Traceable and Explainable Multimodal Large Language Models: An Information-Theoretic View

Zihan Huang[1]*, **Junda Wu**[1]*, **Rohan Surana**[1], **Raghav Jain**[1], **Tong Yu**[2], **Raghavendra Addanki**[2], **David Arbour**[2], **Sungchul Kim**[2], **Julian McAuley**[1]
[1]UC San Diego [2]Adobe Research
{zih043,juw069,rsurana,r6jain,jmcauley}@ucsd.edu
{tyu,raddanki,arbour,sukim}@adobe.com

Abstract

Existing multimodal large language models (MLLMs) often lack traceable and explainable mechanisms for visual-textual alignment, making it challenging to understand how textual instructions shape multimodal representations. To address this shortcoming, we propose an information-theoretic framework that clarifies how MLLMs handle and transform both text and visual inputs. In particular, we measure the visual information gain that arises from textual instructions and multimodal encodings, thereby illuminating how different modalities interact and contribute to the model's overall processing. Our framework decomposes the multimodal encoding process into layer-wise mutual information measures for better explainability, quantifying the visual contribution as the difference between unconditional and text-conditional mutual information. Specifically, inspired by the Information Bottleneck framework, we introduce a Concept Bottleneck that maps high-dimensional multimodal representations into an interpretable space, enabling tractable variational upper bounds on the mutual information between visual inputs and the model's internal states. Furthermore, we quantify the contextual contribution introduced by textual cues via an InfoNCE mechanism that contrasts multimodal representations computed with and without text guidance. This dual perspective, facilitated by tractable variational upper bounds, provides insight into how visual information is encoded and filtered by textual instructions, while also highlighting the contextual information induced and enhanced by MLLMs. Empirical findings demonstrate underexplored dynamics of visual-textual interaction within MLLMs, underscoring how textual instructions distinctly shape visual representations and demonstrating how visual prompts, when effectively paired with instructions, enhance multimodal understanding.

1 Introduction

Multimodal Large Language Models (MLLMs) have advanced in integrating multimodal information, achieving impressive results in tasks such as question answering (Zhang et al., 2023b; Wu et al., 2025c;b), captioning (Liu et al., 2025; Wu et al., 2024c), navigation (Wang et al., 2025; Nguyen et al., 2024) and multimodal retrieval (Bao et al., 2025; Li et al., 2024; Wu et al., 2025a; Huang et al., 2025; Wu et al., 2024a). However, understanding how textual instructions shape multimodal representations within these models remains challenging (Zhao et al., 2024; Wu et al., 2024b; Wang et al., 2024b). Prior works have attempted to analyze multimodal alignment using attention-based visualization techniques (Yeh et al., 2024; DeRose et al., 2021; Jaunet et al., 2022; Wu et al., 2021) and perturbation-based sensitivity analyses (Lundberg & Lee, 2017; Ribeiro et al., 2016; Wu et al., 2025d).

While these methods provide insights into which visual features influence model outputs, they lack a principled and quantifiable framework to distinguish the intrinsic contribution

*These authors contributed equally to this work.

of visual inputs from the modifications introduced by textual guidance in multimodal encoding. Consequently, they struggle to explain how visual information is retained, transformed, or compressed by textual instructions. Recent works propose to capture layer-wise information flow in text-only LLMs, detecting information deficiencies across layers (Kim et al., 2024a; Wu et al., 2022). However, existing techniques fail to disentangle retained visual content from text-driven refinements across different model layers, making it difficult to trace the flow of multimodal information. Without a rigorous framework to track these interactions, current explainability methods remain insufficient for systematically understanding MLLMs. As shown in Figure 1, conventional similarity metrics either exhibit monotonic trends or degenerate discriminative power in later layers, failing to reveal the nuanced processing stages. More critically, they lack theoretical grounding for interpreting multimodal interactions.

To address this limitation, we propose an information-theoretic framework to systematically quantify the interplay between visual and textual modalities in MLLMs. Inspired by the Information Bottleneck principle (Zhu et al., 2025; Yang et al., 2024; Tishby et al., 2000b; Wu et al., 2023), our framework leverages a Concept Bottleneck (Yamaguchi et al., 2025; Lai et al., 2024; Koh et al., 2020a) that maps complex multimodal representations into interpretable latent spaces, enabling a tractable measurement of visual information specifically retained, transformed, or enhanced by textual instructions. Furthermore, we leverage an InfoNCE-based contrastive mechanism to distinguish the contributions of textual instructions explicitly, offering clearer insights into how textual contexts shape multimodal processing.

Combining the above theoretical probes, we conduct comprehensive empirical studies revealing underexplored dynamics of multimodal information processing within MLLMs (Luo & Specia, 2024; Zhao et al., 2024; Zhuang et al., 2022). We uncover that visual information evolves through multi-stage transformations shaped by textual instructions, concerning task-specific contexts (Awal et al., 2025; Hu et al., 2023) and visual complexities (Mu et al., 2024). In addition, we demonstrate that disruptive textual instructions significantly degrade the effectiveness of visual representation processing, highlighting the vulnerability of MLLMs to linguistic distractions, which leads to potential hallucinations (Liu et al., 2024a). We reveal that visual prompting significantly enhances representation quality only when effectively paired with textual guidance (Nguyen et al., 2023; Wu et al., 2024d), emphasizing the critical synergy between textual and visual cues. Finally, our analysis illustrates how different types of instructions shape the multimodal interaction patterns within MLLMs, offering novel insights into designing more robust and explainable multimodal systems. We summarize our contributions as follows:

- We introduce an information-theoretic framework based on the Information Bottleneck principle to systematically quantify and analyze interactions between visual and textual modalities across different layers of MLLMs.
- We leverage a Concept Bottleneck to enable interpretable and traceable analyses of visual representations influenced by textual contexts, while an InfoNCE-based metric is proposed to explicitly measure the contextual contributions of instructions.
- Our extensive empirical analyses provide insights into multimodal information dynamics, revealing underexplored interactions between visual prompting, textual instructions, and their combined effects on representation processing in MLLMs.

2 Related Works

2.1 Explainable Multimodal Large Language Models

Yu & Ananiadou (2025) study Llava's VQA mechanism, showing visual embeddings encode semantic features and proposing a tool to identify key image regions. Neo et al. (2024) use ablations to show LLaVA localizes object info and aligns visual tokens with vocabulary representations across layers. Zhang et al. (2024b) reveal that LLaVA integrates visual and textual info in two stages, general and targeted, before final prediction. Qi et al. (2025) find VLMs lack spatial awareness due to large embedding norms and proposes normalization

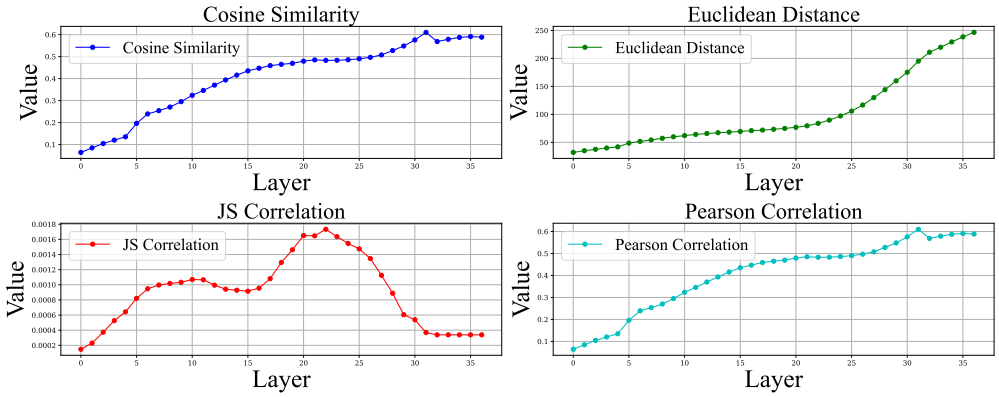


Figure 1: Evaluation result of traditional diagnostic metrics.

to improve spatial reasoning. Zhang et al. (2024a) show shallow layers contain redundant image features and introduce a truncation method to enhance interpretability and accuracy. Unlike prior works that focus on attention patterns or token-level visual attribution, our approach provides a quantifiable, layer-wise decomposition of visual information shaped by textual cues, offering deeper traceability of multimodal interactions.

2.2 Information-Theoretic Approaches in Multimodal Large Language Models

Kim et al. (2024a) propose a layer-wise information framework to detect hallucinations by tracking information flow across model layers. Wang et al. (2024c) analyze semantic vs. linguistic input influence in NLP tasks, showing semantics dominate, especially in sentiment classification. Federici et al. (2021) present an info-theoretic view of distribution shifts, linking test error to data assumptions and generalization objectives. Ton et al. (2024) quantify information gain at each CoT step, enabling unsupervised detection of faulty reasoning in LLMs. Zhang et al. (2025) introduce Entro-duction, using entropy and its variance to dynamically guide reasoning depth in multi-step tasks. While existing methods use information-theoretic tools to detect failures or assess influence, our framework uniquely quantifies visual information gain conditioned on textual instructions, enabling fine-grained analysis of how MLLMs encode, transform, and align modalities.

3 An Information-theoretic View

A central goal of this section is to develop **traceable** and **explainable** mechanisms for analyzing how MLLMs integrate and process visual and textual information. To this end, we introduce a framework that provides a principled and quantifiable way to measure how visual signals are filtered, retained, or transformed in the presence of textual instructions.

3.1 Multimodal Encoding in MLLMs

In a multimodal large language model, given image input V, the text instruction T provides contextual guidance that influences the processing of the visual information. The image is first processed by an image encoder f based on a Vision Transformer (ViT) and a multimodal projector (Liu et al., 2023), which encode the input image as image tokens $f(V)$. These image tokens are subsequently concatenated with language tokens derived from the text input T, before input into the branch of the large language model backbone π,

$$\left(\mathbf{X}^{(l)}\right)_{l=1}^{N} = g_{\pi}\left([f(V), T]\right), \tag{1}$$

where g_{π} is the multi-layer Transformer that encodes multimodal inputs as intermediate multimodal representations $\mathbf{X}^{(l)}$ for each layer l. In addition, we denote the encoded visual

representations as $X^{(l)}_{|f(V)|} \in \mathbf{X}^{(l)}$, while the multimodal representations as $X^{(l)}_{|f(V),T|} \in \mathbf{X}^{(l)}$, at the last token position of the image and the multimodal token sequences, respectively.

3.2 Multimodal Information Flow in MLLMs

Following the information bottleneck (IB) framework (Tishby et al., 2000a; Shwartz-Ziv & Tishby, 2017), we design the layer-wise measurement

$$\mathcal{F}_V^{(l)} = I\Big(X^{(l)}_{|f(V),T|}; V\Big) - I\Big(X^{(l)}_{|f(V),T|}; V \mid T\Big) \tag{2}$$

to capture the portion of visual information in the layer-l representation, specifically extracted in the presence of textual context. Intuitively, although $X^{(l)}_{|f(V),T|}$ may encode a rich set of visual details, not all of these details are pertinent once the textual context is taken into account. By subtracting the conditional mutual information $I\Big(X^{(l)}_{|f(V),T|}; V \mid T\Big)$ that quantifies the extraneous visual information given the text (Tishby et al., 2000a), we isolate those visual features that are enhanced by the accompanying text. This metric thus reflects the model's ability to filter out irrelevant visual noise while preserving features that contribute meaningfully to a multimodal understanding.

To formally track the quantities of the mutual information and the prompt-conditional mutual information in equation 2, we first propose to leverage Concept Bottleneck (Koh et al., 2020b) as a variational information bottleneck (in Lemma 3.1) to track each layer's multimodal representations in a list of candidate visual concepts, which are unbiased to the MLLM. Then, we quantify the mutual information terms in equation 2 by deriving the variational upper bound in the visual conceptual space for the unconditional mutual information (in Lemma 3.2) and the prompt-conditional mutual information (in Lemma 3.3).

Lemma 3.1 (Concept Bottleneck for Mutual Information via Variational IB) *Let $X^{(l)}_{|f(V),T|}$ be the layer-l representation obtained when both the visual input V and the text instruction T are provided. Define the intermediate concept vector by $\hat{C} = f_{CB}\big(X^{(l)}_{|f(V),T|}\big)$, where $f_{CB}(\cdot)$ is a concept bottleneck function (Koh et al., 2020b; Kim et al., 2018). A variational encoder $q(\hat{C} \mid V)$ maps the visual input V to a latent concept distribution, where such concept vector as its unbiased prior distribution $r(\hat{C})$ over the concept space. In addition, the conditional prior $r(\hat{C} \mid T)$ approximates the aggregated posterior of $\hat{C}$ conditioned on the text instruction T.*

Then, the mutual information is approximated by the following variational upper bounds.

Lemma 3.2 (Variational Upper Bound for Mutual Information) *Given the variational encoder $q(\hat{C} \mid V)$ and prior $r(\hat{C})$, the mutual information between the concept representation and the visual input is upper-bounded by*

$$I\Big(X^{(l)}_{|f(V),T|}; V\Big) \leq \mathbb{E}_V\Big[KL\Big(q(\hat{C} \mid V) \,\|\, r(\hat{C})\Big)\Big]. \tag{3}$$

Lemma 3.3 (Variational Upper Bound for Conditional Mutual Information) *Assume that samples can be grouped by text instruction, and let $r(\hat{C} \mid T)$ denote a variational approximation to the distribution of $\hat{C}$ given T. Then, the conditional mutual information is upper-bounded as*

$$I\Big(X^{(l)}_{|f(V),T|}; V \mid T\Big) \leq \mathbb{E}_T\Big[\mathbb{E}_{V|T}\Big[KL\Big(q(\hat{C} \mid V) \,\|\, r(\hat{C} \mid T)\Big)\Big]\Big]. \tag{4}$$

Thus, we define the aligned visual information by integrating equation 3 and equation 4:

$$\mathcal{F}_V^{(l)} \leq \mathbb{E}_V\Big[\mathrm{KL}\big(q(\hat{C} \mid V) \,\|\, r(\hat{C})\big)\Big] - \mathbb{E}_T\Big[\mathbb{E}_{V|T}\big[\mathrm{KL}\big(q(\hat{C} \mid V) \,\|\, r(\hat{C} \mid T)\big)\big]\Big]. \tag{5}$$

By employing the concept bottleneck representation $\hat{C}$ within this variational framework, we obtain an interpretable estimate of both the mutual information $I(\hat{C}; V)$ and the conditional mutual information $I(\hat{C}; V \mid T)$, thereby quantifying the extent to which visual features are aligned with and enhanced by the textual instruction.

3.3 Induced Contextual Information from MLLMs

In multimodal large language models (MLLMs), the text instruction T does more than provide textual guidance to the visual encoder; it also induces higher-level task-related contextual information into the LLM's internal representations. To quantify this induced contextual information, we define the task-related mutual information as the difference between the mutual information computed over multimodal representations (obtained when the text is present) and the mutual information computed over baseline representations (obtained without text). Formally, we write:

$$\mathcal{F}_T^{(l)} = I\Big(X_{|f(V),T|}^{(l)}; X_{|f(V),T|}^{(L)}\Big) - I\Big(X_{|f(V)|}^{(l)}; X_{|f(V)|}^{(L)}\Big),$$

where $X_{|f(V),T|}^{(l)}$ and $X_{|f(V),T|}^{(L)}$ denote the layer-l and the final layer L representations computed with the text instruction T, respectively, and $X_{|f(V)|}^{(l)}$ and $X_{|f(V)|}^{(L)}$ denote the corresponding representations when the textual guidance is absent. This formulation isolates the additional contextual information that is induced by the LLM's language processing branch. To estimate these mutual information terms in practice, we adopt an InfoNCE-based approach. Recall that InfoNCE provides a tractable lower bound for mutual information by contrasting positive pairs against negative pairs. We define two losses for a mini-batch of N samples.

Lemma 3.4 (InfoNCE Lower Bound for Task-Related Mutual Information) *For a mini-batch of N samples, let $X_{|f(V),T|}^{(l),i}$ and $X_{|f(V),T|}^{(L),i}$ denote the layer-l and layer-L representations computed with the text instruction T for the i-th sample, and similarly let $X_{|f(V)|}^{(l),i}$ and $X_{|f(V)|}^{(L),i}$ denote the corresponding baseline representations computed without T. With a similarity function $s(\cdot,\cdot)$ and a temperature parameter $\tau > 0$, we define*

$$\mathcal{L}_{|f(V),T|}^{(l)} = -\frac{1}{N}\sum_{i=1}^{N}\left[\log\exp\left(\frac{s\big(X_{|f(V),T|}^{(l),i},\, X_{|f(V),T|}^{(L),i}\big)}{\tau}\right) - \log\sum_{j=1}^{N}\exp\left(\frac{s\big(X_{|f(V),T|}^{(l),i},\, X_{|f(V),T|}^{(L),j}\big)}{\tau}\right)\right], \tag{6}$$

$$\mathcal{L}_{|f(V)|}^{(l)} = -\frac{1}{N}\sum_{i=1}^{N}\left[\log\exp\left(\frac{s\big(X_{|f(V)|}^{(l),i},\, X_{|f(V)|}^{(L),i}\big)}{\tau}\right) - \log\sum_{j=1}^{N}\exp\left(\frac{s\big(X_{|f(V)|}^{(l),i},\, X_{|f(V)|}^{(L),j}\big)}{\tau}\right)\right], \tag{7}$$

which provide lower bounds on the mutual information between the multimodal representations,

$$I\Big(X_{|f(V),T|}^{(l)}; X_{|f(V),T|}^{(L)}\Big) \geq \log N - \mathcal{L}_{|f(V),T|}^{(l)}, \quad I\Big(X_{|f(V)|}^{(l)}; X_{|f(V)|}^{(L)}\Big) \geq \log N - \mathcal{L}_{|f(V)|}^{(l)}. \tag{8}$$

Finally, by the estimation of the above InfoNCE lower bounds (detailed in Appendix B), we provide an estimate for the task-related mutual information:

$$\mathcal{F}_T^{(l)} \geq \Big[\log N - \mathcal{L}_{|f(V),T|}^{(l)}\Big] - \Big[\log N - \mathcal{L}_{|f(V)|}^{(l)}\Big] = \mathcal{L}_{|f(V)|}^{(l)} - \mathcal{L}_{|f(V),T|}^{(l)}. \tag{9}$$

This result quantifies the benefit of the contexts in injecting task-related information into the model. In particular, a lower $\mathcal{L}_{|f(V),T|}^{(l)}$, indicating higher mutual information in the presence of T, relative to the baseline loss $\mathcal{L}_{|f(V)|}^{(l)}$ suggests that the text instruction effectively aligns and enriches the representations, thereby enhancing the multimodal understanding of the task.

4 Experiments

To validate our framework, we instantiate the two core metrics, $\mathcal{F}_V^{(l)}$ and $\mathcal{F}_T^{(l)}$, introduced in Section 3. These metrics allow us to systematically probe how visual and textual modalities

interact across layers in MLLMs. We conduct extensive experiments to investigate how these interactions manifest across different tasks, input settings, and instruction types, addressing the following research questions (RQs), each targeting a specific aspect of traceability and explainability in multimodal processing:

- **RQ1**: How do textual instructions shape visual information processing?
- **RQ2**: How do textual instructions influence the multimodal information flow?
- **RQ3**: How do visual prompts interact with textual instructions in MLLMs?
- **RQ4**: How do multimodal representations vary across different instructions?

Datasets. We assess the behavioral differences of MLLMs in processing various types of data based on several image-to-text datasets. Specifically, we employ the Visual Question Answering-v2 (VQA-v2) (Goyal et al., 2017) and AOKVQA (Schwenk et al., 2022) datasets to investigate MLLMs' performance in Visual QA challenges. Additionally, we use image data from VQA (Antol et al., 2015) to implement and train our $\mathcal{F}_V$ and $\mathcal{F}_T$ evaluators. We leverage the validation set of COCO-caption (Chen et al., 2015) and the Fine-grained Hallucination Evaluation Framework (Hal-Eval) (Jiang et al., 2024) to explore MLLMs' behavior in Visual Captioning challenges.

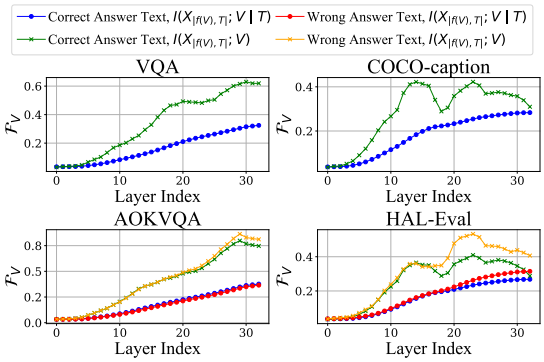
Models. We use LLaVA-1.5 (Liu et al., 2024b) with 7B parameters and Qwen-2.5-VL (Bai et al., 2023) with 3B parameters to extract the visual hidden states information with different settings on various datasets. we use YOLO11-large (Khanam & Hussain, 2024) to annotate object distribution prior to detecting objects on the images. We further illustrate the implementation details in Appendix A.

4.1 (RQ1) How do textual instructions shape visual information processing?

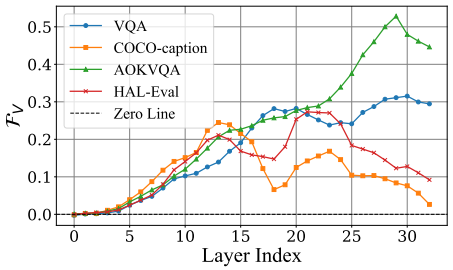
We examine the internal layer-wise processing of MLLMs, specifically focusing on how textual guidance shapes visual representations (Wu et al., 2024e; Niu et al., 2024). Using explainable metrics ($\mathcal{F}_V$, $\mathcal{F}_T$), we analyze the transformation of these representations across layers and investigate variations influenced by task type, visual complexity, and textual precision (Awal et al., 2025; Hu et al., 2023; Mu et al., 2024). Intuitively, an increase in $\mathcal{F}_V$ indicates that visual information is being injected or actively processed by the MLLMs, causing the representations to differ significantly from their initial margin-based states. The metric $\mathcal{F}_V$ measures relevant visual information retained after considering textual context. A decrease indicates effective filtering of irrelevant visual details. The metric $\mathcal{F}_T$ reflects task-related contextual information injected by textual instructions. While filtering visual information, the model integrates task-specific contextual information from the text, enhancing multimodal representations and increasing $\mathcal{F}_T$. The interplay between $\mathcal{F}_V$ and $\mathcal{F}_T$ represents a balancing act. As the model filters out irrelevant visual complexity (lowering $\mathcal{F}_V$), it simultaneously injects relevant textual information, increasing $\mathcal{F}_T$. This dynamic maintains effective multimodal understanding, optimizing task performance.

Finding 1. Multimodal encoding process presents 4 stages. When examining the layer-to-layer representation transformation, we observe that the visual representation is not processed in a straightforward manner with a monotonous trend. Instead, the information undergoes a complex processing sequence across multiple stages, which we categorize as follows: (1) Information Pre-processing: an increase in $\mathcal{F}_V$ and a stable $\mathcal{F}_T$; (2) Information Filtering: a decrease in $\mathcal{F}_V$ and stable or decrease in $\mathcal{F}_T$; (3) Contextual Information Injection and Multimodal Mixture: an increase in $\mathcal{F}_V$ and increase or stable $\mathcal{F}_T$; (4) Information Compression: a decrease in $\mathcal{F}_V$ and increase in $\mathcal{F}_T$, as is shown in Figure 2c, 2d, 3.

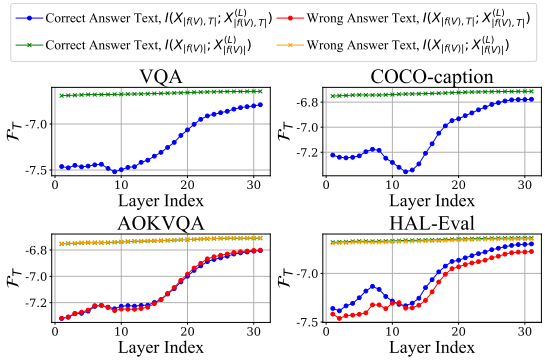
We further validate this staged processing through targeted interventions. As shown in Figure 4, masking image tokens causes the characteristic dynamics of Stages 1-3 to collapse into monotonic trends, while Stage 4 remains unaffected. This demonstrates that: (1) The first three stages are indeed visual-dependent processes, and (2) The final compression stage operates primarily on textual representations, consistent with its role in response generation.



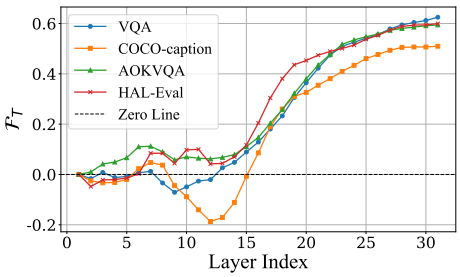
(a) $I(X_{|f(V),T|}; V \mid T)$ and $I(X_{|f(V),T|}; V)$.



(b) $\mathcal{F}_V$ across layers.

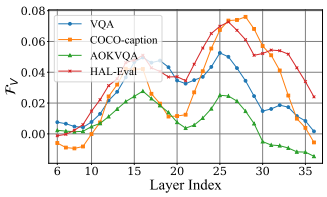


(c) $I(X^{(l)}_{|f(V),T|}; X^{(L)}_{|f(V),T|})$ and $I(X^{(l)}_{|f(V),T|}; X^{(L)}_{|f(V),T|})$.

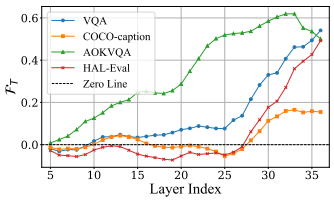


(d) $\mathcal{F}_T$ across layers.

Figure 2: $\mathcal{F}_V$ and $\mathcal{F}_T$ across different layers in LLaVA-1.5-7b.



(a) $\mathcal{F}_V$ across layers.



(b) $\mathcal{F}_T$ across layers.

Figure 3: $\mathcal{F}_V$ and $\mathcal{F}_T$ across layers in Qwen-2.5-VL-3b.

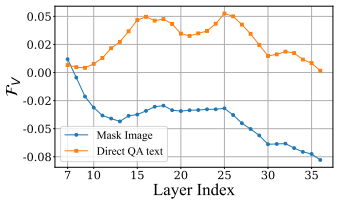


Figure 4: Qwen $\mathcal{F}_V$ across layers after masking images.

Finding 2. In captioning tasks, multimodal information processing and compression dynamically interplay across various layers of MLLMs. As shown in Figure 2b, 3a, we detect two peaks in the metric $\mathcal{F}_V$: around layer 12 and 22 in LLaVA or around 15 and 25 in Qwen. On COCO-caption which excludes imprecise answers, $\mathcal{F}_V$ increases before layer 12 in LLaVA and layer 15 in Qwen, as is on Figure 2b, 3a, while $\mathcal{F}_T$ remains stable or decreases, as is on Figure 2d, 3b. This indicates that the MLLMs are processing visual information from images rather than injecting information to align it with the output results. Between layer 12 to 18 in LLaVA or layer 15 to 21 in Qwen, both $\mathcal{F}_V$ and $\mathcal{F}_T$ decrease, suggesting an information filtering process by the MLLMs, as the visual representation becomes more concrete ($\mathcal{F}_V$) while the injected information ($\mathcal{F}_T$) does not contribute towards generating a text response. From layers 19 to 22 in LLaVA and 22 to 25 in Qwen, both $\mathcal{F}_V$ increase and $\mathcal{F}_T$ increase or remain stable, indicating information injection or multimodal mixture by the MLLMs, which alters the visual representation while incorporating more contextual information for output generation. Finally, the decrease in $\mathcal{F}_V$ and the increase in $\mathcal{F}_T$ suggest information compression, solidifying the visual information towards the textual output.

Finding 3. In QA tasks, MLLMs show an extended Information Pre-processing Stage compared to captioning tasks. We observe two less evident or delayed peaks compared to captioning tasks: around layers 18 and 29 in LLaVA and layer 16 and 25 in Qwen, as shown in Figure 2b, 3a. The delayed decrease in $\mathcal{F}_V$, especially for LLaVA, suggests an extended Information Pre-processing Stage, indicating an extension of information processing to generate a specific response.

Finding 4. LLaVA exhibit greater reliance on injected contextual information in QA tasks than more advanced Qwen. In QA tasks, the overall increasing trend of LLaVA's $\mathcal{F}_V$, as illustrated in Figure 2b, suggests that the visual representation increasingly diverges from its original state. This elevated $\mathcal{F}_V$, coupled with reduced and delayed information compression, indicates a more extensive injection of contextual information and common sense from the MLLM, as well as a corresponding decrease in response certainty. In contrast, Qwen maintains a more stable four-stage trend, as depicted in Figure 3a, highlighting its superior effectiveness in information injection and greater certainty in responses.

Finding 5. Complex visual and inconsistent textual inputs influence MLLMs' efficiencies of information processing in certain stages. HAL-Eval and AOKVQA exhibit higher visual complexity than COCO-caption and VQA, and they provide imprecise answer candidates. We differentiate between imprecise and precise answer candidates to compare the differences in $\mathcal{F}_V$ and $\mathcal{F}_T$ influenced by the textual precision. We observe that the $\mathcal{F}_V$ valley around layer 18 in LLaVA and 21 in Qwen is less pronounced, as illustrated in Figures 2b and 3a. This finding suggests that both information filtering and injection are less intensive when MLLMs overly depend on imprecise answer texts while neglecting the visual representation. In the case of AOKVQA, both LLaVA and Qwen exhibit lower $\mathcal{F}_V$ prior to the first peak, specifically before layer 12 in LLaVA and 15 in Qwen, as shown in Figures 2b and 3a. Moreover, the $\mathcal{F}_T$ values before layer 12 in LLaVA and layer 15 in Qwen are higher than those observed in VQA, as shown in Figures 2d and 3b. This indicates an early task-related contextual information injection during the information pre-processing stage influenced by the complexities of visual and inconsistent textual inputs.

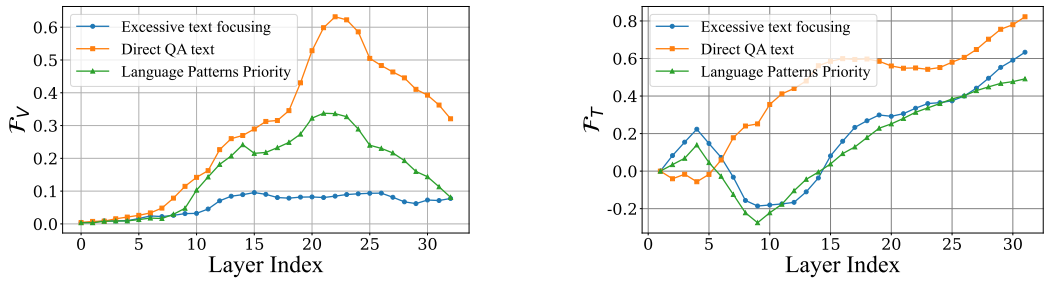
4.2 (RQ2) How do disruptive textual instructions influence multimodal information?

We investigate the hallucinations and errors in MLLMs (Li et al., 2023a), which happen mostly when model focuses excessively on the current segment of generated content while ignoring input visual information (Wang et al., 2024a; Lee et al., 2024) and when model prioritizing language patterns to produce fluent yet inaccurate content in visual-text tasks (Wang et al., 2023). Based on this, we construct several prompts to trigger these two kinds of errors in MLLMs, which is the "Excessive Text Focusing" and "Language Patterns Priority" in Figure 5. We refer to these prompts that trigger errors as "**disruptive prompts**". By comparing multimodal representation changes across layers using original POPE (Li et al., 2023b) queries and these disruptive prompts with LLaVA, we aim to quantify their effects on visual information processing and textual guidance within the models.

Finding 6. Disruptive prompts impairs visual information processing. With disruptive prompts, the visual representation peaks of QA tasks in $\mathcal{F}_V$ become less significant, as shown in Figure 2b, indicating a diminished visual information processing and reduced contextual information injection. The peak of $\mathcal{F}_V$ around layer 22 in Figure 2b diminishes in the Language Patterns Priority prompts pattern, suggesting less contextual information injection by MLLMs. Furthermore, when prompts that trigger excessive focus on text are added, there is almost no change in $\mathcal{F}_V$, indicating a neglect of the visual representation.

Finding 7. Disruptive prompts impairs contextual information injection of MLLMs. Based on the trend of $\mathcal{F}_T$ in Figure 5b, we observe that the difference in $\mathcal{F}_T$ between layer 0 and the final layer is smaller when disruptive prompts are applied, indicating that less contextual information aligns with the final response are injected by the MLLMs. Additionally, we note a fluctuation in $\mathcal{F}_T$ before layer 12. This fluctuation of $\mathcal{F}_T$ prior to layer 12, combined with the reduced difference in $\mathcal{F}_T$ and a stable $\mathcal{F}_V$ in Figure 5a across

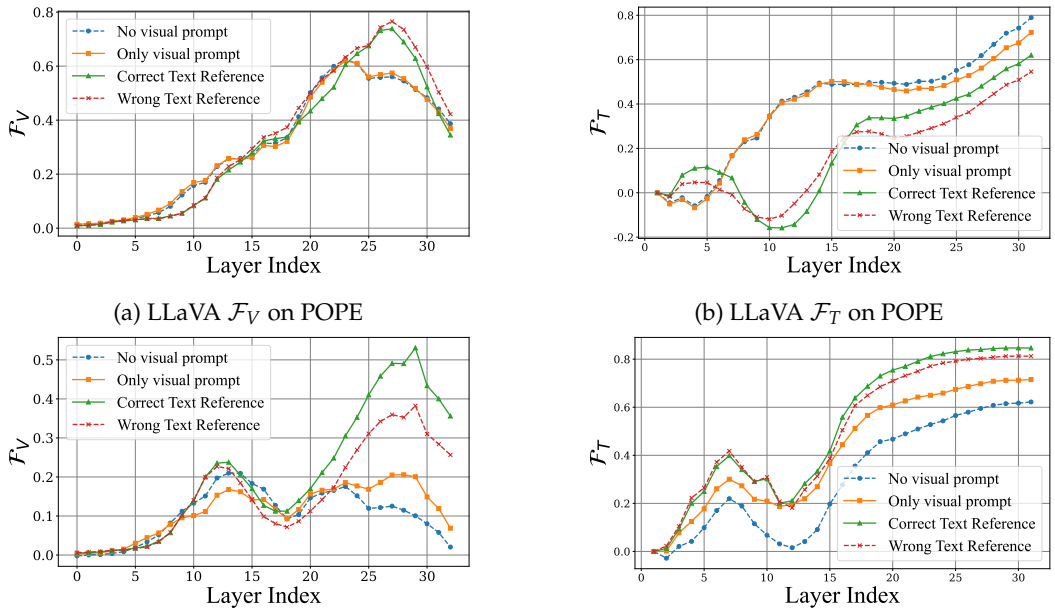
layers, indicates that MLLMs are overly focused on the disruptive textual instructions while neglecting the provided image and visual-to-text question.



(a) (b)

Figure 5: Impacts of disruptive prompts on LLaVA $\mathcal{F}_V$ and $\mathcal{F}_T$ across different layers.

4.3 (RQ3) How do visual prompts interact with textual instructions in MLLMs?



(a) LLaVA $\mathcal{F}_V$ on POPE (b) LLaVA $\mathcal{F}_T$ on POPE

(c) LLaVA $\mathcal{F}_V$ on HAL-Eval (d) LLaVA $\mathcal{F}_T$ on HAL-Eval

Figure 6: Impacts of visual prompt and textual instruction combination on $\mathcal{F}_V$ and $\mathcal{F}_T$.

We investigate the influence of explicit visual prompting (e.g., bounding boxes and digit annotations on detectable objects) on multimodal representations within MLLMs (Zhang et al., 2023a; Yan et al., 2025; Lin et al., 2024; Wu et al., 2024e). We compare four settings: original POPE queries without visual prompted images (baseline), original POPE queries with visual prompted images, textual queries correctly referencing visual prompts, and textual queries incorrectly referencing visual prompts, which are denoted as "No visual prompt", "Only visual prompt" (means having visual prompt without text reference), "Correct Text Reference" and "Wrong Text Reference" in Figure 6 respectively. We aim to quantify the distinct and combined impacts of visual and textual instructions on visual information processing across different layers in MLLMs.

Finding 8. The impact of textual instructions is more significant than that of visual prompts on multimodal information processing. Results in Figure 6 shows that textual prompting alters the trends of $\mathcal{F}_V$ and $\mathcal{F}_T$ across different layers, whereas adding visual prompts does not. Additionally, the differences induced by changing textual prompting on $\mathcal{F}_V$ and $\mathcal{F}_T$ are more significant than those results from changing visual prompting.

Simply adding visual prompts without referencing them through textual instructions does not yield a noticeable impact on visual representation processing. This difference is even less significant in POPE compared to HAL-Eval. This phenomenon suggests that textual instructions dominate visual information processing during inference, indicating that MLLMs are considerably more sensitive to language prompts than to visual prompts.

Finding 9. Textual references to visual prompts significantly enhance contextual information injection from MLLMs. As show in Figure 6a and 6c, before layer 22, regardless of whether we add visual prompts or refer to these visual prompts through textual instructions, $\mathcal{F}_V$ shows similar trends across layers. This indicates that the impact of visual prompting manifests in the later stages of MLLMs inference. Considering $\mathcal{F}_V$ after layer 22 where we previously denote as information injection and compression, when visual prompts are referred by textual instructions, there is a rapid increase in $\mathcal{F}_V$. This indicates a more significant contextual information injection from MLLMs happens when visual prompts are referred by textual reference.

4.4 (RQ4) How do key words in instructions influence MLLM knowledge injection?

To investigate the influence of different instructions on representations in MLLMs (Yan et al., 2024; Kim et al., 2024b; Xia et al., 2024; Qin et al., 2024), we categorized queries from the VQA-v2 dataset into distinct types based on their keyword prompts (e.g., "Is," "What," "Can," "Why," "Where"). By comparing visual representation transformations across these question categories based on $\mathcal{F}_V$ and $\mathcal{F}_T$ in Figure 7, we evaluated how specific linguistic structures and query intents influence visual information processing within MLLMs.

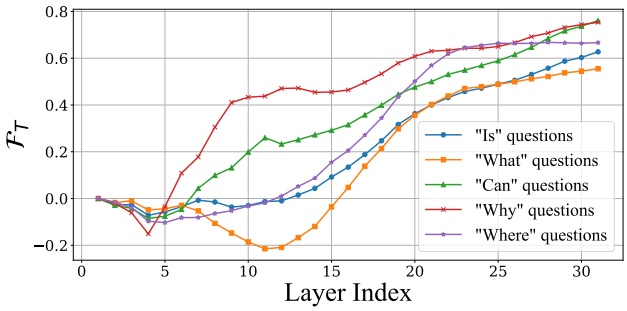


Figure 7: Impacts of instruction types on LLaVA $\mathcal{F}_T$.

Finding 10: Our analysis shows that different types of textual instructions result in distinct patterns of contextual information injection in MLLMs. The processing behavior of visual representations by MLLMs is influenced by keywords in instructions that determine the type of instructions. We select several representative instruction types from those lists in VQA-v2 dataset and use LLMs to infer responses based on these different types of instructions. For "Why" and "Can" questions, MLLMs inject more contextual information aligned with the response, as indicated by $\mathcal{F}_T$. Conversely, MLLMs inject less contextual information when answering "Where" questions compared to "Why" and "Can" questions, but more contextual information than when responding to "What" and "Is" questions.

5 Conclusion

In this work, we introduced an information-theoretic framework to systematically analyze and quantify the interactions between visual and textual modalities in MLLMs. Our metrics provided tractable and interpretable measures for visual information gain and textual contextual contributions across different model layers. Our empirical findings revealed four distinct stages of multimodal representation processing, highlighting the dynamic interplay between visual and textual instructions. We demonstrate that textual instructions dominate visual information processing, with disruptive prompts significantly impairing both visual encoding and contextual information injection. Furthermore, our analysis underscored the critical synergy between visual and textual instructions, showing that effective multimodal understanding relies heavily on their alignment. These insights not only advance our understanding of MLLMs' internal mechanisms but also offer practical guidance for designing more robust and explainable multimodal systems.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 2425–2433. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.279. URL https://doi.org/10.1109/ICCV.2015.279.

Rabiul Awal, Le Zhang, and Aishwarya Agrawal. Investigating prompting techniques for zero- and few-shot visual question answering, 2025. URL https://arxiv.org/abs/2306.09996.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966.

Tong Bao, Che Liu, Derong Xu, Zhi Zheng, and Tong Xu. MLLM-I2W: Harnessing multimodal large language model for zero-shot composed image retrieval. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 1839–1849, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.125/.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015. URL http://arxiv.org/abs/1504.00325.

Joseph F. DeRose, Jiayao Wang, and Matthew Berger. Attention flows: Analyzing and comparing attention mechanisms in language models. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1160–1170, 2021. doi: 10.1109/TVCG.2020.3028976.

Marco Federici, Ryota Tomioka, and Patrick Forré. An information-theoretic approach to distribution shifts, 2021. URL https://arxiv.org/abs/2106.03783.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 6325–6334. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.670. URL https://doi.org/10.1109/CVPR.2017.670.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A. Smith, and Jiebo Luo. Promptcap: Prompt-guided image captioning for VQA with GPT-3. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 2951–2963. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00277. URL https://doi.org/10.1109/ICCV51070.2023.00277.

Chengkai Huang, Junda Wu, Yu Xia, Zixu Yu, Ruhan Wang, Tong Yu, Ruiyi Zhang, Ryan A Rossi, Branislav Kveton, Dongruo Zhou, et al. Towards agentic recommender systems in the era of multimodal large language models. *arXiv preprint arXiv:2503.16734*, 2025.

Theo Jaunet, Corentin Kervadec, Romain Vuillemot, Grigory Antipov, Moez Baccouche, and Christian Wolf. Visqa: X-raying vision and language reasoning in transformers. *IEEE Trans. Vis. Comput. Graph.*, 28(1):976–986, 2022. doi: 10.1109/TVCG.2021.3114683. URL https://doi.org/10.1109/TVCG.2021.3114683.

Chaoya Jiang, Hongrui Jia, Mengfan Dong, Wei Ye, Haiyang Xu, Ming Yan, Ji Zhang, and Shikun Zhang. Hal-eval: A universal and fine-grained hallucination evaluation framework for large vision language models. In Jianfei Cai, Mohan S. Kankanhalli, Balakrishnan Prabhakaran, Susanne Boll, Ramanathan Subramanian, Liang Zheng, Vivek K. Singh, Pablo César, Lexing Xie, and Dong Xu (eds.), *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pp. 525–534. ACM, 2024. doi: 10.1145/3664647.3680576. URL https://doi.org/10.1145/3664647.3680576.

Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements, 2024. URL https://arxiv.org/abs/2410.17725.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Hazel Kim, Adel Bibi, Philip Torr, and Yarin Gal. Detecting llm hallucination through layer-wise information deficiency: Analysis of unanswerable questions and ambiguous prompts, 2024a. URL https://arxiv.org/abs/2412.10246.

Jihoo Kim, Wonho Song, Dahyun Kim, Yunsu Kim, Yungi Kim, and Chanjun Park. Evalverse: Unified and accessible library for large language model evaluation, 2024b. URL https://arxiv.org/abs/2404.00943.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 2020a. URL http://proceedings.mlr.press/v119/koh20a.html.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. PMLR, 2020b.

Songning Lai, Lijie Hu, Junxiao Wang, Laure Berti-Équille, and Di Wang. Faithful vision-language interpretation via concept bottleneck models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=rp0EdI8X4e.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. Volcano: Mitigating multimodal hallucination through self-feedback guided revision. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pp. 391–404. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.23. URL https://doi.org/10.18653/v1/2024.naacl-long.23.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Yongqi Li, Wenjie Wang, Leigang Qu, Liqiang Nie, Wenjie Li, and Tat-Seng Chua. Generative cross-modal retrieval: Memorizing images in multimodal language models for retrieval and beyond. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11851–11861, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.639. URL https://aclanthology.org/2024.acl-long.639/.

Yuanze Lin, Yunsheng Li, Dongdong Chen, Weijian Xu, Ronald Clark, Philip Torr, and Lu Yuan. Rethinking visual prompting for multimodal large language models with external knowledge, 2024. URL https://arxiv.org/abs/2407.04681.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models, 2024a. URL https://arxiv.org/abs/2402.00253.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 26286–26296. IEEE, 2024b. doi: 10.1109/CVPR52733.2024.02484. URL https://doi.org/10.1109/CVPR52733.2024.02484.

Zhihang Liu, Chen-Wei Xie, Bin Wen, Feiwu Yu, Jixuan Chen, Boqiang Zhang, Nianzu Yang, Pandeng Li, Yun Zheng, and Hongtao Xie. What is a good caption? A comprehensive visual caption benchmark for evaluating both correctness and coverage of mllms. *CoRR*, abs/2502.14914, 2025. doi: 10.48550/ARXIV.2502.14914. URL https://doi.org/10.48550/arXiv.2502.14914.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 4765–4774, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Haoyan Luo and Lucia Specia. From understanding to utilization: A survey on explainability for large language models, 2024. URL https://arxiv.org/abs/2401.12874.

Yida Mu, Ben P. Wu, William Thorne, Ambrose Robinson, Nikolaos Aletras, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. Navigating prompt complexity for zero-shot classification: A study of large language models in computational social science. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 12074–12086, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1055/.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models, 2024. URL https://arxiv.org/abs/2410.07149.

Dang Nguyen, Jian Chen, Yu Wang, Gang Wu, Namyong Park, Zhengmian Hu, Hanjia Lyu, Junda Wu, Ryan Aponte, Yu Xia, et al. Gui agents: A survey. *arXiv preprint arXiv:2412.13501*, 2024.

Thao Nguyen, Yuheng Li, Utkarsh Ojha, and Yong Jae Lee. Visual instruction inversion: Image editing via image prompting. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/1e75f7539cbde5de895fab238ff42519-Abstract-Conference.html.

Peisong Niu, Tian Zhou, Xue Wang, Liang Sun, and Rong Jin. Understanding the role of textual prompts in llm for time series forecasting: an adapter view, 2024. URL https://arxiv.org/abs/2311.14782.

Jianing Qi, Jiawei Liu, Hao Tang, and Zhigang Zhu. Beyond semantics: Rediscovering spatial awareness in vision-language models, 2025. URL https://arxiv.org/abs/2503.17349.

Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. Infobench: Evaluating instruction following ability in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pp. 13025–13048. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.772. URL https://doi.org/10.18653/v1/2024.findings-acl.772.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In John DeNero, Mark Finlayson, and Sravana Reddy (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pp. 97–101, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-3020. URL https://aclanthology.org/N16-3020/.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VIII*, volume 13668 of *Lecture Notes in Computer Science*, pp. 146–162. Springer, 2022. doi: 10.1007/978-3-031-20074-8_9. URL https://doi.org/10.1007/978-3-031-20074-8_9.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000a.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method, 2000b. URL https://arxiv.org/abs/physics/0004057.

Jean-Francois Ton, Muhammad Faaiz Taufiq, and Yang Liu. Understanding chain-of-thought in llms through information theory, 2024. URL https://arxiv.org/abs/2411.11984.

Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, and Conghui He. VIGC: visual instruction generation and correction. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pp. 5309–5317. AAAI Press, 2024a. doi: 10.1609/AAAI.V38I6.28338. URL https://doi.org/10.1609/aaai.v38i6.28338.

Jianing Wang, Junda Wu, Yupeng Hou, Yao Liu, Ming Gao, and Julian McAuley. Instructgraph: Boosting large language models via graph-centric instruction tuning and preference alignment. *arXiv preprint arXiv:2402.08785*, 2024b.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, Jitao Sang, and Haoyu Tang. Evaluation and analysis of hallucination in large vision-language models, 2023. URL https://arxiv.org/abs/2308.15126.

Luran Wang, Mark Gales, and Vatsal Raina. An information-theoretic approach to analyze nlp classification tasks, 2024c. URL https://arxiv.org/abs/2402.00978.

Ruoyu Wang, Tong Yu, Junda Wu, Yao Liu, Julian McAuley, and Lina Yao. Weakly-supervised vlm-guided partial contrastive learning for visual language navigation. *arXiv preprint arXiv:2506.15757*, 2025.

Junda Wu, Tong Yu, and Shuai Li. Deconfounded and explainable interactive vision-language retrieval of complex scenes. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 2103–2111, 2021.

Junda Wu, Rui Wang, Tong Yu, Ruiyi Zhang, Handong Zhao, Shuai Li, Ricardo Henao, and Ani Nenkova. Context-aware information-theoretic causal de-biasing for interactive sequence labeling. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3436–3448, 2022.

Junda Wu, Tong Yu, Rui Wang, Zhao Song, Ruiyi Zhang, Handong Zhao, Chaochao Lu, Shuai Li, and Ricardo Henao. Infoprompt: Information-theoretic soft prompt tuning for natural language understanding. *Advances in neural information processing systems*, 36: 61060–61084, 2023.

Junda Wu, Cheng-Chun Chang, Tong Yu, Zhankui He, Jianing Wang, Yupeng Hou, and Julian McAuley. Coral: collaborative retrieval-augmented large language models improve long-tail recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3391–3401, 2024a.

Junda Wu, Xintong Li, Tong Yu, Yu Wang, Xiang Chen, Jiuxiang Gu, Lina Yao, Jingbo Shang, and Julian McAuley. Commit: Coordinated instruction tuning for multimodal large language models. *arXiv preprint arXiv:2407.20454*, 2024b.

Junda Wu, Hanjia Lyu, Yu Xia, Zhehao Zhang, Joe Barrow, Ishita Kumar, Mehrnoosh Mirtaheri, Hongjie Chen, Ryan A Rossi, Franck Dernoncourt, et al. Personalized multimodal large language models: A survey. *arXiv preprint arXiv:2412.02142*, 2024c.

Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A Rossi, Ruiyi Zhang, et al. Visual prompting in multimodal large language models: A survey. *arXiv preprint arXiv:2409.15310*, 2024d.

Junda Wu, Warren Li, Zachary Novack, Amit Namburi, Carol Chen, and Julian McAuley. Collap: Contrastive long-form language-audio pretraining with musical temporal structure augmentation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025a.

Junda Wu, Zachary Novack, Amit Namburi, Hao-Wen Dong, Carol Chen, Jiaheng Dai, and Julian McAuley. Futga-mir: Enhancing fine-grained and temporally-aware music understanding with music information retrieval. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025b.

Junda Wu, Yu Xia, Tong Yu, Xiang Chen, Sai Sree Harsha, Akash V Maharaj, Ruiyi Zhang, Victor Bursztyn, Sungchul Kim, Ryan A Rossi, et al. Doc-react: Multi-page heterogeneous document question-answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 67–78, 2025c.

Junda Wu, Yuxin Xiong, Xintong Li, Yu Xia, Ruoyu Wang, Yu Wang, Tong Yu, Sungchul Kim, Ryan A Rossi, Lina Yao, et al. Mitigating visual knowledge forgetting in mllm instruction-tuning via modality-decoupled gradient descent. *arXiv preprint arXiv:2502.11740*, 2025d.

Yixuan Wu, Yizhou Wang, Shixiang Tang, Wenhao Wu, Tong He, Wanli Ouyang, Philip Torr, and Jian Wu. Dettoolchain: A new prompting paradigm to unleash detection ability of MLLM. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXXII*, volume 15090 of *Lecture Notes in Computer Science*, pp. 164–182. Springer, 2024e. doi: 10.1007/978-3-031-73411-3_10. URL https://doi.org/10.1007/978-3-031-73411-3_10.

Congying Xia, Chen Xing, Jiangshu Du, Xinyi Yang, Yihao Feng, Ran Xu, Wenpeng Yin, and Caiming Xiong. FOFO: A benchmark to evaluate llms' format-following capability. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 680–699. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.40. URL https://doi.org/10.18653/v1/2024.acl-long.40.

Shin'ya Yamaguchi, Kosuke Nishida, Daiki Chijiwa, and Yasutoshi Ida. Zero-shot concept bottleneck models, 2025. URL https://arxiv.org/abs/2502.09018.

An Yan, Zhengyuan Yang, Junda Wu, Wanrong Zhu, Jianwei Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Julian McAuley, Jianfeng Gao, and Lijuan Wang. List items one by one: A new data source and learning paradigm for multimodal llms, 2025. URL https://arxiv.org/abs/2404.16375.

Jianhao Yan, Yun Luo, and Yue Zhang. Refutebench: Evaluating refuting instruction-following for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pp. 13775–13791. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.818. URL https://doi.org/10.18653/v1/2024.findings-acl.818.

Zhao Yang, Yuanzhe Zhang, Pengfei Cao, Cao Liu, Jiansong Chen, Jun Zhao, and Kang Liu. Information bottleneck based knowledge selection for commonsense reasoning. *Information Sciences*, 660:120134, 2024. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2024.120134. URL https://www.sciencedirect.com/science/article/pii/S0020025524000471.

Catherine Yeh, Yida Chen, Aoyu Wu, Cynthia Chen, Fernanda B. Viégas, and Martin Wattenberg. Attentionviz: A global view of transformer attention. *IEEE Trans. Vis. Comput. Graph.*, 30(1):262–272, 2024. doi: 10.1109/TVCG.2023.3327163. URL https://doi.org/10.1109/TVCG.2023.3327163.

Zeping Yu and Sophia Ananiadou. Understanding multimodal llms: the mechanistic interpretability of llava in visual question answering, 2025. URL https://arxiv.org/abs/2411.10950.

Ao Zhang, Hao Fei, Yuan Yao, Wei Ji, Li Li, Zhiyuan Liu, and Tat-Seng Chua. Vpg-trans: Transfer visual prompt generator across llms. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023a. URL http://papers.nips.cc/paper_files/paper/2023/hash/407106f4b56040b2e8dcad75a6e461e5-Abstract-Conference.html.

Jinghan Zhang, Xiting Wang, Fengran Mo, Yeyang Zhou, Wanfu Gao, and Kunpeng Liu. Entropy-based exploration conduction for multi-step reasoning, 2025. URL https://arxiv.org/abs/2503.15848.

Xiaofeng Zhang, Yihao Quan, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. From redundancy to relevance: Information flow in lvlms across reasoning tasks, 2024a. URL https://arxiv.org/abs/2406.06579.

Xin Zhang, Wen Xie, Ziqi Dai, Jun Rao, Haokun Wen, Xuan Luo, Meishan Zhang, and Min Zhang. Finetuning language models for multimodal question answering. In Abdulmotaleb El-Saddik, Tao Mei, Rita Cucchiara, Marco Bertini, Diana Patricia Tobon Vallejo, Pradeep K. Atrey, and M. Shamim Hossain (eds.), *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023-3 November 2023*, pp. 9420–9424. ACM, 2023b. doi: 10.1145/3581783.3612837. URL https://doi.org/10.1145/3581783.3612837.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models, 2024b. URL https://arxiv.org/abs/2411.18620.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.*, 15(2):20:1–20:38, 2024. doi: 10.1145/3639372. URL https://doi.org/10.1145/3639372.

Zhiyu Zhu, Zhibo Jin, Jiayu Zhang, Nan Yang, Jiahao Huang, Jianlong Zhou, and Fang Chen. Narrowing information bottleneck theory for multimodal image-text representations interpretability, 2025. URL https://arxiv.org/abs/2502.14889.

Yong Zhuang, Tong Yu, Junda Wu, Shiqu Wu, and Shuai Li. Spatial-temporal aligned multi-agent learning for visual dialog systems. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 482–490, 2022.

A Implementation Details

A.1 Extracted Visual Information

We employ a three-layer Concept Encoder (CE) to encode features extracted from the l-th layer of LLMs, denoted as $X_{I_t}^{(l)}$, into concept bottleneck information $\hat{C}_l$. This CE is trained on hidden state information of LLaVA on the MSCOCO-2014 dataset, which represent $q(\hat{C} \mid V)$. The concept bottleneck information $\hat{C}_l$ in this context is represented as a 80-dimensional vector, corresponding to 80 object classes detectable in MSCOCO-2014 using YOLOv11. The vector is normalized, with each dimension reflecting the importance score of the respective object. On each dataset YOLOv11 is firstly applied to generate the object distribution prior, denoted as $r(\hat{C})$. To implement margin on text, we use TF-IDF to vectorize and classify the texts input of MLLMs into several clusters and margin the expectation of KL on each cluster.

A.2 Induced Contextual Information from MLLMs

We trained two MLPs for vector similarity evaluation, denoted as $s(.,.)$. When multiple hidden state representations are extracted from same image-text samples, the similarity between the them should be high. In our context, we assess information similarity in MLLMs at each decode layers with the last decode layer. There are two settings for $\mathcal{F}_T$ evaluation, task-aware similarity and baseline similarity, where task-aware similarity is evaluated with MLLMs given instruction texts and baseline similarity is evaluated with MLLMs only given system tokens and image tokens.

B Approximation of InfoNCE

To robustly track the similarity function $s(\cdot, \cdot)$ in high-dimensional representations, we first apply a low-rank projection that maps the original representations into a joint embedding space. Specifically, for any representation X (*i.e.*, $X_T^{(l),i}$, $X_T^{(L),i}$, $X_0^{(l),i}$, or $X_0^{(L),i}$), we define a projection function:

$$f_{\text{proj}}(X) = W^\top X + b,$$

where $W \in \mathbb{R}^{d \times d_j}$ is a learnable projection matrix with $d_j \ll d$ and $b \in \mathbb{R}^{d_j}$ is a bias term. The low-rank projection serves to distill the most salient features from the original representations and align them in a common, lower-dimensional space. The similarity function is then computed as the inner product in this joint space:

$$s(x, y) = \langle f_{\text{proj}}(x), f_{\text{proj}}(y) \rangle = f_{\text{proj}}(x)^\top f_{\text{proj}}(y),$$

which reduces computational complexity and enforces a structured joint distribution, ensuring the computed similarities accurately reflect the alignment between the representations.