SELF-STEERING MULTIMODAL LLMS FOR ENHANCED KNOWLEDGE UTILIZATION VIA DYNAMIC INFORMATION FLOW MODULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Retrieval-Augmented Generation enhances Multimodal Large Language Models (MLLMs) with external knowledge, yet retrieved evidence is often underutilized during generation. Previous studies typically attribute this issue to insufficient attention allocation or conflicts between contextual and parametric knowledge, overlooking how retrieved information is propagated and preserved in MLLMs. In this work, we revisit this problem from the perspective of internal information flow across Transformer layers. Through layer-wise analyses, we reveal that context-grounded information is better captured in intermediate representations than in final-layer representations, suggesting that ineffective information propagation limits the preservation of retrieved knowledge across layers. To address this, we propose Dynamic Contribution Steering (DCS), a training-free framework that dynamically modulates the contributions of attention, MLP, and residual pathways according to token-level attention distributions and neuron activation statistics. By suppressing irrelevant or dominant internal signals and enhancing context-related representations, DCS improves the utilization of retrieved knowledge without requiring additional training, steering vectors, or extra decoding procedures. Furthermore, we introduce Knowledge Layer Amplification (KLA), which identifies intermediate layers with stronger context-grounded decoding capability and adaptively incorporates their representations during generation. Extensive experiments on MRAG benchmarks across various MLLMs show that the proposed approach improves context-grounded answer generation, establishing information-flow steering as an effective paradigm for improving knowledge utilization in MLLMs.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) (Liu et al., 2024a; Bai et al., 2025a;b) have significantly advanced vision-language understanding and generation by integrating visual encoders with Large Language Models (LLMs) (Yang et al., 2025). Despite their impressive performance on general multimodal tasks such as image captioning (Al-Shamayleh et al., 2024) and visual question answering (Marino et al., 2019), they still struggle with knowledge-intensive scenarios that require up-to-date or domain-specific information beyond their parametric memory. To overcome this limitation, Multimodal Retrieval-Augmented Generation (MRAG) (Mei et al., 2025) has emerged as a promising paradigm. By retrieving relevant evidence from external knowledge sources and incorporating it into the model input, MRAG equips MLLMs with access to non-parametric contextual knowledge at inference time.

Ideally, the retrieved context should guide MLLMs toward more factual and context-grounded answers. However, a growing body of work (Shi et al., 2024a; Yuan et al., 2024; An et al., 2026b) has shown that merely providing relevant context does not guarantee effective knowledge utilization. Even when the retrieved context explicitly contains the answer, MLLMs may fail to incorporate it into reasoning, instead relying on parametric knowledge or superficial visual cues. Consequently, the discrepancy between knowledge availability and knowledge utilization has emerged as a fundamental challenge in developing reliable MRAG systems.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

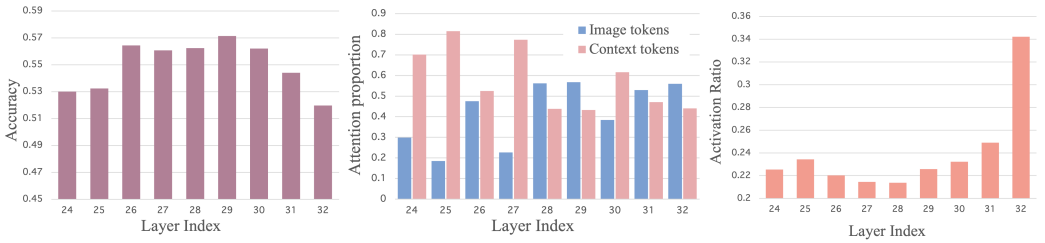

Figure 1: Experiments on the knowledge-intensive benchmark InfoSeek (Chen et al., 2023b) using LLaVA-1.5 (Liu et al., 2023). **Left**: Decoding accuracy obtained from hidden states at different layers (Results are reported from the 24th layer onward, as earlier layers cannot yield valid predictions). **Middle**: Layer-wise attention allocation proportion towards image and contextual tokens. **Right**: Layer-wise MLP neuron activation ratio.

To address this challenge, existing studies have attempted to explain the knowledge utilization issue from perspectives such as attention allocation (An et al., 2026b), knowledge conflict (Shi et al., 2024a; Yuan et al., 2024), and decoding calibration (Yuan et al., 2024). Building on these insights, prior methods typically employ contrastive decoding (Li et al., 2023), context-aware decoding (Shi et al., 2024a), or adaptive logit fusion (An et al., 2026b) strategies to encourage greater reliance on retrieved evidence. While effective in certain scenarios, these approaches generally require additional decoding passes, resulting in increased inference overhead. More importantly, they often characterize knowledge utilization through observable decoding behavior, leaving the internal mechanisms by which retrieved information is preserved, transformed, or suppressed across model layers less fully understood.

To better understand this process, we examine how context-grounded information evolves inside Transformer-based MLLMs. Instead of analyzing only final-layer predictions, we evaluate whether hidden representations from different layers can independently support context-grounded decoding. As illustrated in Fig. 1 (Left), intermediate layers exhibit stronger decoding accuracy than the final layer on knowledge-intensive tasks. This observation suggests that retrieved knowledge may already be encoded in intermediate representations, but can become weakened, overwritten, or misaligned during subsequent computation. Thus, MRAG failures may arise not only from insufficient attention to retrieved context, but also from the way information flows through attention, MLP, and residual pathways across layers.

Motivated by this finding, we propose to view knowledge utilization in MRAG as a dynamic information-flow balancing problem. In Transformer blocks, attention modules mediate contextual information aggregation, MLP modules transform and amplify latent knowledge representations, and residual connections preserve information across layers. As shown in Fig. 1 (Middle and Right), both attention allocation and MLP activation vary substantially across layers, indicating that retrieved knowledge is not propagated uniformly through the network. When the relative contributions of these computational pathways become imbalanced, context-grounded signals can be overwhelmed by parametric knowledge or superficial visual cues, ultimately leading to context-ungrounded and factually incorrect responses.

Building on the above insight, we introduce Dynamic Contribution Steering (DCS), a training-free framework that adaptively steers the contributions of different computational pathways based on context-dependent attention distributions and MLP activation statistics. Specifically, DCS selectively suppresses over-dominant attention or MLP pathways that amplify noisy parametric knowledge and superficial visual cues, while correspondingly enhancing residual pathways to better preserve contextual knowledge signals, thereby steering MLLMs toward context-grounded responses. Furthermore, we propose Knowledge Layer Amplification (KLA), which identifies and amplifies knowledge-utilization layers whose hidden representations exhibit strong context-grounded decoding capability. Unlike prior activation steering methods (Nguyen & Le, 2026a;b), our framework is self-steering and requires neither auxiliary training data nor learned steering vectors. It also differs from contrastive decoding methods (An et al., 2026b) by avoiding multiple decoding passes, making it suitable for efficient inference-time deployment. Extensive experiments on multiple MRAG benchmarks, together with theoretical analysis, demonstrate the effectiveness of our approach.

In summary, our main contributions are threefold: (1) We provide a unified mechanistic perspective on knowledge utilization in MRAG, revealing that the issue arises from the information flow across

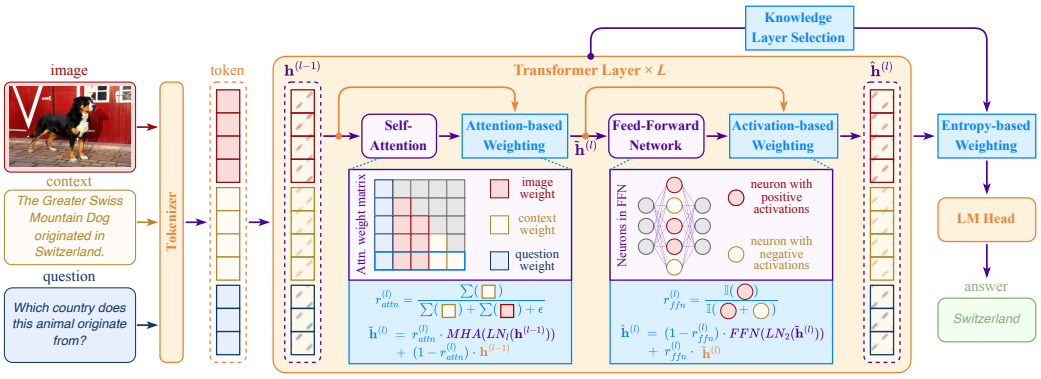

Figure 2: Overview of the proposed framework. Dynamic Contribution Steering adaptively rebalances the contributions of different pathways, while Knowledge Layer Amplification enhances the contribution of context-grounded layers for decoding.

Transformer layers. (2) We propose Dynamic Contribution Steering (DCS), a self-steering, training-free framework that can control pathway contributions and steer MLLMs for context-grounded generation. (3) Extensive experiments and theoretical analyses demonstrate that DCS consistently improves knowledge utilization while introducing no additional training or inference procedures.

2 Related Work

Knowledge Utilization in MRAG. Multimodal Retrieval Augmented Generation (MRAG) (Caffagni et al., 2024; Yan & Xie, 2024) has been widely explored to equip MLLMs with fine-grained and up-to-date knowledge for solving knowledge-intensive tasks. Despite its promise, MRAG still suffers from a persistent knowledge utilization gap: retrieved evidence that explicitly contains the correct answer may not be effectively leveraged during generation (An et al., 2026b; Yuan et al., 2025). Existing studies have investigated this issue from several perspectives. One line of work attributes such failures to insufficient attention allocation to contextual evidence (Liu et al., 2024b; An et al., 2026b), suggesting that models may under-attend to relevant retrieved tokens even when they are available. Another line emphasizes conflicts between retrieved evidence and parametric knowledge (Huang et al., 2026; Shi et al., 2024a), where internal memory can override external contextual signals. Despite these insights, most existing approaches characterize knowledge utilization primarily through observable model patterns, leaving the internal mechanisms by which retrieved information is preserved, transformed, or suppressed across model layers less well understood.

Knowledge Utilization Enhancement. To improve knowledge utilization, prior methods primarily explore inference-time intervention strategies, which can be broadly categorized into three lines. The first line focuses on attention reallocation, encouraging MLLMs to attend more to retrieved contextual evidence (An et al., 2026b). The second line adopts contrastive decoding strategies to suppress interference from parametric knowledge and mitigate hallucinated outputs (Li et al., 2023; Wang et al., 2024; Yuan et al., 2024). The third line leverages activation steering, often learned from additional supervised data, to explicitly guide model representations toward better utilization of external knowledge (Wang et al., 2025). Despite their effectiveness, these methods typically require either additional data or multiple decoding passes, resulting in non-negligible inference overhead.

3 Method

3.1 Preliminaries

MRAG. Let q denote the input question and I the corresponding image. The goal of an MLLM $\mathcal{M}_\theta$ is to autoregressively predict an answer sequence y. In the MRAG framework, external knowledge is introduced to compensate for the model's limited parametric knowledge. A multimodal retriever retrieves a piece of relevant textual context c from an external knowledge base conditioned on (q, I), and the model predicts the next token according to:

$$y_t = \arg\max_{y_t} \mathcal{M}_\theta(y_t \mid q, I, c, y_{<t}), \tag{1}$$

where $y_{<t}$ is the prefix generated before step t.

Transformer Architecture. We consider a decoder consisting of L layers. Given the output hidden states $\mathbf{h}^{(\ell-1)} \in \mathbb{R}^d$ at layer $\ell-1$, each Transformer block is computed as

$$\tilde{\mathbf{h}}^{(\ell)} = \underbrace{\mathrm{MHA}\Big(\mathrm{LN}_1\Big(\mathbf{h}^{(\ell-1)}\Big)\Big)}_{\text{self-attention module}} + \underbrace{\mathbf{h}^{(\ell-1)}}_{\text{residual}}, \quad \mathbf{h}^{(\ell)} = \underbrace{\mathrm{FFN}\Big(\mathrm{LN}_2\Big(\tilde{\mathbf{h}}^{(\ell)}\Big)\Big)}_{\text{feed-forward network}} + \underbrace{\tilde{\mathbf{h}}^{(\ell)}}_{\text{residual}}, \tag{2}$$

where $\mathrm{LN}_1(\cdot)$ and $\mathrm{LN}_2(\cdot)$ denote layer normalization.

Roles of three modules. Eq. 2 shows that each Transformer block updates hidden states through three information pathways: the residual pathway, the self-attention pathway, and the FFN pathway. The residual pathway directly carries previous hidden states to the next layer, serving as an information-preserving channel across layers. The self-attention pathway enables token-level interaction by aggregating information from other tokens, which is essential for extracting and incorporating contextual information in MLLMs (An et al., 2026b). The FFN pathway applies token-wise nonlinear transformations, refining hidden representations and injecting learned parametric knowledge into hidden states through a key-value mechanism (Geva et al., 2021; An et al., 2026a). These three pathways jointly determine how representations evolve across layers. Therefore, their relative contributions provide a useful perspective for analyzing how retrieved knowledge is propagated inside MLLMs. Specifically, if the attention pathway assigns insufficient attention to retrieved context tokens, the model may fail to incorporate external evidence even when the relevant knowledge is present in the input (Liu et al., 2024b; An et al., 2026b). If the FFN pathway becomes overly activated, stored parametric knowledge may dominate the representation update and interfere with context-grounded knowledge utilization (Huang et al., 2026). Conversely, a sufficiently preserved residual pathway can help maintain knowledge-rich intermediate representations across layers. These analyses suggest that modulating the contributions of the three pathways is important for enhancing knowledge utilization in MLLMs.

3.2 Dynamic Contribution Steering

Based on the analysis that external knowledge utilization is closely related to the relative contributions of attention, FFN, and residual pathways, we propose **Dynamic Contribution Steering** (DCS), a training-free framework that adaptively steers the internal information flow of MLLMs. The key idea is to dynamically adjust the contribution of each pathway according to the model's attention distribution and neuron activation statistics, so that retrieved contextual knowledge can be better preserved and utilized during decoding.

Attention-based Modulation. In MRAG, the input sequence usually consists of visual tokens, query tokens, and retrieved context tokens. Let $\mathcal{V}$ and $\mathcal{C}$ denote the index sets of visual tokens and retrieved context tokens, respectively. At layer ℓ, we compute the average attention mass of each head assigned to visual and context tokens as:

$$a_{\mathrm{vis}}^{(\ell)} = \frac{1}{|\mathcal{V}|}\sum_{j\in\mathcal{V}} \mathbf{A}_{nj}^{(\ell)}, \quad a_{\mathrm{ctx}}^{(\ell)} = \frac{1}{|\mathcal{C}|}\sum_{j\in\mathcal{C}} \mathbf{A}_{nj}^{(\ell)}, \tag{3}$$

where $\mathbf{A}^{(\ell)}$ denotes the attention matrix at layer ℓ, n denotes the index of the last token. We then define the context attention ratio as follows:

$$r_{\mathrm{attn}}^{(\ell)} = \frac{a_{\mathrm{ctx}}^{(\ell)}}{a_{\mathrm{ctx}}^{(\ell)} + a_{\mathrm{vis}}^{(\ell)} + \epsilon}, \tag{4}$$

where ϵ is a small constant for numerical stability. A larger $r_{\mathrm{attn}}^{(\ell)}$ indicates that the attention pathway assigns relatively more attention to retrieved context tokens, while a smaller value suggests that visual tokens dominate the attention flow. Based on this ratio, DCS adaptively controls the contribution of attention output and the residual pathway:

$$\tilde{\mathbf{h}}^{(\ell)} = r_{\mathrm{attn}}^{(\ell)} \cdot \mathrm{MHA}\left(\mathrm{LN}_1\left(\mathbf{h}^{(\ell-1)}\right)\right) + (1 - r_{\mathrm{attn}}^{(\ell)}) \cdot \mathbf{h}^{(\ell-1)}. \tag{5}$$

When sufficient attention is directed to the retrieved context, DCS increases the contribution of the attention pathway to facilitate context integration. Conversely, when attention is dominated by visual

tokens, DCS strengthens the residual pathway, preserving context-aware representations accumulated in preceding layers from being diluted by context-irrelevant attention outputs. Consequently, Attention-based Modulation adaptively regulates information flow according to the current attention distribution, ensuring that contextual evidence is effectively retained for subsequent processing.

Activation-based Modulation. The FFN module can be interpreted as a key-value memory (Geva et al., 2021). Given the normalized hidden state $\mathbf{x}^{(\ell)} = \mathrm{LN}_2(\tilde{\mathbf{h}}^{(\ell)})$, the FFN output is computed as:

$$\mathrm{FFN}(\mathbf{x}^{(\ell)}) = \sum_{i=1}^{d_{\mathrm{ff}}} \sigma\left(\mathbf{x}^{(\ell)}\mathbf{k}_i\right)\mathbf{v}_i = \sum_{i=1}^{d_{\mathrm{ff}}} m_i^{(\ell)}\mathbf{v}_i, \tag{6}$$

where d_{ff} denotes the intermediate dimension of the FFN, $\mathbf{k}_i$ and $\mathbf{v}_i$ denote the i-th key and value vectors in the FFN, respectively. $\sigma(\cdot)$ denotes the FFN activation function, and $m_i^{(\ell)}$ denotes the activation coefficient of the corresponding parametric memory entry.

Following the key-value interpretation and previous work (Huang et al., 2026), a neuron is considered activated when $m_i^{(\ell)} > 0$, indicating that the associated parametric knowledge is retrieved and contributes to the current representation update (Fan et al., 2025). We therefore define the FFN activation ratio at layer ℓ as:

$$r_{\mathrm{ffn}}^{(\ell)} = \frac{1}{d_{\mathrm{ff}}} \sum_{i=1}^{d_{\mathrm{ff}}} \mathbb{I}\left[m_i^{(\ell)} > 0\right], \tag{7}$$

where $\mathbb{I}[\cdot]$ is the indicator function. A larger $r_{\mathrm{ffn}}^{(\ell)}$ indicates that more parametric memory entries are activated, implying stronger involvement of the model's internal knowledge in the current update. So DCS uses this activation ratio to adaptively balance the FFN output and the residual pathway. Then the FFN sub-layer is updated as:

$$\hat{\mathbf{h}}^{(\ell)} = r_{\mathrm{ffn}}^{(\ell)} \cdot \tilde{\mathbf{h}}^{(\ell)} + (1 - r_{\mathrm{ffn}}^{(\ell)}) \cdot \mathrm{FFN}\left(\mathrm{LN}_2\left(\tilde{\mathbf{h}}^{(\ell)}\right)\right). \tag{8}$$

When the FFN activation ratio is high, DCS reduces the influence of the FFN pathway and preserves more information through the residual connection, preventing overly activated parametric memories from dominating context-grounded representations. Conversely, when fewer FFN neurons are activated, the FFN pathway is allowed to contribute more, enabling the model to exploit useful parametric knowledge when it does not conflict with retrieved evidence.

Modulation Propagation Analysis. We further analyze how steering the contribution of an intermediate layer influences the final hidden states used for decoding. For simplicity, we omit layer normalization and express the Transformer update at layer ℓ in its additive residual form:

$$\mathbf{h}^{(\ell)} = \mathbf{h}^{(\ell-1)} + \mathbf{a}^{(\ell)} + \mathbf{m}^{(\ell)}, \tag{9}$$

where $\mathbf{a}^{(\ell)}$ and $\mathbf{m}^{(\ell)}$ denote the outputs of the attention and FFN pathways, respectively. Recursively expanding the residual updates yields the final hidden state at layer L:

$$\mathbf{h}^{(L)} = \mathbf{h}^{(\ell)} + \sum_{k=\ell+1}^{L} \mathbf{a}^{(k)} + \sum_{k=\ell+1}^{L} \mathbf{m}^{(k)}. \tag{10}$$

Eq. 10 shows that the final representation is recursively accumulated from the residual stream together with the attention and FFN updates of all subsequent layers. Consequently, modifying the contribution of an intermediate layer alters not only its output representation but also the inputs received by every following layer, thereby influencing the entire downstream computation. In DCS, the hidden-state update at layer ℓ is reweighted as:

$$\hat{\mathbf{h}}^{(\ell)} = \alpha^{(\ell)}\mathbf{h}^{(\ell-1)} + \beta^{(\ell)}\mathbf{a}^{(\ell)} + \gamma^{(\ell)}\hat{\mathbf{m}}^{(\ell)}, \tag{11}$$

where $\alpha^{(\ell)} = r_{\mathrm{ffn}}^{(\ell)} \cdot (1 - r_{\mathrm{attn}}^{(\ell)})$, $\beta^{(\ell)} = r_{\mathrm{ffn}}^{(\ell)} \cdot r_{\mathrm{attn}}^{(\ell)}$, and $\gamma^{(\ell)} = (1 - r_{\mathrm{ffn}}^{(\ell)})$ control the contributions of each pathways. $\hat{\mathbf{m}}^{(\ell)}$ is the modified output of the FFN sub-layer caused by modulation. The final hidden state can then be written as

$$\hat{\mathbf{h}}^{(L)} = \alpha^{(\ell)}\mathbf{h}^{(\ell-1)} + \beta^{(\ell)}\mathbf{a}^{(\ell)} + \gamma^{(\ell)}\hat{\mathbf{m}}^{(\ell)} + \sum_{k=\ell+1}^{L} \hat{\mathbf{a}}^{(k)} + \sum_{k=\ell+1}^{L} \hat{\mathbf{m}}^{(k)}, \tag{12}$$

Table 1: VQA Accuracy on generative freeform VQA datasets over three runs.

Model	Decoding	Human			Validation		
		Unseen Question	Unseen Entity	Overall	Unseen Question	Unseen Entity	Overall
LLaVA-1.5	Regular	$7.59_{(\pm 0.08)}$	$7.90_{(\pm 0.05)}$	$7.74_{(\pm 0.01)}$	$19.98_{(\pm 0.02)}$	$19.59_{(\pm 0.01)}$	$19.78_{(\pm 0.01)}$
	Parametric	$6.27_{(\pm 0.05)}$	$6.26_{(\pm 0.19)}$	$6.26_{(\pm 0.09)}$	$7.14_{(\pm 0.03)}$	$6.28_{(\pm 0.01)}$	$6.68_{(\pm 0.00)}$
	CD	$7.39_{(\pm 0.21)}$	$7.31_{(\pm 0.06)}$	$7.35_{(\pm 0.09)}$	$20.32_{(\pm 0.01)}$	$19.90_{(\pm 0.00)}$	$20.11_{(\pm 0.01)}$
	AdaCAD	$7.81_{(\pm 0.02)}$	$8.07_{(\pm 0.03)}$	$7.94_{(\pm 0.01)}$	$21.23_{(\pm 0.03)}$	$20.91_{(\pm 0.02)}$	$21.07_{(\pm 0.03)}$
	Entropy	$7.98_{(\pm 0.09)}$	$8.34_{(\pm 0.02)}$	$8.15_{(\pm 0.03)}$	$21.97_{(\pm 0.02)}$	$21.85_{(\pm 0.01)}$	$21.91_{(\pm 0.03)}$
	CAD	$8.02_{(\pm 0.05)}$	$8.15_{(\pm 0.03)}$	$8.08_{(\pm 0.02)}$	$21.68_{(\pm 0.01)}$	$20.93_{(\pm 0.02)}$	$21.30_{(\pm 0.02)}$
	COIECD	$8.78_{(\pm 0.04)}$	$8.65_{(\pm 0.01)}$	$8.71_{(\pm 0.01)}$	$22.43_{(\pm 0.05)}$	$21.73_{(\pm 0.05)}$	$22.07_{(\pm 0.05)}$
	AGLA	$8.74_{(\pm 0.28)}$	$9.18_{(\pm 0.00)}$	$8.94_{(\pm 0.08)}$	$22.34_{(\pm 0.02)}$	$21.88_{(\pm 0.01)}$	$22.11_{(\pm 0.02)}$
	VCD	$9.22_{(\pm 0.00)}$	$9.26_{(\pm 0.00)}$	$9.24_{(\pm 0.02)}$	$22.30_{(\pm 0.03)}$	$22.38_{(\pm 0.03)}$	$22.23_{(\pm 0.03)}$
	ALFAR	$12.80_{(\pm 0.09)}$	$11.24_{(\pm 0.00)}$	$11.96_{(\pm 0.02)}$	$23.82_{(\pm 0.00)}$	$23.75_{(\pm 0.01)}$	$23.78_{(\pm 0.01)}$
	Ours	$\mathbf{13.39}_{(\pm 0.06)}$	$\mathbf{11.74}_{(\pm 0.05)}$	$\mathbf{12.51}_{(\pm 0.06)}$	$\mathbf{24.56}_{(\pm 0.02)}$	$\mathbf{24.44}_{(\pm 0.01)}$	$\mathbf{24.50}_{(\pm 0.01)}$
InstructBLIP	Regular	$4.20_{(\pm 0.00)}$	$3.86_{(\pm 0.03)}$	$4.02_{(\pm 0.01)}$	$3.60_{(\pm 0.01)}$	$3.82_{(\pm 0.00)}$	$3.71_{(\pm 0.00)}$
	Parametric	$4.06_{(\pm 0.01)}$	$3.65_{(\pm 0.01)}$	$3.84_{(\pm 0.01)}$	$2.36_{(\pm 0.01)}$	$1.92_{(\pm 0.00)}$	$2.12_{(\pm 0.00)}$
	CD	$4.52_{(\pm 0.03)}$	$3.55_{(\pm 0.01)}$	$3.98_{(\pm 0.01)}$	$3.59_{(\pm 0.01)}$	$4.00_{(\pm 0.00)}$	$3.79_{(\pm 0.00)}$
	AdaCAD	$4.57_{(\pm 0.05)}$	$3.70_{(\pm 0.10)}$	$4.09_{(\pm 0.08)}$	$3.71_{(\pm 0.02)}$	$4.35_{(\pm 0.01)}$	$4.01_{(\pm 0.01)}$
	Entropy	$4.56_{(\pm 0.05)}$	$4.14_{(\pm 0.01)}$	$4.34_{(\pm 0.02)}$	$3.81_{(\pm 0.01)}$	$4.39_{(\pm 0.00)}$	$4.08_{(\pm 0.00)}$
	CAD	$4.52_{(\pm 0.03)}$	$3.55_{(\pm 0.01)}$	$3.98_{(\pm 0.01)}$	$3.77_{(\pm 0.03)}$	$4.43_{(\pm 0.02)}$	$4.08_{(\pm 0.02)}$
	COIECD	$4.64_{(\pm 0.10)}$	$4.08_{(\pm 0.02)}$	$4.33_{(\pm 0.01)}$	$4.07_{(\pm 0.00)}$	$4.54_{(\pm 0.00)}$	$4.30_{(\pm 0.00)}$
	AGLA	$4.80_{(\pm 0.05)}$	$4.28_{(\pm 0.09)}$	$4.52_{(\pm 0.05)}$	$3.74_{(\pm 0.01)}$	$4.10_{(\pm 0.01)}$	$3.91_{(\pm 0.01)}$
	VCD	$4.70_{(\pm 0.01)}$	$4.14_{(\pm 0.01)}$	$4.40_{(\pm 0.00)}$	$3.62_{(\pm 0.01)}$	$4.12_{(\pm 0.00)}$	$3.85_{(\pm 0.00)}$
	ALFAR	$5.98_{(\pm 0.00)}$	$5.67_{(\pm 0.00)}$	$5.82_{(\pm 0.00)}$	$4.55_{(\pm 0.00)}$	$5.68_{(\pm 0.00)}$	$5.05_{(\pm 0.00)}$
	Ours	$\mathbf{6.88}_{(\pm 0.01)}$	$\mathbf{6.57}_{(\pm 0.02)}$	$\mathbf{6.72}_{(\pm 0.02)}$	$\mathbf{5.25}_{(\pm 0.01)}$	$\mathbf{6.08}_{(\pm 0.00)}$	$\mathbf{5.63}_{(\pm 0.01)}$

where $\hat{\mathbf{a}}^{(k)}$ and $\hat{\mathbf{m}}^{(k)}$ denote the attention and FFN outputs after propagating the intervention introduced at layer ℓ. Subtracting Eq. 12 and Eq. 10, we can get:

$$\hat{\mathbf{h}}^{(L)} - \mathbf{h}^{(L)} = \Delta^{(\ell)} + \sum_{k=\ell+1}^{L} \left(\Delta \mathbf{a}^{(k)} + \Delta \mathbf{m}^{(k)} \right). \tag{13}$$

Let $\mathcal{F}_{\ell+1:L} = \mathcal{T}^{(L)} \circ \cdots \circ \mathcal{T}^{(\ell+1)}$ denote the composition of the subsequent Transformer layers. Then:

$$\hat{\mathbf{h}}^{(L)} = \mathcal{F}_{\ell+1:L} \left(\mathbf{h}^{(\ell)} + \Delta^{(\ell)} \right), \quad \mathbf{h}^{(L)} = \mathcal{F}_{\ell+1:L} \left(\mathbf{h}^{(\ell)} \right). \tag{14}$$

Assuming that the perturbation is sufficiently small, a first-order Taylor expansion around $\mathbf{h}^{(\ell)}$ gives

$$\hat{\mathbf{h}}^{(L)} - \mathbf{h}^{(L)} \approx \mathbf{J}_{\ell+1:L} \Delta^{(\ell)}, \tag{15}$$

where $\mathbf{J}_{\ell+1:N}$ is the Jacobian of the subsequent Transformer layers with respect to the hidden state at layer ℓ.

The analysis demonstrates that steering an intermediate layer affects the final hidden states through two complementary mechanisms: it directly rebalances the contributions of the residual, attention, and FFN pathways at the intervention layer, while indirectly reshaping the computations of all subsequent layers through recursive propagation. Therefore, steering knowledge-utilization layers can guide the decoding representation toward hidden states that preserve retrieved contextual knowledge more effectively while suppressing excessive reliance on parametric knowledge.

3.3 Knowledge Layer Amplification

While DCS dynamically rebalances the information flow within each Transformer layer, our analysis reveals that intermediate layers often exhibit stronger knowledge-utilization capability than the final layer. Motivated by this observation, we propose **Knowledge Layer Amplification** (KLA), a training-free strategy that identifies the layer encoding the most context-grounded representations and explicitly integrates its hidden states into the final decoding process.

Knowledge Layer Selection. An ideal knowledge-utilization layer should effectively capture retrieved evidence while avoiding excessive interference from parametric knowledge. Specifically,

such a layer is expected to allocate sufficient attention to retrieved contexts and maintain relatively less dependence on FFN transformations, thereby preserving richer context-grounded information for subsequent decoding. Based on this intuition, we define a layer-wise context utilization score:

$$s^{(\ell)} = r^{(\ell)}_{\text{attn}} \cdot \left(1 - r^{(\ell)}_{\text{ffn}}\right), \tag{16}$$

where higher values indicate that the layer exhibits stronger utilization of retrieved knowledge with less domination from parametric transformations. The knowledge layer is then selected as follows:

$$\ell^* = \arg\max_{\ell} s^{(\ell)}. \tag{17}$$

Entropy-guided Layer Fusion. After identifying the knowledge layer ℓ^*, we further integrate its hidden state with the final-layer representation before decoding. Moreover, since the final layer often exhibits stronger FFN dominance, it tends to encode more parametric knowledge. Fusing the final-layer representation with the selected knowledge layer thus provides a complementary combination of parametric knowledge and retrieved contextual knowledge, which is particularly beneficial when retrieved evidence is incomplete or noisy.

Let $\mathbf{h}^{(L)}$ and $\mathbf{h}^{(\ell^*)}$ denote the hidden states from the final layer and the selected knowledge layer, respectively. We first project both representations into next-token distributions using the language modeling head $\mathbf{W}_o$:

$$\mathbf{p}^{(m)} = \text{Softmax}\left(\mathbf{W}_o \mathbf{h}^{(m)}\right), m \in \{L, \ell^*\}. \tag{18}$$

To adaptively determine the contribution of each layer, we measure the uncertainty of their decoding distributions using entropy, which can reflect the reliability of the prediction (Shi et al., 2024b; An et al., 2026a):

$$\mathcal{H}^{(m)} = -\sum_{v \in \mathcal{V}_{all}} p_v^{(m)} \log p_v^{(m)}, m \in \{L, \ell^*\}. \tag{19}$$

A lower entropy indicates a more confident prediction distribution. Based on this observation, we assign fusion weights according to entropy-aware confidence:

$$w^{(m)} = \frac{\exp\left(-\mathcal{H}^{(m)}\right)}{\sum_{n \in \{L, \ell^*\}} \exp\left(-\mathcal{H}^{(n)}\right)}, m \in \{L, \ell^*\}, \tag{20}$$

The fused representation is then obtained as:

$$\mathbf{h}^{\text{fuse}} = w^{(L)} \mathbf{h}^{(L)} + w^{(\ell^*)} \mathbf{h}^{(\ell^*)}. \tag{21}$$

Finally, the next token y_t is generated based on the fused representation:

$$\mathbf{p}(y_t) = \text{Softmax}\left(\mathbf{W}_o \mathbf{h}^{\text{fuse}}\right). \tag{22}$$

Both DCS and KLA require neither additional training nor auxiliary supervision. All modulation signals are derived solely from the model's intrinsic behaviors, including attention allocation, FFN activation statistics, and decoding uncertainty. By dynamically steering information pathways within Transformer layers and amplifying representations from knowledge-utilization layers, our approach enables MLLMs to preserve and exploit retrieved evidence for knowledge-grounded generation.

4 Experiment

4.1 Experimental Settings

Evaluation Benchmarks. Following previous work (An et al., 2026b), we conduct experiments on seven knowledge-intensive vision-language benchmarks spanning three representative scenarios. For *open-ended generation*, we adopt the **Human** and **Validation** splits of InfoSeek (Chen et al., 2023b). For *multiple-choice question answering*, we evaluate on **InfoSeek** (Chen et al., 2023b) and **ViQuAE** (Lerner et al., 2022). To further assess commonsense and encyclopedic reasoning,

Table 2: VQA Accuracy comparison on discriminative multi-choice VQA datasets over three runs.

Model	Decoding	InfoSeek	ViQuAE	Model	Decoding	InfoSeek	ViQuAE
Shikra	Regular	$19.41_{(\pm 0.12)}$	$17.73_{(\pm 0.01)}$	MiniGPT-4	Regular	$25.83_{(\pm 1.42)}$	$24.06_{(\pm 0.46)}$
	Parametric	$9.65_{(\pm 0.14)}$	$10.90_{(\pm 0.04)}$		Parametric	$19.73_{(\pm 0.57)}$	$20.42_{(\pm 1.22)}$
	CD	$24.83_{(\pm 0.20)}$	$21.38_{(\pm 0.03)}$		CD	$26.55_{(\pm 0.31)}$	$20.46_{(\pm 0.76)}$
	CAD	$24.51_{(\pm 0.06)}$	$21.68_{(\pm 0.15)}$		CAD	$27.58_{(\pm 0.43)}$	$23.39_{(\pm 0.07)}$
	AdaCAD	$23.95_{(\pm 0.18)}$	$21.51_{(\pm 0.01)}$		AdaCAD	$28.01_{(\pm 0.17)}$	$23.05_{(\pm 0.61)}$
	Entropy	$24.12_{(\pm 0.03)}$	$21.99_{(\pm 0.08)}$		Entropy	$28.84_{(\pm 0.68)}$	$22.59_{(\pm 0.07)}$
	COIECD	$24.18_{(\pm 0.13)}$	$21.68_{(\pm 0.15)}$		COIECD	$29.44_{(\pm 0.01)}$	$25.94_{(\pm 0.18)}$
	VCD	$25.76_{(\pm 0.14)}$	$23.11_{(\pm 0.72)}$		VCD	$28.74_{(\pm 0.10)}$	$25.45_{(\pm 0.05)}$
	AGLA	$26.26_{(\pm 0.29)}$	$22.72_{(\pm 0.15)}$		AGLA	$29.28_{(\pm 0.87)}$	$27.87_{(\pm 0.08)}$
	ALFAR	$28.11_{(\pm 0.11)}$	$23.42_{(\pm 0.29)}$		ALFAR	$30.88_{(\pm 0.12)}$	$32.28_{(\pm 0.02)}$
	Ours	$\mathbf{29.67}_{(\pm 0.08)}$	$\mathbf{24.52}_{(\pm 0.22)}$		**Ours**	$\mathbf{32.16}_{(\pm 0.03)}$	$\mathbf{33.57}_{(\pm 0.01)}$

we include **OK-VQA** (Marino et al., 2019), **AOK-VQA** (Schwenk et al., 2022), and **Encyclopedic VQA (E-VQA)** (Mensink et al., 2023).

Compared Models and Baselines. Our evaluation covers seven widely used multimodal large language models, including LLaVA-1.5 (7B,13B) (Liu et al., 2023), InstructBLIP (7B,13B) (Dai et al., 2023), Shikra (7B) (Chen et al., 2023a), MiniGPT-4 (7B) (Zhu et al., 2023), LLaVA-Next (7B) (Liu et al., 2024a), Qwen2.5-VL (3B,7B) (Bai et al., 2025b), and InternVL-2.5 (2B,4B,8B) (Chen et al., 2024). We compare our approach against two groups of training-free decoding methods. The first group consists of context-aware decoding algorithms, including CD (Li et al., 2023), CAD (Shi et al., 2024a), AdaCAD (Wang et al., 2024), Entropy-based Decoding (Qiu et al., 2024), COIECD (Yuan et al., 2024), and ALFAR (An et al., 2026b). The second group contains representative hallucination reduction methods for LVLMs, namely VCD (Leng et al., 2023) and AGLA (An et al., 2024a). Due to space constraints, **more experimental results are included in the Appendix**.

Implementation Details. Following previous work (An et al., 2026b), we use CLIP-ViT-L/14-336 (Radford et al., 2021) as the retriever and prepend the highest-ranked retrieved document to the model input as contextual knowledge. The retrieval corpus is built from Wikipedia (Chen et al., 2023b). During inference, all methods employ multinomial sampling. DCS is applied to the last 5 layers. To maintain consistency with previous decoding-based approaches (An et al., 2026b), adaptive plausibility constraints (Li et al., 2023) are imposed on the output logits. Results of compared methods are cited from ALFAR (An et al., 2026b) for a fair comparison. Experiments are conducted on two 3090 GPUs.

Table 3: VQA Accuracy comparison on the knowledge-based VQA datasets with LLaVA-1.5.

Model	OK-VQA	AOK-VQA	E-VQA
Regular	$46.17_{(\pm 0.12)}$	$44.13_{(\pm 0.00)}$	$19.14_{(\pm 2.83)}$
Parametric	$45.13_{(\pm 0.40)}$	$43.23_{(\pm 1.02)}$	$5.34_{(\pm 0.01)}$
CD	$55.00_{(\pm 0.03)}$	$51.67_{(\pm 0.44)}$	$28.62_{(\pm 0.68)}$
CAD	$56.43_{(\pm 0.46)}$	$53.93_{(\pm 0.56)}$	$28.62_{(\pm 0.76)}$
AdaCAD	$57.10_{(\pm 0.04)}$	$54.40_{(\pm 0.48)}$	$28.33_{(\pm 0.42)}$
Entropy	$56.27_{(\pm 0.05)}$	$53.93_{(\pm 0.26)}$	$29.24_{(\pm 0.11)}$
COIECD	$56.43_{(\pm 0.46)}$	$53.93_{(\pm 0.56)}$	$28.24_{(\pm 0.74)}$
VCD	$57.80_{(\pm 0.13)}$	$57.40_{(\pm 0.12)}$	$27.71_{(\pm 0.33)}$
AGLA	$57.53_{(\pm 0.05)}$	$55.40_{(\pm 0.63)}$	$28.29_{(\pm 0.56)}$
ALFAR	$60.83_{(\pm 0.01)}$	$59.93_{(\pm 0.02)}$	$29.57_{(\pm 0.08)}$
Ours	$\mathbf{62.01}_{(\pm 0.07)}$	$\mathbf{61.33}_{(\pm 0.16)}$	$\mathbf{30.15}_{(\pm 0.06)}$

4.2 Experimental Results

First, Tab. 1 reports the results on two free-form generative benchmarks (Chen et al., 2023b). Across all LVLMs and datasets, our framework consistently achieves the best performance, surpassing both the *Regular* decoding strategy and existing state-of-the-art decoding methods by clear margins. These results demonstrate that dynamically rebalancing information flow substantially improves knowledge utilization in retrieval-augmented LVLMs. **Second**, we further evaluate our framework on two multi-choice discriminative benchmarks (Zhu et al., 2024). As summarized in Tab. 2, our method consistently yields the highest accuracy across different models and datasets, outperforming all competing decoding strategies. The results indicate that the proposed framework generalizes effectively beyond free-form generation to discriminative retrieval-augmented reasoning tasks. **Third**, to evaluate the generality of our approach beyond entity-centric knowledge benchmarks (Chen et al., 2023b; Lerner et al., 2022), we additionally conduct experiments on commonsense reasoning datasets (Marino et al., 2019; Schwenk et al., 2022; Mensink et al., 2023) using

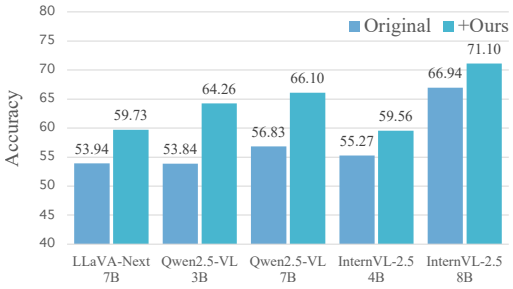

Figure 3: Experimental results with more recent MLLMs on the multi-choice ViQuAE dataset.

Table 4: Average inference time of different methods over 3,000 samples.

Variants	InfoSeek	Human
Regular	0.46s (1.00x)	0.50s (1.00x)
AdaCAD	0.62s (1.35x)	0.73s (1.45x)
COIECD	0.62s (1.35x)	0.73s (1.45x)
VCD	0.63s (1.38x)	0.74s (1.47x)
AGLA	0.83s (1.81x)	1.02s (2.04x)
ALFAR	0.62s (1.35x)	0.73s (1.46x)
Ours	**0.46s (1.00x)**	**0.51s (1.02x)**

LLaVA-1.5 (Liu et al., 2023). As shown in Tab. 3, our framework consistently establishes new state-of-the-art performance, confirming that the information flow rebalancing mechanism effectively improves knowledge utilization across a broader range of knowledge-intensive tasks.

4.3 Discussion

Ablation Study. We conduct ablation studies on two types of benchmarks (Chen et al., 2023b; Schwenk et al., 2022) using the LLaVA-1.5 (Liu et al., 2023) model as our backbone. As reported in Tab. 5, each proposed component contributes positively to the overall performance. Specifically, the two dynamic modulation modules consistently improve knowledge utilization by rebalancing the information flow, while the proposed Knowledge Layer Amplification further boosts performance by leveraging the stronger knowledge-aware representations encoded in intermediate layers.

Table 5: Ablation study with different model variants.

Variants	ViQuAE	OK-VQA
Regular	53.32	46.17
+ Attention-based Modulation	54.61	49.14
+ Activation-based Modulation	56.32	52.06
+ Knowledge Layer Amplification	58.94	62.01

Inference Efficiency. We further evaluate the inference efficiency of the proposed method against representative approaches on both discriminative and generative tasks. As shown in Tab. 4, unlike existing decoding-based methods that require two forward passes to separately estimate parametric and retrieved contextual knowledge, our method operates with a single forward pass and introduces virtually no additional computational overhead. This design enables a more efficient inference process while maintaining superior performance, making it suited for real-world deployment.

Experiments on More MLLMs. To evaluate the generalizability of our approach, we conduct experiments on more recent MLLMs with varying architectures and model scales. As shown in Fig. 3, our method consistently achieves performance improvements across all evaluated models, demonstrating its robustness and effectiveness under different model types and scales.

5 Conclusion

In this work, we revisit the knowledge utilization problem in MRAG from the perspective of internal information flow in Transformer-based MLLMs. Through comprehensive layer-wise analysis, we reveal that intermediate representations preserve stronger context-grounded knowledge than final-layer representations, suggesting that ineffective knowledge propagation and preservation across Transformer layers are key factors behind retrieval utilization failures. Motivated by this finding, we propose Dynamic Contribution Steering (DCS) and Knowledge Layer Amplification (KLA), two training-free and efficient strategies that dynamically modulate information pathways and enhance knowledge-aware representations during generation for MLLMs. Experiments on multiple MRAG benchmarks demonstrate that our approach consistently improves context-grounded answer generation across different MLLMs. These results highlight the importance of understanding and optimizing internal information flow, providing a new perspective for knowledge integration in MLLMs.

AI USE STATEMENT

In this work, we used generative AI tools solely for language polishing, including editing the manuscript to improve readability. We have not used generative AI tools for any tasks with required disclosure, including generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, translation, cleaning or reformatting datasets, qualitative and thematic data analysis, or interpreting results, and the rest of the required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, personal data, or any practices that require institutional review board (IRB) approval. We do not release any new dataset, and all data used in this study are publicly available and properly cited. Our methods and applications are not intended for harmful purposes, and we do not identify any potential conflicts of interest or sponsorship issues. We have considered issues of discrimination, bias, and fairness, and do not find any specific concerns arising from this work. There are no privacy or security issues, legal compliance concerns, or research integrity issues associated with this study.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our results, we provide detailed descriptions of our method and implementation in the main paper and the appendix, which include descriptions of the proposed method, experimental settings, and relevant implementation details.

REFERENCES

Ahmad Sami Al-Shamayleh, Omar Adwan, Mohammad A Alsharaiah, Abdelrahman H Hussein, Qasem M Kharma, and Christopher Ifeanyi Eke. A comprehensive literature review on image captioning methods and metrics based on deep learning technique. *Multimedia Tools and Applications*, 83(12):34219–34268, 2024.

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Guang Dai, Ping Chen, and Shijian Lu. Agla: Mitigating object hallucinations in large vision-language models with assembly of global and local attention. *arXiv preprint arXiv:2406.12718*, 2024a.

Wenbin An, Feng Tian, Jiahao Nie, Wenkai Shi, Haonan Lin, Yan Chen, QianYing Wang, Yaqiang Wu, Guang Dai, and Ping Chen. Knowledge acquisition disentanglement for knowledge-based visual question answering with large language models. *arXiv preprint arXiv:2407.15346*, 2024b.

Wenbin An, Haonan Lin, Jiahao Nie, Feng Tian, Wenkai Shi, Yaqiang Wu, Qianying Wang, and Ping Chen. Unleashing the potential of model bias for generalized category discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 15365–15373, 2025.

Wenbin An, Jiahao Nie, Feng Tian, Mingxiang Cai, Yaqiang Wu, Xiaoqin Zhang, and Shijian Lu. Enhancing retrieval-augmented large vision language models via knowledge conflict mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2318–2326, 2026a.

Wenbin An, Jiahao Nie, Feng Tian, Haonan Lin, Yaqiang Wu, QianYing Wang, Xiaoqin Zhang, Shijian Lu, et al. Boosting knowledge utilization in multimodal large language models via adaptive logits fusion and attention reallocation. *Advances in Neural Information Processing Systems*, 38: 37050–37074, 2026b.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1818–1826, 2024.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023a.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14948–14968, 2023b.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2306.04387*, 2023.

Yuchun Fan, Yongyu Mu, Yilin Wang, Lei Huang, Junhao Ruan, Bei Li, Tong Xiao, Shujian Huang, Xiaocheng Feng, and Jingbo Zhu. Slam: Towards efficient multilingual reasoning via selective language alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 9499–9515, 2025.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, 2021.

Pengcheng Huang, Zhenghao Liu, Yukun Yan, Haiyan Zhao, Xiaoyuan Yi, Hao Chen, Zhiyuan Liu, Maosong Sun, Tong Xiao, Ge Yu, et al. Parammute: Suppressing knowledge-critical ffns for faithful retrieval-augmented generation. *Advances in Neural Information Processing Systems*, 38: 100378–100410, 2026.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*, 2023.

Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G Moreno, and Jesús Lovón Melgarejo. Viquae, a dataset for knowledge-based visual question answering about named entities. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3108–3120, 2022.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12286–12312, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024a.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pp. 125–140. Springer, 2024b.

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

Lang Mei, Siyu Mo, Zhihan Yang, and Chong Chen. A survey of multimodal retrieval-augmented generation. *arXiv preprint arXiv:2504.08748*, 2025.

Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3113–3124, 2023.

Tuc Nguyen and Thai Le. Atlas: Adaptive test-time latent steering with external verifiers for enhancing llms reasoning. *arXiv preprint arXiv:2601.03093*, 2026a.

Tuc Nguyen and Thai Le. Beyond linear activation steering: Invertible latent transformations for controlling llm behavior. *arXiv preprint arXiv:2606.08454*, 2026b.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Zexuan Qiu, Zijing Ou, Bin Wu, Jingjing Li, Aiwei Liu, and Irwin King. Entropy-based decoding for retrieval-augmented large language models. *arXiv preprint arXiv:2406.17519*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pp. 146–162. Springer, 2022.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 783–791, 2024a.

Wenkai Shi, Wenbin An, Feng Tian, Yan Chen, Yaqiang Wu, Qianying Wang, and Ping Chen. A unified knowledge transfer network for generalized category discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 18961–18969, 2024b.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Adacad: Adaptively decoding to balance conflicts between contextual and parametric knowledge. *arXiv preprint arXiv:2409.07394*, 2024.

Yilin Wang, Heng Wang, Yuyang Bai, and Minnan Luo. Continuously steering llms sensitivity to contextual knowledge with proxy models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4682–4698, 2025.

Yibin Yan and Weidi Xie. Echosight: Advancing visual-language models with wiki knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1538–1551, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Xiaowei Yuan, Zhao Yang, Yequan Wang, Shengping Liu, Jun Zhao, and Kang Liu. Discerning and resolving knowledge conflicts through adaptive decoding with contextual information-entropy constraint. *arXiv preprint arXiv:2402.11893*, 2024.

Xiaowei Yuan, Zhao Yang, Ziyang Huang, Yequan Wang, Siqi Fan, Yiming Ju, Jun Zhao, and Kang Liu. Exploiting contextual knowledge in llms through v-usable information based layer enhancement. *arXiv preprint arXiv:2504.15630*, 2025.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Tinghui Zhu, Qin Liu, Fei Wang, Zhengzhong Tu, and Muhao Chen. Unraveling cross-modality knowledge conflicts in large vision-language models. *arXiv preprint arXiv:2410.03659*, 2024.

APPENDIX

A1 LIMITATION AND FUTURE WORK

Despite the improved performance in enhancing Multimodal LLMs for knowledge-intensive tasks, our work could be improved in several aspects. First, we conducted experiments on the most widely used MLLMs due to resource constraints. It will be useful to evaluate our framework on larger (i.e., 70B) MLLMs. Second, the framework requires direct access to the internal parameters of the MLLMs, rendering it inapplicable to black-box, API-based models. Extending the proposed method to operate effectively under black-box constraints represents a valuable direction for future research.

A2 EVALUATION METRICS

VQA Accuracy (ACC): VQA Accuracy assesses the quality of response by comparing it with the ground-truth answers.

Memorization Ratio (MR): Memorization Ratio quantifies the model's reliance on parametric knowledge, which can be calculated as follows:

$$\texttt{MR} = \frac{f_p}{f_p + f_c} \tag{A1}$$

where f_p is the frequency of relying on internal parametric knowledge, and f_c is the frequency of relying on external contextual knowledge. f_p and f_c are computed by comparing the responses with the output of vanilla MLLMs and the retrieved context, respectively.

Context Ratio (CR): Context Ratio indicates the model's preference for contextual knowledge, which can be calculated as follows:

$$\texttt{CR} = \frac{f_c}{f_p + f_c} \tag{A2}$$

By comparing the metrics MR and CR, we can analyze the model's knowledge preference and the sources of response error.

Knowledge Conflict Ratio (KCR): Knowledge Conflict Ratio measures the discrepancy between parametric and contextual knowledge by comparing the output of vanilla MLLMs and the retrieved context, which can be computed as follows:

$$\texttt{KCR} = \frac{f_{p=c}}{f_{p=c} + f_{p\neq c}} \tag{A3}$$

where $f_{p=c}$ is the frequency of the parametric knowledge that aligns with the retrieved contextual knowledge, and $f_{p\neq c}$ denotes the frequency of parametric knowledge that contradicts the retrieved contextual knowledge.

A3 KNOWLEDGE DEPENDENCY ANALYSIS

To verify that the proposed DCS framework improves MRAG performance by enhancing knowledge utilization in MLLMs, we conduct experiments on the InfoSeek dataset with ground-truth contextual knowledge. As shown in Tab. A1, although providing ground-truth knowledge (LLaVA + GT Context) can partially alleviate knowledge deficiencies, the model still achieves limited performance due to its strong reliance on outdated parametric knowledge. In contrast, by applying our method to MLLMs, DCS effectively improves their ability to utilize contextual knowledge with reduced interference from parametric knowledge, leading to a higher context utilization ratio and consequently better answering performance.

Table A1: Inference with parametric knowledge (LLaVA), perfect contextual knowledge (LLaVA + GT Context), and perfect contextual knowledge with our method (LLaVA + DCS).

Model	ACC↑	MR↓	CR↑
LLaVA	39.2	100.0	0.0
LLaVA + GT Context	65.5	46.9	53.1
LLaVA + DCS	**72.0**	**37.3**	**62.7**

Table A3: Statistics of the evaluated datasets and the corresponding knowledge bases used for context retrieval.

Dataset	# VQA pairs	Knowledge Base
Validation Chen et al. (2023b)	73,620	Wikipedia Chen et al. (2023b)
Human Chen et al. (2023b)	8,931	Wikipedia Chen et al. (2023b)
InfoSeek Zhu et al. (2024)	3,000	Wikipedia Chen et al. (2023b)
ViQuAE Zhu et al. (2024)	3,000	Wikipedia Chen et al. (2023b)
OK-VQA Marino et al. (2019)	5,046	GPT-3.5 An et al. (2024b)
AOK-VQA Schwenk et al. (2022)	1,145	GPT-3.5 An et al. (2024b)
E-VQA Mensink et al. (2023)	700	Encyclopedia Mensink et al. (2023)

A4 Influence of Parametric Knowledge

To further substantiate the widespread presence of conflicts between parametric and contextual knowledge, we conduct experiments on two multiple-choice datasets: InfoSeek (Chen et al., 2023b) and ViQuAE (Lerner et al., 2022), using LLaVA-1.5 (Liu et al., 2023). As shown in Tab. A2, the MLLM demonstrates a high conflict ratio, with approximately half of the samples exhibiting conflicts between parametric and contextual knowledge, which means that reducing the effect of parametric knowledge for contextual knowledge is essential for MRAG, reflecting the necessity of suppressing parametric knowledge activation in our proposed Activation-based Modulation.

Table A2: Knowledge Conflict Ratio between the parametric knowledge and the ground-truth contextual knowledge.

	Infoseek	**ViQuAE**
KCR	60.85%	48.94%

A5 Dataset and Knowledge Base

We present the statistics of the evaluated datasets and their corresponding knowledge bases in Tab. A3. Specifically, for the Validation (Chen et al., 2023b) and InfoSeek (Zhu et al., 2024) datasets, we follow prior work (Caffagni et al., 2024) and adopt a knowledge base consisting of 1.7K entities derived from the original Wikipedia-based knowledge base (Chen et al., 2023b). For the Human (Chen et al., 2023b) and ViQuAE (Lerner et al., 2022) datasets, we utilize the full Wikipedia knowledge base (Chen et al., 2023b), selecting 73.6K entities that include associated images to support knowledge retrieval. For OK-VQA (Marino et al., 2019) and AOK-VQA (Schwenk et al., 2022), we employ the knowledge base curated by (An et al., 2024b), which was automatically generated using GPT-3.5 (Ouyang et al., 2022). For E-VQA (Mensink et al., 2023), we select templated questions accompanied by images from the iNaturalist dataset (Van Horn et al., 2018), using the corresponding ground-truth knowledge to support inference. All evaluations are conducted using the official evaluation scripts provided with each benchmark.

A6 Compared Methods

The implementation details of all baseline and compared methods are summarized below. Since most of these methods were originally designed for textual data, we re-implemented them on multimodal datasets to ensure a fair and consistent comparison.

- *Regular* means sampling the response from the output of MLLMs with additional context as input: $p(y) = \mathrm{softmax}(l(y|q, I, c))$, where $l(y|q, I, c)$ is the logits with question q, image I and context c as inputs.
- *Parametric* means sampling the response from the output of vanilla MLLMs: $p(y) = \mathrm{softmax}(l(y|q, I))$, where $l(y|q, I)$ is the logits with question q and image I as inputs.
- *Contrastive Decoding (CD)* mitigates knowledge conflicts by contrasting the output logits with and without context as inputs: $p(y) = \mathrm{softmax}(l(y|q, I, c) - l(y|q, I))$.

- *Adaptive Context-Aware Decoding (AdaCAD)* adaptively adjusts the contribution of the parametric knowledge and contextual knowledge with JS divergence: $p(y) = \text{softmax}((1+\alpha)l(y|q,I,c) - \alpha l(y|q,I))$, where $\alpha = \text{JS}(l(y|q,I,c), l(y|q,I))$.
- *Entropy-based Decoding (Entropy)* adaptively adjusts the contribution of the parametric knowledge and contextual knowledge with the entropy of the output with context as input: $p(y) = \text{softmax}((1+\alpha)l(y|q,I,c) - \alpha l(y|q,I))$, where $\alpha = -\text{E}(l(y|q,I,c))$ is the negative entropy.
- *Context-Aware Decoding (CAD)* balances the contribution of the parametric knowledge and contextual knowledge with a fixed factor: $p(y) = \text{softmax}(2 \cdot l(y|q,I,c) - l(y|q,I))$.
- *Contextual Information-Entropy Constraint Decoding (COIECD)* fuses the parametric knowledge and contextual knowledge with an additional constraint:

$$p(y) = \begin{cases} \text{softmax}(2 \cdot l(y|q,I,c) - l(y|q,I)), & \text{if } y \in C(y) \\ \text{softmax}(2 \cdot l(y|q,I) - l(y|q,I,c)), & \text{otherwise} \end{cases} \tag{A4}$$

  where C is a constraint computed from the entropy of the output probabilities.
- *Assembly of Global and Local Attention (AGLA)* contrasts the output logits with the augmented image and the original image: $p(y) = \text{softmax}(2 \cdot l(y|q,I^{aug},c) - l(y|q,I,c))$, where I^{aug} is the augmented image.
- *Visual Contrastive Decoding (VCD)* contrasts the output logits with the original image and the distorted image: $p(y) = \text{softmax}(2 \cdot l(y|q,I,c) - l(y|q,I^{dis},c))$, where I^{dis} is the distorted image.
- *Adaptive Logits Fusion and Attention Reallocation (ALFAR)* fuses the output logits with the original MLLM and the modified MLLM with attention reallocation: $p(y) = \text{softmax}(l(y|q,I,c) + \alpha \cdot l_{ar}(y|q,I^{dis},c))$, where α is the attention-based weighting factor.

A7 Adaptive Plausibility Constraints

Following prior studies (Shi et al., 2024a; Qiu et al., 2024; Leng et al., 2023), we adopt adaptive plausibility constraints (Li et al., 2023) to ensure fair comparisons. Calibrating the full output distribution can unintentionally penalize plausible tokens from the original distribution while elevating implausible ones from the modified distribution. To address this issue, we implement a selective calibration strategy by retaining only the tokens with high original probabilities and truncating the rest, as described below:

$$\begin{gathered} \mathcal{V}_{\text{token}}\left(y_{<i}\right) = \left\{y_i \in \mathcal{V} : p_\theta\left(y_i\right) \geq \beta \max_w p_\theta\left(w\right)\right\} \\ p\left(y_i\right) = 0, \text{ if } y_i \notin \mathcal{V}_{\text{token}}\left(y_{<i}\right) \end{gathered} \tag{A5}$$

where $\mathcal{V}_{\text{token}}$ is the set of selected tokens and $\mathcal{V}$ is the output vocabulary. Adaptive Plausibility Constraints are applied uniformly across all compared methods to ensure fair and consistent evaluation.

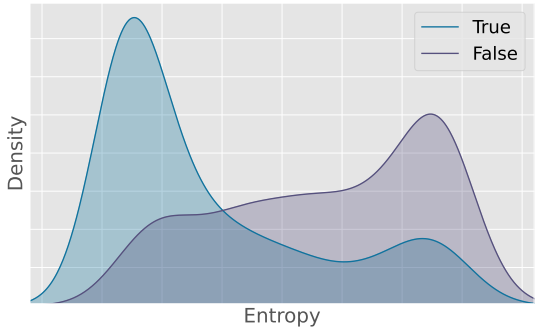

Figure A1: Entropy of the output logits for the true and false responses.

A8 Entropy-based Weighting

Inspired by the findings of prior works (Shi et al., 2024b; An et al., 2025), we utilize the entropy of the output logits as an indicator to assess output quality. As illustrated in Fig. A1, samples yielding correct responses tend to exhibit low entropy, indicating that the model is confident in its predictions, and reflected by a peaked, non-uniform logit distribution. In contrast, samples with incorrect responses typically show high entropy, as the model's uncertainty results in a more uniform distribution over the output logits.

A9 Experiments with Larger MLLMs

Tab. A4 reports the experimental results of the 13B variants of LLaVA-1.5 (Liu et al., 2023) and InstructBLIP (Dai et al., 2023) evaluated on the multi-choice InfoSeek dataset (Chen et al., 2023b; Zhu et al., 2024). The results show that the proposed framework consistently yields performance gains for both models, highlighting its effectiveness and scalability as model size increases.

Table A4: Experimental results with larger LVLMs.

Decoding	LLaVA-13B	InstructBLIP-13B
Parametric	44.40	8.93
Regular	55.83	18.70
Ours	**61.03**	**28.62**

A10 Knowledge Retrieval with different retrievers

We investigate the impact of various retrievers and retrieval strategies on both recall and model performance using the InfoSeek dataset (Chen et al., 2023b; Zhu et al., 2024) with LLaVA-1.5 7B (Liu et al., 2023). As presented in Tab. A5, our framework consistently outperforms baselines across all configurations, demonstrating its robustness to retrieval noise. Notably, retrieval based solely on the input question yields low recall, primarily due to the query's limited information regarding entities present in the images. Nevertheless, our framework maintains strong performance even under low-recall conditions, which is consistent with the conclusion in previous analysis.

Table A5: Experimental results with different retrievers. '-T' means using the input question for knowledge retrieval.

Retriever	Recall	LLaVA	Ours
CLIP-B/16 Radford et al. (2021)	38.27	45.93	**51.74**
CLIP-L/14 Radford et al. (2021)	56.53	51.23	**57.96**
CLIP-L/14-336 Radford et al. (2021)	58.37	51.97	**59.23**
CLIP-B/16-T Radford et al. (2021)	5.56	32.97	**35.80**
CLIP-L/14-T Radford et al. (2021)	6.00	33.40	**36.28**
CLIP-L/14-336-T Radford et al. (2021)	5.73	33.37	**37.19**

A11 Experimental Results with More MLLMs

To supplement the experimental results omitted from Tab. 1 and Tab. 2 in the main paper, we provide additional evaluations on a broader range of MLLMs. Specifically, we report results on generative free-form VQA datasets using Shikra and MiniGPT-4 in Tab. A6, and present results on discriminative multi-choice VQA datasets using LLaVA-1.5 and InstructBLIP in Tab. A7. These additional experiments further demonstrate the effectiveness and generalizability of our proposed method across different model architectures, evaluation paradigms, and VQA task formats.

Table A6: VQA Accuracy comparison on generative freeform VQA datasets over three runs.

Model	Decoding	Human			Validation		
		Unseen Question	Unseen Entity	Overall	Unseen Question	Unseen Entity	Overall
Shikra	Regular	$6.71_{(\pm 0.03)}$	$6.31_{(\pm 0.01)}$	$6.50_{(\pm 0.02)}$	$11.93_{(\pm 0.01)}$	$11.78_{(\pm 0.01)}$	$11.85_{(\pm 0.01)}$
	Parametric	$5.76_{(\pm 0.10)}$	$6.10_{(\pm 0.07)}$	$5.92_{(\pm 0.05)}$	$7.61_{(\pm 0.01)}$	$6.25_{(\pm 0.01)}$	$6.86_{(\pm 0.01)}$
	CD	$8.21_{(\pm 0.00)}$	$7.15_{(\pm 0.01)}$	$7.64_{(\pm 0.01)}$	$12.41_{(\pm 0.00)}$	$11.89_{(\pm 0.00)}$	$12.14_{(\pm 0.00)}$
	AdaCAD	$8.30_{(\pm 0.00)}$	$7.11_{(\pm 0.00)}$	$7.66_{(\pm 0.00)}$	$12.87_{(\pm 0.03)}$	$12.53_{(\pm 0.02)}$	$12.70_{(\pm 0.02)}$
	Entropy	$8.32_{(\pm 0.03)}$	$7.73_{(\pm 0.11)}$	$8.01_{(\pm 0.05)}$	$13.78_{(\pm 0.02)}$	$13.33_{(\pm 0.01)}$	$13.55_{(\pm 0.02)}$
	CAD	$8.16_{(\pm 0.06)}$	$7.16_{(\pm 0.02)}$	$7.62_{(\pm 0.00)}$	$12.99_{(\pm 0.03)}$	$12.51_{(\pm 0.02)}$	$12.75_{(\pm 0.03)}$
	COIECD	$8.32_{(\pm 0.02)}$	$7.73_{(\pm 0.08)}$	$8.01_{(\pm 0.03)}$	$14.46_{(\pm 0.02)}$	$14.21_{(\pm 0.03)}$	$14.33_{(\pm 0.02)}$
	AGLA	$8.24_{(\pm 0.01)}$	$7.56_{(\pm 0.05)}$	$7.88_{(\pm 0.01)}$	$14.29_{(\pm 0.02)}$	$13.91_{(\pm 0.01)}$	$14.08_{(\pm 0.01)}$
	VCD	$8.13_{(\pm 0.05)}$	$7.41_{(\pm 0.03)}$	$7.75_{(\pm 0.04)}$	$13.71_{(\pm 0.03)}$	$13.81_{(\pm 0.03)}$	$13.76_{(\pm 0.03)}$
	ALFAR	$8.61_{(\pm 0.01)}$	$8.04_{(\pm 0.01)}$	$8.31_{(\pm 0.01)}$	$15.25_{(\pm 0.00)}$	$15.11_{(\pm 0.01)}$	$15.18_{(\pm 0.01)}$
	Ours	$\mathbf{9.32}_{(\pm 0.01)}$	$\mathbf{8.84}_{(\pm 0.01)}$	$\mathbf{9.07}_{(\pm 0.01)}$	$\mathbf{16.13}_{(\pm 0.02)}$	$\mathbf{15.96}_{(\pm 0.01)}$	$\mathbf{16.04}_{(\pm 0.02)}$
MiniGPT4	Regular	$4.38_{(\pm 0.07)}$	$3.00_{(\pm 0.02)}$	$3.56_{(\pm 0.02)}$	$12.69_{(\pm 0.02)}$	$12.38_{(\pm 0.02)}$	$12.53_{(\pm 0.02)}$
	Parametric	$2.34_{(\pm 0.01)}$	$2.10_{(\pm 0.01)}$	$2.21_{(\pm 0.00)}$	$4.72_{(\pm 0.01)}$	$3.93_{(\pm 0.01)}$	$4.29_{(\pm 0.01)}$
	CD	$4.28_{(\pm 0.00)}$	$2.59_{(\pm 0.00)}$	$3.22_{(\pm 0.00)}$	$14.49_{(\pm 0.01)}$	$14.43_{(\pm 0.00)}$	$14.46_{(\pm 0.01)}$
	AdaCAD	$4.78_{(\pm 0.00)}$	$3.43_{(\pm 0.01)}$	$3.99_{(\pm 0.00)}$	$14.82_{(\pm 0.02)}$	$14.96_{(\pm 0.02)}$	$14.89_{(\pm 0.02)}$
	Entropy	$4.80_{(\pm 0.07)}$	$2.91_{(\pm 0.00)}$	$3.62_{(\pm 0.00)}$	$14.66_{(\pm 0.02)}$	$14.66_{(\pm 0.01)}$	$14.66_{(\pm 0.02)}$
	CAD	$4.97_{(\pm 0.01)}$	$3.44_{(\pm 0.01)}$	$4.07_{(\pm 0.01)}$	$14.83_{(\pm 0.01)}$	$14.94_{(\pm 0.00)}$	$14.88_{(\pm 0.01)}$
	COIECD	$4.57_{(\pm 0.03)}$	$3.40_{(\pm 0.01)}$	$3.90_{(\pm 0.00)}$	$14.87_{(\pm 0.01)}$	$14.67_{(\pm 0.02)}$	$14.77_{(\pm 0.02)}$
	AGLA	$4.67_{(\pm 0.02)}$	$3.63_{(\pm 0.02)}$	$4.09_{(\pm 0.01)}$	$14.31_{(\pm 0.02)}$	$13.92_{(\pm 0.01)}$	$14.11_{(\pm 0.01)}$
	VCD	$4.52_{(\pm 0.03)}$	$3.43_{(\pm 0.01)}$	$3.90_{(\pm 0.01)}$	$14.46_{(\pm 0.01)}$	$14.30_{(\pm 0.02)}$	$14.38_{(\pm 0.02)}$
	ALFAR	$5.05_{(\pm 0.02)}$	$3.87_{(\pm 0.01)}$	$4.38_{(\pm 0.00)}$	$15.16_{(\pm 0.01)}$	$15.27_{(\pm 0.01)}$	$15.05_{(\pm 0.01)}$
	Ours	$\mathbf{5.24}_{(\pm 0.06)}$	$\mathbf{4.27}_{(\pm 0.03)}$	$\mathbf{4.71}_{(\pm 0.05)}$	$\mathbf{15.71}_{(\pm 0.01)}$	$\mathbf{15.68}_{(\pm 0.01)}$	$\mathbf{15.69}_{(\pm 0.01)}$

Table A7: VQA Accuracy comparison on discriminative multi-choice VQA datasets over three runs.

Model	Decoding	InfoSeek	ViQuAE	Model	Decoding	InfoSeek	ViQuAE
LLaVA-1.5	Regular	$51.97_{(\pm 0.42)}$	$53.32_{(\pm 0.20)}$	InstructBLIP	Regular	$23.44_{(\pm 0.89)}$	$19.82_{(\pm 0.12)}$
	Parametric	$39.15_{(\pm 0.02)}$	$51.06_{(\pm 0.16)}$		Parametric	$8.73_{(\pm 0.23)}$	$6.53_{(\pm 0.35)}$
	CD	$49.95_{(\pm 0.20)}$	$52.56_{(\pm 0.03)}$		CD	$22.95_{(\pm 0.63)}$	$21.76_{(\pm 0.51)}$
	CAD	$52.08_{(\pm 0.16)}$	$52.99_{(\pm 0.23)}$		CAD	$26.56_{(\pm 0.56)}$	$23.18_{(\pm 0.30)}$
	AdaCAD	$52.30_{(\pm 0.04)}$	$52.99_{(\pm 0.30)}$		AdaCAD	$27.07_{(\pm 0.05)}$	$23.64_{(\pm 0.08)}$
	Entropy	$53.33_{(\pm 0.07)}$	$54.26_{(\pm 0.05)}$		Entropy	$27.50_{(\pm 0.07)}$	$23.19_{(\pm 0.39)}$
	COIECD	$52.08_{(\pm 0.21)}$	$52.99_{(\pm 0.23)}$		COIECD	$25.65_{(\pm 0.05)}$	$20.84_{(\pm 0.05)}$
	VCD	$53.87_{(\pm 0.07)}$	$55.13_{(\pm 0.09)}$		VCD	$23.31_{(\pm 0.07)}$	$20.28_{(\pm 0.53)}$
	AGLA	$53.53_{(\pm 0.50)}$	$54.24_{(\pm 0.21)}$		AGLA	$21.24_{(\pm 0.55)}$	$16.77_{(\pm 0.30)}$
	ALFAR	$58.35_{(\pm 0.21)}$	$55.91_{(\pm 0.13)}$		ALFAR	$35.65_{(\pm 0.09)}$	$24.11_{(\pm 0.07)}$
	Ours	$\mathbf{59.23}_{(\pm 0.21)}$	$\mathbf{58.94}_{(\pm 0.07)}$		**Ours**	$\mathbf{42.93}_{(\pm 0.07)}$	$\mathbf{28.17}_{(\pm 0.07)}$