Make LVLMs Focus: Context-Aware Attention Modulation for Better Multimodal In-Context Learning

Yanshu Li[1], Jianjiang Yang[2], Ziteng Yang[1], Bozheng Li[1], Ligong Han[3], Hongyang He[4], Zhengtao Yao[5], Yingjie Victor Chen[6], Songlin Fei[6], Dongfang Liu[7], Ruixiang Tang[8]*

[1]Brown University
[2]University of Bristol
[3]MIT-IBM Watson AI Lab
[4]University of Warwick
[5]University of Southern California
[6]Purdue University
[7]Rochester Institute of Technology
[8]Rutgers University
yanshu_li1@brown.edu, ruixiang.tang@rutgers.edu

Abstract

Multimodal in-context learning (ICL) is becoming a key capability that allows large vision-language models (LVLMs) to adapt to novel tasks without parameter updates, which expands their usefulness in many real-world applications. However, ICL performance remains unstable even when the in-context demonstrations (ICDs) are well matched, showing that LVLMs still struggle to make full use of the provided context. While existing work mainly focuses on prompt engineering or post-hoc logit calibration, we study the attention mechanisms inside LVLMs to address their inherent limitations. We identify two important weaknesses in their self-attention that hinder effective ICL. To address these weaknesses, we propose **Context-Aware Modulated Attention** (CAMA), a training-free and plug-and-play method that dynamically adjusts attention logits based on the input in-context sequence. CAMA uses a two-stage modulation process that strengthens attention to semantically important tokens, especially visual ones. Across four LVLMs and seven benchmarks, CAMA consistently outperforms vanilla models and baselines, showing clear effectiveness and generalization. It can also activate the intended benefits of prompt engineering methods and remains robust across different sequence configurations. Therefore, CAMA opens up new directions for improving multimodal reasoning through a deeper understanding of attention dynamics.

Extended version — https://arxiv.org/abs/2505.17097

Introduction

Large vision-language models (LVLMs) have emerged as powerful tools for multimodal information processing and generation (Zhao et al. 2023). Through large-scale pretraining, they integrate visual and textual signals into the shared representation space of large language models (LLMs) and have achieved notable success across vision-language tasks (Ye et al. 2023; Liu et al. 2023). However, adapting LVLMs to new domains remains challenging due to the high costs of multimodal data preparation and training.

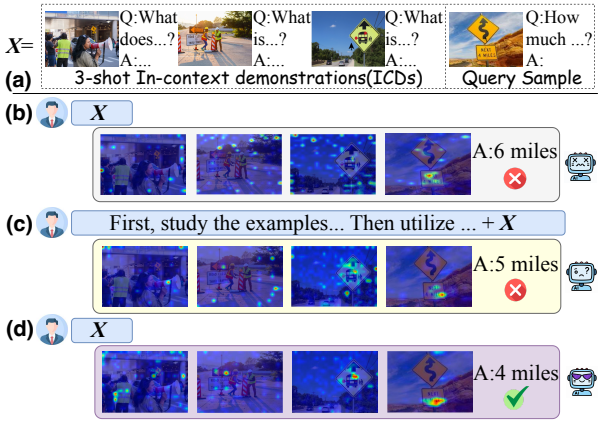

Figure 1: (a) Example of a 3-shot multimodal in-context sequence. (b)-(d) present the vanilla model, adding an instruction to the sequence, and our proposed method, **CAMA**, respectively. All attention heatmaps come from layer 18, and redder regions indicate stronger attention.

To reduce these costs, researchers are applying in-context learning (ICL), a technique widely used in LLMs (Brown et al. 2020; Dong et al. 2024), to LVLMs (Li 2025). In ICL, a few in-context demonstrations (ICDs) are added to the input as reference examples. This allows the model to adapt to new tasks by interpreting the demonstrations without parameter updates. Recent progress in model architectures and training protocols enables LVLMs to process multiple images and perform interleaved reasoning, which makes multimodal ICL practical (Alayrac et al. 2022; Laurençon et al. 2024a; Bai et al. 2025; Chen et al. 2024c). These advances expand the practical use of LVLMs across a wide range of applications (Doveh et al. 2024; Chen et al. 2025; Guo et al. 2024).

However, the benefits of multimodal ICL can be diminished by its pronounced instability. Recent studies have found that LVLMs' ICL is highly sensitive to subtle prompt details.

*Corresponding Author.

Minor changes in the order or formatting of ICDs can cause significant performance swings (Liu et al. 2022; Gao, Fisch, and Chen 2021; Lu et al. 2022). This sensitivity becomes stronger when the number of ICDs increases or when the task requires complex multimodal reasoning (Chen et al. 2023; Li et al. 2025b). To improve multimodal ICL in LVLMs, two main routes have been explored. The first optimizes the prompt. It adds guiding text, highlights key image regions, or selects and orders ICDs with predefined metrics (Li et al. 2024b; Yang et al. 2023). Nonetheless, prompt engineering needs substantial prior knowledge, shows limited stability, and is heavily dependent on the base model. The second route edits the internal logits of the models, as in contrastive decoding (Lee, Tsai, and Chiu 2024). This approach alters reasoning more directly but needs distorted inputs and extra forward passes for calibration (Liu, Zheng, and Chen 2024). Since practical ICL prioritizes efficiency, our work investigates the following questions: *What is the intrinsic limitation underlying the instability of multimodal ICL in LVLMs? Can we address it using a more efficient, training-free method?*

In pursuit of these questions, we examine the attention dynamics of LVLMs during multimodal ICL. Because their inputs interleave images and texts, we move beyond the existing focus on single-image scenarios and analyze LVLMs from two distinct perspectives: (1) text-based visual grounding within each image-text pair, and (2) query-sample-driven attention allocation across ICDs. We visualize attention deficits in both aspects through targeted experiments and verify that these deficits lead to insufficient utilization of the in-context sequence by the LVLM, resulting in unstable multimodal ICL. In response to these deficits, we present Context-Aware Modulated Attention (CAMA), a training-free method that dynamically modulates the model's internal attention logits during inference based on the input context. CAMA addresses the two deficits with a two-stage modulation that targets the shallow and middle layers of the decoder. Stage I applies **intra-ICD grounding**, highlighting the image tokens in each ICD and in the query sample that best align with their accompanying text, which reduces the attention sink introduced by the interleaved image-text format. Stage II performs **query-centric routing**, reallocating attention among ICDs in proportion to their value to produce the desired answer, thus improving the use of context. Experiments demonstrate that CAMA delivers consistent performance gains for multimodal ICL, providing insights for reshaping attention mechanisms to build models with better visual capabilities.

The contributions of this paper are summarized below:

- We analyze LVLMs' attention dynamics in multimodal ICL beyond single-image settings and reveal two deficits: weak vision-text alignment within each pair and misalignment between the query sample and its ICDs.
- Building on the identified deficits, we introduce CAMA, the first training-free and model-agnostic method designed to enhance multimodal ICL. CAMA applies a two-stage modulation of internal attention logits that steers the model toward the tokens most relevant to ICL.
- Extensive experiments demonstrate that CAMA improves multimodal ICL across diverse LVLMs and benchmarks. Ablation studies confirm the necessity of each design and reveal CAMA's broader potential.

Related Work

In-context Learning (ICL). ICL enables models to solve unseen tasks by conditioning on an input sequence of input–output examples (i.e., ICDs) without updating any parameters (Brown et al. 2020; Dong et al. 2024). This capability markedly improves their practicality and is especially valued in resource-intensive multimodal domains (Wies, Levine, and Shashua 2023). Large vision–language models (LVLMs) gain ICL capabilities through targeted pretraining or fine-tuning on interleaved image–text data, as in Flamingo (Alayrac et al. 2022) and LLaVA-NeXT (Li et al. 2024a). Now, it has become a core ability of commercial model families such as QwenVL (Bai et al. 2025) and InternVL (Chen et al. 2024c). Consequently, recent work has begun to explore the mechanisms of multimodal ICL (Doveh et al. 2024; Chen et al. 2025). However, few studies probe internal attention. Therefore, cross-modal interactions in multimodal ICL remain only partially understood, which prevents state-of-the-art (SOTA) models from fully exploiting ICL's potential.

Enhancing Multimodal ICL. Limited use of input-sequence information by LVLMs is considered a major source of instability in multimodal ICL (Liu et al. 2022; Gao, Fisch, and Chen 2021; Li et al. 2023). Three solution paths have thus emerged. First, task-oriented datasets combined with instruction tuning (Jiang et al. 2024; Chen et al. 2024b) or direct preference optimization (DPO) (Jia et al. 2025) train LVLMs to reason across multiple ICDs. Second, prompt-level methods optimize ICD selection with metrics such as similarity scores and information entropy (Li et al. 2024b; Yang et al. 2023; Zhou et al. 2024; Wu et al. 2022), or by training automatic selectors (Li et al. 2025c; Yang et al. 2024). Third, calibration-based methods adjust LVLMs' final logits using contrastive decoding (Kim et al. 2024; Fazli, Wei, and Zhu 2025; Li et al. 2025a). While these approaches improve performance, they all face challenges. The first requires extensive curated data and parameter updates, the second is less adaptable to fixed-prompt scenarios, and both depend on the base model. The third needs to design distorted input sequences, plus multiple forward passes. In contrast, our CAMA sidesteps all these drawbacks by modulating attention logits at inference time without additional data or tuning.

Attention Dynamics in Multimodal ICL

Background and Notation

Current-generation LVLMs generally consist of three core components: a vision encoder that processes images, a projector that converts visual features into embeddings, and an autoregressive LLM that decodes both image and text embeddings to produce output. In multimodal ICL, LVLM typically takes an interleaved image-text sequence as input, as illustrated in Figure 1(a). This work focuses mainly on visual question-answering (VQA), which emphasizes both visual perception and language-based reasoning.

We consider an LVLM $\mathcal{M}$, which generates the answer y given an n-shot sequence X as input. X consists of n in-context demonstrations (ICDs) and a single query sample:

$$y \leftarrow \mathcal{M}(X),$$
$$X = (X_1^I, X_1^T, \ldots, X_n^I, X_n^T, X_{n+1}^I, X_{n+1}^T), \tag{1}$$

where the query sample is indexed as $n+1$ for unified notation. $X_i^I \in \mathbb{R}^{S_i^I \times D}$ denotes the token sequence of the i-th image, $X_i^T \in \mathbb{R}^{S_i^T \times D}$ denotes the token sequence of the i-th text segment and D is the dimensionality of hidden states. S_i^I and S_i^T are the token counts of the i-th image and text segment, respectively. $S = \sum_{i=1}^{n+1}(S_i^I + S_i^T)$ is the total length of the token sequence and $\mathbf{S}$ is its index set. Within each ICD, the text tokens can be further divided into a question part and an answer part, whose counts are S_i^Q and S_i^A. $\mathbf{S}_i^I, \mathbf{S}_i^Q$ and $\mathbf{S}_i^A$ denote the token index sets for the image, question, and answer of the i-th ICD, respectively.

The sequence X is then passed to the LLM of $\mathcal{M}$, which performs an N-layer decoder forward pass. Each layer contains a multi-head attention (MHA) module. The h-th head in the l-th layer maps the hidden states of X to queries $Q^{l,h} \in \mathbb{R}^{S \times D_k}$, keys $K^{l,h} \in \mathbb{R}^{S \times D_k}$, and values $V^{l,h} \in \mathbb{R}^{S \times D_k}$ by linear transformations, where D_k is the head dimension. Attention logits $\mathbf{A}^{l,h} \in \mathbb{R}^{S \times S}$ is given by:

$$\mathbf{A}^{l,h} = \frac{Q^{l,h}(K^{l,h})^\top}{\sqrt{D_k}}, \tag{2}$$

which directly reveals both the direction and intensity of interactions between any two tokens during the forward pass. After applying a causal mask followed by softmax, $\mathbf{A}^{l,h}$ becomes the attention weight matrix that is multiplied with the value $V^{l,h}$ to produce the head's attention output.

Attention Deficits of LVLMs

To understand the instability of multimodal in-context learning in LVLMs, we delve into the model's attention mechanism to analyze its dynamics in both effective and ineffective scenarios. Our objective is to uncover distinct patterns that distinguish these cases. Specifically, we investigate attention dynamics at two levels: (1) Within each ICD: We examine text-based visual grounding to verify whether the model pays attention to relevant objects in the images. (2) Across ICDs: We evaluate whether the model effectively distributes attention among different ICDs according to their relevance to the query. To ensure generalizable experimental results, we experiment with **3**-shot settings on two LVLMs: LLaVA-NeXT-7B (Li et al. 2024a) and Idefics2-8B (Laurençon et al. 2024b). Query samples and ICDs are taken from the validation and training sets of VQAv2, respectively.

Setups. We pair each query with three ICDs of the same question type (e.g., "How many" or "Is there") and process the resulting sequences using both LVLMs. From this pool, we identify 2,500 sequences where LLaVA-NeXT-7B produces a correct answer while Idefics2-8B fails, and another 2,500 sequences with the opposite outcome. These form two distinct sets of 5,000 cases each: a "correct" group (indicating effective ICL) and a "wrong" group (indicating ineffective ICL). For each sequence, we manually annotate bounding boxes on the ICD images based on their corresponding textual descriptions. We then extract attention maps from each model layer, identify the top 20% most-attended regions, and compute their Intersection over Union (IoU) with the annotated boxes to obtain the **intra-ICD alignment score** s_{align}. Additionally, we sample 10,000 sequences from the original pool and replace two of the three ICDs in each sequence with unrelated images. This leaves only one key ICD that matches the query type. For each sequence, we then create three variants by placing the key ICD in the three slots, respectively. Using the same processing pipeline, we build new correct and wrong groups with 5,000 cases each. For these groups, we compute saliency maps (Wang et al. 2023) at each layer and measure the proportion of information flow from the key ICD to the generated answer tokens relative to the total layer-wise flow. This yields the **contribution score** s_{contrib}.

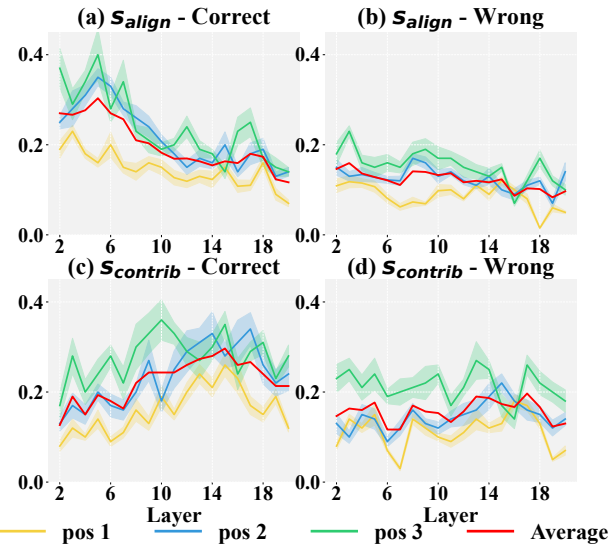

Figure 2: Layer-wise trends of the intra-ICD alignment score s_{align} and and the key ICD contribution score $s_{contrib}$ in effective and ineffective multimodal ICL. Pos 1, 2, and 3 denote the key ICD position in the sequence.

Results. As shown in Figure 2, in the correct group, LVLMs exhibit significantly higher s_{align} in the shallow layers (Layers 2–4), with the largest score gap compared to the wrong group also occurring in these early layers. This suggests that effective ICL relies on aligning visual attention with textual semantics from the outset. Furthermore, in the correct group, s_{contrib} shows a marked increase beginning around Layer 10 and remains consistently high through Layer 20. In contrast, this rise is absent in the wrong group, where scores remain consistently lower. This pattern indicates that attention allocation in the middle layers, driven by the relevance of the query, is another key factor for successful ICL. We refer to the failure of LVLMs to establish either of these attention behaviors as **attention deficits**, which contribute to instability in ICL. We also observe that the earlier an ICD appears in the sequence, the lower its alignment and contribution scores are across all layers and across both groups.

We distill three core findings from the above results:

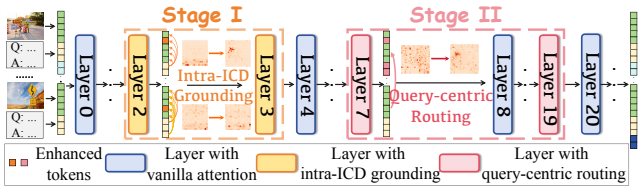

Figure 3: An overview pipeline of CAMA.

- ***Finding 1***: In shallow layers, LVLMs struggle to focus on visual information that aligns with the text semantics within each image-text pair.
- ***Finding 2***: In middle layers, LVLMs struggle to prioritize the key ICDs that match the query.
- ***Finding 3***: Both deficits are amplified by the position of ICDs, with those earlier in the sequence experiencing more severe impacts.

Method

Overview

To mitigate the two attention deficits at inference time without training, we introduce **Context-Aware Modulated Attention** (CAMA). CAMA dynamically reshapes the internal attention logits according to the input sequence during the prefilling process, encouraging the model to focus on the tokens most critical to effective ICL. The overall pipeline is depicted in Figure 3. Building on ***Finding 1*** and ***2***, we divide CAMA into two stages, each attached to a different depth of the LLM and mainly aimed at a specific deficit:

- **Stage I:** In the shallow layers, CAMA performs **Intra-ICD grounding**. At this stage, for each ICD, we first locate the key image tokens that are essential to justify the provided answer. We then amplify the attention paid to these tokens so the model captures the critical visual cues, laying a solid foundation for subsequent ICL.
- **Stage II:** In the middle layers, CAMA performs **query-centric routing**. At this stage, we operate at the attention head level to manage the complex interactions between the query sample and the ICDs. Specifically, we identify the heads that exhibit the strongest query-to-ICD attention and rescale their logits based on the cross-modal similarity between the query and each ICD.

Stage I: Intra-ICD grounding

Stage I operates on the shallow layer set $\mathcal{L}_{\text{stageI}}$, focusing on improving image-text alignment within each ICD to enhance LVLM's early perception of key visual features. This not only mitigates the early attention deficits but also benefits subsequent layers. To achieve this goal, we first need to identify the key visual tokens in each ICD based on the semantics of its paired Q-A text. However, conventional metrics, such as summing attention scores from text to image tokens or relying on embedding similarity, are inadequate for multimodal ICL, as they may exacerbate attention deficits or fail to capture the nuanced semantics of Q-A dialogues (Marino et al. 2021). To address this, we introduce a **dynamic attention increment** strategy tailored to multimodal ICDs.

For the i-th ICD in X, we designate three anchor tokens $\mathbf{T}_i$: the first token of the question $\mathbf{S}_i^Q[0]$, the first token of the answer $\mathbf{S}_i^A[0]$ and the last token of the answer $\mathbf{S}_i^A[-1]$. In our setting, these tokens are typically "Q", "A", and a punctuation mark (as illustrated in Figure 1(a)). After the first layer, the forward pass enables these tokens to capture and summarize the semantics of the preceding tokens. Therefore, they can be treated as proxies for the overall semantics of the image, the question, and the answer. We begin by computing the attention distributions that these anchor tokens assign to every image token $j \in \mathbf{S}_i^I$ in layer $l \in \mathcal{L}_{\text{stageI}}$:

$$P_l(\mathbf{T}_i) = \text{softmax}_j\Big(\tfrac{1}{H}\sum_{h=1}^{H}\mathbf{A}^{l,h}(\mathbf{T}_i, j)\Big) \in \mathbb{R}^{1\times S_i^I}, \tag{3}$$

where $\mathbf{T}_i \in \{\mathbf{S}_i^Q[0], \mathbf{S}_i^A[0], \mathbf{S}_i^A[-1]\}$ denotes the anchor tokens. The inner summation averages the logits over all H heads in l, and the row-wise softmax turns this average into a probability distribution across the image tokens.

Next, we calculate the differences between these distributions as dynamic attention increments. They provide bias-reduced estimates of the image tokens' contributions to both question understanding and answer making within each ICD. We quantify them by defining two non-negative forward gains $c_{l,i}^1, c_{l,i}^2 \in \mathbb{R}^{1\times S_i^I}$ in a divergence-like form:

$$c_{l,i}^1 = \big[P_l(\mathbf{S}_i^A[0]) - P_l(\mathbf{S}_i^Q[0])\big]_+ \cdot \log\frac{P_l(\mathbf{S}_i^A[0])}{P_l(\mathbf{S}_i^Q[0])}, \tag{4}$$

$$c_{l,i}^2 = \big[P_l(\mathbf{S}_i^A[-1]) - P_l(\mathbf{S}_i^A[0])\big]_+ \cdot \log\frac{P_l(\mathbf{S}_i^A[-1])}{P_l(\mathbf{S}_i^A[0])}, \tag{5}$$

where $[\cdot]_+$ keeps only the positive values by setting all negative ones to zero. After summing the two gains, column j is the score of the j-th image token:

$$s_{i,j} = \sum_{l\in\mathcal{L}_{\text{stageI}}} (c_{l,i}^1 + c_{l,i}^2)[j]. \tag{6}$$

A larger $s_{i,j}$ indicates that the token plays a more significant role in the visual perception of the image by LVLM based on the corresponding Q-A pair and is therefore more closely aligned with the text semantics. For each i-th ICD, we take the image token indices with the top-k_I% scores and define them as the key set $\mathcal{K}_{\mathcal{I}i}$. Finally, we modulate the attention logits in Eq.2 to amplify the key image tokens' *incoming* attention:

$$\mathbf{A}^{l,h}(r,j) \leftarrow \mathbf{A}^{l,h}(r,j) + \frac{n-i+1}{n}\frac{s_{i,j}}{\max\limits_{j'\in\mathcal{K}_{\mathcal{I}i}} s_{i,j'} + \epsilon}. \tag{7}$$

where $\epsilon = 1e{-}6$, $j \in \mathcal{K}_{\mathcal{I}i}$ and r denotes any later token in X. The factor $(n-i+1)/n$ is used to offset the position bias noted in ***Finding 3***. Note that during this stage, we also detect and enhance the query sample's key image tokens through only $c_{l,n+1}^1$. With intra-ICD grounding, image tokens that align well with corresponding textual semantics receive greater attention in subsequent layers and during answer generation, providing a bias-corrected basis for deeper reasoning.

Stage II: Query-centric routing

After highlighting the most semantically aligned image tokens in Stage I, Stage II works on the middle layers $\mathcal{L}_{\text{stageII}}$ to refine the global information flow under the guidance of the query. Inspired by (Singh et al. 2024), which shows that specific attention heads dominate the information flow from the query sample to ICDs during ICL, we perform fine-grained and targeted modulation at the head level to lessen the influence of other complex interactions in multimodal ICL.

We first use the attention flow from the query sample to the context (i.e., ICDs) as a signal to identify **query-centric heads**. As information gradually shifts toward later tokens in the middle layers, and text typically exhibits a stronger attention-flow tendency (Gao et al. 2019), we measure the *query→context flow* using only the query sample's text. Recall that $\mathbf{S}_{n+1}^{T}$ denotes the text token index set of the query sample. For each head h in layer $l \in \mathcal{L}_{\text{stageII}}$, we compute:

$$\rho^{l,h} = \frac{1}{|\mathbf{S}_{n+1}^{T}|} \sum_{q \in \mathbf{S}_{n+1}^{T}} \sum_{c \in \mathbf{S}_{ctx}} \mathbf{A}^{l,h}(q, c), \tag{8}$$

where $\mathbf{S}_{\text{ctx}} = \bigcup_{i=1}^{n} \left(\mathbf{S}_i^I \cup \mathbf{S}_i^Q \cup \mathbf{S}_i^A\right)$. $\rho_{l,h}$ aggregates the attention flow from the query sample to the preceding context, allowing us to quantify each head's contribution in that direction. The heads are then ranked by $\rho_{l,h}$, and the top-k_{II}% are chosen as query-centric heads, forming the set $\mathcal{H}_l^{\text{QC}}$.

Next, we perform ICD-level attention modulation within each query-centric head. In the middle layers, token-level attention becomes blurred by aggregation and is suboptimal for separating the contributions of individual ICDs. Thus, we propose a similarity-based method. To balance semantic maturity and clarity, the hidden states from the final layer of Stage I, $\mathcal{L}_{\text{stageI}}[-1]$, are used for subsequent computations.

For the embeddings of the i-th ICD taken from $\mathcal{L}_{\text{stageI}}[-1]$, we compute the mean of the key image tokens $\mathcal{K}_{\mathcal{I}i}$ to obtain a visual vector and the mean of all question-and-answer tokens to obtain a textual vector. We then concatenate these two vectors and apply ℓ_2-normalization to the resulting embedding, yielding the joint representation p_i. We apply the same steps to the query sample to obtain p_{query}. The cosine similarity between p_i and p_{query} is the query-centric score:

$$w_i = \frac{\exp\left(\langle p_i, p_{\text{query}} \rangle\right)}{\sum_{k=1}^{n} \exp\left(\langle p_k, p_{\text{query}} \rangle\right)}. \tag{9}$$

The score w_i quantifies how semantically relevant each ICD is to the query sample. Applying this score in the query-centric heads mitigates the LVLM's difficulty in locating crucial contexts and improves answer generation. Specifically, for each head $h \in \mathcal{H}_l^{\text{QC}}$ in layer $l \in \mathcal{L}_{\text{stageII}}$, we modulate the attention logits in Eq.2 as follows:

$$\mathbf{A}^{l,h}(r, J) \leftarrow \mathbf{A}^{l,h}(r, J) + \frac{n - i + 1}{n} w_i. \tag{10}$$

We apply this modulation to all key image tokens and text tokens of each ICD to maintain completeness, so $J \in (\mathcal{K}_{\mathcal{I}i} \cup \mathbf{S}_i^Q \cup \mathbf{S}_i^A)$ and r denote any subsequent token in X. $(n - i + 1)/n$ applies the same position decay as in Eq.7. Stage II enhances the attention given to each ICD in proportion to its contribution, ensuring that key information is not lost within the extensive context. The two stages of CAMA jointly enable the LVLM to exploit the provided context more effectively. All remaining layers maintain their original attention.

Experiments

Setup

Benchmarks and models. Following the standard multimodal ICL evaluation (Awadalla et al. 2023), we test CAMA on VQAv2 (Goyal et al. 2017), VizWiz (Gurari et al. 2018), and OK-VQA (Marino et al. 2019). To further assess the generalization of CAMA, we also evaluate it on GQA (Hudson and Manning 2019), TextVQA (Singh et al. 2019), the CLEVR subset of VL-ICL bench (Zong, Bohdal, and Hospedales 2024), and MMStar (Chen et al. 2024a). In addition to LLaVA-NeXT-7B and Idefics2-8B, we also report results on two latest LVLMs, InternVL2.5-8B (Chen et al. 2024c) and Qwen2.5VL-7B (Bai et al. 2025).

Baselines. We compare CAMA with five baselines. (1) **Vanilla** denotes the vanilla models. (2) The instruction-augmented method (**+Inst**) add an instruction before each sequence: "First, study the examples we provide. Then utilize what you have learned to answer the new question." (3) Contrastive decoding (**CD**) (Lee, Tsai, and Chiu 2024) replaces each ICD image with a blank one and uses the distorted logits to calibrate the original logits. (4) Visual enhancement (**VE**) (Su et al. 2025) manually draws a red bounding box around the relevant region of each ICD image. (5) SoFt Attention (**SoFA**) (Tian et al. 2025) is a training-free method that inserts a bidirectional attention mask after every two decoder layers, which reduces position bias when multiple images are contained in the input.

Implementation details. For each benchmark, samples in its validation set act as query samples, each paired with **eight** randomly retrieved ICDs from the training split, forming an 8-shot sequence. Stage I is applied to the 2nd and 3rd layers. Stage II is applied to every second layer from the 7th through the 19th. We set $k_{\text{I}} = k_{\text{II}} = 20$. All experiments are conducted on NVIDIA H200 GPUs.

Main Results

CAMA is effective and robust across all VQA benchmarks and LVLMs. Table 1 presents accuracy results across seven VQA benchmarks for four LVLMs with varying input image resolutions and LLM decoders. CAMA achieves the highest accuracy in all 28 experiments, surpassing all baselines. On average, it raises accuracy over the vanilla models by 2.96%. Notably, stronger models benefit even more: InternVL2.5 and Qwen2.5VL see improvements of 3.61% and 3.15%, respectively, compared to 2.35% on LLaVA-NeXT and 2.73% on Idefics2. These findings show that CAMA is both effective and generalizable. With its plug-and-play design, CAMA can consistently enhance emerging open-source LVLMs, which gives it strong practical value.

CAMA can activate the effect of prompt-based methods. We also report results that combine CAMA with a prompt-based baseline, as shown in Table 1 in the rows

LVLM	Method	VQAv2	VizWiz	OK-VQA	GQA	TextVQA	CLEVR	MMStar	Avg.
LLaVA-NeXT	Vanilla	61.86	37.64	57.63	55.38	61.93	16.50	44.72	47.95
	+Inst	61.69	38.52	58.21	55.70	61.74	17.46	43.95	48.18
	CD	61.79	37.70	57.48	55.46	62.07	17.18	41.59	47.61
	VE	61.94	37.58	57.97	55.29	61.78	18.16	44.92	48.23
	SoFA	63.21	38.09	58.14	57.42	62.28	14.29	45.71	48.45
	CAMA	64.46	39.87	59.94	58.60	63.40	18.67	47.16	50.30
	CAMA(+Inst)	64.89	40.23	60.27	58.71	63.61	21.28	47.42	50.92
	CAMA(VE)	**65.24**	**40.67**	**60.58**	**59.04**	**64.07**	**23.17**	**47.71**	**51.50**
Idefics2	Vanilla	57.32	38.46	43.60	57.49	70.02	34.61	42.65	49.16
	+Inst	57.61	38.25	43.75	57.38	70.30	35.48	42.51	49.21
	CD	56.83	38.19	43.47	57.30	68.87	33.70	41.49	48.55
	VE	57.28	38.89	44.26	57.98	71.18	36.41	42.93	49.85
	SoFA	59.04	38.95	46.12	57.75	72.31	34.44	43.29	50.27
	CAMA	60.53	39.90	47.23	59.79	74.38	36.52	44.86	51.89
	CAMA(+Inst)	60.74	**40.37**	47.69	60.00	**76.21**	38.95	44.59	52.40
	CAMA(VE)	**60.82**	40.16	**47.80**	**60.08**	75.72	**39.58**	**45.27**	**52.78**
InternVL2.5	Vanilla	69.58	58.27	62.32	67.21	80.29	56.91	62.70	65.33
	+Inst	69.89	58.92	62.18	67.53	81.10	57.34	62.38	65.53
	CD	69.81	58.79	63.01	67.54	80.17	57.11	62.56	65.57
	VE	69.80	58.96	63.14	67.28	80.39	59.53	62.95	66.00
	SoFA	70.85	59.62	62.48	68.30	82.75	59.12	62.70	66.55
	CAMA	72.54	62.15	66.27	70.68	85.19	61.45	**64.27**	68.94
	CAMA(+Inst)	72.61	62.48	65.58	70.93	85.46	64.67	63.37	68.97
	CAMA(VE)	**72.85**	**64.57**	**66.73**	**71.30**	**85.66**	**64.90**	63.79	**69.97**
Qwen2.5VL	Vanilla	71.94	57.39	65.70	82.31	83.61	62.83	65.18	69.85
	+Inst	72.43	57.80	66.18	83.57	83.72	63.65	66.03	70.43
	CD	72.31	57.35	66.07	82.49	83.76	62.72	65.63	70.05
	VE	72.69	59.30	66.41	84.55	83.92	63.26	66.47	70.94
	SoFA	72.31	59.06	67.28	84.52	84.34	64.85	66.87	71.32
	CAMA	74.96	60.83	68.80	85.32	87.14	66.15	**67.78**	73.00
	CAMA(+Inst)	74.79	61.28	69.21	85.26	86.87	**67.02**	67.39	73.12
	CAMA(VE)	**75.22**	**62.37**	**69.74**	**85.69**	**87.89**	66.81	67.70	**73.63**

Table 1: Accuracy of 8-shot ICL on seven VQA benchmarks for four LVLMs under different enhancement methods. The highest value is highlighted in **bold**, and the second highest is underlined. The cells shaded in light gray present the results of CAMA.

Method	Image captioning		Classification	Storytelling
	Flickr30k	MSCOCO	Hatefulmemes	L-I-VST
Vanilla	69.93	112.46	75.16	38.61
+Inst	71.28	114.36	76.20	38.46
CD	72.35	114.97	74.38	40.32
SoFA	72.65	115.89	76.51	40.41
CAMA	73.88	116.72	77.49	41.79
CAMA(+Inst)	**74.37**	**117.04**	**77.93**	**42.38**

Table 2: Average performance of 8-shot multimodal ICL on four benchmarks spanning three additional tasks, reported using CIDEr↑ , ROC-AUC↑, and L-I-score↑, respectively.

"CAMA(+Inst)" and "CAMA(VE)". We find that a prompt-based method alone gives only a marginal performance gain, which confirms the attention deficits inside LVLMs. When we add CAMA the model not only improves on its own but also **activates** the real benefit of these methods. For example, +Inst exceeds the vanilla model by just 0.27%, whereas CAMA lifts this improvement to 3.28% and adds another 0.32% over using CAMA alone. The performance gain brought by CAMA's activation is most evident on CLEVR, where VE sharply reduces the difficulty of cognizing the original image-text mapping. This result further confirms CAMA's practical promise, as it allows curated prompts to exert their full effect and maximizes model performance.

CAMA exhibits strong cross-task generalization. To fully evaluate the generalization of CAMA, we introduce three additional tasks beyond VQA: image captioning, image classification, and visual storytelling. In these tasks, the text paired with each image is no longer a Q-A pair; instead, it is a caption, a class label, or a narrative sentence, and all three start with the prefix "A:". The text part of the query sample only contains "A:". Therefore, for these tasks, Stage I of CAMA is based solely on attention gains $c^1_{l,i}$ to identify key image tokens. The results are reported in Table 2. CAMA again achieves superior performance on all these tasks and activates the gains of +Inst, showing that its benefits extend beyond VQA to broader interleaved scenarios.

Ablation Study and Analyses

Adaptability to diverse ICD counts. It is non-trivial for enhancement methods to adapt to different shot counts to meet various practical needs. As Figure 4(a) shows, CAMA consistently outperforms the vanilla model in 2, 4, 8, and 16-shot configurations, yielding gains of 2.15% to 6.52%. These results validate CAMA's generalization across diverse scenarios and its ability to meet varying requirements. Meanwhile, CAMA's advantage grows as the sequence length increases, which aligns with ***Finding 3*** because longer contexts intensify positional bias. As LVLMs' context windows expand,

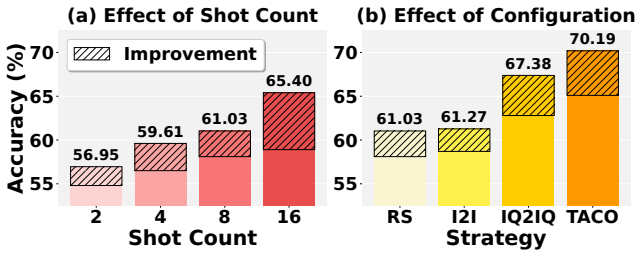

Figure 4: Average performance of CAMA across four LVLMs and seven VQA benchmarks as the count of ICDs and sequence configuration strategies vary.

Method	VQAv2	VizWiz	OK-VQA	GQA	TextVQA
CAMA	68.12	50.69	60.56	68.60	77.53
w/o Stage I	67.05	49.23	59.46	68.09	76.99
w/o Stage II	66.41	48.87	58.40	67.21	76.08
Stage I					
w/o increment	67.14	48.86	59.28	67.39	76.80
w/o top-k_{I}%	67.44	50.10	59.82	67.93	77.14
w/o position	67.91	50.12	60.02	68.05	76.98
Stage II					
w/o top-k_{II}%	67.18	49.25	59.84	68.11	76.69
$L_{\text{stageI}}[0]$	67.85	49.16	60.28	68.20	77.05
w/o position	68.03	49.84	60.19	68.20	77.12

Table 3: Average 8-shot performance of CAMA on four LVLMs under different ablation settings. "w/o increment" disables dynamic attention increment and computes the attention score using only $\mathbf{S}_i^A[-1]$. "w/o top k_{I}%/k_{II}%" applies CAMA to all image tokens or heads. "$L_{\text{stageI}}[0]$" computes similarity with embeddings from $L_{\text{stageI}}[0]$. "w/o position" removes the position-decay factor.

CAMA provides a promising path for further progress.

Adaptability to diverse sequence configurations. The in-context sequence configuration is key to ICL performance. Our main experiments build sequences through uniform random sampling, while some advanced systems use embedding similarity to improve sequence quality. Each candidate ICD and the query sample are encoded using CLIP-ViT-L/14 to compute cosine similarity, and the top eight matches are selected as ICDs. I2I uses only image embeddings, and IQ2IQ concatenates image and question embeddings. We also include TACO (Li et al. 2025c), a state-of-the-art ICD retriever based on language models. As shown in Figure 4(b), CAMA consistently improves accuracy by 2.59% to 5.08%. Stronger retrieval gives larger gains, which shows that the activation effect of CAMA also applies to configuration methods.

The superiority of CAMA designs. We conduct a comprehensive ablation study on the two stages of CAMA and the key components within each stage (Table 3). Removing either Stage I or Stage II leads to a noticeable drop in performance, with the absence of Stage II resulting in a more substantial degradation. This confirms that the two-stage modulation is necessary and that stronger query-guided reasoning is more crucial for multimodal ICL. Disabling each key design of a stage also produces different levels of degradation. We

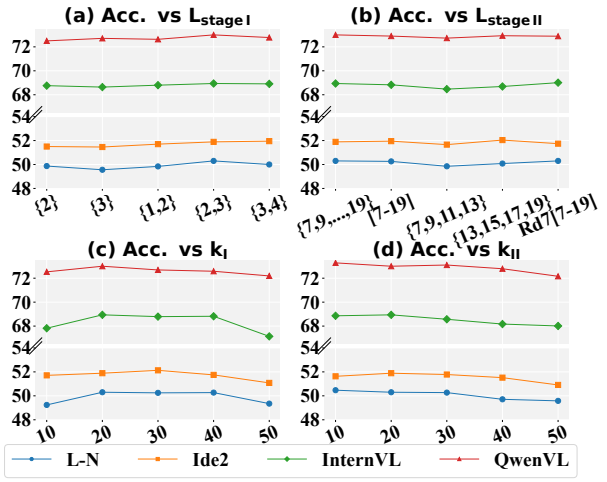

Figure 5: Average performance of CAMA on seven VQA benchmarks while varying L_{stageI}, L_{stageII}, k_{I}, and k_{II}. Random7 means randomly choosing seven layers.

draw the following conclusions: (1) Selecting the key image tokens is essential when enhancing image tokens in shallow layers. It prevents our modulation from being diluted by the later softmax. (2) Query-centric heads play a vital role in multimodal ICL. Targeted enhancement of these heads promotes specific information flows. (3) The dynamic attention increment avoids the bias introduced by the attention sink. (4) The position-decay factor in CAMA effectively alleviates position bias when LVLMs perform ICL with long sequences, preventing crucial information at the beginning of the sequence from being overlooked. This is particularly important in multimodal ICL, where each image consumes a large number of tokens and leads to long contexts, thereby improving overall performance. As the number of ICD shots increases, the position factor becomes more significant.

Impact of the selection of layers and hyperparameters. As shown in Figure 5, the performance of CAMA remains stable as we vary the layers where the two stages are applied. Meanwhile, when k_{I} is set between 20% and 40%, CAMA stays consistent, and when k_{II} is between 10% and 30%, CAMA also stays consistent. This behavior indicates that query-centric heads are more targeted than key image tokens, yet both possess a relatively wide optimal range. The results demonstrate that CAMA is robust to layer and hyperparameter choices and does not require carefully curated tuning strategies, highlighting its practicality and efficiency.

Conclusion

We introduce CAMA, a plug-and-play and training-free attention modulation method that enhances multimodal ICL. We begin by analyzing the attention dynamics of LVLMs in this setting and, guided by the findings, design a two-stage modulation pipeline for CAMA. CAMA effectively corrects attention deficits and guides the model to focus on image regions that contribute the most to multimodal ICL. Experiments across multiple benchmarks and LVLMs show that CAMA achieves superior results. We believe that CAMA can inspire new directions for future advances of LVLM.

Acknowledgments

We acknowledge the computing resources provided by NSF ACCESS.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Awadalla, A.; Gao, I.; Gardner, J.; Hessel, J.; Hanafy, Y.; Zhu, W.; Marathe, K.; Bitton, Y.; Gadre, S.; Sagawa, S.; Jitsev, J.; Kornblith, S.; Koh, P. W.; Ilharco, G.; Wortsman, M.; and Schmidt, L. 2023. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. arXiv:2308.01390.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Wang, J.; Qiao, Y.; Lin, D.; et al. 2024a. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37: 27056–27087.

Chen, S.; Han, Z.; He, B.; Liu, J.; Buckley, M.; Qin, Y.; Torr, P.; Tresp, V.; and Gu, J. 2025. Can Multimodal Large Language Models Truly Perform Multimodal In-Context Learning? In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 6000–6010. IEEE.

Chen, T.; Zhang, E.; Gao, Y.; Li, K.; Sun, X.; Zhang, Y.; Li, H.; and Ji, R. 2024b. Mmict: Boosting multi-modal fine-tuning with in-context examples. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Chen, Y.; Zhao, C.; Yu, Z.; McKeown, K.; and He, H. 2023. On the relation between sensitivity and accuracy in in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 155–167.

Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. 2024c. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Ma, J.; Li, R.; Xia, H.; Xu, J.; Wu, Z.; Chang, B.; et al. 2024. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, 1107–1128.

Doveh, S.; Perek, S.; Mirza, M. J.; Lin, W.; Alfassy, A.; Arbelle, A.; Ullman, S.; and Karlinsky, L. 2024. Towards multimodal in-context learning for vision & language models. *arXiv preprint arXiv:2403.12736*.

Fazli, M.; Wei, B.; and Zhu, Z. 2025. Mitigating Hallucination in Large Vision-Language Models via Adaptive Attention Calibration. *arXiv preprint arXiv:2505.21472*.

Gao, P.; Jiang, Z.; You, H.; Lu, P.; Hoi, S. C.; Wang, X.; and Li, H. 2019. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6639–6648.

Gao, T.; Fisch, A.; and Chen, D. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, 3816–3830.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Guo, Q.; Wang, L.; Wang, Y.; Ye, W.; and Zhang, S. 2024. What makes a good order of examples in in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, 14892–14904.

Gurari, D.; Li, Q.; Stangl, A. J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; and Bigham, J. P. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3608–3617.

Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6700–6709.

Jia, H.; Jiang, C.; Xu, H.; Ye, W.; Dong, M.; Yan, M.; Zhang, J.; Huang, F.; and Zhang, S. 2025. Symdpo: Boosting in-context learning of large multimodal models with symbol demonstration direct preference optimization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 9361–9371.

Jiang, D.; He, X.; Zeng, H.; Wei, C.; Ku, M.; Liu, Q.; and Chen, W. 2024. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.

Kim, Y.; Kim, H. J.; Park, C.; Park, C.; Cho, H.; Kim, J.; Yoo, K. M.; Lee, S.-g.; and Kim, T. 2024. Adaptive contrastive decoding in retrieval-augmented generation for handling noisy contexts. *arXiv preprint arXiv:2408.01084*.

Laurençon, H.; Marafioti, A.; Sanh, V.; and Tronchon, L. 2024a. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*.

Laurençon, H.; Tronchon, L.; Cord, M.; and Sanh, V. 2024b. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37: 87874–87907.

Lee, Y.-L.; Tsai, Y.-H.; and Chiu, W.-C. 2024. Delve into visual contrastive decoding for hallucination mitigation of large vision-language models. *arXiv preprint arXiv:2412.06775*.

Li, F.; Zhang, R.; Zhang, H.; Zhang, Y.; Li, B.; Li, W.; Ma, Z.; and Li, C. 2024a. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Li, J.; Zhang, J.; Jie, Z.; Ma, L.; and Li, G. 2025a. Mitigating hallucination for large vision language model by inter-modality correlation calibration decoding. *arXiv preprint arXiv:2501.01926*.

Li, L.; Peng, J.; Chen, H.; Gao, C.; and Yang, X. 2024b. How to configure good in-context sequence for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26710–26720.

Li, X.; Lv, K.; Yan, H.; Lin, T.; Zhu, W.; Ni, Y.; Xie, G.; Wang, X.; and Qiu, X. 2023. Unified demonstration retriever for in-context learning. *arXiv preprint arXiv:2305.04320*.

Li, Y. 2025. Advancing Multimodal In-Context Learning in Large Vision-Language Models with Task-aware Demonstrations. *arXiv preprint arXiv:2503.04839*.

Li, Y.; He, H.; Cao, Y.; Cheng, Q.; Fu, X.; and Tang, R. 2025b. M2IV: Towards Efficient and Fine-grained Multimodal In-Context Learning in Large Vision-Language Models. *arXiv preprint arXiv:2504.04633*.

Li, Y.; Yun, T.; Yang, J.; Feng, P.; Huang, J.; and Tang, R. 2025c. TACO: Enhancing Multimodal In-context Learning via Task Mapping-Guided Sequence Configuration. *arXiv preprint arXiv:2505.17098*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, J.; Shen, D.; Zhang, Y.; Dolan, W. B.; Carin, L.; and Chen, W. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd workshop on knowledge extraction and integration for deep learning architectures*, 100–114.

Liu, S.; Zheng, K.; and Chen, W. 2024. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, 125–140. Springer.

Lu, Y.; Bartolo, M.; Moore, A.; Riedel, S.; and Stenetorp, P. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8086–8098.

Marino, K.; Chen, X.; Parikh, D.; Gupta, A.; and Rohrbach, M. 2021. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14111–14121.

Marino, K.; Rastegari, M.; Farhadi, A.; and Mottaghi, R. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 3195–3204.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8317–8326.

Singh, A. K.; Moskovitz, T.; Hill, F.; Chan, S. C.; and Saxe, A. M. 2024. What needs to go right for an induction head? a mechanistic study of in-context learning circuits and their formation. *arXiv preprint arXiv:2404.07129*.

Su, Z.; Xia, P.; Guo, H.; Liu, Z.; Ma, Y.; Qu, X.; Liu, J.; Li, Y.; Zeng, K.; Yang, Z.; et al. 2025. Thinking with Images for Multimodal Reasoning: Foundations, Methods, and Future Frontiers. *arXiv preprint arXiv:2506.23918*.

Tian, X.; Zou, S.; Yang, Z.; and Zhang, J. 2025. Identifying and Mitigating Position Bias of Multi-image Vision-Language Models. *arXiv preprint arXiv:2503.13792*.

Wang, L.; Li, L.; Dai, D.; Chen, D.; Zhou, H.; Meng, F.; Zhou, J.; and Sun, X. 2023. Label Words are Anchors: An Information Flow Perspective for Understanding In-Context Learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 9840–9855.

Wies, N.; Levine, Y.; and Shashua, A. 2023. The learnability of in-context learning. *Advances in Neural Information Processing Systems*, 36: 36637–36651.

Wu, Z.; Wang, Y.; Ye, J.; and Kong, L. 2022. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering. *arXiv preprint arXiv:2212.10375*.

Yang, X.; Peng, Y.; Ma, H.; Xu, S.; Zhang, C.; Han, Y.; and Zhang, H. 2024. Lever LM: configuring in-context sequence to lever large vision language models. *Advances in Neural Information Processing Systems*, 37: 100341–100368.

Yang, X.; Wu, Y.; Yang, M.; Chen, H.; and Geng, X. 2023. Exploring diverse in-context configurations for image captioning. *Advances in Neural Information Processing Systems*, 36: 40924–40943.

Ye, Q.; Xu, H.; Xu, G.; Ye, J.; Yan, M.; Zhou, Y.; Wang, J.; Hu, A.; Shi, P.; Shi, Y.; et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Zhou, Y.; Li, X.; Wang, Q.; and Shen, J. 2024. Visual in-context learning for large vision-language models. *arXiv preprint arXiv:2402.11574*.

Zong, Y.; Bohdal, O.; and Hospedales, T. 2024. VL-ICL bench: The devil in the details of multimodal in-context learning. *arXiv preprint arXiv:2403.13164*.