Dynamic Multi-Path Retrieval for Knowledge-based Visual Question Answering

Zeyu Song[1], **Yimin Deng**[1,2], **Yuxin Zhang**[1], **Guoshuai Zhao**[1*], **Chengxu Liu**[1], **Jialie Shen**[3] and **Xueming Qian**[4]
[1]School of Software Engineering, Xi'an Jiaotong University
[2]City University of Hong Kong
[3]City St George's, University of London
[4]School of Information and Communication Engineering, Xi'an Jiaotong University
guoshuai.zhao@xjtu.edu.cn

Abstract

Knowledge-based Visual Question Answering (KB-VQA) requires models to answer visual questions by reasoning over external knowledge beyond the given image. Existing approaches suffer from two main limitations. First, candidate knowledge is often retrieved in a single modality, either textual or visual, which prevents effective use of heterogeneous and complementary evidence. Second, current approaches typically employ fixed fusion weights, ignoring the varying importance of modalities for different queries. Consequently, irrelevant evidence is introduced while critical knowledge may be overlooked. To address these issues, we propose Dynamic Multi-Path Retrieval for KB-VQA (DMRAG). Our framework retrieves candidates through multiple retrieval paths that capture complementary visual and semantic cues. It then performs Question-Adaptive Gated Fusion (QGF) to balance contributions from different modalities according to the query's information need. The fused candidates are further refined via multimodal rearrangement to support accurate answer generation. Experiments on the E-VQA and InfoSeek datasets show that DMRAG improves both retrieval recall and answer accuracy over prior methods, demonstrating its effectiveness for KB-VQA. Our code is available at https://github.com/qwqq335/DMRAG.

1 Introduction

Visual Question Answering (VQA) [Antol *et al.*, 2015; Wu *et al.*, 2023] integrates visual perception and language understanding and has become a standard benchmark for evaluating the reasoning capability of Multimodal Large Language Models (MLLMs) [Liu *et al.*, 2023; Wang *et al.*, 2024]. Recent progress in large-scale multimodal pretraining has led to substantial improvements in both accuracy and efficiency on a range of perception-centric tasks, including object recognition, counting, and attribute prediction. In contrast, Knowledge-based Visual Question Answering (KB-VQA) requires reasoning over external knowledge that is not directly observable in the image, and thus remains a persistent challenge [Wu *et al.*, 2024]. Current models rely heavily on parameterized knowledge acquired during pretraining, which provides limited coverage of long-tail entities and time-sensitive facts. As a result, these models are prone to knowledge hallucination and factual obsolescence during inference, as illustrated in the top row of Figure 1.

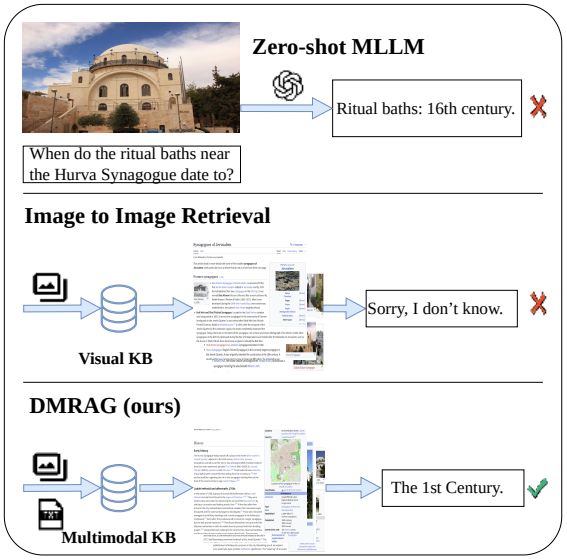

Figure 1: A zero-shot MLLM may answer without sufficient external grounding, resulting in hallucinations or factual mismatch. Image-to-image retrieval over a visual-only knowledge base often returns weakly related candidates and misses question-critical evidence. In contrast, DMRAG retrieves complementary evidence from a multimodal knowledge base and dynamically fuses visual and semantic cues for more reliable answer generation.

Recent studies adopt Retrieval-Augmented Generation (RAG) [Fan *et al.*, 2024] to tackle KB-VQA by supplementing an MLLM's parametric knowledge with evidence retrieved from external knowledge databases [Deng *et al.*, 2026a]. By grounding generation in retrieved documents, RAG expands coverage over long-tail entities and time-

*Corresponding author.

sensitive facts, thereby reducing hallucinations and enabling more accurate answer generation in open-domain scenarios where the answer cannot be inferred from the image alone.

Despite its promise, reliable retrieval in multimodal settings remains challenging. Existing RAG-based KB-VQA approaches often suffer from two issues. First, many methods [Yan and Xie, 2024; Yang *et al.*, 2025] retrieve candidate evidence from a single modality, which fails to fully exploit the heterogeneous and complementary signals in multimodal knowledge bases and may miss question-critical cues. Second, even when multiple modalities are involved [Hong *et al.*, 2025], fusion is frequently static or coarse-grained, and cannot adapt to how different questions rely on visual evidence versus semantic knowledge [Deng *et al.*, 2026b]. Consequently, irrelevant information may dominate the retrieved set, while crucial evidence is overlooked at the initial recall stage, which further propagates to downstream reranking and answer generation and degrades overall performance.

To address these limitations, we propose the Dynamic Multi-Path Retrieval framework (DMRAG). DMRAG explicitly decomposes retrieval into complementary views and coordinates them adaptively [Shang *et al.*, 2023; Zhao *et al.*, 2023; Wu *et al.*, 2024]. The framework comprises two retrieval paths. The first is a fine-grained visual retrieval path that localizes question-relevant image regions and retrieves visually matched knowledge. The second is a generative semantic retrieval path that produces a hypothetical textual description and uses it to perform text-to-text retrieval over an external knowledge base. This design is motivated by the observation that visual similarity alone is unreliable when questions hinge on abstract concepts or factual context not visible in the image, while direct text retrieval suffers from the mismatch between short questions and long encyclopedic passages. To bridge this gap, we adapt Hypothetical Document Embeddings (HyDE) [Gao *et al.*, 2023]: we generate a hypothetical document that approximates the target evidence and use it as the semantic query. A Question-Adaptive Gated Fusion (QGF) network estimates the relative importance of visual and semantic evidence and dynamically fuses the retrieved candidates. The fused candidates are further refined by a multimodal rearrangement module to support answer generation. Overall, DMRAG enables flexible, query-dependent use of multimodal knowledge during retrieval.

The main contributions are summarized as follows:

- We propose DMRAG, a dynamic multi-path retrieval framework for KB-VQA that retrieves candidates through complementary ROI-based visual and HyDE-based semantic pathways, and fuses their evidence via QGF module to improve retrieval relevance and enable more accurate answer generation.
- We introduce hypothetical document generation for text-to-text retrieval in KB-VQA, which improves retrieval relevance over unstructured textual knowledge.
- Extensive experiments on the E-VQA and InfoSeek benchmarks show that DMRAG consistently improves both retrieval recall and answer accuracy in practice compared with existing methods.

2 Related Work

2.1 Multimodal Large Language Models

Recent NLP and LLM studies have explored knowledge-enhanced learning and structural-semantic reasoning to improve model reliability in complex tasks [Deng *et al.*, 2025]. MLLMs integrate visual and language modalities through large-scale visual-language pretraining, demonstrating strong cross-modal understanding and generation capabilities. They support tasks such as visual question answering, image understanding, and cross-modal inference under zero-shot or few-shot settings. With improved scale and pretraining strategies, MLLMs also exhibit open-domain knowledge reasoning capabilities, making them a foundation for knowledge-enhanced visual question answering.

2.2 Knowledge-based Visual Question Answering

Knowledge-based Visual Question Answering (KB-VQA) focuses on problems that cannot be answered solely based on image content. Datasets such as OK-VQA [Marino *et al.*, 2019] and A-OKVQA [Schwenk *et al.*, 2022] involve questions requiring external knowledge, including general commonsense knowledge that is not unique. E-VQA [Mensink *et al.*, 2023] and InfoSeek [Chen *et al.*, 2023] further extend this by including images and textual content from Wikipedia, requiring models to retrieve relevant background knowledge of entities for reasoning and answer generation.

Early research mainly explored incorporating RAG into MLLMs. For example, Wiki LaVA [Caffagni *et al.*, 2024] adopts a hierarchical retrieval strategy from coarse to fine, first performing document-level recall, followed by paragraph-level filtering, and injecting the retrieved text as context, effectively mitigating knowledge noise. EchoSight [Yan and Xie, 2024] proposes a retrieval pipeline of visual first, multimodal rearrangement to balance visual and semantic relevance while maintaining high recall. From the perspective of retrieval granularity, OMGM [Yang *et al.*, 2025] gradually narrows the retrieval space through a multi-granularity collaborative mechanism, achieving a more favorable trade-off between efficiency and accuracy.

Beyond retrieval structure design, some works focus on retrieval noise reduction and knowledge effectiveness. Wiki-PRF [Hong *et al.*, 2025] introduces a three-stage process-retrieval-filter framework, leveraging reinforcement learning to guide query formulation and explicitly filter candidate knowledge. MMKB-RAG [Ling *et al.*, 2025] utilizes the intrinsic knowledge boundaries of the model to generate semantic labels and jointly screen retrieval results. In the generation stage, approaches such as ReflectiVA [Cocchi *et al.*, 2025] and mR^{2}AG [Zhang *et al.*, 2024] incorporate reflection mechanisms, allowing the model to evaluate the adequacy of retrieved content and trigger re-retrieval when necessary, thus enhancing the feedback capability of the RAG system. Additionally, mkg-RAG [Yuan *et al.*, 2025] integrates multimodal knowledge graphs into RAG to enhance multi-hop reasoning.

Overall, existing KB-VQA methods have progressively advanced in hierarchical retrieval, rearrangement and filtering, reflection mechanisms, and knowledge structuring. Nevertheless, fully exploiting multimodal knowledge bases and mod-

eling queries from multiple complementary retrieval perspectives remains an open problem. Motivated by this, this work proposes DMRAG to enable dynamic collaborative fusion of retrieval results from multiple sources.

3 Method

3.1 Overview

The task of KB-VQA requires reasoning over multimodal inputs by leveraging external knowledge. We utilize a multimodal knowledge base comprising paired text articles and associated images as the source for retrieval. As shown in Figure 2, we design two complementary retrieval pathways. The ROI-based visual pathway focuses on fine-grained entities by extracting salient regions and encoding region-level features for precise visual matching. In parallel, the HyDE-based semantic pathway generates a hypothetical encyclopedic paragraph to enrich the query, improving alignment with the knowledge base text and narrowing the visual–semantic gap. We then perform Question-Adaptive Gated Fusion (QGF) to merge candidates from both pathways into a unified evidence pool. Finally, a cross-modal reranker refines the fused set and selects the most relevant evidence for answer generation.

3.2 ROI-based Visual Retrieval

Global image embeddings often include background cues that dilute the evidence of the entity in the question [Zhu *et al.*, 2015]. Thus, retrieval based only on global features may return irrelevant images. To address this issue, we adopt an Entity-Driven Salient Region Retrieval strategy. The core idea is to localize the Region of Interest (ROI) corresponding to the query entity and perform region-level image retrieval.

Entity Extraction

We first identify the key subject in the question Q. We use spaCy to parse Q and extract named entities and noun phrases. To cover the main entities in the question, we retain the top-M entities, where $M = 3$:

$$\mathcal{E} = \{e_i \in \psi_{\text{spaCy}}(Q) \mid i \leq M\}. \tag{1}$$

ROI Localization

To ground the extracted entities in the visual content, we use Grounding DINO [Liu *et al.*, 2024b] as an open-set object detector. Given the image I and the entity set $\mathcal{E}$, the detector outputs candidate bounding boxes with confidence scores:

$$\mathcal{O} = \bigcup_{e \in \mathcal{E}} \text{GroundingDINO}(I, e) = \{(b_k, s_k)\}_{k=1}^{N}, \tag{2}$$

where b_k denotes the coordinates of the k-th box and s_k is its confidence score. We select the ROI as the box with the highest confidence:

$$b^* = \arg\max_{b_k \in \mathcal{O}}(s_k). \tag{3}$$

Visual Retrieval

We crop I using b^* to suppress background distractors. The cropped region is encoded by the frozen EVA-CLIP image encoder f_v, and the feature is ℓ_2 normalized:

$$\mathbf{v}_{\text{ROI}} = \frac{f_v(\text{Crop}(I, b^*))}{\|f_v(\text{Crop}(I, b^*))\|_2}. \tag{4}$$

All knowledge base images are pre-encoded into $\{\mathbf{v}_j^{\text{KB}}\}$ with the same encoder and indexed by FAISS. We compute cosine similarity to score each candidate:

$$s_{\text{ROI}}(k_j) = \mathbf{v}_{\text{ROI}}^{\top} \mathbf{v}_j^{\text{KB}}. \tag{5}$$

We then retrieve the top-K visual candidates:

$$\mathcal{R}_{\text{ROI}} = \underset{k_j \in \mathcal{K}}{\text{TopK}}\Big(s_{\text{ROI}}(k_j)\Big). \tag{6}$$

3.3 HyDE-based Semantic Retrieval

Visual similarity alone is unreliable when the question depends on abstract concepts or factual context that is not visible in the image. Direct text retrieval also suffers from a mismatch between short questions and long encyclopedic passages. We bridge this mismatch using a generated hypothetical passage as the semantic query. Specifically, we generate a hypothetical document that approximates the target evidence, and we use it as the semantic query.

Hypothetical Document Generation

We prompt a Multimodal Large Language Model, denoted as $\mathcal{M}$, to synthesize a paragraph d_{hypo}. The generation is conditioned on the image I, the question Q, and an instruction prompt $\mathcal{P}_{\text{HyDE}}$ that enforces a Wikipedia-style description:

$$d_{\text{hypo}} = \mathcal{M}(I, Q, \mathcal{P}_{\text{HyDE}}). \tag{7}$$

This paragraph rewrites the user intent into a form that is closer to the knowledge base text distribution.

Semantic Retrieval

We encode d_{hypo} with the frozen EVA-CLIP text encoder f_t and normalize it to obtain the semantic query vector:

$$\mathbf{q}_{\text{HyDE}} = \frac{f_t(d_{\text{hypo}})}{\|f_t(d_{\text{hypo}})\|_2}. \tag{8}$$

The knowledge base texts are pre-encoded as $\{\mathbf{t}_j\}$. We score each entry by cosine similarity:

$$s_{\text{HyDE}}(k_j) = \mathbf{q}_{\text{HyDE}}^{\top} \mathbf{t}_j. \tag{9}$$

We then retrieve the top-K semantic candidates:

$$\mathcal{R}_{\text{HyDE}} = \underset{k_j \in \mathcal{K}}{\text{TopK}}\Big(s_{\text{HyDE}}(k_j)\Big). \tag{10}$$

3.4 QGF module

The ROI-based path emphasizes visual attributes, while the HyDE-based path captures factual context. Their relative importance varies across questions, so fixed fusion weights are suboptimal. We therefore learn an adaptive fusion module [Mi *et al.*, 2025] to predict question-specific weights.

Explicit Supervision Construction

To supervise the gating module, we build oracle weights based on retrieval quality. Let k^* be the ground-truth knowledge entry. For each path, we compute the rank of k^* among its top-K results:

$$r_{\text{path}} = \begin{cases} \text{Rank}(k^*, \mathcal{R}_{\text{path}}) & \text{if } k^* \in \mathcal{R}_{\text{path}}, \\ K + 1 & \text{otherwise}, \end{cases} \tag{11}$$

where path $\in \{\text{ROI}, \text{HyDE}\}$. We convert ranks into relevance scores via $v_{\text{path}} = (r_{\text{path}} + \epsilon)^{-1}$ and normalize them with Softmax to obtain oracle weights:

$$\mathbf{w}^* = [w^*_{\text{ROI}}, w^*_{\text{HyDE}}] = \text{Softmax}([v_{\text{ROI}}, v_{\text{HyDE}}]). \tag{12}$$

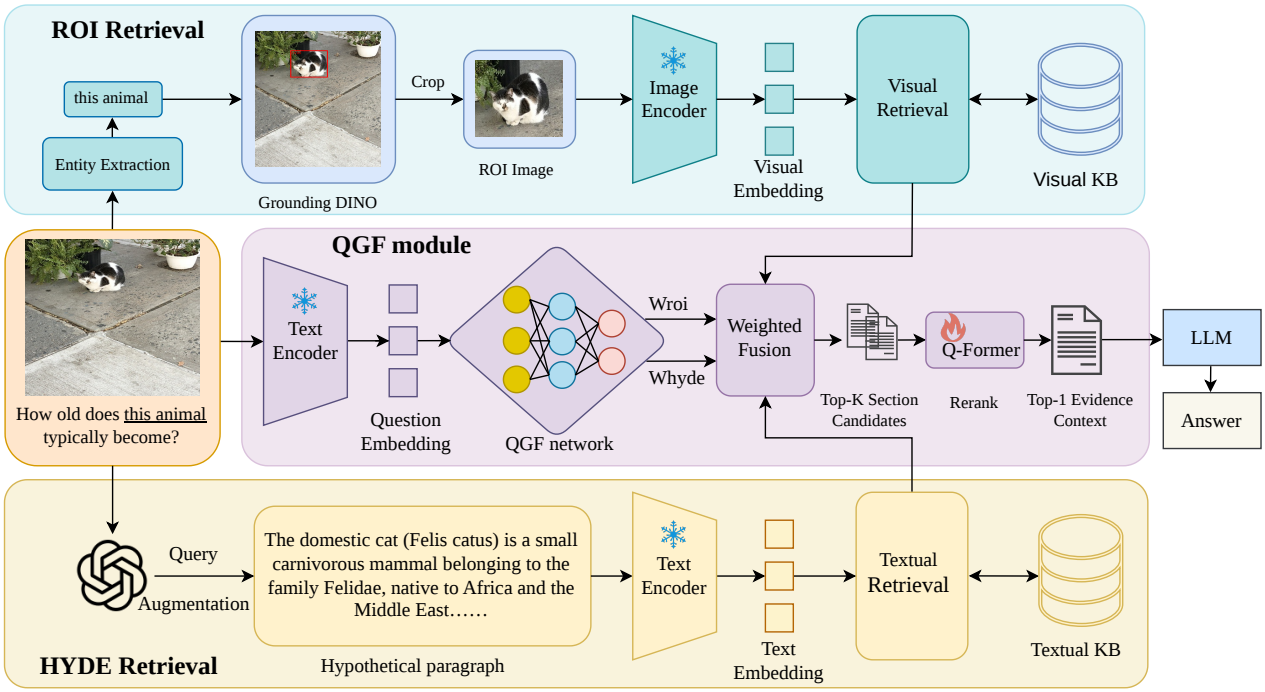

Figure 2: **Overview of the DMRAG framework.** The architecture consists of three primary components. (1) The ROI Retrieval Path identifies entities derived from the question using Grounding DINO and encodes the cropped regions for fine-grained image-to-image retrieval. (2) The HyDE Retrieval Path utilizes an MLLM to generate hypothetical encyclopedic paragraphs, enabling text-to-text retrieval to align visual context with external knowledge. (3) QGF module and Generation employs a gating network to dynamically predict importance weights (W_{roi}, W_{hyde}) based on the question intent. The system fuses the candidate lists using these weights and uses a Q-Former to rerank the results, selecting the most relevant evidence for the LLM to generate the final answer.

Gating Network and Optimization

We use a lightweight MLP G_θ as the gating network. It maps the question embedding $\mathbf{q}$ to predicted fusion weights:

$$\mathbf{w}_{\text{pred}} = G_\theta(\mathbf{q}). \tag{13}$$

We train G_θ by minimizing the KL divergence between $\mathbf{w}^*$ and $\mathbf{w}_{\text{pred}}$, where w_p denotes the p-th element of $\mathbf{w}_{\text{pred}}$:

$$\mathcal{L}_{\text{gate}} = \mathcal{D}_{\text{KL}}(\mathbf{w}^* \| \mathbf{w}_{\text{pred}}) = \sum_p w_p^* \log\left(\frac{w_p^*}{w_p + \epsilon}\right). \tag{14}$$

Score Normalization and Fusion

At inference time, the two retrieval scores may follow different distributions, which makes them hard to combine directly. We apply Min-Max normalization within each path:

$$\tilde{s}_p(k_j) = \frac{s_p(k_j) - \min(\mathcal{S}_p)}{\max(\mathcal{S}_p) - \min(\mathcal{S}_p)}. \tag{15}$$

We then compute the fused score using the predicted weights:

$$s_{\text{fused}}(k_j) = w_{\text{ROI}} \cdot \tilde{s}_{\text{ROI}}(k_j) + w_{\text{HyDE}} \cdot \tilde{s}_{\text{HyDE}}(k_j). \tag{16}$$

Candidates are ranked by s_{fused}, and the top-K entries are passed to reranking.

3.5 Reranking and Answer Generation

The retrieval stage is designed for high recall, so its top results can still include distractors. To further improve precision, we adopt a Q-Former-based reranking module following EchoSight [Yan and Xie, 2024]. The reranker takes the image, the question, and each retrieved text candidate as input and computes fine-grained relevance through cross-attention. The highest-ranked knowledge entry is then used as context for the Large Language Model, which produces an answer explicitly grounded in the selected evidence.

4 Experiments

4.1 Datasets

We evaluate our method on two KB-VQA benchmark datasets, E-VQA [Mensink *et al.*, 2023] and InfoSeek [Chen *et al.*, 2023]. E-VQA contains 221,000 question–answer pairs associated with 16,700 fine-grained Wikipedia entities. The images in this dataset are sourced from iNaturalist 2021 (iNat21) and the Google Landmarks Dataset v2 (GLDv2). E-VQA is equipped with a knowledge base consisting of approximately 2 million Wikipedia documents, each containing a title, textual paragraphs, and related images. In our experiments, we use the full knowledge base provided by E-VQA.

InfoSeek contains approximately 1.3 million visual information-seeking questions, covering more than 11K visual entities from the OVEN dataset [Hu *et al.*, 2023]. We use a subset of 100,000 documents constructed in [Yan and Xie, 2024] as our knowledge base.

4.2 Evaluation Metrics

We evaluate our method from two aspects: retrieval performance and question answering performance.

Method	E-VQA				InfoSeek			
	R@1	R@5	R@10	R@20	R@1	R@5	R@10	R@20
CLIP I-T [Radford *et al.*, 2021]	3.3	7.7	12.1	16.5	32.0	54.0	61.6	68.2
Wiki-LLaVA [Caffagni *et al.*, 2024]	3.3	-	9.9	13.2	36.9	-	66.1	71.9
LLM-RA [Jian *et al.*, 2024]	-	-	-	-	47.3	53.8	-	-
mR^{2}AG [Zhang *et al.*, 2024]	-	-	-	-	38.0	-	65.0	71.0
ReflectiVA [Cocchi *et al.*, 2025]	15.6	36.1	-	49.8	56.1	77.6	-	**86.4**
EchoSight [Yan and Xie, 2024]								
w/o. reranking	13.3	31.3	41.0	48.8	45.6	67.1	73.0	77.9
w. reranking	36.5	47.9	48.8	48.8	53.2	74.0	77.4	77.9
OMGM [Yang *et al.*, 2025]								
w/o. reranking	19.1	41.2	49.8	58.7	52.6	73.9	80.0	84.8
w. reranking	42.8	55.7	58.1	58.7	64.0	80.8	83.6	84.8
DMRAG (ours)								
w/o. reranking	49.0	64.6	71.5	77.7	57.5	77.3	82.4	85.7
w. reranking	**55.0**	**70.3**	**74.8**	**77.7**	**65.5**	**81.7**	**84.0**	85.7

Table 1: Comparison of different methods on E-VQA and InfoSeek datasets.

Model	LLM	Retriever	Feature	E-VQA		InfoSeek		
				Single-Hop	All	Unseen-Q	Unseen-E	All
Zero-shot MLLMs								
BLIP-2 [Li *et al.*, 2023b]	Flan-T5$_{XL}$	-	-	12.6	12.4	12.7	12.3	12.5
InstructBLIP [Dai *et al.*, 2023]	Flan-T5$_{XL}$	-	-	11.9	12.0	8.9	7.4	8.1
LLaVA-v1.5 [Liu *et al.*, 2024a]	Vicuna-7B	-	-	16.3	16.9	9.6	9.4	9.5
LLaVA-v1.5 [Liu *et al.*, 2024a]	LLaMA-3.1-8B	-	-	16.0	16.9	8.3	8.9	7.8
Qwen2-VL-Instruct [Wang *et al.*, 2024]	Qwen-2-7B	-	-	16.4	16.4	17.9	17.8	17.9
GPT-4V [Achiam *et al.*, 2023]	-	-	-	26.9	28.1	15.0	14.3	14.6
Qwen2.5-VL-3B (Base) [Bai *et al.*, 2025]	Qwen2.5-3B	-	-	17.9	19.6	20.4	21.9	21.4
Qwen2.5-VL-7B (Base) [Bai *et al.*, 2025]	Qwen2.5-7B	-	-	21.7	20.3	22.8	24.1	23.7
Retrieval-Augmented Models								
Wiki-LLaVA [Caffagni *et al.*, 2024]	Vicuna-7B	CLIP ViT-L/14+Contriever	Textual	17.7	20.3	30.1	27.8	28.9
Wiki-LLaVA [Caffagni *et al.*, 2024]	LLaMA-3.1-8B	CLIP ViT-L/14+Contriever	Textual	18.3	19.6	28.6	25.7	27.1
EchoSight [Yan and Xie, 2024]	Mistral-7B/LLaMA-3-8B	EVA-CLIP-8B	Visual	19.4	-	-	-	27.7
EchoSight [Yan and Xie, 2024]	LLaMA-3.1-8B	EVA-CLIP-8B	Textual	22.4	21.7	30.0	30.7	30.4
EchoSight [Yan and Xie, 2024]	LLaMA-3.1-8B	EVA-CLIP-8B	Visual	26.4	24.9	18.0	19.8	18.8
mR^{2}AG [Zhang *et al.*, 2024]	LLaMA-3.1-8B	EVA-CLIP-8B	Textual	-	-	40.6	39.8	40.2
ReflectiVA [Cocchi *et al.*, 2025]	LLaMA-3.1-8B	EVA-CLIP-8B	Textual	28.0	29.2	40.4	39.8	40.1
ReflectiVA [Cocchi *et al.*, 2025]	LLaMA-3.1-8B	EVA-CLIP-8B	Visual	35.5	35.5	28.6	28.1	28.3
MMKB-RAG [Ling *et al.*, 2025]	Qwen-2-7B	EVA-CLIP-8B	Visual	39.7	35.9	36.4	36.3	36.4
DMRAG (**Ours**)	LLaMA-3.1-8B	EVA-CLIP-8B	Visual+Textual	**48.4**	**42.2**	**41.9**	**41.7**	**41.4**

Table 2: VQA accuracy scores on E-VQA test set and InfoSeek validation set where all results from retrieval-augmented models are reported without considering any re-ranking stage to reorder retrieved web pages. **Bold** indicates state-of-the-art, underline denotes second-best. results that are not directly comparable due to different knowledge bases. All retrieval-augmented results exclude re-ranking.

For the retrieval stage, we adopt Recall@K as the primary evaluation metric. A retrieval is considered successful in this setting only when the URL of the retrieved Wikipedia article exactly matches the target article.

For the question answering stage, to ensure fair comparison, we follow prior work by using VQA Accuracy [Goyal *et al.*, 2017; Marino *et al.*, 2019] and Relaxed Accuracy [Methani *et al.*, 2020; Masry *et al.*, 2022] for InfoSeek, and the BEM score [Zhang *et al.*, 2019] for E-VQA.

4.3 Implementation Details

Initial Retrieval

In our experiments, we primarily employ EVA-CLIP-8B [Sun *et al.*, 2023] for multimodal retrieval. For the ROI-based path, we adopt an image-to-image retrieval strategy, comparing the query image with images embedded in Wikipedia documents to retrieve the corresponding pages. For the HyDE path, we perform text-to-text retrieval: generates hypothetical paragraphs using Qwen2.5-VL-7B-Instruct [Bai *et al.*, 2025], which are then embedded using EVA-CLIP-8B. Both ROI and HyDE features are encoded in the same embedding space as the knowledge base's textual and visual vectors to ensure alignment. The ROI path additionally leverages SpaCy for text preprocessing and Grounding DINO [Liu *et al.*, 2024b] for region proposals and ROI extraction.

Gating Network

The gating network is trained on a randomly sampled 50K subset from the InfoSeek and E-VQA training sets. For each training sample, the initial retrieval results from both ROI and HyDE paths are processed to compute oracle fusion weights based on the ranks of the ground-truth knowledge entries. These oracle weights serve as supervision for the gating network. The network itself is a simple and lightweight MLP with two hidden layers of size $[512, 256]$, ReLU activations,

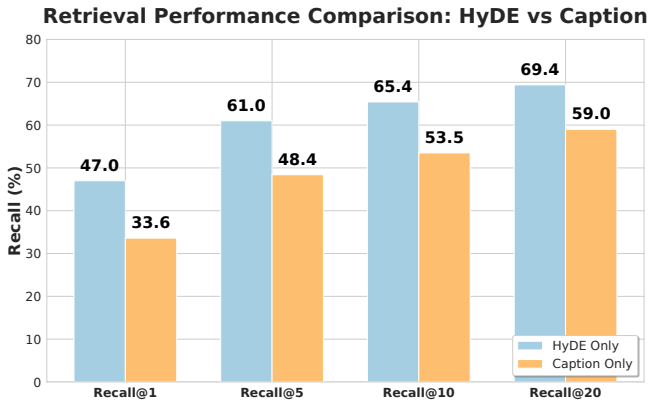

Figure 3: Retrieval performance comparison between HyDE and Image Captioning.

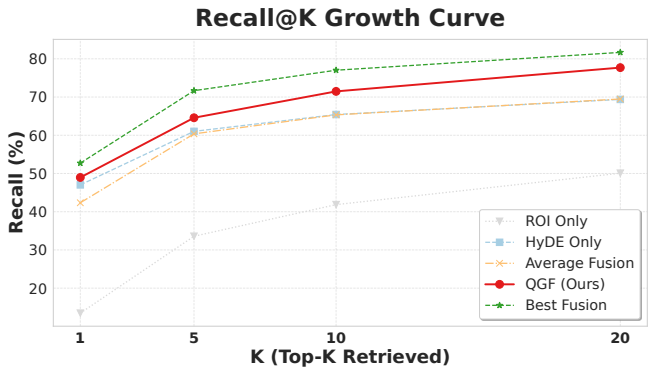

Figure 4: Recall@K growth curves of different retrieval strategies.

and Dropout rate 0.1. It takes 1280-dimensional question embeddings extracted by EVA-CLIP-8B as input and outputs a 2-dimensional softmax vector representing the fusion weights for ROI and HyDE. The network is optimized using Adam with learning rate 10^{-3}, weight decay 10^{-5}, batch size 64, for a total of 50 epochs. The loss function is the KL divergence between the predicted and oracle fusion weights. A ReduceLROnPlateau scheduler with factor 0.5 and patience 5 is applied, and early stopping is used if the validation loss does not decrease for 10 consecutive epochs.

Reranker

Following the EchoSight paradigm [Yan and Xie, 2024], we adopt a Q-Former-based reranker to refine the top-K candidates obtained from the QGF module. The encoder is initialized from pre-trained Q-Former weights provided by the LAVIS library [Li *et al.*, 2023a], and the first 32 embeddings are selected as the multimodal feature matrix corresponding to query token positions. Since InfoSeek training samples generally do not contain labeled evidence, the Q-Former is trained on the E-VQA training subset and then evaluated on both E-VQA and InfoSeek test sets. All data processing and hyperparameter settings are kept consistent with EchoSight to allow fair comparison.

4.4 Experimental Results

Retrieval Performance

Table 1 compares the recall performance of different methods on E-VQA and InfoSeek datasets. Overall, our method achieves consistent and substantial gains on both datasets, indicating that our multi-path retrieval, adaptive fusion, and cross-modal reranking align questions with knowledge entries more effectively. On E-VQA, even without reranking, our method reaches R@1=49.0% with only the two retrieval pathways and adaptive weighted fusion, surpassing all baselines and even exceeding the strong OMGM baseline with reranking (42.8%). This suggests that our method is already strong at the first-stage retrieval, pushing the target entry to higher ranks and thus providing higher quality evidence candidates for downstream answer generation. After applying the reranker, R@1 further increases to 55.0%. Similar trends are observed on InfoSeek, where our method achieves R@1 of 65.5% with reranking, surpassing previous state-of-the-art methods such as OMGM (64.0%) and ReflectiVA [Cocchi *et al.*, 2025] (56.1%). Overall, the combination of ROI+HyDE with adaptive weighting and reranking enables our method to improve both retrieval coverage and precision.

VQA Accuracy

Table 2 reports the VQA accuracy of various zero-shot and retrieval-augmented models, considering only results without reranking. DMRAG consistently outperforms all baselines across both datasets, showing that improvements stem from stronger first-stage retrieval and better utilization of retrieved evidence rather than reranking. On E-VQA, our method achieves 48.4% in Single-Hop and 42.2% in All-question settings, higher than previous retrieval-augmented models [Caffagni *et al.*, 2024; Yan and Xie, 2024; Ling *et al.*, 2025; Cocchi *et al.*, 2025], reflecting the benefits of combining ROI-based visual retrieval and HyDE-based semantic retrieval with the query-adaptive gating module. This enables DMRAG to capture both fine-grained visual cues and abstract textual knowledge, resulting in more accurate answers. On InfoSeek, our method achieves the highest scores across all categories, including 41.9% on unseen questions and 41.7% on unseen entities, highlighting strong generalization to unseen content and handling of long-tail entities. Compared with baselines, DMRAG narrows the gap between zero-shot and retrieval-augmented approaches, indicating that multi-path retrieval with adaptive fusion prioritizes relevant evidence for each query, mitigates noise from irrelevant candidates, and improves VQA accuracy. These results demonstrate that our framework strengthens first-stage retrieval and enhances answer generation even without reranking, establishing a foundation for further gains when incorporating post-retrieval refinements, and underscoring the effectiveness of dynamic multi-path retrieval and query-dependent evidence fusion in knowledge-based visual question answering.

4.5 Ablation Study

All ablation experiments are conducted on the E-VQA dataset to analyze the contribution of individual components in our retrieval framework. We first present a component-wise ablation by progressively introducing each module into the sys-

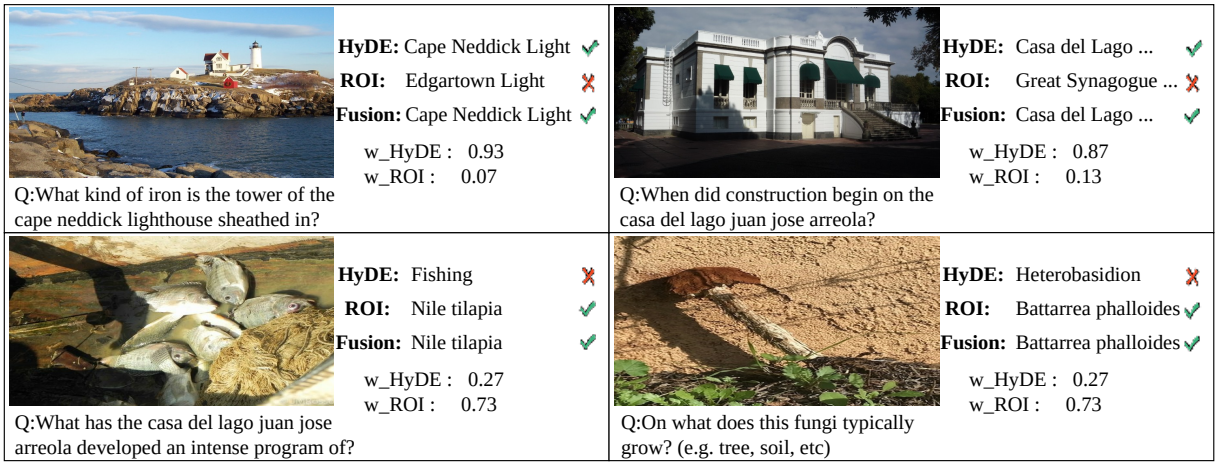

Figure 5: Qualitative examples of DMRAG. Top-1 retrieved documents (Recall@1) from HyDE, ROI, and gated fusion are shown with fusion weights. ✓indicates a Recall@1 hit (counted as a correct answer), and ✗indicates a miss.

Setting	ROI	HyDE	QGF	Rerank	R@1	R@5	R@10	R@20
DMRAG w/o HyDE	✓				13.46	33.58	41.85	50.08
DMRAG w/o ROI		✓			47.02	61.02	65.43	69.38
DMRAG w/o QGF	✓	✓	✓		48.96	64.58	71.48	77.70
DMRAG	✓	✓	✓	✓	**54.99**	**70.25**	**74.80**	**77.70**

Table 3: Component-wise ablation study of the proposed retrieval framework.

tem, with results summarized in Table 3. Based on this main ablation, we further investigate two key design choices: the effectiveness of HyDE for semantic retrieval and the necessity of QGF module.

Ablation of retrieval framework components

Table 3 reports the retrieval performance as different components are progressively introduced. Among single-path baselines, ROI-only performs the worst (R@20=50.08), indicating that pure visual matching is insufficient for encyclopedic KB-VQA. In contrast, HyDE-only achieves substantially higher recall at high K (R@20=69.38), further highlighting the importance of semantic retrieval. Building directly upon this, introducing the QGF module further improves high-K recall (R@20=77.70), demonstrating that query-dependent weighting effectively integrates complementary retrieval paths and increases candidate coverage. Finally, adding the reranking module leads to additional gains, particularly at lower K values, confirming that refining candidates after fusion improves overall ranking quality.

Impact of HyDE Compared with Image Caption

We justify the use of HyDE for the text retrieval branch by comparing it with a naive image-caption querying baseline, as shown in Figure 3. HyDE delivers substantially better retrieval performance across Recall@K, especially at higher K values. This is because image captions mainly describe visible content in the image, whereas HyDE generates a hypothetical encyclopedic paragraph that anticipates what answer-relevant evidence should look like in Wikipedia, resulting in stronger semantic alignment with knowledge-base text.

Impact of adaptive fusion

Finally, we analyze the necessity of adaptive gated fusion using the Recall@K growth curves in Figure 4 together with the results in Table 3. Although fixed 50%–50% fusion improves overall recall compared to ROI-only, it remains limited in high-K coverage, with R@20 reaching only 69.49, which is comparable to HyDE-only (69.38). This indicates that naive averaging fails to fully exploit the complementary strengths of different paths and may introduce noise from weaker signals. In contrast, adaptive gated fusion achieves a substantial improvement at high K (R@20=77.70), and its Recall@K curve gradually approaches the Best Fusion upper bound, demonstrating that the gating mechanism effectively approximates per-query optimal path selection.

4.6 Qualitative Results

We provide qualitative examples to illustrate the effect of multi-path retrieval and adaptive fusion in DMRAG. Figure 5 shows the top-1 retrieved document (Recall@1) from HyDE, ROI, and the QGF module, together with the fusion weights. In our evaluation, a Recall@1 hit indicates a correct answer: ✓ denotes a successful retrieval and ✗ denotes a retrieval failure. These examples suggest that HyDE and ROI are complementary, and the gating module adaptively selects the more relevant source for each question.

5 Conclusion

We propose DMRAG, a dynamic multi-path retrieval framework for KB-VQA. By effectively combining ROI-based visual retrieval and HyDE-based semantic retrieval with the QGF module, the framework enables flexible and query-dependent use of multimodal knowledge. Despite its effectiveness, DMRAG remains dependent on the coverage and quality of external knowledge sources; future work will explore more robust retrieval strategies and improved integration between retrieval and reasoning.

Acknowledgments

This work is in part funded by the National Natural Science Foundation of China (Grant No. 62372364) and the Technical Innovation Guidance Plan of Shaanxi Province, China (Grant No. 2024QCY-KXJ-199).

Contribution Statement

The first two authors, Zeyu Song and Yimin Deng, contributed equally to this work.

References

[Achiam *et al.*, 2023] 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.

[Antol *et al.*, 2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[Bai *et al.*, 2025] 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.

[Caffagni *et al.*, 2024] 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.

[Chen *et al.*, 2023] 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? *arXiv preprint arXiv:2302.11713*, 2023.

[Cocchi *et al.*, 2025] Federico Cocchi, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Augmenting multimodal llms with self-reflective tokens for knowledge-based visual question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9199–9209, 2025.

[Dai *et al.*, 2023] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

[Deng *et al.*, 2025] Yimin Deng, Yuxia Wu, Yejing Wang, Guoshuai Zhao, Li Zhu, Qidong Liu, Derong Xu, Zichuan Fu, Xian Wu, Yefeng Zheng, et al. A multi-expert structural-semantic hybrid framework for unveiling historical patterns in temporal knowledge graphs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20553–20565, 2025.

[Deng *et al.*, 2026a] Yimin Deng, Zhenxi Lin, Yejing Wang, Guoshuai Zhao, Pengyue Jia, Zichuan Fu, Derong Xu, Yefeng Zheng, Xiangyu Zhao, Li Zhu, et al. Multidx: A multi-source knowledge integration framework towards diagnostic reasoning. *arXiv preprint arXiv:2604.24186*, 2026.

[Deng *et al.*, 2026b] Yimin Deng, Yejing Wang, Zhenxi Lin, Zichuan Fu, Guoshuai Zhao, Derong Xu, Yefeng Zheng, Xiangyu Zhao, Xian Wu, Li Zhu, et al. Adaptime: Enabling adaptive temporal reasoning in large language models. *arXiv preprint arXiv:2604.24175*, 2026.

[Fan *et al.*, 2024] Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6491–6501, 2024.

[Gao *et al.*, 2023] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, 2023.

[Goyal *et al.*, 2017] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi 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*, pages 6904–6913, 2017.

[Hong *et al.*, 2025] Yuyang Hong, Jiaqi Gu, Qi Yang, Lubin Fan, Yue Wu, Ying Wang, Kun Ding, Shiming Xiang, and Jieping Ye. Knowledge-based visual question answer with multimodal processing, retrieval and filtering. *arXiv preprint arXiv:2510.14605*, 2025.

[Hu *et al.*, 2023] Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12065–12075, 2023.

[Jian *et al.*, 2024] Pu Jian, Donglei Yu, and Jiajun Zhang. Large language models know what is key visual entity: An llm-assisted multimodal retrieval for vqa. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10939–10956, 2024.

[Li *et al.*, 2023a] Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven CH Hoi. Lavis: A one-stop library for language-vision intelligence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 31–41, 2023.

[Li *et al.*, 2023b] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[Ling *et al.*, 2025] Zihan Ling, Zhiyao Guo, Yixuan Huang, Yi An, Shuai Xiao, Jinsong Lan, Xiaoyong Zhu, and Bo Zheng. Mmkb-rag: A multi-modal knowledge-based retrieval-augmented generation framework. *arXiv preprint arXiv:2504.10074*, 2025.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[Liu *et al.*, 2024a] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024.

[Liu *et al.*, 2024b] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024.

[Marino *et al.*, 2019] 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.

[Masry *et al.*, 2022] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279, 2022.

[Mensink *et al.*, 2023] 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.

[Methani *et al.*, 2020] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*, pages 1527–1536, 2020.

[Mi *et al.*, 2025] Yunqi Mi, Boyang Yan, Guoshuai Zhao, Jialie Shen, and Xueming Qian. A plug-and-play model-agnostic embedding enhancement approach for explainable recommendation. *IEEE Transactions on Multimedia*, 2025.

[Radford *et al.*, 2021] 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.

[Schwenk *et al.*, 2022] 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.

[Shang *et al.*, 2023] Heng Shang, Guoshuai Zhao, Jing Shi, and Xueming Qian. A multiview text imagination network based on latent alignment for image-text matching. *IEEE Intelligent Systems*, 38(3):41–50, 2023.

[Sun *et al.*, 2023] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[Wang *et al.*, 2024] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[Wu *et al.*, 2023] Sen Wu, Guoshuai Zhao, and Xueming Qian. Resolving zero-shot and fact-based visual question answering via enhanced fact retrieval. *IEEE Transactions on Multimedia*, 26:1790–1800, 2023.

[Wu *et al.*, 2024] Feng Wu, Guoshuai Zhao, Tengjiao Li, Jialie Shen, and Xueming Qian. Improving conversational recommendation system through personalized preference modeling and knowledge graph. *IEEE Transactions on Knowledge and Data Engineering*, 36(12):8529–8540, 2024.

[Yan and Xie, 2024] Yibin Yan and Weidi Xie. Echosight: Advancing visual-language models with wiki knowledge. *arXiv preprint arXiv:2407.12735*, 2024.

[Yang *et al.*, 2025] Wei Yang, Jingjing Fu, Rui Wang, Jinyu Wang, Lei Song, and Jiang Bian. Omgm: Orchestrate multiple granularities and modalities for efficient multimodal retrieval. *arXiv preprint arXiv:2505.07879*, 2025.

[Yuan *et al.*, 2025] Xu Yuan, Liangbo Ning, Wenqi Fan, and Qing Li. mkg-rag: Multimodal knowledge graph-enhanced rag for visual question answering. *arXiv preprint arXiv:2508.05318*, 2025.

[Zhang *et al.*, 2019] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[Zhang *et al.*, 2024] 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.

[Zhao *et al.*, 2023] Guoshuai Zhao, Chaofeng Zhang, Heng Shang, Yaxiong Wang, Li Zhu, and Xueming Qian. Generative label fused network for image–text matching. *Knowledge-Based Systems*, 263:110280, 2023.

[Zhu *et al.*, 2015] Lei Zhu, Jialie Shen, Hai Jin, Ran Zheng, and Liang Xie. Content-based visual landmark search via multimodal hypergraph learning. *IEEE Transactions on Cybernetics*, 45(12):2756–2769, 2015.