# FG-TRACER: Tracing Information Flow in Multimodal Large Language Models in Free-Form Generation

Alessia Saporita[1,2], Vittorio Pipoli[1,3], Federico Bolelli[1]✉,
Lorenzo Baraldi[1], Andrea Acquaviva[2], and Elisa Ficarra[1]

[1]University of Modena and Reggio Emilia, Italy
[2]University of Bologna, Italy [3]University of Pisa, Italy

{name.surname}@unimore.it

## Abstract

*Multimodal Large Language Models (MLLMs) have achieved impressive performance across a variety of vision–language tasks. However, their internal working mechanisms remain largely underexplored. In this work, we introduce* FG-TRACER*, a framework designed to analyze the information flow between visual and textual modalities in MLLMs in free-form generation. Notably, our numerically stabilized computational method enables the first systematic analysis of multimodal information flow in underexplored domains such as image captioning and chain-of-thought (CoT) reasoning. We apply* FG-TRACER *to three state-of-the-art MLLMs—LLaVA 1.5, LLaMA 3.2-Vision, and Qwen 2.5-VL—across three vision–language benchmarks—TextVQA, COCO 2014, and ChartQA—and we conduct a word-level analysis of multimodal integration. Our findings uncover distinct patterns of multimodal fusion across models and tasks, demonstrating that fusion dynamics are both model- and task-dependent. Overall,* FG-TRACER *offers a robust methodology for probing the internal mechanisms of MLLMs in free-form settings, providing new insights into their multimodal reasoning strategies. Our source code is publicly available at* https://github.com/AImageLab-zip/FG-TRACER

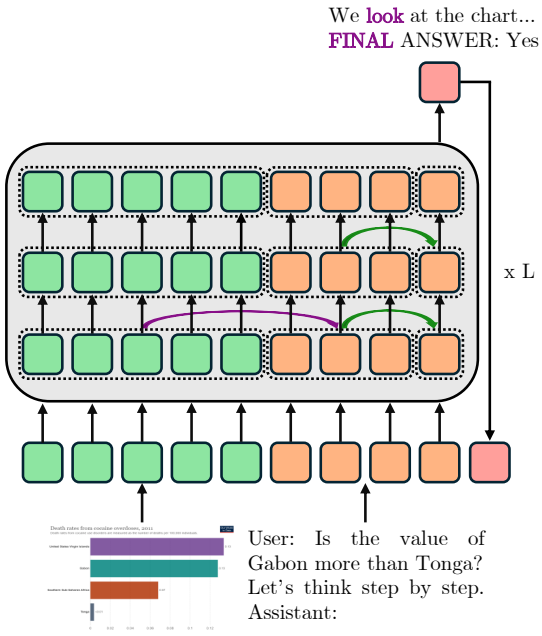


Figure 1. Illustration of the inner mechanism of MLLMs in chain-of-thought reasoning, with $L$ indicating the answer length. The model first interprets the prompt and fuses visual and textual information, then propagates it to the final position for answer prediction. Words expressing temporal, spatial or procedural relations—such as *look* and *final*—exhibit strong visual grounding.

## 1. Introduction

Multimodal Large Language Models (MLLMs) [2, 14, 27, 44, 45] represent a significant advancement in artificial intelligence, as they extend the capabilities of traditional language models [3, 10, 33, 34] to support reasoning over non-textual modalities [8, 15, 37, 46]. Typically, they comprise a vision encoder [12, 13, 29, 39] for extracting visual features, a fusion module that integrates visual and textual information into a shared representation space, and a language model backbone. This multimodal fusion enables MLLMs to perform more complex reasoning and exhibit capabilities that are essential for a wide range of real-world vision-language tasks, including text understanding [4, 21], image captioning [7, 35], and chart understanding [1]. Despite their impressive performance, the internal mechanisms of MLLMs remain insufficiently understood, particularly in complex free-form generation settings.

Recent studies have investigated the inner mechanisms

✉ Corresponding author: federico.bolelli@unimore.it

of MLLMs by analyzing information storage in model parameters [6], knowledge encoding [22], object-level visual grounding [24, 25], visual signal decay across layers [38], token-level redundancy [42], and the dynamics of cross-modal interactions [30, 43]. Basu *et al.* [6] reveal that specific visual tokens are responsible for transferring information from the image to causal blocks within the architecture. Zhang *et al.* [43] investigate the internal mechanisms of multimodal integration in the LLaVA model series for visual question answering (VQA) [17] tasks involving single-token answers. However, the dynamics of multimodal information flow in MLLMs during complex, free-form tasks remain largely underexplored.

In this paper, we address this gap by introducing FG-TRACER, a framework for analyzing visual–textual interactions in MLLMs in free-form generation. FG-TRACER employs an attention masking mechanism and a novel numerically stabilized formulation that enables reliable quantification of the information flow between visual and textual modalities in tasks requiring long-answer generation, such as chain-of-thought (CoT) reasoning and image captioning. A visualization of the internal mechanism underlying CoT reasoning is presented in Fig. 1.

Leveraging the capabilities of FG-TRACER, we examine three state-of-the-art MLLMs—LLaVA 1.5 [16], LLaMA 3.2-Vision [14], and Qwen 2.5-VL [5]—which employ different strategies for multimodal integration. LLaMA 3.2-Vision exploits a cross-attention mechanism to integrate visual features into the language model's internal representations, while LLaVA 1.5 and Qwen 2.5-VL concatenate visual and textual embeddings and rely on self-attention to achieve fusion. These distinctions make them well-suited for studying how different architectural designs influence visual–textual interactions in multimodal reasoning. Moreover, we study the information flow on three distinct visual-language tasks: text understanding on TextVQA [4] dataset, image captioning on COCO 2014 [35], and chart understanding on ChartQA [1]. TextVQA is a benchmark for text-based VQA, which requires Optical Character Recognition (OCR) capabilities to reason over textual context embedded in the image. COCO 2014 is a widely used dataset to evaluate models' ability to generate descriptive, free-form captions. ChartQA is a benchmark designed to assess a model's ability to answer questions about chart images, where the questions typically require multi-step, chain-of-thought reasoning to infer information from visual data. Our analysis yields several key findings: (*i*) multimodal fusion predominantly occurs in early-to-mid layers, but how visual and textual information are fused is model- and task-dependent; (*ii*) in OCR-based VQA tasks, visual information directly influences response generation, as the model requires fine-grained visual features during decoding to interpret structured textual elements embedded within the image; (*iii*) in image captioning, MLLMs directly rely on the visual input after multimodal integration, reflecting the need for persistent visual access for generating accurate descriptions; and (*iv*) in CoT reasoning, visual information contributes throughout the step-by-step reasoning process, while the final answer is derived primarily from the generated linguistic chain, demonstrating that MLLMs utilize intermediate reasoning steps to encode the necessary information for answer generation. Furthermore, our word-level analysis of multimodal integration reveals that terms associated with spatial, temporal, compositional, or procedural semantics demonstrate strong visual grounding, whereas structural words—such as pronouns, articles, and prepositions—do not rely on visual information and are instead driven by linguistic priors. To the best of our knowledge, FG-TRACER is the first methodology capable of tracing and quantifying the information flow between visual and textual modalities in long, free-form generation tasks, such as image captioning and CoT reasoning, thereby offering novel insights into the inner mechanisms of MLLMs.

**Contributions.** The contributions of the paper can be summarized as follows:

- We introduce FG-TRACER, a novel framework that combines attention masking with a new reliable quantification of the information flow between visual and textual modalities in MLLMs in free-form generation;
- Our approach enables analysis of the cross-modal information flow in MLLMs in complex tasks—such as CoT reasoning and image captioning—that were previously inaccessible using existing methods;
- We apply the methodology to three state-of-the-art MLLMs—LLaVA 1.5, LLaMA 3.2-Vision, and Qwen 2.5-VL—spanning three distinct vision–language tasks—text-based VQA, image captioning, and chart VQA, which involves CoT reasoning—uncovering that multimodal fusion dynamics are both model- and task-dependent;
- We perform a word-level analysis in image captioning and CoT reasoning, showing that terms expressing spatial, temporal, contextual, or procedural relations exhibit strong visual grounding.

## 2. Related Work

**Multimodal Large Language Models (MLLMs).** MLLMs [2, 14, 44, 45] represent a major advancement in artificial intelligence by integrating the reasoning capabilities of Large Language Models (LLMs) [3, 10, 33, 34] with the ability to process and generate content across multiple modalities [8, 15, 37, 46]. These models typically comprise modality-specific encoders—such as vision encoders [12, 13, 29]—paired with a large language

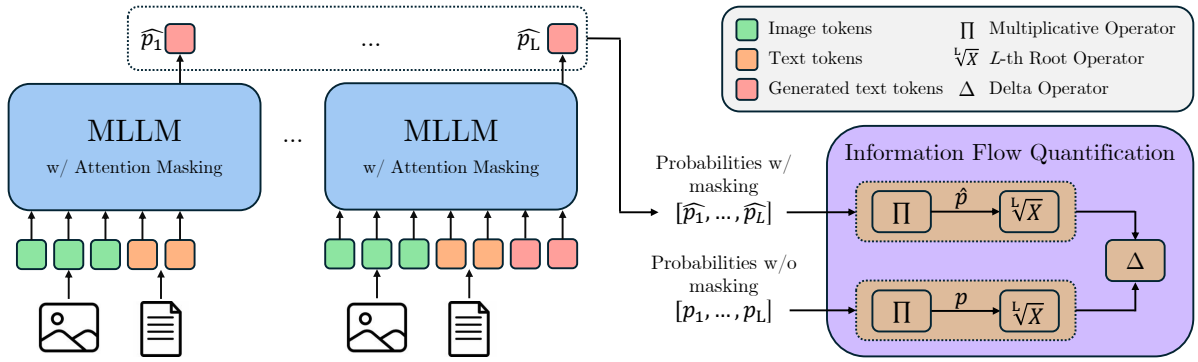


Figure 2. Overview of the FG-TRACER framework. For each output token, FG-TRACER computes the probabilities of the correct answer tokens with and without attention masking applied—$\hat{p_1}, \ldots, \hat{p_L}$ and $p_1, \ldots, p_L$, respectively—between selected token groups across a window of layers. The token-level probabilities $[p_1, \ldots, p_L]$ and $[\hat{p}_1, \ldots, \hat{p}_L]$ are then multiplied, and normalized via the $L$-th root, where $L$ is the answer token length, to avoid length-induced distortions. The information flow between the selected token groups is then quantified as the relative change in output probabilities ($\Delta$) between the normalized probabilities computed with and without the attention masking.

model backbone, enabling cross-modal understanding and generation. Prominent examples include Flamingo [2], BLIP-2 [19], MiniGPT-4 [45], and the more recent LLaVA 1.5 [20], LLaMA 3.2-Vision [14], and Qwen 2.5-VL [5]. However, the internal working mechanisms of MLLMs remain poorly understood.

**Analyzing Internal Mechanisms of MLLMs.** While significant progress has been made in interpreting the internal behavior of large language models [11, 23, 26, 41], the multimodal domain remains underexplored. Recent studies have begun to investigate the internal states of MLLMs by examining information storage [6], knowledge encoding [22], object-level visual grounding [24, 25, 31], visual signal degradation across layers [38], token-level redundancy [42], and cross-modal interaction dynamics [30, 43]. Basu *et al.* [6] analyze how visual information is stored and transferred within MLLM architectures. Yin *et al.* [38] investigates how visual signals shift across layers to enable more efficient inference. Palit *et al.* [25] demonstrates that representations in later layers of the model have a causal impact on the generated VQA token outputs, offering a visualization-based tool for tracing how visual input influences language generation in a multimodal model. More recently, Tao *et al.* [32] characterizes how global and local semantic representations are encoded across layers. Jiahao *et al.* [18] reveals how individual neurons contribute to either visual, textual, or cross-modal understanding. Zhang *et al.* [43] examine the dynamics of multimodal integration in the LLaVA model series on VQA tasks involving single-token answers [17]. Despite recent advances, the dynamics of multimodal information flow in MLLMs in complex, free-form tasks—such as image captioning and CoT reasoning—remain underexplored. Hence, this work introduces a framework capable to analyze multimodal information flow in these tasks, overcoming the limitations of previous approaches.

## 3. Method

In this work, we present FG-TRACER, a methodology that combines attention masking with a novel formulation for quantifying cross-modal information flow between visual and textual modalities in MLLMs in free-form generation. The approach iteratively masks attention between specific target and source token groups across selected layers, and computes the probability of generating the correct answer token. To account for multi-token outputs, we multiply the individual token probabilities and normalize them by the sequence length to mitigate distortions from long responses. The information flow is then quantified as the change in output probability, where significant deviations indicate that the intervention disrupted a critical information pathway, thus revealing the contribution of each source token group.

### 3.1. Preliminaries

**MLLMs.** MLLMs are designed to process and reason over visual and textual inputs within a unified architecture. A visual encoder extracts high-level features from the input image, which are then projected into the language model's embedding space, allowing the LLM backbone to integrate and reason over visual and textual information. Specifically, the LLM backbone is a stack of $N_L$ transformer layers, each comprising a masked multi-head attention (MHA) [36] module and a feed-forward network (FFN). The hidden representation at layer $l$, denoted as $H^l = [h_i^l]_{i=1}^N \in \mathbb{R}^{N \times d}$, where $N$ is the total number of input tokens, is defined as follows:

$$H^l = \mathrm{FFN}(Z^l) + Z^l, \quad Z^l = \mathrm{MHA}(H^{l-1}) + H^{l-1}, \quad (1)$$

where $H^{l-1}$ represents the hidden representation of the previous layer and $Z^l$ denotes the vector obtained by adding the MHA output at layer $l$ to the residual input $H^{l-1}$. The MHA module projects $H^{l-1}$ into query ($Q^l$), key ($K^l$), and

value ($V^l$) matrices using learned weights $W_Q^l, W_K^l, W_V^l \in \mathbb{R}^{d\times d}$, and partitions them across $N_H$ heads, $\{Q^{l,j}\}_{j=1}^{N_H}$, $\{K^{l,j}\}_{j=1}^{N_H}$, and $\{V^{l,j}\}_{j=1}^{N_H}$, each of dimension $\mathbb{R}^{N\times\frac{d}{N_H}}$. The MHA output is computed as:

$$\mathrm{MHA}(Q^l, K^l, V^l) = [head_1^l, \ldots, head_{N_H}^l]W_O^l, \quad (2)$$

$$head_j^l = \mathrm{softmax}\left(\frac{Q^{l,j}(K^{l,j})^T}{\sqrt{d/N_H}} + M^{l,j}\right)V^{l,j}, \quad (3)$$

where $[\cdot]$ denotes concatenation, $W_O^l \in \mathbb{R}^{d\times d}$ is the output projection matrix, and $M^{l,j}$ is the attention mask, indexed by layer $l$ and head $j$, controlling token-to-token visibility. The FFN component computes the layer output via:

$$\mathrm{FFN}(Z^l) = \sigma\big(Z^l W_B^l\big)W_U^l, \quad (4)$$

where $W_U^l, W_B^l \in \mathbb{R}^{d\times d_{ff}}$, $d_{ff}$ is the inner hidden dimension, and $\sigma$ is a nonlinear activation function.

MLLMs typically employ one of two main strategies for integrating visual and textual information within the LLM backbone: (*i*) self-attention-based fusion or (*ii*) cross-attention-based fusion. Initially, the input image $x$ is divided into fixed-size patches and passed through a vision encoder to produce a sequence of visual embeddings $V = [v_i]_{i=1}^{N_V}$, where $N_V$ is the number of visual tokens and each $v_i \in \mathbb{R}^d$ is a $d$-dimensional feature vector corresponding to a specific image patch. Simultaneously, a text sequence $t$, composed of $N_T$ tokens, is embedded using a learned word embedding table, yielding $T = [t_i]_{i=1}^{N_T}$, $t_i \in \mathbb{R}^d$. In the self-attention paradigm, visual and textual tokens are concatenated into a sequence $H^0 = [V, T] \in \mathbb{R}^{(N_v+N_T)\times d}$ and jointly processed by the $N_L$ transformer layers. In this case, the attention mask is strictly upper triangular, preventing each token from attending to future positions. In the cross-attention paradigm, the MLLM receives as input the textual sequence $T$, and multimodal fusion is performed using additional multi-head cross-attention layers, in which textual hidden states $H^l$ serve as queries, while visual embeddings $V$ act as keys and values. In the cross-attention layers, a fully permissive attention mask is employed, enabling all textual tokens to attend to all visual tokens w/o restriction.

**Autoregressive Decoding.** The hidden representation $h_N^{N_L}$ corresponding to the final position $N$ in the input sequence at the top transformer layer $N_L$ is projected through an output embedding matrix $E \in \mathbb{R}^{|\mathcal{V}|\times d}$, where $|\mathcal{V}|$ denotes the vocabulary. The resulting probability distribution over vocabulary tokens is given by:

$$P_N = \mathrm{softmax}\left(h_N^L E^T\right), \quad (5)$$

where the token with the highest probability in $P_N$ is selected as the predicted output. For multi-token responses ($e.g.$, phrases or full sentences), decoding proceeds in an autoregressive manner: the predicted token is appended to the input sequence, and the model reprocesses the extended sequence to generate the next token. This iterative process continues until an end-of-sequence token is produced or a maximum length is reached.

### 3.2. FG-TRACER

In this work, we introduce FG-TRACER, a framework for quantifying multimodal information flow between visual and textual modalities in MLLMs in free-form generation. An overview of the method is presented in Fig. 2. FG-TRACER comprises two components: an attention masking mechanism that selectively blocks attention between specific token groups, and a novel information flow formulation designed to enable reliable analysis in long, free-form answers by addressing the challenges posed by sequence length in probability computations. Unlike prior work [43], which is restricted to single-token answers and does not generalize to free-form outputs, FG-TRACER supports tasks such as image captioning and chain-of-thought (CoT) reasoning by accurately quantifying information flow across multi-token sequences. We analyze two types of information flow: (*i*) from each modality (text or image) to the last input token, responsible for the prediction, capturing direct modality influence; and (*ii*) from visual to textual tokens (excluding the last input token), to examine the multimodal integration process.

**Attention Masking Mechanism.** The multi-head attention is the only module that enables communication across different groups of tokens, thus integrating contextual information. To investigate the aforementioned information flows, we intervene in each layer $l$ by selectively preventing the last token from attending to preceding textual tokens (text-to-last flow), visual tokens (image-to-last flow), or itself (last-to-last flow) and by blocking the attention from textual tokens (excluding the last token) to the visual tokens (image-to-text flow). Each intervention is applied within a local window of $k = 9$ layers centered around layer $l$. Results with other window sizes are reported in the supplementary material. To block the information flow from a source token group ($i.e.$, text, image, or last token) to a target group ($i.e.$, text, or last token), we set the corresponding entries in the self- or cross-attention mask to $-\infty$.

**Information Flow Quantification.** We propose a novel formulation to quantify the information flow between source and target token groups. Given an image-question pair, the model generates the $i$-th answer token with probability $p_i$. After masking the attention between the source and target groups, the model generates the same token with probability $\hat{p}_i$. Since free-form answers typically consist of multiple tokens, the overall sequence probability—denoted as $p$ without masking and $\hat{p}$ with masking—is computed as

the product of individual token probabilities:

$$p = \prod p_i = p_1 \cdot p_2 \cdot \ldots \cdot p_L, \tag{6}$$

$$\hat{p} = \prod \hat{p_i} = \hat{p_1} \cdot \hat{p_2} \cdot \ldots \cdot \hat{p_L}, \tag{7}$$

where $L$ is the number of tokens in the generated response and each $p_i, \hat{p_i} \in [0, 1]$. Then, we quantify the information flow between the source and target token groups as the relative change in output probability. Specifically, we first normalize the output probabilities $p$ and $\hat{p}$ by computing the $L$-th root of the probabilities, and then we calculate the relative change $\Delta$ as:

$$\Delta = \left( \frac{\hat{p}^{\frac{1}{L}} - p^{\frac{1}{L}}}{p^{\frac{1}{L}}} \right) \cdot 100. \tag{8}$$

By accounting for sequence length, this formulation enables the detection of patterns in long answers. Normalizing probabilities via $L$-th root is crucial for long-form answers, as otherwise small changes in individual token probabilities would result in exponentially large changes in the overall answer probability when the number of tokens $L$ increases.

Consider a scenario in which the probability of each token undergoes a small relative change $\delta$, such that $\hat{p}_i = (1 - \delta) \cdot p_i$ for $i = 1, \ldots, L$. Under this perturbation, the probability of the entire sequence becomes $\hat{p} = (1-\delta)^L \cdot p$, and the relative change in sequence-level probability is given by $\Delta = \left(\frac{\hat{p}-p}{p}\right) \cdot 100 = ((1 - \delta)^L - 1) \cdot 100$. This relative change decreases exponentially with increasing $L$ and asymptotically approaches $-100\%$. For instance, in the ChartQA dataset, the average answer length is 125 tokens. If each token's probability decreases by 5% (*i.e.*, $\delta = 0.05$), the resulting relative change in sequence-level probability is $\left((1 - 0.05)^{125} - 1\right) \cdot 100 = -99.84\%$. However, if the probabilities are normalized—specifically by taking the $L$-th root—the relative change in probability becomes $((1 - 0.05) - 1) \cdot 100 = -5\%$. This example shows that, in long answers, without normalization, even minor perturbations at the token-level probabilities can result in exponentially large and misleading relative changes in sequence-level probability. To mitigate this effect and ensure comparability across different sequence lengths, we normalize by the number of tokens $L$.

# 4. Experimental Results

**Multimodal Backbone.** In our analysis, we study three state-of-the-art MLLM backbones: LLaVA 1.5–13B [16], LLaMA 3.2-11B-Vision [14], and Qwen 2.5-VL-7B [5]. LLaVA 1.5 builds upon the Vicuna-13B [9] language model and employs a pre-trained CLIP [29] visual encoder to extract visual embeddings, and a linear projection layer to align image and textual features. It follows the self-attention paradigm to fuse modalities. Qwen 2.5-VL augments the Qwen 2.5 [28] language backbone with a dynamic-resolution Vision Transformer [12] enhanced by windowed attention, and leverages the self-attention paradigm to integrate visual and textual modalities. In contrast, LLaMA 3.2–11B-Vision, based on the LLaMA 3.2 language backbone, employs the cross-attention paradigm to fuse modalities, inserting cross-attention layers at fixed intervals. These distinctions enable us to investigate how different architectural design choices influence cross-modal information flow within MLLMs.

**Datasets.** We evaluate our framework on three distinct vision–language tasks: text understanding on TextVQA [4], image captioning on COCO 2014 [35], and chart understanding on ChartQA [1]. TextVQA is an open-ended VQA dataset that requires reasoning over text embedded within natural images. This task poses significant challenges due to its reliance on accurate OCR capabilities. COCO 2014 is a widely used benchmark for image captioning, consisting of high-quality natural images annotated with free-form descriptive captions. ChartQA is a representative benchmark for chart understanding that involves reasoning over charts. Due to the inherent complexity of interpreting charts and performing quantitative reasoning, we employ Chain-of-Thought (CoT) prompting to support multi-step inference. To ensure a reliable analysis, we restrict our study to correctly answered samples, thereby avoiding noise introduced by disrupted or irrelevant attention pathways associated with incorrect predictions. Additional experiments on incorrect samples are reported in the supplementary materials. For the COCO 2014 dataset, we randomly sample 2,000 image–question pairs that were accurately answered. For TextVQA and ChartQA, we include all correctly answered instances—yielding approximately 2,000 samples for TextVQA in all MLLMs, 400 samples for ChartQA in LLaVA 1.5, and 2,000 samples in LLaMA 3.2-Vision and Qwen 2.5-VL. The data is sourced from the validation sets of TextVQA and COCO 2014, and the test set of ChartQA.

## 4.1. Results

In our analysis, we examine whether each modality directly contributes to the prediction and how the model integrates visual and textual representations. To this end, we mask attention from a target token group to a source group at each layer $l$, within a sliding window of $k = 9$ layers, and we quantify the information flow as the relative change in the answer probability with the formulation described in Eq. (8). We apply this methodology to LLaVA 1.5, LLaMA 3.2-Vision, and Qwen 2.5-VL on the TextVQA, COCO 2014, and ChartQA datasets. The results are presented in Fig. 3, which illustrates the patterns of the different information flows—namely last-to-last, image-to-last, text-to-last, and image-to-text—across the datasets. The

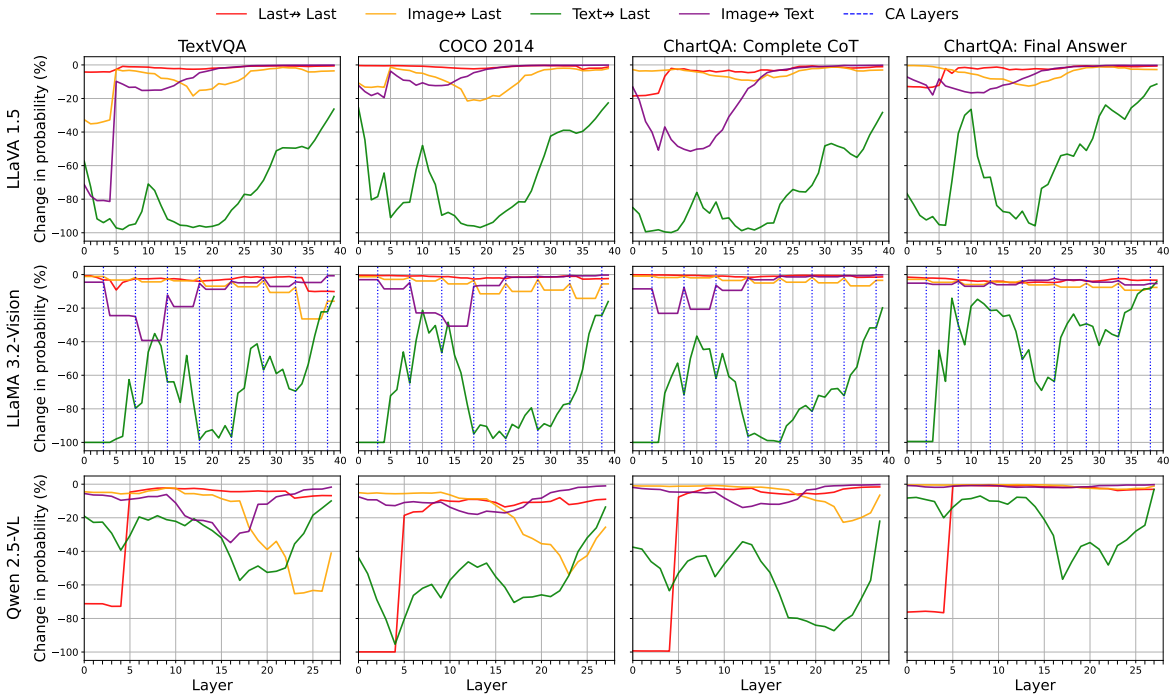


Figure 3. Relative changes in answer probability for LLaVA 1.5, LLaMA 3.2-Vision, and Qwen 2.5-VL across TextVQA, COCO 2014, and ChartQA datasets. ChartQA is analyzed under two settings: complete chain-of-thought responses—ChartQA: Complete CoT—and final answer generation given the reasoning chain as input—ChartQA: Final Answer. Dotted blue lines indicate cross-attention layers.

x-axis denotes the layer index, while the y-axis represents the resulting relative change in the answer probability. Values near zero suggest minimal information flow, whereas larger negative values indicate a stronger dependence on the masked interaction. Additionally, we conduct a word-level analysis in long, free-form generation tasks—specifically, image captioning and CoT reasoning—to identify which words in the generated responses exhibit the highest dependency on the visual input and which are instead driven by linguistic context.

**General Information Flow Patterns.** As shown in Fig. 3, the substantial image-to-text information flow is present in the early to mid layers, suggesting that multimodal fusion occurs early in the MLLMs. Nevertheless, distinct patterns of multimodal fusion emerge across models and datasets, indicating that visual–linguistic integration is both task- and model-dependent. In LLaVA-1.5, fusion begins in the earliest layers and unfolds in two stages—consistent with the findings of Zhang et al. [43] on the GQA dataset [17]. In contrast, LLaMA-3.2-Vision exhibits a single-stage fusion that begins in the first cross-attention layers, whereas in Qwen-2.5-VL fusion is delayed until the mid layers and likewise occurs in a single stage. In addition, we observe that in LLaMA 3.2-Vision and LLaVA 1.5, the last-to-last information flow is negligible across all datasets, *i.e.*, the last token does not integrate information from its own hidden representation during answer prediction. Instead, in Qwen 2.5-VL the flow is strong in the earliest layers, suggesting that the final token primarily integrates information from itself in the first layers. Moreover, Qwen-2.5-VL exhibits stronger image-to-last information flow in the final layers across all datasets, indicating a greater reliance on direct visual information for generating the answer. Furthermore, all MLLMs exhibit a two-stage text-to-last information flow. The first stage emerges in the early layers, where a strong flow indicates that they are initially interpreting the semantic content of the prompt. The second stage arises after multimodal integration is largely complete, likely corresponding to a task-specific reasoning phase during which the grounded multimodal representation is utilized to infer the answer. In the final layers, this information flow decreases, marking the concluding phase of answer generation. This reasoning process is consistent across all three datasets.

**Results on Text Understanding.** With respect to the text understanding task, we observe in Fig. 3 a stronger image-to-text information flow compared to the other datasets, indicating that the models rely more heavily on the visual input. This behavior is likely caused by the nature of the task, which requires the extraction of fine-grained visual features to accurately recognize and understand text embedded within the image. Additionally, we observe a relevant image-to-last information flow after multimodal integration, suggesting that MLLMs directly draw information

from the image to answer the question, likely due to the task's dependence on OCR-specific content.

**Results on Image Captioning.** In the context of image captioning, as shown in Fig. 3, we observe a substantial image-to-last information flow following multimodal integration in all models. This result indicates that MLLMs require direct access to visual features during decoding to generate coherent and semantically accurate captions. While in LLaMA 3.2-Vision and in Qwen 2.5-VL, the image-to-last information flow remains consistent throughout the answer generation phase, in LLaVA 1.5, it decreases in the last layers, suggesting that the model has already integrated all necessary information.

**Results on CoT Reasoning.** We analyze multimodal information flow during CoT reasoning on the ChartQA dataset. CoT prompting enables MLLMs to decompose complex visual questions into a sequence of reasoning steps, thus facilitating the inference of the final answer. Examples of complete CoT responses are reported in the supplementary material. We perform the analysis under two settings: (*i*) by examining the information flow across the entire chain-of-thought response, and (*ii*) by assessing the information flow in the concise final answer, while providing the whole reasoning chain as input to the MLLM. The corresponding results are shown in the third and fourth columns of Fig. 3, respectively. In the first scenario, we observe a strong text-to-last information flow across all MLLMs, as the models heavily rely on the textual modality to guide the reasoning process. In LLaVA 1.5, the image-to-text information flow is comparatively stronger than in the other MLLMs, suggesting a greater dependence on visual input during reasoning. This difference may reflect its limited capabilities in CoT reasoning, as demonstrated also by Zeng *et al.* in [40]. In the other setting, we observe a marked reduction in the image-to-text information flow across all MLLMs. This decline indicates that all necessary information to answer the question is contained within the provided reasoning chain. In the case of LLaMA 3.2-Vision and Qwen 2.5-VL, the minimal image-to-text flow suggests that the models rely almost entirely on textual representations to infer the final answer. In contrast, LLaVA exhibits a residual degree of multimodal fusion, indicating that the model retains a limited reliance on visual input—likely due to its comparatively weaker capacity to generate an effective reasoning chain, as previously discussed. Overall, these findings highlight the effectiveness of CoT reasoning in decomposing complex visual questions into sequential steps that encapsulate the necessary information for accurate answer prediction.

**Word-Level Analysis.** Within the contexts of image captioning and CoT reasoning, which require long answers, we investigate which words rely most heavily on visual input and which can instead be inferred from linguistic context alone, without depending on multimodal integration. The results of this word-level analysis on LLaVA 1.5, LLaMA 3.2-Vision and Qwen 2.5-VL are presented in Tab. 1. We quantify word-level information flow by applying Eq. (8), where, in this analysis, $L$ denotes the number of tokens in each word, and $p$ and $\hat{p}$ are computed by multiplying the probabilities of the word's constituent tokens. Specifically, for each dataset, the table lists at the top the most frequently generated words with the highest image-to-text drop value (denoted as *Drop*)—corresponding to the minimum relative change in the answer probability attributed to the image-to-text information flow—and at the bottom the most frequently generated words with the lowest drop value. We select the most frequently generated words by retaining words whose occurrence across the dataset exceeds the $90^{\text{th}}$-percentile frequency, thereby ensuring that our analysis accurately reflects the models' generation behavior. In both datasets, the words requiring more visual grounding differ between the MLLMs. Nonetheless, these words exhibit shared characteristics. In ChartQA, they are typically action verbs that describe cognitive or procedural steps in the reasoning process—such as *examine, look, arrange, identify, subtract, corresponding, find, associated*—or words that organize the steps temporally or logically—such as *first, shown, since, final, following, section*. We report examples of CoT responses containing these words in the supplementary material to illustrate their usage in context. In COCO 2014, they typically serve to describe relationships—such as *near, situated, another, next*—provide contextual details—such as *showcasing, setting,*

Table 1. Results of the word-level analysis on LLaVA 1.5, LLaMA 3.2-Vision and Qwen 2.5-VL on ChartQA and COCO 2014. For each MLLM and dataset, the top five words are the most frequently generated terms with the highest image-to-text drop value—denoted as *Drop*—while the bottom five are the most frequently generated words with the lowest image-to-text drop value.

| | LLaVA 1.5 | | LLaMA 3.2-Vision | | Qwen 2.5-VL | |
|---|---|---|---|---|---|---|
| | **Word** | **Drop** | **Word** | **Drop** | **Word** | **Drop** |
| ChartQA | first | -66.20 | examine | -76.91 | examine | -62.22 |
| | corresponding | -62.30 | look | -69.93 | following | -38.59 |
| | shown | -60.43 | arrange | -64.44 | find | -37.14 |
| | since | -53.73 | subtract | -57.71 | section | -34.20 |
| | identify | -50.09 | final | -57.04 | associated | -34.03 |
| | need | -2.03 | need | -2.61 | the | -3.13 |
| | a | -2.95 | of | -3.15 | of | -3.75 |
| | we | -3.85 | at | -5.21 | to | -5.73 |
| | to | -4.85 | we | -6.68 | need | -5.80 |
| | of | -5.57 | the | -7.12 | in | -8.96 |
| COCO 2014 | that | -81.77 | showcasing | -79.46 | while | -65.50 |
| | another | -70.15 | setting | -79.21 | setting | -64.82 |
| | using | -59.67 | features | -65.01 | enjoying | -56.08 |
| | near | -55.39 | situated | -63.75 | another | -54.52 |
| | while | -50.14 | featuring | -50.80 | next | -53.39 |
| | food | -1.25 | image | -0.34 | image | -0.60 |
| | of | -2.48 | of | -3.53 | by | -3.48 |
| | to | -4.84 | depicts | -4.65 | of | -4.90 |
| | his | -6.42 | to | -9.07 | are | -7.41 |
| | a | -9.98 | a | -9.72 | to | -7.53 |

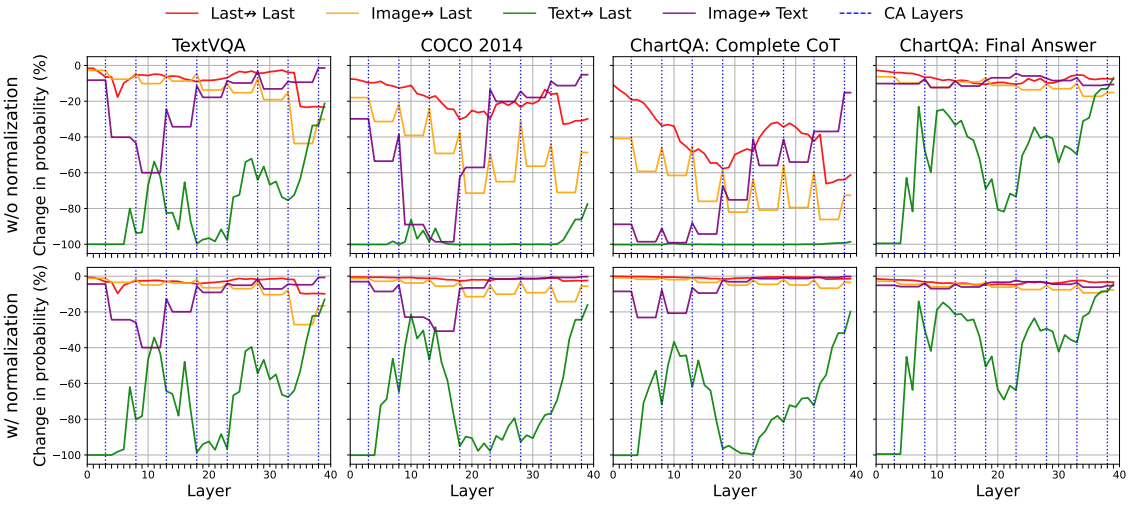


Figure 4. Relative changes in prediction probability for LLaMA 3.2-Vision across TextVQA, COCO 2014, and ChartQA datasets, computed with and without the $L$-th root normalization in Eq. (8). Dotted blue lines indicate cross-attention layers.

*features, featuring, enjoying*—or link elements within the image—such as *that, using, while*. In contrast, words that do not rely on visual input are typically structural words—such as pronouns, articles, and prepositions—which support the syntactic structure of the sentence rather than contributing to substantive reasoning or conveying multimodal content. The word *"need"* does not depend on visual input, as it functions to organize reasoning rather than to reference or interpret visual elements. In LLaMA 3.2-Vision and in Qwen 2.5-VL, the words *"image"* and *"depicts"* frequently appear at the beginning of captions—$i.e.$ *"The image depicts . . . "*— indicating that their occurrence is structurally driven rather than visually informed. Similarly, in LLaVA 1.5, the word *'food"* is usually generated in phrases such as *"a plate of food"* or *"a bowl of food"*, indicating that its generation is driven by captioning patterns rather than by explicit visual identification. Overall, our findings suggest that content words that convey spatial, temporal, procedural, or contextual relationships are typically visually grounded words. In contrast, structural words such as pronouns, articles, and prepositions, as well as terms commonly used in template-based expressions, exhibit minimal visual dependence, indicating that they are primarily generated from linguistic priors rather than visual content. Visualizations of the information flows previously analyzed—text-to-image, text-to-last, image-to-last, and last-to-last—for both highly and minimally visually grounded words are provided in the supplementary materials.

### 4.2. Ablation

**Effectiveness of the Normalization.** We demonstrate the effectiveness of our formulation defined in Eq. (8) for long, free-form answers by quantifying the information flows with and without the normalization factor, $i.e.$ the $L$-th root operator. This analysis is conducted on LLaMA 3.2-Vision across the TextVQA, COCO 2014, and ChartQA datasets, with ChartQA evaluated under the two previously described settings. Results are presented in Fig. 4. In TextVQA and in the final answer generation setting of ChartQA—both involving short answers—the previously observed patterns remain evident even without normalization, as the information flows are less affected by length-induced distortion. In contrast, in image captioning and CoT reasoning, which require longer responses, the information flow patterns become obscured without normalization as the unnormalized sequence probabilities induce length distortion in the information flows by diminishing exponentially with increasing output length. These findings highlight that normalization is particularly critical in long answers, as it enables fair and interpretable comparisons of multimodal information flows across tasks with varying output lengths, by preventing length-induced distortions.

## 5. Conclusion

We introduced FG-TRACER, a new framework that combines attention masking with a novel stable quantification of the information flow enabling the analysis of cross-modal information flow in MLLMs in free-form generation tasks. Our experiments on TextVQA, COCO 2014, and ChartQA reveal that multimodal fusion is both model- and task-dependent, uncovering distinct reasoning patterns in MLLMs. Our word-level analysis further indicates that procedural and relational terms are typically visually grounded words, whereas structural words are usually linguistically driven. Overall, FG-TRACER provides a robust approach for investigating reasoning dynamics in MLLMs and contributes to advancing multimodal interpretability.

**Acknowledgements.** This work was supported by Key Digital Technologies Joint Undertaking (KDT JU) in EdgeAI "Edge AI Technologies for Optimised Performance Embedded Processing" project, GA No. 101097300 and by the University of Modena and Reggio Emilia and Fondazione di Modena through the "Fondo di Ateneo per la Ricerca - FAR 2024" (CUP E93C24002080007). This work also received funding from DECIDER, the European Union's Horizon 2020 research and innovation programme under GA No. 965193 and "AIDA: explAinable multImodal Deep learning for personAlized oncology" (Project Code 20228MZFAA).

**Disclosure of Interests.** The authors have no conflicts of interest to declare.

## References

[1] Masry Ahmed et al. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning, 2022. 1, 2, 5

[2] Jean-Baptiste Alayrac et al. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*, 2022. 1, 2, 3

[3] Ebtesam Almazrouei et al. The Falcon Series of Open Language Models. *arXiv preprint arXiv:2311.16867*, 2023. 1, 2

[4] Singh Amanpreet et al. Towards VQA Models That Can Read, 2019. 1, 2, 5

[5] Shuai Bai et al. Qwen2.5-vl technical report, 2025. 2, 3, 5

[6] Samyadeep Basu et al. Understanding Information Storage and Transfer in Multi-modal Large Language Models. *arXiv preprint arXiv:2406.04236*, 2024. 2, 3

[7] Federico Bolelli et al. A Hierarchical Quasi-Recurrent approach to Video Captioning. In *IEEE International Conference on Image Processing, Applications and Systems (IPAS)*, 2018. 1

[8] Zhang Boqiang et al. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding, 2025. 1, 2

[9] Wei-Lin Chiang et al. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90% ChatGPT Quality. https://lmsys.org/blog/2023-03-30-vicuna/, 2023. 5

[10] Aakanksha Chowdhery et al. PaLM: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research*, 24(240), 2023. 1, 2

[11] Kevin Clark et al. What Does BERT Look At? An Analysis of BERT's Attention. *arXiv preprint arXiv:1906.04341*, 2019. 3

[12] Alexey Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2, 5

[13] Yuxin Fang et al. EVA: Exploring the Limits of Masked Visual Representation Learning at Scale. In *Computer Vision and Pattern Recognition*, 2023. 1, 2

[14] Aaron Grattafiori et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024. 1, 2, 3, 5

[15] Jiaming Han et al. OneLLM: One Framework to Align All Modalities with Language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2

[16] Liu Haotian et al. Improved Baselines with Visual Instruction Tuning, 2024. 2, 5

[17] Drew A. Hudson and Christopher D. Manning. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering, 2019. 2, 3, 6

[18] Jiahao Huo et al. MMNeuron: Discovering Neuron-Level Domain-Specific Interpretation in Multimodal Large Language Model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 6801–6816. Association for Computational Linguistics, 2024. 3

[19] Junnan Li et al. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning*, 2023. 3

[20] Haotian Liu et al. Visual Instruction Tuning, 2023. 3

[21] Minesh Mathew et al. DocVQA: A Dataset for VQA on Document Images. In *Winter Conference on Applications of Computer Vision*, 2021. 1

[22] Kevin Meng et al. Mass-Editing Memory in a Transformer. *arXiv preprint arXiv:2210.07229*, 2022. 2, 3

[23] Kevin Meng et al. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems*, 2022. 3

[24] Clement Neo et al. Towards Interpreting Visual Information Processing in Vision-Language Models. *arXiv preprint arXiv:2410.07149*, 2024. 2, 3

[25] Vedant Palit et al. Towards Vision-Language Mechanistic Interpretability: A Causal Tracing Tool for BLIP. In *International Conference on Computer Vision*, 2023. 2, 3

[26] Kiho Park et al. The Linear Representation Hypothesis and the Geometry of Large Language Models. *arXiv preprint arXiv:2311.03658*, 2023. 3

[27] Vittorio Pipoli et al. MissRAG: Addressing the Missing Modality Challenge in Multimodal Large Language Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3215–3224, 2025. 1

[28] Qwen et al. Qwen2.5 technical report, 2025. 5

[29] Alec Radford et al. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*, 2021. 1, 2, 5

[30] Alessia Saporita et al. Tracing Information Flow in LLaMA Vision: A Step Toward Multimodal Understanding. In *21st International Conference in Computer Analysis of Images and Patterns*, 2025. 2, 3

[31] Sarah Schwettmann et al. Multimodal Neurons in Pretrained Text-Only Transformers, 2023. 3

[32] Mingxu Tao et al. Probing Multimodal Large Language Models for Global and Local Semantic Representations, 2024. 3

[33] Gemma Team et al. Gemma: Open Models Based on Gemini Research and Technology. *arXiv preprint arXiv:2403.08295*, 2024. 1, 2

[34] Hugo Touvron et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*, 2023. 1, 2

[35] Lin Tsung-Yi et al. Microsoft COCO: Common Objects in Context, 2015. 1, 2, 5

[36] Ashish Vaswani et al. Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30, 2017. 3

[37] Jin Xu et al. Qwen2.5-omni technical report, 2025. 1, 2

[38] Hao Yin et al. Lifting the Veil on Visual Information Flow in MLLMs: Unlocking Pathways to Faster Inference. In *Computer Vision and Pattern Recognition Conference*, 2025. 2, 3

[39] Liu Ze et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, 2021. 1

[40] Xingchen Zeng, Haichuan Lin, Yilin Ye, and Wei Zeng. Advancing Multimodal Large Language Models in Chart Question Answering with Visualization-Referenced Instruction Tuning, 2024. 7

[41] Lin Zhang et al. Mechanistic Unveiling of Transformer Circuits: Self-Influence as a Key to Model Reasoning. *arXiv preprint arXiv:2502.09022*, 2025. 3

[42] Xiaofeng Zhang et al. From Redundancy to Relevance: Information Flow in LVLMs Across Reasoning Tasks. *arXiv preprint arXiv:2406.06579*, 2024. 2, 3

[43] Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal Information Flow in Multimodal Large Language Models, 2025. 2, 3, 4, 6

[44] Peng Zhiliang et al. Kosmos-2: Grounding multimodal large language models to the world, 2023. 1, 2

[45] Deyao Zhu et al. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 2, 3

[46] Zhao Zijia et al. Chatbridge: Bridging modalities with large language model as a language catalyst, 2023. 1, 2

# FG-TRACER: Tracing Information Flow in Multimodal Large Language Models in Free-Form Generation - *Supplementary Material*

## 1. ChartQA: Word Information Flows

We analyze the information flows—last-to-last, image-to-last, text-to-last, and image-to-text—related to the content words (*examine, look, arrange, subtract, final*) and structural words (*need, of, at, we, the*) of ChartQA [1] dataset, using LLaMA 3.2-Vision [4]. The results are presented in Fig. 5. We observe that content words exhibit a strong image-to-text information flow, with multimodal fusion occurring primarily in the early to mid layers. These results demonstrate the strong coupling between these words and visual input. In contrast, for structural words, the image-to-text flow is negligible, while the text-to-last flow remains strong, indicating that their activation is driven more by linguistic priors rather than visual context.

## 2. ChartQA: Responses with CoT Reasoning

Additionally, we provide four examples of question–answer pairs from the ChartQA dataset, generated using chain-of-thought (CoT) prompting with LLaMA3.2-Vision, to illustrate the structure of the model's reasoning process and contextualize the content words (highlighted in bold).

**Question 1:** *What's the value of the largest yellow bar?*
**Response 1:** To determine the value of the largest yellow bar, we must **examine** the chart and locate the bar with the highest percentage. The chart displays the share of representatives among each group who have served 1-3 terms in the House (114th Congress).
The yellow bars represent the percentage of representatives in each group.
Key information from the chart: the largest yellow bar corresponds to "Freedom Republicans" with a value of 72.
**FINAL** ANSWER: 72.

**Question 2:** *What is the number of deaths per 100,000 in 2002 in Colombia?*
**Response 2:** To find the number of deaths per 100,000 in 2002 in Colombia, we can **look** at the bar graph provided. The x-axis lists the countries, and the y-axis represents the number of deaths per 100,000 persons.
Looking at the bar for Colombia, we can see that it is the second bar from the left.
The y-axis starts at 0 and increases in increments of 50%, with labels at 50, 100, 150, 200, and 250.
Given the scale, the bar for Colombia appears to be slightly below the 175 mark. Therefore, based on the bar graph, the number of deaths per 100,000 in 2002 in Colombia is approximately 175.
**FINAL** ANSWER: 175.

**Question 3:** *Is the median value of the light green bars greater than 45?*
**Response 3:** To determine the median value of the light green bars, we need to identify the values represented by these bars in the chart. The light green bars correspond to the percentage of people who trust information from the Chinese government regarding the coronavirus outbreak.
To find the median, we **arrange** these values in ascending order: 13%, 40%, 51%. Since there are an odd number of values (three), the middle value is the median. In this case, the median value is 40%.
Comparing this median value to 45%, we can see that the median value of the light green bars (40%) is less than 45%.
**FINAL** ANSWER: No.

**Question 4:** *What is the difference in the value of High blood sugar and High blood pressure?*
**Response 4:** To determine the difference between High blood sugar and High Blood pressure, we **subtract** the value of High Blood pressure from the value of High blood sugar. Values:
High blood sugar: 879
High Blood pressure: 676
Calculation: Difference = High blood sugar - High Blood pressure = 879 - 676 = 203
**FINAL** ANSWER: 203.

## 3. COCO 2014: Word Information Flows

We also analyze the information flows—last-to-last, image-to-last, text-to-last, and image-to-text—associated with the content words (*showcasing, setting, features, situated, featuring*) and the structural words (*image, of, depicts, to, a*) of COCO 2014 [6] dataset (Tab. 1 of the main paper), using LLaMA 3.2-Vision. The results are presented in Fig. 6. For the content words, a consistent image-to-text information flow is observed across all terms, suggesting a close association between these words and visual input. Conversely,

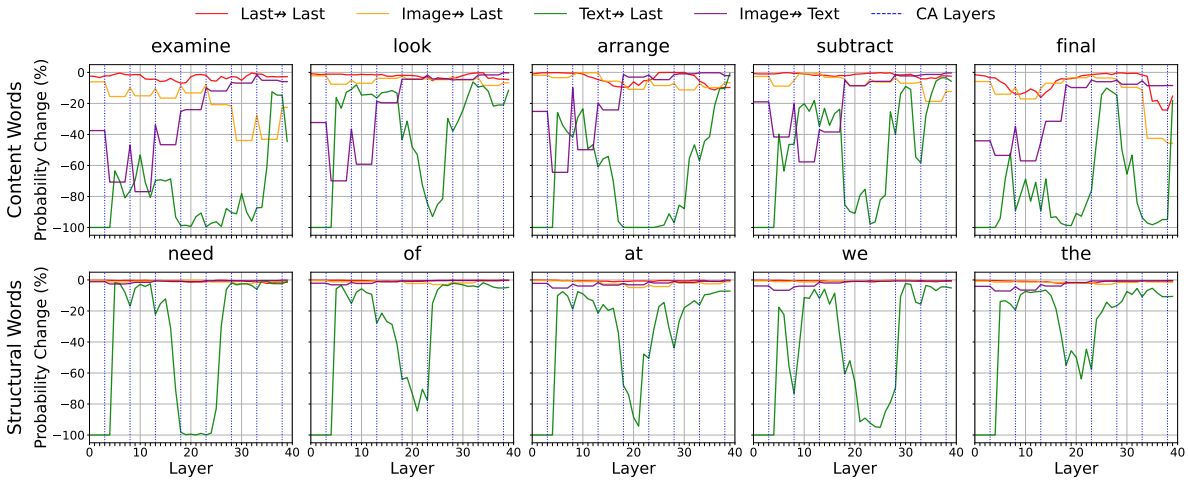


Figure 5. Information flow patterns—last-to-last, image-to-last, text-to-last, and image-to-text—for LLaMA 3.2-Vision on the ChartQA dataset, analyzed for content and structural words. Dotted blue lines indicate cross-attention layers.

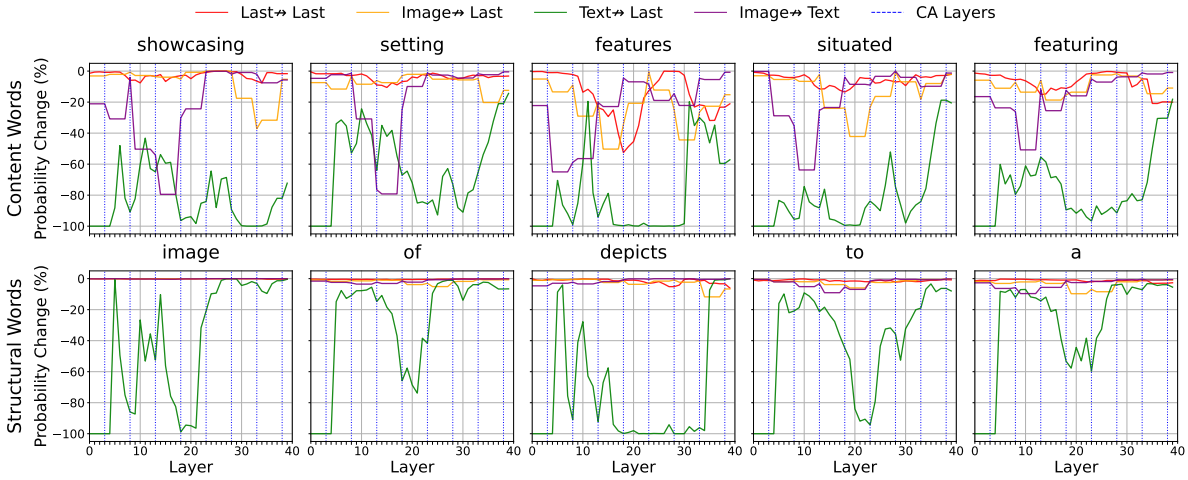


Figure 6. Information flow patterns—last-to-last, image-to-last, text-to-last, and image-to-text—for LLaMA 3.2-Vision on the COCO 2014 dataset, analyzed for content and structural words. Dotted blue lines indicate cross-attention layers.

structural words exhibit negligible image-to-text information flow, but show a strong text-to-last information flow, indicating that their activation is largely influenced by linguistic structure rather than visual cues.

## 4. COCO 2014: Image Captions

Furthermore, we present four representative COCO captions processed by LLaMA 3.2-Vision to illustrate the contextual usage of the content words (highlighted in bold) reported in Tab. 1 of the main paper.

**Caption 1.** This image is a close-up photograph of a toucan in a cage, **showcasing** its vibrant colors and unique beak.

**Caption 2.** The image **features** a plate of food, a bottle of wine, and a plant on a table in an outdoor **setting**.

**Caption 3.** The image depicts a traffic light with a green smiley face on it, **situated** in a suburban area.

**Caption 4.** The image depicts a fire hydrant spraying water into the air on a city street, with a brick building in the background **featuring** graffiti and a sign indicating a parking restriction.

**Caption 5.** The image depicts a Starbucks coffee cup, a muffin, and a banana on a table, showcasing a casual breakfast or snack **setting**.

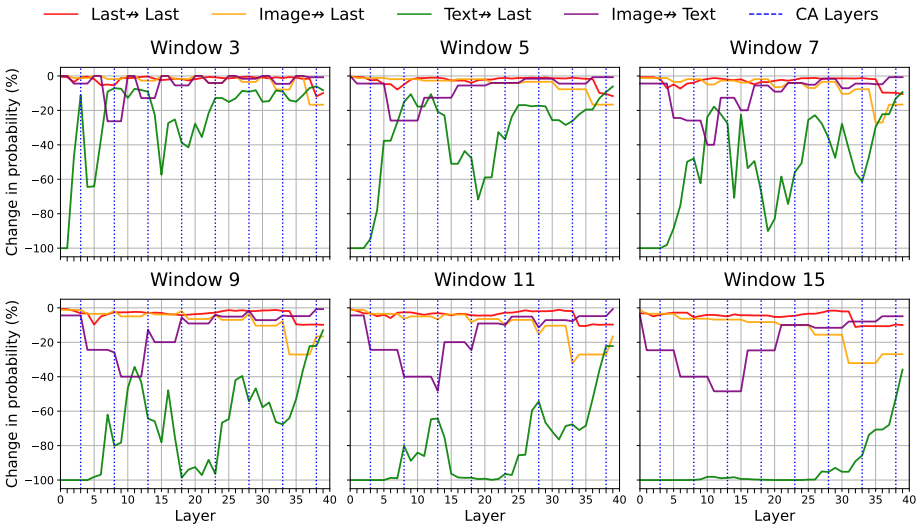


Figure 7. Relative changes in prediction probabilities for LLaMA 3.2-Vision on the TextVQA dataset across different window sizes.


Figure 8. Relative changes in prediction probabilities for LLaMA 3.2-Vision on the COCO 2014 dataset across different window sizes.

## 5. Information Flows for Different Window Sizes

In the experiments reported in the main paper, we employed a window size of $k = 9$. To evaluate the effect of window size on model behavior, we analyze the relative change in answer probabilities in LLaMA 3.2-Vision on the TextVQA [2] and COCO 2014 [6] datasets across a range of window sizes: $k$ = 3, 5, 7, 9, 11, 15. The results are presented in Fig. 7 and Fig. 8. Overall, the patterns of the information flows—last-to-last, image-to-last, text-to-last, and image-to-text—remain consistent across different values of $k$ in both datasets, with multimodal integration occurring primarily in the early to mid layers. However, the magnitude of change in output probability increases with larger window sizes. This trend is expected, as more restrictive attention windows reduce the number of accessible attention edges, thereby limiting the model's capacity to contextualize the input effectively. Notably, with small $k$ ($e.g.$, 3 or 5), masking applies only to a few layers, and we mainly observe local information flow patterns, which limits our ability to study broader multimodal dynamics. Instead, with large $k$ ($e.g.$, 11, 15), masking spans many layers, overly suppressing cross-modal interactions. The choice of using k=9 provides a balanced setting: it is large enough to reveal meaningful effects beyond local patterns, yet not so large that it disrupts information flow excessively.

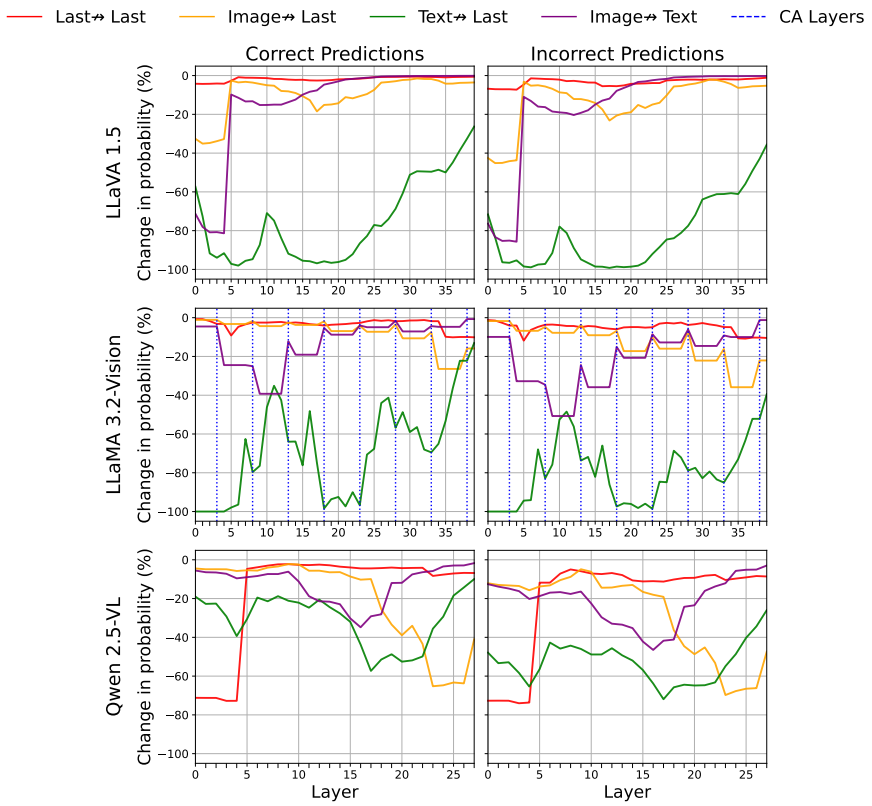


Figure 9. Relative changes in prediction probabilities for LLaVA 1.5, LLaMA 3.2-Vision, and Qwen 2.5-VL on the TextVQA dataset using only correct or wrong samples.

## 6. Information Flows for Wrong Samples

In the main paper, we reported results based exclusively on correctly answered samples for each model and dataset, in order to minimize potential noise introduced by incorrect predictions. To examine whether errors might reveal distinct patterns of multimodal fusion, we conducted an additional analysis on the subset of incorrect samples from the TextVQA dataset—*i.e.*, cases where the models produced wrong predictions—on LLaVA 1.5 [5], LLaMA 3.2-Vision, and Qwen 2.5-VL [3]. The experimental methodology was identical to that employed in the main study on correct samples. Specifically, we computed the same information flows—last-to-last, image-to-last, text-to-last, and image-to-text—thereby enabling a direct comparison between correct and incorrect subsets. The corresponding results are shown in Fig. 9. Notably, across all three MLLMs, the results obtained for incorrect predictions remain consistent with those observed for the correct subset: the same patterns of information flow emerge. These findings indicate that the underlying reasoning dynamics are a general property of model behavior, and that failures are not attributable to disruptions or breakdowns in multimodal fusion. These results highlight the robustness of information flow patterns across both successful and unsuccessful predictions.

## References


[1] Masry Ahmed et al. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning, 2022. 1

[2] Singh Amanpreet et al. Towards VQA Models That Can Read, 2019. 3

[3] Shuai Bai et al. Qwen2.5-vl technical report, 2025. 4

[4] Aaron Grattafiori et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024. 1

[5] Liu Haotian et al. Improved Baselines with Visual Instruction Tuning, 2024. 4

[6] Lin Tsung-Yi et al. Microsoft COCO: Common Objects in Context, 2015. 1, 3