Visual-Prior Guided and MLLM-Enhanced Dual Retrieval for Knowledge-based VQA

Ji Yan[1,2], Jingzi Gu[*1,2], Ruicheng Xiong[1,2], Gengqi Yang[1,2],
Guimiao Yang[1,2], Dayan Wu[1,2], Peng Fu[1,2], Zheng Lin[1,2]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
{yanji, gujingzi, xiongruicheng, yanggengqi, yangguimiao, wudayan, fupeng, linzheng}@iie.ac.cn

***Abstract*—Knowledge-based Visual Question Answering (KB-VQA) necessitates retrieving external knowledge to answer questions about images. Despite the progress of Retrieval-Augmented Generation (RAG), existing paradigms face significant limitations. In particular, visual-only retrieval often neglects valuable semantic information in questions, whereas employing Multimodal Large Language Models (MLLMs) to generate semantically relevant queries from images is prone to severe entity hallucinations due to the lack of grounding. To address these challenges, we propose a novel dual-stream retrieval framework, where the core module of the text-to-text stream is Visual-Prior Guided Retrieval-Augmented Captioning (ViP-RAC). Specifically, ViP-RAC leverages retrieval results from the image stream as visual priors to anchor MLLMs, effectively mitigating hallucinations and generating factually aligned search queries. Furthermore, we employ a Weighted Reciprocal Rank Fusion (RRF) algorithm to synergistically integrate the visual-based and text-based retrieval streams, ensuring robust knowledge acquisition. Extensive experiments on the Encyclopedic-VQA and InfoSeek benchmarks demonstrate that our approach achieves new state-of-the-art performance.**

***Index Terms*—visual question answering, retrieval-augmented generation, multimodal large language model**

I. Introduction

Multimodal Large Language Models (MLLMs) [1]–[5] have emerged as unified foundations for various vision-language tasks, demonstrating remarkable capabilities in standard Visual Question Answering (VQA) [6]–[8]. However, despite their impressive visual reasoning skills, MLLMs continue to face significant challenges in addressing Knowledge-based Visual Question Answering (KB-VQA) [9]–[13]. Unlike standard VQA, which relies primarily on visual perception to interpret image content, KB-VQA necessitates accessing extensive external encyclopedic knowledge grounded in visual cues, extending beyond the model's internal parametric knowledge. To overcome the limitations of MLLMs in memorizing long-tail or up-to-date information, Retrieval-Augmented Generation (RAG) [14] has become a widely adopted paradigm. By retrieving relevant documents from external knowledge bases (KB) to augment the generation process, RAG effectively bridges the gap between visual understanding and world knowledge.

Despite these advancements, existing methods [15]–[20] still face challenges in achieving high retrieval recall. A primary limitation is their predominant reliance on visual modality matching. While effective for capturing visual appearance, this paradigm fails to utilize the rich semantic information embedded in the original question. To leverage this information, recent works like Wiki-PRF [20] attempt to fuse image and question inputs using MLLMs to generate descriptive queries. However, this strategy introduces a critical issue: Hallucination in Ungrounded Query Generation. Lacking external context, MLLMs often fabricate specific entities based on superficial visual features. As shown in Fig. 1, when asked about a railway station, a vanilla MLLM misidentifies *"Montreal Central Station"* as *"Union Station"* due to architectural similarities. Such hallucinations are catastrophic for RAG systems, as they generate misleading search queries that retrieve entirely irrelevant documents, causing errors to propagate inevitably into the final answer.

To address these challenges, we propose a novel dual-stream retrieval framework, where the core module of the text retrieval stream is ViP-RAC (**Vi**sual-**P**rior Guided **R**etrieval **A**ugmented **C**aptioning). The design of ViP-RAC is motivated by our observation that while coarse-grained visual retrieval results may not always pinpoint the exact target, they reliably capture visually similar entities or scene contexts, thereby serving as powerful anchors for the MLLM. Specifically, by utilizing the textual information associated with the top-k retrieved candidates from the image retrieval stream as visual priors, ViP-RAC effectively calibrates the MLLM's generation process. This mechanism enables the production of factually aligned captions with significantly reduced hallucinations. As illustrated in Fig. 1, even if the generated query does not immediately pinpoint the exact entity name, our method successfully corrects the geographic scope (identifying *"Montreal"* instead of *"Toronto"*), ultimately leading to the successful recall of the target document.

Furthermore, our framework enhances retrieval robustness through superior modality alignment and synergized fusion. Distinct from prior methods like Wiki-PRF [20] that typically perform cross-modal retrieval, our approach utilizes the generated captions to retrieve the textual content of documents. This Text-to-Text (T2T) paradigm ensures superior semantic alignment by operating within a unified modality space. Finally, to integrate the results from both streams, we employ a Weighted Reciprocal Rank Fusion (RRF) [21] algorithm. This strategy

*Corresponding author.

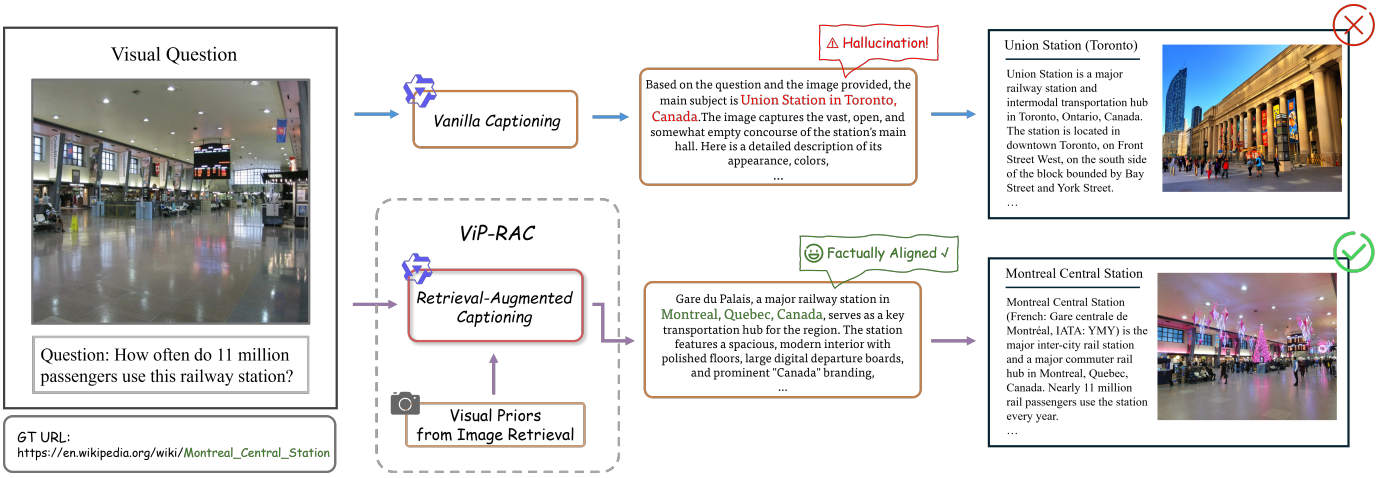

Fig. 1. **Illustration of vanilla captioning and our ViP-RAC.** By leveraging retrieved candidates from the visual stream as visual priors, ViP-RAC effectively mitigates the hallucination issues inherent in vanilla MLLM-generated queries, ultimately leading to the successful retrieval of the correct document.

prioritizes candidates that rank highly across both modalities, ensuring that the final candidate set remains robust against noise from any single stream.

Our main contributions are summarized as follows:

- We propose a dual-stream retrieval framework synergized via Weighted RRF, where textual retrieval complements the visual stream to address semantic neglect and enhance overall recall, ensuring single-modal noise robustness.
- We introduce ViP-RAC as the core module of the text-to-text stream, which leverages image retrieval results as visual priors to anchor MLLMs, thereby mitigating hallucinations in ungrounded query generation. Notably, our design is generator-agnostic and can be seamlessly integrated into existing RAG systems.
- Extensive experiments on E-VQA and InfoSeek benchmarks demonstrate the superiority of our approach, establishing new state-of-the-art performance of 40.8 and 43.4, respectively.

II. Related Work

Knowledge-based VQA [22] transcends standard recognition by requiring external knowledge, evolving from commonsense reasoning [9], [10] to fine-grained, Wikipedia-scale settings like Encyclopedic-VQA [12] and InfoSeek [13]. To address the memory limitations of MLLMs in these scenarios, RAG has become the standard paradigm, typically utilizing contrastive image-text encoders [23], [24] to align visual inputs with large-scale knowledge bases. Recent advancements focus on refining this pipeline, such as employing hierarchical verification [16] or Q-Former-based reranking [17], or introducing model-level reasoning controls via learnable tokens [18], [20]. Despite progress, methods such as Wiki-PRF [20] that utilize generic MLLMs for query generation suffer from a critical limitation whereby ungrounded query generation often leads to entity hallucinations. Such hallucinations cause the retriever to fetch entirely irrelevant documents, which inevitably introduces noise into the context window and degrades the final answer accuracy. In contrast, our approach uniquely leverages retrieval results from the image stream as visual priors to ground MLLMs, effectively rectifying these hallucinations before query generation.

III. Method

A. Task Definition

Given a visual question consisting of an image I and a natural language question Q, the goal of KB-VQA is to generate an accurate answer A. Unlike standard VQA tasks, KB-VQA requires the generator model G to leverage additional snippets $\mathcal{C}$ retrieved from an external knowledge base $\mathcal{K}$ as context. The objective can therefore be formulated as:

$$A = \arg\max_{A} G(A \mid I, Q, \mathcal{C}). \tag{1}$$

The external knowledge base $\mathcal{K}$ consists of a large-scale collection of multimodal documents, denoted as $\mathcal{K} = \{D_i\}_{i=1}^{N}$. Formally, each document D_i is defined as a triplet:

$$D_i = (T_i, S_i, V_i), \tag{2}$$

where T_i represents the metadata (e.g., title, summary, and section titles) of the i-th document, S_i denotes the full textual content, and V_i is the associated visual content.

The overall architecture of our proposed framework is illustrated in Fig. 2. It consists of three core components: Image Retriever, Text Retriever, and Answer Generator. Initially, the Image Retriever performs image-to-image retrieval to identify candidate images, which are subsequently mapped to their corresponding documents in the knowledge base. Subsequently, the Text Retriever incorporates our proposed ViP-RAC module. Instead of relying on ungrounded query generation, ViP-RAC leverages the retrieved candidates from the visual stream as visual priors to anchor the MLLM, effectively correcting entity hallucinations and generating factually aligned queries for precise textual retrieval. Finally, retrieval results from both streams are synergistically integrated via a

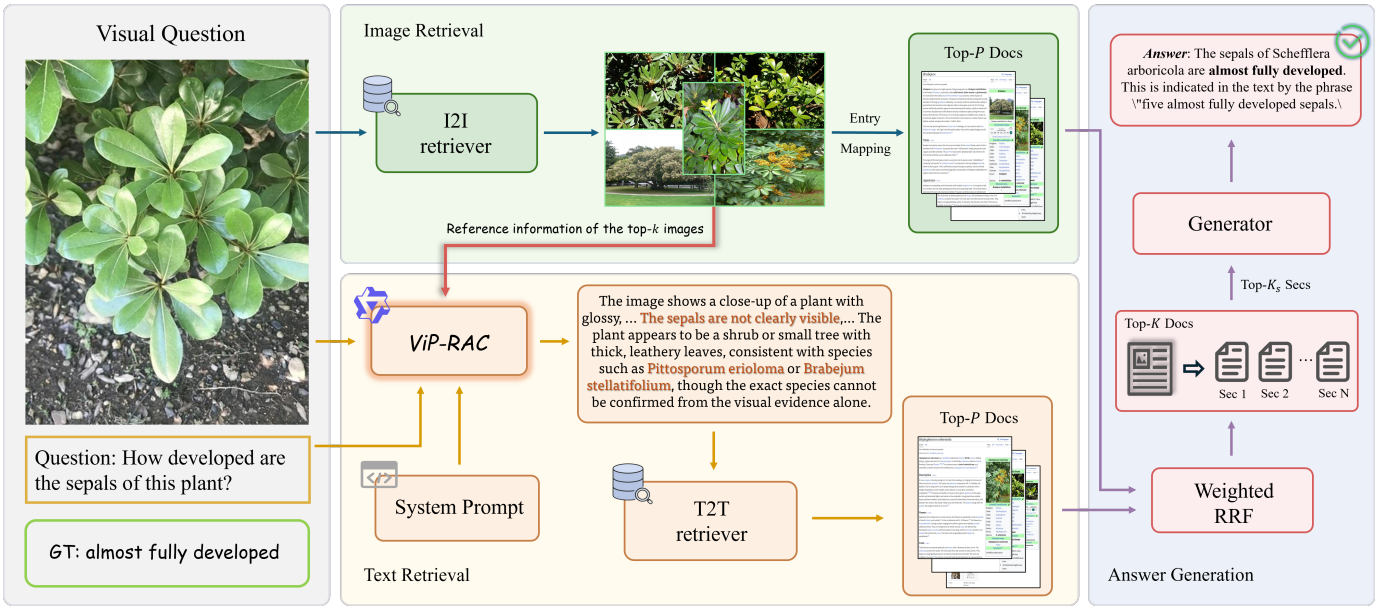

Fig. 2. **The overall view of our proposed dual-stream retrieval framework.** (1) Given a visual question, the Image Retriever performs coarse-grained Image-to-Image (I2I) retrieval to obtain top-P images, where the relevant information from the top-5 images serves as visual priors for the text-stream MLLM. (2) In the Text Retriever stream, the **ViP-RAC** module utilizes the textual descriptions of these visual priors to provide a factual basis for the MLLM, generating a factually aligned caption (query) for the Text-to-Text (T2T) retriever. (3) Candidate documents from both streams are fused via a Weighted RRF algorithm, and the top-K_s relevant sections are fed into the Generator to derive the final answer.

weighted RRF algorithm to support the Generator in deriving the final answer.

B. Retrieval Stage

To effectively recall relevant knowledge from the massive external knowledge base, we employ a dual-stream retrieval strategy that synergizes visual appearance matching with semantic textual search.

Image Retrieval. Given the query image I_q, the objective of the visual stream is to retrieve candidate documents based on visual similarity. We adopt the visual encoder from EVA-CLIP [24], denoted as Φ_{vis}, to extract visual vectors. The similarity score S_{vis} between the query image and the visual content V_i of the i-th document in the knowledge base is computed via cosine similarity:

$$S_{\text{vis}}(I_q, V_i) = \frac{\Phi_{\text{vis}}(I_q) \cdot \Phi_{\text{vis}}(V_i)}{\|\Phi_{\text{vis}}(I_q)\| \|\Phi_{\text{vis}}(V_i)\|}. \tag{3}$$

Based on these scores, we retrieve the top-P_1 most relevant images and map them to their corresponding documents, forming the visual candidate set $\mathcal{D}_{\text{vis}}$. Beyond serving as candidate answers, the associated textual information of the top-ranked images in $\mathcal{D}_{\text{vis}}$ provides coarse-grained but vital visual cues. We leverage the textual content of the top-k results from this set to construct the visual priors, denoted as $\mathcal{P}_{vis}$, which are subsequently utilized to support the text retrieval stream.

Text Retrieval. This stream focuses on extracting the visual content pertinent to the question as well as the key semantic entities inherent in the question. To bridge the gap between visual perception and textual search, we deploy our **ViP-RAC** module. Specifically, we employ a MLLM, denoted as $\mathcal{M}$, to serve as the query generator. The model takes the query image I_q, the user question Q, and the relevant information of the visual priors $\mathcal{P}_{vis}$ as inputs to generate a refined search query $\hat{Q}$:

$$\hat{Q} = \mathcal{M}(I_q, Q, \mathcal{P}_{vis}). \tag{4}$$

By conditioning on $\mathcal{P}_{vis}$, the MLLM is grounded in the retrieved visual context, significantly reducing hallucinations. Subsequently, we perform dense retrieval using the generated query $\hat{Q}$. We encode both $\hat{Q}$ and the textual content S_i of each document into dense vectors using a text embedding model, denoted as Ψ_{text}. The textual relevance score is computed as:

$$S_{\text{text}}(\hat{Q}, S_i) = \frac{\Psi_{\text{text}}(\hat{Q}) \cdot \Psi_{\text{text}}(S_i)}{\|\Psi_{\text{text}}(\hat{Q})\| \|\Psi_{\text{text}}(S_i)\|}. \tag{5}$$

This process retrieves the top-P_2 textually relevant documents, denoted as $\mathcal{D}_{\text{text}}$, which complement the visual candidates by addressing semantic details.

C. Generation Stage

Weighted Reciprocal Rank Fusion. Since the visual and textual retrieval streams operate in distinct embedding spaces, their similarity scores S_{vis} and S_{text} may follow different distributions, making direct summation suboptimal. To robustly integrate the results of our dual-stream retrieval, we employ the weighted RRF algorithm [21]. RRF relies on the rank information rather than absolute scores, ensuring stability against outliers. Let $r_{\text{vis}}(d)$ and $r_{\text{text}}(d)$ be the rank of

TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS ON E-VQA AND INFOSEEK DATASETS. BEST RESULTS ARE IN **BOLD**, AND SECOND BEST ARE UNDERLINED.

Method	Generator	Retriever	E-VQA		InfoSeek		
			Single-hop	All	Unseen-Q	Unseen-E	All
Zero-shot MLLMs							
BLIP-2 [25]	–	–	12.6	12.4	12.7	12.3	12.5
LLaVA-v1.5-7B [26]	–	–	16.3	16.9	9.6	9.4	9.5
GPT-4V [27]	–	–	26.9	28.1	15.0	14.3	14.6
Qwen2.5-VL-3B [2]	–	–	17.9	19.6	20.4	21.9	21.4
Qwen2.5-VL-7B [2]	–	–	21.7	20.3	22.8	24.1	23.7
Retrieval-Augmented Models							
$\text{DPR}^{\dagger}_{\text{V+T}}$ [28]	Multi-passage BERT	CLIP ViT-B/32	29.1	–	–	–	12.4
RORA-VLM† [29]	Vicuna-7B	CLIP+Google Search	–	20.3	25.1	27.3	–
Wiki-LLaVA [16]	Vicuna-7B	CLIP ViT-L/14	17.7	20.3	30.1	27.8	28.9
EchoSight† [17]	Mistral-7B	EVA-CLIP-8B	41.8	–	–	–	31.3
EchoSight* [17]	Mistral-7B	EVA-CLIP-8B	41.7	38.2	25.2	26.2	25.7
ReflectiVA [18]	LLaMA-3.1-8B	EVA-CLIP-8B	35.5	35.5	40.4	39.8	40.1
MMKB-RAG [19]	Qwen2-7B	EVA-CLIP-8B	39.7	35.9	36.4	36.3	36.4
Wiki-PRF [20]	Qwen-2.5VL-7B	EVA-CLIP-8B	37.1	36.0	43.3	42.7	42.8
	InternVL3-8B	EVA-CLIP-8B	40.1	39.2	43.5	42.1	42.5
ViP-RAC (Ours)	Mistral-7B | LLaMA-3.1-8B	EVA-CLIP-8B+Qwen3-Emb-0.6B	43.5	39.6	41.5	41.4	41.3
	Qwen-2.5VL-7B	EVA-CLIP-8B+Qwen3-Emb-0.6B	**44.6**	**40.8**	**43.7**	**43.2**	**43.4**

* Represents our reproductions.
† Indicates results that are not directly comparable because different knowledge bases were utilized.

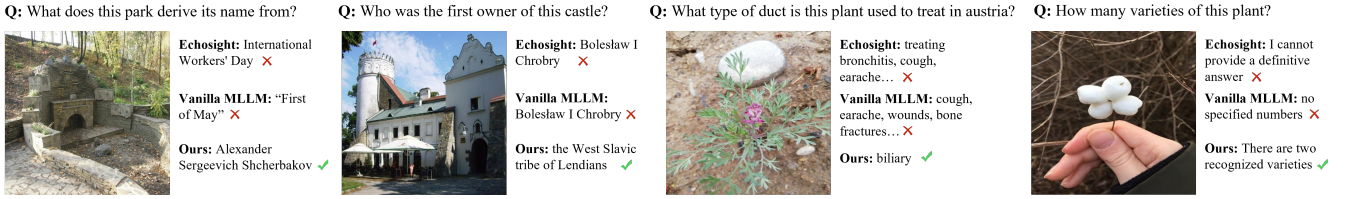

Fig. 3. **Qualitative examples of our method.**The examples illustrate that our method accurately identifies specific entities and retrieves factually correct knowledge across diverse domains, including landmarks (left two) and natural species (right two).

document d in the retrieval lists $\mathcal{D}_{\text{vis}}$ and $\mathcal{D}_{\text{text}}$, respectively. The fusion score $S_{\text{RRF}}(d)$ is calculated as:

$$S_{\text{RRF}}(d) = w_v \cdot \frac{1}{\eta + r_{\text{vis}}(d)} + w_t \cdot \frac{1}{\eta + r_{\text{text}}(d)}, \quad (6)$$

where η is a smoothing constant, and w_v, w_t are hyperparameters that assign specific weights to the visual and textual streams, respectively. Based on S_{RRF}, we re-rank the union of retrieved documents and retain the top-K candidates, denoted as $\mathcal{D}_{\text{merged}}$.

Answer Generation. Following the generation strategy in EchoSight [17], we employ a pre-trained Q-Former to perform section-level re-ranking on the candidates $\mathcal{D}_{\text{merged}}$, retaining the top-$K_s$ most relevant sections as the final context $\mathcal{C}_{\text{final}}$. Subsequently, the answer is generated by the generator G. For LLM-based generators, the input consists of the question Q and $\mathcal{C}_{\text{final}}$, whereas for MLLM-based generators, the query image I is additionally included to provide visual evidence.

IV. EXPERIMENTS

A. Experimental Setup

Datasets and Metrics. We evaluate on E-VQA [12] and InfoSeek [13]. Following established protocols [18], [20], we use the E-VQA test split with a 2M-document Knowledge Base and the InfoSeek validation split with a 100k-document subset. Retrieval is measured via Recall@K, while answer quality is evaluated using the BEM score [30] for E-VQA and VQA accuracy [7] for InfoSeek.

Baselines. We compare ViP-RAC against two categories of previous methods: (1) Zero-shot MLLMs: General-purpose models including BLIP-2 [25], LLaVA-v1.5 [26], GPT-4V [27], and Qwen2.5-VL [2], which rely solely on internal parametric knowledge. (2) Retrieval-Augmented Models: Specialized KB-VQA frameworks such as Wiki-LLaVA [16], EchoSight [17], ReflectiVA [18], and Wiki-PRF [20]. These baselines represent diverse retrieval and reranking strategies for performance comparison.

TABLE II
RETRIEVAL PERFORMANCE COMPARISON ON E-VQA AND INFOSEEK.

Method	E-VQA			InfoSeek		
	R@1	R@5	R@20	R@1	R@5	R@20
Wiki-LLaVA	3.3	–	13.2	36.9	–	71.9
EchoSight	13.3	31.3	48.8	45.6	67.1	77.9
EchoSight*	13.4	31.8	48.9	46.5	64.3	73.4
ReflectiVA	15.6	36.1	49.8	56.1	**77.6**	86.4
Wiki-PRF	–	–	–	54.9	–	–
ViP-RAC (Ours)	**28.1**	**43.0**	**58.2**	**60.2**	77.4	**86.7**

* represents our reproductions.

TABLE III
ABLATION STUDY ON E-VQA. **VIP** DENOTES THE VISUAL-PRIOR GUIDED RAC MODULE.

Text Encoder	Components			E-VQA	
	I2I	T2T	ViP	Single-hop	All
Single Stream Retrieval					
(Visual Only)	✓			41.73	38.15
Qwen3-Emb-0.6B		✓		31.47	29.32
Qwen3-Emb-0.6B		✓	✓	41.18	37.44
Dual Stream Fusion					
BM25	✓	✓		33.49	31.10
EVA-CLIP Text	✓	✓		42.17	38.42
Qwen3-Emb-0.6B	✓	✓		43.12	39.25
Qwen3-Emb-0.6B (Ours)	✓	✓	✓	**43.54**	**39.63**

Implementation Details. For the Image Retriever, we utilize the EVA-CLIP-8B [24] as the visual encoder to compute image-to-image similarity. For the Text Retriever, we employ Qwen3-VL-8B-Instruct as the backbone for the ViP-RAC module to generate search queries, and Qwen3-Embedding-0.6B as the dense encoder for vectorizing both queries and knowledge base documents. For the Answer Generator, we evaluate several backbones, including Mistral-7B [31], LLaMA-3.1-8B [32], and Qwen2.5-VL-7B [2].

During retrieval, each stream fetches the top 50 candidates ($P_1 = P_2 = 50$), utilizing the top-5 visual results as priors ($k = 5$). In the fusion stage, the Weighted RRF algorithm uses a smoothing constant $\eta = 60$, with weights $w_v = 1.2$ and $w_t = 1.0$ assigned to the visual and text streams, respectively. We retain the top 20 documents ($K = 20$) for section-level re-ranking. Based on dataset characteristics, the number of input sections K_s fed to the generator is set to 1 for E-VQA and 3 for InfoSeek. All vector similarity searches are accelerated using the FAISS [33] library.

B. Main Results

VQA Results. As shown in Table I, ViP-RAC significantly improves upon zero-shot MLLMs, boosting Qwen2.5-VL-7B from 20.3 to 40.8 on E-VQA, confirming the necessity of external knowledge. Compared to existing RAG frameworks, ViP-RAC consistently achieves superior performance. On E-VQA, it surpasses state-of-the-art methods like EchoSight and Wiki-PRF. Notably, even with less powerful generators like Mistral-7B, ViP-RAC reaches 39.6, indicating that performance gains stem primarily from high-quality retrieved context rather than the inherent capacity of the generator. On InfoSeek, ViP-RAC establishes a new benchmark with a score of 43.4, outperforming Wiki-PRF. This result demonstrates that guidance via visual priors can effectively mitigate the performance degradation caused by hallucination in ungrounded query generation.

Retrieval Results. Then we evaluate the effectiveness of our retrieval module in identifying relevant knowledge. Table II compares the retrieval recall of ViP-RAC against state-of-the-art methods. On E-VQA, ViP-RAC substantially outperforms ReflectiVA, achieving an R@1 of 28.1%, representing a nearly two-fold increase over the 15.6% baseline. This confirms that visual priors effectively anchor the MLLM, facilitating precise query generation for target entities. Similarly, on InfoSeek, our method consistently leads in both R@1 (60.2%) and R@20 (86.7%), illustrating that our visual-prior guided dual-stream architecture robustly handles diverse queries and suppresses retrieval noise.

Qualitative Results. Fig. 3 compares our method with EchoSight and a vanilla MLLM. Guided by visual priors, our method accurately identifies fine-grained entities where competitors either hallucinate generic answers or provide incorrect details. These cases confirm that our dual-stream framework effectively anchors query generation to retrieve precise, factually aligned knowledge for complex reasoning.

C. Ablation Study

To validate the effectiveness of the proposed components, we conduct ablation studies on the E-VQA dataset. Table III details the performance under different retrieval configurations.

Impact of Visual-Prior Guidance. We first evaluate the individual components in a single-stream retrieval setting. As shown in the first section of Table III, the visual-only retrieval already achieves a competitive score of 38.15, establishing a strong baseline. In contrast, relying solely on the T2T retriever yields unsatisfactory results (29.32) due to the hallucination issues inherent in ungrounded query generation. However, incorporating the ViP module leads to a substantial performance leap of over 8 points, boosting the overall score to 37.44. This empirical evidence confirms that visual priors effectively ground the MLLM, correcting hallucinations and ensuring factually aligned query generation.

Effectiveness of Dual-Stream Fusion. We further analyze the synergy between visual and textual streams. In the dual-stream setting, naive lexical matching (BM25) introduces noise, degrading performance compared to the visual-only baseline. Conversely, dense retrievers (EVA-CLIP text encoder and Qwen3-Embedding-0.6B) provide complementary semantic information, improving overall accuracy. Notably, our full framework achieves a top score of 39.63 by combining I2I retrieval with ViP-enhanced T2T retrieval. This demonstrates

that our dual-stream architecture successfully integrates visual appearance cues with explicit semantic constraints, yielding robust retrieval results.

V. Conclusion

In this paper, we presented a novel dual-stream retrieval framework for KB-VQA, incorporating ViP-RAC as the core module of the text-to-text stream. By leveraging coarse-grained retrieval results as visual priors, ViP-RAC anchors MLLMs to generate factually aligned queries, effectively mitigating zero-shot hallucinations. Furthermore, we employ a Weighted RRF algorithm to synergize visual and textual streams for robust knowledge acquisition. Extensive experiments on E-VQA and InfoSeek benchmarks demonstrate that our approach significantly outperforms existing methods, establishing new state-of-the-art results.

VI. Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62472420.

References

[1] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 716–23 736, 2022.

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

[3] D. Caffagni, F. Cocchi, L. Barsellotti, N. Moratelli, S. Sarto, L. Baraldi, M. Cornia, and R. Cucchiara, "The revolution of multimodal large language models: a survey," *arXiv preprint arXiv:2402.12451*, 2024.

[4] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

[5] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 26 296–26 306.

[6] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[7] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6904–6913.

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

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

[10] D. Schwenk, A. Khandelwal, C. Clark, K. Marino, and R. Mottaghi, "A-okvqa: A benchmark for visual question answering using world knowledge," in *European conference on computer vision*. Springer, 2022, pp. 146–162.

[11] S. Shah, A. Mishra, N. Yadati, and P. P. Talukdar, "Kvqa: Knowledge-aware visual question answering," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 8876–8884.

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

[13] Y. Chen, H. Hu, Y. Luan, H. Sun, S. Changpinyo, A. Ritter, and M.-W. Chang, "Can pre-trained vision and language models answer visual information-seeking questions?" *arXiv preprint arXiv:2302.11713*, 2023.

[14] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[15] Y. Ding, J. Yu, B. Liu, Y. Hu, M. Cui, and Q. Wu, "Mukea: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5089–5098.

[16] D. Caffagni, F. Cocchi, N. Moratelli, S. Sarto, M. Cornia, L. Baraldi, and R. Cucchiara, "Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1818–1826.

[17] Y. Yan and W. Xie, "Echosight: Advancing visual-language models with wiki knowledge," *arXiv preprint arXiv:2407.12735*, 2024.

[18] F. Cocchi, N. Moratelli, M. Cornia, L. Baraldi, and R. Cucchiara, "Augmenting multimodal llms with self-reflective tokens for knowledge-based visual question answering," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 9199–9209.

[19] Z. Ling, Z. Guo, Y. Huang, Y. An, S. Xiao, J. Lan, X. Zhu, and B. Zheng, "Mmkb-rag: A multi-modal knowledge-based retrieval-augmented generation framework," *arXiv preprint arXiv:2504.10074*, 2025.

[20] Y. Hong, J. Gu, Q. Yang, L. Fan, Y. Wu, Y. Wang, K. Ding, S. Xiang, and J. Ye, "Knowledge-based visual question answer with multimodal processing, retrieval and filtering," *arXiv preprint arXiv:2510.14605*, 2025.

[21] G. V. Cormack, C. L. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms condorcet and individual rank learning methods," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 2009, pp. 758–759.

[22] J. Deng, Z. Wu, H. Huo, and G. Xu, "A comprehensive survey of knowledge-based vision question answering systems: The lifecycle of knowledge in visual reasoning task," *arXiv preprint arXiv:2504.17547*, 2025.

[23] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[24] Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao, "Eva-clip: Improved training techniques for clip at scale," *arXiv preprint arXiv:2303.15389*, 2023.

[25] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[26] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 26 296–26 306.

[27] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[28] P. Lerner, O. Ferret, and C. Guinaudeau, "Cross-modal retrieval for knowledge-based visual question answering," in *European Conference on Information Retrieval*. Springer, 2024, pp. 421–438.

[29] J. Qi, Z. Xu, R. Shao, Y. Chen, J. Di, Y. Cheng, Q. Wang, and L. Huang, "Rora-vlm: Robust retrieval-augmented vision language models," *arXiv preprint arXiv:2410.08876*, 2024.

[30] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[31] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," 2023. [Online]. Available: https://arxiv.org/abs/2310.06825

[32] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv e-prints*, pp. arXiv–2407, 2024.

[33] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou, "The faiss library," *IEEE Transactions on Big Data*, 2025.