Enhancing Retrieval-Augmented Large Vision Language Models via Knowledge Conflict Mitigation

Wenbin An[1,2,*], Jiahao Nie[3,*], Feng Tian[1,2,†], Mingxiang Cai[4], Yaqiang Wu[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, shijian.lu@ntu.edu.sg

Abstract

Multimodal Retrieval-Augmented Generation (MRAG) has recently been explored to empower Large Vision Language Models (LVLMs) with more comprehensive and up-to-date contextual knowledge, aiming to compensate for their limited and coarse-grained parametric knowledge in knowledge-intensive tasks. However, the retrieved contextual knowledge is usually not aligned with LVLMs' internal parametric knowledge, leading to knowledge conflicts and further unreliable responses. To tackle this issue, we design KCM, a training-free and plug-and-play framework that can effectively mitigate knowledge conflicts while incorporating MRAG for more accurate LVLM responses. KCM enhances contextual knowledge utilization by modifying the LVLM architecture from three key perspectives. First, KCM adaptively adjusts attention distributions among multiple attention heads, encouraging LVLMs to focus on contextual knowledge with reduced distraction. Second, KCM identifies and prunes knowledge-centric LVLM neurons that encode coarse-grained parametric knowledge, thereby suppressing interferences and enabling more effective integration of contextual knowledge. Third, KCM amplifies the information flow from the input context by injecting supplementary context logits, reinforcing its contribution to the final output. Extensive experiments over multiple LVLMs and benchmarks show that KCM outperforms the state-of-the-art consistently by large margins, incurring neither extra training nor external tools.

Code — https://github.com/Lackel/KCM

Introduction

By integrating the perception capabilities of vision encoders (Radford et al. 2021) with the generative power of Large Language Models (LLMs) (Touvron et al. 2023; Chiang et al. 2023; Bai et al. 2023a), Large Vision Language Models (LVLMs) (Bai et al. 2023b; Dai et al. 2023; Liu et al. 2023; Zhu et al. 2023) have achieved great success in various visual understanding tasks such as image captioning (Al-Shamayleh et al. 2024) and semantic segmentation (Lai et al. 2024). However, the parametric knowledge encoded in LVLMs is typically general and coarse-grained (Guu et al. 2020), constraining their effectiveness on knowledge-intensive tasks that demand domain-specific information (Chen et al. 2023b). To overcome this limitation, recent studies have explored Multimodal Retrieval-Augmented Generation (MRAG) (Caffagni et al. 2024) to retrieve contextual knowledge from an external database, thereby empowering LVLMs with domain-specific information for better handling of knowledge-intensive tasks.

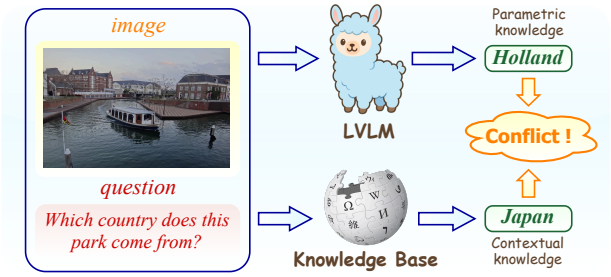

Figure 1: The parametric knowledge encoded in LVLMs conflicts with the retrieved contextual knowledge, which can confuse LVLMs and lead to unreliable responses.

Despite the improved quality of retrieved knowledge, MRAG often suffers from knowledge conflicts (Xu et al. 2024; Jin et al. 2024), where the retrieved contextual knowledge is not aligned with the parametric knowledge acquired during LVLM training. As illustrated in Fig. 1, the conflicts arise when the LVLM's internal understanding of an image contradicts external contextual information. Such a discrepancy interferes with the integration of parametric and contextual knowledge and leads to confusion in response generation, undermining the reliability of LVLMs.

We further investigate the limitations of LVLMs' parametric knowledge and how knowledge conflicts undermine the effectiveness of MRAG. In the first setting, we evaluate the vanilla LLaVA, which relies exclusively on its internal parametric knowledge. As Tab. 1 shows, the model performs poorly due to its encoded broad and coarse-grained parametric knowledge. In the second setting, we test MRAG

*Equal contribution.
†Corresponding authors.

Model	ACC↑	MR↓	CR↑
LLaVA	39.2	100.0	0.0
GT Context	100.0	0.0	100.0
LLaVA + GT Context	65.5	46.9	53.1
LLaVA + KCM	**71.2**	**39.7**	**60.3**

Table 1: Experiments with parametric knowledge (LLaVA) and perfect contextual knowledge (GT Context). Accuracy (ACC) evaluates response quality. Memorization Ratio (MR) and Context Ratio (CR) quantify the model's reliance on parametric and contextual knowledge, respectively.

with *perfect ground-truth* contextual knowledge to eliminate variability of context quality and isolate the effects of knowledge conflicts. Ideally, the model should exploit such ground-truth contextual knowledge and generate perfect responses under such a setup (denoted as GT Context). However, the LVLM with the ground-truth contextual knowledge (denoted as LLaVA+GT Context) still produces a large portion of false responses. We examine the source of failures and identify that LVLMs demonstrate a persistent tendency to rely on their internal parametric memory (Jin et al. 2024), as reflected by a high memorization ratio, even in the presence of perfect contextual knowledge. These findings suggest that integrating parametric and contextual knowledge can lead to conflicts and confuse the model regarding which source to trust, ultimately hindering the effective use of contextual knowledge and leading to sub-optimal responses.

We propose KCM, a training-free and plug-and-play framework designed to mitigate knowledge conflicts and enhance the integration of contextual knowledge in LVLMs. KCM dives deep into the mechanism of LVLM architectures and enhances contextual knowledge utilization through systematic redesign of three key components. From the perspective of **information extraction**, since LVLMs extract and model contextual dependencies through multi-head attention (Vaswani 2017), KCM adaptively adjusts attention distributions across multiple attention heads to encourage the model to prioritize more relevant contextual knowledge. From the perspective of **knowledge activation**, KCM analyses neuron activation patterns in LVLMs (Fan et al. 2025) and prunes neurons associated with coarse-grained parametric knowledge, thereby reducing interference and enabling better utilization of the contextual knowledge. From the perspective of **response generation**, KCM amplifies the influence of contextual knowledge by injecting additional context logits into the final prediction, boosting the contribution of contextual knowledge. Finally, KCM combines enriched contextual knowledge with parametric knowledge via an **entropy-based weighting**. This allows the model to dynamically decide which source to rely on, enhancing robustness and adaptability. As Tab. 1 shows, KCM achieves substantial performance improvements by enabling more effective knowledge utilization. With neither extra training nor external tools, KCM outperforms the SOTA models across a variety of LVLMs and benchmarks.

The contributions of this work can be summarized in three major aspects: *First*, we examine knowledge conflicts comprehensively while introducing MRAG into LVLMs, revealing how such conflicts could undermine the reliability and effectiveness of MRAG. *Second*, we design KCM, a training-free and plug-and-play framework that systematically modifies LVLMs to improve the integration and utilization of contextual knowledge. *Third*, extensive experiments over both generative and discriminative benchmarks show that KCM outperforms SOTA methods by clear margins.

Related Work

Large Vision Language Models. LLMs (Touvron et al. 2023; Bai et al. 2023a) have demonstrated remarkable performance and adaptability across a wide range of tasks (An et al. 2024a, 2025), which has in turn accelerated the progress of LVLMs (Bai et al. 2023b; Dai et al. 2023; Gong et al. 2023; Liu et al. 2023; Zhu et al. 2023; Chen et al. 2023a). Most LVLMs follow a two-stage training paradigm: a pre-training stage that aligns visual and textual modalities using either projectors(Liu et al. 2023; Chen et al. 2023a) or Q-formers (Li et al. 2023a; Dai et al. 2023), followed by an instruction-tuning stage to enhance performance on multimodal tasks (Sun et al. 2024; An et al. 2024c). Leveraging the generative capabilities of LLMs and the perceptual strength of vision encoders, LVLMs have achieved impressive performance on perception-driven and commonsense-based multimodal benchmarks (Marino et al. 2019; Schwenk et al. 2022). However, LVLMs often perform poorly on knowledge-intensive tasks (Chen et al. 2023b), largely due to their broad and coarse-grained parametric knowledge as required during training.

Knowledge Conflicts in MRAG. MRAG (Caffagni et al. 2024) has emerged as a promising paradigm, aiming to empower LVLMs with fine-grained and contextual knowledge while handling knowledge-intensive tasks. It empowers LVLMs by fusing their internal parametric knowledge with contextual knowledge as retrieved from external multimodal data. However, MRAG tends to introduce knowledge conflicts (Xu et al. 2024), where the retrieved contextual knowledge may contradict the models' internal parametric knowledge. Such conflicts can confuse LVLMs during inference, ultimately degrading the reliability and consistency of MRAG in LVLM response generation.

Mitigating knowledge Conflicts in LLMs. Several approaches have been developed to address similar knowledge conflict issues in the LLM domain. For example, CAD (Shi et al. 2024) leverages contrastive decoding (Li et al. 2023c) to enhance factual consistency. AdaCAD (Wang et al. 2024), Entropy (Qiu et al. 2024), and COIECD (Yuan et al. 2024) attempt to mitigate the knowledge conflicts by exploring JS divergence, entropy regularization, and information constraints, respectively. However, these approaches focus on linguistic data only, which cannot be directly applied to LVLMs with multimodal data as inputs. We design KCM to fill this gap by explicitly targeting knowledge conflict mitigation under the multimodal setting, enabling more reliable response generation in MRAG-enhanced LVLMs.

Preliminary

Multimodal Retrieval-Augmented Generation

Given a textual question q and an image I, an LVLM $\mathcal{M}_\theta$ is expected to generate a reliable response y. To handle knowledge-intensive tasks, MRAG has been investigated to empower LVLMs with external knowledge. Specifically, it adopts a multimodal retriever to retrieve relevant textual knowledge c from a knowledge base. Hence, the objective of MRAG at time step t can be formulated by:

$$y_t = \arg\max_{y_t} M_\theta(y_t|q, I, c, y_{<t}) \tag{1}$$

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

LVLM Architecture

LVLMs typically encode the input image into visual tokens, which are then concatenated with text tokens and fed into LLMs for response generation. In the following sections, we delve into the internal architecture of LLMs to understand how they process input tokens and generate responses.

Multi-Head Attention To model contextual dependencies and extract salient information, LVLMs apply Multi-Head Attention (MHA) over the input token sequence. Specifically, for the i-th attention head, the self-attention mechanism computes the relevance of each token to all tokens and produces the output $\mathbf{O}^i$ as follows:

$$\mathbf{O}^i = \text{softmax}(\frac{\mathbf{Q}^i \cdot (\mathbf{K}^i)^\top}{\sqrt{d_k}} + \mathbf{M}) \cdot \mathbf{V}^i = \text{softmax}(\mathbf{A}^i) \cdot \mathbf{V}^i \tag{2}$$

where $\mathbf{A}^i$ is the attention weight matrix of the i-th head, $\mathbf{M}$ is a causal mask, and d_k is the feature dimension. The query $\mathbf{Q}^i$, the key $\mathbf{K}^i$, and the value $\mathbf{V}^i$ are vectors obtained through three linear layers. The matrix $\mathbf{A}^i$ captures the relative importance of different tokens during generation, thereby offering a meaningful basis to analyze the contribution of different token types (*e.g.,* image and context) to the generated tokens. Since different attention heads are known to specialize in capturing distinct aspects of the input (Li et al. 2023b), we categorize them into two types under the MRAG setting. The first is **parametric heads**, which primarily attend to image tokens to acquire parametric knowledge. The second is **contextual heads**, which focus on context tokens to acquire contextual knowledge.

Feed-Forward Network Following the MHA layer, a Feed-Forward Network (FFN) processes the output features with two linear layers. The FFN adopts a key-value memory mechanism (Geva et al. 2021; Huang et al. 2025), where parametric knowledge is primarily stored and selectively activated by input features. The FFN layer can be formulated as follows (bias terms omitted for brevity):

$$\text{FFN}(\mathbf{x}) = f(\mathbf{x} \cdot \mathbf{K}^\top) \cdot \mathbf{V} = \sum_{i=1}^{d} f(\mathbf{x} \cdot \mathbf{k}_i) \cdot \mathbf{v}_i = \sum_{i=1}^{d} \alpha_i \cdot \mathbf{v}_i \tag{3}$$

where $\mathbf{x}$ is the input of the FFN layer, $\mathbf{K} = \{\mathbf{k}_1, ..., \mathbf{k}_d\}$ and $\mathbf{V} = \{\mathbf{v}_1, ..., \mathbf{v}_d\}$ denote the sets of key vectors and value vectors, respectively. f is a nonlinear activation function, and the coefficient α_i represents the activation level of the corresponding value vector $\mathbf{v}_i$. Specifically, if $\alpha_i = 0$, the corresponding value vector $\mathbf{v}_i$ is not activated and thus contributes nothing to the output, indicating that the associated parametric knowledge remains unused (Fan et al. 2025). Conversely, $\alpha_i > 0$ signifies the activation of the parametric knowledge encoded in $\mathbf{v}_i$. This offers the probability of deactivating knowledge-centric neurons, thereby mitigating interference from the coarse-grained parametric knowledge.

Token Generation

By stacking multiple decoder blocks with MHA layers and FFN layers, LVLMs compute the probability of the next token based on the hidden states of the last token as follows:

$$p(y) = \text{softmax}(\mathbf{W_{out}}\mathbf{h_n} + \mathbf{b_{out}}) \tag{4}$$

where $\mathbf{W_{out}}$ and $\mathbf{b_{out}}$ are weight and bias vectors that map the hidden states into logits, and $\mathbf{h_n}$ is the hidden state of the last token. Although the last token can attend to previous tokens via self-attention, the contribution of context tokens is still indirect and limited due to the attention sink phenomenon (Xiao et al. 2024), where the model only focuses on a small subset of input tokens. This restricts the full utilization of contextual knowledge in the generation process.

Method

The proposed KCM mitigates knowledge conflicts and enhances contextual knowledge utilization by modifying three key components in LVLMs, as illustrated in Fig. 2. First, **Attention Adjustment** adaptively redistributes attention across different attention heads, encouraging the model to focus on the input context and extract significant contextual knowledge. Second, **Neuron Pruning** identifies and deactivates neurons associated with parametric knowledge, reducing interference from the coarse-grained parametric knowledge. Third, **Context-enhanced Logits Generation** amplifies the influence of contextual knowledge by injecting additional context logits into the final prediction, thereby increasing its contribution to the generated response. Finally, the enriched contextual knowledge is integrated with parametric knowledge via an **entropy-based weighting**, which allows the model to dynamically decide which source to rely on and enhance the robustness of KCM.

Attention Adjustment

As analyzed before, different attention heads specialize in capturing distinct aspects of input information (Li et al. 2023b; Vaswani 2017). We examine this feature by analyzing the attention distribution over different types of tokens. For each attention head, we compute the average attention it allocates to the visual and context tokens as follows:

$$R_I^i = \frac{1}{|S_I|} \sum_{j \in S_I} \mathbf{A}_{nj}^i, \quad R_c^i = \frac{1}{|S_c|} \sum_{j \in S_c} \mathbf{A}_{nj}^i \tag{5}$$

where i denotes the index of an attention head, n denotes the index of the last input token, S_I and S_c represent the index sets of image and context tokens, respectively. R_I^i and

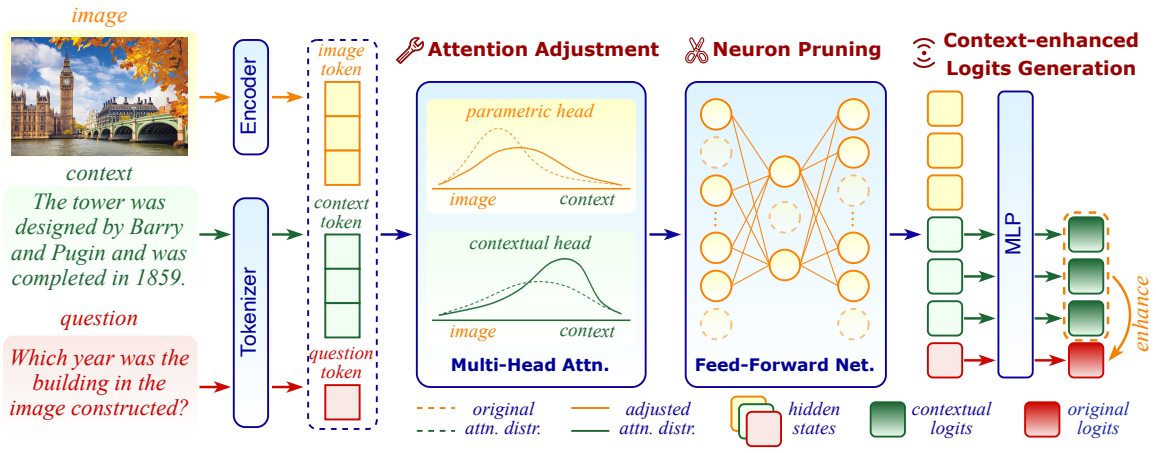

Figure 2: Improving contextual knowledge utilization via modifying three components in LVLMs.

R_c^i thus reflect the relative importance of image and context tokens for the i-th attention head. We define a ratio $\beta^i = R_I^i / R_c^i$ to classify attention heads into two categories: heads with $\beta^i > 1$ are deemed as *parametric heads*, as they predominantly attend to image tokens and rely on parametric knowledge, conversely, heads with $\beta^i < 1$ are categorized as *contextual heads*, as they focus more on contextual information. To encourage the model to better utilize contextual knowledge while preserving feature extraction capability, we adjust the attention weights by decreasing image token attention in parametric heads and increasing context token attention in contextual heads as follows:

$$\mathbf{A}_{nj}^i = \begin{cases} \mathbf{A}_{nj}^i / \beta^i, & \text{if } \beta^i > 1 \text{ and } j \in S_I \\ \mathbf{A}_{nj}^i / \beta^i, & \text{if } \beta^i < 1 \text{ and } j \in S_c \\ \mathbf{A}_{nj}^i, & \text{otherwise} \end{cases} \tag{6}$$

The ratio β^i can be interpreted as a temperature parameter that adaptively increases or decreases the attention weights, guiding the model to focus on contextual knowledge.

Neuron Pruning

As discussed before, we exploit neuron activation patterns to identify coarse-grained parametric knowledge and mitigate its conflicts with the retrieved contextual knowledge. To preserve LVLMs' general capabilities, we only prune the FFN layers with the highest number of activated neurons by randomly deactivating a portion of their neurons as follows:

$$\text{FFN}^l(\mathbf{x}) = \sum_{i=1}^{d} D(\alpha_i, \gamma) \cdot \mathbf{v}_i \tag{7}$$

where l denotes the layer index, D is the deactivation function that randomly sets γ percent of the α_i to zero.

Context-enhanced Logits Generation

As analyzed in the previous section, the attention sink phenomenon (Xiao et al. 2024), where the model predominantly focuses on a small subset of input tokens, can hinder the contribution of context tokens to the final output, thereby restricting the utilization of contextual knowledge. We address this issue by connecting the logits from context tokens to the original logits, explicitly integrating contextual information and enhancing the model's faithfulness to the contextual knowledge. Specifically, we construct the context logits l_{cont} using an exponentially weighted average of the logits from all context tokens to enhance semantic consistency in the logits. The overall logits l_c that model contextual knowledge are obtained by combining the original logits l_{ori} from the last token and the context logits l_{cont} as follows:

$$l_{cont} = \sum_{i \in S_c} \exp(-d_i) \cdot l_i, \quad l_c = l_{ori} + l_{cont} \tag{8}$$

where d_i is the positional distance between the i-th context token and the last token, and l_i represents the logits computed from the i-th context token.

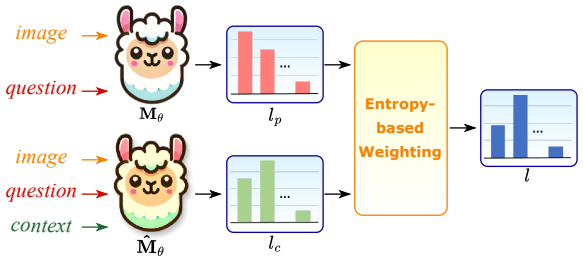

Figure 3: KCM Integrates parametric (up) and contextual (bottom) knowledge through entropy-based weighting.

Finally, as illustrated in Fig. 3, KCM fuses parametric knowledge from the vanilla LVLM M_θ, and contextual knowledge from our modified LVLM $\hat{M}_\theta$ through an entropy-based weighting for the two logits, which can dynamically decide which knowledge to rely on and improve the robustness of KCM when the retrieved context is noisy.

$$l = (1 + H_p) \cdot l_c + (1 + H_c) \cdot l_p \tag{9}$$

where l_c and $H(c)$ are output logits and entropy of next-token probabilities computed by our modified LVLM $\hat{M}_\theta$,

Model	Decoding	Human			Validation		
		Unseen Question	Unseen Entity	Overall	Unseen Question	Unseen Entity	Overall
LLaVA-1.5	Regular	$7.59_{(\pm 0.08)}$	$7.90_{(\pm 0.05)}$	$7.74_{(\pm 0.01)}$	$19.98_{(\pm 0.02)}$	$19.59_{(\pm 0.01)}$	$19.78_{(\pm 0.01)}$
	Parametric	$6.27_{(\pm 0.05)}$	$6.26_{(\pm 0.19)}$	$6.26_{(\pm 0.09)}$	$7.14_{(\pm 0.03)}$	$6.28_{(\pm 0.01)}$	$6.68_{(\pm 0.00)}$
	CD	$7.39_{(\pm 0.21)}$	$7.31_{(\pm 0.06)}$	$7.35_{(\pm 0.09)}$	$20.32_{(\pm 0.01)}$	$19.90_{(\pm 0.00)}$	$20.11_{(\pm 0.01)}$
	AdaCAD	$7.81_{(\pm 0.02)}$	$8.07_{(\pm 0.03)}$	$7.94_{(\pm 0.01)}$	$21.23_{(\pm 0.03)}$	$20.91_{(\pm 0.02)}$	$21.07_{(\pm 0.03)}$
	Entropy	$7.98_{(\pm 0.09)}$	$8.34_{(\pm 0.02)}$	$8.15_{(\pm 0.03)}$	$21.97_{(\pm 0.02)}$	$21.85_{(\pm 0.01)}$	$21.91_{(\pm 0.03)}$
	CAD	$8.02_{(\pm 0.05)}$	$8.15_{(\pm 0.03)}$	$8.08_{(\pm 0.02)}$	$21.68_{(\pm 0.01)}$	$20.93_{(\pm 0.02)}$	$21.30_{(\pm 0.02)}$
	COIECD	$8.78_{(\pm 0.04)}$	$8.65_{(\pm 0.01)}$	$8.71_{(\pm 0.01)}$	$22.43_{(\pm 0.05)}$	$21.73_{(\pm 0.05)}$	$22.07_{(\pm 0.05)}$
	AGLA	$8.74_{(\pm 0.28)}$	$9.18_{(\pm 0.00)}$	$8.94_{(\pm 0.08)}$	$22.34_{(\pm 0.02)}$	$21.88_{(\pm 0.01)}$	$22.11_{(\pm 0.02)}$
	VCD	$9.22_{(\pm 0.00)}$	$9.26_{(\pm 0.00)}$	$9.24_{(\pm 0.02)}$	$22.30_{(\pm 0.03)}$	$22.38_{(\pm 0.03)}$	$22.23_{(\pm 0.03)}$
	Ours	$\mathbf{13.09}_{(\pm 0.09)}$	$\mathbf{11.81}_{(\pm 0.01)}$	$\mathbf{12.42}_{(\pm 0.02)}$	$\mathbf{23.43}_{(\pm 0.00)}$	$\mathbf{23.74}_{(\pm 0.01)}$	$\mathbf{23.58}_{(\pm 0.01)}$
InstructBLIP	Regular	$4.20_{(\pm 0.00)}$	$3.86_{(\pm 0.03)}$	$4.02_{(\pm 0.01)}$	$3.60_{(\pm 0.01)}$	$3.82_{(\pm 0.00)}$	$3.71_{(\pm 0.00)}$
	Parametric	$4.06_{(\pm 0.01)}$	$3.65_{(\pm 0.01)}$	$3.84_{(\pm 0.01)}$	$2.36_{(\pm 0.01)}$	$1.92_{(\pm 0.00)}$	$2.12_{(\pm 0.00)}$
	CD	$4.52_{(\pm 0.03)}$	$3.55_{(\pm 0.01)}$	$3.98_{(\pm 0.01)}$	$3.59_{(\pm 0.01)}$	$4.00_{(\pm 0.00)}$	$3.79_{(\pm 0.00)}$
	AdaCAD	$4.57_{(\pm 0.05)}$	$3.70_{(\pm 0.10)}$	$4.09_{(\pm 0.08)}$	$3.71_{(\pm 0.02)}$	$4.35_{(\pm 0.01)}$	$4.01_{(\pm 0.01)}$
	Entropy	$4.56_{(\pm 0.05)}$	$4.14_{(\pm 0.01)}$	$4.34_{(\pm 0.02)}$	$3.81_{(\pm 0.01)}$	$4.39_{(\pm 0.00)}$	$4.08_{(\pm 0.00)}$
	CAD	$4.52_{(\pm 0.03)}$	$3.55_{(\pm 0.01)}$	$3.98_{(\pm 0.01)}$	$3.77_{(\pm 0.03)}$	$4.43_{(\pm 0.02)}$	$4.08_{(\pm 0.02)}$
	COIECD	$4.64_{(\pm 0.10)}$	$4.08_{(\pm 0.02)}$	$4.33_{(\pm 0.01)}$	$4.07_{(\pm 0.00)}$	$4.54_{(\pm 0.00)}$	$4.30_{(\pm 0.00)}$
	AGLA	$4.80_{(\pm 0.05)}$	$4.28_{(\pm 0.09)}$	$4.52_{(\pm 0.05)}$	$3.74_{(\pm 0.01)}$	$4.10_{(\pm 0.01)}$	$3.91_{(\pm 0.01)}$
	VCD	$4.70_{(\pm 0.01)}$	$4.14_{(\pm 0.01)}$	$4.40_{(\pm 0.00)}$	$3.62_{(\pm 0.01)}$	$4.12_{(\pm 0.00)}$	$3.85_{(\pm 0.00)}$
	Ours	$\mathbf{6.74}_{(\pm 0.02)}$	$\mathbf{6.80}_{(\pm 0.02)}$	$\mathbf{6.77}_{(\pm 0.02)}$	$\mathbf{5.61}_{(\pm 0.01)}$	$\mathbf{6.27}_{(\pm 0.00)}$	$\mathbf{5.92}_{(\pm 0.01)}$
Shikra	Regular	$6.71_{(\pm 0.03)}$	$6.31_{(\pm 0.01)}$	$6.50_{(\pm 0.02)}$	$11.93_{(\pm 0.01)}$	$11.78_{(\pm 0.01)}$	$11.85_{(\pm 0.01)}$
	Parametric	$5.76_{(\pm 0.10)}$	$6.10_{(\pm 0.07)}$	$5.92_{(\pm 0.05)}$	$7.61_{(\pm 0.01)}$	$6.25_{(\pm 0.01)}$	$6.86_{(\pm 0.01)}$
	CD	$8.21_{(\pm 0.00)}$	$7.15_{(\pm 0.01)}$	$7.64_{(\pm 0.01)}$	$12.41_{(\pm 0.00)}$	$11.89_{(\pm 0.00)}$	$12.14_{(\pm 0.00)}$
	AdaCAD	$8.30_{(\pm 0.00)}$	$7.11_{(\pm 0.00)}$	$7.66_{(\pm 0.00)}$	$12.87_{(\pm 0.03)}$	$12.53_{(\pm 0.02)}$	$12.70_{(\pm 0.02)}$
	Entropy	$8.32_{(\pm 0.03)}$	$7.73_{(\pm 0.11)}$	$8.01_{(\pm 0.05)}$	$13.78_{(\pm 0.02)}$	$13.33_{(\pm 0.01)}$	$13.55_{(\pm 0.02)}$
	CAD	$8.16_{(\pm 0.06)}$	$7.16_{(\pm 0.02)}$	$7.62_{(\pm 0.00)}$	$12.99_{(\pm 0.03)}$	$12.51_{(\pm 0.02)}$	$12.75_{(\pm 0.03)}$
	COIECD	$8.32_{(\pm 0.02)}$	$7.73_{(\pm 0.08)}$	$8.01_{(\pm 0.03)}$	$14.46_{(\pm 0.02)}$	$14.21_{(\pm 0.03)}$	$14.33_{(\pm 0.02)}$
	AGLA	$8.24_{(\pm 0.01)}$	$7.56_{(\pm 0.05)}$	$7.88_{(\pm 0.01)}$	$14.29_{(\pm 0.02)}$	$13.91_{(\pm 0.01)}$	$14.08_{(\pm 0.01)}$
	VCD	$8.13_{(\pm 0.05)}$	$7.41_{(\pm 0.03)}$	$7.75_{(\pm 0.04)}$	$13.71_{(\pm 0.03)}$	$13.81_{(\pm 0.03)}$	$13.76_{(\pm 0.03)}$
	Ours	$\mathbf{8.42}_{(\pm 0.02)}$	$\mathbf{8.41}_{(\pm 0.02)}$	$\mathbf{8.41}_{(\pm 0.02)}$	$\mathbf{15.27}_{(\pm 0.02)}$	$\mathbf{14.64}_{(\pm 0.01)}$	$\mathbf{14.95}_{(\pm 0.02)}$
MiniGPT4	Regular	$4.38_{(\pm 0.07)}$	$3.00_{(\pm 0.02)}$	$3.56_{(\pm 0.02)}$	$12.69_{(\pm 0.02)}$	$12.38_{(\pm 0.02)}$	$12.53_{(\pm 0.02)}$
	Parametric	$2.34_{(\pm 0.01)}$	$2.10_{(\pm 0.01)}$	$2.21_{(\pm 0.00)}$	$4.72_{(\pm 0.01)}$	$3.93_{(\pm 0.01)}$	$4.29_{(\pm 0.01)}$
	CD	$4.28_{(\pm 0.00)}$	$2.59_{(\pm 0.00)}$	$3.22_{(\pm 0.00)}$	$14.49_{(\pm 0.01)}$	$14.43_{(\pm 0.00)}$	$14.46_{(\pm 0.01)}$
	AdaCAD	$4.78_{(\pm 0.00)}$	$3.43_{(\pm 0.01)}$	$3.99_{(\pm 0.00)}$	$14.82_{(\pm 0.02)}$	$14.96_{(\pm 0.02)}$	$14.89_{(\pm 0.02)}$
	Entropy	$4.80_{(\pm 0.07)}$	$2.91_{(\pm 0.00)}$	$3.62_{(\pm 0.00)}$	$14.66_{(\pm 0.02)}$	$14.66_{(\pm 0.01)}$	$14.66_{(\pm 0.02)}$
	CAD	$4.97_{(\pm 0.01)}$	$3.44_{(\pm 0.01)}$	$4.07_{(\pm 0.01)}$	$14.83_{(\pm 0.01)}$	$14.94_{(\pm 0.00)}$	$14.88_{(\pm 0.01)}$
	COIECD	$4.57_{(\pm 0.03)}$	$3.40_{(\pm 0.01)}$	$3.90_{(\pm 0.00)}$	$14.87_{(\pm 0.01)}$	$14.67_{(\pm 0.02)}$	$14.77_{(\pm 0.02)}$
	AGLA	$4.67_{(\pm 0.02)}$	$3.63_{(\pm 0.02)}$	$4.09_{(\pm 0.01)}$	$14.31_{(\pm 0.02)}$	$13.92_{(\pm 0.01)}$	$14.11_{(\pm 0.01)}$
	VCD	$4.52_{(\pm 0.03)}$	$3.43_{(\pm 0.01)}$	$3.90_{(\pm 0.01)}$	$14.46_{(\pm 0.01)}$	$14.30_{(\pm 0.02)}$	$14.38_{(\pm 0.02)}$
	Ours	$\mathbf{4.99}_{(\pm 0.02)}$	$\mathbf{3.84}_{(\pm 0.03)}$	$\mathbf{4.34}_{(\pm 0.02)}$	$\mathbf{14.90}_{(\pm 0.03)}$	$\mathbf{15.00}_{(\pm 0.02)}$	$\mathbf{14.95}_{(\pm 0.03)}$

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

modeling the enriched contextual knowledge with less interference. l_p and $H(p)$ are output logits and entropy of next-token probabilities from the vanilla LVLM M_θ with no context provided, modeling the parametric knowledge.

Experiment

Experimental Settings

Datasets. We evaluate our method across three types of knowledge-intensive datasets. *Free-form generative datasets*: **Human** (Chen et al. 2023b) and **Validation** (Chen et al. 2023b). *Multi-choice discriminative datasets*: **Infoseek** (Chen et al. 2023b) and **ViQuAE** (Lerner et al. 2022). *Commonsense knowledge datasets*: **OK-VQA** (Marino et al. 2019), **AOK-VQA** (Schwenk et al. 2022), and **Encyclopedic VQA** (E-VQA) (Mensink et al. 2023).

LVLMs and SOTA methods. We evaluate six representative LVLMs: LLaVA-1.5 (7B) (Liu et al. 2023), InstructBLIP (7B) (Dai et al. 2023), Shikra (7B) (Chen et al. 2023a), and MiniGPT-4 (7B) (Zhu et al. 2023). We then compare knowledge conflict mitigation methods in the LLM domain, including Contrastive Decoding (CD) (Li et al. 2023c), Adaptive Context-Aware Decoding (AdaCAD) (Wang et al. 2024), Entropy-based Decoding (Entropy) (Qiu et al. 2024), Context-Aware Decoding (CAD) (Shi et al. 2024), and COntextual Information-Entropy Constraint Decoding (COIECD) (Yuan et al. 2024). Additionally, we benchmark two hallucination mitigation methods: Visual Contrastive Decoding (VCD) (Leng et al. 2023) and Assembly of Global and Local Attention (AGLA) (An et al. 2024b).

Implementation details. We employ the vision encoder of CLIP-ViT-L/14-336 (Radford et al. 2021) to retrieve knowledge and append the top retrieved item to the input as contextual knowledge. The Wikipedia dumps with associated images provided by (Chen et al. 2023b) are selected as the

Model	Decoding	InfoSeek	ViQuAE	Model	Decoding	InfoSeek	ViQuAE
LLaVA-1.5	Regular	$51.97_{(\pm 0.42)}$	$53.32_{(\pm 0.20)}$	InstructBLIP	Regular	$23.44_{(\pm 0.89)}$	$19.82_{(\pm 0.12)}$
	Parametric	$39.15_{(\pm 0.02)}$	$51.06_{(\pm 0.16)}$		Parametric	$8.73_{(\pm 0.23)}$	$6.53_{(\pm 0.35)}$
	CD	$49.95_{(\pm 0.20)}$	$52.56_{(\pm 0.03)}$		CD	$22.95_{(\pm 0.63)}$	$21.76_{(\pm 0.51)}$
	CAD	$52.08_{(\pm 0.16)}$	$52.99_{(\pm 0.23)}$		CAD	$26.56_{(\pm 0.56)}$	$23.18_{(\pm 0.30)}$
	AdaCAD	$52.30_{(\pm 0.04)}$	$52.99_{(\pm 0.30)}$		AdaCAD	$27.07_{(\pm 0.05)}$	$23.64_{(\pm 0.08)}$
	Entropy	$53.33_{(\pm 0.07)}$	$54.26_{(\pm 0.05)}$		Entropy	$27.50_{(\pm 0.07)}$	$23.19_{(\pm 0.39)}$
	COIECD	$52.08_{(\pm 0.21)}$	$52.99_{(\pm 0.23)}$		COIECD	$25.65_{(\pm 0.05)}$	$20.84_{(\pm 0.05)}$
	VCD	$53.87_{(\pm 0.07)}$	$55.13_{(\pm 0.09)}$		VCD	$23.31_{(\pm 0.07)}$	$20.28_{(\pm 0.53)}$
	AGLA	$53.53_{(\pm 0.50)}$	$54.24_{(\pm 0.21)}$		AGLA	$21.24_{(\pm 0.55)}$	$16.77_{(\pm 0.30)}$
	Ours	$\mathbf{58.67}_{(\pm 0.41)}$	$\mathbf{57.80}_{(\pm 0.01)}$		**Ours**	$\mathbf{31.93}_{(\pm 0.03)}$	$\mathbf{25.20}_{(\pm 0.05)}$
Shikra	Regular	$19.41_{(\pm 0.12)}$	$17.73_{(\pm 0.01)}$	MiniGPT-4	Regular	$25.83_{(\pm 1.42)}$	$24.06_{(\pm 0.46)}$
	Parametric	$9.65_{(\pm 0.14)}$	$10.90_{(\pm 0.04)}$		Parametric	$19.73_{(\pm 0.57)}$	$20.42_{(\pm 1.22)}$
	CD	$24.83_{(\pm 0.20)}$	$21.38_{(\pm 0.03)}$		CD	$26.55_{(\pm 0.31)}$	$20.46_{(\pm 0.76)}$
	CAD	$24.51_{(\pm 0.06)}$	$21.68_{(\pm 0.15)}$		CAD	$27.58_{(\pm 0.43)}$	$23.39_{(\pm 0.07)}$
	AdaCAD	$23.95_{(\pm 0.18)}$	$21.51_{(\pm 0.01)}$		AdaCAD	$28.01_{(\pm 0.17)}$	$23.05_{(\pm 0.61)}$
	Entropy	$24.12_{(\pm 0.03)}$	$21.99_{(\pm 0.08)}$		Entropy	$28.84_{(\pm 0.68)}$	$22.59_{(\pm 0.07)}$
	COIECD	$24.18_{(\pm 0.13)}$	$21.68_{(\pm 0.15)}$		COIECD	$29.44_{(\pm 0.01)}$	$25.94_{(\pm 0.18)}$
	VCD	$25.76_{(\pm 0.14)}$	$23.11_{(\pm 0.72)}$		VCD	$28.74_{(\pm 0.10)}$	$25.45_{(\pm 0.05)}$
	AGLA	$26.26_{(\pm 0.29)}$	$22.72_{(\pm 0.15)}$		AGLA	$29.28_{(\pm 0.87)}$	$27.87_{(\pm 0.08)}$
	Ours	$\mathbf{27.90}_{(\pm 0.09)}$	$\mathbf{23.50}_{(\pm 0.02)}$		**Ours**	$\mathbf{31.29}_{(\pm 0.05)}$	$\mathbf{30.62}_{(\pm 0.00)}$

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

knowledge base. Multinomial sampling is used as the default decoding strategy. To ensure fair comparisons with prior studies (Shi et al. 2024; Leng et al. 2023), we apply adaptive plausibility constraints (Li et al. 2023c) to the final logits. The deactivation hyperparameter γ is set to 0.3 and the deactivation layer is the last layer.

Experimental Results

Experiments on free-form datasets. Tab. 2 presents the experimental results of four representative LVLMs (Liu et al. 2023; Dai et al. 2023; Chen et al. 2023a; Zhu et al. 2023) evaluated on two free-form generative datasets (Chen et al. 2023b). The proposed framework consistently outperforms both the *Regular* decoding strategy and other state-of-the-art decoding methods by a notable margin across all LVLMs and datasets, demonstrating its effectiveness in enhancing retrieval-augmented LVLM performance.

Experiments on multi-choice datasets. In addition to free-form generative datasets, we evaluate our framework on two multi-choice discriminative datasets (Chen et al. 2023b; Lerner et al. 2022; Zhu et al. 2024) using four representative LVLMs (Liu et al. 2023; Dai et al. 2023; Chen et al. 2023a; Zhu et al. 2023). As shown in Tab. 3, our framework yields an average improvement of 6.4% over the *Regular* decoding strategy and consistently outperforms state-of-the-art decoding methods by substantial margins, demonstrating its effectiveness across diverse retrieval-augmented tasks.

Experiments on commonsense knowledge datasets. Beyond the entity knowledge-centric datasets (Chen et al. 2023b; Lerner et al. 2022), we further evaluate our framework on widely-used commonsense knowledge-based benchmarks, including OK-VQA (Marino et al. 2019), AOK-VQA (Schwenk et al. 2022), and Encyclopedic VQA (E-VQA) (Mensink et al. 2023), using LLaVA-1.5 (Liu et al. 2023). As shown in Tab. 4, the proposed framework significantly outperforms the *Regular* decoding and consistently exceeds the performance of state-of-the-art methods, highlighting its effectiveness across a broader spectrum of knowledge-intensive tasks.

Model	OK-VQA	AOK-VQA	E-VQA
Regular	$46.17_{(\pm 0.12)}$	$44.13_{(\pm 0.00)}$	$19.14_{(\pm 2.83)}$
Parametric	$45.13_{(\pm 0.40)}$	$43.23_{(\pm 1.02)}$	$5.34_{(\pm 0.01)}$
CD	$55.00_{(\pm 0.03)}$	$51.67_{(\pm 0.44)}$	$28.62_{(\pm 0.68)}$
CAD	$56.43_{(\pm 0.46)}$	$53.93_{(\pm 0.56)}$	$28.62_{(\pm 0.76)}$
AdaCAD	$57.10_{(\pm 0.04)}$	$54.40_{(\pm 0.48)}$	$28.33_{(\pm 0.42)}$
Entropy	$56.27_{(\pm 0.05)}$	$53.93_{(\pm 0.26)}$	$29.24_{(\pm 0.11)}$
COIECD	$56.43_{(\pm 0.46)}$	$53.93_{(\pm 0.56)}$	$28.24_{(\pm 0.74)}$
VCD	$57.80_{(\pm 0.13)}$	$57.40_{(\pm 0.12)}$	$27.71_{(\pm 0.33)}$
AGLA	$57.53_{(\pm 0.05)}$	$55.40_{(\pm 0.63)}$	$28.29_{(\pm 0.56)}$
Ours	$\mathbf{60.90}_{(\pm 0.01)}$	$\mathbf{60.57}_{(\pm 0.06)}$	$\mathbf{29.95}_{(\pm 0.03)}$

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

Discussion

Ablation Study

We perform ablation studies on both multi-choice and commonsense knowledge-based datasets (Chen et al. 2023b; Schwenk et al. 2022) to evaluate the contribution of each component in the proposed framework based on LLaVA-1.5 (Liu et al. 2023). As illustrated in Fig. 4 (Left), the incorporation of Attention Adjustment (**+Attention**) facilitates improved utilization of contextual knowledge by LVLMs, thereby enhancing overall performance. The Neuron Pruning mechanism (**+Pruning**) reduces interference from coarse-grained parametric knowledge, further enabling more effective use of external knowledge and contributing to performance improvements. The context-enhanced Logits module (**+Context**) strengthens the model's alignment with contextual knowledge, thereby increasing its faithfulness to the input context. Additionally, the integration of parametric and retrieved knowledge through entropy-based weighting (**+Entropy**) balances the benefits of both knowledge types

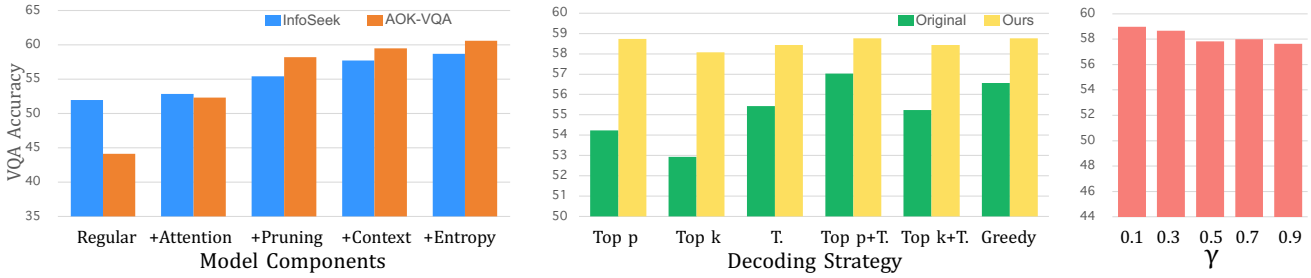

Figure 4: **Left:** Ablation studies with different model variants on the InfoSeek and AOK-VQA dataset. **Middle:** Results with different decoding strategies on the InfoSeek dataset. **Right:** Results with different hyperparameters γ on the InfoSeek dataset.

Model	Recall↑	ACC↑	MR↓	CR↑
GT Know. + LLaVA	100	65.5	46.9	53.1
GT Know. + Ours	100	**71.2**	**39.7**	**60.3**
1st Know. + LLaVA	58.4	52.0	56.2	43.8
1st Know. + Ours	58.4	**58.7**	**49.8**	**50.2**
2nd Know. + LLaVA	10.6	38.8	53.2	46.8
2nd Know. + Ours	10.6	**44.9**	**50.7**	**49.3**

Table 5: Evaluation results using ground-truth contextual knowledge (GT Know.), first (1st Know.), and second (2nd Know.) retrieved knowledge. Recall measures the quality of contextual knowledge. Accuracy (ACC) assesses response quality. Memorization Ratio (MR) quantifies the model's reliance on parametric knowledge, while Context Ratio (CR) indicates its preference for contextual knowledge.

while mitigating the influence of noisy retrieved content, resulting in further performance gains.

Effect of Different Decoding Strategies

In addition to the multinomial sampling decoding strategy discussed in this study, we further investigate the effectiveness of several alternative decoding methods using LLaVA-1.5 (Liu et al. 2023) on the multi-choice InfoSeek dataset (Chen et al. 2023b). Specifically, we evaluate another six decoding strategies including Top-P sampling (Holtzman et al. 2019) ($p = 0.5$), Top-K sampling (Fan, Lewis, and Dauphin 2018) ($k = 50$), greedy decoding (DeVore and Temlyakov 1996), temperature sampling (Ackley, Hinton, and Sejnowski 1985) ($t = 0.5$), Top-P sampling with temperature ($p = 0.5$ and $t = 0.5$) and Top-K sampling with temperature ($k = 50$ and $t = 0.5$). As illustrated in Fig. 4 (Middle), the proposed framework yields substantial performance improvements across all decoding strategies, highlighting its robustness in enhancing LVLM capabilities.

Effect of Neuron Pruning Ratio

We examine the impact of the neuron pruning ratio γ, as defined in Eq. 7, on the multi-choice InfoSeek dataset (Chen et al. 2023b) using LLaVA-1.5 (Liu et al. 2023). As shown in Fig. 4 (Right), the proposed framework is not sensitive to the variations in γ, demonstrating its robustness.

Effect of Contextual Knowledge Quality

We investigate the influence of contextual knowledge quality on performance using the multi-choice InfoSeek dataset (Chen et al. 2023b) with LLaVA-1.5 (Liu et al. 2023), as summarized in Tab. 5. The results indicate that the proposed framework consistently achieves substantial performance improvements across varying levels of contextual knowledge quality, demonstrating its robustness and general effectiveness in the presence of retrieval noise. Furthermore, our framework exhibits a lower memorization ratio and a higher context ratio compared to the original LVLM, suggesting that it can leverage the retrieved contextual knowledge more effectively to produce accurate responses. Notably, the entropy-based weighting mechanism enables the model to dynamically adjust its reliance on parametric knowledge when the quality of contextual input deteriorates, further contributing to its adaptability.

Conclusion

In this work, we present KCM, a training-free and plug-and-play framework designed to enhance the reliability and accuracy of LVLMs in knowledge-intensive tasks. KCM enhances LVLM performance by strategically modifying the model architecture through three key mechanisms: adaptive attention redistribution, neuron pruning, and contextual logit amplification. These modifications enable more reliable integration of retrieved contextual knowledge while reducing interference from coarse-grained parametric knowledge, thereby supporting more accurate response generation. Extensive experiments demonstrate that KCM consistently outperforms state-of-the-art approaches across diverse LVLMs and benchmark datasets, all without requiring additional training or external resources. These results underscore the effectiveness and practicality of KCM in advancing the knowledge utilization capabilities of LVLMs, contributing to the development of more trustworthy and context-aware retrieval-augmented systems.

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

Ackley, D. H.; Hinton, G. E.; and Sejnowski, T. J. 1985. A learning algorithm for Boltzmann machines. *Cognitive science*, 9(1): 147–169.

Al-Shamayleh, A. S.; Adwan, O.; Alsharaiah, M. A.; Hussein, A. H.; Kharma, Q. M.; and Eke, C. I. 2024. A comprehensive literature review on image captioning methods and metrics based on deep learning technique. *Multimedia Tools and Applications*, 83(12): 34219–34268.

An, W.; Nie, J.; Wu, Y.; Tian, F.; Lu, S.; and Zheng, Q. 2025. Empowering Multimodal LLMs with External Tools: A Comprehensive Survey. *arXiv preprint arXiv:2508.10955*.

An, W.; Shi, W.; Tian, F.; Lin, H.; Wang, Q.; Wu, Y.; Cai, M.; Wang, L.; Chen, Y.; Zhu, H.; et al. 2024a. Generalized Category Discovery with Large Language Models in the Loop. In *Findings of the Association for Computational Linguistics ACL 2024*, 8653–8665.

An, W.; Tian, F.; Leng, S.; Nie, J.; Lin, H.; Wang, Q.; Dai, G.; Chen, P.; and Lu, S. 2024b. AGLA: Mitigating Object Hallucinations in Large Vision-Language Models with Assembly of Global and Local Attention. *arXiv preprint arXiv:2406.12718*.

An, W.; Tian, F.; Nie, J.; Shi, W.; Lin, H.; Chen, Y.; Wang, Q.; Wu, Y.; Dai, G.; and Chen, P. 2024c. Knowledge Acquisition Disentanglement for Knowledge-based Visual Question Answering with Large Language Models. *arXiv preprint arXiv:2407.15346*.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023b. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Caffagni, D.; Cocchi, F.; Moratelli, N.; Sarto, S.; Cornia, M.; Baraldi, L.; and Cucchiara, R. 2024. Wiki-LLaVA: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1818–1826.

Chen, K.; Zhang, Z.; Zeng, W.; Zhang, R.; Zhu, F.; and Zhao, R. 2023a. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. *arXiv preprint arXiv:2306.15195*.

Chen, Y.; Hu, H.; Luan, Y.; Sun, H.; Changpinyo, S.; Ritter, A.; and Chang, M.-W. 2023b. 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*, 14948–14968.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv preprint arXiv:2306.04387*.

DeVore, R. A.; and Temlyakov, V. N. 1996. Some remarks on greedy algorithms. *Advances in computational Mathematics*, 5(1): 173–187.

Fan, A.; Lewis, M.; and Dauphin, Y. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.

Fan, Y.; Mu, Y.; Wang, Y.; Huang, L.; Ruan, J.; Li, B.; Xiao, T.; Huang, S.; Feng, X.; and Zhu, J. 2025. SLAM: Towards Efficient Multilingual Reasoning via Selective Language Alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, 9499–9515.

Geva, M.; Schuster, R.; Berant, J.; and Levy, O. 2021. Transformer Feed-Forward Layers Are Key-Value Memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 5484–5495.

Gong, T.; Lyu, C.; Zhang, S.; Wang, Y.; Zheng, M.; Zhao, Q.; Liu, K.; Zhang, W.; Luo, P.; and Chen, K. 2023. MultiModal-GPT: A Vision and Language Model for Dialogue with Humans. *arXiv preprint arXiv:2305.04790*.

Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; and Chang, M. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, 3929–3938. PMLR.

Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi, Y. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Huang, P.; Liu, Z.; Yan, Y.; Yi, X.; Chen, H.; Liu, Z.; Sun, M.; Xiao, T.; Yu, G.; and Xiong, C. 2025. Pip-kag: Mitigating knowledge conflicts in knowledge-augmented generation via parametric pruning. *arXiv preprint arXiv:2502.15543*.

Jin, Z.; Cao, P.; Chen, Y.; Liu, K.; Jiang, X.; Xu, J.; Qiuxia, L.; and Zhao, J. 2024. Tug-of-War between Knowledge: Exploring and Resolving Knowledge Conflicts in Retrieval-Augmented Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 16867–16878.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9579–9589.

Leng, S.; Zhang, H.; Chen, G.; Li, X.; Lu, S.; Miao, C.; and Bing, L. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*.

Lerner, P.; Ferret, O.; Guinaudeau, C.; Le Borgne, H.; Besançon, R.; Moreno, J. G.; and Lovón Melgarejo, J. 2022. 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*, 3108–3120.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Li, K.; Patel, O.; Viégas, F.; Pfister, H.; and Wattenberg, M. 2023b. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36: 41451–41530.

Li, X. L.; Holtzman, A.; Fried, D.; Liang, P.; Eisner, J.; Hashimoto, T. B.; Zettlemoyer, L.; and Lewis, M. 2023c. 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)*, 12286–12312.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. *arXiv preprint arXiv:2304.08485*.

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

Mensink, T.; Uijlings, J.; Castrejon, L.; Goel, A.; Cadar, F.; Zhou, H.; Sha, F.; Araujo, A.; and Ferrari, V. 2023. Encyclopedic VQA: Visual questions about detailed properties of fine-grained categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3113–3124.

Qiu, Z.; Ou, Z.; Wu, B.; Li, J.; Liu, A.; and King, I. 2024. Entropy-based decoding for retrieval-augmented large language models. *arXiv preprint arXiv:2406.17519*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Schwenk, D.; Khandelwal, A.; Clark, C.; Marino, K.; and Mottaghi, R. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, 146–162. Springer.

Shi, W.; Han, X.; Lewis, M.; Tsvetkov, Y.; Zettlemoyer, L.; and Yih, W.-t. 2024. 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)*, 783–791.

Sun, S.; An, W.; Tian, F.; Nan, F.; Liu, Q.; Liu, J.; Shah, N.; and Chen, P. 2024. A review of multimodal explainable artificial intelligence: Past, present and future. *arXiv preprint arXiv:2412.14056*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Wang, H.; Prasad, A.; Stengel-Eskin, E.; and Bansal, M. 2024. AdaCAD: Adaptively Decoding to Balance Conflicts between Contextual and Parametric Knowledge. *arXiv preprint arXiv:2409.07394*.

Xiao, G.; Tian, Y.; Chen, B.; Han, S.; and Lewis, M. 2024. Efficient Streaming Language Models with Attention Sinks. In *The Twelfth International Conference on Learning Representations*.

Xu, R.; Qi, Z.; Guo, Z.; Wang, C.; Wang, H.; Zhang, Y.; and Xu, W. 2024. Knowledge Conflicts for LLMs: A Survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 8541–8565.

Yuan, X.; Yang, Z.; Wang, Y.; Liu, S.; Zhao, J.; and Liu, K. 2024. Discerning and Resolving Knowledge Conflicts through Adaptive Decoding with Contextual Information-Entropy Constraint. *arXiv preprint arXiv:2402.11893*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv preprint arXiv:2304.10592*.

Zhu, T.; Liu, Q.; Wang, F.; Tu, Z.; and Chen, M. 2024. Unraveling Cross-Modality Knowledge Conflicts in Large Vision-Language Models. *arXiv preprint arXiv:2410.03659*.