background of the background area area of a country.
Self-Cap (Ours): Three sheep are grazing in a grassy field and one is looking at the camera.

Figure 3.3: Qualitative results on COCO sample images, comparing Self-Cap with a model trained using PAC-S as reward.

3.2.10 Qualitative Results

To validate the quality of captions generated by our approach, Figure 3.3 shows some qualitative samples from the COCO test set. In this case, we compare captions generated by Self-Cap with those generated by a captioning model trained with PAC-S reward. As it can be seen, Self-Cap can generate more descriptive and complex captions while minimizing repetitions and grammatical errors often encountered when combining SCST with CLIP-based rewards.

3.2.11 Conclusion

We present Self-Cap, a novel fine-tuning method for image captioning which entails a two-phase training procedure. It leverages a discriminator to provide feedback by learning directly from the errors of the captioner. In a setting utilizing a CLIP-based reward, the proposed solution demonstrates state-of-the-art performance in supervised metrics. Additionally, we showcase the out-of-domain capabilities of our approach on three different datasets. Self-Cap generates captions that are not only more complex and semantically richer but also yield superior

grammatical accuracy compared to competitors.

Takeaways

Self-Cap shows that replacing CIDEr with a learnable multimodal reward is beneficial only if the reward model is *adapted* to captioning-specific failure modes. By injecting self-generated hard negatives during discriminator fine-tuning, the reward becomes sensitive not only to image–text semantic alignment but also to linguistic well-formedness, mitigating the verbosity and repetition that arise when optimizing naïve CLIP-based scores. At the same time, the approach still inherits the intrinsic fragility of policy-gradient fine-tuning, which motivates the second part of the chapter.

From reward learning to optimization design. The two approaches presented in this chapter are connected by a controlled progression of design choices. In the first part, the reinforcement learning paradigm is kept fixed while intervening on the reward signal itself, isolating the effect of reward learning on training stability and caption quality. In the second part, the evaluator is kept fixed and the optimization objective is redesigned, isolating the role of the optimization paradigm independently of the reward definition. This progression enables a principled analysis of reward mismatch and optimization mismatch as distinct, yet interacting, sources of failure.

3.3 DiCO: Direct Preference Optimization

3.3.1 Limits of Reinforcement Learning for Captioning

The first part of this chapter argued that modern vision–language evaluators provide a compelling alternative to hand-crafted captioning metrics, but that naively optimizing them within a classical SCST pipeline can severely deteriorate linguistic quality and training stability. Even when the reward function is improved–for instance, by shaping it to account for the captioner's characteristic failure modes—the overall learning dynamics are still governed by a reinforcement-

learning loop whose gradient estimates are high-variance and whose behavior is notoriously sensitive to reward scale, sampling strategy, and baseline design.

This raises a natural follow-up question that shifts the focus from *what* reward to use to *how* to optimize it: if the reward signal is extracted from a strong multimodal evaluator with high human correlation, is reinforcement learning still the right tool for fine-tuning a captioner? In other words, is the instability observed with CLIP-based rewards a side effect of imperfect rewards, or is it rooted in the optimization paradigm itself?

The second contribution presented in this chapter takes the latter perspective seriously. Rather than stabilizing SCST further, it questions whether the captioner should be fine-tuned through policy-gradient reinforcement learning at all. The key idea is to convert the problem of reward maximization into a *direct* learning objective that: (i) preserves the preference signal induced by a CLIP-based evaluator, (ii) prevents reward hacking and collapse by construction, and (iii) enables stable optimization through standard gradient descent. This change of viewpoint produces a training paradigm that is closer in spirit to modern preference-based alignment methods, but tailored to the structural constraints and evaluation practices of image captioning.

3.3.2 Preference Distillation from CLIP

At a high level, the method described in the following section (DiCO) reframes captioner fine-tuning as *preference distillation* from an external evaluator. Instead of treating the evaluator as a scalar reward within a reinforcement-learning loop, it is used to induce pairwise (or listwise) preferences among candidate captions generated for the same image. The captioner is then optimized to increase the likelihood of evaluator-preferred captions relative to less preferred alternatives, while remaining anchored to the pre-trained model distribution.

Importantly, the external evaluator should not be interpreted as a reward model in the RLHF sense. CLIP-based metrics such as CLIP-S or PAC-S are contrastive similarity models that assign compatibility scores to image–caption pairs. In our formulation, these scores are not directly optimized as rewards; in-

stead, they are used to induce relative preferences among candidate captions generated for the same image. The distilled reward signal is therefore implicit in the preference structure derived from the evaluator, rather than given by a separately trained reward model based on human annotations.

This formulation yields two important conceptual advantages for a thesis-level narrative. First, it makes explicit the mechanism through which the evaluator constrains the captioner: not by inflating a potentially hackable scalar, but by enforcing *relative* quality relations across candidates. Second, it separates the roles of (a) the evaluator, which defines the preference structure, and (b) the optimizer, which performs stable likelihood-based learning under a principled regularization term. The remainder of this section provides the technical development of this idea and empirically demonstrates that direct optimization can retain the semantic gains of modern evaluators while eliminating the collapse modes typical of SCST-based CLIP reward maximization.
The next subsection restates the problem and contributions in the standard paper format; in the thesis narrative, it should be read as a more detailed instantiation of the above motivation and as the entry point for the formal derivation of the direct optimization objective.

3.3.3 Problem Formulation

The task of image captioning [257, 280, 306, 133] requires an algorithm to describe a visual input in natural language. As a captioner should ideally match the level of detail and precision desired by the user, over time there has been an increasing interest in developing training strategies for aligning the behavior of a captioner to mimic a desired style and quality level.

Traditionally, the quality of captions has been measured with textual similarity metrics, so captioners have been trained to maximize a non-differentiable metric like CIDEr [279] during a fine-tuning stage based on reinforcement learning, *i.e.* Self-Critical Sequence Training (SCST) [236, 231, 181]. As this strategy requires the availability of multiple reference captions and tends to produce less distinctive descriptions that ignore the fine detailed aspects of an image, re-

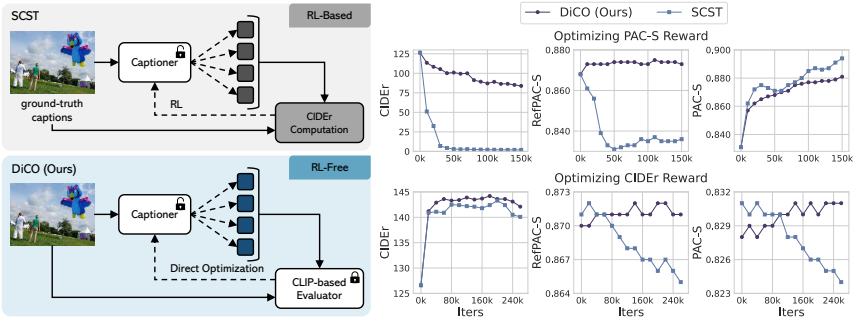


Figure 3.4: Comparison between SCST [236] and our *Direct CLIP-Based Optimization* (DiCO). DiCO distills a reward model from a learnable CLIP-based captioning evaluator, without requiring reinforcement learning and preventing reward hacking and divergence.

cently there have been preliminary attempts to optimize higher-quality image captioning metrics based on embedding spaces that do not require human references [62, 141, 342], like CLIP-Score [105] and PAC-Score [240]. Besides, these metrics also consider the actual multi-modal alignment between the generated text and the visual content of the input image rather than just comparing texts. Most importantly, they also showcase a superior alignment with human judgment, making them ideal candidates for tuning the behavior of captioners towards a higher quality of generation.

Unfortunately, optimizing modern metrics with pre-existing strategies like SCST results in instability and model collapse [62]. We showcase this in Fig. 3.4, where we employ SCST for optimizing either PAC-S or CIDEr (light blue lines). When we try to optimize PAC-S, the fine-tuned captioner hacks the metric and deviates from a fluent and high-quality generation, resulting in a rapid decrease according to all other metrics and leading to repetitions and grammatical errors. To solve these issues, we propose DiCO, a novel training methodology that can align a captioner towards better quality captions by distilling from an external contrastive-based evaluator like CLIP-S or PAC-S, without incurring model collapse and without employing a reinforcement learning objective. Our approach achieves this goal by learning a reward model directly into the captioner and mimicking pairwise quality relations expressed by the external evaluator. This ensures a high degree of alignment with human preferences while avoiding reward hack-

ing. This is visually represented in Fig. 3.4 (dark blue lines): DiCO can optimize both a modern metric like PAC-S and a traditional one like CIDEr by maintaining good scores across all metrics.

We assess the quality of the proposed training methodology by conducting extensive experiments on the COCO dataset [171]. Furthermore, in the supplementary materials, we prove the generalization capabilities of DiCO over other six image captioning benchmarks. Our experimental results demonstrate that DiCO features state-of-the-art quality in the generated captions and improved training stability. This also results in a better performance in terms of modern captioning metrics, while also balancing with competitive performances on traditional handcrafted metrics. On the other hand, when adopted to maximize standard captioning metrics like CIDEr [279], DiCO achieves state-of-the-art results also in this setting. Going beyond automatic image captioning metrics, we confirm the effectiveness of our approach by also employing human-based evaluation.

To sum up, our proposal markedly differs from all fine-tuning strategies in the current image captioning literature. Presently, this field remains closely tied to traditional techniques, employing the classic SCST algorithm with rewards based on ground-truth captions, while overlooking a concerted emphasis on semantic and syntactic richness, as well as alignment with human cognition. Extensive experiments on standard image captioning datasets demonstrate the effectiveness of the proposal.

3.3.4 Method

Preliminaries. SCST [236] optimizes sequence-level rewards through policy-gradient methods and has become a standard paradigm for image captioning. However, when combined with modern multimodal evaluators, this optimization strategy exposes intrinsic stability issues. DiCO is designed to address these limitations by unifying reward estimation and caption generation within a single, differentiable training objective. Background on SCST, RLHF, and CLIP-based evaluation metrics is provided in Appendix B.

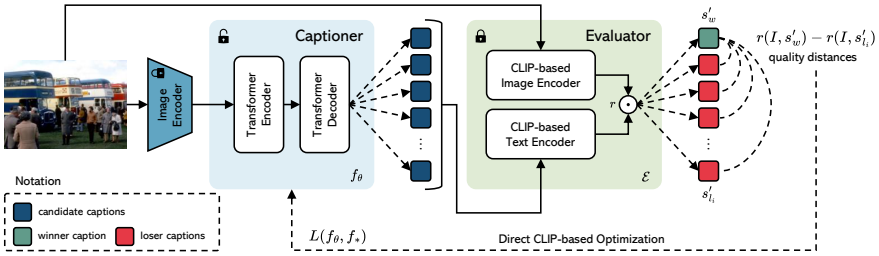


Figure 3.5: Overview of our approach. Given an image and candidate generations, the figure shows the process for captioner fine-tuning by distilling from a CLIP-based evaluator.

MOTIVATION. While adopting significantly different technical choices, there are striking conceptual similarities between the modern RLHF paradigm employed in LLMs and the traditional SCST approach employed in image captioning. Both approaches, indeed, employ reinforcement learning to optimize a reward function, which nevertheless in SCST is a hand-crafted metric, while in RLHF is a learned function from human data. While using RLHF in captioning is impracticable due to the insufficient amount of human preference data to train the reward model, contrastive-based learnable metrics offer a compelling alternative to it, as they show a significant alignment with human judgment [240]. Our proposal solves this issue by distilling a reward model from a pre-trained captioning evaluator, considering pairwise relationships from candidate captions. In addition, it also avoids model collapse which is frequent in SCST (cf. Fig. 3.4).

DERIVING THE FINE-TUNING OBJECTIVE. Following recent works on LLM alignment [208], we aim at fine-tuning a captioner f_θ with a Proximal Policy Optimization (PPO) objective [245], where given an image I and a caption s' sampled from the model, the environment produces a reward $r(s', I)$ through a reward model. In addition, we add a per-token KL penalty with the output of the pre-trained model to mitigate overoptimization of the fine-tuned captioner to the reward model. Our objective is therefore defined as

$$\max_{f_\theta} \mathbb{E}_{I\sim\mathcal{D}, s'\sim f_\theta(\cdot|I)} \left[r(s', I) \right] - \beta \mathbb{D}_{\mathrm{KL}} \left[f_\theta(s'|I) || f_*(s'|I) \right], \tag{3.4}$$

where β controls the deviation from the pre-trained model, termed as f_*. As it can

be seen, the second term has a crucial role, as it prevents the fine-tuned model f_θ from deviating from the distribution on which the reward model is accurate, and prevents the captioner from hacking it, *i.e.* collapsing to high-rewarded answers.

Let f_r denote the optimal policy induced by the reward function r under the KL-regularized fine-tuning objective in Eq. 3.4. Under this objective, it can be shown [227] that the optimal solution to the fine-tuning problem is given by

$$f_r(s'|I) = \frac{1}{Z(I)} f_*(s'|I) \exp\left(\frac{1}{\beta} r(s', I)\right), \tag{3.5}$$

where $Z(I) = \sum_s f_*(s|I) \exp\left(\frac{1}{\beta} r(s, I)\right)$ is the partition function over possible captions. Although the partition function is difficult to estimate, we can still manipulate Eq. 3.5 to express the reward function in terms of the optimal captioner, the pre-trained captioner, and the partition function, as follows:

$$r(s', I) = \beta \log \frac{f_r(s'|I)}{f_*(s'|I)} + \beta \log Z(I). \tag{3.6}$$

Defining a distilled reward model. Since sufficiently large datasets of human preference annotations are not available for image captioning, we do not train a reward model in the standard RLHF sense. Instead, we use a contrastive captioning evaluator $\mathcal{E}$ as a source of preference signals. Given an image and a candidate caption, $\mathcal{E}$ produces a compatibility score measuring image–text alignment. These scores are then converted into relative preferences among captions generated for the same image, which allows us to define a distilled reward structure without requiring explicit human preference labels.

Given a dataset $\mathcal{D}$ comprising images, we let the captioner generate $k+1$ candidate captions (*e.g.* through beam search). Then, for each image, we select the caption with the highest score according to $\mathcal{E}$ and denote it as s'_w (*i.e.* "winner"). The others, instead, are denoted as $\{s'_{l_i}\}_{i=1}^k$ (*i.e.* "losers"). Based on the evaluator, we define a reward model which distinguishes between the winner caption s'_w and the loser captions $\{s'_{l_i}\}_i$. To make the reward model more robust and accurate, we also impose that it can predict the *relative quality distances* between the win-

ner and the loser captions. Formally, we define our reward model through the following objective:

$$\mathcal{L}_R(r) = -\mathbb{E}\left[\log\sigma\left(\sum_{i=1}^{k}\gamma_i(r(I,s'_w) - r(I,s'_{l_i}))\right)\right], \tag{3.7}$$

where the expectation is taken over images in the dataset and winner and loser captions. Also, γ_i weights the relative distance between the winner caption s'_w and the i-th loser caption s'_{l_i} according to the evaluator $\mathcal{E}$. Specifically, it is computed as a normalized probability distribution between score distances, as follows:

$$\gamma_i = \text{softmax}_{s'_{l_1},\dots,s'_{l_k}}\left(\frac{\mathcal{E}(I,s'_w) - \mathcal{E}(I,s'_{l_i})}{\tau}\right), \tag{3.8}$$

where τ is a temperature parameter. Clearly, considering that γ_i sum up to 1, the reward model objective can be rewritten as

$$\mathcal{L}_R(r) = -\mathbb{E}\left[\log\sigma\left(r(I,s'_w) - \sum_{i=1}^{k}\gamma_i r(I,s'_{l_i})\right)\right]. \tag{3.9}$$

Overall loss function. Following [227], we learn the reward model directly into the captioner. Recalling that the Bradley-Terry model depends only on the difference in rewards between two completions and that γ_i are a valid probability distribution, we replace the definition of $r(s', I)$ as a function of the optimal fine-tuned and pre-trained captioner (Eq. 3.6) into the reward model objective (Eq. 3.9), and obtain the final fine-tuning loss of DiCO as

$$L(f_\theta, f_*) = -\mathbb{E}\left[\log\sigma\left(\beta\log\frac{f_\theta(s'_w|I)}{f_*(s'_w|I)} - \beta\sum_{i=1}^{k}\gamma_i\log\frac{f_\theta(s'_{l_i}|I)}{f_*(s'_{l_i}|I)}\right)\right], \tag{3.10}$$

where, noticeably, the unknown partition function $Z(I)$ has been cancelled out. Furthermore, the obtained fine-tuning loss, while being derived from the optimal solution to a PPO objective, can be directly optimized through gradient descent, without the need of employing reinforcement learning techniques.

Comparing DiCO with SCST and RLHF. DiCO fine-tunes a captioning model by aligning it to a contrastive-based evaluator while avoiding overparametrization and model collapse. In comparison with SCST and RLHF, its unique feature is that of *distilling a reward model from an external evaluator* by *learning it directly inside of the captioner*. Further, this is done by avoiding the usage of reinforcement learning at fine-tuning time, which is common to both SCST and RLHF. In comparison with RLHF, the analogy lies in learning from preferences over model-generated candidates, but the source of supervision is different. Standard RLHF relies on human-annotated preference labels to train a reward model that ranks candidate outputs sampled from the policy. In our setting, candidate captions are also sampled from the model, but their relative preferences are provided by an external evaluator rather than by human annotators. The key distinction is therefore not the sampling process itself, but the replacement of human preference annotations with evaluator-derived preference signals.

3.3.5 Experimental Setup

Datasets. All experiments are performed on the COCO dataset [171], using the standard splits defined in [133] with 5,000 images for both test and validation and the rest for training. We report our experimental results on the test set of COCO.

Evaluation metrics. In addition to the standard image captioning metrics like BLEU [209], METEOR [28], and CIDEr [279], we employ two CLIP-based scores, namely CLIP-S [105] and PAC-S [240], in both their reference-free and reference-based versions, using the ViT-B/32 backbone for both metrics (also see Appendix B). Moreover, following recent works [141, 43], we measure the quality of generated captions in distinguishing images in a dataset and compute the percentage of the times the image corresponding to each generated caption is retrieved among the first K retrieved items. This is done by ranking the images in terms of CLIP similarity between visual and textual embeddings, using the CLIP ViT-B/32 model, and computing recall at K with $K = 1, 5, 10$. We also compute the mean reciprocal rank (MRR) for each generated caption: higher

MRR scores indicate that captions are more discriminative and therefore usually more detailed.

Implementation and training details. Our baseline architecture is a standard Transformer model with 3 layers in both encoder and decoder, a hidden dimensionality equal to 512, and 8 attention heads. To extract visual features, we use either RN50, ViT-B/32, or ViT-L/14 pre-trained with a CLIP-based objective [225]. Our code is based on the popular Hugging Face Transformers[†] library. All experiments are performed using the Adam optimizer, initially pre-training all the models with cross-entropy. During fine-tuning, we use a batch size of 16, a fixed learning rate equal to $1 \cdot 10^{-6}$, and a beam size of 5 (*i.e.* the number k of loser captions is set to 4). For efficiency, we train with ZeRo memory offloading and mixed-precision [196]. Unless otherwise specified, the β parameter is set to 0.2 and the ViT-L/14 backbone is used to extract visual features. The temperature parameter τ defined in Eq. 3.8 is set to $1/(3 \cdot 10^2)$. Early stopping is performed according to the reference-based version of the CLIP metrics used as reward.

3.3.6 Comparison with Prior Work

Results on COCO test set. We compare our model trained with the proposed DiCO strategy with other state-of-the-art solutions. We restrain the comparison by only considering captioning models that use CLIP-based visual features to encode images, which have proven to be the most widely employed choice in recent works. In particular, we include some recent standard image captioning models exclusively trained on the COCO dataset with a standard XE+SCST training paradigm like CLIP-VL [248], COS-Net [166], and PMA-Net [29]. Moreover, we compare with LLM-based captioning models focused on zero-shot generation capabilities such as ZeroCap [265], lightweight architectures like ClipCap [197] and SmallCap [230], or large-scale training paradigms such as the recently proposed MiniGPT-v2 [49] and BLIP-2 [158] models. To directly compare our solution with other CLIP-based optimization strategies, we also report the results of our base model trained with SCST using CLIP-S or PAC-

[†] https://huggingface.co/docs/transformers

Model			Reference-based Metrics					Reference-free Metrics					
			B-4	M	C	RefCLIP-S	RefPAC-S	CLIP-S	PAC-S	R@1	R@5	R@10	MRR
Standard Captioners		**Backbone**											
CLIP-VL [248]		RN50×4	40.2	29.7	134.2	0.820	0.862	0.770	0.826	24.0	48.9	61.5	34.8
COS-Net [166]		RN101	42.0	30.6	141.1	0.814	0.870	0.758	0.832	25.8	52.3	64.9	37.1
PMA-Net [29]		ViT-L/14	43.0	30.6	144.1	0.814	0.869	0.755	0.821	-	-	-	-
LLM-based Captioners		**Backbone**											
ZeroCap [265]		ViT-B/32	2.3	10.1	15.1	0.771	0.800	0.810	0.816	-	-	-	-
ClipCap [197]		ViT-B/32	32.3	28.1	108.5	0.809	0.862	0.766	0.833	27.1	53.3	65.5	38.3
SmallCap [230]		ViT-B/32	37.0	27.9	119.7	0.804	0.863	0.748	0.826	23.1	48.2	60.0	33.7
MiniGPT-v2 [49]		ViT-g/14	18.8	24.6	80.4	0.795	0.848	0.752	0.818	27.4	52.0	63.0	37.9
BLIP-2 [158]		ViT-g/14	43.7	32.0	145.8	0.823	0.877	0.767	0.837	31.4	57.5	69.1	42.7
CLIP-based Optimization	**Reward**	**Backbone**											
Cho *et al.* (SCST) [62]	CLIP-S	RN50	6.3	19.7	11.2	0.786	0.823	**0.843**	0.837	43.2	71.9	82.3	55.5
Cho *et al.* (SCST) [62]	CLIP-S+Gr.	RN50	16.9	24.9	71.0	0.792	0.849	0.779	0.839	35.3	63.4	75.2	47.4
SCST	CLIP-S	RN50	14.3	24.7	3.1	0.765	0.830	0.804	0.837	36.9	64.9	75.9	48.7
SCST	PAC-S	RN50	18.5	26.5	32.2	0.785	0.849	0.799	0.860	**44.3**	73.2	83.4	56.5
DiCO (Ours)	CLIP-S	RN50	20.7	25.7	78.9	**0.811**	0.852	0.815	0.842	37.5	66.6	78.1	49.8
DiCO (Ours)	PAC-S	RN50	**22.7**	**27.0**	**79.8**	0.801	**0.865**	0.797	**0.869**	**44.3**	**73.9**	**84.2**	**56.8**
SCST	CLIP-S	ViT-B/32	11.4	23.1	1.1	0.778	0.830	**0.851**	0.846	43.4	70.8	81.1	55.1
SCST	PAC-S	ViT-B/32	20.3	27.1	40.7	0.796	0.858	0.810	0.870	50.0	77.6	87.0	61.8
DiCO (Ours)	CLIP-S	ViT-B/32	22.6	27.0	81.7	**0.817**	0.861	0.825	0.858	46.3	74.0	83.7	58.0
DiCO (Ours)	PAC-S	ViT-B/32	**23.7**	**27.3**	**84.8**	0.810	**0.872**	0.814	**0.882**	**52.9**	**80.8**	**89.5**	**64.8**
SCST	CLIP-S	ViT-L/14	10.2	23.0	1.1	0.793	0.827	**0.865**	0.834	43.3	70.7	80.5	55.0
SCST	PAC-S	ViT-L/14	22.3	**28.4**	51.1	0.801	0.861	0.805	0.862	46.7	74.7	84.8	58.8
DiCO (Ours)	CLIP-S	ViT-L/14	21.4	27.1	82.6	**0.824**	0.863	0.837	0.856	46.5	74.7	84.7	58.4
DiCO (Ours)	PAC-S	ViT-L/14	**25.2**	**28.4**	**89.1**	0.815	**0.875**	0.812	**0.877**	**50.9**	**78.7**	**87.6**	**62.9**

Table 3.5: Comparison with state-of-the-art models on the COCO test set. Bold font indicates the best results among captioners optimized via CLIP-based rewards with comparable backbones, while underlined indicates the overall best results.

S as reward and those of the model proposed in [62] in which a standard Transformer is optimized via SCST with a CLIP-based reward, eventually regularizing the training with an additional score that considers the grammatical correctness of generated sentences.

Results are shown in Table 3.5, including our model trained with DiCO using RN50, ViT-B/32, and ViT-L/14 as visual backbones. Notably, all versions of our model achieve better results than other methods optimized with CLIP-based rewards on almost all evaluation metrics. For example, when comparing our solution optimized via PAC-S reward with SCST and the model proposed in [62], we can notice how not only DiCO improves the performance in terms of standard metrics (*e.g.* 79.8 CIDEr points using RN50 features vs. 32.2 and 71.0 respectively obtained by SCST and [62]), but also obtains increased retrieval-based scores indicating that captions generated by our model are more discriminative and detailed than those generated by competitors. Additionally, DiCO leads to the overall best results on reference-free metrics also surpassing huge models

SCST: A rusted rusty rusty rusty rusty rusty rusty scissors in a garden with plants in the background.
DiCO (Ours): A rusted scissors sticking out of a metal fence with plants in the background.

SCST: Several cows grazing in a field on a mountain range with a mountain in distance under mountain range in background.
DiCO (Ours): A herd of cattle grazing on a lush green hill next to a large mountain range.

SCST: A little boy with headphones sitting in front of a computer with headphones on a desk in corner with headphones on background.
DiCO (Ours): A little boy with headphones sitting at a desk using a computer.

SCST: A person holding up a clear plastic plastic container filled with sugar covered donuts with people in background in background on background.
DiCO (Ours): A hand holding a plastic container filled with sugar covered donuts.

Figure 3.6: Qualitative results on COCO sample images, using PAC-S as reward.

trained on millions or even billions of data like MiniGPT-v2 and BLIP-2, further confirming the effectiveness of our training strategy. To validate the quality of generated captions, we report in Fig. 3.6 some qualitative results on sample images from the COCO dataset. DiCO can generate more descriptive and detailed captions while reducing repetitions and grammatical errors typically generated using SCST.

Human-based and LLM-based evaluations. As a complement of standard metrics, we also perform a user study and an evaluation based on a widely used LLM (*i.e.* GPT-3.5). To perform the user study, we present the users with an image and a pair of captions, one generated by our model and the other generated by a competitor, and ask them to select the preferred caption judging in terms of (1) *helpfulness* (*i.e.* which caption is most helpful to someone who can not see the image), and (2) *correctness* (*i.e.* which caption is more correct both in terms of grammar and consistency with the image). Users could also state that captions are equivalent on one or both evaluation axes. In this case, 0.5 points are given to both captions. To perform LLM-based evaluation, instead, we leverage the Turbo version of GPT-3.5 and directly ask it to evaluate a pair of captions taking into account the corresponding reference sentences. In particular, we ask the LLM to return a score between 0 and 100 for each caption between the two in the prompt, where one is generated by our model and the other by a competitor, and use this score to compute the number of times GPT-3.5 prefers our solution

	Humans		
	Helpfulness	Correctness	**GPT-3.5**
ZeroCap [265]	20.3	27.8	20.8
SmallCap [230]	27.8	36.1	50.0
MiniGPT-v2 [49]	33.3	42.9	44.8
BLIP-2 [158]	49.1	48.6	51.2
Cho *et al*. (CLIP-S Reward) [62]	11.2	17.9	21.5
Cho *et al*. (CLIP-S+Gr. Reward) [62]	41.3	36.7	43.0
SCST (PAC-S Reward)	44.6	40.6	48.5

Table 3.6: Percentage of times a caption from a competitor is preferred against that generated by our proposal, using either human-based evaluations or GPT-3.5. Our solution is preferred more than 50% of the time in almost all cases.

			Semantic			Grammar					
Model	**Reward**	**Backbone**	C	CLIP-S	PAC-S	n_1	n_2	n_3	n_4	RE	%Correct
SCST	CLIP-S	RN50	3.1	0.804	0.837	11.762	5.168	2.809	1.518	6.0	24.7
SCST	PAC-S	RN50	32.2	0.799	0.860	5.453	1.588	0.645	0.288	1.6	71.6
DiCO	CLIP-S	RN50	78.9	**0.815**	0.842	**1.583**	**0.143**	**0.039**	**0.015**	**0.1**	**96.1**
DiCO	PAC-S	RN50	**79.8**	0.797	**0.869**	2.051	0.219	0.055	0.017	**0.1**	94.4
SCST	CLIP-S	ViT-B/32	1.1	**0.851**	0.846	11.166	3.566	1.232	0.395	1.5	3.9
SCST	PAC-S	ViT-B/32	40.7	0.810	0.870	5.078	1.443	0.584	0.260	1.6	73.3
DiCO	CLIP-S	ViT-B/32	81.7	0.825	0.858	**1.938**	0.230	0.071	0.026	0.2	94.8
DiCO	PAC-S	ViT-B/32	**84.8**	0.814	**0.882**	1.939	**0.190**	**0.048**	**0.014**	**0.1**	**96.4**
SCST	CLIP-S	ViT-L/14	1.1	**0.865**	0.834	8.788	2.114	0.716	0.248	1.0	2.6
SCST	PAC-S	ViT-L/14	51.1	0.805	0.862	8.788	4.611	1.200	0.479	1.3	72.6
DiCO	CLIP-S	ViT-L/14	82.6	0.837	0.856	**1.710**	**0.142**	**0.039**	**0.014**	**0.1**	**95.4**
DiCO	PAC-S	ViT-L/14	**89.1**	0.812	**0.877**	2.107	0.218	0.056	0.017	**0.1**	94.3

Table 3.7: Comparison on semantic and grammar metrics. n_i means i-gram repetitions. Results are reported on the COCO test set.

against a competitor and vice versa. If the score is the same for both captions, we give 0.5 points to both of them. To force the model to produce a more accurate evaluation, we also ask it to produce a reason for each score, which has been shown to lead to ratings that correlate well with human judgment [42].

Table 3.6 shows one-to-one comparisons between our model and one of the considered competitors in terms of both human-based and LLM-based evaluations. Results are reported on a subset of 1,000 images randomly taken from the COCO test set. For each comparison, we report the percentage of times a caption generated by one of the competitors is preferred against the one generated by our solution with PAC-S reward. As it can be seen, DiCO is almost always pre-

Model	Reference-based							Reference-free	
	B-4	M	R	C	S	RefCLIP-S	RefPAC-S	CLIP-S	PAC-S
Up-Down [22]	36.3	27.7	56.9	120.1	21.4	0.787	0.848	0.723	0.803
SGAE [312]	39.0	28.4	58.9	129.1	22.2	0.796	0.855	0.734	0.812
AoANet [117]	38.9	29.2	58.8	129.8	22.4	0.797	0.857	0.737	0.815
$\mathcal{M}^2$ Transformer [68]	39.1	29.8	58.3	131.3	22.6	0.793	0.852	0.734	0.813
COS-Net [166]	42.0	30.6	60.6	141.1	**24.6**	0.814	0.870	**0.758**	**0.832**
PMA-Net [29]	43.0	30.6	61.1	144.1	24.0	0.814	0.869	0.755	0.821
Transformer (SCST)	43.6	30.8	61.0	143.3	23.2	0.809	0.866	0.750	0.826
Transformer (DiCO w/ $\beta = 0.05$)	43.2	31.2	61.1	**144.2**	24.4	0.815	0.871	0.756	0.831
Transformer (DiCO w/ $\beta = 0.1$)	**43.7**	31.2	61.2	143.8	24.5	**0.817**	**0.872**	0.757	**0.832**
Transformer (DiCO w/ $\beta = 0.2$)	**43.7**	**31.3**	**61.3**	143.5	24.4	0.816	**0.872**	0.756	0.831

Table 3.8: Comparison with standard captioners using CIDEr-based optimization.

ferred more than 50% of the time, having a comparable number of preferences only when compared with BLIP-2. When instead considering other CLIP-based optimized models, captions generated by our solution are selected in a considerable number of cases from both human evaluators and GPT-3.5 (*e.g.* more than 55-60% compared to [62] with CLIP-S+Grammar reward).

ADDITIONAL RESULTS ON GRAMMAR METRICS. Besides the semantic coherence between images and their descriptions, we compare our method against SCST from the point of view of the fluency and grammatical correctness of the generated captions. To this end, in Table 3.7 we report the average number of n-gram repetitions per caption (*i.e.* n_i with $i = 1, 2, 3, 4$), computed using the `nltk` language toolkit[‡]. We also include the Repetition Evaluation (RE) proposed in [304], which measures the redundancy of n-grams inside a caption (where $n = 4$ as in the original paper). Additionally, we employ the text encoder from [62] and present the percentage of captions classified as grammatically correct (*i.e.* %Correct). Experiments across different backbones confirm that SCST reaches high scores on the optimized metrics, but collapses to predictions that exhibit many repetitions, undermining the fluency of the generated text. DiCO does not suffer from the same problem, keeping low values for repetitions while showcasing state-of-the-art performance over the reward metrics.

CIDER-BASED OPTIMIZATION. Finally, we assess whether our training paradigm can also be applied using the CIDEr score as reward, as usually done in

[‡] https://www.nltk.org/

standard image captioning approaches. Results are reported in Table 3.8, showing the performance of a standard Transformer model fine-tuned with the classical SCST procedure and that of other captioners. For completeness, we also include the results in terms of ROUGE [169] and SPICE [21] which are typically used in standard image captioning evaluation. In this case, we apply DiCO with different β values on the same baseline architecture used in previous experiments (*i.e.* a vanilla Transformer with 3 layers in both encoder and decoder). Interestingly, our solution achieves better results than SCST also in this setting, with 144.2 CIDEr points vs. 143.3 obtained by SCST. As an additional result, DiCO reaches better or comparable performance to that obtained by recent captioning models based on more complex architectures and optimized via SCST, thus proving to be a valid alternative also in a standard CIDEr-based setting.

3.3.7 Conclusion

We presented DiCO, a novel fine-tuning strategy for image captioning which aligns a model to a learnable evaluator with high human correlation. Our approach optimizes a distilled reward model by solving a weighted classification problem directly inside the captioner, which allows it to capture fine-grained differences between multiple candidate captions. Experimental results on several datasets, conducted through automatic metrics and human evaluations, validate the effectiveness of our approach, which can generate more descriptive and detailed captions than competitors. At the same time, it achieves state-of-the-art results when trained to optimize traditional reference-based metrics.

3.4 Discussion

Reward modeling versus optimization paradigms

Taken together, Self-Cap and DiCO isolate two complementary sources of failure when moving beyond hand-crafted metrics: (i) *reward mismatch*, where a generic vision–language evaluator provides a preference signal that underpenalizes syntactic degeneration, and (ii) *optimization mismatch*, where high-

variance policy gradients amplify reward biases and encourage reward hacking. Self-Cap primarily addresses (i) by reshaping the reward model using self-generated negatives, while DiCO primarily addresses (ii) by replacing scalar reward maximization with preference-based likelihood learning.

Practical Implications

When a practitioner must remain within an SCST-style pipeline (*e.g.*, for compatibility with existing codebases), Self-Cap provides a robust way to benefit from multimodal rewards without sacrificing fluency. When training stability and controllability are paramount, DiCO offers a more principled default by eliminating reinforcement learning and anchoring learning to a pre-trained distribution.

Limitations and Future Work

Both methods inherit the limitations of the underlying evaluator: if the evaluator systematically prefers a specific style (*e.g.*, longer captions), that bias can still influence optimization. A natural next step is to combine evaluator distillation with explicit style/length control or multi-objective preference aggregation, so that semantic richness can be increased without drifting toward verbosity.

3.5 Chapter Summary

This chapter examined image captioning as a controlled setting for studying multimodal alignment, highlighting the interplay between reward design and optimization stability. The next chapter extends these insights to Multimodal Large Language Models, where alignment must hold across long contexts and open-ended multimodal reasoning.

4

Multimodal Large Language Models: Design Space and Backbone Analysis

Multimodal Large Language Models (MLLMs) lie at the intersection of vision and language. They extend large language models with a visual encoder and a lightweight interface that maps visual representations into the language model's token space, enabling instruction following and reasoning over mixed image–text inputs. This capability supports a wide range of vision–language tasks, from visual understanding and grounding to domain-specific applications such as medical report generation.

Despite rapid progress, the MLLM landscape remains difficult to interpret and compare. Most systems follow the same high-level template—vision encoder, language backbone, and vision-to-language adapter—yet differ substantially in how these components are connected, trained, and evaluated. The community

This chapter is related to the publications "D. Caffagni *et al*., The Revolution of Multimodal Large Language Models: A Survey, Findings of ACL 2024" [2] and "F. Cocchi *et al*., LLaVA-MORE: A Comparative Study of LLMs and Visual Backbones for Enhanced Visual Instruction Tuning, ICCV Workshops 2025" [4]. See the list of Publications on page 149 for more details.

has largely converged on a narrow set of backbones (often LLaMA-derived LLMs and CLIP-style encoders), while alternative vision models and training protocols are explored unevenly. As a result, it is still unclear which architectural and training choices are essential for robust multimodal alignment, instruction following, and generalization, and which differences are simply artifacts of data and evaluation.

From captioning to MLLMs. This chapter continues the trajectory developed in the earlier part of this thesis. Previous chapters focused on image captioning, where relatively compact models were optimized to describe visual content with increasing fluency and fidelity, and where we studied improved training objectives and self-trained rewards to better align vision and language at the level of single images. The emergence of large language models with strong instruction-following and reasoning capabilities shifted the focus from generating a single sentence to supporting open-ended, interactive multimodal behavior. In this sense, MLLMs can be seen as a generalization of captioning systems: alignment and grounding must hold across long, compositional exchanges rather than a single output. Accordingly, this chapter adopts an analytical and empirical perspective aimed at clarifying the MLLM design space and identifying which architectural and training choices genuinely matter for multimodal reasoning. This shift exposes new challenges in multimodal alignment that cannot be studied in isolation at the level of single-image caption generation. This chapter thus marks the transition from closed-book multimodal models, whose reasoning relies solely on internal representations, to the open-book multimodal setting explored in the next chapter, where external knowledge retrieval becomes a first-class component.

Scope and focus of this chapter. This chapter focuses on MLLMs as systems for vision–language understanding, reasoning, and instruction following. Extensions to visual synthesis (image generation and editing) are mentioned only to contextualize architectural trends, and are not treated as a core objective of this thesis.

Research question. This chapter addresses the following question: *Which architectural and training choices are responsible for effective multimodal reason-*

ing? In particular, we study how visual backbones, language model scale, and training data interact to shape multimodal alignment, instruction following, and generalization. The chapter combines (i) a survey that consolidates the current design space and open challenges and (ii) a controlled empirical study that directly evaluates the impact of backbone and training choices, enabling principled and actionable insights.

4.1 Approach Overview

This chapter follows a two-part structure that mirrors the conceptual progression of the thesis: first, we analyze and systematize the existing design space of MLLMs; second, we isolate key architectural factors through a controlled empirical study. To provide structure to this rapidly evolving field, the chapter is organized around two complementary contributions that combine synthesis with controlled empirical analysis.

Comprehensive survey. We first consolidate the recent literature on vision-based MLLMs, focusing on architectural design (visual encoders, adapter modules, and language backbones), alignment strategies, instruction tuning, datasets, and evaluation practices. Rather than cataloguing models in isolation, the survey highlights recurring design patterns, sources of inconsistency across works, and open challenges that hinder fair comparison and reproducibility.

Backbone analysis. Building on the survey insights, we then present a controlled empirical study that systematically pairs multiple language models with diverse visual backbones under matched training protocols and datasets. This isolates the impact of key design choices—vision backbone family, model scale, image resolution, and pre-training data—on multimodal alignment, instruction following, and downstream generalization. The resulting findings provide practical guidance for designing MLLMs and motivate more unified evaluation frameworks.

4.2 Design Space of Multimodal Large Language Models

This section synthesizes the current design space of Multimodal Large Language Models, based on our prior survey work, and reorganizes existing approaches around architectural, training, and evaluation choices. This section serves as the first contribution of the chapter. Rather than proposing a new model, it systematically analyzes the current MLLM design space to identify dominant architectural patterns, sources of confounding, and open challenges that hinder principled comparison.

4.2.1 Problem Introduction

We begin by introducing the general architectural and historical context that led to the emergence of Multimodal Large Language Models. The introduction of the attention operation and the Transformer architecture [278] has enabled the creation of models capable of handling various modalities on an increasingly large scale. This advancement is largely attributed to the versatility of the operator and the adaptability of the architecture. Initially, this breakthrough was leveraged for language-specific models [76, 34] but quickly extended to support diverse modalities [160, 187] and facilitate their integration within unified embedding spaces [225].

The surge in sophisticated Large Language Models (LLMs), particularly their capacity for in-context learning, has encouraged researchers to broaden the scope of these models to encompass multiple modalities, both as inputs and outputs. This expansion has led to the development of cutting-edge models such as GPT-4V [15] and Gemini [23], showcasing state-of-the-art performance.

The development of MLLMs entails merging single-modality architectures for vision and language, establishing effective connections between them through vision-to-language adapters, and devising innovative training approaches. These methodologies are crucial for ensuring modality alignment and the ability to follow instructions accurately.

In a context marked by the rapid release of new models, our goal is to offer an exhaustive overview of the MLLM landscape, with a focus on models exploiting the visual modality. This overview serves as both an update on the current state and a source of inspiration for future developments. We identify three core aspects that define these models: their architecture, training methodologies, and the tasks they are designed to perform. We begin by detailing the prevalent choices for vision encoders and adapter modules that equip LLMs with cross-modal capabilities. Following this, we delve into the training processes and data utilized. We then explore the range of tasks addressed by MLLMs. The review concludes with a discussion of the persisting challenges in the field and the promising directions for future research.

The motivation behind this survey stems from an emerging scientific interest in the field of MLLMs, as evidenced by the constant increase in published works. In comparison with existing surveys on the topic [321, 298, 116], our work exhibits substantial differences. Notably, it addresses several critical areas that were overlooked in prior works, including visual grounding; topics such as generation and editing are discussed in the original survey but are outside the scope of this thesis chapter. Furthermore, our survey details the main components utilized by each discussed MLLM, such as the visual encoders and the specific LLM employed. Additionally, our analysis offers a comparative perspective on the performance and hardware requirements of the discussed papers, incorporating both quantitative results and detailed information on benchmarks. Through this comprehensive approach, our survey aims to fill the existing gaps and provide a more nuanced understanding of the current landscape in the field.

4.2.2 Preliminaries

Large Language Models. Brown et al. [34] discovered that in-context learning, *i.e.*, prepending the prompt with a few examples demonstrating the desired output of an LLM [63, 107, 263], improves its performance, especially over unseen tasks. Generalization can be further enhanced by providing the LLM with the natural language description of the desired task for each training sample. This

technique, called instruction-tuning [65, 291, 290, 126], turns out to be critical for aligning the behavior of an LLM with that of humans and currently empowers the most advanced LLMs, eventually boosted via reinforcement learning from human feedback (RLHF) [208, 15, 58, 26].

PEFT. When a pre-trained LLM needs to be adapted to a specific domain or application, parameter-efficient fine-tuning (PEFT) schemes represent an important alternative to train the entire LLM, since these strategies only introduce a few new parameters. Among these, prompt-tuning [101, 150, 164, 182] learns a small set of vectors to be fed to the model as soft prompts before the input text. By contrast, LoRA [111] constrains the number of new weights by learning low-rank matrices. This technique is orthogonal to quantization methods such as QLoRA [75], which further decreases the memory footprint of the LLM compared to the usual half-precision weights.

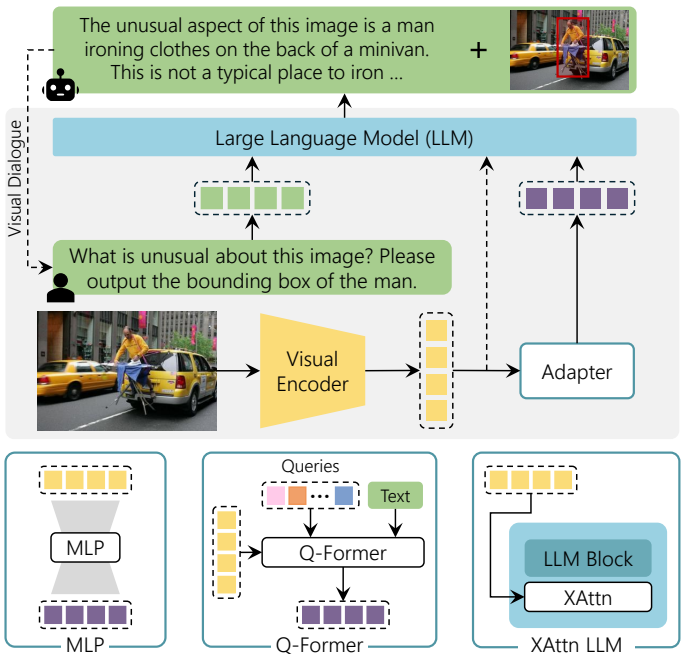


Figure 4.1: Representative architecture of multimodal large language models (*e.g.*, LLaVA-style architectures), composed of a visual encoder, a language model, and an adapter module that connects visual inputs to the textual space.

Towards Multimodal LLMs. The development of MLLMs follows a similar path to that of LLMs, with Flamingo [18] being the first to explore in-context learning at scale in the vision-language field. Then, visual instruction-tuning [179] quickly became the most prominent training paradigm also in the multimodal domain, as well as the use of PEFT techniques to fine-tune the LLM. Any MLLM contains at least three components (Fig. 4.1): an LLM backbone serving as an interface with the user, one (or more) visual encoders, and one or more vision-to-language adapter modules. Popular choices for the LLM backbone often fall into the LLaMA family [271, 272], given that their weights are freely accessible, they have been trained on public data solely, and they boast different sizes to accommodate various use cases. In addition, their derivative versions are popular as well, such as Alpaca [262] and Vicuna [61]. The former fine-tunes LLaMA on instructions written using GPT-3, while the latter exploits user-shared conversations with ChatGPT [204]. Alternatives are OPT [339], Magneto [284], MPT [199], and the instruction-tuned [65] or multilingual [309] flavors of T5 [228], an encoder-decoder language model pre-trained for multiple tasks.

Pre-Training of Model Components. The main components of MLLMs are the visual encoder and the language model. The visual encoder is designed to provide LLMs with visual information and the most used ones are CLIP-based architectures [225, 297] whose pre-training objective is the alignment between CLIP embeddings, obtained thanks to a contrastive loss that aligns the correct image-text pairs. An exception is the EVA-CLIP models family [85], which exploits a MAE pre-training strategy [103] to reconstruct the masked-out image-text aligned visual features, conditioned on visible image patches. On the other hand, LLMs primarily rely on the widely employed Transformer model, although the Mamba architecture [97] has also emerged in recent times. This proposes to make a State-Space Model (SSM) time-dependent, effectively creating a selective SSM with favorable properties: (i) inference costs and memory requirements that scale linearly with the sequence length, and (ii) efficient parallel training thanks to a smart GPU implementation of the algorithm. Similar to Transformers, Mamba models for language modeling are pre-trained using the next

token prediction task. Very recent studies propose MLLMs featuring Mamba as the language backbone [223, 345].

A summary of the MLLMs covered in this survey is reported in Table 4.1, indicating for each model the LLM on which it is based, the visual encoder, the adapter used to connect visual and language components, whether the MLLM is trained with visual instruction tuning or not, and a short list of the main tasks and capabilities.

4.2.3 Visual Encoders

In MLLMs, one of the key components is a visual encoder, which is specifically designed to provide the LLM with extracted visual features. It is common to employ a frozen pre-trained visual encoder while training only a learnable interface that connects visual features with the underlying LLM. While this is usually done using low-resolution images with fixed aspect ratios, some attempts [307, 168] involve adapting pre-trained visual backbones to handle images of different resolutions and aspect ratios.

The most often employed visual encoders are based on pre-trained Vision Transformer (ViT) models with a CLIP-based objective to exploit the inherent alignment of CLIP embeddings. Popular choices are the ViT-L model from CLIP [225], the ViT-H backbone from OpenCLIP [297], and the ViT-g version from EVA-CLIP [85].

As shown in [158], a stronger image encoder leads to better performance. Building on this insight, Lin et al. [174] and Gao et al. [90] propose an ensemble of frozen visual backbones to capture robust visual representations and different levels of information granularity. Concurrently, PaLI models [56, 53], noticing an imbalance between language and visual parameters, propose scaling the visual backbone respectively to a 4- and 22-billion parameter ViT.

Model	LLM	Visual Encoder	V2L Adapter	VInstr. Tuning	Main Tasks & Capabilities
BLIP-2 [158]	FlanT5-XXL-11B★	EVA ViT-g	Q-Former	✗	Visual Dialogue, VQA, Captioning, Retrieval
FROMAGe [140]	OPT-6.7B★	CLIP ViT-L	Linear	✗	Visual Dialogue, Captioning, Retrieval
Kosmos-1 [118]	Magneto-1.3B◊	CLIP ViT-L	Q-Former*	✗	Visual Dialogue, VQA, Captioning
LLaMA-Adapter V2 [89]	LLaMA-7B▲	CLIP ViT-L	Linear	✗	VQA, Captioning
OpenFlamingo [25]	MPT-7B★	CLIP ViT-L	XAttn LLM	✗	VQA, Captioning
Flamingo [18]	Chinchilla-70B★	NFNet-F6	XAttn LLM	✗	Visual Dialogue, VQA, Captioning
PaLI [56]	mT5-XXL-13B♦	ViT-e	XAttn LLM	✗	Multilingual, VQA, Captioning, Retrieval
PaLI-X [53]	UL2-32B♦	ViT-22B	XAttn LLM	✗	Multilingual, VQA, Captioning
LLaVA [179]	Vicuna-13B♦	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning
MiniGPT-4 [352]	Vicuna-13B★	EVA ViT-g	Linear	✓	VQA, Captioning
mPLUG-Owl [316]	LLaMA-7B▲	CLIP ViT-L	Q-Former*	✓	Visual Dialogue, VQA
InstructBLIP [70]	Vicuna-13B★	EVA ViT-g	Q-Former	✓	Visual Dialogue, VQA, Captioning
MultiModal-GPT [92]	LLaMA-7B▲	CLIP ViT-L	XAttn LLM	✓	Visual Dialogue, VQA, Captioning
LaVIN [190]	LLaMA-13B▲	CLIP ViT-L	MLP	✓	Visual Dialogue, VQA, Captioning
Otter [155]	LLaMA-7B★	CLIP ViT-L	XAttn LLM	✓	VQA, Captioning
Kosmos-2 [213]	Magneto-1.3B◊	CLIP ViT-L	Q-Former*	✓	Visual Dialogue, VQA, Captioning, Referring, REC
Shikra [50]	Vicuna-13B♦	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning, Referring, REC, GroundCap
Clever Flamingo [46]	LLaMA-7B▲	CLIP ViT-L	XAttn LLM	✓	Visual Dialogue, VQA, Captioning
SVIT [344]	Vicuna-13B♦	CLIP ViT-L	MLP	✓	Visual Dialogue, VQA, Captioning
BLIVA [113]	Vicuna-7B★	EVA ViT-g	Q-Former+Linear	✓	Visual Dialogue, VQA, Captioning
IDEFICS [146]	LLaMA-65B★	OpenCLIP ViT-H	XAttn LLM	✓	Visual Dialogue, VQA, Captioning
Qwen-VL [27]	Qwen-7B♦	OpenCLIP ViT-bigG	Q-Former*	✓	Visual Dialogue, Multilingual, VQA, Captioning, REC
StableLLaVA [167]	Vicuna-13B♦	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning
Ferret [323]	Vicuna-13B♦	CLIP ViT-L	Linear	✓	Visual Dialogue, Captioning, Referring, REC, GroundCap
LLaVA-1.5 [177]	Vicuna-13B♦	CLIP ViT-L	MLP	✓	Visual Dialogue, VQA, Captioning
MiniGPT-v2 [49]	LLaMA-2-7B▲	EVA ViT-g	Linear	✓	Visual Dialogue, VQA, Captioning, Referring, REC, GroundCap
Pink [308]	Vicuna-7B▲	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning, Referring, REC, GroundCap
CogVLM [288]	Vicuna-7B♦	EVA ViT-E	MLP	✓	Visual Dialogue, VQA, Captioning, REC
DRESS [58]	Vicuna-13B▲	EVA ViT-g	Linear	✓	Visual Dialogue, VQA, Captioning
LION [47]	FlanT5-XXL-11B★	EVA ViT-g	Q-Former+MLP	✓	Visual Dialogue, VQA, Captioning, REC
mPLUG-Owl2 [318]	LLaMA-2-7B♦	CLIP ViT-L	Q-Former*	✓	Visual Dialogue, VQA, Captioning
SPHINX [174]	LLaMA-2-13B♦	Mixture	Linear	✓	Visual Dialogue, VQA, Captioning, Referring, REC, GroundCap
Honeybee [41]	Vicuna-13B♦	CLIP ViT-L	ResNet blocks	✓	Visual Dialogue, VQA, Captioning
VILA [170]	LLaMA-2-13B♦	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning
SPHINX-X [90]	Mixtral-8×7B♦	Mixture	Linear	✓	Visual Dialogue, Multilingual, VQA, Captioning, Referring, REC

Table 4.1: Summary of generalist MLLMs for vision-to-language tasks. For each model, we indicate the LLM used in its best configuration as shown in the original paper (◊: LLM training from scratch; ♦: LLM fine-tuning; ▲: LLM fine-tuning with PEFT techniques; ★: frozen LLM). The ✻ marker indicates variants to the reported vision-to-language adapter, while gray color indicates models not publicly available.

The utilization of such large and powerful models is made feasible by the common practice of maintaining the visual encoder frozen during training, as observed in [158, 118, 89, 50]. However, this design also has limitations. First, a frozen visual encoder cannot adapt its representations to the downstream alignment requirements of the language model, which may weaken multimodal integration. Second, visual encoders typically output dense patch-level representations, which preserve fine-grained spatial detail but also translate into long token sequences when fed into the language model. This increases computational cost and places a heavier burden on the adapter to compress and reorganize local vi-

sual evidence into a form that is useful for language reasoning. To mitigate these issues, other approaches [316, 318] employ a two-stage training paradigm. In the first stage, they incorporate a trainable visual backbone while keeping the pre-trained LLM frozen. According to their findings, enabling the vision encoder to be trainable enhances performance on tasks such as visual question answering or visual description. However, it may lead to performance degradation in other tasks, indicating a degree of forgetting and degradation of the general visual representation.

4.2.4 Vision-to-Language Adapters

The simultaneous presence of inputs from different modalities emphasizes the need to incorporate a module capable of delineating latent correspondences within these unimodal domains. These modules, termed as "adapters", are intended to facilitate interoperability between the visual and textual domains. A spectrum of different adapters are used in common MLLMs, ranging from elementary architectures such as linear layers or MLP to advanced methodologies such as Transformer-based solutions, exemplified by the Q-Former model, and conditioned cross-attention layers added to the LLM.

Linear and MLP Projections. The most straightforward approach for projecting visual inputs into textual embeddings involves learning a linear mapping, which translates visual features to the same dimensionality as the textual counterpart. Some approaches like LLaMA-Adapter [89] and FROMAGe [140] only employ a single linear layer to perform the multimodal connection, while LLaVA-1.5 [177] adopts a two-layer MLP, showing improved multimodal capabilities. Despite its widespread adoption in early MLLMs, the use of linear projections has proven highly effective even in recent methods with a more advanced understanding of the visual input [50, 170, 288, 323, 344]. It is, therefore, a simple yet effective technique for aligning visual features with textual counterparts. A different approach [41] proposes to replace linear layers with convolutional ones, demonstrating moderate improvements.

Q-Former. It is a Transformer-based model proposed in BLIP-2 [158] and then

used in several other approaches [47, 70, 113]. It is characterized by its adaptable architecture, which consists of two Transformer blocks that share mutual self-attention layers, facilitating the alignment process between visual and textual representations. It involves a set of learnable queries that interact within the self-attention layers and interface with visual features via a cross-attention mechanism. Textual and visual elements communicate via shared self-attention within the modules.

Drawing inspiration from the Q-Former, various modified versions have been introduced. In this regard, mPLUG-Owl models [316, 318] simplify the Q-Former architecture and propose a visual abstractor component that operates by condensing visual information into distinct learnable tokens to derive more semantically enriched visual representations. On the same line, Qwen-VL [27] compresses visual features using a single-layer cross-attention module with learnable queries also incorporating 2D positional encodings.

Additional Cross-Attention Layers. This approach has been proposed in Flamingo [18] with the integration of dense cross-attention blocks among the existing pre-trained layers of the LLM. The newly added layers are often combined with a zero-initialized tanh-gating mechanism to ensure that, upon initialization, the conditioned model acts as its original version. The use of additional cross-attention layers imposes the need to train them from scratch, increasing the number of trainable parameters compared to other alternatives. To reduce computational complexity, this strategy is usually paired with a Perceiver-based component [124] that reduces the number of visual tokens before they are fed to the LLM. Since its introduction, several models [25, 46, 146, 155] employ this technique to connect the visual modality with the underlying LLM, demonstrating enhanced training stability and improved performance.

4.2.5 Training and Alignment

Starting from a pre-trained LLM, the training of an MLLM undergoes a single-stage or a two-stage process. In both cases, a standard cross-entropy loss is utilized for predicting the next token, serving as an auto-regressive objective.

Single-Stage Training. This possibility is explored by LLaMA-Adapter [89] which introduces additional trainable parameters to encapsulate the visual knowledge and manage text-only instruction learning at the same time. To achieve this, the model undergoes joint training using image-text pairs and instructions, operating on separate parameters. Concurrently, the model proposed in [140] adapts the final loss function by incorporating two contrastive losses for image-text retrieval. During the training, only three linear layers are updated. On a different line, Kosmos-1 [118] considers a frozen visual backbone and trains the language model of 1.3B parameters from scratch.

Flamingo [18] and its open source variants [25, 146], instead, train the cross-attention layers and the Perceiver-based component to connect the visual features with the frozen LLM blocks. Additionally, Otter [155] extends Flamingo's training to increment its in-context capabilities. Given the amount of training data currently available, approaches like SPHINX-X [90] opt to perform a single all-in-one training stage in which to update all model components, possibly also using text-only data to preserve the conversational capabilities of the LLM.

Two-Stage Training. In the first of the two training stages, the objective is to align the image features with the text embedding space. After this stage, the outputs tend to be fragmented and not coherent. Therefore, a second step is done to improve multimodal conversational capabilities. LLaVA [179, 177] is among the first to introduce a visual instruction-following training scheme, which is performed as a second training stage updating the parameters of both the multimodal adapter and LLM. During the first stage, instead, only the multimodal adapter is trainable. By contrast, MiniGPT-4 [352] is notable for training solely the linear layer responsible for multimodal alignment across both stages. In the second stage, it uses filtered data, collected and refined through the model itself after the first stage.

Another approach, as demonstrated in InstructBLIP [70], involves the freezing of the visual encoder and LLM. In both training stages, only the Q-Former and the connection module are trainable. In contrast to previous approaches where the visual backbone remains frozen, mPLUG-Owl [316, 318] updates it in the initial stage, facilitating the capture of both low- and high-level visual infor-

mation. Also, in the second stage text-only and multimodal data are used jointly to increase alignment. By contrast, Shikra [50] updates all weights in both stages, with the only exception of the visual backbone which is kept frozen.

Training Data. During the first (or single) training stage, the datasets predominantly consist of large-scale, publicly available, and uncurated data. For instance, the Conceptual Captions 3M (CC3M) dataset [247] is composed of 3M images paired with textual captions specifically designed for image captioning systems. Unlike the widely-used and curated MS-COCO [171] dataset, which serves similar purposes, images and captions in CC3M are gathered from the web, showcasing a broader spectrum of styles and content. Similarly, the LAION family [244, 243] represents an extended collection of non-curated image-text pairs sourced from web pages, providing a rich resource for pre-training multimodal language models. Additionally, the COYO-700M [38] dataset stands out as a significant resource, containing 747M image-text pairs. Notably, each alt-text in COYO-700M is linked to an image within HTML documents. Furthermore, DataComp [88] presents an extensive pool of 12.8B filtered image-text pairs sourced from common crawl.

It is important to highlight the distinction between datasets used in the initial phase of training, which typically comprise large-scale, uncurated data, and those selected for refinement in subsequent stages. While the former emphasizes diversity and scale, the latter focuses on specificity and task relevance, facilitating a more tailored approach to model optimization. Especially in single-training stage approaches, certain methods [18, 146] also leverage interleaved datasets, which contain images interleaved with text coming from the web, aiming to augment the dataset size for large models [107]. Images within these datasets can be positioned at the beginning or in the middle of a sentence, allowing models to support arbitrarily interleaved sequences of images and text as input, thereby enhancing flexibility in input formats by blending textual and visual elements. Among these datasets, the most used are WebLI [56], composed of 10B images and image-text pairs, and MMC4 [355], an extension of the text-only C4 [228] dataset composed of 365M documents and 156B tokens relatives to different concepts, and OBELICS [146], an open and curated collection of interleaved image-text web

documents, containing 141M documents, 115B text tokens, and 353M images.

In the context of visual instruction tuning, which constitutes the second training stage for MLLMs, the available amount of data is limited. This limitation is mainly due to the creation process which is time-consuming and less well-defined. In this phase, different datasets are used to improve performances on a series of downstream tasks. Among them, LLaVA-Instruct [179] is a collection of GPT-4 generated multimodal instruction-following data. It comprises 158k unique language-image descriptions, spanning various types of tasks including 58k conversations, 23k detailed descriptions, and 77k complex reasoning. Similarly, LRV-Instruction [176] initially consisted of 400k visual instructions generated by GPT-4, and more recently, it has been updated with an additional set of 300k visual instructions. To enhance robustness in instruction tuning, LRV-Instruct also includes negative instructions organized across three semantic levels, showing that instruct-tuned MLLMs on this dataset suffer less from hallucination compared to the original versions. Moreover, LLaVAR [343] considers publicly available OCR tools to collect results on 422k text-rich images from the LAION dataset. The pipeline first collects 422k noisy text-rich images and then extracts the text through OCR models. With the help of GPT-4, the results and captions are used to create 16k conversations, also including specific questions to create complex instructions which can be helpful to boost performance on new tasks.

Standard MLLMs can tackle visual understanding tasks such as VQA, captioning, and multi-turn conversation. More recently, research has moved toward finer-grained capabilities that require region-level understanding and grounding.

4.2.6 Visual Grounding

The visual grounding capabilities of an MLLM correspond to the ability to carry a dialogue with the user that includes the positioning of the content, also referred to as a referential dialogue [50]. In particular, You et al. [323] introduce *referring* as the ability to understand the content of an input region and can be evaluated on tasks such as region captioning and referring expression generation. Con-

Model	LLM	Visual Encoder	Supporting Model	Main Tasks & Capabilities
ContextDET [328]	OPT-6.7B★	Swin-B	-	Visual Dialogue, VQA, Captioning, Detection, REC, RES
DetGPT [215]	Vicuna-13B★	EVA ViT-g	G-DINO★	Visual Dialogue, Detection
VisionLLM [287]	Alpaca-7B▲	Intern-H	Deformable-DETR▲	VQA, Captioning, Detection, Segmentation, REC
BuboGPT [348]	Vicuna-7B★	EVA ViT-g	RAM, G-DINO, SAM★	Visual Dialogue, Audio Understanding, Captioning, GroundCap
ChatSpot [346]	Vicuna-7B♦	CLIP ViT-L	-	Visual Dialogue, VQA, Captioning, Referring
GPT4RoI [340]	LLaVA-7B♦	OpenCLIP ViT-H	-	Visual Dialogue, VQA, Captioning, Referring
ASM [289]	Husky-7B▲	EVA ViT-g	-	VQA, Captioning, Referring
LISA [145]	LLaVA-13B▲	CLIP ViT-L	SAM♦	Visual Dialogue, Captioning, RES
PVIT [45]	LLaVA-7B♦	CLIP ViT-L	RegionCLIP★	Visual Dialogue, VQA, Captioning, Referring
GLaMM [233]	Vicuna-7B▲	OpenCLIP ViT-H	SAM♦	Visual Dialogue, Captioning, Referring, REC, RES, GroundCap
Griffon [331]	LLaVA-13B♦	CLIP ViT-L	-	REC, Detection, Phrase Grounding
LLaFS [354]	CodeLLaMA-7B▲	CLIP RN50	-	Few-Shot Segmentation
NExT-Chat [332]	Vicuna-7B♦	CLIP ViT-L	SAM♦	Visual Dialogue, Captioning, Referring, REC, RES, GroundCap
GSVA [302]	LLaVA-13B▲	CLIP ViT-L	SAM♦	VQA, Segmentation, REC, RES
Lenna [294]	LLaVA-7B▲	CLIP ViT-L	G-DINO♦	VQA, Captioning, REC
LISA++ [311]	LLaVA-13B▲	CLIP ViT-L	SAM♦	Visual Dialogue, Captioning, RES
LLaVA-G [335]	Vicuna-13B♦	CLIP ViT-L	OpenSeeD, S-SAM♦	Visual Dialogue, REC, RES, Grounding
PixelLLM [305]	FlanT5-XL-3B▲	EVA ViT-L	SAM★	Referring, REC, RES, GroundCap
PixelLM [234]	LLaVA-7B▲	CLIP ViT-L	-	Visual Dialogue, RES
VistaLLM [220]	Vicuna-13B♦	EVA	-	Visual Dialogue, VQA, Referring, REC, RES, GroundCap
ChatterBox [267]	LLaVA-13B▲	CLIP ViT-L	iTPN-B★, DINO♦	Visual Dialogue, Referring, REC, GroundCap
GELLA [222]	LLaVA-13B▲	CLIP ViT-L	Mask2Former♦	Segmentation, RES, GroundCap
PaLI-3 [55]	UL2-3B♦	SigLIP ViT-g	VQ-VAE♦	VQA, Captioning, Retrieval, RES

Table 4.2: Summary of MLLMs with components specifically designed for visual grounding and region-level understanding. For each model, we indicate the LLM used in its best configuration, in some cases initialized with the weights of a pre-trained MLLM, and any supporting models used to perform the task (♦: fine-tuning; ▲: fine-tuning with PEFT techniques; ★: frozen). Gray color indicates models not publicly available.

versely, *grounding* is associated with localizing regions of a given textual description and corresponds to tasks such as referring expression comprehension (REC), referring expression segmentation (RES), phrase grounding, and grounded captioning. Two main components are required to equip MLLMs with these capabilities: a region-to-sequence method to process input regions and a sequence-to-region method to ground nouns and phrases. A summary of the MLLMs with visual grounding capabilities is reported in Table 4.2.

Region-as-Text. The most common way to output regions is to directly insert them into generated text as a series of coordinates, represented as numbers or as special tokens dedicated to location bins. Shikra [50], Kosmos-2 [213], MiniGPT-v2 [49], Ferret [323], CogVLM [288], SPHINX [174], Qwen-VL [27], and Griffon [331] convert bounding boxes into text by indicating two points. VisionLLM [287], VistaLLM [220], LLaFS [354], and ChatSpot [346] allow the MLLM to handle polygons by representing them as a series of points.

Embedding-as-Region. Another solution is to read input regions through region encoders and provide the output regions as embeddings extracted from

the last layer of the MLLM to a decoder. For input regions, GLaMM [233], GPT4RoI [340], ASM [289] and ChatterBox [267] leverage features of the image encoder to perform ROI align on the bounding box, whereas PVIT [45] exploits RegionCLIP [350]. PixelLLM [305] and LLaVA-G [335] use the prompt encoder of SAM [138] and Semantic-SAM [156] respectively. For output regions, LISA [145], GLaMM, GSVA [302], NeXt-Chat [332], and LISA++ [311] send the embedding corresponding to special tokens to the mask decoder of SAM, LLaVA-G to OpenSeeD [334], Lenna [294] to Grounding-DINO [180], and PixelLM [234] to a custom lightweight pixel decoder.

By contrast, ContextDET [328] introduces a decoder that receives the latent embedding of the noun with learnable queries, performs a cross-attention with image features, and then uses a segmentation head. ChatterBox [267] combines features from the iTPN-B encoder [268] and the MLLM and provides them to the DINO detector [333]. GELLA [222] presents a fusion module in Mask2Former [60] to propose masks based on multi-modal image features and an association module to assign latent embeddings to them. PaLI-3 [55] converts embeddings into segmentation masks through a VQ-VAE [276] decoder.

Text-to-Grounding. Other approaches are based on open-vocabulary models that accept textual categories as input. DetGPT [215] generates a list of categories for Grounding-DINO. BuboGPT [348] leverages a combination of RAM, Grounding-DINO, and SAM and matches tags with nouns in the output sequence.

4.2.7 Other Modalities and Applications

Beyond images, extensions to video, audio, and domain-specific settings have been explored. While these directions highlight the versatility of MLLMs, they introduce additional challenges (temporal modeling, modality fusion, domain constraints) that are orthogonal to the controlled architectural comparisons pursued in this chapter. We therefore omit a detailed discussion and focus on vision–language instruction following.

4.2.8 Summary and Open Challenges

We have provided a comprehensive overview of the recent evolution of MLLMs, first focusing on how to equip LLMs with multimodal capabilities and then exploring the main tasks addressed by these models. Based on the analysis presented, we outline key open challenges and promising research directions to further empower MLLMs. Overall, the survey reveals a recurring tension: architectural diversity has grown faster than our ability to compare approaches under controlled and reproducible settings. This observation motivates the controlled empirical study in the next section, which isolates the contribution of individual design choices under a fixed training and evaluation protocol.

Multimodal Retrieval-Augmented Generation. While retrieval-augmented generation (RAG) is a consolidated technique in LLMs [152, 24], its application in MLLMs is still under-explored. We believe that the emergence of VQA datasets that require external retrieved knowledge [57, 194] may enable the development of MLLMs with RAG capabilities [115, 3].

Correction of Hallucinations. Several studies [175, 352] show that MLLMs tend to exhibit high hallucination rates, especially when generating longer captions. While some solutions are emerging to mitigate this problem [175, 283, 299, 322, 130], understanding and correcting the underlying causes of hallucinations remains an important open challenge that is worth addressing to allow the application of these models in more critical contexts (*e.g.*, medicine) and guarantee their accuracy and trustworthiness.

Prevent Harmful and Biased Generation. Ensuring the safety and fairness of large-scale models is of fundamental interest in the community. Recent works show that models trained on web-crawled data are prone to generate inappropriate and biased content. Although recent efforts are being made to reduce this phenomenon in text-to-image generative models [242, 86, 219], further exploration is needed to prevent the same behavior in MLLMs [216].

Reduce Computational Load. MLLMs are highly computationally demanding. Effective strategies [64] are needed to reduce computational requirements and enable more accessible development of MLLMs. Possible directions

entail reducing training requirements both in terms of model scale and data quantity and optimizing the inference stage.

4.3 From Architectural Diversity to Controlled Comparison

The previous section highlights a core limitation of the current MLLM literature: architectural innovations are often evaluated under different datasets, training regimes, and benchmark settings, making it difficult to disentangle the contribution of individual components. In particular, comparisons across models are frequently confounded by changes in data scale, instruction tuning recipes, or evaluation protocols. The following section addresses this issue through a controlled empirical study. By fixing the overall training pipeline and evaluation suite, we isolate how language backbone choice, visual encoder family, input resolution, and pre-training data affect multimodal alignment, instruction following, and downstream generalization. This transforms the descriptive insights of the survey into concrete and actionable design evidence. In this sense, the following study plays the same role as controlled ablations in classical representation learning: it transforms a fragmented literature into testable design hypotheses.

4.4 A Comparative Study of LLMs and Visual Backbones for Enhanced Visual Instruction Tuning

The goal of this study is not to introduce yet another MLLM architecture, but to disentangle the effect of individual design choices under a fixed and reproducible training protocol.

The emergence of LLMs with remarkable expressive capabilities has revolutionized the way diverse language-related tasks are approached [61, 264, 271, 13]. This advancement has inspired the Computer Vision and Multimedia communities to move beyond traditional text-only paradigms and adopt multiple modali-

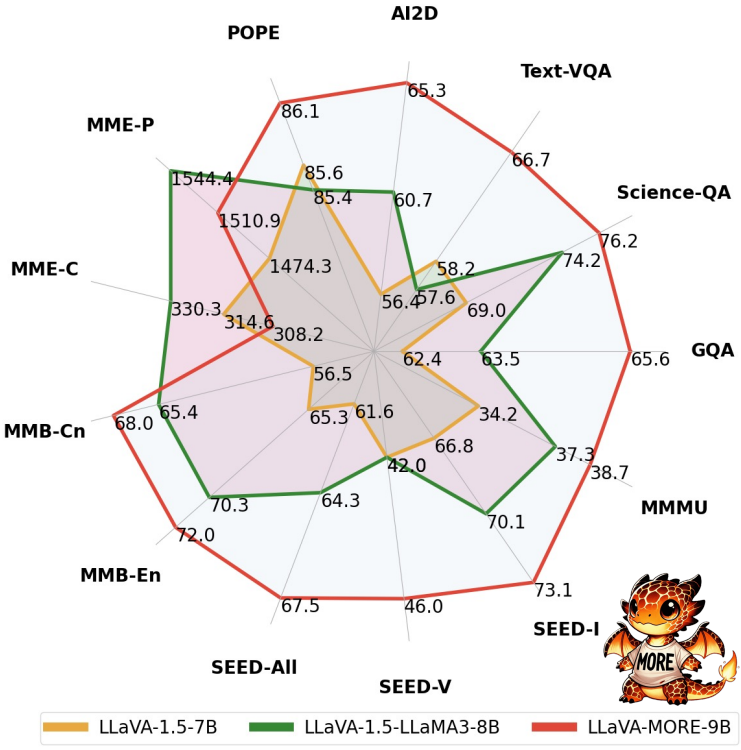


Figure 4.2: Performance comparison of the best version of LLaVA-MORE with other LLaVA variants across different benchmarks for multimodal reasoning and visual question answering.

ties, including vision, audio, and beyond. Consequently, this shift has led to the emergence of MLLMs [2], which establish sophisticated relationships between concepts across different embedding spaces, enabling richer multimodal understanding.

Current MLLMs [18, 179, 177, 17, 59, 127] typically integrate a language model with a visual backbone using specialized adapters that bridge the gap between modalities. While these systems demonstrate impressive performance, the field has converged around a somewhat narrow technical approach, with most implementations leveraging LLaMA-derived LLMs and LLaVA-based training protocols. Additionally, visual encoders based on contrastive training such as CLIP [225] and its derivatives [273, 85, 330] have become the default choice for extracting visual features. These encoders are specifically trained to gener-

ate embeddings that seamlessly integrate with language models, further driving their widespread adoption. While contrastive learning has been highly effective in aligning images and text within a shared space, other vision models [39, 206] learn robust visual features in a purely self-supervised scheme, without relying on weak supervision from text. These models showcase intriguing emerging properties, and yet their application to MLLMs is relatively understudied.

To address this, our work conducts a comprehensive empirical study that systematically pairs diverse LLMs –ranging from efficient models [13] to significantly larger architectures [271, 264] – with various visual backbones [225, 330, 206, 273]. By exploring different architectural combinations, we aim to uncover the strengths and limitations of various vision-language integration strategies, shedding light on overlooked design choices and their impact on multimodal learning. Fig. 4.2 illustrates a comparison of our best-performing model (*i.e.*, LLaVA-MORE-9B, based on SigLIP2 as visual encoder and Gemma-2-9B as LLM) against LLaVA-based competitors (*i.e.*, LLaVA-1.5-7B [177] and LLaVA-1.5-LLaMA3-8B [232], both using a CLIP-based visual encoder and Vicuna-7B and LLaMA3-8B as LLM, respectively).

To ensure experimental consistency, we follow the established LLaVA [177] methodology, pre-training models on natural language description tasks before applying visual instruction fine-tuning to improve cross-domain generalization and human alignment. However, recent works not only introduce architectural enhancements but also incorporate specially curated datasets [73, 27, 184], making fair comparisons across models challenging. To isolate the role of data, in this work we compare LLaVA models pre-trained on different datasets against the same evaluation protocol.

To further explore the impact of training data, we conduct an additional study examining how different types of pre-training data influence multimodal alignment, reasoning ability, and generalization. Specifically, we compare models trained on web-scale image-text corpora, such as LAION [243], against those leveraging task-specific datasets with richer structural supervision [177]. Additionally, we explore the impact of using Recap-DataComp-1B [162], a recaptioned variant designed to enhance image-text alignment. Our results demon-

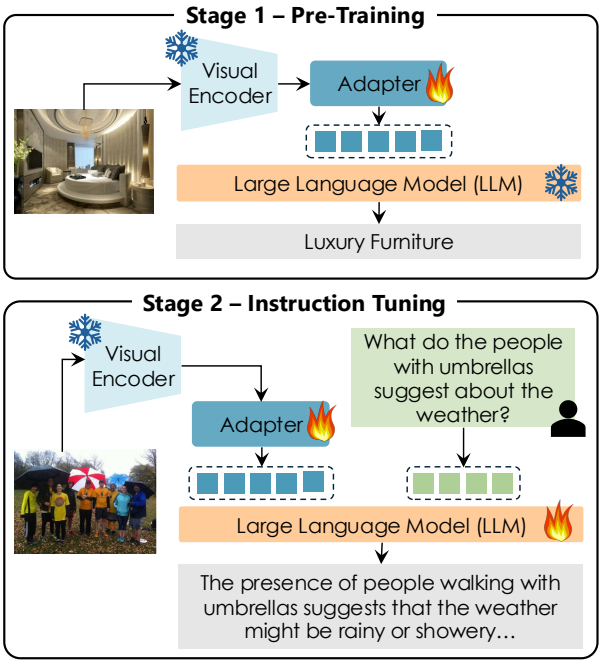


Figure 4.3: Overview of the LLaVA-MORE architecture. Our approach follows the standard LLaVA framework with a two-stage training process. The first stage aligns the visual features to the underlying LLM, ensuring effective cross-modal representation. The second stage enhances the MLLM conversational capabilities through visual instruction tuning. Following this paradigm, we systematically compare different LLM and visual encoder choices to evaluate their impact on various multimodal tasks.

strate that dataset selection significantly affects recognition capabilities and cross-domain transfer performance.

To summarize, this work provides key insights into cross-modal representation learning and offers a practical guide for developing more efficient and effective MLLMs, while challenging conventional assumptions about visual and textual backbones and dataset requirements for optimizing pre-training strategies.

4.4.1 Overall Architecture

Following the LLaVA architecture, a typical MLLM consists of three fundamental components: a large language model backbone for user interaction and generating text, one or more visual encoders to process visual input and extract fea-

tures, and at least one vision-to-language adapter to bridge the gap between visual and textual modalities [2]. The visual encoder provides essential global visual features to the LLM, with CLIP-based architectures [225, 297] being the most commonly used. In this setup, the visual encoder is typically pre-trained on another task and kept frozen for the entire training of the MLLM.

In this thesis, we propose LLaVA-MORE, a new family of models that extend the standard LLaVA architecture by combining the visual encoder with various LLMs, ranging from small- to medium-scale models. As small-scale models, we utilize Gemma-2 2B [264] and Phi-4-Mini [13] (with 3.8B parameters), both designed for strong reasoning scalability, effectively challenging larger models. For medium-scale models, we select the variant of Gemma-2 [264] with 9B parameters alongside two recent architectures: the original LLaMA-3.1 [96] LLM with 8B parameters and its distilled DeepSeek-R1 version [98] (*i.e.*, DeepSeek-R1-Distill-LLaMA-8B). For each LLM category, we assess the impact of varying the visual backbone, with the aim to identify the optimal configuration. Specifically, our study compares the standard CLIP ViT-L/14 encoder employed in the LLaVA-1.5 model [177] with two variants of DINOv2 differentiated by the presence or the absence of visual register tokens [206, 71], known for their strong, semantically rich visual features, as well as SigLIP [330] and its more advanced successor, SigLIP2 [273].

To train our models, we follow the two-stage training paradigm commonly used in the literature. However, our approach stands out by applying a consistent training and evaluation strategy across all models, ensuring fairness in comparisons. In the first stage, only the vision-to-language adapter is optimized to align the image features with the text embedding space. In the second stage, visual instruction-following training is conducted to enhance multimodal conversational capabilities. During this phase, the parameters of both the multimodal adapter and the LLM are updated. An overview of the overall architecture and the two-stage training process is shown in Fig. 4.3.

4.4.2 Implementation Details

Architectural Components. The LLaVA-MORE family follows LLaVA architecture, employing CLIP ViT-L/14@336 as the visual backbone while varying the underlying language model. We categorize the selected LLMs into two groups based on scale: small-scale models, including Gemma-2-2B and Phi-4-3.8B, and medium-scale models, such as LLaMA-3.1-8B, DeepSeek-R1-Distill-LLaMA-8B, and Gemma-2-9B. To further investigate the impact of the visual backbone, we conduct experiments on the best-performing models, replacing CLIP [225] with alternative vision encoders, including DINOv2 [206], SigLIP [330], and SigLIP2 [273]. Additionally, we examine the effects of applying Scaling on Scales (S^2) [249] to both the CLIP and SigLIP2 architectures. Finally, motivated by the LLaVA-1.5 framework and insights from [52, 54], which highlight the advantages of using MLPs over linear projections in self-supervised learning, we adopt a two-layer MLP as the vision-language adapter to enhance multimodal fusion.

Training Details. Considering the LLaVA framework, we adopt a two-stage training strategy. In the first stage, only the weights of the adapter are updated using image-caption pairs as training data. Specifically, the caption style follows the alt-text structure as in web-scale multimodal datasets. The models are trained for one full epoch, covering a total of 558k samples from a combination of different sources (*i.e.*, LAION [243], CC3M [44], and SBU [207]). In the second step, we fine-tune the model on high-quality visual instruction-following data to improve its multimodal reasoning capabilities. This sequential training approach has been shown to significantly improve performance on downstream tasks [179]. Notably, the next token prediction is used as the loss function in both training phases. LLaVA-MORE models are trained with the same set of hyperparameters as LLaVA-1.5 to ensure consistency and comparability. In particular, we employ a global batch size of 256 during pre-training and 128 during the visual instruction tuning phase. All experiments are run in a multi-GPU, multi-node configuration with a total of 16 A100 64GB NVIDIA GPUs.

	VQA Benchmarks				MLLM Benchmarks								
	GQA	Science-QA	TextVQA	AI2D	POPE	MME-P	MME-C	MMB-Cn	MMB-En	SEED-All	SEED-V	SEED-I	MMMU
Small-Scale MLLMs													
LLaVA-Phi-2.7B [356]	-	68.4	-	-	85.0	1335.1	-	-	59.8	-	-	-	-
LLaVA-MORE (Ours)													
Gemma-2-2B [264]	**62.4**	71.1	**54.4**	57.1	**86.0**	**1401.1**	**337.8**	**65.8**	53.3	62.2	41.9	67.6	33.4
Phi-4-3.8B [13]	62.1	**71.3**	54.0	**61.1**	85.9	1372.2	281.1	64.2	**69.2**	**63.5**	**42.3**	**69.1**	**38.8**
Medium-Scale MLLMs													
LLaVA-1.5-7B [177]	62.4	69.0	58.2	56.4	85.6	1474.3	314.6	56.5	65.3	61.6	42.0	66.8	34.2
LLaVA-1.5-LLaMA3-8B [232]	63.5	74.2	57.6	60.7	85.4	**1544.4**	330.3	65.4	70.3	64.3	42.0	**70.1**	37.3
LLaVA-MORE (Ours)													
LLaMA-3.1-8B [96]	63.6	**76.3**	58.4	61.8	85.1	1531.5	**353.3**	**68.2**	**72.4**	64.1	42.4	69.8	**39.4**
DeepSeek-R1-Distill-LLaMA-8B [98]	63.0	74.5	56.3	58.8	85.1	1495.1	295.0	66.8	61.3	63.5	43.5	68.6	38.1
Gemma-2-9B [264]	**64.2**	75.4	**60.7**	**64.8**	**86.8**	1522.5	307.5	65.9	71.9	**64.5**	**44.1**	**69.9**	37.9

Table 4.3: Performance analysis when changing the underlying LLMs. Results are reported considering both small- and medium-scale LLMs, comparing LLaVA-MORE with existing LLaVA-based variants. All models employ the CLIP ViT-L/14@336 as the visual backbone.

4.4.3 Evaluation Benchmarks

We evaluate the LLaVA-MORE family on a diverse set of task-oriented and instruction-following benchmarks.

VQA Benchmarks. These primarily assess the model ability to answer questions based on visual inputs. For the VQA setting, we consider the following datasets:

- **GQA** [120] is based on Visual Genome scene graph annotations [142] and comprises 113k images and 22M questions focusing on scene understanding and compositionality. Results are reported on the test split, which represents 10% of the total image set.

- **ScienceQA** [189] evaluates models with challenging multimodal multiple-choice questions across three diverse domain subjects (*i.e.*, natural science, language science, and social science), 26 topics, 127 categories, and 379 skills. Each question is annotated with explanations linked to relevant lectures from elementary and high school science curricula. We report results on the test set which includes 4,241 examples.

- **TextVQA** [252] is a dataset built on Open Images [144] designed to evaluate the OCR capabilities of vision-and-language models. In our experiments, we employ the validation set which comprises 5,734 samples.

- **AI2D** [134] is a comprehensive collection of diagrams specifically designed for

educational and research purposes. It consists of over 5,000 grade school science diagrams, which cover a wide range of topics and are accompanied by more than 15,000 diverse and richly formulated multiple-choice questions and answers.

MLLM Benchmarks. These evaluate broader multimodal language understanding and reasoning capabilities. Specifically, we consider the following benchmarks:

- **POPE** [165] is a benchmark for evaluating object hallucinations in MLLM generations. It includes several subsets, namely random, popular, and adversarial, which are generated using various sampling methodologies. The dataset consists of 8,910 binary classification queries, enabling thorough analysis of the object hallucination phenomena in MLLMs.

- **MME** [87] is designed to measure proficiency in various communication modalities through 14 diverse tasks. These tasks assess comprehension and manipulation in areas such as quantification, spatial reasoning, and color identification. Overall, it contains 2,374 samples.

- **MMBench (MMB)** [183] includes approximately 3,000 multiple-choice questions, distributed across a collection of 20 distinct domains and understanding capability. Questions are designed to assess the effectiveness of MLLMs across various task paradigms. These capabilities are systematically structured into a hierarchical taxonomy, encompassing broad categories like perception and reasoning, as well as finer-grained skills such as object localization and attribute inference.

- **SEED-Bench (SEED)** [154] evaluates MLLMs across 12 dimensions, including scene understanding, OCR, and action recognition. The dataset comprises 19k multiple-choice questions curated by human annotators.

- **MMMU** [326] is a challenging benchmark for multimodal models, focusing on massive multi-discipline tasks demanding college-level subject knowledge.

It consists of 900 validation samples drawn from university textbooks or online courses spanning six main disciplines. Questions may include multiple images interleaved with text. Evaluation includes exact and word matching for multiple-choice and open-ended questions, and models are tested in zero or few-shot settings.

4.4.4 Effect of the Language Backbone

LLaVA-MORE extends the widely recognized LLaVA architecture by incorporating small- and medium-scale LLMs. As shown in Table 4.3, our experiments first investigate the relationship between model size and performance across various benchmarks. Among the small-scale LLMs, we compare the LLaVA version based on Phi-2.7B [356] with our versions based on Phi-4-3.8B and Gemma-2-2B. Overall, both versions of LLaVA-MORE consistently outperform the existing baseline across multiple benchmarks. Notably, Phi-4-3.8B achieves the highest scores on most benchmarks, particularly excelling in the MMMU performance, where it surpasses Gemma-2-2B by 5.4%. Similar improvements are also evident in the SEED dataset, where Phi-4-3.8B achieved significantly higher performance than other small-scale LLMs. These results underscore Phi-4-3.8B's superior reasoning and generalization capabilities.

Conversely, among medium-scale models, LLaVA-1.5-7B and LLaVA-1.5-LLaMA3-8B provide stronger baselines, with LLaVA-1.5-LLaMA3-8B achieving the highest score in MME-P (1544.4). However, our LLaVA-MORE models consistently surpass these baselines, demonstrating superior performance across both VQA and MLLM benchmarks. In particular, Gemma-2-9B emerges as the best-performing model, especially excelling in VQA benchmarks. It achieves the highest scores on GQA and AI2D, significantly outperforming both baselines and other models within the LLaVA-MORE family. In MLLM benchmarks, LLaVA-MORE combined with LLaMA-3.1-8B demonstrates instead strong capabilities on the MMB dataset, showing good results in multiple-choice question settings. Notably, our models exhibit less sensitivity to object hallucinations, as seen by the results achieved on the POPE benchmark, and demonstrate superior

	Resolution	# Tokens	VQA Benchmarks				MLLM Benchmarks								
			GQA	Science-QA	TextVQA	AI2D	POPE	MME-P	MME-C	MMB-Cn	MMB-En	SEED-All	SEED-V	SEED-I	MMMU
LLaVA-MORE-3.8B (Ours)															
CLIP ViT-L/14 [225]	336^2	576	62.1	71.3	54.0	61.1	85.9	1372.2	281.1	64.2	69.2	63.5	42.3	69.1	38.8
DINOv2 ViT-L/14 [206]	224^2	256	60.9	66.6	41.4	58.2	85.5	1236.6	281.1	53.8	58.9	59.8	40.6	64.8	37.9
DINOv2$_{reg}$ ViT-L/14 [71]	224^2	256	60.4	69.0	41.3	56.4	85.2	1263.2	**288.2**	57.4	51.4	58.7	41.4	63.2	38.6
SigLIP ViT-L/14 [330]	384^2	729	**63.6**	**73.8**	57.6	**62.9**	86.4	1379.0	282.9	66.5	**71.4**	65.7	46.4	70.8	**40.0**
SigLIP2 ViT-L/14 [273]	384^2	729	63.4	71.8	**59.7**	**62.9**	**86.5**	**1406.7**	282.5	**66.8**	69.8	**66.4**	**47.4**	**71.4**	38.8
LLaVA-MORE-9B (Ours)															
CLIP ViT-L/14 [225]	336^2	576	64.2	75.4	60.7	64.8	**86.8**	**1522.5**	307.5	65.9	71.9	64.5	44.1	69.9	37.9
DINOv2 ViT-L/14 [206]	224^2	256	63.1	71.5	48.1	61.3	85.3	1394.4	**334.3**	56.4	63.8	61.0	40.6	66.4	38.7
DINOv2$_{reg}$ ViT-L/14 [71]	224^2	256	62.8	69.1	47.9	59.1	84.0	1413.9	295.4	60.1	53.8	60.1	42.3	64.7	38.3
SigLIP ViT-L/14 [330]	384^2	729	64.8	**76.3**	63.9	64.7	86.1	1487.9	299.3	**69.1**	**74.4**	66.6	**46.6**	71.9	**39.7**
SigLIP2 ViT-L/14 [273]	384^2	729	**65.6**	76.2	**66.7**	**65.3**	86.1	1510.9	308.2	68.0	72.0	**67.5**	46.0	**73.1**	38.7

Table 4.4: Performance analysis with varying visual backbones. Results are reported for the best small- and medium-scale LLaVA-MORE configurations using Phi-4-3.8B and Gemma-2-9B, respectively. Input resolution and the number of visual tokens are also included.

	S^2	Resolution	# Tokens	VQA Benchmarks				MLLM Benchmarks								
				GQA	Science-QA	TextVQA	AI2D	POPE	MME-P	MME-C	MMB-Cn	MMB-En	SEED-All	SEED-V	SEED-I	MMMU
LLaVA-MORE-3.8B (Ours)																
CLIP ViT-L/14 [225]	✗	336^2	576	62.1	71.3	54.0	61.1	85.9	**1372.2**	**281.1**	64.2	**69.2**	63.5	42.3	69.1	38.8
CLIP ViT-L/14 [249]	✓	$336^2 \times 14$	576	**62.7**	**71.8**	**54.9**	**61.7**	**86.9**	1366.8	263.9	**68.6**	64.4	**64.2**	**43.1**	**69.8**	**39.8**
SigLIP2 ViT-L/14 [273]	✗	384^2	729	63.4	71.8	59.7	62.9	86.5	1406.7	282.5	66.8	69.8	66.4	47.4	71.4	**38.8**
SigLIP2 ViT-L/14 [249]	✓	$384^2 \times 14$	729	**64.1**	**72.2**	**61.5**	**63.2**	**87.4**	**1466.7**	**331.1**	**69.2**	**70.7**	**67.0**	**47.6**	**72.1**	38.7
LLaVA-MORE-9B (Ours)																
CLIP ViT-L/14 [225]	✗	336^2	576	64.2	**75.4**	60.7	**64.8**	**86.8**	**1522.5**	307.5	65.9	**71.9**	64.5	44.1	69.9	37.9
CLIP ViT-L/14 [249]	✓	$336^2 \times 14$	576	**65.2**	73.7	**63.4**	64.2	86.1	1495.9	**331.8**	**68.6**	70.2	**65.1**	**45.2**	**70.4**	**39.0**
SigLIP2 ViT-L/14 [273]	✗	384^2	729	65.6	**76.2**	66.7	**65.3**	86.1	1510.9	308.2	68.0	72.0	**67.5**	**46.0**	**73.1**	38.7
SigLIP2 ViT-L/14 [249]	✓	$384^2 \times 14$	729	**65.9**	74.9	**68.1**	64.1	**86.7**	**1557.6**	**320.7**	**68.6**	**73.5**	67.2	45.6	72.9	**40.2**

Table 4.5: Performance analysis when applying the S^2 multi-scale visual processing [249]. Results are reported considering the best small- and medium-scale LLaVA-MORE configurations with Phi-4-3.8B and Gemma-2-9B, using both CLIP and SigLIP2 visual encoders.

performance on the SEED dataset, further highlighting their robustness in multimodal reasoning tasks.

Comparing small- and medium-scale models, it is noteworthy that some small-scale LLaVA-MORE models outperform even medium-scale baselines like LLaVA-1.5-7B. For instance, in Science-QA and AI2D, LLaVA-MORE with Phi-4-3.8B surpasses both the baseline and LLaVA-MORE with DeepSeek-R1-Distill-LLaMA-8B.

This trend extends to MLLM benchmarks, where LLaVA-MORE models demonstrate competitive performance in MMB and SEED, further highlighting their efficiency and strong reasoning capabilities despite their smaller size. Overall, our results emphasize that scaling up is not the only path to better performance, as architectural choices and fine-tuning strategies significantly impact model effectiveness across different benchmarks.

4.4.5 Effect of the Visual Backbone

We then investigate which visual backbone is more promising for building MLLMs. To this end, we select the best small- and medium-scale models resulting from Table 4.3 and study how their performance is affected when the visual encoder is varied. In detail, we opt for Phi-4-3.8B as the small-scale LLM (*i.e.*, LLaVA-MORE-3.8B), and Gemma-2-9B as the medium-scale one (*i.e.*, LLaVA-MORE-9B). As per the visual backbone, in Table 4.4, we include four pre-trained ViT-based models, in addition to the standard CLIP used by LLaVA. All models share the ViT-L/14 architecture, and yet there are striking differences in terms of training data, input image resolutions, and on the pre-training strategies. In particular, we can delineate two pre-training paradigms: DINOv2 [206], eventually enhanced with registers [71], has been pre-trained with self-supervision and knowledge distillation on 142M images, while the other visual encoders leverage the weak supervision of noisy image-text pairs during pre-training. Specifically, CLIP [225] learns robust visual models via contrastive learning, while SigLIP [330] exploits the sigmoid loss to improve cross-modal alignment. The recent SigLIP2 [273] builds upon SigLIP by incorporating additional training objectives, including image captioning, self-distillation, and image-masked prediction.

From Table 4.4, we observe that image-text pre-trained visual backbones consistently outperform DINOv2 backbones at both small and medium scales. Moreover, adding register tokens in DINOv2 does not help in reducing the gap. We hypothesize that, thanks to their pre-training, CLIP and SigLIP provide visual features that are readily aligned with text, simplifying the role of the multimodal adapter of making them understandable by the LLM.

The only exception is MME-C, where DINOv2 with registers scores the highest at the 3.8B scale, and DINOv2 is the best at the 9B scale with 334.3 points. Among the models that benefited from image-text pre-training, SigLIP shows a substantial improvement over CLIP at both scales, highlighting its effectiveness despite introducing a computational overhead—specifically, it forces the LLM to process approximately 26% more visual tokens due to the use of higher-resolution

	VQA Benchmarks				MLLM Benchmarks								
	GQA	Science-QA	TextVQA	AI2D	POPE	MME-P	MME-C	MMB-Cn	MMB-En	SEED-All	SEED-V	SEED-I	MMMU
LLaVA-MORE-3.8B (Ours)													
LLaVA Pre-Train LCS (558k)	63.4	71.8	59.7	62.9	86.5	1406.7	282.5	66.8	69.8	66.4	47.4	71.4	38.8
LAION (558k)	64.3	**72.5**	**62.3**	**65.2**	**86.8**	**1453.2**	287.1	**67.2**	**72.3**	66.6	46.4	71.9	**39.7**
Recap (558k)	**64.6**	71.7	61.4	64.5	86.5	1428.7	**297.9**	67.1	71.6	67.3	**47.7**	72.5	39.0
LAION+Recap (558k)	**64.6**	71.8	61.3	63.9	86.6	1425.8	297.5	65.8	71.7	**67.6**	47.5	**72.9**	39.2
LLaVA-MORE-9B (Ours)													
LLaVA Pre-Train LCS (558k)	65.6	76.2	66.7	**65.3**	86.1	1510.9	308.2	68.0	72.0	67.5	46.0	73.1	38.7
LAION (558k)	65.6	76.0	67.0	65.1	**86.9**	**1579.8**	**350.7**	68.5	73.9	67.4	45.4	**73.2**	41.1
Recap (558k)	**65.9**	76.2	67.2	64.3	**86.9**	1540.8	318.2	**69.4**	**74.7**	**67.6**	**47.2**	73.0	40.2
LAION+Recap (558k)	65.8	**77.1**	**67.8**	65.2	86.7	1537.8	335.4	65.9	73.5	67.4	45.2	73.1	**41.2**

Table 4.6: Performance analysis when changing the training dataset during the first pre-training stage. Results are reported for the best small- and medium-scale LLaVA-MORE configurations using Phi-4-3.8B and Gemma-2-9B.

inputs. Despite its more complicated training recipe, the new SigLIP2 performs on par with the original SigLIP at the 3.8B scale but records an average 0.4% gain over SigLIP at the 9B scale. For this reason, we will always include SigLIP2 as the visual backbone in all the subsequent experiments. One key finding from our evaluation is that SigLIP-based visual backbones establish a new performance frontier for MLLMs across most benchmarks. Beyond the advantage conferred by increased resolution, we attribute much of this success to the massive billion-scale image-text pre-training enabled by the sigmoid loss of SigLIP, compared to the 400M image-text pairs seen by CLIP during pre-training.

4.4.6 Effect of Image Resolution

The choice of a robust visual backbone and the appropriate LLM are critical factors in enhancing model performance. However, the resolution of the input image plays a similarly important role in visual understanding. Higher image resolutions provide more fine-grained visual information, which can significantly aid the MLLM in better interpreting the content and thereby improving performance.

Table 4.5 presents an evaluation of the LLaVA-MORE models that analyzes the impact of increasing image resolution when using the CLIP and SigLIP2 visual backbones. To tackle this, we leverage the S^2 scheme [249], a widely adopted method designed to enhance image resolution. Specifically, we interpolate the input image to create additional copies with $2\times$ and $3\times$ the standard resolution accepted by the visual encoder. The copies are chunked into 4 and 9 squared im-

ages respectively, of the same resolution as the original one. In total, 14 images are generated per sample, each processed independently by the visual encoder. The resulting visual tokens are spatially pooled and then channel-wise concatenated, resulting in the same number of visual tokens, but with 3× more channels.

Notably, Table 4.5 (top) shows that S^2 generally improves LLaVA-MORE-3.8B when using CLIP as the visual backbone. The improvements are even more consistent when switching to SigLIP2, which already works at a higher resolution than CLIP. However, when scaling up to LLaVA-MORE-9B (Table 4.5 bottom), the benefits of S^2 vanish for some benchmarks. For instance, on ScienceQA and AI2D, S^2 appears detrimental at the 9B scale, while it is advantageous with LLaVA-MORE-3.8B.

From these results, we can conclude that small-scale MLLMs may greatly benefit from working with high-resolution images. However, the positive impact of higher resolution appears to diminish as model size increases, and, ultimately, the benefits of increasing image resolution seem to be highly task-dependent. For instance, with TextVQA, a benchmark that heavily relies on detecting and recognizing text within images, S^2 consistently brings improvements. A similar pattern is found on GQA, which challenges MLLMs with questions requiring in-depth scene understanding. Curiously, we find substantial gains by applying S^2 on the Chinese version of MMBench (MMB-Cn), especially at the 3.8B scale, while its behavior on the English version is conflicting: S^2 improves with SigLIP2, but it is detrimental with CLIP.

4.4.7 Effect of Pre-Training Data

Finally, we analyze the effect of using different data sources for the pre-training phase. Specifically, we select 558K samples from three different sources for image-caption pairs: (i) samples directly derived from the LAION dataset, (ii) samples from Recap-DataComp-1B [162] where captions are generated by an MLLM and provide a dense description of the scene, and (iii) a balanced combination of the first two configurations (*i.e.*, LAION+Recap). Table 4.6 compares these three new datasets against the original LLaVA pre-training recipe

(*i.e.*, LLaVA Pre-Train LCS), consisting of a mixture of image-caption pairs from LAION [243], CC3M [44], and SBU [207].

Interestingly, LLaVA-MORE-3.8B never achieves the top score with the original mixture. Conversely, exclusively training on LAION stands out as the winning choice in 8 out of 13 benchmarks. This may be because LAION's noisy web-sourced pairs closely resemble the nature of the training data used by the SigLIP2 visual encoder behind LLaVA-MORE-3.8B. In contrast, LLaVA-MORE-9B shows more robust performance across the different configurations. Notably, adding Recap samples boosts LLaVA-MORE-9B's Chinese fluency, leading to the best score of 69.4 on MMB-Cn and improving other MLLM results.

4.4.8 Qualitative Results

To better understand the behavior of different MLLMs beyond their quantitative performance on standard MLLM and VQA benchmarks, we also present qualitative results on the image captioning task*, as shown in Fig. 4.4. In particular, we compare LLaVA-1.5-7B [177], which serves as our baseline, against our model based on Phi-4-3.8B and Gemma-2-9B using SigLip2 as visual encoder. As it can be seen, LLaVA-MORE versions can effectively describe input images and provide detailed and rich textual descriptions. For example, only the two versions of LLaVA-MORE successfully avoid hallucinations and provide an accurate description of the scene. In contrast, the LLaVA-1.5-7B model incorrectly identifies the object in question (*i.e.*, erroneously recognizing the snowboard in the image as skis) highlighting a notable failure in visual understanding and alignment with textual output.

4.4.9 Conclusion

This work presents a quantitative analysis of the integration of different LLMs and visual backbones into the LLaVA architecture, aiming to systematically and comparably evaluate the contribution of each component. To the best of our

*For these qualitative results, we use the prompt: "`Describe this image.`".

Figure 4.4: Qualitative of image descriptions generated by three MLLMs (LLaVA-1.5-7B, LLaVA-MORE-3.8B, and LLaVA-MORE-9B). The example shows differences in detail, context, and narrative style across scenarios like sports, transport, toys, and portraits.

knowledge, this is the first comparative analysis of different backbones conducted under consistent experimental settings, using the same dataset across all training stages and the same evaluation protocol. Moreover, this work also aims to develop a framework for training and evaluating different versions of MLLM.

4.5 Discussion

The empirical findings presented above provide concrete evidence for several design trade-offs in MLLMs. We now synthesize these results into broader lessons for multimodal model design.

Visual Backbone Choice Matters More Than Scale

Our results indicate that visual encoders pre-trained with image–text objectives (*e.g.*, CLIP and SigLIP) consistently outperform self-supervised alternatives when integrated into MLLMs. This highlights the importance of semantic alignment at the representation level. Consistently across both the 3.8B and 9B settings, SigLIP/SigLIP2 improves a broad set of benchmarks relative to CLIP, while DINOv2 variants lag despite strong semantic features. This suggests that, in LLaVA-style architectures, the backbone's pre-alignment to language-like semantics reduces the burden on the adapter and stabilizes instruction tuning.

Scaling Is Not a Silver Bullet

Across multiple benchmarks, carefully chosen small- and medium-scale models rival or outperform larger baselines. This suggests that architectural compatibility and training strategy can outweigh raw parameter count. Several results indicate that parameter count alone does not determine downstream behavior: carefully matched small models can outperform medium-scale baselines when the visual backbone and training recipe are well aligned. Practically, this points to a design strategy where backbone compatibility and data choices are optimized first, before resorting to indiscriminate scaling.

Implications for Multimodal Reasoning

Taken together, these observations point toward a shift from indiscriminate scaling toward principled model design. In later chapters, these insights motivate the integration of external knowledge and compositional structure to further enhance multimodal reasoning. Once representation-level alignment is improved (via backbone choice, effective resolution, and data), remaining failure modes increasingly resemble reasoning and knowledge limitations rather than purely perceptual gaps. This observation motivates the thesis direction pursued in later chapters: complementing MLLMs with mechanisms for external knowledge integration and more compositional evaluation to better characterize "true" multimodal reasoning.

From Closed-Book MLLMs to Multimodal RAG

The results of this chapter suggest that, once representation-level alignment is addressed through appropriate backbone and training choices, remaining limitations increasingly concern factual coverage, long-tail knowledge, and multi-hop reasoning. These limitations cannot be resolved by architectural scaling alone. The next chapter builds on this observation by introducing multimodal retrieval-augmented generation, where external knowledge sources complement MLLMs to enable more robust and interpretable visual question answering.

4.6 Chapter Summary

This chapter examined Multimodal Large Language Models through a combined survey and controlled empirical analysis. The survey highlighted the rapid growth and architectural diversity of MLLMs, together with the lack of principled comparisons across models due to confounding differences in training data, evaluation protocols, and design choices. To address this gap, we conducted a controlled study isolating the impact of language model scale, visual backbone, input resolution, and pre-training data within a fixed LLaVA-style pipeline. The results showed that representation alignment and architectural compatibility play a more critical role than indiscriminate scaling, with carefully chosen small and medium models matching or surpassing larger baselines. Despite these improvements, remaining limitations increasingly relate to factual coverage and multi-step reasoning rather than perceptual alignment alone. This observation motivates the transition to multimodal retrieval-augmented generation, explored in the following chapter.

5

Retrieval-Augmented Generation for Multimodal Large Language Models

The previous chapters traced the evolution of multimodal learning from classical image captioning to large-scale multimodal language models capable of instruction following and multimodal reasoning. While these models exhibit strong perceptual grounding and fluent generation, they remain fundamentally *parametric*: all factual knowledge is implicitly encoded in their weights. As a consequence, multimodal LLMs struggle with queries that require long-tail, encyclopedic, or up-to-date information, particularly when answering questions involving rare entities or world knowledge not sufficiently represented during training. In such cases, models often hallucinate or fall back to generic responses, revealing a fundamental limitation of purely parametric knowledge representations.

This chapter is related to the publications "D. Caffagni *et al.*, Wiki-LLaVa: Hierarchical retrieval-augmented generation for multimodal llms, CVPR Workshops 2024" [3] and "F. Cocchi *et al.*, Augmenting multimodal llms with self-reflective tokens for knowledge-based visual question answering, CVPR 2025" [5]. See the list of Publications on page 149 for more details.

This limitation becomes increasingly evident in knowledge-intensive multimodal tasks, such as encyclopedic visual question answering. As shown in the previous chapter, improvements in visual backbones and architectural alignment substantially enhance perceptual understanding, but do not address failures caused by missing or outdated factual knowledge. As multimodal reasoning benchmarks evolve toward more knowledge-demanding settings, access to external information sources becomes a necessary extension rather than an optional enhancement.

Retrieval-Augmented Generation (RAG) offers a principled solution to this problem by decoupling knowledge storage from reasoning. Instead of memorizing all facts during training, a model can dynamically retrieve relevant information from an external knowledge base and incorporate it into the generation process. While retrieval-augmented methods have demonstrated significant gains in text-only language models, extending this paradigm to multimodal settings introduces new challenges. Multimodal LLMs lack native mechanisms for sourcing, integrating, and reasoning over external visual and textual evidence, and naïvely appending retrieved content can overwhelm the model or degrade performance.

Beyond accessing external knowledge, multimodal RAG systems must also determine *when* retrieval is necessary and *which* retrieved information is relevant. Current embedding-based retrievers often operate at a coarse granularity and struggle to accurately rank passages for complex multimodal queries. Without mechanisms for selective retrieval and relevance estimation, external knowledge may introduce noise rather than improving reasoning. The central challenge is therefore to design multimodal systems that can both acquire external knowledge and selectively integrate it, while preserving the strengths of parametric multimodal models.

Scope and role of this chapter. This chapter studies retrieval-augmented generation as the next step in the evolution of multimodal language models. While previous chapters focused on improving perceptual alignment and architectural design, here we address the limitation of parametric knowledge by enabling multimodal models to access external information sources. The chapter investigates how retrieval mechanisms can be integrated into multimodal LLMs

without disrupting their instruction-following behavior, and how models can be endowed with the ability to reason about the necessity and relevance of retrieved knowledge. Together, these contributions position retrieval as a key ingredient for advancing multimodal reasoning beyond perceptual understanding.

Research question. This chapter addresses the following research question: *How can multimodal large language models be augmented with external knowledge in a way that improves factual accuracy and reasoning, while preserving fluency, stability, and generalization?* In particular, we study how retrieval pipelines can be designed for multimodal queries, how retrieved information should be structured and integrated, and how models can selectively decide when external knowledge is required.

5.1 Approach Overview

This chapter explores two complementary approaches to multimodal retrieval-augmented generation, building upon the architectural and empirical insights developed in the previous chapters. While both approaches extend multimodal LLMs with access to external knowledge, they target different aspects of the retrieval problem.

Hierarchical retrieval for knowledge-based visual question answering. The first approach introduces a retrieval-augmented multimodal model designed for visual question answering tasks that require external factual knowledge. It augments a standard multimodal LLM with a hierarchical retrieval pipeline operating over a large document collection. Given an image–question pair, the system first retrieves relevant documents using a contrastive image–text encoder, and then refines retrieval at the passage level using a text-based retriever. The selected passages are appended to the model input as additional context tokens, enabling the multimodal LLM to generate answers grounded in external knowledge without modifying its internal architecture.

Reflective tokens for selective retrieval. The second approach addresses the question of when and how retrieval should occur. It extends a pretrained multimodal LLM with self-reflective tokens that allow the model to rea-

son explicitly about retrieval during generation. These tokens enable the model to predict whether external knowledge is needed and to assess the relevance of retrieved passages. Training follows a staged procedure in which relevance supervision is first learned at the document level and then distilled into the multimodal model. During inference, the model can dynamically decide to invoke retrieval and selectively integrate external information, while relying on parametric knowledge when retrieval is unnecessary.

Together, these two methods define a progressive view of multimodal retrieval-augmented generation. The first establishes how external knowledge can be accessed and injected into multimodal models, while the second investigates how retrieval itself can be controlled and reasoned about. This progression mirrors the broader thesis trajectory, moving from architectural augmentation to principled mechanisms for knowledge-aware multimodal reasoning.

5.2 Hierarchical Retrieval for Multimodal Question Answering

This section presents Wiki-LLaVA, a retrieval-augmented multimodal framework based on hierarchical document and passage retrieval.

5.2.1 Problem Introduction

Recently, LLMs have demonstrated impressive performance in zero-shot textual tasks. Specifically, recent literature has devised models capable of tackling diverse tasks, as instructed by the user [208, 61, 271]. In this context, the classical approach is that of fine-tuning a model on varied tasks that are described through natural language [228, 65], thus empowering the model to assimilate externally provided instructions and facilitating robust generalization across multiple domains. Following these advancements, the computer vision community has started to investigate the extension of such models to vision-and-language contexts, thus generating MLLMs. On this line, the fusion of visual features into LLM backbones through vision-to-language adapters [177, 158, 18, 352] has in-

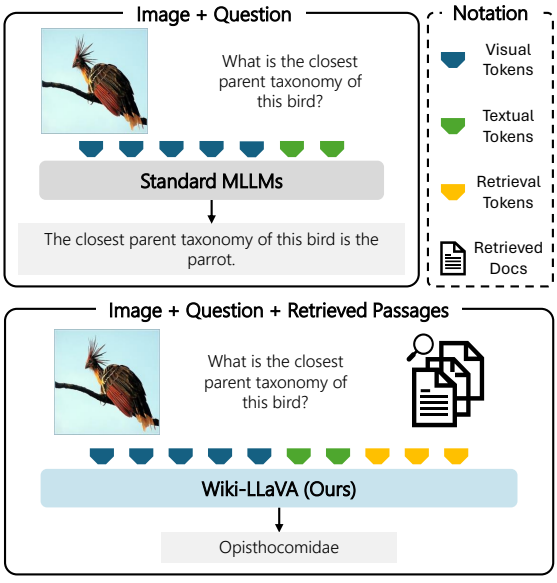


Figure 5.1: Comparison between a standard multimodal LLM and Wiki-LLaVa. Our model integrates knowledge retrieved from an external knowledge base of documents through a hierarchical retrieval pipeline. As a result, it provides more precise answers when tasked with questions that require external knowledge.

duced notable performance improvements, enabling extensive generalization to vision-and-language tasks requiring elaborate visual descriptions.

In this context, MLLMs excel by simply including a small module (*i.e.*, an adapter) that aligns visual features with textual ones. However, despite these models being built upon LLMs trained on large-scale data, they exhibit notable limitations when confronted with highly specific user queries or when a certain degree of compositional reasoning is required to formulate the response. Moreover, certain types of knowledge are difficult to encode within the parameters of an MLLM, due to the scarcity of long-tail information in the training data. In response to this challenge, different benchmarks have been recently introduced for evaluating the capabilities of MLLM to tackle queries related to external data, such as InfoSeek [57] and Encyclopedic-VQA [194]. While different works [224, 148, 70, 158] have been testing on these benchmarks, underscoring the significance of this area, none of them has developed architectures specifically

designed for tackling external knowledge.

Motivated by these observations, we introduce Wiki-LLaVA, a retrieval-augmented extension of LLaVA designed for knowledge-based visual question answering. The core idea is to keep the multimodal generation architecture unchanged, while augmenting its input context with evidence retrieved from an external document collection. We implement this through a hierarchical retrieval pipeline: an entity-level stage first identifies candidate Wikipedia pages using a contrastive vision–language retriever, and a second passage-level stage selects the most relevant textual chunks using a dedicated text retriever. The retrieved evidence is then appended to the multimodal prompt, enabling the model to ground its answers in external knowledge while preserving the instruction-following interface of LLaVA. We evaluate this approach on Encyclopedic-VQA and InfoSeek, analyzing both retrieval quality and end-to-end answering accuracy. Results show that retrieval substantially improves performance when questions require entity-specific facts, while also highlighting retrieval as the main bottleneck at large knowledge-base scale. This study therefore establishes a simple but effective baseline for multimodal RAG, and motivates the need for more selective and model-aware retrieval mechanisms, which we address in the second part of this chapter.

5.2.2 Proposed Method

The design of Wiki-LLaVA follows two principles: (i) augment knowledge without modifying the MLLM architecture, and (ii) retrieve evidence at the appropriate granularity to reduce noise in the prompt. Concretely, we implement a two-level retrieval pipeline (entity, then passage) over a document collection, and inject the selected passages as additional context tokens for the MLLM. Our goal is to equip MLLMs with the ability to answer complex and specific questions that cannot be addressed solely through the image content and pre-trained knowledge. To achieve this, we propose Wiki-LLaVA, which integrates external knowledge derived from an external memory into the LLaVA model, without significantly altering its design. Instead, we augment the capabilities of the model by incor-

porating retrieval information as additional input context. Overall, Wiki-LLaVA comprises three components, as shown in Fig. 5.2: a visual encoder, which is employed to provide the MLLM with visual context and as a query to retrieve from an external knowledge base, the knowledge base itself (*e.g.*, Wikipedia), and a hierarchical retrieval module which retrieves relevant documents and passages from the external knowledge base, to be employed as additional context for the MLLM.

5.2.3 Knowledge-based Augmentation

Multimodal integration and autoregressive generation. An MLLM usually takes as input a multimodal input query, comprising both image and text, and generates a textual output in an autoregressive manner. Formally, the architecture is trained to model a probability distribution $p(w_t|I, w_0, w_1, ..., w_{t-1}, \theta)$, where θ denotes the parameters of the model, I represents an input image, and $w_0, .., w_{t-1}$ denotes the textual prompt. The textual prompt usually includes a pre-defined system-level prompt and a question related to the input image, given by the user. Clearly, a standard MLLM can only rely on the user prompt, the input image, and the knowledge stored in its internal parameters (*i.e.*, θ) to accommodate requests, thus limiting its ability to answer questions that rely on external knowledge.

We employ LLaVA [179] as our reference MLLM. LLaVA exploits the capabilities of a pre-trained LLM (*i.e.*, Vicuna [61]) and a pre-trained visual model (*i.e.*, a CLIP-based visual encoder [225]), which are interconnected through an MLP adapter, in charge of converting CLIP features to dense input tokens. For an input image I, therefore, LLaVA utilizes a pre-trained CLIP visual encoder E_v, extracts a dense grid of visual features $Z_v = E_v(I)$, which is then projected via a learnable MLP to produce a sequence of dense embedding tokens $v_o, v_1, ..., v_N$. Finally, these are prepended to the system prompt, and the full sequence of visual and textual tokens is then given as input to the LLM component of the model.

Augmentation with external knowledge. To augment the MLLM with external knowledge, we enrich the input context by injecting relevant tex-

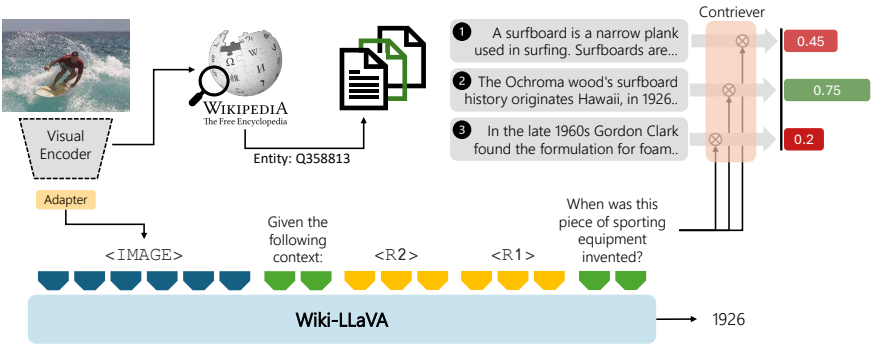


Figure 5.2: Overview of the architecture of Wiki-LLaVA, which augments a multimodal LLM with external knowledge through a hierarchical retrieval pipeline.

tual data from an external memory composed of documents. Formally, the distribution of the MLLM is conditioned on additional textual retrieval-knowledge tokens, leading to

$$p(w_t | \overbrace{v_o, v_1, ..., v_N}^{\text{Visual tokens}}, \underbrace{w_0, w_1, ..., w_{t-1}}_{\text{System + user prompt}}, \overbrace{e_0, e_1, ..., e_\tau}^{\text{External memory tokens}}), \tag{5.1}$$

where $e_0, ..., e_\tau$ represents the added tokens retrieved from the external memory. In contrast to the standard formulation of MLLMs, by enriching the input context we allow the model to generate more specific answers by exploiting tokens retrieved from the memory.

HIERARCHICAL RETRIEVAL FROM AN EXTERNAL MEMORY. The external memory comprises a collection of (document, image, text-title) triplets taken from documents, denoted as $\mathcal{D} = \{(d_i, t_i)_i\}$. Within this memory, we conduct a hierarchical two-step search to retrieve appropriate information. Initially, we locate the most pertinent document, followed by identifying the relevant passage inside a particular document, which is subsequently exploited as additional input context in the MLLM.

In the first stage, given an input query image I we perform an approximate k-nearest neighbor search into the external memory, using document titles as re-

trievable keys. The similarity between the query image and the text titles is modeled as the inner product between their respective embeddings, which are computed through the visual and textual CLIP encoders (*i.e.*, E_v and E_t), as follows:

$$\text{sim}(I_i, t_i) = E_v(I) \cdot E_t(t_i)^T. \tag{5.2}$$

Then, the knowledge retriever returns the top-k documents associated with the most relevant items retrieved using the aforementioned procedure.

Retrieving document passages. In the second step, we analyze each of the retrieved documents to identify the most relevant passages corresponding to the user's question. Each document is defined as a sequence of chunks, denoted as $d_i = [c_{i_0}, .., c_{i_T}]$, and, given the input question, we retrieve the chunks with the highest similarity to the question. We employ the Contriever architecture [122] to embed each chunk of the selected document, along with the query (*i.e.*, the question provided by the user), and compute the similarity as an inner product between embeddings. By retrieving the n most appropriate passages inside each of the retrieved documents, overall we obtain $k \cdot n$ passages.

Context enrichment. Once we find the most relevant chunks, we employ their raw contents as an additional input to the MLLM. Specifically, the final prompt that we employ includes the image tokens, the retrieved raw chunks, the system-level prompt, and the user question. Formally, considering three retrieved passages, the final prompt is defined as follows:

```
<IMAGE>\nGiven the following context:\n
<R1>\n<R2>\n<R3>\n <QUESTION>
Give a short answer. ASSISTANT:
```
(5.3)

5.2.4 Training

While the aforementioned approach could work in a zero-shot fashion, using the original weights θ of the pre-trained MLLM, we also investigate the case of fine-tuning the model to augment its capabilities of exploiting retrieved passages. In

particular, in this case, the model is trained on pairs of questions and ground-truth answers requiring external knowledge. As this would potentially reduce the capabilities of the MLLM on tasks not requiring external knowledge (*i.e.*, all the other tasks on which the model has been originally trained), we apply a data mixing approach in which ground-truth pairs requiring external knowledge are mixed with ground-truth pairs not requiring external knowledge in the same mini-batch.

5.2.5 Experimental Setup

In this section, we first introduce the experimental settings, describing the datasets employed, the evaluation protocol, and the implementation and training details used to perform the experiments. Then, we present our experimental results, analyzing the effectiveness of CLIP fine-tuning and evaluating how it is possible to incorporate retrieved knowledge in an MLLM. Finally, limitations of the proposed approach and possible future works are reported.

Encyclopedic-VQA [194]. The dataset contains around 221k question-answer pairs associated with 16.7k different fine-grained entities, with up to 5 images representing the same entity. Overall, there are more than 1M triplets composed of an image, a question, and the corresponding answer. Fine-grained entities and related images are extracted from iNaturalist 2021 [277] and Google Landmarks Dataset V2 [295], which are associated with the corresponding Wikipedia article. Questions are divided into four different categories, namely single-hop, automatically generated, multi-answer, and two-hop. In particular, single-hop questions have been manually annotated and a single Wikipedia article is needed to answer them. Automatically generated questions are similar to the single-hop questions but have been generated by automatic models. Multi-answer questions, instead, can be answered with a list of terms, but always refer to a single fine-grained entity. Finally, two-hop questions require two retrieval steps to answer them. The dataset also comes with a knowledge base composed of 2M Wikipedia articles, suitable for answering dataset questions.

Dataset triplets are divided into training, validation, and test splits respectively

composed of 1M, 13.6k, and 5.8k samples. In our experiments, we employ the training split to fine-tune the LLaVA model and report the results on the test set of the dataset. During testing, we filter out two-hop questions resulting in 4,750 test triplets.

InfoSeek [57]. The dataset contains 1.3M image-question-answer triplets corresponding to around 11k different entities (*i.e.*, Wikipedia articles). The vast majority of questions have been obtained with an almost entirely automatic procedure, by filling human-authored templates with knowledge triples from Wikidata. In this case, images are derived from the OVEN dataset [112]. Triplets are divided into training, validation, and test sets, with around 934k, 73k, and 348k samples respectively. At the time of the submission, the ground-truth answers and entities from the test set were not available. Therefore, we report our results on the validation split. Both validation and test sets contain questions related to new entities not included in the training split and questions not seen during training.

Along with image-question-answer triplets, a knowledge base composed of 6M Wikipedia entities is provided. In our experiments, we consider a randomly extracted subset of 100k entities, in which we guarantee the presence of the 6,741 entities associated with questions from the training and validation splits.

5.2.6 Implementation Details

LLaVA fine-tuning. We employ two distinct fine-tuning approaches, with each being exclusively applied to one of the datasets. In order to maintain the performance of the LLaVA model on well-established MLLM datasets, we supplement fine-tuning data with samples from the LLaVA-Instruct dataset [179]. Specifically, given its size of 158k, we double the probability of having examples from this dataset in each mini-batch. To reduce the number of trainable parameters, we train using low-rank adapters [111] with a total batch size of 512 samples.

Retrieval. Textual documents sourced from Wikipedia content are embedded using the Contriever architecture [122], segmenting the text into chunks of 600 characters each. Furthermore, for streamlined efficiency, the process in-

volves utilizing a single visual encoder. Specifically, following the LLaVA architecture [179], we employ the CLIP ViT-L/14@336 backbone to embed images to give as input to the MLLM, while simultaneously leveraging it to extract query visual features in the initial hierarchical retrieval step, facilitating the integration of an external memory component.

To perform entity retrieval, we employ approximate kNN search rather than exact kNN search because it significantly improves the computational speed of the entire pipeline. To this aim, we employ the Faiss library [131] and a graph-based HNSW index with 32 links per vertex.

5.2.7 Evaluation Protocol

We evaluate our models in two settings: without external knowledge base and with external knowledge base. The former means that we ask the model to directly answer a visual question, by solely relying on the competencies learned during pre-training and/or fine-tuning. On the other hand, in the latter setting, we leverage the proposed hierarchical retrieval method to search for additional information in the external knowledge base. In practice, this is represented by two dumps of Wikipedia comprehending 2M and 100k pages, respectively for Encyclopedic-VQA and InfoSeek. Concerning the evaluation metrics, we report the accuracy over the Encyclopedic-VQA test split and the InfoSeek validation split, following the official evaluation scripts provided along with the datasets.

5.2.8 Experimental Results

Analyzing CLIP performance. We start by evaluating entity retrieval results using CLIP. In this setting, we consider images from the Encyclopedic-VQA test set and InfoSeek validation set and measure the CLIP ability to find the correct entity within the knowledge base of each respective dataset (*i.e.*, composed of 2M entries for Encyclopedic-VQA and 100k entries for InfoSeek). As previously mentioned, we perform retrieval using images as queries and Wikipedia titles as retrievable items.

Results are reported in Table 5.1 in terms of recall@k (R@k) with k =

Dataset	KB	R@1	R@10	R@20	R@50
Encyclopedic-VQA	2M	3.3	9.9	13.2	17.5
InfoSeek	100k	36.9	66.1	71.9	78.4

Table 5.1: Entity retrieval results on the Encyclopedic-VQA test set and InfoSeek validation set. To comply with the visual encoder employed in LLaVA, all results are obtained using CLIP ViT-L/14@336.

$1, 10, 20, 50$ which measures the percentage of times the correct entity is found in the top-k retrieved elements. Notably, correctly retrieving the Wikipedia entity associated with the input image strongly depends on the size of the employed knowledge base. In fact, when using 100k items, as in the case of InfoSeek, the correct entity is retrieved as the first item 36.9% of the time and among the top-10 66.1% of the time. Instead, when using a significantly larger knowledge base as in the case of Encyclopedic-VQA, which contains 2M items, retrieval results are significantly lower with 3.3% and 9.9% respectively in terms of R@1 and R@10. These results highlight a central limitation of straightforward multimodal RAG: entity-level retrieval quality degrades sharply as the knowledge base scales, and retrieval errors propagate directly to the generation stage by injecting irrelevant evidence into the prompt. This motivates hierarchical retrieval and relevance-aware selection as first-class design choices, rather than secondary implementation details.

Results on Encyclopedic-VQA and InfoSeek. We then report visual question-answering results in Table 5.2. We include the performance of zero-shot models like BLIP-2 [158], InstructBLIP [70], and the LLaVA-1.5 baseline model [179], which are not fine-tuned on the considered datasets and that do not leverage the external knowledge base. Moreover, we consider the accuracy results of LLaVA-1.5 when fine-tuned on the training set of Encyclopedic-VQA and InfoSeek, but not augmented with retrieved context. The results of our approach (*i.e.*, Wiki-LLaVA) are reported both in the standard setting in which CLIP is used to retrieve the most representative entity from the knowledge base and in its *oracle* version, which employs the entity corresponding to the input image-question pair. For both cases, we consider a different number n of retrieved textual chunks, all corresponding to the top-1 (or ground-truth) entity. When em-

Model	LLM	KB	k	n	Enc-VQA		InfoSeek		
					Single-Hop	All	Unseen-Q	Unseen-E	All
Zero-shot Models									
BLIP-2 [158]	Flan-T5$_{XL}$	✗	-	-	12.6	12.4	12.7	12.3	12.5
InstructBLIP [70]	Flan-T5$_{XL}$	✗	-	-	11.9	12.0	8.9	7.4	8.1
LLaVA-1.5 [177]	Vicuna-7B	✗	-	-	16.3	16.9	9.6	9.4	9.5
Fine-tuned Models									
LLaVA-1.5 [177]	Vicuna-7B	✗	-	-	23.3	28.5	19.4	16.7	17.9
Wiki-LLaVA	Vicuna-7B	✓	1	1	21.8	26.4	26.6	24.6	25.5
Wiki-LLaVA	Vicuna-7B	✓	1	2	19.9	23.2	29.1	26.3	27.6
Wiki-LLaVA	Vicuna-7B	✓	1	3	17.7	20.3	30.1	27.8	28.9
Wiki-LLaVA	Vicuna-7B	✓	2	1	21.3	25.4	27.8	24.6	26.1
Wiki-LLaVA	Vicuna-7B	✓	3	1	20.5	24.3	27.4	24.5	25.3
Wiki-LLaVA	Vicuna-7B	✓	1	1	34.7	37.2	41.1	41.1	41.1
Wiki-LLaVA	Vicuna-7B	✓	1	2	39.2	40.2	49.1	46.5	47.8
Wiki-LLaVA	Vicuna-7B	✓	1	3	38.5	38.6	52.7	50.3	51.5

Table 5.2: Accuracy results on the Encyclopedic-VQA test set and InfoSeek validation set. **Yellow color** indicates models employing the CLIP model to perform entity retrieval, while **gray color** indicates the use of ground-truth entities (*i.e.*, oracle). k denotes the number of retrieved entities, and n represents the number of textual chunks retrieved for each entity that are given to the MLLM as additional context.

ploying CLIP, we also vary the number k of retrieved entities (*i.e.*, $k = 1, 2, 3$) using $n = 1$ when k is greater than 1. This choice is given by the maximum context length that Vicuna takes as input, which is set to 2,048 tokens.

As it can be seen, zero-shot MLLMs face difficulties in correctly answering the given questions as these models can only rely on the knowledge embedded inside the LLM.

While Table 5.2 reports the performance of zero-shot MLLMs without external knowledge, we do not include zero-shot models augmented with retrieved passages. In practice, models such as BLIP-2, InstructBLIP, or LLaVA are not trained to interpret raw retrieved documents within their prompts. As also observed in prior work on multimodal RAG, naïvely injecting retrieved passages into the prompt often fails to improve performance and may even degrade answer quality. For this reason, most retrieval-augmented multimodal approaches rely on task-specific fine-tuning to enable the model to effectively exploit retrieved evidence. Consequently, our comparison focuses on models that have been explicitly adapted to operate in a retrieval-augmented setting.

When instead using an external knowledge base, the accuracy results signifi-

	Enc-VQA		InfoSeek		
Fine-tuning	Single-Hop	All	Unseen-Q	Unseen-E	All
✗	16.3	16.9	9.6	9.4	9.5
✓	23.4	29.0	17.1	15.0	16.0
✓ + LLaVA-Instruct	23.3	28.5	19.4	16.7	17.9

Table 5.3: Performance analysis when using the LLaVA-Instruct dataset during fine-tuning. All results are obtained without external knowledge retrieval.

cantly increase especially on the InfoSeek dataset with 100k retrievable items.

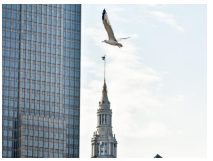
Overall, retrieving passages from different entities does not always help increase the results. Instead, using more than one textual chunk as additional context for the MLLM generally improves the final accuracy on the InfoSeek validation set with an overall improvement of 2.1 and 3.4 accuracy points with $n = 2$ and $n = 3$ respectively. Furthermore, it is worth noting that employing oracle entities significantly boosts the final accuracy. In particular, oracle entities lead to an improvement of 13.8% on Encyclopedic-VQA and 22.6% on InfoSeek, comparing the best-performing configuration with CLIP-based entity retrieval (*i.e.*, $k = 1$ and $n = 1$ for Encyclopedic-VQA and $k = 1$ and $n = 3$ for InfoSeek) with the best performing oracle-based version (*i.e.*, $k = 1$ and $n = 2$ for Encyclopedic-VQA and $k = 1$ and $n = 3$ for InfoSeek). These results confirm the effectiveness of directly employing retrieved passages to augment a pre-trained MLLM and further highlight the importance of having a good entity retrieval model to limit the possibility of feeding the MLLM with irrelevant content.

Some qualitative results on sample image-question pairs from Encyclopedic-VQA (first row) and InfoSeek (second row) are reported in Fig. 5.3, comparing the answers given by Wiki-LLaVA with those coming from the original LLaVA-1.5 model. For completeness, we also report some failure cases (third row) in which both models are not able to correctly answer the given question.

Evaluating the importance of the fine-tuning datasets. As described in Sec. 5.2.4 and Sec. 5.2.6, the MLLM fine-tuning is done with a mixture of data containing image-question-answer triples from the Encyclopedic-VQA or InfoSeek training set and visual instruction tuning data from LLaVA-Instruct [179], which has been used to originally fine-tune the LLaVA model. In

In what state is this building located?

LLaVA-1.5:
California ✗
Wiki-LLaVA:
Washington ✓



When was this building constructed?

LLaVA-1.5:
1970 ✗
Wiki-LLaVA:
1927 ✓

Which geographic area is this fish found?

LLaVA-1.5:
Gulf of Mexico ✗
Wiki-LLaVA:
Brazil ✓

What is the oldest age of this animal?

LLaVA-1.5:
10 years ✗
Wiki-LLaVA:
24.9 ✓

Which culture is associated with this place?
`Ancient Greek`

LLaVA-1.5:
Roman ✗
Wiki-LLaVA:
Nuragic Civilization ✗

What is the name of the main club of this stadium?
`FC Rotor`

LLaVA-1.5:
Real Madrid ✗
Wiki-LLaVA:
FC Dynamo Kyiv ✗

Figure 5.3: Qualitative results on sample image-question pairs from Encyclopedic-VQA (first row) and InfoSeek (second row) comparing the proposed approach with the original LLaVA-1.5 model. Some failure cases are shown in the third row with the corresponding ground-truth.

	MME		MMMU	MMB	POPE	
Fine-tuning	Cogn	Perc	Acc	Acc	Acc	F1
-	355.7	1513.3	35.1	71.6	86.9	85.8
Enc-VQA	200.7	802.8	36.6	67.7	72.9	63.4
Enc-VQA + LLaVA-Instruct	290.0	1170.1	36.6	70.4	87.2	86.6
InfoSeek	296.8	1377.2	35.2	71.7	82.0	79.6
InfoSeek + LLaVA-Instruct	341.3	1438.9	35.6	71.1	85.8	84.2

Table 5.4: Performance preservation analysis with respect to the original LLaVA-1.5 model (first row) on diverse benchmarks for MLLM evaluation.

Table 5.3, we evaluate the effect of mixing fine-tuning data for the knowledge-based VQA task. In this setting, we only report the results of the fine-tuned models without external knowledge retrieval. Notably, using visual instruction tuning data can help to regularize the fine-tuning phase on the InfoSeek dataset, leading to an overall improvement of 1.9 accuracy points compared to the model fine-tuned only on image-question-answer triplets from the training set of the dataset. On Encyclopedic-VQA, instead, training with instruction tuning data does not lead to performance improvement although without degrading the original results.

Preservation of LLaVA performance. Finally, we analyze the impact of LLaVA fine-tuning on knowledge-based VQA datasets when evaluating the model on common MLLM evaluation benchmarks [2]. In particular, we include results on MME [87] which contains image-question pairs covering 14 different tasks grouped in two macro-categories (*i.e.*, cognition and perception), MMMU [326] that is composed of multiple-choice and open-ended questions possibly interleaved with one or more images and extracted from diverse university textbooks and online courses, MMBench (MMB) [183] that includes multiple-choice questions across 20 different domains, and POPE [165] that is focused on evaluating object hallucinations and comprises binary classification entries, each related to an image. More details about the evaluation metrics and number of samples can be found in the original paper of each dataset.

Results are shown in Table 5.4 comparing the original LLaVA model with the two fine-tuned versions on Encyclopedic-VQA and InfoSeek, with and without the use of visual instruction tuning data. Overall, employing samples from the LLaVA-Instruct dataset can better preserve the results of the original model, only partially degrading the performance on the considered benchmarks compared to the original model. While the most significant deterioration is achieved on the MME dataset, in the other settings the original performances are better preserved, also leading to a slight improvement on MMMU and POPE benchmarks compared to the LLaVA-1.5 results.

5.2.9 Limitations and Future Works

While our work represents an important first step toward retrieval-augmented multimodal language models, it also highlights several limitations that point to promising directions for future research.

First, the effectiveness of the proposed hierarchical pipeline is strongly constrained by the quality of the underlying embedding spaces. Entity-level retrieval remains brittle at scale, especially when relying on generic image–text embeddings that are not explicitly optimized for fine-grained entity discrimination. Improving cross-modal representations tailored to large-scale knowledge retrieval

remains an open challenge.

Second, the proposed framework treats retrieval and generation as largely decoupled processes. Retrieved passages are injected into the model input as raw context, leaving it to the language model to implicitly determine which information is useful. As a result, retrieval noise can easily propagate to the generation stage, particularly when irrelevant or weakly related passages are retrieved. This highlights a fundamental limitation of prompt-based RAG approaches: the model lacks an explicit mechanism to reason about the necessity and relevance of external knowledge.

More broadly, our approach does not endow the multimodal model with the ability to decide *when* retrieval should be performed, nor to selectively attend to or discard retrieved evidence. Addressing this limitation likely requires moving beyond passive context injection toward architectures or training strategies that allow the model to explicitly reason about retrieval itself. More generally, this work exposes a key trade-off in multimodal RAG: increasing knowledge access improves factuality, but also increases the risk of distraction and error propagation when retrieval is imperfect. Addressing this trade-off requires retrieval to become a controllable part of the model's reasoning process, rather than a fixed pre-processing step.

These observations motivate the need for retrieval-aware multimodal models, in which the decision to retrieve and the assessment of retrieved content are integrated into the model's reasoning process. In the following section, we address this limitation by introducing self-reflective mechanisms that allow multimodal LLMs to explicitly decide *whether* to retrieve and *which* retrieved evidence to trust during generation.

5.2.10 Conclusion

We have presented Wiki-LLaVA, an architecture for augmenting an existing MLLM with external knowledge. Our proposal leverages an external knowledge source of documents to improve the effectiveness of an MLLM when tasked with questions and dialogues. In particular, we devise a hierarchical architecture for

retrieving documents and eliciting selected parts to be included in the MLLM input context. Extensive experiments demonstrate the effectiveness of the proposed solution, and its capability to maintain the proficiency of the MLLM across different tasks.

5.3 Retrieval-Aware Multimodal Reasoning

This section introduces Reflective LLaVA, which equips multimodal LLMs with self-reflective tokens to reason about the necessity and relevance of retrieved knowledge.

5.3.1 Problem Introduction

While the past few years have seen a surge of LLMs with increasing fluency and reasoning capabilities [34, 228, 271], thanks to the availability of large-scale training data and novel training techniques [65, 208, 126], the Computer Vision community has recently started extending the capabilities of such models beyond pure text, with the inclusion of additional modalities like images, video, and 3D data. The resulting emergence of MLLMs has been characterized by the development of models targeting multiple tasks [2, 102, 260] – ranging from visual dialogue to image generation –, architectural innovations [158, 18, 317], and novel training recipes [177, 146].

What most existing MLLMs share, though, is their exclusive reliance on the knowledge learned at training time – an issue that severely limits their practical applicability to cases that adhere to the training distribution. While this issue is also common to LLMs, it becomes more pressing in the case of MLLMs, where obtaining high-quality and large-scale multimodal data becomes even more difficult. Ideally, indeed, an MLLM should be capable of engaging in dialogues concerning specific visual details, long-tail knowledge, fine-grained categories, and instances [194, 57]. However, this type of knowledge makes it hard for MLLMs to encode in their parameters because such long-tail information occurs rarely in the training data. Additionally, this lack of precise knowledge can also lead the

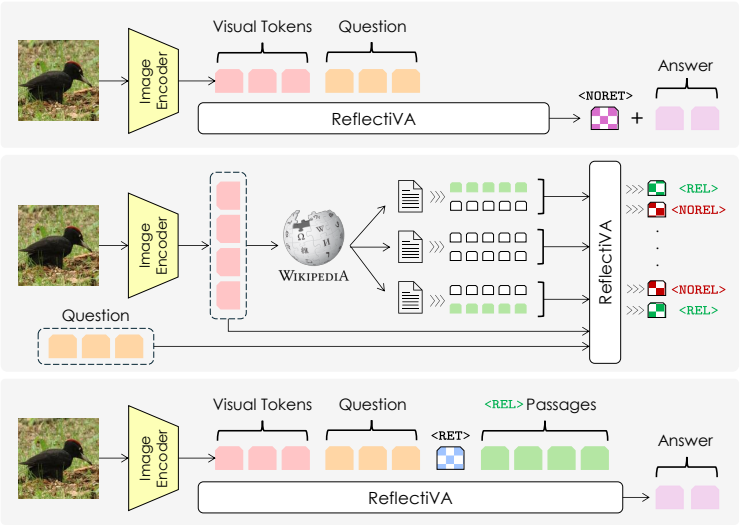


Figure 5.4: Overview of ReflectiVA, which employs reflective tokens for knowledge-based visual question answering. Our model learns to predict the need for retrieval data from an external knowledge source (top), classifies the relevance of each retrieved item (middle) and generates the final answer based on relevant items (bottom).

MLLM to generate incorrect answers, thus again limiting their usage in practical cases.

A viable solution to this issue is to rely on a non-parametric approach, where content from external knowledge sources is incorporated into the MLLM context, conditioning the generation on accurate and relevant information [3, 310]. However, building retrieval-augmented MLLMs presents unique challenges, ranging from the difficulty of retrieving appropriate evidence, to the brittleness of integrating noisy multimodal context into autoregressive generation. Further, an MLLM should also be capable of identifying *when* external knowledge is needed, as opposed to answering questions that simply do not need external knowledge, *e.g.* purely visual questions. Lastly, retrieving relevant items from a multimodal knowledge source is an open problem, with state-of-the-art embedding spaces [225] exhibiting limited performance [293, 172]. This further outlines the need to discover *which* retrieved items may be relevant to answering a given user query.

Our approach is conceptually related to Self-RAG [24], which introduces special tokens that allow a language model to decide when to retrieve external knowledge and to critique retrieved evidence during generation. Self-RAG enables LLMs to interleave generation, retrieval, and self-critique through dedicated control tokens. However, Self-RAG is designed for text-only language models, where both queries and retrieved information are purely textual. In contrast, our setting involves multimodal inputs and multimodal knowledge sources, where retrieval must be conditioned on visual information and relevance judgments must consider both visual and textual context. Reflective LLaVA therefore extends the self-reflective paradigm to multimodal reasoning, enabling models to reason jointly about visual inputs, textual queries, and retrieved knowledge, and to assess the relevance of retrieved passages conditioned on visual evidence.

Drawing inspiration from these challenges, we propose a multimodal model for knowledge-based visual question answering which can jointly determine the need for accessing external knowledge and the relevance of items retrieved from an external knowledge base. Our model termed Reflective LLaVA (ReflectiVA), employs *reflective tokens* to augment the capabilities of a pre-trained MLLM [177] for knowledge-based generation. In particular, the vocabulary of the model is extended to generate additional tokens with which the model can decide whether retrieval is needed or not (Fig. 5.4, top), and if a retrieved sample is relevant or not for the input query (Fig. 5.4, middle). Training is conducted with a two-stage procedure employing two learnable models. We first train an in-article discriminator that can discriminate relevant passages from irrelevant ones found inside the same article. We then employ data annotated synthetically with this model, together with a mixture of existing datasets, to train the final model on all reflective tokens.

Experimentally, we evaluate the performance of the proposed approach on the Encyclopedic-VQA [194] and InfoSeek [57] datasets, which contain question-answer pairs linked with two knowledge bases derived from Wikipedia pages. Additionally, we assess the zero-shot generalization capabilities to two other VQA datasets [149, 125] which may require external knowledge to answer questions correctly. With extensive experiments, we demonstrate that our model outper-

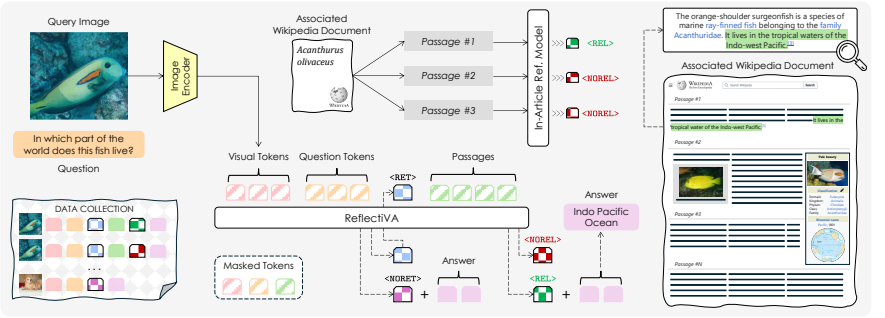


Figure 5.5: Training approach for ReflectiVA. An in-article model is trained to predict the relevance of passages extracted from the ground-truth document corresponding to an (I, q) pair. The in-article model then generates training data for ReflectiVA, which is trained to predict the need for external knowledge and the relevance of passages, along with the answer, using positive, soft- and hard-negative passages.

forms previous works, and provides increased answer accuracy on all considered datasets and settings. Further, we demonstrate that the proposed approach maintains high performance on standard MLLM benchmarks [87, 326, 165] as well as on traditional VQA datasets [120, 252] that do not require external knowledge during generation. Building upon the limitations identified in the previous section, we now move from passive retrieval augmentation to retrieval-aware multimodal reasoning.

Design principles. Our approach is guided by three principles: (i) retrieval should be *optional*, since many multimodal queries are purely perceptual; (ii) retrieved evidence should be *explicitly validated* to reduce noise propagation; and (iii) the resulting system should preserve the general-purpose capabilities of the base MLLM.

5.3.2 Proposed Method

Task Definition. In retrieval-augmented generation, given a textual query q and a query image I, an MLLM is expected to generate an answer y by possibly leveraging additional snippets $\mathbf{S}$ retrieved from an external knowledge source as context. The objective of multimodal retrieval-augmented generation can there-

fore be written as

$$y = \arg\max_{y} \text{MLLM}(y|I, q, \mathbf{S}). \tag{5.4}$$

In our setting, the external knowledge source $\mathcal{S}$ is composed of multimodal documents, each endowed with metadata (*e.g.*, title and summary), textual passages organized in sections, and possibly visual content. Formally, the external database can be defined as a collection

$$\mathcal{S} = \{(\tilde{t}_i, \tilde{\mathbf{P}}_i, \tilde{I}_i)\}_{i=1}^{N}, \tag{5.5}$$

where $\tilde{t}_i$ represents the metadata of a multimodal document, $\tilde{\mathbf{P}}_i$ the set of its textual passages, and $\tilde{I}_i$ its visual content.

Summary of the Approach. To address the limitations of existing retrieval-augmented methods, our approach introduces two innovative strategies. Firstly, we enable the model to determine the optimal timing for retrieval, specifically when generating with an empty retrieval set $\mathbf{S}$ is advantageous because the query does not require external information. Secondly, after the retrieval process, we empower the model to identify the relevance of retrieved passages for generation. Both abilities are enabled through the incorporation of reflective tokens into the vocabulary of the model. These tokens are trained following a two-step two-model training recipe, which ultimately enables the MLLM with the ability to determine whether retrieval is needed and to select pertinent passages from the external database. An overview of our methodology is illustrated in Fig. 5.5.

5.3.3 Adding Reflective Tokens

Given a pre-trained MLLM, we augment its vocabulary $\mathcal{V}_0$ with a set of four additional tokens, *i.e.* {<RET>, <NORET>, <REL>, <NOREL>}, which will enable the model to distinguish whether retrieval is needed (<RET>, <NORET>) and whether a retrieved sample is relevant to the input query (<REL>, <NOREL>).

Generation Protocol. At test time, the MLLM is prompted with an input image I and a query q, and is asked to produce either the <RET> or the <NORET> token. If the <NORET> token is sampled, the MLLM will be asked to directly gen-

erate the answer y without relying on additional snippets. In this case, the generation process follows a schema

$$\begin{aligned} \texttt{"<NORET>"} &\sim \text{MLLM}\left([I, q], \{\texttt{<RET>}, \texttt{<NORET>}\}, 1\right), \\ y &\sim \text{MLLM}([I, q, \texttt{<NORET>}], \mathcal{V}_0), \end{aligned} \tag{5.6}$$

where $y \sim \text{MLLM}(p, V, t)$ indicates that a sequence of tokens y is sampled (*e.g.* through beam search) from the MLLM when prompted with an input p and constrained to emit tokens belonging to a vocabulary V and up to a length of t tokens*. Finally, $[\cdot, \cdot]$ indicates concatenation.

If instead the <RET> token is sampled, we firstly retrieve a set of candidate textual passages $\mathbf{S}_0 = \{s_0, ..., s_k\}$ from the external knowledge base, and then ask the MLLM to evaluate the relevance of each of them through the emission of <REL> and <NOREL> tokens. In this case, the generation follows a protocol in the form

$$\begin{aligned} \texttt{"<RET>"} &\sim \text{MLLM}([I, q], \{\texttt{<RET>}, \texttt{<NORET>}\}, 1), \\ r_i &\sim \text{MLLM}([I, q, \texttt{<RET>}, s_i], \{\texttt{<REL>}, \texttt{<NOREL>}\}, 1), \end{aligned} \tag{5.7}$$

where the second generation step is repeated for each item in $\mathbf{S}_0$ and r_i indicates the relevance token sampled for each retrieved snippet.

After sampling relevance tokens, the MLLM is then asked to generate the final answer using the set of snippets that have been judged relevant, *i.e.* $\mathbf{S} = \{s_i \in \mathbf{S}_0, r_i = \texttt{<REL>}\}$. Formally, this generation stage is defined as

$$y \sim \text{MLLM}([I, q, \texttt{<RET>}, \mathbf{S}], \mathcal{V}_0), \tag{5.8}$$

where, for readability, the concatenation of all items in $\mathbf{S}$ is not made explicit.

Coarse-Grained Retrieval. To obtain the set of candidate passages $\mathbf{S}_0$, we utilize the input query image I as an anchor to retrieve a set of candidate documents from the knowledge source. We encode either the metadata or the im-

*When t is omitted, we let the model generate until an "end of sequence" token is sampled.

age† of each document in the database using a CLIP-based textual or visual encoder [225, 259] and build a dense vector-search database. The encoded features $\mathbf{z}_i \in \mathbb{R}^d$, where d is the dimensionality of the CLIP embedding, act as a search index $\mathbf{Z} = \{\mathbf{z}_i\}_i$. The retriever then employs a non-parametric function to compute the cosine similarity between the embedding of the query image and all search indexes.

Based on this similarity search, the coarse-grained retriever retrieves the top-k articles that are most similar to the query image. The set of candidate passages $\mathbf{S}_0$ is then built as the union of the passages belonging to the top-k articles.

5.3.4 Training an In-Article Reflective Model

To make retrieval a controllable part of multimodal reasoning, we decompose the learning problem into two stages. The first stage focuses on learning fine-grained relevance judgments within a single document, while the second stage transfers this capability to a full retrieval-augmented multimodal model operating across documents. Clearly, a coarse-grained retrieval at the document level is not precise enough to retrieve the exact passage containing the answer. Indeed, this retrieval step is expected to have limited recall at lower values of k. Further, a more detailed examination of relevant documents is necessary to identify passages that can be utilized by the MLLM.

To this aim, we train the MLLM to emit the relevance tokens `<REL>` and `<NOREL>`. This is done with a two-stage, two-model training pipeline. Initially, we train an in-article reflective MLLM capable of distinguishing between relevant passages and negative passages from the same article. Subsequently, we employ predictions from that model to train the final MLLM with the ability to cope with negative passages taken from the same articles and from different articles.

Automatic Data Construction. Most datasets for knowledge-based VQA do not provide human-labeled annotations of ground-truth passages. Therefore, we employ a proprietary LLM to automatically annotate positive and negative passages. Given a query image I, a question q and the set of passages from the asso-

†Depending on the test case, see Sec. 5.3.6 for details.

ciated article $\mathbf{P}$, we caption the image I with three captioning models (*i.e.* LLaVA-v1.5 [177], BLIP-2 [158], and InstructBLIP [70]) and prompt the LLM to assess whether a passage $s \in \mathbf{P}$ can answer q given the textual description of I. To help identify positive passages, for each sample, we select the two passages that have the highest similarity with their respective question according to the Contriever embedding space [122]. We also ensure to have at least one positive and one negative passage for each (I, q) sample in the dataset.

Model Training. Having collected positive and negative passages, the in-article reflective model is trained on a mixture of samples associated with positive and negative passages, using sequences in the form

$$\begin{aligned} &[I, q, \texttt{<RET>}, s, \texttt{<REL>}, y] \text{ and} \\ &[I, q, \texttt{<RET>}, \tilde{s}, \texttt{<NOREL>}, y], \end{aligned} \tag{5.9}$$

where s refers to a passage predicted as positive, and $\tilde{s}$ to a negative passage from the same page. The model is trained using a time-wise cross-entropy loss over the reflective tokens and over the answer.

5.3.5 Training the Overall Model

In the second stage, predictions from the in-article reflective model are employed to construct the dataset for training the overall model. The capabilities of the in-article model are indeed used to automatically annotate textual passages from existing datasets that require external knowledge. This is also complemented by negative passages taken from other pages of the knowledge base, plus samples that do not need an external knowledge base. The ultimate result of this stage is an MLLM capable of both answering questions and generating special tokens to assess whether additional information retrieval is necessary and whether it would be beneficial for answering.

Data Curation. The training split of the Encyclopedic-VQA [194] and InfoSeek [57] datasets are employed to create the data collection to train the second stage. Each sample (I, q) is expanded with three distinct passages: a positive, a

hard negative coming from the ground-truth page, and a soft negative coming from a different page.

To construct the first two cases, each sample is processed by the in-article reflective model to label each passage as either relevant or not relevant. After this, the passage with the highest probability of containing the <REL> token is taken as positive. Instead, the hard negative sample is randomly chosen from one of the sections predicted as <NOREL>. For the soft negative case, the image I is used to retrieve inside the top-1 page, excluding the ground-truth page. From the retrieved page, then, a random passage is considered a soft negative. Additionally, data from the LLaVA-Instruct dataset [177] are also included. These samples are labeled as cases where no retrieval is necessary.

Model Training. Finally, the model is trained using a balanced mixture of sequences in the form

$$\begin{aligned} &[I, q, \texttt{<NORET>}, y], \\ &[I, q, \texttt{<RET>}, s, \texttt{<REL>}, y], \\ &[I, q, \texttt{<RET>}, \tilde{s}, \texttt{<NOREL>}, y] \text{ and} \\ &[I, q, \texttt{<RET>}, \bar{s}, \texttt{<NOREL>}, y], \end{aligned} \tag{5.10}$$

where $\bar{s}$ refers to a soft-negative passage. We then employ a time-wise cross-entropy loss over all ground-truth tokens, with the exception of those related to I, q, and retrieved passages $s, \tilde{s}, \bar{s}$.

5.3.6 Experimental Setup

Datasets. Our experiments are conducted on Encyclopedic-VQA [194] and InfoSeek [57], which contain question-answer pairs linked to documents from an external knowledge base (*e.g.* Wikipedia). Encyclopedic-VQA consists of 221k pairs associated with 16.7k fine-grained entities (*i.e.* Wikipedia pages). Questions are divided into single-hop and two-hop types: the former indicates that a single Wikipedia page is required to answer them, while the latter requires a sequential retrieval process across multiple documents. Dataset samples are split into

training, validation, and test sets with 1M, 13.6k, and 5.8k items respectively. Experiments are reported on the test set, where single-hop questions correspond to 4.8k samples. The InfoSeek dataset, instead, contains 1.3M image-question pairs associated with around 11k Wikipedia pages. The dataset comprises 934k training, 73k validation, and 348k test samples. Following existing literature [3, 310], experimental results are reported on the validation set which includes questions not contained in the training split and questions associated with unseen entities.

External Knowledge Bases. Both datasets come with an external knowledge base composed of Wikipedia documents. In particular, Encyclopedic-VQA contains a knowledge base of 2M Wikipedia pages. Each page includes the Wikipedia title, the corresponding textual sections, and associated images. InfoSeek, instead, provides a knowledge base composed of 6M Wikipedia entities. In our experiments, we use the original 2M knowledge base for Encyclopedic-VQA, while we extract a subset of 100k pages[‡] from the original 6M for InfoSeek, following recent works [3, 310].

Evaluation Metrics. We follow the evaluation protocol provided along with the datasets. Generated answers for Encyclopedic-VQA are evaluated according to the BERT matching score (BEM) [36] between predicted and ground-truth answers. Instead, when evaluating answers for image-question pairs from InfoSeek, we use VQA accuracy [95] and relaxed accuracy [195] depending on the question type.

Architectural and Training Details. Our model is based on the LLaVA-v1.5 MLLM [177] with LLaMA-3.1-8B [84] as language model[§]. It employs CLIP ViT-L/14@336 as visual encoder and an MLP as the vision-to-language connector.

For both training phases, we fine-tune the LLaVA architecture to learn how to generate the introduced reflective tokens. To do that, we modify the original vocabulary of the LLaMA-3.1 LLM, replacing the final four reserved special tokens with our custom tokens. During training, we employ a learning rate of 2×10^{-5} and a global batch size of 128, updating the weights of both the MLP and LLM.

[‡]The knowledge base used for InfoSeek contains the same entities as [3].

[§]https://huggingface.co/aimagelab/LLaVA_MORE-llama_3_1-8B-finetuning

		E-VQA			InfoSeek		
Model	**Ret. Mode**	R@1	R@5	R@20	R@1	R@5	R@20
CLIP ViT-L/14	Textual (T)	3.4	8.7	14.0	36.9	59.9	71.9
CLIP ViT-L/14	Textual (T+S)	0.7	2.3	6.4	19.3	40.9	57.1
CLIP ViT-L/14	Visual	9.9	22.0	31.7	22.5	40.4	44.1
EVA-CLIP-8B	Textual (T)	7.5	15.2	20.7	52.5	71.2	79.2
EVA-CLIP-8B	Textual (T+S)	10.1	20.5	29.4	**56.1**	**77.6**	**86.4**
EVA-CLIP-8B	Visual	**15.6**	**36.1**	**49.8**	29.6	41.4	46.6

Table 5.5: Retrieval performance on Encyclopedic-VQA test set and the InfoSeek validation set. "Textual" refers to image-to-text retrieval using either title only (T) or title with summary (T+S), and "Visual" corresponds to image-to-image retrieval.

Training Data Collection. We train our model on a collection of data from different sources. For both training phases, we include samples from the Encyclopedic-VQA and InfoSeek training splits, as well as data from LLaVA-Instruct [177] to retain the generative capabilities of the MLLM. To automatically annotate positive and negative passages used to train the in-article reflective model, we employ GPT-4, prompted with the query question, the textual description of the associated image, the corresponding answer, and the passage to annotate. We also employ few-shot examples in the prompt to facilitate the annotation task. Additional details on the training data mixture are reported in the supplementary material C.2.

Coarse-Grained Retrieval Details. To identify the set of documents most relevant to the query image and associated question, we evaluate two CLIP-based retrieval models, namely CLIP ViT-L/14@336 [225] and EVA-CLIP-8B [259]. We explore two retrieval configurations for both models: (i) image-to-text retrieval, which computes similarity between the query image and document metadata (either the title alone or the title with the summary of the page), and (ii) image-to-image retrieval, which assesses similarity between the query image and images within Wikipedia documents.

Retrieval results for each variant are detailed in Table 5.5. Notably, EVA-CLIP demonstrates superior results across all configurations. However, the optimal retrieval configuration varies between datasets. Specifically, image-to-image retrieval yields the highest accuracy on Encyclopedic-VQA, while image-to-text retrieval proves the most effective for InfoSeek. This behavior can be attributed to

Model	LLM	Retrieval Mode		E-VQA		InfoSeek		
				Single-Hop	All	Unseen-Q	Unseen-E	All
Zero-shot LLMs								
Vanilla	Vicuna-7B	-		2.1	2.0	0.3	0.0	0.0
Vanilla	LLaMA-3-8B	-		16.3	17.3	1.5	0.0	0.0
Vanilla	LLaMA-3.1-8B	-		16.5	16.6	2.1	0.0	0.0
Zero-shot MLLMs								
BLIP-2 [158]	Flan-T5$_{XL}$	-		12.6	12.4	12.7	12.3	12.5
InstructBLIP [70]	Flan-T5$_{XL}$	-		11.9	12.0	8.9	7.4	8.1
LLaVA-v1.5 [177]	Vicuna-7B	-		16.3	16.9	9.6	9.4	9.5
LLaVA-v1.5 [177]	LLaMA-3.1-8B	-		16.0	16.9	8.3	8.9	7.8
Retrieval-Augmented Models								
DPR$_{V+T}$ [148]	Multi-passage BERT	CLIP ViT-B/32	Visual+Textual	29.1	-	-	-	12.4
RORA-VLM [221]	Vicuna-7B	CLIP+Google Search	Visual+Textual	-	20.3	25.1	27.3	-
Wiki-LLaVA [3]	Vicuna-7B	CLIP ViT-L/14+Contriever	Textual	17.7	20.3	30.1	27.8	28.9
Wiki-LLaVA [3]$^{\diamond}$	LLaMA-3.1-8B	CLIP ViT-L/14+Contriever	Textual	18.3	19.6	28.6	25.7	27.1
EchoSight [310]	Mistral-7B/LLaMA-3-8B	EVA-CLIP-8B	Visual	19.4	-	-	-	27.7
EchoSight [310]$^{\diamond}$	LLaMA-3.1-8B	EVA-CLIP-8B	Textual	22.4	21.7	30.0	30.7	30.4
EchoSight [310]$^{\diamond}$	LLaMA-3.1-8B	EVA-CLIP-8B	Visual	26.4	24.9	18.0	19.8	18.8
ReflectiVA (Ours)	LLaMA-3.1-8B	CLIP ViT-L/14	Textual	24.9	26.7	34.5	32.9	33.7
ReflectiVA (Ours)	LLaMA-3.1-8B	EVA-CLIP-8B	Textual	28.0	29.2	**40.4**	**39.8**	**40.1**
ReflectiVA (Ours)	LLaMA-3.1-8B	EVA-CLIP-8B	Visual	**35.5**	**35.5**	28.6	28.1	28.3

Table 5.6: VQA accuracy scores on the Encyclopedic-VQA test set and the InfoSeek validation set, where all results from retrieval-augmented models are reported without considering any re-ranking stage to reorder retrieved documents. Gray color indicates results that are not directly comparable due to different knowledge bases, and the marker $\diamond$ represents our reproductions with different LLMs.

the distinct characteristics and structural composition of each dataset and their respective knowledge bases. Consequently, unless specified otherwise, we adopt image-to-image retrieval for Encyclopedic-VQA and image-to-text retrieval (using title-only for CLIP ViT-L and title with summary for EVA-CLIP) for InfoSeek, with the number k of retrieved documents equal to 5.

5.3.7 Comparison with the State of the Art

Results on Encyclopedic-VQA and InfoSeek. We evaluate our model on the aforementioned datasets, comparing it to various zero-shot LLMs, MLLMs, and retrieval-augmented competitors. Specifically, we report results of three LLMs – Vicuna [61], LLaMA-3, and LLaMA-3.1 [84] – each prompted with both the query question and a description of the query image generated by an image captioning model. Additionally, we assess the performance of BLIP-2 [158], InstructBLIP [70], and LLaVA-v1.5 [177], without external retrieval augmentation and using only the query image and question as input. As direct competitors, we include DPR [148], RORA-VLM [221], Wiki-LLaVA [3], and EchoSight [310], which all leverage external knowledge retrieval. To ensure a fair

Model	LLM	E-VQA	
		Single-Hop	All
Textual Retrieval Mode			
EchoSight [310]♢	LLaMA-3.1-8B	26.8	26.0
ReflectiVA (Ours)	LLaMA-3.1-8B	**33.6**	**33.9**
Visual Retrieval Mode			
EchoSight [310]	Mistral-7B/LLaMA-3-8B	**41.8**	-
EchoSight [310]♢	LLaMA-3.1-8B	36.3	34.2
ReflectiVA (Ours)	LLaMA-3.1-8B	40.6	**39.7**

Table 5.7: VQA accuracy scores on Encyclopedic-VQA when models are equipped with a document re-ranking component. Gray color indicates results that are not directly comparable due to different knowledge bases, and the marker ♢ represents our reproductions.

comparison with the considered methods, for both Wiki-LLaVA and EchoSight we develop a variant based on LLaMA-3.1, employing the same knowledge bases used in our solution.

Results are shown in Table 5.6, in which we report for each retrieval-augmented model the details of the retrieval pipeline used (*i.e.* the retrieval model and modality). As it can be seen, both zero-shot LLMs and MLLMs fail to correctly answer the given questions due to the lack of external knowledge during the generation. This is particularly evident on the InfoSeek dataset where LLMs exhibit accuracy scores close to zero, highlighting the need for the visual inputs to generate correct answers for this dataset. For knowledge-based models, the proposed ReflectiVA exhibits state-of-the-art results on both the Encyclopedic-VQA and InfoSeek datasets, outperforming all evaluated competitors by a substantial margin and highlighting the benefits of employing reflective tokens for the task.

Integrating a Re-Ranking Stage. Recent studies focusing on text-only LLMs [229, 325, 310] have shown that incorporating a re-ranking stage within a retrieval-augmented generation pipeline can enhance performance. Following this approach, we evaluate our model with the re-ranking component proposed in [310] which reorders retrieved textual passages prior to model input. Specifically, in our experiments, we first retrieve the top-k relevant Wikipedia pages from the knowledge base using the retrieval model previously described. The retrieved passages are then processed by the re-ranking component, and we input the top-k_p re-ranked passages into our model to assess their relevance to the query and

Model	LLM	E-VQA	InfoSeek		
		Single-Hop	Un-Q	Un-E	All
KB Article					
Vanilla	Vicuna-7B	34.1	5.3	4.3	4.7
Vanilla	LLaMA-3-8B	72.9	10.0	7.9	8.8
Vanilla	LLaMA-3.1-8B	73.6	15.2	13.9	14.5
LLaVA-v1.5 [177]	Vicuna-7B	42.9	14.2	13.4	13.8
LLaVA-v1.5 [177]	LLaMA-3.1-8B	54.1	20.1	17.7	18.8
KB Passages					
Wiki-LLaVA [3]	Vicuna-7B	38.5	52.7	50.3	51.5
Wiki-LLaVA [3]$^{\diamond}$	LLaMA-3.1-8B	46.8	51.2	50.6	50.9
ReflectiVA (Ours)	LLaMA-3.1-8B	**75.2**	**57.8**	**57.4**	**57.6**

Table 5.8: VQA accuracy scores on Encyclopedic-VQA and InfoSeek with oracle Wikipedia entities. The top of the table shows results using the full Wikipedia article as input to the LLM/M-LLM, while the bottom of the table shows performance using model-specific strategies to identify relevant text passages.

	k	E-VQA	InfoSeek		
		Single-Hop	Un-Q	Un-E	All
Effectiveness of Reflective Tokens and Training					
In-Article Reflective Model	5	21.1	25.5	23.8	24.6
ReflectiVA (Overall Model)	5	**35.5**	**40.4**	**39.8**	**40.1**
always w/ <RET> token	5	35.3	40.2	**39.8**	40.0
w/o <REL>/<NOREL> tokens	5	23.6	32.2	30.6	31.4
w/o KB (always with <NORET> token)	-	21.3	17.7	15.3	16.4
Varying the Number of Retrieved Documents					
	1	29.0	**40.6**	**41.0**	**40.8**
ReflectiVA (Overall Model)	5	35.5	40.4	39.8	40.1
	10	**36.0**	37.2	37.0	37.1
	20	35.7	30.6	31.3	30.9

Table 5.9: Ablation study results demonstrating the effectiveness of the proposed reflective tokens and training strategy, along with the impact of different numbers of retrieved documents.

generate a response.

Results are shown in Table 5.7 comparing our results to those of EchoSight on the Encyclopedic-VQA dataset[¶]. As it can be seen, performing a re-ranking step can further improve the results of our model which achieves, in its best configuration, 40.6 accuracy points compared to 35.5 without re-ranking. These results consistently outperform those obtained by EchoSight, with the same LLM and knowledge base used in our setting, in both retrieval configurations.

Results using Oracle Documents. To thoroughly evaluate the perfor-

[¶]For this setting, we do not include results on InfoSeek since the re-ranker proposed in [310] is trained on samples from Encyclopedic-VQA.

Model	LLM	ViQuAE F1	ViQuAE EM	S3VQA GPT-4
LLaVA-v1.5 [177]	Vicuna-7B	15.1	26.6	23.9
LLaVA-v1.5 [177]	LLaMA-3.1-8B	15.0	25.6	24.4
Wiki-LLaVA (E-VQA) [3]♢	LLaMA-3.1-8B	10.5	16.7	22.7
Wiki-LLaVA (InfoSeek) [3]♢	LLaMA-3.1-8B	12.7	21.8	21.8
ReflectiVA (w/o KB)	LLaMA-3.1-8B	16.6	27.6	26.9
ReflectiVA (Ours)	LLaMA-3.1-8B	**23.2** (52.0%)	**38.1**	**29.3** (16.8%)

Table 5.10: Zero-shot performance on additional knowledge-based VQA datasets. The percentage of samples in which our model incorporates external knowledge is highlighted in green.

mance of our model, we conduct experiments under an oracle setting, where the ground-truth entity (*i.e.* the Wikipedia page associated with the query) is provided. In this configuration, all text passages from the oracle entity are input to ReflectiVA, which then selects the relevant passages before generating an answer. Table 5.8 presents the results of this analysis, directly comparing ReflectiVA with Wiki-LLaVA, which leverages a Contriever model [122] to retrieve the most relevant passages within the oracle document. We also report the performance of standard LLMs and MLLMs when prompted with the entire Wikipedia article. Notably, ReflectiVA achieves the highest performance across both the Encyclopedic-VQA and InfoSeek benchmarks, surpassing Wiki-LLaVA and standard models, further highlighting its effectiveness in isolating the passages most pertinent to the given image-question pair. It is also noteworthy that while vanilla LLMs achieve high accuracy on Encyclopedic-VQA when prompted with the entire oracle Wikipedia page, in the case of InfoSeek, employing a strategy to select the most relevant passages leads to significantly better performance, with ReflectiVA always reaching the best results.

Qualitative Results. Fig. 5.6 provides a qualitative comparison on sample image-question pairs from Encyclopedic-VQA (top row) and InfoSeek (bottom row).

5.3.8 Ablation Studies and Analyses

Effectiveness of Two-Stage Training. We analyze the impact of our two-stage, two-model training strategy. Specifically, we assess the effectiveness of the

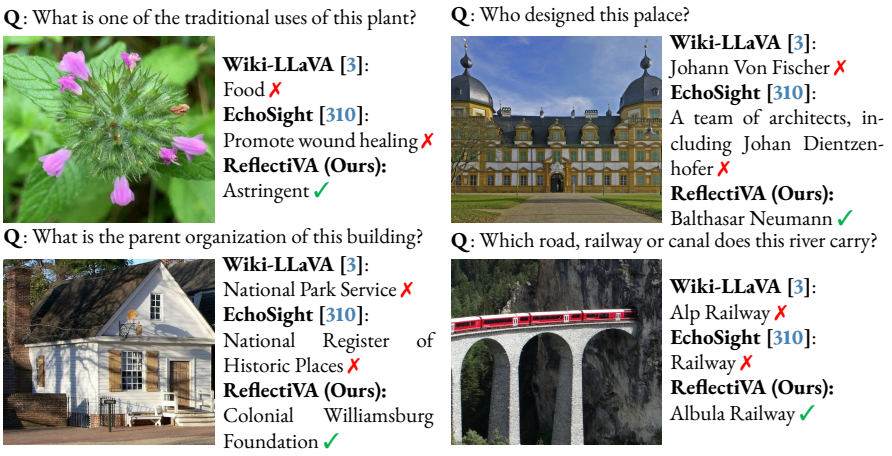


Figure 5.6: Sample qualitative results on image-question pairs from Encyclopedic-VQA (top row) and InfoSeek (bottom row), where we compare the answers provided by ReflectiVA with those from WikiLLaVA and EchoSight.

Model	LLM	MMMU	MMB (EN)	POPE	SEED-Img	MME (P)	MME (C)	GQA	TextVQA	Science-QA	AI2D
LLaVA-v1.5 [177]	Vicuna-7B	34.2	65.3	85.6	66.8	1474.3	314.6	62.4	58.2	69.0	56.4
LLaVA-v1.5 [177]	LLaMA-3.1-8B	39.4	72.4	85.1	69.8	1531.5	353.3	63.6	58.4	76.3	61.8
Wiki-LLaVA (E-VQA) [3]	Vicuna-7B	36.6	70.4	86.6	-	1170.1	290.0	-	-	-	-
Wiki-LLaVA (InfoSeek) [3]	Vicuna-7B	35.6	71.1	84.2	-	1438.9	341.3	-	-	-	-
Wiki-LLaVA (E-VQA) [3]$^{\diamond}$	LLaMA-3.1-8B	32.2	60.9	84.6	59.2	1350.7	306.8	56.6	49.1	67.5	55.1
Wiki-LLaVA (InfoSeek) [3]$^{\diamond}$	LLaMA-3.1-8B	35.9	52.0	85.7	60.5	1417.8	349.6	58.6	50.1	69.1	54.3
ReflectiVA (Ours)	LLaMA-3.1-8B	38.9	69.9	85.1	68.6	1564.5	355.7	62.1	56.8	75.4	60.6

Table 5.11: Performance preservation analysis on standard benchmarks for MLLM evaluation and traditional VQA datasets.

in-article reflective model, which is trained using both positive and negative passages drawn from the same Wikipedia pages, on Encyclopedic-VQA and InfoSeek. Table 5.9 (top) demonstrates that training only the first model alone does not yield high accuracy scores. Notably, the complete model (*i.e.* ReflectiVA) consistently achieves higher accuracy.

Effectiveness of Reflective Tokens. To evaluate the effectiveness of the proposed reflective tokens, we conduct ablation studies that isolate the contributions of both retrieval and relevance tokens. We design three variants of the inference pipeline: (i) without retrieval from the knowledge base (*i.e.* always enforcing `<NORET>` during generation), (ii) without assessing the relevance of retrieved passages, where two random passages are selected from each of the top-5 documents, and (iii) always enforcing the `<RET>` token while using the standard

<REL>/<NOREL> pipeline. Results are summarized in the top part of Table 5.9 for Encyclopedic-VQA and InfoSeek. As shown, consistently performing retrieval causes only a minor performance degradation, given that both datasets are designed to require external knowledge. In contrast, omitting the proposed relevance tokens leads to a substantial accuracy drop (*e.g.* from 35.5 to 23.6 on Encyclopedic-VQA and from 40.1 to 31.4 on InfoSeek), underscoring the critical role of identifying relevant passages before answer generation. Furthermore, the lowest scores occur when retrieval from the external knowledge base is entirely bypassed, emphasizing the need for retrieval.

Varying the Number of Retrieved Documents. In Table 5.9 (bottom), we further analyze the effect of varying the number k of retrieved documents. While setting k equal to 10 achieves optimal results on Encyclopedic-VQA, using only the top-1 retrieved document generally performs better for InfoSeek. This discrepancy can be attributed to differences in the knowledge base sizes (*i.e.* significantly larger for Encyclopedic-VQA) and retrieval performance across the two datasets (cf. Table 5.5). Overall, using the top-5 retrieved documents provides the best trade-off across both datasets, leading us to set k equal to 5 in our experiments.

Results on Other Knowledge-based Datasets. We also validate the generalization capabilities of ReflectiVA to zero-shot settings which always require knowledge retrieval. Specifically, we report the results on two additional knowledge-based VQA datasets, *i.e.* ViQuAE [149] and S3VQA [125]. For the former, we include performance in terms of F1 score and exact matching, while for the latter we ask GPT-4 to compute a score between 0 and 100 of the correctness of the given answers, following [42, 300]. From Table 5.10, it can be seen that even in these challenging settings, ReflectiVA achieves the best performance, demonstrating the usefulness of predicting reflective tokens.

Performance Preservation on Standard Benchmarks. Finally, we evaluate whether the proposed approach impacts performance on standard MLLM and VQA benchmarks that do not require external knowledge. In Table 5.11, we compare ReflectiVA with Wiki-LLaVA, which fine-tunes a LLaVA

model for similar purposes, and include results from the original LLaVA-v1.5 model, tested with both Vicuna-7B and LLaMA-3.1. ReflectiVA incurs only a minor performance reduction relative to the original LLaVA model, while significantly outperforming Wiki-LLaVA.

5.3.9 Conclusion

We proposed ReflectiVA, a multimodal LLM with retrieval-augmented generation. Our method employs reflective tokens, trained in a two-stage two-model pipeline. Extensive experiments, conducted on both VQA datasets requiring external knowledge and standard datasets, demonstrate the efficacy of the proposed solution.

5.4 Discussion

The results presented in this chapter highlight both the potential and the limitations of retrieval-augmented generation in multimodal large language models.

A first key observation emerging from this chapter is that extending multimodal language models beyond purely parametric knowledge is necessary for addressing knowledge-intensive visual reasoning tasks. Hierarchical retrieval, as demonstrated by Wiki-LLaVA, enables multimodal models to ground their answers in external documents without modifying the underlying generation architecture. This confirms that retrieval is an effective mechanism for overcoming the limitations of purely parametric knowledge in MLLMs.

However, the experiments also reveal retrieval as a major bottleneck. Entity-level retrieval quality degrades rapidly as the knowledge base scales, and retrieval errors directly propagate to the generation stage by injecting irrelevant or misleading context. These findings show that naïvely expanding the input context is insufficient, and that retrieval noise can be as harmful as missing knowledge.

The second contribution of this chapter addresses this limitation by making retrieval an explicit object of reasoning. By introducing self-reflective tokens, Reflective LLaVA enables multimodal models to decide when external knowledge is required and to assess the relevance of retrieved evidence before generation. This

transforms retrieval from a fixed pre-processing step into a controllable component of multimodal reasoning, significantly improving robustness and factual grounding.

Taken together, these results suggest a shift in how multimodal RAG systems should be designed. Rather than treating retrieval as an external utility, effective multimodal reasoning requires models to reason about knowledge access itself, balancing the benefits of external information against the risks of noise and distraction. This perspective positions retrieval awareness as a central design principle for scalable and reliable multimodal intelligence.

5.5 Chapter Summary

This chapter investigated retrieval-augmented generation as a key mechanism for overcoming the limitations of purely parametric multimodal language models. We showed that while naïve multimodal RAG pipelines can substantially improve factual accuracy on knowledge-intensive tasks, they remain brittle at scale due to imperfect retrieval and unfiltered context injection.

We introduced Wiki-LLaVA as a hierarchical retrieval framework that enables multimodal models to access external knowledge without architectural changes, and ReflectiVA as a retrieval-aware paradigm in which models explicitly reason about when and how external evidence should be used.

Taken together, these contributions identify retrieval-aware reasoning as a necessary step toward scalable, factual, and robust multimodal intelligence. While this chapter focused on grounding multimodal models through external knowledge access, the next chapter addresses an orthogonal limitation of current systems: their difficulty in reasoning compositionally over visual and linguistic structures.

6

Causal Compositionality for Vision–Language Reasoning

Compositional reasoning represents a fundamental challenge for vision–language models that goes beyond both perceptual grounding and factual knowledge access. Even when visual content is correctly perceived and external information is available, models may still fail to understand *who does what to whom*, confusing agents, attributes, and relations. Such failures expose a deeper limitation in how vision–language models represent and combine linguistic structure with visual evidence.

The previous chapters addressed complementary aspects of multimodal reasoning. Image captioning models were studied through the lens of training objectives and evaluation metrics, highlighting the role of optimization in shaping linguistic quality. Multimodal large language models were then analyzed in terms of architectural alignment and scaling, revealing the strengths and limits of para-

This Chapter is related to the publication "F. Parascandolo *et al.*, Causal graphical models for vision-language compositional understanding, ICLR 2025" [11]. See the list of Publications on page 149 for more details.

metric multimodal representations. Retrieval-augmented generation further extended this analysis by addressing the limits of parametric knowledge through external memory access. However, none of these approaches directly confronts the problem of *compositional structure* itself.

This chapter focuses on a different and orthogonal dimension of multimodal intelligence: the ability to model and exploit the causal and syntactic relationships that govern language. Specifically, we investigate whether explicitly encoding linguistic structure into vision–language models can improve compositional generalization, independently of additional data curation or retrieval mechanisms. To this end, we introduce a causal perspective on vision–language modeling, in which dependency relations between words are treated as causal constraints that guide both training and inference. We argue that the absence of such structured inductive biases constitutes a fundamental bottleneck for robust multimodal reasoning, independent of scale or data availability.

6.1 Approach Overview

In this chapter we investigate a principled approach to vision–language compositionality grounded in *Causal Graphical Models* (CGMs). Rather than relying on extra annotated data, our method imposes structure on the captioning process itself. For each caption, we obtain a *dependency tree* using an off-the-shelf parser; its nodes correspond to words and its edges capture head–dependent relations. We combine this tree with the image's visual features to construct a CGM that specifies how textual tokens and image patches causally relate. The joint probability of a caption is then factorized along the tree, ensuring that each token is predicted only from its ancestors and the relevant visual context. This factorization discourages the model from learning spurious correlations induced by sequential order and emphasises genuine syntactic dependencies.

An example of the dependency tree and the corresponding causal graphical model is shown in Fig. 6.1.

To realise this causal factorization, we replace the text encoder of a pre-trained VLM with a small decoder trained under a *causally ordered generative training*

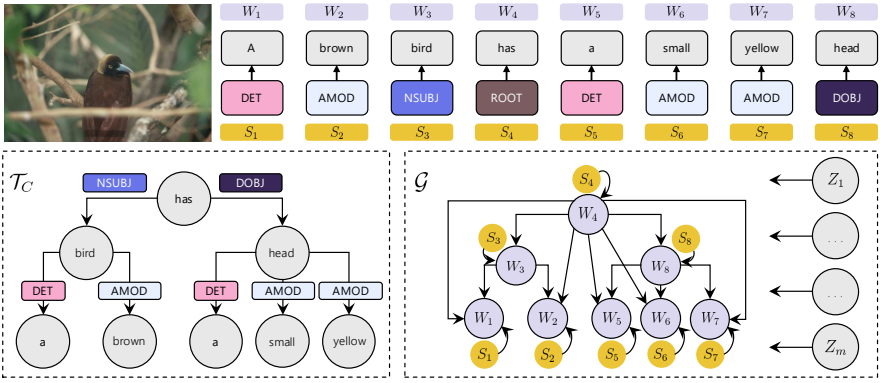


Figure 6.1: Dependency relations between words in a sentence. On the left, the DT ($\mathcal{T}$) extracted from the caption shown above using a dependency parser [82]. On the right, the corresponding CGM ($\mathcal{G}$). To improve readability, in $\mathcal{G}$ we use different colors for different variable types and we omit the causal dependencies between visual (Z) and textual (W) variables.

(COGT) regime. At each step, the decoder operates on two embeddings for every word: a *masked token* indicating the word's syntactic category and a *visible token* representing the word itself. A dependency-guided attention mechanism restricts the masked token's attention to its parents in the dependency tree, while cross-attention allows both token types to incorporate visual features. During training, we maximise the log-likelihood of this causally factorized distribution, which encourages the model to learn only those conditional dependencies specified by the CGM and to ignore irrelevant co-occurrences.

Inference follows the same structured ordering: candidate captions are scored by generating tokens in a semi-parallel fashion according to their depth in the dependency tree, with words at the same depth predicted simultaneously while conditioning on their ancestors. This semi-parallel generation harnesses the CGM to evaluate captions without biasing predictions toward any particular linearisation. Extensive experiments demonstrate that this approach substantially improves compositional reasoning on multiple benchmarks compared to existing methods, even when trained on substantially less data. By embedding causal structure into the generation process, the proposed model advances the understanding of compositionality in vision–language models without requiring curated training datasets.

6.2 Causal Modeling for Vision–Language Compositional Reasoning

6.2.1 Problem Introduction

Vision-Language Models (VLMs) have shown impressive results in different tasks such as zero-shot classification, image-text retrieval, vision-question answering, image-captioning, and many others [225, 158, 251, 179]. However, despite this success, most VLMs still struggle to understand the compositional nature of the human language. For instance, Yuksekgonul et al. [327] empirically showed that common VLMs usually do not consider the order and the syntactic/semantic relations of words in a sentence, which is effectively treated as a *bag of words* representation, where "the horse is eating the grass" and "the grass is eating the horse" can easily be confused. Similar observations have been reported in several recent studies on vision–language compositionality. For example, Hsieh et al. [110] and Zhao et al. [347] show that state-of-the-art VLMs often fail when object attributes, relations, or word order are perturbed, while Burapacheep et al. [37] demonstrate that models frequently confuse attributes associated with different objects. These results suggest that many VLMs rely on shallow correlations between words and visual features rather than on structured reasoning about linguistic relations. These findings have motivated the introduction of several compositional benchmarks designed to systematically probe such failures in VLMs [110, 347, 37]. One contributing factor behind this bag-of-words behavior is the contrastive loss used in CLIP [225] (and in other VLMs), which compares a single vector representing the textual encoder's output with a single vector representing the visual encoder's output, sacrificing textual and visual details [327, 132, 30]. Another contributing factor may be the limited descriptive detail of many captions used for VLM pre-training, which are usually noisy or do not describe the details of the image and the interactions among its objects [80].

Most of the compositional methods that have been recently proposed to alleviate this problem focus on creating annotations with a richer compositional structure, used to fine-tune a VLM [327, 80, 40]. While these approaches improve

compositional robustness, they remain largely data-centric: they attempt to correct compositional failures by exposing the model to additional examples rather than by modifying the underlying generative structure of the model. As a result, they require carefully curated training data and may not generalize beyond the compositional patterns observed during training. For instance, NegCLIP [327] creates *hard negatives*, in which the original caption is modified swapping the positions of some words, and these hard negatives are used jointly with common negatives to fine-tune CLIP using the standard contrastive loss. However, the automatic creation of hard negatives is itself noisy, leading to captions which often do not have a correct syntactic/semantic meaning (this problem is inherited by some compositional benchmarks, see 6.2.4). In [274], a VLM is pre-trained from scratch using a captioning strategy and a huge private dataset. Specifically, the authors propose both Cap, where the pre-training strategy is a standard *AutoRegressive* (AR) next-token prediction, and CapPa, where the AR training is mixed with a *parallel* training, in which all the textual tokens are simultaneously predicted. Tschannen et al. [274] show that both Cap and CapPa achieve excellent results on compositional tasks, and argue that a *generative training* encourages the VLM to focus on fine-grained descriptions of the visual content.

To build intuition, consider the caption "A brown bird has a small yellow head". In a standard left-to-right autoregressive captioning model, the adjective "brown" must be predicted before the noun "bird" has been generated. This makes the prediction ambiguous, since multiple objects in the image may share the same attribute. Ideally, however, the generation of an attribute should depend on the entity it describes. Dependency structures naturally capture this relationship, motivating the use of dependency trees to guide caption generation.

In this chapter, inspired by Cap and CapPa, we propose a VLM adaptation approach for compositional reasoning which is based on a decoder trained with a captioning strategy. However, unlike the standard fully-sequential AR and the parallel predictions used in [274], we propose a partially ordered, semi-parallel AR prediction strategy which is guided by the dependency relations of a *Causal Graphical Model* (CGM) [241]. In more detail, we use an off-the-shelf *dependency parser* [82], which creates a syntactic tree from a given textual sentence.

Specifically, given a caption, a dependency parser automatically builds a *Dependency Tree* (DT), in which each node is associated with a caption word and each edge represents a syntactic dependency relation between two words (see Fig. 6.1). The DT, jointly with the visual features extracted from the image using a frozen visual encoder, are used to build a CGM, which describes the dependency relations among image patches and textual tokens. Our token prediction strategy is based on the dependency relations contained in this CGM. Using the example in Fig. 6.1, the adjective "brown" depends on the noun "bird". In contrast, using a standard AR approach, where the token prediction order follows the English grammar, the captioning model should predict "brown" before knowing that this adjective refers to "bird", which is a quite ambiguous task, since many objects may be brown in the image. Conversely, when our model predicts the adjective ("brown"), it knows the noun ("bird") it refers to, thus the word generation can be specific to the entities, the attributes and the relations contained in the input image. Generally speaking, we factorize the joint distribution of all the caption words following the *disentangled factorization* of a CGM [241], and our semi-parallel AR model predicts a token conditioned only on the tokens on which it depends. For instance, in the example of 6.1, "small" and "yellow" are predicted in parallel and they are conditionally independent given "head", thus no statistical dependence is learned between these two words. The advantage of this strategy is that the decoder can focus on learning only the main causal dependency relations, ignoring possible spurious associations [212] induced by the sequential order of the words in a natural language sentence. Moreover, we use the same prediction strategy also at inference time, when we compute the likelihood of a candidate caption. In this case too, the use of the CGM makes the likelihood estimation independent of spurious associations due to the sequential order of the words.

We validate our method using different VLMs (CLIP, XVLM [329] and InstructBLIP [70]). Using extensive experiments with five compositional datasets, we show that our approach largely outperforms all previous works, setting a new *state of the art* in all the evaluated benchmarks, and that it also improves on Cap and CapPa, despite being trained on much less data. In the following, we formal-

ize this limitation from the perspective of vision–language modeling and review existing compositional benchmarks and training strategies, before introducing a causal formulation that directly constrains the generative process.

6.2.2 Method

Given an image-caption pair (X, C), our goal is to define a set of conditional distributions over the random variables associated with the image features and the caption words. For this purpose, as anticipated in 6.2.1 and 2.3, we use an off-the-shelf dependency parser [82] which, for a specific $C = [w_1, ..., w_n]$, returns a DT $\mathcal{T}^*$ (6.1), where each node corresponds to a word and each edge (i, j) connects the "dependent" word w_j with its "head" w_i (2.3). $\mathcal{T}$ contains the syntactic and semantic dependencies between the words in C [202], and we make this dependency explicit by connecting each word to all the words it transitively depends on in the tree. Specifically, we define a CGM $\mathcal{G}$ by associating each word w_j with a random variable W_j, corresponding to a node of $\mathcal{G}$. Moreover, we connect the node corresponding to W_j with all the variables corresponding to the ancestors of w_j in $\mathcal{T}$ (6.1). Formally, if $w_{i_1}, ..., w_{i_k}$ are the ancestors of w_j in $\mathcal{T}$, then we assume a causal dependence between the corresponding variables: $W_{i_1} \rightarrow W_j, ..., W_{i_k} \rightarrow W_j$. Furthermore, the parser labels each word in $\mathcal{T}$ with a syntactic type using a prefixed vocabulary V [250, 341]. For instance, if $type(w_j) = \texttt{nsubj} \in V$, it means that w_j is a noun and it plays the role of the subject in the sentence. Intuitively, we can think of these syntactic types as categorical syntactic features extracted from C, which we formally describe using n random variables $S_1, ..., S_n$, where each S_j ranges over V. In $\mathcal{G}$, we assume that each W_j depends on its corresponding syntactic variable S_j: $S_j \rightarrow W_j$.

Finally, we extract a set of features from X using the VLM visual encoder $\mathcal{E}$: $\mathcal{Z} = \mathcal{E}(X) = \{\boldsymbol{z}_1, ..., \boldsymbol{z}_m\}$ (details in 6.2.3), and, similarly to the textual case, we associate a random variable Z_k to each feature $\boldsymbol{z}_k \in \mathcal{Z}$. In $\mathcal{G}$, we assume that W_j depends on all the visual variables: $Z_1 \rightarrow W_j, ..., Z_m \rightarrow W_j$. Using the above assumptions, we define the *parents* [241] of W_j as: $\mathbf{PA}(W_j) =$

*Note that, for each C in the training/testing set, $\mathcal{T}$ needs to be extracted only once and can be done offline.

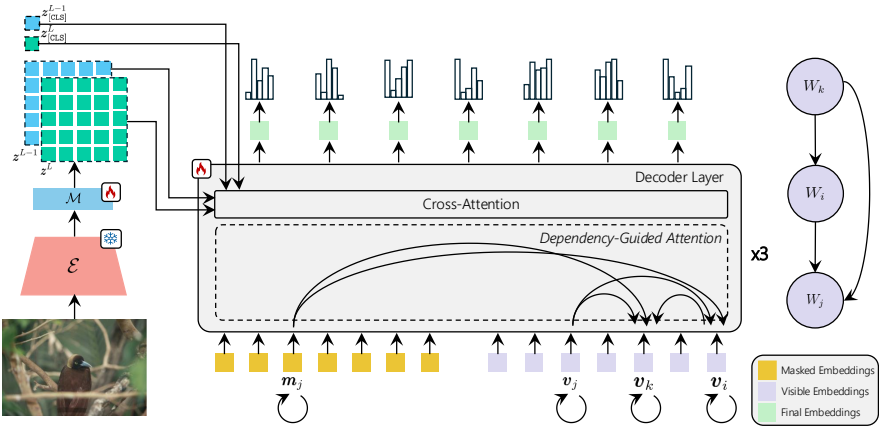


Figure 6.2: A schematic illustration of our decoder.

$\{W_{i_1}, ..., W_{i_k}, S_j, Z_1, ..., Z_m\}$, and we model the conditional joint distribution of the textual variables given the visual and the syntactic variables as:

$$P(W_1, ..., W_n | S_1, ..., S_n, Z_1, ..., Z_m) = \prod_{j=1}^{n} P(W_j | \mathbf{PA}(W_j)), \tag{6.1}$$

where the right side of 6.1 is obtained using the *disentangled factorization* of CGMs [241, 214] and assuming that $S_1, ..., S_n$ and $Z_1, ..., Z_m$ are independent of each other (see D.1.1 for more details). In 6.2.3 we show how a VLM can be adapted to predict this disentangled factorization both at training and at inference time.

Discussion. Tschannen et al. [274] formulate the joint distribution of the words in C using the standard AR prediction strategy commonly adopted by image captioning methods (6.2.1):

$$P(W_1, ..., W_n | Z_1, ..., Z_m) = \prod_{j=1}^{n} P(W_j | W_1, ..., W_{j-1}, Z_1, ..., Z_m). \tag{6.2}$$

The advantage of our formulation (6.1) over 6.2 is that, in our case, the model

needs to learn only the inter-variable conditional distributions indicated by the dependency parser, reducing the risk of overfitting on the training data [241]. Specifically, the dependency parser helps in discarding those spurious *associations* [211, 212] contained in 6.2 which depend on the sequence of words in C but do not correspond to a strict semantic/syntactic relation (*e.g.*, "small" and "yellow" in the example of 6.2.1). In contrast, we interpret the dependency relations extracted by a dependency parser as *causal relations* because they directly model the (linguistic) influence of the "head" variable with respect to the generation of the "dependent" variable. For instance, the probability values of an adjective are directly influenced by the noun it refers to, because the adjective describes an attribute *of that noun*, thus the corresponding conditional probability is not a spurious association.

While the causal dependency relations in C may not be exhaustively described by $\mathcal{G}$ and there may be other relations between words in C, we follow [93], and we assume that, in a symbolic domain like the natural language, the joint distribution over the words of a sentence should be *sparse*. This is also in line with very recent work which shows that sparse attention in Transformers helps the network focus on the most relevant context and improves its performance removing noise [151, 319]. Thus, we prefer sparseness to completeness and we assume that the word dependencies extracted by a dependency parser are causally sufficient (see 2.3). Finally, Tschannen et al. [274] propose also a parallel prediction strategy, which corresponds to:

$$P(W_1, ..., W_n | Z_1, ..., Z_m) = \prod_{j=1}^{n} P(W_j | Z_1, ..., Z_m). \tag{6.3}$$

In 6.3, each W_j is assumed to be conditionally independent from all the other textual variables given the visual variables. The empirical results reported in [274] do not show a clear winner between the AR and the parallel prediction, and the authors use a mixture of the two strategies in training their VLM (2.3). Conversely, in our experiments (6.2.5 and6.2.6) we show that our proposed disentangled factorization (6.1) is a better trade-off between the conditional independence of 6.3 and the standard image captioning factorization of 6.2, and it also improves over

the mixed strategy adopted in [274].

6.2.3 Using a decoder for causal prediction

In this section, we show how textual tokens can be generated using our CGM. Note that our goal is not image captioning, but we use our method, which we call Causally-Ordered Generative Training (COGT), for vision-language compositional understanding. Since CLIP is the most commonly adopted backbone by previous works on compositionality (2.3), in the following we use CLIP as an example VLM, and in 6.2.6 we show additional results obtained with other VLMs.

We freeze the CLIP visual encoder ($\mathcal{E}$) and, from a given image X, we extract a set of features from the last (L) and the penultimate ($L-1$) layer of $\mathcal{E}$: $\mathcal{Z} = \{\boldsymbol{z}^L_{[\text{CLS}]}, \boldsymbol{z}^L_1 \ldots, \boldsymbol{z}^L_p, \boldsymbol{z}^{L-1}_{[\text{CLS}]}, \boldsymbol{z}^{L-1}_1 \ldots, \boldsymbol{z}^{L-1}_p\}$. For the l-th layer of the encoder, $\boldsymbol{z}^l_{[\text{CLS}]}$ is the embedding vector of the class token [79], while $\boldsymbol{z}^l_1 \ldots, \boldsymbol{z}^l_p$ are the embedding vectors of the patch tokens. Using a grid of p patch tokens, we have $m = 2p + 2$. We use the embedding vectors of the penultimate layer jointly with the last layer features to help the model reasoning about smaller resolution objects. Indeed, previous work [91, 301] showed that there is usually a decrease in the amount of spatial information represented in the last layer of CLIP. Moreover, we use a mapping network $\mathcal{M}$ (6.2) to reduce the dimensionality of the visual features to match the decoder embedding size. $\mathcal{M}$ is composed of a linear layer, preceded and followed by LayerNorm, and all features in $\mathcal{Z}$ are obtained as output of $\mathcal{M}$. The parameters of $\mathcal{M}$ are learned jointly with our decoder (see below) and $\mathcal{M}$ is shared by all features in $\mathcal{Z}$ and both layers of $\mathcal{E}$ (L and $L-1$).

We replace the CLIP textual encoder with our decoder $\mathcal{D}$, a relatively small network, composed of only three blocks with ∼64M total parameters, which is the module we use to adapt CLIP to solve compositional tasks. 6.2 shows the architecture of $\mathcal{D}$, which takes as input $2n$ tokens. The first sequence of n tokens are masked tokens, while the others are visible tokens, and we represent each w_j with both a masked and a visible token. Specifically, to condition $\mathcal{D}$ with respect to the event $S_j = t\,(t \in V)$ in 6.1, we use masked tokens specific for each syntactic type t in V. In more detail, V is composed of the 45 standard syntactic categories

defined in [250] (see D.4). We associate each category t with a masked token $\texttt{MSK}_t$. Then, for each word $w_j \in C$, if $type(w_j) = t$, then the masked token used for w_j is $\texttt{MSK}_t$. This is simply implemented using a lookup table of masked token embeddings, composed of 45 different initial embedding vectors (learned using standard backpropagation) and which extends the (single) masked token used in common masked-token prediction tasks [135]. The other n tokens are visible, standard textual tokens, one for each $w_j \in C$. In this way, w_j is represented both as a visible token and as a masked token of type t. In a given layer of $\mathcal{D}$, these two tokens are respectively represented by the masked-token embedding vector $\boldsymbol{m}_j$ and the visible-token embedding vector $\boldsymbol{v}_j$.

Each block of $\mathcal{D}$ is composed of two layers. In the first layer, we compute the self-attention of each masked embedding $\boldsymbol{m}_j$ with itself, jointly with the attention of $\boldsymbol{m}_j$ with all the visible embeddings $\boldsymbol{v}_{i_1}, ..., \boldsymbol{v}_{i_k}$, where $\mathbf{PA}(W_j) = \{W_{i_1}, ..., W_{i_k}, S_j, Z_1, ..., Z_m\}$. Note that there is no attention between $\boldsymbol{m}_{j_1}$ and $\boldsymbol{m}_{j_2}$, with $j_1 \neq j_2$. In the same layer, we compute the self-attention of each visible embedding $\boldsymbol{v}_j$ with itself, jointly with the attention of $\boldsymbol{v}_j$ with $\boldsymbol{v}_{i_1}, ..., \boldsymbol{v}_{i_k}$ (6.2). Note that there is no information leak, since $\boldsymbol{m}_j$, later used for the final prediction, has no direct or indirect access to $\boldsymbol{v}_j$. We call this *Dependency Guided Attention* to differentiate it from the standard self-attention (6.2). In the second layer of each block of $\mathcal{D}$, both the masked ($\boldsymbol{m}_j$) and the visible ($\boldsymbol{v}_j$) embeddings pay attention to the visual features in $\mathcal{Z}$ using cross-attention, in this way implementing the dependence between W_j and $Z_1, ..., Z_m$. Finally, after the last block of $\mathcal{D}$ we discard the visible-token embeddings and we fed each masked-token final embedding to a linear layer computing a posterior distribution over the vocabulary of textual terms. $\mathcal{D}$ is trained from scratch using as the only objective the maximization of the log-likelihood of the disentangled factorization:

$$\mathcal{L} = \log \left[\prod_{j=1}^{n} P(W_j | \mathbf{PA}(W_j)) \right] = \sum_{j=1}^{n} \log(P(W_j | \mathbf{PA}(W_j))). \tag{6.4}$$

Inference Most of the compositional tasks are modeled as image-to-text re-

trieval tasks. In case of COGT, given a testing image X, we compute the log-likelihood of all the candidate testing captions and we select the highest scoring sentence. Note that computing $\mathcal{Z}$ is independent of the specific caption C, and it can be done once for each image in the dataset. The log-likelihood is computed using 6.4 and a semi-parallel AR prediction strategy which follows the partial order induced by the DT. Specifically, using the dependency parser we extract $\mathcal{T}$ from a candidate caption C. Then we proceed using a *level order traversal* of $\mathcal{T}$, in which, starting from the root, we predict in parallel all the tokens of a given level of the tree and then we move to the next level.

6.2.4 Experiments

In our evaluation we use four common compositional benchmarks: ARO [327], SugarCrepe [110], VL-CheckList [347] and ColorSwap [37], and an additional benchmark FG-OVD [31] which we propose in this chapter. Most of them are composed of different tasks and datasets, and we report both the task-specific and the average accuracy across all tasks. Following [336], we do not use COCO Order and Flickr Order (two of the ARO tasks) because it has been previously showed that a "blind" LM, *with no access to the image*, can achieve about99% accuracy on these tasks [274]. The reason for this is the grammatical and semantic errors introduced in the negatives when swapping or replacing caption words (see 6.2.1). For instance, the sentence "with man is wearing ears the an glasses pierced orange hat and" (Flickr Order) can be easily detected as false by an LM without any visual knowledge. In D.2.3 we show additional results with Winoground [266], which, however, is often not used by other methods based on CLIP [336] because requires the VLM to be able to detect out-of-focus objects in low-resolution images [77]. In contrast, we propose to use FG-OVD [31], a benchmark originally proposed to evaluate the ability of open-vocabulary object detectors to discern fine-grained object properties. In FG-OVD, negative captions are created starting from the object-specific captions by replacing attributes referring to the object's color, material, texture, etc. We crop the objects' bounding boxes which we use jointly with positive and negative captions and an image-

to-text retrieval task.

	ARO			SugarCrepe				VL-Checklist				ColorSwap	FG-OVD	Avg
Model	Relation	Attribute	Avg	Add	Replace	Swap	Avg	Attribute	Object	Relation	Avg	ITT	Avg	
Fully-Parallel	76.37	49.24	62.81	98.98	77.98	68.81	81.92	83.66	67.55	74.7	75.3	25.24	41.84	**57.42**
Mixed	84.83	69.12	76.98	99.01	84.17	78.39	87.19	85.89	76.96	91.44	84.76	41.33	45.21	**67.10**$_{+9.67}$
Sequential-AR	84.86	77.87	81.37	98.96	83.62	81.50	88.02	86.82	75.45	91.11	84.45	46.33	46.24	**69.28**$_{+2.18}$
COGT	87.56	90.26	88.91	98.26	87.10	83.14	89.50	86.07	78.91	89.37	84.78	61.33	51.48	**75.20**$_{+5.92}$

Table 6.1: Comparison between different generative training strategies. The value **+x** reported in the i-th row, column Average, refers to the average improvement across all datasets with respect to the method in row $i-1$.

6.2.5 Ablations

In the experiments of this section we follow a widely adopted protocol, first proposed by [327], in which the VLM backbone is CLIP and the only training dataset is COCO [171]. However, we *do not* use the hard negatives of [327] for training because of their frequent semantic and syntactic errors (see above). In 6.1 we compare to each other the different word prediction strategies described in 6.2.2 using CLIP as the VLM. Specifically, we indicate with *Sequential-AR* the replacement of our decoder with a standard AR decoder (cross-attention over $\mathcal{Z}$ and standard causal attention over the past words of the caption), trained using the common image captioning objective defined in 6.2. Similarly to COGT, we freeze the CLIP encoder and we use the visual features ($\mathcal{Z}$) extracted from both the last and the penultimate layer of $\mathcal{E}$ (6.2.3). We use a decoder of the same size, which takes only visible words as input. *Sequential-AR* can be considered as our re-implementation of Cap[†] [274], trained on COCO and with a frozen visual encoder, which can be directly compared with the CGM-based strategy of COGT. Similarly, we indicate with *Fully-Parallel* our re-implementation of the parallel prediction strategy proposed in [274], using a decoder which takes as input only masked tokens (only cross-attention over $\mathcal{Z}$ with a frozen visual encoder), trained using 6.3. Finally, in *Mixed* we use the sequential-parallel mixed strategy adopted in CapPa, in which, following [274], we use 75% of the training samples with a parallel prediction (6.3) and 25% of the samples with an AR prediction (6.2).

[†]There are no publicly available network weights for [274].

The results show that COGT outperforms all the other prediction strategies in all the datasets, often with a significant margin. For instance, COGT achieves an average accuracy improvement of +17.77 points across all datasets with respect to *Fully-Parallel*, which arguably shows that the conditional independence assumption in 6.3 is too strong. Overall, these results confirm that an off-the-shelf dependency parser provides a priori knowledge which can be exploited to model the conditional dependencies between words in a sentence.

We further investigate the role of the dependency parser in 6.2. Specifically, the column *Parser* refers to the adopted dependency parser, where we compare 3 different methods: Deep Biaffine [82], CRFPar [341] and Deep Biaffine + RoBERTa [82]. Note that we use the parsers as black boxes, without any training or fine-tuning, and the differences in the corresponding rows of 6.2 are based only on the use of a different external parser for COGT. 6.2 shows that the best results correspond to the use of Deep Biaffine + RoBERTa [82], which is aligned with the higher accuracy of this parser compared to the other two according to the linguistic leaderboard Penn Tree Bank [192]. Note also that, according to this widely adopted parser ranking [192], there are higher performing parsers (*e.g.*, [200]), however their code is not publicly available or it is not easy to use. Thus, we opted for Deep Biaffine + RoBERTa (used in all the other experiments of this chapter). However, the results in 6.2 show that, using a better parser, COGT can most likely achieve even better results.

The *Mask-Specific* column in 6.2 indicates the use of a dedicated masked token for each of the 45 syntactic categories of V (6.2.3), which is compared with a generic BERT-like masked token [135]. In the latter case, we use the same masked token initial embedding vector for all the n masked tokens fed to $\mathcal{D}$ (replicated n times), thus dropping any conditioning on S_j in 6.1. The results in 6.2 show that this corresponds to a -2.69 point drop in accuracy averaged across all five datasets.

Finally, the *Layers* column in 6.2 indicates the number of layers of CLIP we use to extract the visual features $\mathcal{Z}$: *Layers* = 1 means only the last layer (L); *Layers* = 2 means that we use also the penultimate layer (6.2.3). When the last layer only is used, the average accuracy drop is -4.75, showing the importance of using

			ARO			SugarCrepe				VL-Checklist				ColorSwap	FG-OVD	Avg
Parser	Mask-Specific	Layers	Relation	Attribute	Avg	Add	Replace	Swap	Avg	Attribute	Object	Relation	Avg	ITT	Avg	
CRFPar	✓	2	85.68	88.34	87.01	98.16	84.94	80.30	87.80	86.99	77.68	87.09	83.92	56.33	43.74	71.76
Deep Biaffine	✓	2	86.56	89.10	87.83	98.11	85.80	81.49	88.46	87.02	78.30	87.75	84.35	61.33	44.74	73.34
Deep Biaffine + RoBERTa	✗	2	84.75	86.16	85.46	98.86	84.37	80.25	87.82	83.79	78.24	90.84	84.29	58.00	46.99	72.51
Deep Biaffine + RoBERTa	✓	1	86.82	89.67	88.25	98.26	86.56	82.33	89.05	84.41	78.94	89.54	84.30	45.00	45.63	70.45
Deep Biaffine + RoBERTa	✓	2	87.56	90.26	88.91	98.26	87.10	83.14	89.50	86.07	78.91	89.37	84.78	61.33	51.48	75.20

Table 6.2: Empirical contribution of different components of COGT.

lower-level features in compositionality tasks where the VLM needs to consider small, non-foreground objects.

6.2.6 Main experiments

Setting. In this section we compare COGT with state-of-the-art compositional methods. Since different works are based on different VLMs and use different training data, to make the comparison as fair as possible, we split our evaluation based on both the VLM backbone and the used training set. Specifically, in 6.3 we group the approaches based on CLIP [225] and in 6.4 those which adopt a different VLM, while 6.5 is dedicated to methods based on VLMs pre-trained using a language based decoder. In the first category, our approach is indicated by COGT-CLIP. In the second group, we use XVLM [329] as our backbone (COGT-XVLM). Finally, we use InstructBLIP [70] for the VLM category with a language-based decoder (COGT-InstructBLIP). COGT-CLIP, COGT-XVLM and COGT-InstructBLIP are trained on COCO *only* ($\sim$ 100K training samples, see 6.2.5). Moreover, following [336], we present additional results training on a combination of three datasets: COCO, CC3M [247], and Visual Genome [142], and we call the corresponding methods as COGT-CLIP+, COGT-XVLM+ and COGT-InstructBLIP+. In this case, we use a decoder $\mathcal{D}$ with four blocks. Note that we use only $\sim 50K$ samples from Visual Genome because we *removed those training data which overlap with ARO and VL-Checklist*. On the other hand, CC3M ($\sim$3.3M training samples) is a much larger but also noisier dataset, since its captions are obtained from the Alt-text HTML attribute associated with web images, and we use it also to indirectly evaluate the robustness of COGT to noisy textual descriptions (see D.1). For each compared baseline, the results shown in the tables refer to the values reported in the original article (when available) or to

our reproduction using the (possibly available) public code. The results on FG-OVD are averaged over all tasks and we report in D.2.1 the task-specific values.

	ARO			SugarCrepe				VL-Checklist				ColorSwap	FG-OVD	Avg
Model	Relation	Attribute	Avg	Add	Replace	Swap	Avg	Attribute	Object	Relation	Avg	ITT	Avg	
Zero-shot														
CLIP [225]	59.00	62.00	60.50	85.58	80.76	70.83	79.05	67.93	82.83	64.19	71.65	35.67*	47.33	58.84
Training on COCO only														
CLIP Fine-Tuned [327]	63.00	65.00	64.00	.	.	.	.	.	.	.	.	.	.	.
NegCLIP [327]	81.00	71.00	76.00	87.29	85.36	75.30	82.65	72.24	87.00	71.39	76.87	35.67*	41.69	62.57
CE-CLIP [336]	83.00	76.40	79.70	92.90	87.00	74.90	84.94	72.60	84.60	71.80	76.30	18.67	41.97	60.31
Structure-CLIP [119]	85.10*	83.50*	84.30*	.	.	.	.	.	.	.	.	.	.	.
GNM [239]	65.00	65.00	65.00	82.85	80.95	66.71	76.83	70.15	85.91	64.10	73.38	13.00	38.79	53.40
Plausible Adj. Neg [35]	65.07	67.94	66.51	89.64	85.37	70.88	81.96	76.51	88.13	69.90	78.17	17.67	44.98	57.86
SDS-CLIP [30]	55.00	66.00	60.50	.	.	.	.	.	.	.	.	.	.	.
COGT-CLIP	87.56	90.26	88.91	98.26	87.10*	83.14	89.50	86.07	78.91	89.37*	84.78	61.33	51.48	75.20
Training on datasets larger than COCO														
CE-CLIP+ [336]	83.60	77.10	80.35	94.40	**89.30**	78.00*	87.23*	76.70	86.30	74.70	79.23	.	.	.
IL-CLIP [349]	.	.	.	73.80	73.00	62.90	69.90	.	.	.	.	.	.	.
syn-CyCLIP [40]	69.00	63.65	66.33	.	.	.	.	68.06	.	65.73	.	.	.	.
DAC-SAM [80]	77.16	70.50	73.83	92.87	86.18	71.06	83.37	75.80	**88.50**	89.80	84.70*	16.33	48.36	61.31
DAC-LLM [80]	81.28	73.91	77.60	95.83*	88.09	72.48	85.47	77.30*	87.30*	86.40	83.66	18.33	49.60*	62.93*
COGT-CLIP+	**90.67**	**96.01**	**93.34**	**98.42**	87.05	**84.21**	**89.89**	**90.71**	84.91	**92.33**	**89.31**	**81.66**	**69.96**	**84.83**

Table 6.3: Comparison with compositional methods based on CLIP. For each baseline, we report the values published in the original article. In case a given dataset was not used by that baseline, but a public code is available, we report the results obtained by our reproduction. With **x**, $\underline{x}$ and x^* we indicate the first, the second and the third best result, respectively.

CLIP based methods. 6.3 shows that COGT-CLIP *largely* outperforms all the other approaches trained only on COCO **and it also outperforms all the methods trained on datasets larger than COCO**. For instance, using the average across all the datasets, COGT-CLIP outperforms the second best result in 6.3 (DAC-LLM [80]) by a remarkable 12.27 points. Note that DAC-LLM was trained on CC3M, a dataset an order of magnitude larger than COCO, and using high-quality LLM-based annotations (2.3). Moreover, even considering the average computed on the individual datasets, COGT-CLIP outperforms all the other methods in 6.3 (including those trained on datasets larger than COCO). We believe that these results show that our CGM-based training strategy can better generalize by leveraging available training data, most likely because we remove spurious inter-variable associations from the learning objective (6.2.2). Moreover, COGT-CLIP+ achieves even better results, with an average across all benchmarks that is almost 22 points more than the second best result (DAC-LLM).

Other VLMs. The results in 6.3 are confirmed by those reported in 6.4, where COGT-XVLM largely outperforms the other methods, included CE-XVLM [336], which uses our same VLM (XVLM) and the same training data (COCO).

Model	ARO			SugarCrepe				VL-Checklist				ColorSwap	FG-OVD	Avg
	Relation	Attribute	Avg	Add	Replace	Swap	Avg	Attribute	Object	Relation	Avg	ITT	Avg	
Zero-shot														
XVLM [329]	73.40	86.80	80.10	.	.	.	.	75.10*	85.80	70.40	76.50	.	.	.
Training on COCO only														
CE-XVLM [336]	73.90*	89.30*	81.60*	.	.	.	.	74.80	**86.90**	79.70*	78.60*	.	.	.
HardNeg-DiffusionITM [143]	52.30	67.60	59.95	.	.	.	.	.	.	.	.	.	.	.
COGT-XVLM	87.64	92.30	89.97	**98.65**	**89.17**	84.37	90.73	85.87	80.49	88.74	85.03	69.67	50.12	77.10
Training on datasets larger than COCO														
COGT-XVLM+	**91.71**	**96.59**	**94.15**	98.30	88.97	**86.49**	**91.25**	**91.54**	84.73*	**92.33**	**89.53**	**82.33**	**74.22**	**86.30**

Table 6.4: Comparison with methods based on other VLMs. Similar to 6.3, the baseline results are either taken from the original paper or reproduced using the public code.

Model	ARO			SugarCrepe				VL-Checklist				ColorSwap	FG-OVD	Avg
	Relation	Attribute	Avg	Add	Replace	Swap	Avg	Attribute	Object	Relation	Avg	ITT	Avg	
BLIP [159]	59.00	88.00	73.50	.	.	.	.	75.20	82.20	70.50	75.70	.	.	.
BLIP2 [158]	41.20	71.30	56.25	.	.	.	.	77.80	84.90	70.60	77.80	.	.	.
InstructBLIP (FlanT5XL) [70]	69.20	50.83	60.02	65.43	72.59	63.41	67.14	56.37	80.33	53.34	63.35	40.33*	26.80*	51.53*
MiniGPT-4 [353]	46.90	55.70	51.30	.	.	.	.	71.30	84.20	.	.	.	.	.
GPT-4V [205]	.	.	.	91.68	**93.37**	86.61	90.55	.	.	.	.	.	.	.
LLaVA-1.5-13B [179]	.	.	.	.	.	80.95	.	.	.	.	.	.	.	.
LLaVA-1.5-13B+CRG [281]	.	.	.	.	.	87.90	.	.	.	.	.	.	.	.
LLaVA-1.6-34B [178]	.	.	.	.	.	81.25	.	.	.	.	.	.	.	.
LLaVA-1.6-34B+CRG [281]	.	.	.	.	.	90.75	.	.	.	.	.	.	.	.
BLIP-VisualGPTScore ($\alpha = 0$) [173] †	89.10*	95.30	92.20*	91.00	93.30	**91.00**	91.77*	78.70*	**92.60**	90.80	87.37	.	.	.
BLIP2-VisualGPTScore ($\alpha = 0$) [173] †	90.70	94.30*	92.50	92.70	93.00*	91.24	92.31	73.90	90.10	89.90*	84.63	.	.	.
Cap [274]	86.60	88.90	87.75	98.94	88.21	84.00	90.38	.	.	.	.	.	.	.
CapPa [274]	86.70	85.70	86.20	**99.13**	87.67	83.11	89.97	.	.	.	.	.	.	.
COGT-InstructBLIP	87.63	88.93	88.28	98.55*	90.61	88.12	**92.42**	85.77	79.96	89.14	84.96*	72.66	51.26	77.87
COGT-InstructBLIP+	**91.12**	**95.64**	**93.38**	98.45	90.27	88.22*	92.31	**90.80**	85.17*	**92.80**	**89.60**	**83.33**	**70.72**	**85.87**

Table 6.5: Comparison with methods based on encoder-decoder VLM architectures pre-trained with a textual token prediction task. † Lin et al. [173] show additional results with α set using dataset-specific cross-validation data, which we do not report, however, to make the comparison fair to other methods that do not have access to benchmark data.

Model	CIFAR10	CIFAR100	ImageNet1K (top 1)	ImageNet1K (top 5)
CLIP [225]	94.2	79.0	75.0	93.2
CLIP Fine-Tuned [327]	95.0	80.0	74.0	-
NegCLIP [327]	94.0	79.0	72.0	-
CE-CLIP [336]	93.8	78.0	-	92.6
CE-CLIP+ [336]	93.8	78.1	-	92.7
COGT-CLIP	96.7	84.9	74.4	93.3
COGT-CLIP+	**96.8**	**85.4**	**75.3**	**93.8**

Table 6.6: Comparison of CLIP-based models using image classification tasks and linear probing.

COGT-XVLM+ further improves these results and it also outperforms COGT-CLIP+. This is probably because the XVLM encoder can better represent small-scale objects than the CLIP encoder, and these objects are often referenced in the captions of these compositional benchmarks [327].

Language-decoding based VLMs. In 6.5 we compare to each other VLMs pre-trained using a decoder and a generative word prediction task. The compositional skills of these methods are generally much higher than the other VLMs,

which indirectly confirms that a word-prediction training helps the VLM to understand the compositional nature of the human language (2.3). However, a direct comparison with VLMs such as CLIP, XVLM or Stable Diffusion [143] is difficult, since each of these backbones has been pre-trained on datasets with a huge difference in size. For instance, Cap and CapPa [274] were pre-trained with a private dataset composed of 1B image/Alt-text pairs [274], which is different orders of magnitude larger than the dataset used to pre-train XVLM ($\sim$ 16M training samples [329]). Despite that, COGT-XVLM+ (6.4) outperforms both Cap and CapPa and all the other methods in 6.5. Similarly, COGT-InstructBLIP+ significantly outperforms the zero-shot accuracy of InstructBLIP and, jointly with COGT-XVLM+, achieves *state-of-the-art performance on the considered compositional datasets*.

6.2.7 Downstream tasks

Doveh et al. [81] show that most compositional methods usually deteriorate the VLM skills on non-compositional, standard tasks. We analyze this aspect using the protocol adopted by [327, 336], which is based on linear probing the fine-tuned CLIP visual encoder on CIFAR10, CIFAR100 and ImageNet. Since in COGT $\mathcal{E}$ is frozen, we use $\mathcal{E}$ jointly with our mapping network $\mathcal{M}$ and, specifically, the feature $\boldsymbol{z}^L_{[\text{CLS}]}$. The results shown in 6.6 show that COGT does not degrade CLIP's features and can even slightly improve them.

6.2.8 Limitations and Future Directions

While the proposed causal compositional framework substantially improves generalization on a wide range of compositional benchmarks, it also presents several limitations that outline research directions beyond the scope of this thesis.

First, the approach relies on an off-the-shelf dependency parser to define the causal structure over linguistic variables. Although modern parsers achieve high accuracy, parsing errors inevitably propagate to the causal graph and may affect downstream predictions. Future work could explore joint or weakly supervised learning of dependency structures, or parser-free approaches in which the causal

graph is inferred directly from data.

Second, the causal interpretation adopted in this chapter is grounded in syntactic dependencies, which capture an important but incomplete subset of linguistic causality. Certain semantic phenomena, such as pragmatic inference, discourse-level dependencies, or implicit relations, are not explicitly modeled in the current formulation. Extending the causal graph to incorporate higher-level semantic or discourse structures represents an interesting avenue for further research.

Third, the current framework focuses on image–text pairs and caption-based compositional reasoning. While the causal factorization naturally generalizes to longer texts, dialogues, or multi-image settings, scaling the approach to large multimodal language models with free-form generation remains an open challenge. Integrating causal compositional constraints into autoregressive multimodal LLMs is a particularly promising direction.

Finally, the proposed model assumes a fixed causal structure during both training and inference. Relaxing this assumption to allow adaptive or context-dependent causal graphs may further improve robustness and flexibility, especially in open-world or out-of-distribution scenarios.

6.3 Chapter Summary

This chapter addressed the problem of compositional reasoning in vision–language models from a causal perspective. We showed that many failures observed in existing models stem not from missing knowledge or insufficient perception, but from the lack of structured representations capable of capturing syntactic and causal relationships between linguistic elements and visual content.

By grounding caption generation in causal graphical models derived from dependency trees, the proposed approach enforces a principled factorization of the joint distribution over words. This causal ordering enables semi-parallel generation that respects linguistic dependencies while avoiding spurious correlations induced by linear word order. Unlike data-centric solutions based on curated negatives or large-scale captioning datasets, the method injects compositional struc-

ture directly into the modeling and training process. Experimental results across multiple benchmarks demonstrate that causal compositional modeling substantially improves generalization on compositional tasks, even when trained with limited data and without modifying the visual backbone. Together with the previous chapters, this analysis highlights that robust multimodal reasoning requires not only perception and knowledge access, but also representations capable of capturing the structured relationships that govern language and visual concepts.

More broadly, the results of this chapter suggest that incorporating structural inductive biases into vision–language models can improve compositional generalization. While these findings are grounded in a specific modeling approach and experimental setting, they indicate that explicitly modeling the causal and compositional organization of language is a promising direction for moving beyond purely scale-driven improvements toward more systematic and robust multimodal reasoning.

7 Conclusions

This dissertation has explored the design of multimodal vision–language systems through a coherent research trajectory that spans image captioning, multimodal large language models, retrieval-augmented generation, and compositional modeling. Rather than treating these topics as independent problems, the thesis has investigated how they relate to one another and how progress in one area naturally motivates advances in the next. This final chapter summarizes the main contributions, synthesises insights across chapters, and outlines promising directions for future research.

7.1 Summary of Contributions

The primary contributions of this thesis can be grouped into four thematic areas, corresponding to the core chapters of the dissertation.

Image Captioning. This thesis revisited image captioning with the goal of improving descriptiveness and alignment with human judgment. The presented works demonstrated that traditional reference-based training objectives often lead to generic and underspecified captions. To address this limitation, the the-

sis introduced learned, self-supervised reward models and preference optimization strategies that guide caption generation using adaptive multimodal signals. These contributions demonstrate that moving beyond handcrafted metrics enables captioning models to generate richer, more informative, and better-aligned descriptions.

Multimodal Large Language Models. The thesis contributed to the understanding of multimodal large language models by analyzing their architectural design and empirical behavior. A comprehensive survey clarified the landscape of MLLMs, highlighting common training paradigms and open challenges. Complementarily, an extensive ablation study revealed the critical role of visual backbones in shaping multimodal reasoning capabilities. Together, these contributions provided insight into how architectural choices affect performance, efficiency, and generalization in large-scale multimodal systems.

Multimodal Retrieval-Augmented Generation. A central contribution of this dissertation lies in the integration of external knowledge into multimodal models through retrieval-augmented generation. Focusing on knowledge-intensive visual question answering, the proposed methods demonstrated that structured multimodal retrieval substantially improves factual accuracy and robustness. The thesis explored hierarchical retrieval pipelines, relevance estimation, and reflective mechanisms that enable models to decide when retrieval is necessary. These results highlight retrieval as a key component for grounding multimodal reasoning beyond parametric knowledge.

Compositional Vision–Language Modeling. The final technical contribution addressed the lack of compositional understanding in existing vision–language models. By introducing dependency-based and causal graphical modeling into the multimodal generation process, the proposed approach enforced syntactic structure and disentangled genuine semantic relations from spurious correlations. The resulting model achieved strong compositional generalization without relying on curated negative data, demonstrating the effectiveness of explicit structure in multimodal reasoning.

7.2 Cross-Chapter Synthesis

While each chapter addresses a distinct research question, the contributions of this thesis are tightly interconnected. Image captioning served as an initial testbed for studying multimodal alignment and revealed the limitations of traditional training objectives. These limitations motivated the transition toward larger and more general multimodal architectures, culminating in the study of MLLMs.

As the scope of multimodal models expanded, the need for external knowledge became increasingly evident, particularly in knowledge-intensive visual question answering. This observation naturally led to the investigation of retrieval-augmented generation as a mechanism to ground multimodal reasoning in external evidence. Finally, the analysis of compositional failures across tasks highlighted the importance of linguistic structure, motivating the introduction of causal and syntactic modeling.

Overall, this body of work suggests that progress in multimodal intelligence is likely to benefit from the joint integration of scalable architectures, principled training objectives, and mechanisms for accessing external knowledge. These elements are supported across multiple contributions of this thesis.

In addition, the compositional modeling study presented in Chapter 6 provides evidence that explicit structural inductive biases can further improve multimodal reasoning, particularly for compositional generalization. While this evidence is grounded in a specific modeling framework and smaller-scale experiments, it indicates that structure-aware modeling may represent a promising complementary direction for future multimodal systems.

7.3 Future Directions

The results presented in this thesis open several promising avenues for future research.

Unified multimodal training objectives. Future work may explore training paradigms that jointly integrate learned rewards, retrieval signals, and structural constraints within a single optimization framework. Such objectives

could provide more holistic supervision while reducing reliance on task-specific heuristics.

Scalable and efficient multimodal retrieval. As multimodal knowledge bases grow in size and complexity, efficient retrieval becomes increasingly important. Future research may investigate scalable indexing strategies, adaptive retrieval policies, and tighter integration between retrieval and generation modules.

Structure-aware multimodal large language models. The success of causal and dependency-based modeling suggests that future MLLMs could benefit from built-in mechanisms that explicitly encode linguistic or semantic structure. Extending these ideas to larger architectures and broader tasks remains an open challenge.

Evaluation beyond benchmark accuracy. Finally, there is a need for evaluation protocols that better capture factuality, reasoning correctness, and compositional understanding in multimodal systems. Developing such benchmarks will be essential for measuring real progress in multimodal intelligence.

7.4 Final Remarks

This dissertation has presented a comprehensive investigation into multimodal vision–language modeling, combining theoretical insights with practical contributions across multiple tasks. By progressively moving from image captioning to large-scale reasoning, knowledge integration, and compositional structure, the thesis reflects both the evolution of the field and the research trajectory developed during the Ph.D.

More broadly, this work suggests that future advances in multimodal artificial intelligence will likely depend not only on larger models and datasets, but also on principled mechanisms for alignment and knowledge access that enable models to connect perception with external information.

In addition, the results of Chapter 6 indicate that incorporating explicit structural reasoning may represent a promising complementary direction, particularly for improving compositional generalization in multimodal systems.

List of Publications

[1] Bucciarelli, D., Moratelli, N., Cornia, M., Baraldi, L., and Cucchiara, R. (2024). Personalizing multimodal large language models for image captioning: an experimental analysis. In *Proceedings of the European Conference on Computer Vision*.

[2] Caffagni, D., Cocchi, F., Barsellotti, L., Moratelli, N., Sarto, S., Baraldi, L., Baraldi, L., Cornia, M., and Cucchiara, R. (2024a). The Revolution of Multimodal Large Language Models: A Survey. In *Findings of the Association for Computational Linguistics*.

[3] Caffagni, D., Cocchi, F., Moratelli, N., Sarto, S., Cornia, M., Baraldi, L., and Cucchiara, R. (2024b). Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

[4] Cocchi, F., Moratelli, N., Caffagni, D., Sarto, S., Baraldi, L., Cornia, M., and Cucchiara, R. (2025a). LLaVA-MORE: A Comparative Study of LLMs and Visual Backbones for Enhanced Visual Instruction Tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*.

[5] Cocchi, F., Moratelli, N., Cornia, M., Baraldi, L., and Cucchiara, R. (2025b). Augmenting multimodal llms with self-reflective tokens for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[6] Compagnoni, A., Caffagni, D., Moratelli, N., Baraldi, L., Cornia, M., and Cucchiara, R. (2025). Mitigating hallucinations in multimodal llms via object-aware preference optimization. In *Proceedings of the British Machine Vision Conference*.

[7] Moratelli, N., Barraco, M., Cornia, M., Baraldi, L., and Cucchiara, R. (2024a). Are learnable prompts the right way of prompting? adapting vision-and-language models with memory optimization. *IEEE Intelligent Systems*, 39:26–34.

[8] Moratelli, N., Barraco, M., Morelli, D., Cornia, M., Baraldi, L., and Cucchiara, R. (2023). Fashion-oriented image captioning with external knowledge retrieval and fully attentive gates. *Sensors*, 23:1286.

[9] Moratelli, N., Caffagni, D., Cornia, M., Baraldi, L., and Cucchiara, R. (2024b). Revisiting image captioning training paradigm via direct clip-based optimization. In *Proceedings of the British Machine Vision Conference*.

[10] Moratelli, N., Cornia, M., Baraldi, L., and Cucchiara, R. (2025). Fluent and accurate image captioning with a self-trained reward model. In *Proceedings of the International Conference on Pattern Recognition*.

[11] Parascandolo, F., Moratelli, N., Sangineto, E., Baraldi, L., and Cucchiara, R. (2025). Causal graphical models for vision-language compositional understanding. In *Proceedings of the International Conference on Learning Representations*.

[12] Sarto, S., Moratelli, N., Cornia, M., Baraldi, L., and Cucchiara, R. (2025). Positive-Augmented Contrastive Learning for Vision-and-Language Evaluation and Training. *International Journal of Computer Vision*, 133:1–25.

Bibliography

[13] Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R. J., Javaheripi, M., Kauffmann, P., et al. (2024). Phi-4 Technical Report. *arXiv preprint arXiv:2412.08905*.

[14] Abouelenin, A., Ashfaq, A., Atkinson, A., Awadalla, H., Bach, N., Bao, J., Benhaim, A., Cai, M., Chaudhary, V., Chen, C., et al. (2025). Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*.

[15] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

[16] Agrawal, H., Desai, K., Chen, X., Jain, R., Batra, D., Parikh, D., Lee, S., and Anderson, P. (2019). nocaps: novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[17] Agrawal, P., Antoniak, S., Hanna, E. B., Bout, B., Chaplot, D., Chudnovsky, J., Costa, D., De Monicault, B., Garg, S., Gervet, T., et al. (2024). Pixtral 12b. *arXiv preprint arXiv:2410.07073*.

[18] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. (2022). Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*.

[19] Allal, L. B., Lozhkov, A., Bakouch, E., Blázquez, G. M., Penedo, G., Tunstall, L., Marafioti, A., Kydlíček, H., Lajarín, A. P., Srivastav, V., et al. (2025). SmolLM2: When Smol Goes Big–Data-Centric Training of a Small Language Model. *arXiv preprint arXiv:2502.02737*.

[20] Anderson, P., Fernando, B., Johnson, M., and Gould, S. (2016a). SPICE: Semantic Propositional Image Caption Evaluation. In *Proceedings of the European Conference on Computer Vision*.

[21] Anderson, P., Fernando, B., Johnson, M., and Gould, S. (2016b). SPICE: Semantic Propositional Image Caption Evaluation. In *Proceedings of the European Conference on Computer Vision*.

[22] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[23] Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. (2023). Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805*.

[24] Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H. (2023). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. *arXiv preprint arXiv:2310.11511*.

[25] Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., et al. (2023). OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. *arXiv preprint arXiv:2308.01390*.

[26] Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al. (2023a). Qwen technical report. *arXiv preprint arXiv:2309.16609*.

[27] Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. (2023b). Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*.

[28] Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics Workshops*.

[29] Barraco, M., Sarto, S., Cornia, M., Baraldi, L., and Cucchiara, R. (2023). With a Little Help from your own Past: Prototypical Memory Networks for Image Captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[30] Basu, S., Hu, S. X., Sanjabi, M., Massiceti, D., and Feizi, S. (2024). Distilling knowledge from text-to-image generative models improves visio-linguistic reasoning in CLIP. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[31] Bianchi, L., Carrara, F., Messina, N., Gennaro, C., and Falchi, F. (2024). The devil is in the fine-grained details: Evaluating open-vocabulary object detectors for fine-grained understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[32] Bolelli, F., Borghi, G., and Grana, C. (2018). XDOCS: an Application to Index Historical Documents. In *Digital Libraries and Multimedia Archives*.

[33] Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Van Den Driessche, G. B., Lespiau, J.-B., Damoc, B., Clark, A., et al. (2022). Improving Language Models by Retrieving from Trillions of Tokens. In *Proceedings of the International Conference on Machine Learning*.

[34] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.

[35] Buettner, K. and Kovashka, A. (2024). Investigating the role of attribute context in vision-language models for object recognition and detection. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.

[36] Bulian, J., Buck, C., Gajewski, W., Boerschinger, B., and Schuster, T. (2022). Tomayto, Tomahto. Beyond Token-level Answer Equivalence for Question Answering Evaluation. *arXiv preprint arXiv:2202.07654*.

[37] Burapacheep, J., Gaur, I., Bhatia, A., and Thrush, T. (2024). Colorswap: A color and word order dataset for multimodal evaluation. In *arXiv:2402.04492*.

[38] Byeon, M., Park, B., Kim, H., Lee, S., Baek, W., and Kim, S. (2022). COYO-700M: Image-Text Pair Dataset. https://github.com/kakaobrain/coyo-dataset.

[39] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[40] Cascante-Bonilla, P., Shehada, K., Smith, J. S., Doveh, S., Kim, D., Panda, R., Varol, G., Oliva, A., Ordonez, V., Feris, R., and Karlinsky, L. (2023). Going beyond nouns with vision & language models using synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[41] Cha, J., Kang, W., Mun, J., and Roh, B. (2023). Honeybee: Locality-enhanced Projector for Multimodal LLM. *arXiv preprint arXiv:2312.06742*.

[42] Chan, D., Petryk, S., Gonzalez, J. E., Darrell, T., and Canny, J. (2023a). CLAIR: Evaluating Image Captions with Large Language Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[43] Chan, D. M., Myers, A., Vijayanarasimhan, S., Ross, D. A., and Canny, J. (2023b). IC^3: Image Captioning by Committee Consensus. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[44] Changpinyo, S., Sharma, P., Ding, N., and Soricut, R. (2021). Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[45] Chen, C., Qin, R., Luo, F., Mi, X., Li, P., Sun, M., and Liu, Y. (2023a). Position-Enhanced Visual Instruction Tuning for Multimodal Large Language Models. *arXiv preprint arXiv:2308.13437*.

[46] Chen, D., Liu, J., Dai, W., and Wang, B. (2023b). Visual Instruction Tuning with Polite Flamingo. *arXiv preprint arXiv:2307.01003*.

[47] Chen, G., Shen, L., Shao, R., Deng, X., and Nie, L. (2023c). LION: Empowering Multimodal Large Language Model with Dual-Level Visual Knowledge. *arXiv preprint arXiv:2311.11860*.

[48] Chen, J., Guo, H., Yi, K., Li, B., and Elhoseiny, M. (2022a). VisualGPT: Data-Efficient Adaptation of Pretrained Language Models for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[49] Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., and Elhoseiny, M. (2023d). MiniGPT-v2: Large Language Model as a Unified Interface for Vision-Language Multi-task Learning. *arXiv preprint arXiv:2310.09478*.

[50] Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., and Zhao, R. (2023e). Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. *arXiv preprint arXiv:2306.15195*.

[51] Chen, Q., Deng, C., and Wu, Q. (2022b). Learning distinct and representative modes for image captioning. In *Advances in Neural Information Processing Systems*.

[52] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020a). A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the International Conference on Machine Learning*.

[53] Chen, X., Djolonga, J., Padlewski, P., Mustafa, B., Changpinyo, S., Wu, J., Ruiz, C. R., Goodman, S., Wang, X., Tay, Y., et al. (2023f). PaLI-X: On Scaling up a Multilingual Vision and Language Model. *arXiv preprint arXiv:2305.18565*.

[54] Chen, X., Fan, H., Girshick, R., and He, K. (2020b). Improved Baselines with Momentum Contrastive Learning. *arXiv preprint arXiv:2003.04297*.

[55] Chen, X., Wang, X., Beyer, L., Kolesnikov, A., Wu, J., Voigtlaender, P., Mustafa, B., Goodman, S., Alabdulmohsin, I., Padlewski, P., et al. (2023g). PaLI-3 Vision Language Models: Smaller, Faster, Stronger. *arXiv preprint arXiv:2310.09199*.

[56] Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., et al. (2023h). PaLI: A Jointly-Scaled Multilingual Language-Image Model. In *Proceedings of the International Conference on Learning Representations*.

[57] Chen, Y., Hu, H., Luan, Y., Sun, H., Changpinyo, S., Ritter, A., and Chang, M.-W. (2023i). Can Pre-trained Vision and Language Models Answer Visual Information-Seeking Questions? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[58] Chen, Y., Sikka, K., Cogswell, M., Ji, H., and Divakaran, A. (2023j). DRESS: Instructing Large Vision-Language Models to Align and Interact with Humans via Natural Language Feedback. *arXiv preprint arXiv:2311.10081*.

[59] Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al. (2024). Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[60] Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., and Girdhar, R. (2022). Masked-Attention Mask Transformer for Universal Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[61] Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. (2023). Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. https://lmsys.org/blog/2023-03-30-vicuna/.

[62] Cho, J., Yoon, S., Kale, A., Dernoncourt, F., Bui, T., and Bansal, M. (2022). Fine-grained image captioning with clip reward. In *North American Chapter of the Association for Computational Linguistics*.

[63] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2023). Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24:1–113.

[64] Chu, X., Qiao, L., Zhang, X., Xu, S., Wei, F., Yang, Y., Sun, X., Hu, Y., Lin, X., Zhang, B., et al. (2024). MobileVLM V2: Faster and Stronger Baseline for Vision Language Model. *arXiv preprint arXiv:2402.03766*.

[65] Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al. (2022). Scaling Instruction-Finetuned Language Models. *arXiv preprint arXiv:2210.11416*.

[66] Clark, K. and Jaini, P. (2023). Text-to-image diffusion models are zero-shot classifiers. In *arXiv:2303.15233*.

[67] Cornia, M., Baraldi, L., and Cucchiara, R. (2022). Explaining Transformer-based Image Captioning Models: An Empirical Analysis. *AI Communications*, 35:111–129.

[68] Cornia, M., Stefanini, M., Baraldi, L., and Cucchiara, R. (2020). Meshed-Memory Transformer for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[69] Dai, B. and Lin, D. (2017). Contrastive learning for image captioning. *Advances in Neural Information Processing Systems*.

[70] Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. (2023). InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*.

[71] Darcet, T., Oquab, M., Mairal, J., and Bojanowski, P. (2024). Vision Transformers Need Registers. In *Proceedings of the International Conference on Learning Representations*.

[72] Deguchi, H., Tamura, A., and Ninomiya, T. (2019). Dependency-based self-attention for transformer NMT. In *Recent Advances in Natural Language Processing*.

[73] Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J. S., Salehi, M., Muennighoff, N., Lo, K., Soldaini, L., et al. (2024). Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models. *arXiv preprint arXiv:2409.17146*.

[74] Dessì, R., Bevilacqua, M., Gualdoni, E., Rakotonirina, N. C., Franzon, F., and Baroni, M. (2023). Cross-Domain Image Captioning with Discriminative Finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[75] Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2024). Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*.

[76] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

[77] Diwan, A., Berry, L., Choi, E., Harwath, D., and Mahowald, K. (2022). Why is winoground hard? investigating failures in visuolinguistic compositionality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[78] Dong, H., Li, J., Wu, B., Wang, J., Zhang, Y., and Guo, H. (2024). Benchmarking and Improving Detail Image Caption. *arXiv preprint arXiv:2405.19092*.

[79] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*.

[80] Doveh, S., Arbelle, A., Harary, S., Herzig, R., Kim, D., Cascante-Bonilla, P., Alfassy, A., Panda, R., Giryes, R., Feris, R., Ullman, S., and Karlinsky, L. (2023a). Dense and aligned captions (DAC) promote compositional reasoning in VL models. In *Advances in Neural Information Processing Systems*.

[81] Doveh, S., Arbelle, A., Harary, S., Schwartz, E., Herzig, R., Giryes, R., Feris, R., Panda, R., Ullman, S., and Karlinsky, L. (2023b). Teaching structured vision & language concepts to vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[82] Dozat, T. and Manning, C. D. (2016). Deep biaffine attention for neural dependency parsing. In *arXiv:1611.01734*.

[83] Du, W., Lin, Z., Shen, Y., O'Donnell, T. J., Bengio, Y., and Zhang, Y. (2020). Exploiting syntactic structure for better language modeling: A syntactic distance approach. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

[84] Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. (2024). The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

[85] Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., and Cao, Y. (2023). Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[86] Friedrich, F., Schramowski, P., Brack, M., Struppek, L., Hintersdorf, D., Luccioni, S., and Kersting, K. (2023). Fair Diffusion: Instructing Text-to-Image Generation Models on Fairness. *arXiv preprint arXiv:2302.10893*.

[87] Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al. (2023). MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*.

[88] Gadre, S. Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al. (2023). Datacomp: In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems*.

[89] Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., et al. (2023). LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model. *arXiv preprint arXiv:2304.15010*.

[90] Gao, P., Zhang, R., Liu, C., Qiu, L., Huang, S., Lin, W., Zhao, S., Geng, S., Lin, Z., Jin, P., Zhang, K., Shao, W., Xu, C., He, C., He, J., Shao, H., Lu, P., Li, H., and Qiao, Y. (2024). SPHINX-X: Scaling Data and Parameters for a Family of Multi-modal Large Language Models. *arXiv preprint arXiv:2402.05935*.

[91] Ghiasi, A., Kazemi, H., Borgnia, E., Reich, S., Shu, M., Goldblum, M., Wilson, A. G., and Goldstein, T. (2022). What do vision transformers learn? A visual exploration. In *arXiv:2212.06727*.

[92] Gong, T., Lyu, C., Zhang, S., Wang, Y., Zheng, M., Zhao, Q., Liu, K., Zhang, W., Luo, P., and Chen, K. (2023). MultiModal-GPT: A Vision and Language Model for Dialogue with Humans. *arXiv preprint arXiv:2305.04790*.

[93] Goyal, A. and Bengio, Y. (2020). Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, 478.

[94] Goyal, A., Lamb, A., Hoffmann, J., Sodhani, S., Levine, S., Bengio, Y., and Schölkopf, B. (2021). Recurrent independent mechanisms. In *Proceedings of the International Conference on Learning Representations*.

[95] 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*.

[96] Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. (2024). The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

[97] Gu, A. and Dao, T. (2023). Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv preprint arXiv:2312.00752*.

[98] Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*.

[99] Gurari, D., Zhao, Y., Zhang, M., and Bhattacharya, N. (2020). Captioning Images Taken by People Who Are Blind. In *Proceedings of the European Conference on Computer Vision*.

[100] Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M. (2020). Retrieval Augmented Language Model Pre-Training. In *Proceedings of the International Conference on Machine Learning*.

[101] Hambardzumyan, K., Khachatrian, H., and May, J. (2021). Warp: Word-level adversarial reprogramming. *arXiv preprint arXiv:2101.00121*.

[102] Han, J., Gong, K., Zhang, Y., Wang, J., Zhang, K., Lin, D., Qiao, Y., Gao, P., and Yue, X. (2024). OneLLM: One Framework to Align All Modalities with Language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[103] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[104] Herzig, R., Mendelson, A., Karlinsky, L., Arbelle, A., Feris, R., Darrell, T., and Globerson, A. (2023). Incorporating structured representations into pretrained vision & language models using scene graphs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[105] Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., and Choi, Y. (2021). CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In

Proceedings of the Conference on Empirical Methods in Natural Language Processing.

[106] Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*.

[107] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. (2022). Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

[108] Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python.

[109] Hou, S., Kai, J., Xue, H., Zhu, B., Yuan, B., Huang, L., Wang, X., and Lin, Z. (2022). Syntax-guided localized self-attention by constituency syntactic distance. In *arXiv:2210.11759*.

[110] Hsieh, C.-Y., Zhang, J., Ma, Z., Kembhavi, A., and Krishna, R. (2024). SugarCrepe: fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems*.

[111] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

[112] Hu, H., Luan, Y., Chen, Y., Khandelwal, U., Joshi, M., Lee, K., Toutanova, K., and Chang, M.-W. (2023a). Open-domain Visual Entity Recognition: Towards Recognizing Millions of Wikipedia Entities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[113] Hu, W., Xu, Y., Li, Y., Li, W., Chen, Z., and Tu, Z. (2024). BLIVA: A Simple Multimodal LLM for Better Handling of Text-Rich Visual Questions. In *Proceedings of the Conference on Artificial Intelligence*.

[114] Hu, X., Gan, Z., Wang, J., Yang, Z., Liu, Z., Lu, Y., and Wang, L. (2022). Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[115] Hu, Z., Iscen, A., Sun, C., Wang, Z., Chang, K.-W., Sun, Y., Schmid, C., Ross, D. A., and Fathi, A. (2023b). REVEAL: Retrieval-Augmented Visual-Language Pre-Training With Multi-Source Multimodal Knowledge Memory.

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

[116] Huang, J., Zhang, J., Jiang, K., Qiu, H., and Lu, S. (2023a). Visual Instruction Tuning towards General-Purpose Multimodal Model: A Survey. *arXiv preprint arXiv:2312.16602*.

[117] Huang, L., Wang, W., Chen, J., and Wei, X.-Y. (2019). Attention on Attention for Image Captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[118] Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O. K., Liu, Q., et al. (2023b). Language Is Not All You Need: Aligning Perception with Language Models. *arXiv preprint arXiv:2302.14045*.

[119] Huang, Y., Tang, J., Chen, Z., Zhang, R., Zhang, X., Chen, W., Zhao, Z., Zhao, Z., Lv, T., Hu, Z., and Zhang, W. (2024). Structure-CLIP: Towards scene graph knowledge to enhance multi-modal structured representations. In *Proceedings of the Conference on Artificial Intelligence*.

[120] 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*.

[121] Iyer, S., Lin, X. V., Pasunuru, R., Mihaylov, T., Simig, D., Yu, P., Shuster, K., Wang, T., et al. (2022). OPT-IML: Scaling Language Model Instruction Meta Learning through the Lens of Generalization. *arXiv preprint arXiv:2212.12017*.

[122] Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., and Grave, E. (2021). Unsupervised Dense Information Retrieval with Contrastive Learning. *arXiv preprint arXiv:2112.09118*.

[123] Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Dwivedi-Yu, J., Joulin, A., Riedel, S., and Grave, E. (2023). Atlas: Few-shot Learning with Retrieval Augmented Language Models. *Journal of Machine Learning Research*, 24:1–43.

[124] Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., and Carreira, J. (2021). Perceiver: General perception with iterative attention. In *Proceedings of the International Conference on Machine Learning*.

[125] Jain, A., Kothyari, M., Kumar, V., Jyothi, P., Ramakrishnan, G., and Chakrabarti, S. (2021). Select, Substitute, Search: A New Benchmark for Knowledge-Augmented Visual Question Answering. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[126] Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D. d. l., Hanna, E. B., Bressand, F., et al. (2024a). Mixtral of Experts. *arXiv preprint arXiv:2401.04088*.

[127] Jiang, D., He, X., Zeng, H., Wei, C., Ku, M., Liu, Q., and Chen, W. (2024b). MANTIS: Interleaved Multi-Image Instruction Tuning. *arXiv preprint arXiv:2405.01483*.

[128] Jiang, K., He, X., Xu, R., and Wang, X. E. (2024c). ComCLIP: training-free compositional image and text matching. In *arXiv:2211.13854*.

[129] Jiang, Z., Xu, F. F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., and Neubig, G. (2023). Active Retrieval Augmented Generation. *arXiv preprint arXiv:2305.06983*.

[130] Jing, L., Li, R., Chen, Y., Jia, M., and Du, X. (2023). FAITHSCORE: Evaluating Hallucinations in Large Vision-Language Models. *arXiv preprint arXiv:2311.01477*.

[131] Johnson, J., Douze, M., and Jégou, H. (2019). Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7:535–547.

[132] Kamath, A., Hessel, J., and Chang, K.-W. (2023). Text encoders bottleneck compositionality in contrastive vision-language models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[133] Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[134] Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., and Farhadi, A. (2016). A diagram is worth a dozen images. In *Proceedings of the European Conference on Computer Vision*.

[135] Kenton, J. D. M.-W. C. and Toutanova, L. K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

[136] Kingma, D. P. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations*.

[137] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. (2023a). Segment anything. In *arXiv:2304.02643*.

[138] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023b). Segment Anything. *arXiv preprint arXiv:2304.02643*.

[139] Koh, J. Y., Fried, D., and Salakhutdinov, R. (2023a). Generating Images with Multimodal Language Models. In *Advances in Neural Information Processing Systems*.

[140] Koh, J. Y., Salakhutdinov, R., and Fried, D. (2023b). Grounding Language Models to Images for Multimodal Inputs and Outputs. In *Proceedings of the International Conference on Machine Learning*.

[141] Kornblith, S., Li, L., Wang, Z., and Nguyen, T. (2023). Guiding Image Captioning Models Toward More Specific Captions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[142] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. (2017). Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 123:32–73.

[143] Krojer, B., Poole-Dayan, E., Voleti, V., Pal, C., and Reddy, S. (2023). Are diffusion models vision-and-language reasoners? In *Advances in Neural Information Processing Systems*.

[144] Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al. (2020). The Open Images Dataset V4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale. *International Journal of Computer Vision*, 128:1956–1981.

[145] Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., and Jia, J. (2023). LISA: Reasoning Segmentation via Large Language Model. *arXiv preprint arXiv:2308.00692*.

[146] Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A., Kiela, D., et al. (2023). OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents. In *Advances in Neural Information Processing Systems*.

[147] Laurençon, H., Tronchon, L., Cord, M., and Sanh, V. (2024). What matters when building vision-language models? In *Advances in Neural Information Processing Systems*.

[148] Lerner, P., Ferret, O., and Guinaudeau, C. (2024). Cross-modal Retrieval for Knowledge-based Visual Question Answering. *arXiv preprint arXiv:2401.05736*.

[149] Lerner, P., Ferret, O., Guinaudeau, C., Le Borgne, H., Besançon, R., Moreno, J. G., and Lovón Melgarejo, J. (2022). ViQuAE, a dataset for knowledge-based visual question answering about named entities. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[150] Lester, B., Al-Rfou, R., and Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[151] Leviathan, Y., Kalman, M., and Matias, Y. (2024). Selective attention improves transformer. In *arXiv:2410.02703*.

[152] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*.

[153] Li, A. C., Prabhudesai, M., Duggal, S., Brown, E., and Pathak, D. (2023a). Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[154] Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., and Shan, Y. (2023b). SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension. *arXiv preprint arXiv:2307.16125*.

[155] Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., and Liu, Z. (2023c). Otter: A Multi-Modal Model with In-Context Instruction Tuning. *arXiv preprint arXiv:2305.03726*.

[156] Li, F., Zhang, H., Sun, P., Zou, X., Liu, S., Yang, J., Li, C., Zhang, L., and Gao, J. (2023d). Semantic-SAM: Segment and recognize anything at any granularity. *arXiv preprint arXiv:2307.04767*.

[157] Li, J., Chen, D., Hong, Y., Chen, Z., Chen, P., Shen, Y., and Gan, C. (2024a). CoVLM: Composing visual entities and relationships in large language models via communicative decoding. In *Proceedings of the International Conference on Learning Representations*.

[158] Li, J., Li, D., Savarese, S., and Hoi, S. (2023e). BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the International Conference on Machine Learning*.

[159] Li, J., Li, D., Xiong, C., and Hoi, S. (2022a). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the International Conference on Machine Learning*.

[160] Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. (2019). VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv preprint arXiv:1908.03557*.

[161] Li, R., Wu, Y., and He, X. (2024b). Learning by correction: Efficient tuning task for zero-shot generative vision-language reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[162] Li, X., Tu, H., Hui, M., Wang, Z., Zhao, B., Xiao, J., Ren, S., Mei, J., Liu, Q., Zheng, H., et al. (2024c). What If We Recaption Billions of Web Images with LLaMA-3? *arXiv preprint arXiv:2406.08478*.

[163] Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al. (2020). Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *Proceedings of the European Conference on Computer Vision*.

[164] Li, X. L. and Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

[165] Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J.-R. (2023f). Evaluating Object Hallucination in Large Vision-Language Models. *arXiv preprint arXiv:2305.10355*.

[166] Li, Y., Pan, Y., Yao, T., and Mei, T. (2022b). Comprehending and ordering semantics for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[167] Li, Y., Zhang, C., Yu, G., Wang, Z., Fu, B., Lin, G., Shen, C., Chen, L., and Wei, Y. (2023g). StableLLaVA: Enhanced Visual Instruction Tuning with Synthesized Image-Dialogue Data. *arXiv preprint arXiv:2308.10253*.

[168] Li, Z., Yang, B., Liu, Q., Ma, Z., Zhang, S., Yang, J., Sun, Y., Liu, Y., and Bai, X. (2023h). Monkey: Image Resolution and Text Label Are Important Things for Large Multi-modal Models. *arXiv preprint arXiv:2311.06607*.

[169] Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics Workshops*.

[170] Lin, J., Yin, H., Ping, W., Lu, Y., Molchanov, P., Tao, A., Mao, H., Kautz, J., Shoeybi, M., and Han, S. (2023a). VILA: On Pre-training for Visual Language Models. *arXiv preprint arXiv:2312.07533*.

[171] 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*.

[172] Lin, W., Mei, J., Chen, J., and Byrne, B. (2024a). PreFLMR: Scaling Up Fine-Grained Late-Interaction Multi-modal Retrievers. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

[173] Lin, Z., Chen, X., Pathak, D., Zhang, P., and Ramanan, D. (2024b). Revisiting the role of language priors in vision-language models. In *Proceedings of the International Conference on Machine Learning*.

[174] Lin, Z., Liu, C., Zhang, R., Gao, P., Qiu, L., Xiao, H., Qiu, H., Lin, C., Shao, W., Chen, K., et al. (2023b). SPHINX: The Joint Mixing of Weights, Tasks, and Visual Embeddings for Multi-modal Large Language Models. *arXiv preprint arXiv:2311.07575*.

[175] Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., and Wang, L. (2023a). Aligning Large Multi-Modal Model with Robust Instruction Tuning. *arXiv preprint arXiv:2306.14565*.

[176] Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., and Wang, L. (2023b). Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning. *arXiv preprint arXiv:2306.14565*.

[177] Liu, H., Li, C., Li, Y., and Lee, Y. J. (2024a). Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[178] Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., and Lee, Y. J. (2024b). LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/.

[179] Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023c). Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*.

[180] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al. (2023d). Grounding DINO: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

[181] Liu, S., Zhu, Z., Ye, N., Guadarrama, S., and Murphy, K. (2017). Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[182] Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., and Tang, J. (2023e). GPT understands, too. *AI Open*.

[183] Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al. (2023f). MMBench: Is Your Multi-modal Model an All-around Player? *arXiv preprint arXiv:2307.06281*.

[184] Liu, Z., Zhu, L., Shi, B., Zhang, Z., Lou, Y., Yang, S., Xi, H., Cao, S., Gu, Y., Li, D., et al. (2024c). NVILA: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*.

[185] Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

[186] Lu, H., Liu, W., Zhang, B., Wang, B., Dong, K., Liu, B., Sun, J., Ren, T., et al. (2024). DeepSeek-VL: Towards Real-World Vision-Language Understanding. *arXiv preprint arXiv:2403.05525*.

[187] Lu, J., Batra, D., Parikh, D., and Lee, S. (2019). ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems*.

[188] Lu, J., Xiong, C., Parikh, D., and Socher, R. (2017). Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[189] Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. (2022). Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In *Advances in Neural Information Processing Systems*.

[190] Luo, G., Zhou, Y., Ren, T., Chen, S., Sun, X., and Ji, R. (2023). Cheap and Quick: Efficient Vision-Language Instruction Tuning for Large Language Models. *arXiv preprint arXiv:2305.15023*.

[191] Luo, R., Price, B., Cohen, S., and Shakhnarovich, G. (2018). Discriminability objective for training descriptive captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[192] Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

[193] Marino, K., Rastegari, M., Farhadi, A., and Mottaghi, R. (2019). OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[194] Mensink, T., Uijlings, J., Castrejon, L., Goel, A., Cadar, F., Zhou, H., Sha, F., Araujo, A., and Ferrari, V. (2023). Encyclopedic VQA: Visual questions about detailed properties of fine-grained categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[195] Methani, N., Ganguly, P., Khapra, M. M., and Kumar, P. (2020). PlotQA: Reasoning over Scientific Plots. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.

[196] Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., and Wu, H. (2018). Mixed Precision Training. In *Proceedings of the International Conference on Learning Representations*.

[197] Mokady, R., Hertz, A., and Bermano, A. H. (2021). ClipCap: CLIP Prefix for Image Captioning. *arXiv preprint arXiv:2111.09734*.

[198] Momeni, L., Caron, M., Nagrani, A., Zisserman, A., and Schmid, C. (2023). Verbs in action: Improving verb understanding in video-language models. In *arXiv:2304.06708*.

[199] MosaicML (2023). Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs.

[200] Mrini, K., Dernoncourt, F., Tran, Q., Bui, T., Chang, W., and Nakashole, N. (2019). Rethinking self-attention: Towards interpretability in neural parsing. In *arXiv:1911.03875*.

[201] Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., et al. (2021). WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

[202] Nivre, J. (2005). Dependency grammar and dependency parsing. https://api.semanticscholar.org/CorpusID:16318436.

[203] Oh, Y., Cho, J. W., Kim, D.-J., Kweon, I. S., and Kim, J. (2024). Preserving multi-modal capabilities of pre-trained VLMs for improving vision-linguistic compositionality. In *arXiv:2410.05210*.

[204] OpenAI (2022). Introducing ChatGPT.

[205] OpenAI (2023). GPT-4V(ision) System Card. https://api.semanticscholar.org/CorpusID:263218031.

[206] 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*, pages 1–31.

[207] Ordonez, V., Kulkarni, G., and Berg, T. (2011). Im2Text: Describing Images Using 1 Million Captioned Photographs. In *Advances in Neural Information Processing Systems*.

[208] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

[209] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

[210] Parascandolo, G., Rojas-Carulla, M., Kilbertus, N., and Schölkopf, B. (2018). Learning independent causal mechanisms. In *Proceedings of the International Conference on Machine Learning*.

[211] Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146.

[212] Pearl, J. and Verma, T. S. (1995). A theory of inferred causation. *Logic, Methodology and Philosophy of Science IX*, 134:789–811.

[213] Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., and Wei, F. (2023). Kosmos-2: Grounding Multimodal Large Language Models to the World. *arXiv preprint arXiv:2306.14824*.

[214] Perry, R., von Kügelgen, J., and Schölkopf, B. (2022). Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. In *Advances in Neural Information Processing Systems*.

[215] Pi, R., Gao, J., Diao, S., Pan, R., Dong, H., Zhang, J., Yao, L., Han, J., Xu, H., and Zhang, L. K. T. (2023). DetGPT: Detect What You Need via Reasoning. *arXiv preprint arXiv:2305.14167*.

[216] Pi, R., Han, T., Xie, Y., Pan, R., Lian, Q., Dong, H., Zhang, J., and Zhang, T. (2024). MLLM-Protector: Ensuring MLLM's Safety without Hurting Performance. *arXiv preprint arXiv:2401.02906*.

[217] Pollastri, F., Maronas, J., Bolelli, F., Ligabue, G., Paredes, R., Magistroni, R., and Grana, C. (2021a). Confidence calibration for deep renal biopsy immunofluorescence image classification. In *Proceedings of the International Conference on Pattern Recognition*.

[218] Pollastri, F., Parreño, M., Maroñas, J., Bolelli, F., Paredes, R., Ramos, D., and Grana, C. (2021b). A deep analysis on high-resolution dermoscopic image classification. *IET Computer Vision*, 15:514–526.

[219] Poppi, S., Poppi, T., Cocchi, F., Cornia, M., Baraldi, L., and Cucchiara, R. (2024). Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models. In *Proceedings of the European Conference on Computer Vision*.

[220] Pramanick, S., Han, G., Hou, R., Nag, S., Lim, S.-N., Ballas, N., Wang, Q., Chellappa, R., and Almahairi, A. (2023). Jack of All Tasks, Master of Many: Designing General-purpose Coarse-to-Fine Vision-Language Model. *arXiv preprint arXiv:2312.12423*.

[221] Qi, J., Xu, Z., Shao, R., Chen, Y., Di, J., Cheng, Y., Wang, Q., and Huang, L. (2024a). RoRA-VLM: Robust Retrieval-Augmented Vision Language Models. *arXiv preprint arXiv:2410.08876*.

[222] Qi, L., Chen, Y.-W., Yang, L., Shen, T., Li, X., Guo, W., Xu, Y., and Yang, M.-H. (2024b). Generalizable Entity Grounding via Assistance of Large Language Model. *arXiv preprint arXiv:2402.02555*.

[223] Qiao, Y., Yu, Z., Guo, L., Chen, S., Zhao, Z., Sun, M., Wu, Q., and Liu, J. (2024). VL-Mamba: Exploring State Space Models for Multimodal Learning. *arXiv preprint arXiv:2403.13600*.

[224] Qiu, J., Madotto, A., Lin, Z., Crook, P. A., Xu, Y. E., Dong, X. L., Faloutsos, C., Li, L., Damavandi, B., and Moon, S. (2024). SnapNTell: Enhancing Entity-Centric Visual Question Answering with Retrieval Augmented Multimodal LLM. *arXiv preprint arXiv:2403.04735*.

[225] 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*.

[226] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI blog*.

[227] Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. (2024). Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*.

[228] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:5485–5551.

[229] Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., and Shoham, Y. (2023). In-Context Retrieval-Augmented Language Models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

[230] Ramos, R., Martins, B., Elliott, D., and Kementchedjhieva, Y. (2023). Smallcap: Lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[231] Ranzato, M., Chopra, S., Auli, M., and Zaremba, W. (2015). Sequence Level Training with Recurrent Neural Networks. In *Proceedings of the International Conference on Learning Representations*.

[232] Rasheed, H., Maaz, M., Khan, S., and Khan, F. S. (2024). LLaVA++: Extending Visual Capabilities with LLaMA-3 and Phi-3. https://github.com/mbzuai-oryx/LLaVA-pp.

[233] Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Cholakkal, H., Anwer, R. M., Xing, E., Yang, M.-H., and Khan, F. S. (2023). GLaMM : Pixel Grounding Large Multimodal Model. *arXiv preprint arXiv:2311.03356*.

[234] Ren, Z., Huang, Z., Wei, Y., Zhao, Y., Fu, D., Feng, J., and Jin, X. (2023). PixelLM: Pixel Reasoning with Large Multimodal Model. *arXiv preprint arXiv:2312.02228*.

[235] Ren, Z., Wang, X., Zhang, N., Lv, X., and Li, L.-J. (2017). Deep reinforcement learning-based image captioning with embedding reward. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[236] 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*.

[237] 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*.

[238] Rotstein, N., Bensaid, D., Brody, S., Ganz, R., and Kimmel, R. (2024). FuseCap: Leveraging Large Language Models to Fuse Visual Data into Enriched Image Captions. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.

[239] Sahin, U., Li, H., Khan, Q., Cremers, D., and Tresp, V. (2024). Enhancing multimodal compositional reasoning of visual language models with generative negative mining. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.

[240] Sarto, S., Barraco, M., Cornia, M., Baraldi, L., and Cucchiara, R. (2023). Positive-augmented contrastive learning for image and video captioning evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[241] Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, 109:612–634.

[242] Schramowski, P., Brack, M., Deiseroth, B., and Kersting, K. (2023). Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[243] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. (2022). LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*.

[244] Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. (2021). LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. In *NeurIPS Workshops*.

[245] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*.

[246] Schwenk, D., Khandelwal, A., Clark, C., Marino, K., and Mottaghi, R. (2022). A-OKVQA: A Benchmark for Visual Question Answering Using World Knowledge. In *Proceedings of the European Conference on Computer Vision*.

[247] Sharma, P., Ding, N., Goodman, S., and Soricut, R. (2018). Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

[248] Shen, S., Li, L. H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.-W., Yao, Z., and Keutzer, K. (2022). How Much Can CLIP Benefit Vision-and-Language Tasks? In *Proceedings of the International Conference on Learning Representations*.

[249] Shi, B., Wu, Z., Mao, M., Wang, X., and Darrell, T. (2024). When Do We Not Need Larger Vision Models? In *Proceedings of the European Conference on Computer Vision*.

[250] Silveira, N., Dozat, T., de Marneffe, M.-C., Bowman, S., Connor, M., Bauer, J., and Manning, C. (2014). A gold standard dependency corpus for English. In *International Conference on Language Resources and Evaluation*.

[251] Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., and Kiela, D. (2021). FLAVA: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[252] Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., and Rohrbach, M. (2019). Towards VQA Models That Can Read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[253] Singh, D., Khan, Z., and Tapaswi, M. (2024). FiGCLIP: Fine-Grained CLIP Adaptation via Densely Annotated Videos. In *arXiv:2401.07669*.

[254] Singh, H., Zhang, P., Wang, Q., Wang, M., Xiong, W., Du, J., and Chen, Y. (2023). Coarse-to-fine contrastive learning in image-text-graph space for improved vision-language compositionality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[255] Socher, R. and Fei-Fei, L. (2010). Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[256] Song, H., Dong, L., Zhang, W.-N., Liu, T., and Wei, F. (2022). CLIP Models are few-shot learners: Empirical studies on VQA and visual entailment. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

[257] Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., and Cucchiara, R. (2022). From Show to Tell: A Survey on Deep Learning-based Image Captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:539–559.

[258] Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. (2020). Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*.

[259] Sun, Q., Wang, J., Yu, Q., Cui, Y., Zhang, F., Zhang, X., and Wang, X. (2024). EVA-CLIP-18B: Scaling CLIP to 18 Billion Parameters. *arXiv preprint arXiv:2402.04252*.

[260] Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., and Wang, X. (2023a). Generative Pretraining in Multimodality. *arXiv preprint arXiv:2307.05222*.

[261] Sun, Z., Shen, Y., Zhou, Q., Zhang, H., Chen, Z., Cox, D., Yang, Y., and Gan, C. (2023b). Principle-driven self-alignment of language models from scratch with minimal human supervision. In *Advances in Neural Information Processing Systems*.

[262] Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. (2023). Stanford Alpaca: An Instruction-Following LLaMA Model. https://github.com/tatsu-lab/stanford_alpaca.

[263] Tay, Y., Dehghani, M., Tran, V. Q., Garcia, X., Bahri, D., Schuster, T., Zheng, H. S., Houlsby, N., and Metzler, D. (2022). Unifying Language Learning Paradigms. *arXiv preprint arXiv:2205.05131*.

[264] Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., et al. (2024). Gemma 2: Improving Open Language Models at a Practical Size. *arXiv preprint arXiv:2408.00118*.

[265] Tewel, Y., Shalev, Y., Schwartz, I., and Wolf, L. (2022). ZeroCap: Zero-Shot Image-to-Text Generation for Visual-Semantic Arithmetic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[266] Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., and Ross, C. (2022). Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*.

[267] Tian, Y., Ma, T., Xie, L., Qiu, J., Tang, X., Zhang, Y., Jiao, J., Tian, Q., and Ye, Q. (2024). ChatterBox: Multi-round Multimodal Referring and Grounding. *arXiv preprint arXiv:2401.13307*.

[268] Tian, Y., Xie, L., Wang, Z., Wei, L., Zhang, X., Jiao, J., Wang, Y., Tian, Q., and Ye, Q. (2023). Integrally Pre-Trained Transformer Pyramid Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[269] Tong, P., Brown, E., Wu, P., Woo, S., IYER, A. J. V., Akula, S. C., Yang, S., Yang, J., Middepogu, M., Wang, Z., et al. (2024a). Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. In *Advances in Neural Information Processing Systems*.

[270] Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., and Xie, S. (2024b). Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[271] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023a). LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

[272] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023b). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

[273] Tschannen, M., Gritsenko, A., Wang, X., Naeem, M. F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al. (2025). SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *arXiv preprint arXiv:2502.14786*.

[274] Tschannen, M., Kumar, M., Steiner, A. P., Zhai, X., Houlsby, N., and Beyer, L. (2023). Image captioners are scalable vision learners too. In *Advances in Neural Information Processing Systems*.

[275] Tsimpoukelli, M., Menick, J. L., Cabi, S., Eslami, S., Vinyals, O., and Hill, F. (2021). Multimodal Few-Shot Learning with Frozen Language Models. In *Advances in Neural Information Processing Systems*.

[276] Van Den Oord, A., Vinyals, O., et al. (2017). Neural Discrete Representation Learning. In *Advances in Neural Information Processing Systems*.

[277] Van Horn, G., Cole, E., Beery, S., Wilber, K., Belongie, S., and Mac Aodha, O. (2021). Benchmarking Representation Learning for Natural World Image Collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[278] 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*.

[279] Vedantam, R., Lawrence Zitnick, C., and Parikh, D. (2015). CIDEr: Consensus-based Image Description Evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[280] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[281] Wan, D., Cho, J., Stengel-Eskin, E., and Bansal, M. (2024). Contrastive region guidance: Improving grounding in vision-language models without training. In *arXiv:2403.02325*.

[282] Wang, B., Ping, W., McAfee, L., Xu, P., Li, B., Shoeybi, M., and Catanzaro, B. (2024). InstructRetro: Instruction Tuning post Retrieval-Augmented Pretraining. In *Proceedings of the International Conference on Machine Learning*.

[283] Wang, B., Wu, F., Han, X., Peng, J., Zhong, H., Zhang, P., Dong, X., Li, W., Li, W., Wang, J., et al. (2023a). VIGC: Visual instruction generation and correction. *arXiv preprint arXiv:2308.12714*.

[284] Wang, H., Ma, S., Huang, S., Dong, L., Wang, W., Peng, Z., Wu, Y., Bajaj, P., Singhal, S., Benhaim, A., et al. (2023b). Magneto: A Foundation Transformer. In *Proceedings of the International Conference on Machine Learning*.

[285] Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., and Wang, L. (2022a). GIT: A Generative Image-to-text Transformer for Vision and Language. *arXiv preprint arXiv:2205.14100*.

[286] Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., and Yang, H. (2022b). OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. In *Proceedings of the International Conference on Machine Learning*.

[287] Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., et al. (2023c). VisionLLM: Large Language Model is also an Open-Ended Decoder for Vision-Centric Tasks. *arXiv preprint arXiv:2305.11175*.

[288] Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., et al. (2023d). CogVLM: Visual Expert for Pretrained Language Models. *arXiv preprint arXiv:2311.03079*.

[289] Wang, W., Shi, M., Li, Q., Wang, W., Huang, Z., Xing, L., Chen, Z., Li, H., Zhu, X., Cao, Z., et al. (2023e). The All-Seeing Project: Towards Panoptic Visual Recognition and Understanding of the Open World. *arXiv preprint arXiv:2308.01907*.

[290] Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. (2023f). Self-Instruct: Aligning Language Models with Self-Generated Instructions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

[291] Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Arunkumar, A., Ashok, A., Dhanasekaran, A. S., Naik, A., Stap, D., et al. (2022c). Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[292] Wazni, H., Lo, K. I., and Sadrzadeh, M. (2024). VerbCLIP: Improving verb understanding in vision-language models with compositional structures. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*.

[293] Wei, C., Chen, Y., Chen, H., Hu, H., Zhang, G., Fu, J., Ritter, A., and Chen, W. (2023a). Uniir: Training and Benchmarking Universal Multimodal Information Retrievers. *arXiv preprint arXiv:2311.17136*.

[294] Wei, F., Zhang, X., Zhang, A., Zhang, B., and Chu, X. (2023b). Lenna: Language enhanced reasoning detection assistant. *arXiv preprint arXiv:2312.02433*.

[295] Weyand, T., Araujo, A., Cao, B., and Sim, J. (2020). Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[296] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[297] Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith, S., Roelofs, R., Lopes, R. G., Hajishirzi, H., Farhadi, A., Namkoong, H., and Schmidt, L. (2022). Robust Fine-Tuning of Zero-Shot Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[298] Wu, J., Gan, W., Chen, Z., Wan, S., and Philip, S. Y. (2023a). Multimodal Large Language Models: A Survey. In *International Conference on Big Data*.

[299] Wu, T.-H., Biamby, G., Chan, D., Dunlap, L., Gupta, R., Wang, X., Gonzalez, J. E., and Darrell, T. (2023b). See, Say, and Segment: Teaching LMMs to Overcome False Premises. *arXiv preprint arXiv:2312.08366*.

[300] Wu, T.-H., Gonzalez, J. E., Darrell, T., and Chan, D. M. (2024). CLAIR-A: Leveraging Large Language Models to Judge Audio Captions. *arXiv preprint arXiv:2409.12962*.

[301] Wysoczańska, M., Siméoni, O., Ramamonjisoa, M., Bursuc, A., Trzciński, T., and Pérez, P. (2024). CLIP-DINOiser: Teaching CLIP a few DINO tricks for open-vocabulary semantic segmentation. In *arXiv:2312.12359*.

[302] Xia, Z., Han, D., Han, Y., Pan, X., Song, S., and Huang, G. (2023). GSVA: Generalized Segmentation via Multimodal Large Language Models. *arXiv preprint arXiv:2312.10103*.

[303] Xiao, Z., Gong, M., Cascante-Bonilla, P., Zhang, X., Wu, J., and Ordonez, V. (2024). Grounding Language Models for Visual Entity Recognition. *arXiv preprint arXiv:2402.18695*.

[304] Xiong, Y., Dai, B., and Lin, D. (2018). Move Forward and Tell: A Progressive Generator of Video Descriptions. In *Proceedings of the European Conference on Computer Vision*.

[305] Xu, J., Zhou, X., Yan, S., Gu, X., Arnab, A., Sun, C., Wang, X., and Schmid, C. (2023). Pixel Aligned Language Models. *arXiv preprint arXiv:2312.09237*.

[306] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning*.

[307] Xu, R., Yao, Y., Guo, Z., Cui, J., Ni, Z., Ge, C., Chua, T.-S., Liu, Z., Sun, M., and Huang, G. (2024). LLaVA-UHD: an LMM Perceiving Any Aspect Ratio and High-Resolution Images. *arXiv preprint arXiv:2403.11703*.

[308] Xuan, S., Guo, Q., Yang, M., and Zhang, S. (2023). Pink: Unveiling the Power of Referential Comprehension for Multi-modal LLMs. *arXiv preprint arXiv:2310.00582*.

[309] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2020). mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

[310] Yan, Y. and Xie, W. (2024). EchoSight: Advancing Visual-Language Models with Wiki Knowledge. *arXiv preprint arXiv:2407.12735*.

[311] Yang, S., Qu, T., Lai, X., Tian, Z., Peng, B., Liu, S., and Jia, J. (2023). LISA++: An Improved Baseline for Reasoning Segmentation with Large Language Model. *arXiv preprint arXiv:2312.17240*.

[312] Yang, X., Tang, K., Zhang, H., and Cai, J. (2019). Auto-Encoding Scene Graphs for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[313] Yang, Z. and Wan, X. (2022). Dependency-based mixture language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

[314] Yao, B. Z., Yang, X., Lin, L., Lee, M. W., and Zhu, S.-C. (2010). I2t: Image parsing to text description. *Proceedings of the IEEE*, 98.

[315] Yao, T., Pan, Y., Li, Y., and Mei, T. (2018). Exploring Visual Relationship for Image Captioning. In *Proceedings of the European Conference on Computer Vision*.

[316] Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al. (2023a). mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality. *arXiv preprint arXiv:2304.14178*.

[317] Ye, Q., Xu, H., Ye, J., Yan, M., Hu, A., Liu, H., Qian, Q., Zhang, J., and Huang, F. (2024a). mPLUG-Owl2: Revolutionizing Multi-modal Large Language Model with Modality Collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[318] Ye, Q., Xu, H., Ye, J., Yan, M., Liu, H., Qian, Q., Zhang, J., Huang, F., and Zhou, J. (2023b). mPLUG-Owl2: Revolutionizing Multi-modal Large Language Model with Modality Collaboration. *arXiv preprint arXiv:2311.04257*.

[319] Ye, T., Dong, L., Xia, Y., Sun, Y., Zhu, Y., Huang, G., and Wei, F. (2024b). Differential transformer. In *arXiv:2410.05258*.

[320] Yellinek, N., Karlinsky, L., and Giryes, R. (2023). 3VL: using trees to teach vision & language models compositional concepts. In *arXiv:2312.17345*.

[321] Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., and Chen, E. (2023a). A Survey on Multimodal Large Language Models. *arXiv preprint arXiv:2306.13549*.

[322] Yin, S., Fu, C., Zhao, S., Xu, T., Wang, H., Sui, D., Shen, Y., Li, K., Sun, X., and Chen, E. (2023b). Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*.

[323] You, H., Zhang, H., Gan, Z., Du, X., Zhang, B., Wang, Z., Cao, L., Chang, S.-F., and Yang, Y. (2023). Ferret: Refer and Ground Anything Anywhere at Any Granularity. *arXiv preprint arXiv:2310.07704*.

[324] Yu, Y., Chung, J., Yun, H., Hessel, J., Park, J., Lu, X., Ammanabrolu, P., Zellers, R., Bras, R. L., Kim, G., and Choi, Y. (2022). Multimodal knowledge alignment with reinforcement learning. *arXiv preprint arXiv:2205.12630*.

[325] Yu, Y., Ping, W., Liu, Z., Wang, B., You, J., Zhang, C., Shoeybi, M., and Catanzaro, B. (2024). RankRAG: Unifying Context Ranking with Retrieval-Augmented Generation in LLMs. *arXiv preprint arXiv:2407.02485*.

[326] Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al. (2023). MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. *arXiv preprint arXiv:2311.16502*.

[327] Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., and Zou, J. Y. (2023). When and why vision-language models behave like bags-of-words, and what to do about it? In *Proceedings of the International Conference on Learning Representations*.

[328] Zang, Y., Li, W., Han, J., Zhou, K., and Loy, C. C. (2023). Contextual Object Detection with Multimodal Large Language Models. *arXiv preprint arXiv:2305.18279*.

[329] Zeng, Y., Zhang, X., and Li, H. (2022). Multi-grained vision language pre-training: Aligning texts with visual concepts. In *Proceedings of the International Conference on Machine Learning*.

[330] Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. (2023). Sigmoid Loss for Language Image Pre-Training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[331] Zhan, Y., Zhu, Y., Chen, Z., Yang, F., Tang, M., and Wang, J. (2023). Griffon: Spelling out All Object Locations at Any Granularity with Large Language Models. *arXiv preprint arXiv:2311.14552*.

[332] Zhang, A., Zhao, L., Xie, C.-W., Zheng, Y., Ji, W., and Chua, T.-S. (2023a). NExT-Chat: An LMM for Chat, Detection and Segmentation. *arXiv preprint arXiv:2311.04498*.

[333] Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L. M., and Shum, H.-Y. (2022a). DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. *arXiv preprint arXiv:2203.03605*.

[334] Zhang, H., Li, F., Zou, X., Liu, S., Li, C., Yang, J., and Zhang, L. (2023b). A simple framework for open-vocabulary segmentation and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[335] Zhang, H., Li, H., Li, F., Ren, T., Zou, X., Liu, S., Huang, S., Gao, J., Zhang, L., Li, C., et al. (2023c). LLaVA-Grounding: Grounded Visual Chat with Large Multimodal Models. *arXiv preprint arXiv:2312.02949*.

[336] Zhang, L., Awal, R., and Agrawal, A. (2024). Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[337] Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. (2021). VinVL: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[338] Zhang, S., Lijie, W., Xiao, X., and Wu, H. (2022b). Syntax-guided contrastive learning for pre-trained language model. In *Findings of the Association for Computational Linguistics*.

[339] Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. (2022c). Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

[340] Zhang, S., Sun, P., Chen, S., Xiao, M., Shao, W., Zhang, W., Chen, K., and Luo, P. (2023d). GPT4RoI: Instruction Tuning Large Language Model on Region-of-Interest. *arXiv preprint arXiv:2307.03601*.

[341] Zhang, Y., Li, Z., and Zhang, M. (2020). Efficient second-order TreeCRF for neural dependency parsing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

[342] Zhang, Y., Wang, J., Wu, H., and Xu, W. (2022d). Distinctive Image Captioning via CLIP Guided Group Optimization. In *Proceedings of the European Conference on Computer Vision*.

[343] Zhang, Y., Zhang, R., Gu, J., Zhou, Y., Lipka, N., Yang, D., and Sun, T. (2023e). LLaVAR: Enhanced Visual Instruction Tuning for Text-Rich Image Understanding. *arXiv preprint arXiv:2306.17107*.

[344] Zhao, B., Wu, B., and Huang, T. (2023a). SVIT: Scaling up Visual Instruction Tuning. *arXiv preprint arXiv:2307.04087*.

[345] Zhao, H., Zhang, M., Zhao, W., Ding, P., Huang, S., and Wang, D. (2024). Cobra: Extending Mamba to Multi-Modal Large Language Model for Efficient Inference. *arXiv preprint arXiv:2403.14520*.

[346] Zhao, L., Yu, E., Ge, Z., Yang, J., Wei, H., Zhou, H., Sun, J., Peng, Y., Dong, R., Han, C., et al. (2023b). ChatSpot: Bootstrapping Multimodal LLMs via Precise Referring Instruction Tuning. *arXiv preprint arXiv:2307.09474*.

[347] Zhao, T., Zhang, T., Zhu, M., Shen, H., Lee, K., Lu, X., and Yin, J. (2022). Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. In *arXiv:2207.00221*.

[348] Zhao, Y., Lin, Z., Zhou, D., Huang, Z., Feng, J., and Kang, B. (2023c). BuboGPT: Enabling Visual Grounding in Multi-Modal LLMs. *arXiv preprint arXiv:2307.08581*.

[349] Zheng, C., Zhang, J., Kembhavi, A., and Krishna, R. (2024). Iterated learning improves compositionality in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[350] Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L. H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al. (2022). RegionCLIP: Region-based Language-Image Pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[351] Zhu, D., Chen, J., Haydarov, K., Shen, X., Zhang, W., and Elhoseiny, M. (2023a). ChatGPT Asks, BLIP-2 Answers: Automatic Questioning Towards Enriched Visual Descriptions. *arXiv preprint arXiv:2303.06594*.

[352] Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. (2023b). MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv preprint arXiv:2304.10592*.

[353] Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. (2024a). MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *Proceedings of the International Conference on Learning Representations*.

[354] Zhu, L., Chen, T., Ji, D., Ye, J., and Liu, J. (2023c). LLaFS: When Large-Language Models Meet Few-Shot Segmentation. *arXiv preprint arXiv:2311.16926*.

[355] Zhu, W., Hessel, J., Awadalla, A., Gadre, S. Y., Dodge, J., Fang, A., Yu, Y., Schmidt, L., Wang, W. Y., and Choi, Y. (2023d). Multimodal c4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*.

[356] Zhu, Y., Zhu, M., Liu, N., Xu, Z., and Peng, Y. (2024b). LLaVA-Phi: Efficient Multi-Modal Assistant with Small Language Model. In *Proceedings of the ACM International Conference on Multimedia Workshops*.

[357] Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. (2019). Fine-Tuning Language Models from Human Preferences. *arXiv preprint arXiv:1909.08593*.

A

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.

A.0.1 Research Internships

May–October 2025: Research Intern at **Amazon Science** (Cambridge, United Kingdom). Research activities on large-scale multimodal foundation models and vision–language reasoning.

A.0.2 Participation in National and International Research Projects

MUCES (Multimedia Platform for Content Enrichment and Search in Audiovisual Archives): Research activities on multimodal deep learning for large-scale audiovisual understanding and retrieval, with a focus on integrating vision and language for searchable cultural heritage archives.

FAIR (Future Artificial Intelligence Research): Participation in the FAIR national initiative (Spoke 8, Pervasive AI) and in cross-cutting research on

vision–language and multimodal challenges, with an emphasis on scalable multimodal learning and reasoning.

ELLIOT (European Large Open Multi-Modal Foundation Models): Research contributions within an EU Horizon Europe project on open, large-scale multimodal foundation models, focusing on vision–language modeling, multimodal representation learning, and robust generalization across heterogeneous data streams.

ELIAS (European Lighthouse of AI for Sustainability): Research activities within the ELIAS network, with a focus on scalable learning algorithms, multimodal modeling, and large-scale machine learning systems.

A.0.3 Teaching Activities

2024: Lecturer in "Multimodal (Large) Language Models", Intensive Master for "AI and ML for Smart Factory".

2023–2024: Guest Lecturer in the M.Sc. course "Scalable AI".

A.0.4 Conference Attendance

29 September–4 October 2024: The 18th European Conference on Computer Vision (ECCV 2024), *Milan, Italy*.

25–28 November 2024: The 35th British Machine Vision Conference (BMVC 2024), *Glasgow, Scotland*.

1–5 December 2024: The 27th International Conference on Pattern Recognition (ICPR 2024), *Kolkata, India*.

24–28 April 2025: The Thirteenth International Conference on Learning Representations (ICLR 2025), *Singapore*.

A.0.5 Seminars and Workshops

November 2022: Attendance at "Digital Humanities and Artificial Intelligence for humans in today's society" seminar by Prof. Rita Cucchiara.

November 2022: Attendance at "Graph Signal Processing for Machine Learning: Challenges and Use Cases" seminar by Prof. Laura Toni.

December 2022: Attendance at "From Handcrafted to End-to-End Learning, and Back: a Journey for Multi-Object Tracking" seminar by Prof. Laura Leal-Taixé.

December 2022: Attendance at "3D Computer Vision for Animals" seminar by Prof. Silvia Zuffi.

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

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

A.0.6 Schools

4–8 September 2023: Attendance and completion of the "International Summer School on Machine Vision 2023 (VISMAC)".

18–22 September 2023: Attendance and completion of the "ELLIS Summer School on Large-Scale AI for Research and Industry 2023".

23–27 September 2024: Attendance and completion of the "2024 IEEE-EURASIP Summer School on Signal Processing (S3P-2024)".

A.0.7 Reviewing Service

Conferences.

IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR).

European Conference on Computer Vision (ECCV).

British Machine Vision Conference (BMVC).

IEEE International Conference on Pattern Recognition (ICPR).

ACM International Conference on Multimedia (ACM Multimedia).

Journals.

IAPR Pattern Recognition Letters (PRL).

Computer Vision and Image Understanding (CVIU).

Workshops.

Workshop on Automatically Domain-Adapted and Personalized Document Analysis (ADAPDA).

Workshop on Computational Aspects of Deep Learning (CADL).

A.0.8 Co-supervision of M.Sc. Theses

2025: *Alberto Compagnoni*, "Reducing Hallucinations in Multimodal LLMs through Direct Preference Optimization".

2024: *Davide Bucciarelli*, "Personalizing Multimodal Large Language Models for Image Captioning: An Experimental Analysis".

2024: *Antonio De Blasi*, "Multimodal Retrieval for Large Multimodal Models: an Experimental Analysis".

2023: *Noemi Scarlino*, "Improving Image Captioning by incorporating Negative Samples into an Enhanced CLIP Architecture".

B

Supplementary material for Chapter 3

This appendix provides concise complementary material for Chapter 3. We focus on (i) the key design choices behind DiCO and the corresponding ablations, and (ii) the essential details required to reproduce the experimental protocol. Unless otherwise stated, extended quantitative results, plots, qualitative examples, and interface or prompt artefacts are reported in the associated publication and its supplementary material.

B.1 Key design choices and ablations

B.1.1 Early-stopping criterion

When fine-tuning with a reference-free reward, we select the checkpoint using the validation value of the corresponding *reference-based* variant of the reward. Specifically, CLIP-S and PAC-S optimization are early-stopped according to RefCLIP-S and RefPAC-S on the validation set, respectively. This choice mitigates reward exploitation while preserving caption quality and alignment with human references.

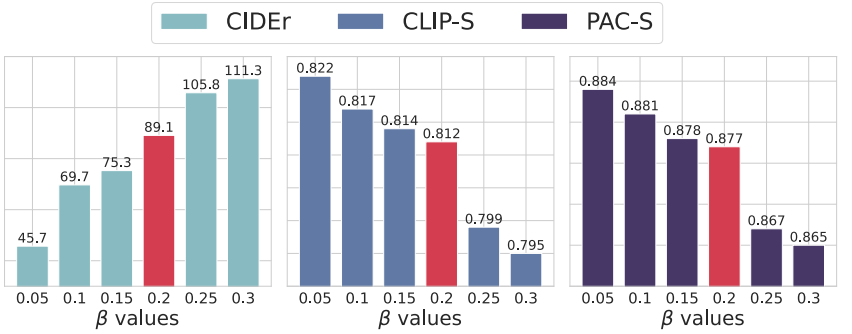


Figure B.1: Effect of β on CIDEr, CLIP-S, and PAC-S (ViT-L/14 backbone).

B.1.2 Quality-aware reward weighting

We weight per-sample rewards using quality distances (Eq. 4 in Chapter 3) to stabilize optimization. Compared to uniform weighting, quality-aware weighting consistently improves the overall trade-off between reference-based captioning metrics and reference-free retrieval/semantic scores. For completeness, the full quantitative ablation is reported in the accompanying publication.

B.1.3 Regularization strength β

The parameter β controls the deviation from the XE-pretrained captioner during preference optimization. Larger values constrain the model closer to the XE distribution, typically preserving higher CIDEr but limiting improvements on reference-free metrics. Smaller values encourage stronger reward optimization at the cost of reference-based performance. Across experiments, $\beta = 0.2$ provides a robust compromise and is used by default in this thesis.

B.1.4 Number of negative samples k

At each training step, DiCO samples $k{+}1$ captions and selects the k lowest-scoring candidates as negatives under the target reward. In practice, $k = 4$ offers a good balance between optimization strength and stability, and is used in all experiments unless stated otherwise. Extended results for different values of k are available in the publication.

B.2 Additional evaluation evidence

B.2.1 Fine-grained captioning

We additionally evaluate on FineCapEval, which emphasizes attributes, relations, and background details. Results are consistent with COCO: DiCO improves the trade-off between standard captioning metrics and CLIP-based metrics across backbones. We refer to the publication for the full table.

B.2.2 Out-of-domain robustness

To assess generalization, we consider nocaps (novel objects), VizWiz (images from blind users), TextCaps (text-rich scenes), and CC3M (web-scale pairs). DiCO improves robustness under distribution shift while maintaining competitive caption quality. Full per-dataset results are reported in the publication.

B.3 Reproducibility notes

B.3.1 Training setup

Cross-entropy pre-training uses gradient accumulation over two GPUs (1,024 samples per batch) with linear warm-up to a peak learning rate of $2.5 \cdot 10^{-4}$. Fine-tuning starts from the XE checkpoint with best validation CIDEr, using 2 GPUs and a global batch size of 16. Unless otherwise stated, we use $\beta = 0.2$ and $k = 4$.

B.3.2 Human evaluation (summary)

We conduct pairwise comparisons between captions generated by DiCO and competing models. Participants select the preferred caption according to (i) correctness and (ii) helpfulness. To reduce presentation bias, the order of the two candidates is randomized. The complete interface instructions and forms are available in the publication supplementary material.

B.3.3 LLM-based evaluation (summary)

We use GPT-3.5 Turbo as an automated judge under a fixed prompt and deterministic decoding whenever supported (temperature $= 0$). We randomize the order of candidate captions to mitigate position bias and discard malformed outputs (*e.g.*, invalid JSON) when computing aggregate statistics. The full prompt template is reported in the publication.

C

Supplementary material for Chapter 5

This appendix provides concise complementary material for Chapter 5. We focus on (i) the key methodological choices underlying the proposed reflective retrieval mechanism, and (ii) a small set of additional analyses that directly support the main claims of the chapter. Unless otherwise stated, extended quantitative results, large-scale sweeps, qualitative examples, competitor reproductions, and implementation artefacts (*e.g.*, prompts, UI elements, full templates) are reported in the associated publications and their supplementary material.

C.1 Key design choices

C.1.1 Why reflective tokens?

A central limitation of prompt-based multimodal RAG is that retrieval is treated as a fixed preprocessing step: retrieved text is appended to the prompt regardless of whether it is needed, and the model is left to implicitly filter noise during generation. This often leads to brittle behavior when retrieval is imperfect or when the query is purely perceptual.

To address this issue, the proposed model introduces explicit *reflective con-*

trol tokens that make retrieval a controllable part of multimodal reasoning. Concretely, the model predicts a <RET> token when external knowledge is required and <NORET> otherwise. When retrieval is performed, the model additionally predicts <REL> or <NOREL> tokens to assess the relevance of each retrieved passage before generating the final answer. This design makes the retrieval decision explicit, reduces blind context injection, and enables selective use of evidence.

C.1.2 Inference protocol (summary)

At inference time, the model is prompted with an image–question pair and is first asked to decide whether retrieval is needed. If <NORET> is produced, the model answers directly from its parametric and visual knowledge. If <RET> is produced, a retriever collects candidate passages from the external knowledge base; each passage is then scored by the model via <REL>/<NOREL> predictions, and the final answer is generated by conditioning on the subset of passages deemed relevant. This protocol operationalizes the chapter's core claim: effective multimodal RAG requires models to decide *when* to retrieve and *what* to trust, rather than merely increasing the context length.

C.1.3 Two-stage training strategy

Training is performed in two stages to stabilize the acquisition of reflective behavior.

Stage 1 (in-article supervision). The model learns to discriminate relevant versus non-relevant passages within the same document. This stage focuses on fine-grained relevance judgments, reducing spurious correlations and encouraging stable <REL>/<NOREL> predictions.

Stage 2 (cross-article supervision). The model is trained with passages that include hard negatives from the same document and soft negatives from other documents. This stage teaches the model to reject plausible but incorrect evidence and improves robustness when retrieval is imperfect.

Together, the curriculum enables reflective tokens to act as reliable intermediate decisions, rather than brittle heuristics learned from noisy retrieval.

	First Stage		Second Stage	
	# Samples	Passages	# Samples	Passages
E-VQA	43.6k	In-Article	2.9M	In- and Cross-Article
InfoSeek	41.0k	In-Article	2.5M	In- and Cross-Article
LLaVA-Instruct	665.3k	-	665.3k	-

Table C.1: Training data mixture employed during the two phases of the proposed training strategy.

C.2 Training data mixture

Table C.1 summarizes the data mixture used during the two training stages. At each stage, samples from different sources are balanced to ensure stable optimization and to preserve general-purpose instruction-following behavior. The first stage focuses on in-article supervision for relevance prediction, while the second stage introduces large-scale cross-article supervision to encourage robust selective integration of retrieved evidence.

C.3 Additional analyses supporting the chapter claims

C.3.1 Reliability of reflective token predictions

We assess the accuracy of the reflective tokens on controlled validation subsets to verify that they correspond to stable and interpretable decisions. Across datasets, reflective token prediction achieves consistently high accuracy (often above 85%), indicating that the model reliably distinguishes (i) queries that require external knowledge from those that do not, and (ii) relevant evidence from irrelevant or misleading retrieved passages. This supports the interpretation of reflective tokens as meaningful intermediate reasoning steps that mediate retrieval-augmented generation.

C.3.2 Implicit passage re-ranking via token probabilities

Beyond discrete `<REL>`/`<NOREL>` predictions, the model's token probabilities can be used to score retrieved passages. In particular, ranking passages by the

log-probability difference between <REL> and <NOREL> provides an implicit reranking mechanism. This analysis shows that reflective tokens do not merely gate retrieval, but also encode a graded notion of evidence utility that can be exploited to prioritize informative passages without introducing a separate reranking model.

C.3.3 Limitations

Despite its effectiveness, the approach remains sensitive to evaluation regimes that require strict formatting or exact-match answers. In such settings, semantically correct answers expressed in alternative surface forms may be penalized. Moreover, extremely fine-grained numerical questions may still require higher-precision evidence selection or post-hoc normalization. These limitations suggest future directions such as confidence-aware decoding, answer normalization, and tighter coupling between retrieval and generation.

C.4 Reproducibility notes

C.4.1 Model and optimization (summary)

The model is based on LLaVA-v1.5 with LLaMA-3.1-8B as the language backbone. Training hyperparameters, optimizer settings, and implementation details follow the associated publication. Here we report only design choices that directly affect reflective retrieval behavior (token vocabulary extension and the two-stage curriculum).

C.4.2 Evaluation protocol (summary)

We follow the official evaluation protocols for Encyclopedic-VQA and InfoSeek. When LLM-based evaluation is used in auxiliary analyses, prompts are fixed and decoding is deterministic whenever supported. Extended evaluation details, prompts, and templates are available in the publication supplementary material.

D

Supplementary material for Chapter 6

This appendix provides concise complementary material for Chapter 6. We focus on (i) the key design choices behind COGT (Causally-Ordered Generative Training, COGT) and the corresponding ablations, and (ii) the essential details required to reproduce the experimental protocol. Unless otherwise stated, extended quantitative results, full ablation grids, qualitative examples, and additional benchmark tables are reported in the associated publication and its supplementary material.

D.1 Key design choices and ablations

D.1.1 Causal factorization and sparsity

We model linguistic dependencies as a causal graph over caption tokens, and train the decoder to maximize the *disentangled* (causal) factorization:

$$P(W_1, \ldots, W_n) = \prod_{j=1}^{n} P(W_j \mid \mathbf{PA}(W_j)), \tag{D.1}$$

where $\mathbf{PA}(W_j)$ denotes the parents of W_j in the induced DAG. In our setting, parents are defined by dependency-tree ancestry (plus syntactic type and visual features, as in Chapter 6). Compared to standard left-to-right autoregression, this induces *sparser* conditionals: under mild assumptions (*e.g.*, approximately balanced trees), the number of conditioning tokens per node scales as $O(\log n)$, instead of $O(n)$ for sequential AR. This sparsity is a key mechanism for reducing spurious correlations induced by linear word order.

D.1.2 Dependency parser choice

The dependency parser determines the causal structure used at both training and inference. We treat parsers as fixed black boxes (no fine-tuning) and observe a consistent trend that more accurate parsers tend to yield stronger compositional performance. In the main experiments we use *Deep Biaffine + RoBERTa*, which offers a strong trade-off between accuracy and usability among the considered options. Full parser ablations (including CRF-based alternatives) are reported in the publication.

D.1.3 Mask-specific syntactic tokens and leakage prevention

The decoder conditions on syntactic roles via dedicated masked tokens $\texttt{MSK}_t$ (one per dependency relation type t). Operationally, each word is represented by (i) a *masked* token encoding its syntactic type and (ii) a *visible* token encoding its lexical identity; predictions are produced from masked states only, so that a token cannot directly attend to its own visible form. Replacing relation-specific masks with a single generic mask consistently reduces performance, confirming that explicit syntactic typing stabilizes learning and improves compositional generalization. For completeness, we refer to the publication for the full ablation grid.

D.1.4 Multi-layer visual features (last + penultimate)

We extract visual tokens from both the last and the penultimate layer of the frozen visual encoder (when applicable). This choice is especially important for fine-grained benchmarks (*e.g.*, color/texture swaps and small-object cues), where penultimate features retain more spatial detail. Using only the last layer typically degrades accuracy; extended results per benchmark are reported in the publication.

D.1.5 Decoder depth and larger-data variants

COCO-only models use a three-block decoder, while the larger-data variants (COGT+ in the chapter) use four blocks. Across benchmarks, scaling training data (COCO $\rightarrow$ COCO+CC3M+VG with overlap filtering) yields consistent improvements, indicating that the causal training objective benefits from more data even when captions are noisy. We leave dataset-quality improvements via LLM recaptioning as future work.

D.2 Benchmark notes

D.2.1 FG-OVD benchmark construction

FG-OVD is used as an image-to-text retrieval benchmark built from object-centric crops. Each annotated object is cropped using its bounding box, resized to 224×224, and paired with its object caption as the positive text. Negatives follow the original FG-OVD protocol: *Trivial* negatives are captions sampled from unrelated objects, while *Easy/Medium/Hard* negatives are created by replacing respectively 3/2/1 attribute mentions in the true caption. Harder settings therefore preserve more of the original caption, requiring finer-grained compositional discrimination. Task-specific results are reported in the publication.

D.2.2 ARO protocol

Following common practice, we evaluate on Visual Genome *Relation* and *Attribution*. We do not include the COCO-Order and Flickr-Order splits, as they can be solved with strong language priors due to frequent grammatical artifacts in negatives.

D.2.3 Additional benchmarks

We additionally report results on Winoground and MMVP in the extended experiments. Because MMVP is small (135 samples), we treat it as supportive evidence rather than a high-confidence estimate. Winoground is included to stress out-of-focus and low-resolution cues that are challenging for many CLIP-like encoders. Full tables are reported in the publication.

D.3 Reproducibility notes

D.3.1 Architectures

Backbones. For COGT-CLIP we use ViT-B/32 CLIP; for COGT-XVLM we use the XVLM Swin-based visual encoder; and for COGT +InstructBLIP we use the InstructBLIP Q-Former output as visual tokens. All visual encoders are frozen.

Visual mapping network. Visual tokens are projected to the decoder dimension through a shared mapping network $\mathcal{M}$ (a linear layer preceded and followed by LayerNorm), applied to tokens from the last and penultimate encoder layers (where applicable).

Decoder. The decoder alternates (i) dependency-guided attention over text tokens and (ii) cross-attention over visual tokens. Unless stated otherwise: 3 blocks for COCO-only, 4 blocks for larger-data variants; 8 attention heads and 512 hidden size (COCO-only), and 12 heads with 768 hidden size (larger-data variants); dropout $= 0.1$ on residuals, attention weights, and embeddings.

D.3.2 Tokenization and tree alignment

Dependency parsers operate at the word level, while the decoder uses subword tokenization. When a word is split into multiple sub-tokens, we modify the dependency tree by replacing the word node with a chain of sub-token nodes. We introduce a dedicated relation (*e.g.*, `comp`) between consecutive sub-tokens so that: the first sub-token inherits the original parents, and subsequent sub-tokens additionally depend on the previous sub-token. This avoids information leakage while preserving the intended causal ancestry.

D.3.3 Training setup

We train only the decoder and $\mathcal{M}$ in mixed precision (FP16). Unless otherwise stated: batch size $= 128$ on a single RTX A5000 (24GB), 10 epochs, Adam with initial learning rate $5 \cdot 10^{-4}$, cosine annealing schedule with 50 warm-up steps. We select checkpoints using the ARO validation protocol adopted by prior work. When supported, we use fixed random seeds for reproducibility. Dependency trees are extracted offline once per caption; the preprocessing cost is modest relative to training.

D.3.4 Inference protocol and efficiency

Compositional benchmarks are evaluated as image-to-text retrieval: visual features for each image are computed once; candidate captions are scored by the causal log-likelihood using a semi-parallel level-order traversal of the dependency tree (tokens at the same depth are predicted in parallel). At test time, dependency parsing adds a small constant overhead per caption; overall inference time remains comparable to CLIP-style scoring. For representative wall-clock and memory figures, see the publication.

D.4 Syntactic relation inventory

Our relation set follows the standard dependency relation inventory used by the parser. For brevity, we omit the full list here and refer to the publication and the

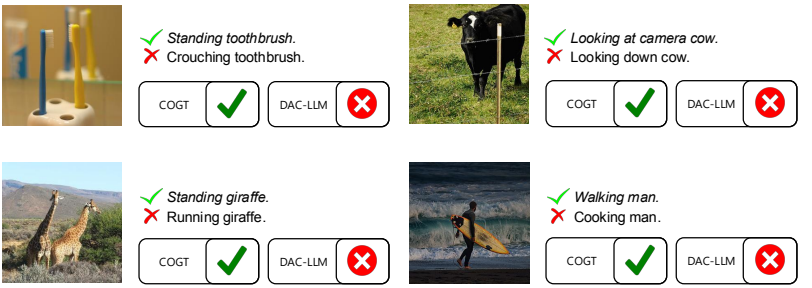


Figure D.1: Qualitative results on sample images of VL-CheckList.

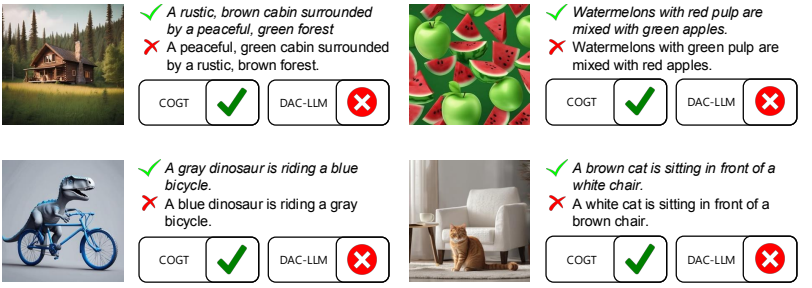


Figure D.2: Qualitative results on sample images of ColorSwap.

parser documentation; the thesis chapter and this appendix only require knowing that each relation type t is mapped to a dedicated masked token $\texttt{MSK}_t$.

D.4.1 Qualitative results

In Figures D.1 and D.2, we present qualitative results comparing COGT-CLIP+ with the second-best approach reported in Table 6.3 (DAC-LLM). These figures also illustrate the different tasks addressed by the benchmarks.