Boosting Knowledge Utilization in Multimodal Large Language Models via Adaptive Logits Fusion and Attention Reallocation

Wenbin An[1,2,*] **Jiahao Nie**[3,*] **Feng Tian**[1,2,†] **Haonan Lin**[1,2] **Mingxiang Cai**[4]
Yaqiang Wu[4] **Qianying Wang**[4,†] **Xiaoqin Zhang**[5] **Shijian Lu**[3,†]

[1]Xi'an Jiaotong University [2]National Engineering Laboratory for Big Data Analytics
[3]Nanyang Technological University [4]Lenovo Research [5]Zhejiang University of Technology
wenbinan@stu.xjtu.edu.cn, jiahao007@e.ntu.edu.sg
fengtian@mail.xjtu.edu.cn, wangqya@lenovo.com, shijian.lu@ntu.edu.sg

Abstract

Despite their recent progress, Multimodal Large Language Models (MLLMs) often struggle in knowledge-intensive tasks due to the limited and outdated parametric knowledge acquired during training. Multimodal Retrieval Augmented Generation addresses this issue by retrieving contextual knowledge from external databases, thereby enhancing MLLMs with expanded knowledge sources. However, existing MLLMs often fail to fully leverage the retrieved contextual knowledge for response generation. We examine representative MLLMs and identify two major causes, namely, attention bias toward different tokens and knowledge conflicts between parametric and contextual knowledge. To this end, we design *Adaptive Logits Fusion and Attention Reallocation (ALFAR)*, a training-free and plug-and-play approach that improves MLLM responses by maximizing the utility of the retrieved knowledge. Specifically, ALFAR tackles the challenges from two perspectives. First, it alleviates attention bias by adaptively shifting attention from visual tokens to relevant context tokens according to query-context relevance. Second, it decouples and weights parametric and contextual knowledge at output logits, mitigating conflicts between the two types of knowledge. As a plug-and-play method, ALFAR achieves superior performance across diverse datasets without requiring additional training or external tools. Extensive experiments over multiple MLLMs and benchmarks show that ALFAR consistently outperforms the state-of-the-art by large margins. Our code and data are available at https://github.com/Lackel/ALFAR.

1 Introduction

Building upon powerful Large Language Models (LLMs) [1, 2, 3, 4, 5, 6, 7, 8], Multimodal Large Language Models (MLLMs) [9, 10, 11, 12, 13, 14, 15, 16, 17] have achieved impressive performance over a wide range of vision-centric tasks such as image captioning [18, 19, 20], visual question answering [21, 22], *etc*. Nevertheless, MLLMs often struggle to handle knowledge-intensive vision-language tasks [23, 24], primarily due to the limited and outdated parametric knowledge acquired during training [25, 26]. Multimodal Retrieval Augmented Generation (MRAG) [27, 28, 29], a prevalent approach that attempts to resolve this issue, retrieves contextual knowledge from external

[*]Equal contribution
[†]Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

data to empower MLLMs for accurate response generation. However, the way of exploiting the contextual knowledge remains under-explored, undermining the effectiveness of MRAG.

We examined representative MLLMs with MRAG and found that while MRAG can improve MLLM performance when high-quality contextual knowledge is retrieved (as shown in Fig. 1), MLLMs often fail to make full use of the retrieved knowledge, even when ground-truth knowledge is available. We identify two primary causes, namely, attention bias among visual and context tokens and conflicts between MLLMs' parametric knowledge and retrieved contextual knowledge. For the attention bias, MLLMs tend to allocate more attention to image tokens over context tokens, especially in shallow layers that are critical for knowledge extraction and exchange [30]. Since images often do not provide sufficient information for knowledge-intensive questions [23, 31], the attention bias hinders the effective utilization of contextual knowledge and leads to inaccurate MLLM responses. In addition, MLLMs allocate attention uniformly across context tokens without prioritization, which dilutes the contributions of query-relevant knowledge and tends to introduce inaccurate MLLM responses.

Knowledge conflicts typically arise from the discrepancy between contextual and parametric knowledge. We observe that MLLMs tend to rely excessively on their parametric knowledge even when accurate contextual knowledge is present, leading to under-utilization of contextual knowledge and counterfactual responses. Such a phenomenon is well aligned with observations in previous LLM studies [33, 34] and findings in psychology research [35, 36], both underscoring a clear preference toward intrinsic instead of retrieved knowledge. On the other end, the preference for the parametric knowledge does help when the contextual knowledge is unreliable [28, 37, 38]. This can be observed in Fig. 1, where low-quality contextual knowledge significantly degrades performance. Therefore, striking a balance between parametric and contextual knowledge while leveraging their complementary strengths is critical for generating accurate responses.

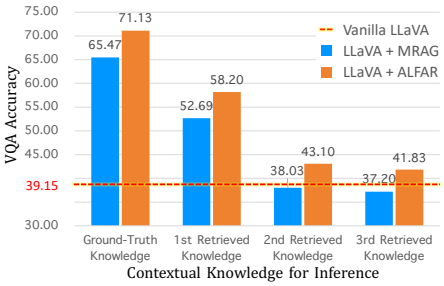

Figure 1: VQA accuracy of LLaVA-1.5 [14] on the multi-choice InfoSeek dataset [32] with respect to the quality of contextual knowledge. ALFAR fully exploits the retrieved knowledge, consistently improving MRAG performance regardless of knowledge quality.

In this work, we propose *Adaptive Logits Fusion and Attention Reallocation (ALFAR)*, a training-free and plug-and-play approach that enables effective utilization of MRAG-retrieved contextual knowledge for accurate MLLM responses. ALFAR addresses attention bias and knowledge conflicts by dynamically adjusting attention allocation and balancing the parametric and contextual knowledge, respectively. Specifically, ALFAR adaptively shifts attention from image tokens to relevant context tokens based on retrieval scores and query-context relevance, enabling MLLMs to focus on more pertinent information. In addition, ALFAR decouples parametric and contextual knowledge at output logits and weights them according to the attention distribution, enabling a balanced and synergistic integration of the two types of knowledge. Extensive experiments across multiple representative MLLMs and benchmarks demonstrate ALFAR's superior and broad applicability without involving additional training or external tools.

The contributions of this work can be summarized in three major aspects. **First**, we dive deeply into knowledge utilization in MLLMs, identifying attention bias and knowledge conflicts as two key factors that impede the effective utilization of the retrieved knowledge. These findings provide valuable insights for advancing knowledge utilization in MLLMs. **Second**, we design ALFAR, a training-free and plug-and-play approach that reallocates attention and balances parametric and contextual knowledge effectively. **Third**, Extensive experiments over multiple generative and discriminative benchmarks validate ALFAR's effectiveness and versatility, demonstrating its superior performance and broad applicability across various multimodal tasks.

2 Related Work

2.1 Multimodal Large Language Models

The rapid advancements in Large Language Models (LLMs) [1, 2, 3, 4, 5, 6, 7, 8, 39] have greatly propelled the development of Multimodal Large Language Models (MLLMs) [9, 10, 11, 13, 14, 15,

16, 40]. To align visual and textual modalities, prior studies explore different approaches such as visual encoders with linear projectors [14, 41, 40], Q-former [18, 10], and Perceivers [42], which transform image patches into visual tokens that are compatible with LLMs. Most MLLMs conduct training in two stages, namely, pre-training for feature alignment and instruction-based fine-tuning [14, 43, 10], enabling impressive performance across diverse multimodal tasks [44, 45, 46, 47]. Despite these advancements, MLLMs often face challenges in knowledge-intensive tasks [23, 31], due to the limitations of their parametric knowledge acquired during training.

2.2 Multimodal Retrieval Augmented Generation

Inspired by the concept of Retrieval Augmented Generation (RAG) for LLMs [48, 49], Multimodal Retrieval Augmented Generation (MRAG) has been widely explored for enhancing MLLMs with more comprehensive and up-to-date knowledge [27, 28, 29, 50]. MRAG retrieves relevant knowledge from a multimodal database and incorporates the retrieved knowledge as the context of the input. For instance, Wiki-LLaVA [27] broadens the knowledge scope of LLaVA [14] by incorporating the retrieved Wikipedia articles in training. EchoSight [29] leverages a fine-tuned Q-Former [18] to filter retrieved knowledge and enhance retrieval recall. ReflectiVA [28] and MR^{2}AG [50] introduce a trainable reflection mechanism to assess the necessity of retrieval and the relevance of retrieved knowledge. In the LLM domain, several training-free methods have been proposed to better utilize the retrieved knowledge to enhance generation quality. For instance, CAD [26] employs contrastive decoding [51] to increase the faithfulness of generation. Moreover, AdaCAD [52], Entropy [53], and COIECD [54] extend CAD [26] by introducing JS divergence, entropy, and information constraints, respectively. Despite the improved faithfulness toward the retrieved context, these methods struggle to balance parametric and contextual knowledge [55], resulting in sub-optimal performance when the contextual knowledge is noisy.

3 Preliminary and Motivation

3.1 Multimodal Retrieval Augmented Generation

Given a textual query q and a query image I, an MLLM $\mathcal{M}_\theta$ parameterized by θ is expected to generate a reliable answer y. To enrich MLLMs with external knowledge, MRAG employs a multimodal retriever $\mathcal{R}_\phi$ to fetch relevant knowledge from a multimodal knowledge base $\mathcal{C} = \{(\widetilde{I}_i, c_i)\}_{i=1}^{M}$, where $\widetilde{I}_i$ and c_i represent an image and its corresponding textual knowledge, respectively. The retriever $\mathcal{R}_\phi$ measures the similarity between the query pair (q, I) and a multimodal knowledge pair $(\widetilde{I}, c)$ based on the cosine similarity between their image embeddings:

$$\alpha = \frac{\mathcal{R}_\phi(I) \cdot \mathcal{R}_\phi(\widetilde{I})}{||\mathcal{R}_\phi(I)|| \cdot ||\mathcal{R}_\phi(\widetilde{I})||} \tag{1}$$

The textual knowledge c with the highest retrieval similarity α is selected as the input context for MLLMs. Consequently, the output distributions of the MLLM with MRAG at the time step t are:

$$p(y_t) \sim \text{softmax}(\mathcal{M}_\theta(y_t|q, I, c, y_{<t})) \tag{2}$$

where $y_{<t}$ represents the sequence of generated tokens before the time step t.

3.2 Self-attention Mechanism in MLLMs

MLLMs generate responses auto-regressively using Transformer blocks [56]. Specifically, the input image, query, and context tokens are concatenated and projected into three distinct vectors: the query vector $\mathbf{Q}$, the key vector $\mathbf{K}$, and the value vector $\mathbf{V}$, through three linear layers, W_q, W_k, and W_v. The self-attention mechanism computes the relevance of each token to other tokens as follows:

$$\mathbf{A} = \frac{\mathbf{Q} \cdot \mathbf{K}^\top}{\sqrt{d}} + \mathbf{M} \tag{3}$$

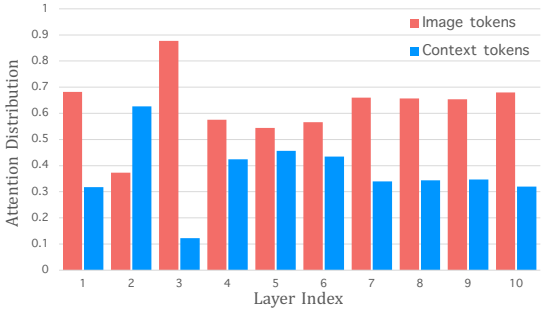

Figure 2: Proportions of attention weights that are assigned to image and context tokens at different shallow layers of LLaVA-1.5 [14].

where $\mathbf{A} \in \mathcal{R}^{n \times n}$ is the attention weight matrix, $\mathbf{M}$ is a causal mask, d is the feature dimension, and n is the number of input tokens. The output $\mathbf{O}$ can then be calculated by:

$$\mathbf{O} = \mathrm{softmax}(\mathbf{A}) \cdot \mathbf{V} \tag{4}$$

The attention weight matrix $\mathbf{A}$ reflects the importance of different input tokens in generating new tokens. This property makes it a valuable tool to analyze the contribution of different types of tokens to the generated responses.

3.3 Attention Bias in MRAG

To generate the response token y_{n+1}, MLLMs perform self-attention over all input tokens which correspond to the n-th row of the attention weight matrix $\mathbf{A}$. We analyze the contributions of the input image and context using an importance score, defined as the total attention weights assigned to these tokens. For the input image, the importance score at layer i is calculated by: $S^i(I) = \sum_{j \in I} A^i_{nj}$[1]. Similarly, for the input context, the importance score at layer i is determined by: $S^i(c) = \sum_{j \in c} A^i_{nj}$. As illustrated in Fig. 2, MLLMs tend to allocate more attention to image tokens than context tokens, particularly in shallow layers that are pivotal for extracting and exchanging information from distinct tokens [30]. This issue affects the effective utilization of contextual knowledge since images often do not capture sufficient information for knowledge-intensive questions [23, 31]. Moreover, MLLMs assign attention uniformly across different parts of the context without highlighting query-relevant segments. Such indiscriminate distribution increases the distraction of irrelevant knowledge, ultimately leading to inaccurate or misleading responses for MLLMs.

3.4 Knowledge Conflicts in MRAG

After knowledge retrieval, MLLMs integrate the retrieved contextual knowledge with their internal parametric knowledge to generate responses. However, similar to LLMs [33, 34, 52], MLLMs often encounter knowledge conflicts due to the discrepancy between the two types of knowledge, affecting the effectiveness of the model in various practical tasks. Worse still, MLLMs tend to prioritize their parametric knowledge even when perfect contextual knowledge is provided, leading to under-utilization of contextual knowledge and factually inconsistent answers. By assuming the accessibility of the ground-truth contextual knowledge, we evaluate the conflict rate of the two types of knowledge and the resultant performance degradation on the multi-choice InfoSeek [32] and ViQuAE [24, 32] datasets with LLaVA-1.5 [14]. As shown in Tab. 1, around half of the samples exhibit knowledge conflicts, which lead to an up to 30% performance drop, highlighting the necessity of mitigating such conflicts to enhance MLLM performance on knowledge-intensive tasks.

Table 1: Experiments with LLaVA-1.5 [14]. Conflict Ratio: Ratio of discrepancy between parametric and contextual knowledge. Performance drop: Accuracy decline due to knowledge conflicts. Details of the two metrics are provided in Appendix A3.

	Infoseek	**ViQuAE**
Conflict Ratio	60.85%	48.94%
Performance Drop	28.87%	27.02%

4 Method

The proposed framework consists of two branches for effective handling of parametric and contextual knowledge as illustrated in Fig. 4. Within the contextual branch, we design an attention reallocation mechanism that tackles the attention bias and improves the utilization of contextual knowledge by adaptively adjusting model attention toward relevant context tokens based on query-context relevance (Sec. 4.1). In addition, the network fuses the parametric and contextual knowledge adaptively in the output logits, mitigating knowledge conflicts under the guidance of the model attention that dynamically captures the relative importance of the two types of knowledge (Sec. 4.2).

4.1 Attention Reallocation

As analyzed in Sec. 3.3, the attention bias results from two major factors, namely, attention preference toward image tokens and uniform attention to context tokens. We address the attention preference

[1]We average all attention heads and omit the notation for simplicity.

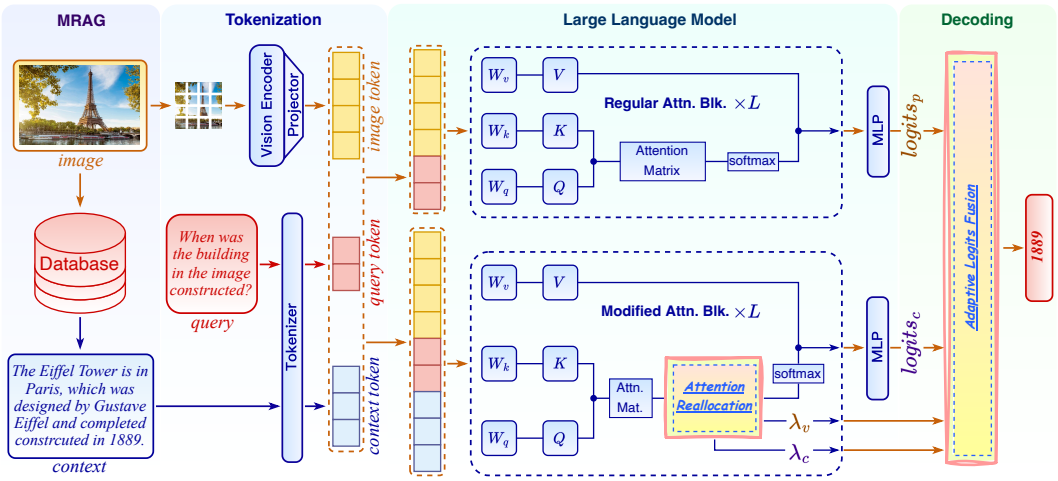

Figure 4: The overview of Adaptive Logits Fusion and Attention Reallocation (ALFAR).

by adaptively adjusting model attention as illustrated in Fig. 3, based on the retrieval similarity α in Eq. 1 that reflects the reliability of the retrieved context. Specifically, less attention is allocated to image tokens if the context is more reliable with a high retrieval similarity. The attention reallocation can be formulated as follows:

$$\hat{\mathbf{A}}_{ni} = (1-\beta)\cdot \mathbf{A}_{ni},\ \text{s.t. } i \in S_I \tag{5}$$

where $\beta = k \cdot \alpha$ is the scaled retrieval similarity with a scaling factor k. $\hat{\mathbf{A}}$ is the modified attention weight matrix, n is the number of all input tokens and S_I is the index set of image tokens. $\mathbf{A}_{ni}$ corresponds to the attention weight in the n-th row and i-th column of $\mathbf{A}$.

In addition, we introduce a query-context relevance score to mitigate the uniform attention to all context tokens and allow MLLMs to focus more on query-relevant context. We derive the relevance score from attention weights assigned to query tokens by j-th context token as follows:

$$\omega_j = \frac{\sum_{k\in S_q} \mathbf{A}_{jk}}{\sum_{l\in S_c}\sum_{k\in S_q} \mathbf{A}_{lk}},\ \text{s.t. } j \in S_c \tag{6}$$

where S_q and S_c are index sets of query and context tokens, respectively. With the relevance scores, we adaptively increase MLLMs' attention to context tokens:

$$\hat{\mathbf{A}}_{nj} = (1+\gamma_j)\cdot \mathbf{A}_{nj},\ \text{s.t. } j \in S_c \tag{7}$$

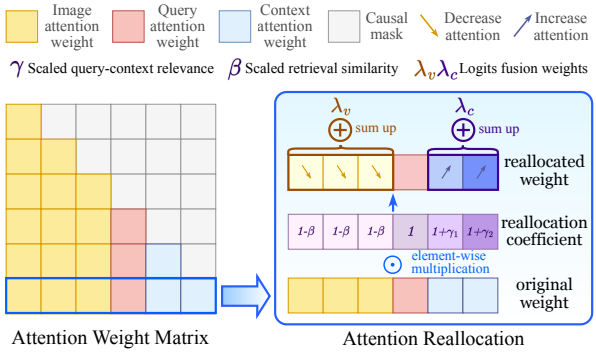

Figure 3: Attention reallocation by adjusting the last row of the attention weight matrix (i.e., $\mathbf{A_n}$).

where $\gamma_j = k \cdot \omega_j$ is the scaled query-context relevance with a scaling factor k. After attention reallocation, we apply softmax to redistribute the attention as in Eq. 4 to compute the output hidden states. This repeats auto-regressively for each subsequent token prediction.

4.2 Adaptive Knowledge Fusion

As analyzed in Sec. 3.4, parametric knowledge could hinder the utilization of contextual knowledge, leading to inaccurate responses. However, parametric knowledge brings benefits when the retrieved context is unreliable. Therefore, striking a balance between parametric and contextual knowledge based on their reliability is essential for generating accurate and reliable responses for MLLMs. Nevertheless, parametric knowledge is implicitly embedded and the two types of knowledge are entangled during inference, making it hard to explicitly represent and utilize them separately. We address this issue by disentangling the two types of knowledge and fuse them at output logits. Specifically, we represent parametric knowledge at step t by using the output logits that have only query and image as inputs:

$$\text{logit}_\text{p} = \mathcal{M}_\theta(y_t|q, I, y_{<t}) \tag{8}$$

Table 2: VQA Accuracy comparison on generative freeform VQA datasets over three runs. *Regular* and *Parametric* denote that MLLMs generate answers with and without retrieved knowledge, respectively. The best performance is marked in **bold**.

Model	Decoding	**Human** [23]			**Validation** [23]		
		Unseen Question	Unseen Entity	Overall	Unseen Question	Unseen Entity	Overall
LLaVA-1.5	Regular [14]	$7.59_{(\pm 0.08)}$	$7.90_{(\pm 0.05)}$	$7.74_{(\pm 0.01)}$	$19.98_{(\pm 0.02)}$	$19.59_{(\pm 0.01)}$	$19.78_{(\pm 0.01)}$
	Parametric [14]	$6.27_{(\pm 0.05)}$	$6.26_{(\pm 0.19)}$	$6.26_{(\pm 0.09)}$	$7.14_{(\pm 0.03)}$	$6.28_{(\pm 0.01)}$	$6.68_{(\pm 0.00)}$
	CD [51]	$7.39_{(\pm 0.21)}$	$7.31_{(\pm 0.06)}$	$7.35_{(\pm 0.09)}$	$20.32_{(\pm 0.01)}$	$19.90_{(\pm 0.00)}$	$20.11_{(\pm 0.01)}$
	AdaCAD [52]	$7.81_{(\pm 0.02)}$	$8.07_{(\pm 0.03)}$	$7.94_{(\pm 0.01)}$	$21.23_{(\pm 0.03)}$	$20.91_{(\pm 0.02)}$	$21.07_{(\pm 0.03)}$
	Entropy [53]	$7.98_{(\pm 0.09)}$	$8.34_{(\pm 0.02)}$	$8.15_{(\pm 0.03)}$	$21.97_{(\pm 0.02)}$	$21.85_{(\pm 0.01)}$	$21.91_{(\pm 0.03)}$
	CAD [26]	$8.02_{(\pm 0.05)}$	$8.15_{(\pm 0.03)}$	$8.08_{(\pm 0.02)}$	$21.68_{(\pm 0.01)}$	$20.93_{(\pm 0.02)}$	$21.30_{(\pm 0.02)}$
	COIECD [54]	$8.78_{(\pm 0.04)}$	$8.65_{(\pm 0.01)}$	$8.71_{(\pm 0.01)}$	$22.43_{(\pm 0.05)}$	$21.73_{(\pm 0.05)}$	$22.07_{(\pm 0.05)}$
	AGLA [57]	$8.74_{(\pm 0.28)}$	$9.18_{(\pm 0.00)}$	$8.94_{(\pm 0.08)}$	$22.34_{(\pm 0.02)}$	$21.88_{(\pm 0.01)}$	$22.11_{(\pm 0.02)}$
	VCD [58]	$9.22_{(\pm 0.00)}$	$9.26_{(\pm 0.00)}$	$9.24_{(\pm 0.02)}$	$22.30_{(\pm 0.03)}$	$22.38_{(\pm 0.03)}$	$22.23_{(\pm 0.03)}$
	ALFAR (ours)	$\mathbf{12.80}_{(\pm 0.09)}$	$\mathbf{11.24}_{(\pm 0.00)}$	$\mathbf{11.96}_{(\pm 0.02)}$	$\mathbf{23.82}_{(\pm 0.00)}$	$\mathbf{23.75}_{(\pm 0.01)}$	$\mathbf{23.78}_{(\pm 0.01)}$
InstructBLIP	Regular [10]	$4.20_{(\pm 0.00)}$	$3.86_{(\pm 0.03)}$	$4.02_{(\pm 0.01)}$	$3.60_{(\pm 0.01)}$	$3.82_{(\pm 0.00)}$	$3.71_{(\pm 0.00)}$
	Parametric [10]	$4.06_{(\pm 0.01)}$	$3.65_{(\pm 0.01)}$	$3.84_{(\pm 0.01)}$	$2.36_{(\pm 0.01)}$	$1.92_{(\pm 0.00)}$	$2.12_{(\pm 0.00)}$
	CD [51]	$4.52_{(\pm 0.03)}$	$3.55_{(\pm 0.01)}$	$3.98_{(\pm 0.01)}$	$3.59_{(\pm 0.01)}$	$4.00_{(\pm 0.00)}$	$3.79_{(\pm 0.00)}$
	AdaCAD [52]	$4.57_{(\pm 0.05)}$	$3.70_{(\pm 0.10)}$	$4.09_{(\pm 0.08)}$	$3.71_{(\pm 0.02)}$	$4.35_{(\pm 0.01)}$	$4.01_{(\pm 0.01)}$
	Entropy [53]	$4.56_{(\pm 0.05)}$	$4.14_{(\pm 0.01)}$	$4.34_{(\pm 0.02)}$	$3.81_{(\pm 0.01)}$	$4.39_{(\pm 0.00)}$	$4.08_{(\pm 0.00)}$
	CAD [26]	$4.52_{(\pm 0.03)}$	$3.55_{(\pm 0.01)}$	$3.98_{(\pm 0.01)}$	$3.77_{(\pm 0.03)}$	$4.43_{(\pm 0.02)}$	$4.08_{(\pm 0.02)}$
	COIECD [54]	$4.64_{(\pm 0.10)}$	$4.08_{(\pm 0.02)}$	$4.33_{(\pm 0.01)}$	$4.07_{(\pm 0.00)}$	$4.54_{(\pm 0.00)}$	$4.30_{(\pm 0.00)}$
	AGLA [57]	$4.80_{(\pm 0.05)}$	$4.28_{(\pm 0.09)}$	$4.52_{(\pm 0.05)}$	$3.74_{(\pm 0.01)}$	$4.10_{(\pm 0.01)}$	$3.91_{(\pm 0.01)}$
	VCD [58]	$4.70_{(\pm 0.01)}$	$4.14_{(\pm 0.01)}$	$4.40_{(\pm 0.00)}$	$3.62_{(\pm 0.01)}$	$4.12_{(\pm 0.00)}$	$3.85_{(\pm 0.00)}$
	ALFAR (ours)	$\mathbf{5.98}_{(\pm 0.00)}$	$\mathbf{5.67}_{(\pm 0.00)}$	$\mathbf{5.82}_{(\pm 0.00)}$	$\mathbf{4.55}_{(\pm 0.00)}$	$\mathbf{5.68}_{(\pm 0.00)}$	$\mathbf{5.05}_{(\pm 0.00)}$
Shikra	Regular [40]	$6.71_{(\pm 0.03)}$	$6.31_{(\pm 0.01)}$	$6.50_{(\pm 0.02)}$	$11.93_{(\pm 0.01)}$	$11.78_{(\pm 0.01)}$	$11.85_{(\pm 0.01)}$
	Parametric [40]	$5.76_{(\pm 0.10)}$	$6.10_{(\pm 0.07)}$	$5.92_{(\pm 0.05)}$	$7.61_{(\pm 0.01)}$	$6.25_{(\pm 0.01)}$	$6.86_{(\pm 0.01)}$
	CD [51]	$8.21_{(\pm 0.00)}$	$7.15_{(\pm 0.01)}$	$7.64_{(\pm 0.01)}$	$12.41_{(\pm 0.00)}$	$11.89_{(\pm 0.00)}$	$12.14_{(\pm 0.00)}$
	AdaCAD [52]	$8.30_{(\pm 0.00)}$	$7.11_{(\pm 0.00)}$	$7.66_{(\pm 0.00)}$	$12.87_{(\pm 0.03)}$	$12.53_{(\pm 0.02)}$	$12.70_{(\pm 0.02)}$
	Entropy [53]	$8.32_{(\pm 0.03)}$	$7.73_{(\pm 0.11)}$	$8.01_{(\pm 0.05)}$	$13.78_{(\pm 0.02)}$	$13.33_{(\pm 0.01)}$	$13.55_{(\pm 0.02)}$
	CAD [26]	$8.16_{(\pm 0.06)}$	$7.16_{(\pm 0.02)}$	$7.62_{(\pm 0.00)}$	$12.99_{(\pm 0.03)}$	$12.51_{(\pm 0.02)}$	$12.75_{(\pm 0.03)}$
	COIECD [54]	$8.32_{(\pm 0.02)}$	$7.73_{(\pm 0.08)}$	$8.01_{(\pm 0.03)}$	$14.46_{(\pm 0.02)}$	$14.21_{(\pm 0.03)}$	$14.33_{(\pm 0.02)}$
	AGLA [57]	$8.24_{(\pm 0.01)}$	$7.56_{(\pm 0.05)}$	$7.88_{(\pm 0.01)}$	$14.29_{(\pm 0.02)}$	$13.91_{(\pm 0.01)}$	$14.08_{(\pm 0.01)}$
	VCD [58]	$8.13_{(\pm 0.05)}$	$7.41_{(\pm 0.03)}$	$7.75_{(\pm 0.04)}$	$13.71_{(\pm 0.03)}$	$13.81_{(\pm 0.03)}$	$13.76_{(\pm 0.03)}$
	ALFAR (ours)	$\mathbf{8.61}_{(\pm 0.01)}$	$\mathbf{8.04}_{(\pm 0.01)}$	$\mathbf{8.31}_{(\pm 0.01)}$	$\mathbf{15.25}_{(\pm 0.00)}$	$\mathbf{15.11}_{(\pm 0.01)}$	$\mathbf{15.18}_{(\pm 0.01)}$
MiniGPT4	Regular [16]	$4.38_{(\pm 0.07)}$	$3.00_{(\pm 0.02)}$	$3.56_{(\pm 0.02)}$	$12.69_{(\pm 0.02)}$	$12.38_{(\pm 0.02)}$	$12.53_{(\pm 0.02)}$
	Parametric [16]	$2.34_{(\pm 0.01)}$	$2.10_{(\pm 0.01)}$	$2.21_{(\pm 0.00)}$	$4.72_{(\pm 0.01)}$	$3.93_{(\pm 0.01)}$	$4.29_{(\pm 0.01)}$
	CD [51]	$4.28_{(\pm 0.00)}$	$2.59_{(\pm 0.00)}$	$3.22_{(\pm 0.00)}$	$14.49_{(\pm 0.01)}$	$14.43_{(\pm 0.00)}$	$14.46_{(\pm 0.01)}$
	AdaCAD [52]	$4.78_{(\pm 0.00)}$	$3.43_{(\pm 0.01)}$	$3.99_{(\pm 0.00)}$	$14.82_{(\pm 0.02)}$	$14.96_{(\pm 0.02)}$	$14.89_{(\pm 0.02)}$
	Entropy [53]	$4.80_{(\pm 0.07)}$	$2.91_{(\pm 0.00)}$	$3.62_{(\pm 0.00)}$	$14.66_{(\pm 0.02)}$	$14.66_{(\pm 0.01)}$	$14.66_{(\pm 0.02)}$
	CAD [26]	$4.97_{(\pm 0.01)}$	$3.44_{(\pm 0.01)}$	$4.07_{(\pm 0.01)}$	$14.83_{(\pm 0.01)}$	$14.94_{(\pm 0.00)}$	$14.88_{(\pm 0.01)}$
	COIECD [54]	$4.57_{(\pm 0.03)}$	$3.40_{(\pm 0.01)}$	$3.90_{(\pm 0.00)}$	$14.87_{(\pm 0.01)}$	$14.67_{(\pm 0.02)}$	$14.77_{(\pm 0.02)}$
	AGLA [57]	$4.67_{(\pm 0.02)}$	$3.63_{(\pm 0.02)}$	$4.09_{(\pm 0.01)}$	$14.31_{(\pm 0.02)}$	$13.92_{(\pm 0.01)}$	$14.11_{(\pm 0.01)}$
	VCD [58]	$4.52_{(\pm 0.03)}$	$3.43_{(\pm 0.01)}$	$3.90_{(\pm 0.01)}$	$14.46_{(\pm 0.01)}$	$14.30_{(\pm 0.02)}$	$14.38_{(\pm 0.02)}$
	ALFAR (ours)	$\mathbf{5.05}_{(\pm 0.02)}$	$\mathbf{3.87}_{(\pm 0.01)}$	$\mathbf{4.38}_{(\pm 0.00)}$	$\mathbf{15.16}_{(\pm 0.01)}$	$\mathbf{15.27}_{(\pm 0.01)}$	$\mathbf{15.05}_{(\pm 0.01)}$

Similarly, we represent contextual knowledge by using the output logits that have context as additional inputs and perform attention reallocation to better utilize the context:

$$\text{logit}_\text{c} = \hat{\mathcal{M}}_\theta(y_t|q, I, c, y_{<t}) \tag{9}$$

where $\hat{\mathcal{M}}_\theta$ is the MLLM with attention reallocation. The reliability of the parametric and contextual knowledge can thus be measured by the attention weights that are assigned to the image and context tokens capturing the correlation among tokens [59, 60, 61]:

$$\lambda_v^t = \sum_{i \in S_I} \mathbf{A}_{ti} \ , \quad \lambda_c^t = \sum_{j \in S_c} \mathbf{A}_{tj} \tag{10}$$

Table 3: VQA Accuracy comparison on discriminative multi-choice VQA datasets over three runs.

Model	Decoding	InfoSeek	ViQuAE	Model	Decoding	InfoSeek	ViQuAE
LLaVA-1.5	Regular [14]	$51.97_{(\pm 0.42)}$	$53.32_{(\pm 0.20)}$	InstructBLIP	Regular [10]	$23.44_{(\pm 0.89)}$	$19.82_{(\pm 0.12)}$
	Parametric [14]	$39.15_{(\pm 0.02)}$	$51.06_{(\pm 0.16)}$		Parametric [10]	$8.73_{(\pm 0.23)}$	$6.53_{(\pm 0.35)}$
	CD [51]	$49.95_{(\pm 0.20)}$	$52.56_{(\pm 0.03)}$		CD [51]	$22.95_{(\pm 0.63)}$	$21.76_{(\pm 0.51)}$
	CAD [26]	$52.08_{(\pm 0.16)}$	$52.99_{(\pm 0.23)}$		CAD [26]	$26.56_{(\pm 0.56)}$	$23.18_{(\pm 0.30)}$
	AdaCAD [52]	$52.30_{(\pm 0.04)}$	$52.99_{(\pm 0.30)}$		AdaCAD [52]	$27.07_{(\pm 0.05)}$	$23.64_{(\pm 0.08)}$
	Entropy [53]	$53.33_{(\pm 0.07)}$	$54.26_{(\pm 0.05)}$		Entropy [53]	$27.50_{(\pm 0.07)}$	$23.19_{(\pm 0.39)}$
	COIECD [54]	$52.08_{(\pm 0.21)}$	$52.99_{(\pm 0.23)}$		COIECD [54]	$25.65_{(\pm 0.05)}$	$20.84_{(\pm 0.05)}$
	VCD [58]	$53.87_{(\pm 0.07)}$	$55.13_{(\pm 0.09)}$		VCD [58]	$23.31_{(\pm 0.07)}$	$20.28_{(\pm 0.53)}$
	AGLA [57]	$53.53_{(\pm 0.50)}$	$54.24_{(\pm 0.21)}$		AGLA [57]	$21.24_{(\pm 0.55)}$	$16.77_{(\pm 0.30)}$
	ALFAR (ours)	$\mathbf{58.35}_{(\pm 0.21)}$	$\mathbf{55.91}_{(\pm 0.13)}$		**ALFAR (ours)**	$\mathbf{35.65}_{(\pm 0.09)}$	$\mathbf{24.11}_{(\pm 0.07)}$
Shikra	Regular [40]	$19.41_{(\pm 0.12)}$	$17.73_{(\pm 0.01)}$	MiniGPT-4	Regular [16]	$25.83_{(\pm 1.42)}$	$24.06_{(\pm 0.46)}$
	Parametric [40]	$9.65_{(\pm 0.14)}$	$10.90_{(\pm 0.04)}$		Parametric [16]	$19.73_{(\pm 0.57)}$	$20.42_{(\pm 1.22)}$
	CD [51]	$24.83_{(\pm 0.20)}$	$21.38_{(\pm 0.03)}$		CD [51]	$26.55_{(\pm 0.31)}$	$20.46_{(\pm 0.76)}$
	CAD [26]	$24.51_{(\pm 0.06)}$	$21.68_{(\pm 0.15)}$		CAD [26]	$27.58_{(\pm 0.43)}$	$23.39_{(\pm 0.07)}$
	AdaCAD [52]	$23.95_{(\pm 0.18)}$	$21.51_{(\pm 0.01)}$		AdaCAD [52]	$28.01_{(\pm 0.17)}$	$23.05_{(\pm 0.61)}$
	Entropy [53]	$24.12_{(\pm 0.03)}$	$21.99_{(\pm 0.08)}$		Entropy [53]	$28.84_{(\pm 0.68)}$	$22.59_{(\pm 0.07)}$
	COIECD [54]	$24.18_{(\pm 0.13)}$	$21.68_{(\pm 0.15)}$		COIECD [54]	$29.44_{(\pm 0.01)}$	$25.94_{(\pm 0.18)}$
	VCD [58]	$25.76_{(\pm 0.14)}$	$23.11_{(\pm 0.72)}$		VCD [58]	$28.74_{(\pm 0.10)}$	$25.45_{(\pm 0.05)}$
	AGLA [57]	$26.26_{(\pm 0.29)}$	$22.72_{(\pm 0.15)}$		AGLA [57]	$29.28_{(\pm 0.87)}$	$27.87_{(\pm 0.08)}$
	ALFAR (ours)	$\mathbf{28.11}_{(\pm 0.01)}$	$\mathbf{23.42}_{(\pm 0.29)}$		**ALFAR (ours)**	$\mathbf{30.88}_{(\pm 0.12)}$	$\mathbf{32.28}_{(\pm 0.02)}$
LLaVA-Next	Regular [62]	$53.90_{(\pm 0.26)}$	$55.44_{(\pm 0.12)}$	Qwen2.5-VL	Regular [63]	$53.00_{(\pm 0.18)}$	$53.84_{(\pm 0.01)}$
	Parametric [62]	$42.50_{(\pm 0.12)}$	$53.94_{(\pm 0.11)}$		Parametric [63]	$45.96_{(\pm 0.32)}$	$53.28_{(\pm 0.15)}$
	CD [51]	$51.43_{(\pm 0.23)}$	$51.38_{(\pm 0.01)}$		CD [51]	$56.23_{(\pm 0.23)}$	$59.33_{(\pm 0.15)}$
	CAD [26]	$53.83_{(\pm 0.13)}$	$52.84_{(\pm 0.27)}$		CAD [26]	$58.67_{(\pm 0.25)}$	$60.92_{(\pm 0.20)}$
	AdaCAD [52]	$55.60_{(\pm 0.34)}$	$54.87_{(\pm 0.13)}$		AdaCAD [52]	$57.87_{(\pm 0.05)}$	$60.89_{(\pm 0.28)}$
	Entropy [53]	$54.43_{(\pm 0.20)}$	$55.60_{(\pm 0.15)}$		Entropy [53]	$58.67_{(\pm 0.13)}$	$60.66_{(\pm 0.19)}$
	COIECD [54]	$53.93_{(\pm 0.11)}$	$54.17_{(\pm 0.33)}$		COIECD [54]	$59.23_{(\pm 0.10)}$	$60.59_{(\pm 0.02)}$
	VCD [58]	$53.27_{(\pm 0.27)}$	$54.32_{(\pm 0.03)}$		VCD [58]	$58.56_{(\pm 0.32)}$	$60.12_{(\pm 0.33)}$
	AGLA [57]	$54.03_{(\pm 0.25)}$	$54.65_{(\pm 0.34)}$		AGLA [57]	$59.13_{(\pm 0.25)}$	$61.08_{(\pm 0.13)}$
	ALFAR (ours)	$\mathbf{58.47}_{(\pm 0.21)}$	$\mathbf{59.29}_{(\pm 0.21)}$		**ALFAR (ours)**	$\mathbf{61.57}_{(\pm 0.18)}$	$\mathbf{63.22}_{(\pm 0.08)}$

Finally, based on the adaptive weights, the two types of knowledge are fused dynamically at each decoding step t:

$$p(y_t) \sim \text{softmax}\left[(1+\frac{\lambda_c^t}{\lambda_v^t})\,\text{logit}_\text{c} - (1-\frac{\lambda_v^t}{\lambda_c^t})\,\text{logit}_\text{p}\right] \tag{11}$$

5 Experiment

5.1 Experimental Settings

Datasets. We conduct experiments over three types of knowledge-intensive datasets: (1) Free-form generative datasets including **Human** [23], a high-quality info-seeking dataset curated and verified by experts, and INFOSEEK$_{wiki}$ [23] which encompasses diverse entities from Wikidata. For INFOSEEK$_{wiki}$, we adopt its **Validation** set for evaluations to be aligned with prior studies [27, 28]. (2) Multi-choice discriminative datasets including **Infoseek** [23, 32] and **ViQuAE** [24, 32] which are both multi-choice knowledge-intensive datasets that are collected for assessing cross-modality knowledge conflicts as described in [32]. (3) Knowledge-based datasets including **OK-VQA** [64], **AOK-VQA** [65] and **Encyclopedic VQA** (E-VQA) [31], which are widely adopted for evaluations of tasks that require commonsense knowledge. Additional details about the datasets and knowledge bases are listed in the Appendix A5.

MLLM baselines and SOTA methods. We perform evaluations by using four representative MLLMs as backbones: LLaVA-1.5 (7B and 13B) [14], InstructBLIP (7B and 13B) [10], Shikra (7B) [40], MiniGPT-4 (7B) [16], LLaVA-Next (7B) [62], and Qwen2.5-VL (3B) [63]. For benchmarking, we select several SOTA training-free decoding methods that aim to mitigate knowledge conflicts in LLMs: Contrastive Decoding (CD) [51], Adaptive Context-Aware Decoding (AdaCAD) [52], Entropy-based decoding (Entropy) [53], Context-Aware Decoding (CAD) [26] and COntextual Information-Entropy Constraint Decoding (COIECD) [54]. In addition, we also benchmark with two representative

hallucination mitigation methods, including Visual Contrastive Decoding (VCD) [58] and Assembly of Global and Local Attention (AGLA) [57].

Implementation details. For knowledge retrieval, we employ the vision encoder of CLIP-ViT-L/14-336 [66] as the retriever and append the first retrieved knowledge to the prompt as context. The scaled factor k is set to 0.4 to avoid excessive adjustment. For the knowledge base, we use Wikipedia dumps provided by [23] and select items with associated images for retrieval. Multinomial sampling serves as the decoding strategy. We denote MLLM inference with retrieved knowledge as *Regular* and without retrieved knowledge as *Parametric*. We follow prior studies [26, 53, 58] and adopt adaptive plausibility constraints [51] for fair comparisons. All experiments are conducted on four NVIDIA RTX 3090 GPUs. All compared methods are reproduced by us according to their released codes or original papers.

5.2 Experimental Results

Experiments on free-form datasets. Tab. 2 shows experimental results of four representative MLLMs [14, 10, 40, 16] over two free-form generative knowledge-intensive datasets [23]. We can see that the proposed ALFAR consistently outperforms the *Regular* decoding strategy by substantial margins (averaged around 2.5% in overall accuracy) across all MLLMs and datasets. Additionally, ALFAR surpasses state-of-the-art decoding methods as well, demonstrating its effectiveness in the better utilization of contextual knowledge.

Experiments on multi-choice datasets. Tab. 3 presents experimental results of six MLLMs [14, 10, 40, 16, 62, 63] over two multi-choice discriminative datasets [32, 24]. Notably, ALFAR achieves an average improvement of 6.6% over Regular decoding and consistently surpasses state-of-the-art decoding strategies by substantial margins, underscoring its effectiveness in diverse tasks. Moreover, we observe that LLaVA-1.5 [14] demonstrates a stronger instruction-following capability compared to other models, enabling it to produce more correctly formatted outputs.

Experiments on knowledge-based datasets. In addition to entity knowledge-based datasets [32, 24], we conduct experiments on commonsense knowledge-based datasets, OK-VQA [64], AOK-VQA [65] and Encyclopedic VQA (E-VQA) [31] with LLaVA-1.5 [14]. As shown in Tab. 4, ALFAR surpasses Regular decoding by 15.2% and consistently outperforms state-of-the-art decoding strategies, underscoring its effectiveness in addressing a broader range of knowledge-intensive tasks.

Table 4: VQA Accuracy comparison on the knowledge-based VQA datasets with LLaVA-1.5 [14] over three runs.

Model	OK-VQA	AOK-VQA	E-VQA
Regular [14]	$46.17_{(\pm 0.12)}$	$44.13_{(\pm 0.00)}$	$19.14_{(\pm 2.83)}$
Parametric [14]	$45.13_{(\pm 0.40)}$	$43.23_{(\pm 1.02)}$	$5.34_{(\pm 0.01)}$
CD [51]	$55.00_{(\pm 0.03)}$	$51.67_{(\pm 0.44)}$	$28.62_{(\pm 0.68)}$
CAD [26]	$56.43_{(\pm 0.46)}$	$53.93_{(\pm 0.56)}$	$28.62_{(\pm 0.76)}$
AdaCAD [52]	$57.10_{(\pm 0.04)}$	$54.40_{(\pm 0.48)}$	$28.33_{(\pm 0.42)}$
Entropy [53]	$56.27_{(\pm 0.05)}$	$53.93_{(\pm 0.26)}$	$29.24_{(\pm 0.11)}$
COIECD [54]	$56.43_{(\pm 0.46)}$	$53.93_{(\pm 0.56)}$	$28.24_{(\pm 0.74)}$
VCD [58]	$57.80_{(\pm 0.13)}$	$57.40_{(\pm 0.12)}$	$27.71_{(\pm 0.33)}$
AGLA [57]	$57.53_{(\pm 0.05)}$	$55.40_{(\pm 0.63)}$	$28.29_{(\pm 0.56)}$
ALFAR (ours)	$\mathbf{60.83}_{(\pm 0.01)}$	$\mathbf{59.93}_{(\pm 0.02)}$	$\mathbf{29.57}_{(\pm 0.08)}$

6 Discussion

6.1 Ablation study

We conduct ablation studies on both multi-choice and commonsense knowledge-based datasets [32, 65] to assess the effectiveness of each design in the proposed ALFAR model with LLaVA-1.5 [14]. As shown in Tab. 5, the **Attention Reallocation** enables MLLMs to better utilize retrieved knowledge, thereby enhancing overall performance. The **Logits Fusion** mitigates knowledge conflicts, allowing MLLMs to integrate retrieved knowledge and improve overall performance effectively. Moreover, applying **Adaptive Weights** during logits fusion helps MLLMs better leverage both parametric and contextual knowledge while reducing the impact of noise from the retrieved context, further contributing to performance gains.

Table 5: Experimental results of ablation study with different model variants.

Variants	InfoSeek	AOK-VQA
Regular [14]	51.97	44.13
+ Attention Reallocation	53.42	46.30
+ Logits Fusion	55.83	55.90
+ Adaptive Weights	58.35	59.93

6.2 Effect of different retrievers

Table 6: Experimental results with different CLIP retrievers. '-T' means using the input query for knowledge retrieval.

Retriever	Recall	LLaVA	Ours
CLIP-B/16 [66]	38.27	45.93	**51.60**
CLIP-L/14 [66]	56.53	51.23	**57.60**
CLIP-L/14-336 [66]	58.37	51.97	**58.35**
CLIP-B/16-T [66]	5.56	32.97	**35.43**
CLIP-L/14-T [66]	6.00	33.40	**36.23**
CLIP-L/14-336-T [66]	5.73	33.37	**36.63**

We investigate the impact of different retrievers and retrieval strategies on recall and model performance on the InfoSeek dataset [32]. As shown in Tab. 6, our model consistently enhances performance across all configurations, demonstrating its robustness to retrieval noise. Additionally, retrieving knowledge based on the input query yields low recall due to its limited information about entities in the image. Despite this limitation, our model achieves consistent performance improvements even under low retrieval recall, owing to its adaptive fusion strategy that balances parametric and contextual knowledge.

6.3 Inference Efficiency

Table 7: Inference time of different methods over one sample with LLaVA-1.5 [14].

Variants	InfoSeek	Human
Regular [14]	0.46s (1.00x)	0.50s (1.00x)
CD [51]	0.62s (1.35x)	0.72s (1.44x)
CAD [26]	0.62s (1.34x)	0.72s (1.44x)
AdaCAD [52]	0.62s (1.35x)	0.73s (1.45x)
Entropy [53]	0.63s (1.37x)	0.73s (1.46x)
COIECD [54]	0.62s (1.35x)	0.73s (1.45x)
VCD [58]	0.63s (1.38x)	0.74s (1.47x)
AGLA [57]	0.83s (1.81x)	1.02s (2.04x)
ALFAR (ours)	0.62s (1.35x)	0.73s (1.46x)

In this section, we conduct a detailed analysis of the inference efficiency of the proposed model in comparison with representative baseline methods. Specifically, we examine how the incorporation of parametric knowledge modeling affects the inference time under both discriminative and generative tasks. Compared with the baseline MRAG, our model introduces only a modest computational overhead, which mainly arises from the additional operations required to encode and integrate parametric knowledge. Nevertheless, this extra cost remains limited, as the parametric knowledge component without contextual input is relatively concise. Moreover, all compared methods except the baseline also require modeling of both contextual and parametric knowledge. Consequently, the overall inference cost of our model is comparable to that of these methods. We evaluate all models on both the discriminative multi-choice dataset Infoseek [32] and the free-form generative dataset Human [23] using a single input sample. The results summarized in Tab. 7 demonstrate that our method achieves similar inference times.

6.4 Qualitative Examples

Fig. 6 shows a qualitative comparison of four decoding approaches. It can be observed that Vanilla LLaVA [14] without context produces a false response due to the lack of knowledge about the entity in the image. LLaVA with MRAG [27] incorporates the contextual knowledge but still produces the same incorrect response as Vanilla LLaVA [14], primarily due to knowledge conflicts that cause MLLMs to favor their parametric knowledge. LLaVA with CAD [26] can mitigate knowledge conflicts effectively by reducing the influence of parametric knowledge. However, it produces a false response as well, mainly due to attention bias with excessive focus on images and uniform attention toward context. In contrast, ALFAR introduces attention reallocation and adaptive logits fusion, enabling MLLMs to prioritize query-relevant contextual knowledge and produce accurate responses.

6.5 ALFAR on MLLM Scalability

Table 8: Experiments with larger MLLMs.

Decoding	LLaVA-13B	InstructBLIP-13B
Parametric	44.40	8.93
Regular	55.83	18.70
ALFAR (ours)	**59.63**	**27.63**

Tab. 8 presents experimental results of the 13B variants of LLaVA-1.5 [14] and InstructBLIP [10] over the InfoSeek dataset [32]. Notably, ALFAR consistently improves performance across both models, demonstrating its superior scalability with respect to the model size. Interestingly, we observe a performance decline in

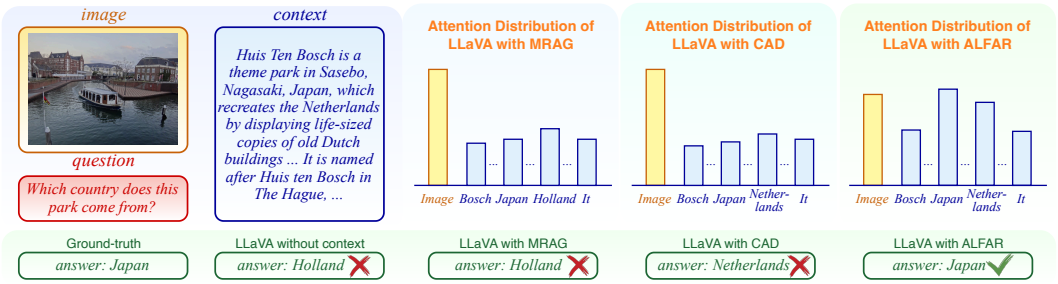

Figure 6: Illustration of generated responses and corresponding attention distributions across different decoding methods, using LLaVA-1.5 [14] as the backbone model.

InstructBLIP 13B compared to the 7B variant, which may be attributed to the model's increased reliance on parametric knowledge.

6.6 Effect of Different Decoding Strategies

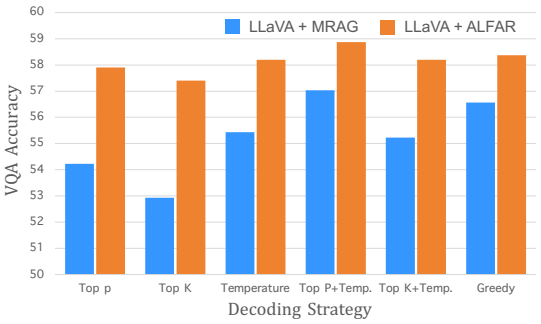

Figure 5: Experimental results with different decoding strategies for response generation.

Beyond the multinomial sampling strategy discussed in this paper, we further investigate the robustness of the proposed ALFAR framework under diverse decoding settings. To this end, we conduct experiments using LLaVA-1.5 [14] on the multi-choice InfoSeek dataset [32], and evaluate six additional decoding strategies commonly adopted in MLLM generation. These strategies include Top-P sampling [67] with $p = 0.7$, Top-K sampling [68] with $k = 50$, greedy decoding [69], temperature sampling [70] with $t = 0.5$, Top-P sampling with temperature ($p = 0.7$, $t = 0.5$), and Top-K sampling with temperature ($k = 50$, $t = 0.5$). As illustrated in Fig. 5, ALFAR consistently enhances model performance across all decoding strategies. This consistent gain highlights the generalizability of ALFAR's learning principle and suggests that it effectively complements various decoding schemes. Moreover, the results indicate that ALFAR can serve as a robust and plug-and-play enhancement to existing MLLMs, ensuring stable performance regardless of decoding configuration.

7 Conclusion

In this paper, we examine representative MLLMs and find that they often struggle to fully utilize retrieved knowledge for knowledge-intensive tasks. We attribute this limitation to two key factors: *attention bias* toward different tokens and *knowledge conflicts* between parametric and contextual knowledge. To address these challenges, we introduce *Adaptive Logits Fusion and Attention Reallocation (ALFAR)*, a training-free and plug-and-play approach that enhances MLLM performance by dynamically reallocating attention and harmonizing parametric and contextual knowledge. Specifically, ALFAR mitigates attention bias by adaptively shifting focus from visual tokens to context tokens based on query-context relevance. Furthermore, it decouples and balances parametric and contextual knowledge at the output logits, effectively resolving conflicts. Experiments across multiple MLLMs and benchmarks show that ALFAR consistently surpasses state-of-the-art methods by substantial margins without requiring additional training or external tools, highlighting its versatility and effectiveness for various knowledge-intensive tasks.

8 Acknowledgments

This work was supported by the National Science and Technology Major Project (2022ZD0117102), National Natural Science Foundation of China (62177038, 62293551, 62277042, 62377038), Project of China Knowledge Centre for Engineering Science and Technology, "LENOVO-XJTU" Intelligent Industry Joint Laboratory Project, The Youth Al Talents Fund of the Chinese Association of Automation under Major Program (HBRC-JKYZD-2024-311).

References

[1] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[2] Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, et al. Ul2: Unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations*, 2022.

[3] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[5] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[6] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model, 2023.

[7] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[8] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[9] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[10] 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.

[11] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023.

[12] Wenbin An, Jiahao Nie, Yaqiang Wu, Feng Tian, Shijian Lu, and Qinghua Zheng. Empowering multimodal llms with external tools: A comprehensive survey. *arXiv preprint arXiv:2508.10955*, 2025.

[13] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023.

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

[15] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[16] 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.

[17] Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12401–12430, 2024.

[18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[19] 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.

[20] Shilin Sun, Wenbin An, Feng Tian, Fang Nan, Qidong Liu, Jun Liu, Nazaraf Shah, and Ping Chen. A review of multimodal explainable artificial intelligence: Past, present and future. *arXiv preprint arXiv:2412.14056*, 2024.

[21] 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*, 2024.

[22] Jusung Lee, Sungguk Cha, Younghyun Lee, and Cheoljong Yang. Visual question answering instruction: Unlocking multimodal large language model to domain-specific visual multitasks. *arXiv preprint arXiv:2402.08360*, 2024.

[23] 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*, pages 14948–14968, 2023.

[24] 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*, pages 3108–3120, 2022.

[25] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020.

[26] 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)*, pages 783–791, 2024.

[27] 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*, pages 1818–1826, 2024.

[28] Federico Cocchi, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Augmenting multimodal llms with self-reflective tokens for knowledge-based visual question answering. *arXiv preprint arXiv:2411.16863*, 2024.

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

[30] Amit Artzy and Roy Schwartz. Attend first, consolidate later: On the importance of attention in different llm layers. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 177–184, 2024.

[31] 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*, pages 3113–3124, 2023.

[32] 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.

[33] Mehrdad Farahani and Richard Johansson. Deciphering the interplay of parametric and non-parametric memory in retrieval-augmented language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16966–16977, 2024.

[34] Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö Arık. Astute rag: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. *arXiv preprint arXiv:2410.07176*, 2024.

[35] Justin Kruger and David Dunning. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, 77(6):1121, 1999.

[36] Aniket Kumar Singh, Suman Devkota, Bishal Lamichhane, Uttam Dhakal, and Chandra Dhakal. The confidence-competence gap in large language models: A cognitive study. *arXiv preprint arXiv:2309.16145*, 2023.

[37] Weizhe Lin, Jingbiao Mei, Jinghong Chen, and Bill Byrne. Preflmr: Scaling up fine-grained late-interaction multi-modal retrievers. *arXiv preprint arXiv:2402.08327*, 2024.

[38] Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, pages 387–404. Springer, 2025.

[39] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*, 2023.

[40] 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*, 2023.

[41] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[42] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024.

[43] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[44] Xiaofeng Zhang, Yihao Quan, Chaochen Gu, Chen Shen, Xiaosong Yuan, Shaotian Yan, Hao Cheng, Kaijie Wu, and Jieping Ye. Seeing clearly by layer two: Enhancing attention heads to alleviate hallucination in lvlms. *arXiv preprint arXiv:2411.09968*, 2024.

[45] Haonan Lin, Mengmeng Wang, Yan Chen, Wenbin An, Yuzhe Yao, Guang Dai, Qianying Wang, Yong Liu, and Jingdong Wang. Dreamsalon: A staged diffusion framework for preserving identity-context in editable face generation. *arXiv preprint arXiv:2403.19235*, 2024.

[46] Xiaofeng Zhang, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. From redundancy to relevance: Enhancing explainability in multimodal large language models. *arXiv e-prints*, pages arXiv–2406, 2024.

[47] Jiahao Nie, Gongjie Zhang, Wenbin An, Yap-Peng Tan, Alex C Kot, and Shijian Lu. Mmrel: A relation understanding dataset and benchmark in the mllm era. *arXiv preprint arXiv:2406.09121*, 2024.

[48] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

[49] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[50] Tao Zhang, Ziqi Zhang, Zongyang Ma, Yuxin Chen, Zhongang Qi, Chunfeng Yuan, Bing Li, Junfu Pu, Yuxuan Zhao, Zehua Xie, et al. mr^{2}ag: Multimodal retrieval-reflection-augmented generation for knowledge-based vqa. *arXiv preprint arXiv:2411.15041*, 2024.

[51] 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)*, pages 12286–12312, 2023.

[52] 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.

[53] 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.

[54] 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.

[55] Yanwen Huang, Yong Zhang, Ning Cheng, Zhitao Li, Shaojun Wang, and Jing Xiao. Dynamic attention-guided context decoding for mitigating context faithfulness hallucinations in large language models. *arXiv preprint arXiv:2501.01059*, 2025.

[56] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[57] 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*, 2024.

[58] 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.

[59] Shiyu Zhao, Zhenting Wang, Felix Juefei-Xu, Xide Xia, Miao Liu, Xiaofang Wang, Mingfu Liang, Ning Zhang, Dimitris N Metaxas, and Licheng Yu. Accelerating multimodel large language models by searching optimal vision token reduction. *arXiv preprint arXiv:2412.00556*, 2024.

[60] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021.

[61] Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. Snapkv: Llm knows what you are looking for before generation. *arXiv preprint arXiv:2404.14469*, 2024.

[62] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[63] 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*, 2025.

[64] 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*, pages 3195–3204, 2019.

[65] 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*, pages 146–162. Springer, 2022.

[66] 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*, pages 8748–8763. PMLR, 2021.

[67] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

[68] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018.

[69] Ronald A DeVore and Vladimir N Temlyakov. Some remarks on greedy algorithms. *Advances in computational Mathematics*, 5(1):173–187, 1996.

[70] David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.

[71] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[72] 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.

[73] 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*, pages 8769–8778, 2018.

[74] Feilong Tang, Zile Huang, Chengzhi Liu, Qiang Sun, Harry Yang, and Ser-Nam Lim. Intervening anchor token: Decoding strategy in alleviating hallucinations for MLLMs. In *The Thirteenth International Conference on Learning Representations*, 2025.

[75] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *URL https://arxiv.org/abs/2307.03172*, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We accurately pinpoint the contributions and scope of this paper in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of the work in the Section Limitations.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all the information needed to reproduce the main experiments for reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will provide open access to the data and code to reproduce the main experiments in our paper.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify all the training and testing details necessary to understand the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the mean and variance of results over three runs for the main experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide sufficient information on the computer resources needed to reproduce the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential societal impacts in Section Broader Impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators used in the paper are properly credited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

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 [71]. 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:

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

where $\texttt{Err}$ 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:

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

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 [23, 32] in Tab. 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 Tab. A2. Specifically, for Validation [23] and InfoSeek [32], we follow previous works [27] and adopt a knowledge

Table A1: Retrieval recall with Ground-Truth knowledge (GT) and knowledge from different retrieval rankings on the multi-choice InfoSeek dataset [23, 32].

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 [23]	73,620	Wikipedia [23]
Human [23]	8,931	Wikipedia [23]
InfoSeek [32]	3,000	Wikipedia [23]
ViQuAE [32]	3,000	Wikipedia [23]
OK-VQA [64]	5,046	GPT-3.5 [21]
AOK-VQA [65]	1,145	GPT-3.5 [21]
E-VQA [31]	700	Encyclopedia [31]

base containing 1.7K entities derived from the original Wikipedia knowledge base [23]. For Human [23] and ViQuAE [24], we use the original Wikipedia knowledge base [23], selecting 73.6K entities accompanied by images for knowledge retrieval. For OK-VQA [64] and AOK-VQA [65], we utilize the knowledge base provided by [21], which was generated using GPT-3.5 [72]. For E-VQA [31], we select templated questions with images from the iNaturalist dataset [73] and use the corresponding ground-truth knowledge for inference. All evaluations are conducted using the official scripts.

A6 Adaptive Plausibility Constraints

We follow prior studies [26, 53, 58] and adopt adaptive plausibility constraints [51] 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:

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

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.

A7 Effect of Intervention Layers

We investigate the impact of attention reallocation at different MLLM layers on the InfoSeek dataset [23] using LLaVA-1.5 [14], as summarized in Tab. 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 [30]. In contrast, applying attention reallocation in middle layers (layers 17-24) yields smaller gains, as these layers primarily handle multimodal alignment and feature aggregation [74], where attention bias is less severe. Notably, reallocating attention in late layers (layers 25-32) leads to the most substantial performance gains, as these layers are responsible for reasoning and directly affect output generation [74]. Furthermore, leveraging attention reallocation in both shallow (layers 1–8) and deep (layer 32) layers yields the best performance.

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**

A8 Effect of different numbers of knowledge

We examine the effect of varying the amount of knowledge provided to MLLMs on retrieval recall and model performance on the InfoSeek dataset [23]. As shown in Tab. A3, appending additional knowledge to the prompt improves retrieval recall but has limited impact on model performance, as MLLMs often struggle to effectively utilize information from lengthy input contexts [75]. Our model addresses this limitation by guiding MLLMs based on query-context relevance. However, the modest performance gains underscore the need for future research on enhancing MLLMs' ability to process and leverage extended contexts.

Table A3: Experimental results with different numbers of knowledge using LLaVA-1.5 [14].

#Knowledge	Recall	LLaVA [14]	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**