

A1 Limitation and Future Work

Despite its strong performance in enhancing knowledge utilization for MLLMs through adaptive logits fusion and attention reallocation, our approach has certain limitations. First, our framework requires access to MLLM parameters, making it inapplicable to black-box API-based models such as GPT-4 [63]. Extending our framework to black-box MLLMs represents a promising direction for future research. Additionally, we observe that MLLMs struggle to effectively extract relevant information from long contexts. Addressing this limitation by improving MLLMs’ ability to leverage extended contexts will be another focus of future work.

A2 Broader Impacts

The proposed model for enhancing knowledge utilization in MLLMs carries significant broader impacts. First, by addressing the critical issue of MRAG, our method enhances the reliability and trustworthiness of MLLMs. This improvement is essential for deploying these models in sensitive and high-stakes applications such as autonomous driving, medical diagnostics, and surveillance systems. Second, the insights and methods introduced in this paper contribute to the broader field of MLLMs, particularly in understanding and improving the knowledge utilization mechanisms within MLLMs. This advancement can spur further research and innovation in integrating visual and textual data, leading to more robust and versatile AI models.

A3 Conflict Rate and Performance Drop

To quantify the conflict between parametric and contextual knowledge and its impact on model performance, we introduce two metrics: *Conflict Rate* and *Performance Drop*.

Conflict Rate measures the proportion of instances where parametric and contextual knowledge provide different information, and *Performance Drop* quantifies the decline in model performance due to knowledge conflict. Since parametric knowledge is implicitly embedded in model parameters and is not directly observable, we approximate its correctness by evaluating the model’s outputs. Specifically, if the model (without external context) produces the correct answer, we assume its parametric knowledge is correct; otherwise, it is considered incorrect. Given access to ground-truth contextual knowledge, the *Conflict Rate* can be defined as the error rate of parametric knowledge, i.e., the proportion of incorrect responses generated by the vanilla model without input context:

$$\text{Conflict Rate} = \text{Err}(\mathcal{M}_\theta(y|q, I), \hat{y}) \quad (\text{A1})$$

where Err is a function that calculates the error rate of the output, $\mathcal{M}_\theta(y|q, I)$ is the output with only images and questions as inputs, and \hat{y} is the ground-truth answer.

When correct contextual knowledge is available, the ideal model should achieve 100% accuracy in the absence of knowledge conflicts. However, influenced by knowledge conflicts, the model cannot achieve 100% accuracy, then we can define *Performance Drop* as the error rate of outputs when both parametric and contextual knowledge are used:

$$\text{Performance Drop} = \text{Err}(\mathcal{M}_\theta(y|q, I, c), \hat{y}) \quad (\text{A2})$$

where $\mathcal{M}_\theta(y|q, I, c)$ is the output with ground-truth context as additional inputs.

A4 Retrieval Recall

To investigate the retrieval recall from different retrieval rankings, we present the recall with Ground-Truth knowledge and knowledge from various retrieval rankings on the multi-choice InfoSeek dataset [20, 50] in Table A1. The low recall negatively impacts performance on knowledge-intensive VQA tasks, highlighting the necessity of developing a more effective retriever.

A5 Dataset

We present statistics of different datasets and the corresponding knowledge bases in Table A2. Specifically, for Validation [20] and InfoSeek [50], we follow previous works [24] and adopt a

Table A1: Retrieval recall with Ground-Truth knowledge (GT) and knowledge from different retrieval rankings on the multi-choice InfoSeek dataset [20, 50].

Index	GT	1	2	3	4
Recall	100	58.37	10.57	5.07	3.07

Table A2: Statistics of the datasets and details of the knowledge bases used.

Dataset	# VQA pairs	Knowledge Base
Validation [20]	73,620	Wikipedia [20]
Human [20]	8,931	Wikipedia [20]
InfoSeek [50]	3,000	Wikipedia [20]
ViQuAE [50]	3,000	Wikipedia [20]
OK-VQA [56]	5,046	GPT-3.5 [18]
AOK-VQA [57]	1,145	GPT-3.5 [18]
E-VQA [28]	700	Encyclopedia [28]

knowledge base containing 1.7K entities derived from the original Wikipedia knowledge base [20]. For Human [20] and ViQuAE [21], we use the original Wikipedia knowledge base [20], selecting 73.6K entities accompanied by images for knowledge retrieval. For OK-VQA [56] and AOK-VQA [57], we utilize the knowledge base provided by [18], which was generated using GPT-3.5 [64]. For E-VQA [28], we select templated questions with images from the iNaturalist dataset [65] and use the corresponding ground-truth knowledge for inference. All evaluations are conducted using the official scripts.

A6 Attention Distribution for Context Tokens

We visualize the attention distribution toward context tokens with LLaVA-1.5 [13] in Fig. A1. From the figure we can see that MLLMs tend to allocate attention uniformly across context tokens without prioritization, which dilutes the contributions of query-relevant knowledge and tends to introduce inaccurate MLLM responses as well.

A7 Adaptive Plausibility Constraints

We follow prior studies [23, 46, 52] and adopt adaptive plausibility constraints [44] for fair comparisons. Specifically, calibrating the entire output distribution may penalize valid outputs from the original distribution and promote implausible outputs from the modified distribution. To mitigate this issue, we selectively consider tokens with high original probabilities and truncate other tokens as follows:

$$\begin{aligned} \mathcal{V}_{\text{token}}(y_{<i}) &= \{y_i \in \mathcal{V} : p_{\theta}(y_i) \geq \beta \max_w p_{\theta}(w)\} \\ p(y_i) &= 0, \text{ if } y_i \notin \mathcal{V}_{\text{token}}(y_{<i}) \end{aligned} \tag{A3}$$

where $\mathcal{V}_{\text{token}}$ is the set of selected tokens and \mathcal{V} is the output vocabulary. We select $\beta = 0.7$ to retain only high-probability tokens.

A8 Effect of Intervention Layers

We investigate the impact of attention reallocation at different MLLM layers on the InfoSeek dataset [20] using LLaVA-1.5 [13], as summarized in Table A4. The results show that reallocating attention in shallow layers (layers 1-16) enhances model performance by mitigating attention bias toward image tokens, thereby improving the extraction of low-level features [27]. In contrast, applying attention reallocation in middle layers (layers 17-24) yields smaller gains, as these layers primarily

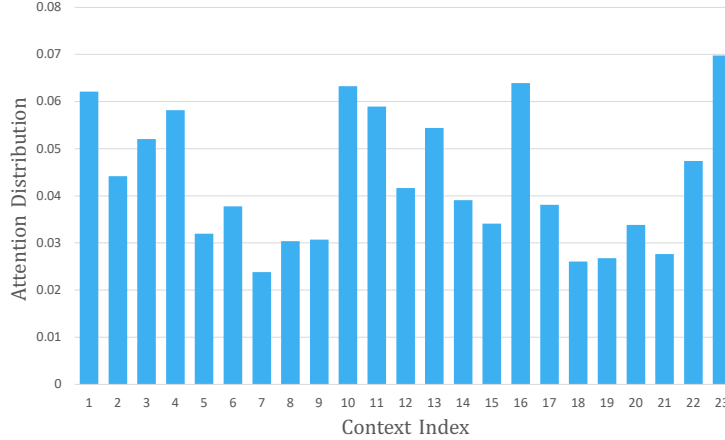


Figure A1: Attention weight distribution toward different context tokens with LLaVA-1.5 [13].

Table A4: An ablation study with different layers for attention reallocation.

Intervention Layers	VQA Accuracy
None	56.70
[1 – 8]	58.08
[9 – 16]	58.17
[17 – 24]	57.57
[25 – 32]	58.10
[1 – 8] \cup [32]	58.67

937 handle multimodal alignment and feature aggregation [66], where attention bias is less severe. Notably,
 938 reallocating attention in late layers (layers 25-32) leads to the most substantial performance gains, as
 939 these layers are responsible for reasoning and directly affect output generation [66]. Furthermore,
 940 leveraging attention reallocation in both shallow (layers 1–8) and deep (layer 32) layers yields the
 941 best performance.

942 A9 Effect of different numbers of knowledge

943 We examine the effect of varying the amount
 944 of knowledge provided to MLLMs on retrieval
 945 recall and model performance on the InfoSeek
 946 dataset [20]. As shown in Table A3, appending
 947 additional knowledge to the prompt improves
 948 retrieval recall but has limited impact on model
 949 performance, as MLLMs often struggle to ef-
 950 fectively utilize information from lengthy input
 951 contexts [67]. Our model addresses this limita-
 952 tion by guiding MLLMs based on query-context
 953 relevance. However, the modest performance gains underscore the need for future research on
 954 enhancing MLLMs’ ability to process and leverage extended contexts.

Table A3: Experimental results with different numbers of knowledge using LLaVA-1.5 [13].

#Knowledge	Recall	LLaVA [13]	Ours
1	58.37	51.97	58.35
2	68.93	49.90	58.70
3	74.00	50.13	58.57
4	76.77	50.23	58.20

955 A10 Adaptive Plausible Constraint Factor

956 We investigate the influence of the adaptive plausible constraint factor β in Eq. A3 on the multi-choice
 957 InfoSeek [20] with LLaVA-1.5 [13] in Table A6. Larger β indicates keeping only high-probability
 958 tokens. The results illustrate that our model is robust to the change of β .

Table A6: An ablation study on different constraint factors β

β	0.4	0.5	0.6	0.7	0.8
Accuracy	56.88	57.50	58.20	58.35	58.33

A11 Effect of the scaled factor

We investigate the influence of the scaled factor k for α and ω_j on the multi-choice InfoSeek [20] with LLaVA-1.5 [13] in Table A5. Larger k means a more severe penalty for the image tokens and a higher enhancement for the context tokens. The results illustrate that our model is robust to the change of k .

Table A5: An ablation study on different scaled factors k

k	0.2	0.3	0.4	0.5	0.6
Accuracy	58.19	58.22	58.35	58.92	58.29

A12 Attention Distribution over All Layers

We visualize the distribution of attention assigned to image and context tokens across all layers of LLaVA-1.5 [13], both before and after attention reallocation, in Fig. A2 and Fig. A3, respectively. The results indicate that the proposed attention reallocation effectively mitigates attention bias and enhances MLLMs' focus on contextual knowledge.

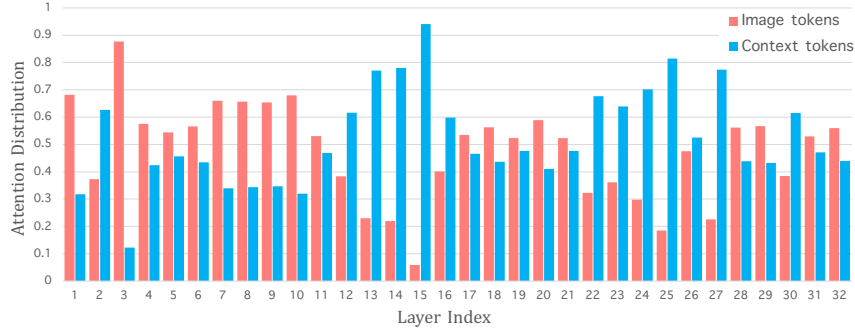


Figure A2: Proportions of attention that are assigned to image and context tokens at all layers.

A13 Prompt Templates for MLLMs

Tables A7 and A8 present the prompt templates for free-form and multi-choice questions across different MLLMs. For vanilla models, prompts do not include contextual knowledge, whereas for models with MRAG, prompts are concatenated with the retrieved knowledge.

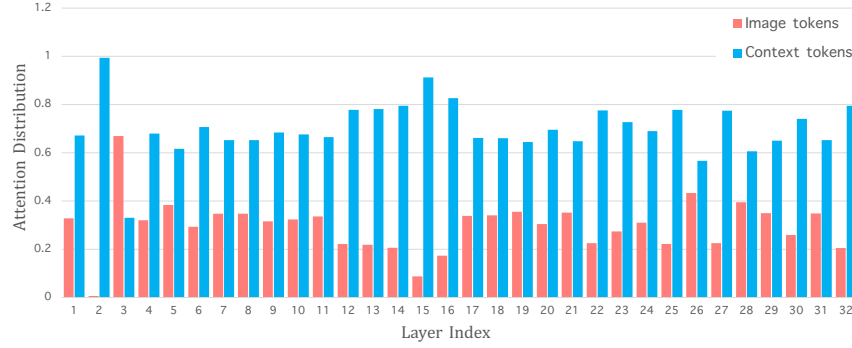


Figure A3: Proportions of attention that are assigned to image and context tokens at all layers of LLaVA-1.5 [13] after attention reallocation.

Prompt for LLaVA [13]

Vanilla LLaVA for free-form questions:

<Image>, <Question>. Answer the question using a single word based on your knowledge.

LLaVA with MRAG for free-form questions:

<Image>, <Question>. Answer the question using a single word based on the given context.
Context: <Context>.

Vanilla LLaVA for multi-choice questions:

<Image>, <Question>. Answer the question using a single word based on your knowledge.
Option: <Option>.

LLaVA with MRAG for multi-choice questions:

<Image>, <Question>. Answer the question using a single word or phrase based on the given context. Context: <Context>. Option: <Option>.

Prompt for InstructBLIP [10]

Vanilla InstructBLIP for free-form questions:

<Image>. You are an expert at question answering. Given the question, please output the answer. No explanation or further questions. Question: <Question>. Short answer:

InstructBLIP with MRAG for free-form questions:

<Image>. You are an expert at question answering. Given the question, please output the answer. No explanation or further questions. Question: <Question>. Context: <Context>. Short answer:

Vanilla InstructBLIP for multi-choice questions:

<Image>. You are an expert at question answering. Given the question, please output the answer. No explanation or further questions. Question: <Question>. Option: <Option>. Short answer:

InstructBLIP with MRAG for multi-choice questions:

<Image>. You are an expert at question answering. Given the question, please output the answer. No explanation or further questions. Question: <Question>. Context: <Context>. Option: <Option>. Short answer:

Table A7: The input prompts used for LLaVA [13] and InstructBLIP [10] to generate responses on free-form and multi-choice datasets.

Prompt for Shikra [36]

Vanilla Shikra for free-form questions:

<Image>. You are an expert at question answering. Given the question, please output the answer. No explanation or further questions. Question: <Question>. Short answer:

Shikra with MRAG for free-form questions:

<Image>. You are an expert at question answering. Given the question, please output the answer. No explanation or further questions. Question: <Question>. Context: <Context>. Short answer:

Vanilla Shikra for multi-choice questions:

<Image>. You are an expert at question answering. Given the question, please output the answer. No explanation or further questions. Question: <Question>. Option: <Option>. The answer should be a single letter from A, B, C or D.

Shikra with MRAG for multi-choice questions:

<Image>. You are an expert at question answering. Given the question, please output the answer. No explanation or further questions. Question: <Question>. Context: <Context>. Option: <Option>. The answer should be a single letter from A, B, C or D.

Prompt for MiniGPT4 [15]

Vanilla MiniGPT4 for free-form questions:

<Image>. You are an expert at question answering. Given the question, please output the answer. No explanation or further questions. Question: <Question>. Short answer with one word:

MiniGPT4 with MRAG for free-form questions:

<Image>. You are an expert at question answering. Given the question, please output the answer. No explanation or further questions. Question: <Question>. Context: <Context>. Short answer with one word:

Vanilla MiniGPT4 for multi-choice questions:

<Image>. You are an expert at question answering. Given the question, please output the answer. No explanation or further questions. Question: <Question>. Option: <Option>. Short answer with one word:

MiniGPT4 with MRAG for multi-choice questions:

<Image>. You are an expert at question answering. Given the question, please output the answer. No explanation or further questions. Question: <Question>. Context: <Context>. Option: <Option>. Short answer with one word:

Table A8: The input prompts used for Shikra [36] and MiniGPT4 [15] to generate responses on free-form and multi-choice datasets.