MMAgent-R^2: Learning to Rerank and Reject for Agentic mRAG

Tao Zhang[*1,2,3,4], Ziqi Zhang[*1,3], Zongyang Ma[1,3], Yuxin Yang[1,2,3], Bing Li[†1,3,5], Chunfeng Yuan[1,3], Kang Rong[4], Fengyun Rao[4], Jing LYU[4], and Weiming Hu[1,2,3,6]

[1] State Key Laboratory of Multimodal Artificial Intelligence Systems, CASIA
{zhangtao2023,mazongyang2020,yangyuxin2023}@ia.ac.cn
{ziqi.zhang,bli,cfyuan,wmhu}@nlpr.ia.ac.cn
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] Beijing Key Laboratory of Super Intelligent Security of Multi-Modal Information
[4] WeChat Vision, Tencent Inc.
{rickrong,fengyunrao,eckolv}@tencent.com
[5] PeopleAI Inc.
[6] School of Information Science and Technology, ShanghaiTech University
[*]Equal contribution [†]Corresponding author

Abstract. Knowledge-based Visual Question Answering (KB-VQA) requires models to retrieve visual entities matching the query image from large-scale encyclopedic knowledge bases and answer related questions. Existing multimodal Retrieval Augmented Generation (mRAG) methods rely on global visual features to match candidate entities, yet when the knowledge base contains numerous visually similar entities, the retriever struggles to distinguish them, populating the candidate set with visually similar but factually mismatched distractors. Since subsequent processing steps such as noise filtering are also confined to this fixed candidate set, errors from failed retrieval inevitably propagate to the final answer. To address these challenges, we propose MMAgent-R^2, an agentic mRAG framework that integrates visual reranking and active rejection as its internal verification mechanism. Visual reranking directly compares query and candidate images, capturing discriminative details beyond textual descriptions to precisely identify the target entity among similar candidates; active rejection discards unreliable results and retrieves additional candidates when no confident match is found, moving beyond the fixed candidate pool. We design a composite reward function with step-level verification rewards and achieve joint optimization of external retrieval, internal verification, and answer generation via GRPO training. Experiments on InfoSeek, E-VQA, and MMhops demonstrate that MMAgent-R^2 achieves state-of-the-art performance, with particularly notable advantages in challenging retrieval scenarios and complex multi-image multi-hop reasoning tasks.

Keywords: Knowledge-based VQA · Retrieval-Augmented Generation · Visual Agent · Reinforcement Learning

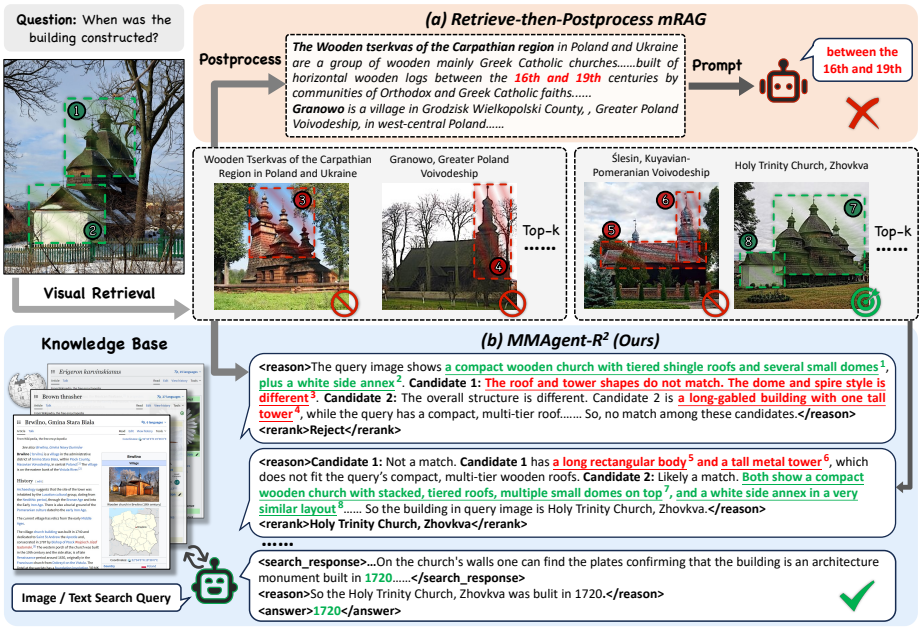

Fig. 1: Comparison between the Retrieve-then-Postprocess mRAG and MMAgent-R^2. **(a)** With a fixed candidate set retrieved via global visual features, the model can be misled by visually similar yet mismatched entries. **(b)** MMAgent-R^2 compares the query image against candidates (①–⑧) to rerank them, and rejects unreliable results to retrieve additional candidates when no confident match is found, enabling correct entity identification and answering.

1 Introduction

Knowledge-based Visual Question Answering (KB-VQA) [6,20,35] requires models to recognize visual entities in images and answer questions about their attributes. As illustrated in Fig. 1, answering "When was this building constructed?" hinges on precisely identifying the building and subsequently associating it with external knowledge (its construction year). Recently, multimodal Retrieval Augmented Generation (mRAG) [5,8,10,12,18,21,31,34,36] methods have emerged as a promising approach for such tasks. These methods generally adopt a "Retrieve-then-Postprocess" paradigm, where the retrieval stage attempts to locate candidate entries visually similar to the query entity from a large-scale visual-textual encyclopedia knowledge base, and the post-processing stage then filters from a small, fixed set of candidates to identify content that accurately matches the query entity, thereby reasoning to generate the answer.

Despite their potential, existing mRAG methods encounter a critical identification bottleneck when distinguishing between visually similar entities. In the retrieval stage, existing methods primarily rely on global features from the full query image or cropped regions [12,22] for similarity matching, which struggle to

capture the key visual differences between similar entities (*e.g.*, the tiered roof and dome configuration of the wooden churches in Fig. 1), populating candidates with visually similar yet factually mismatched entries. Although the post-processing stage reranks [8, 31, 32] textual paragraphs associated with candidate entities, textual descriptions inherently cannot represent subtle visual distinctions, rendering them ineffective at selecting the correct entity. Moreover, the reasoning scope is confined to a fixed candidate pool from the initial retrieval. If the correct entity is not recalled, the model is forced to reason over incorrect entities, inevitably propagating errors to answer generation. Meanwhile, such retrieve-then-postprocess pipelines typically depend on additional modules or predefined processing steps, hindering joint optimization and struggling with complex scenarios such as multi-image inputs and multi-hop reasoning.

To address these challenges, we propose MMAgent-R^2, an agentic mRAG framework integrating **visual reranking** and **active rejection**. These two capabilities constitute the model's internal verification mechanism, performing in-depth visual verification on each retrieved candidate batch. Visual reranking directly compares query and candidate images, capturing critical visual differences that textual descriptions fail to convey, thereby precisely identifying the target entity among similar candidates. When no reliable match exists, active rejection allows the model to discard the current results and retrieve additional candidates, moving beyond the fixed candidate pool and avoiding reasoning over incorrect entities. We unify reranking, rejection, and the invocation of image and text retrieval tools into a single reasoning model via reinforcement learning, eliminating dependence on additional modules. This agentic architecture naturally accommodates complex scenarios such as multi-image inputs and multi-hop reasoning, acquiring accurate knowledge through iterative retrieval and verification.

We train MMAgent-R^2 via Group Relative Policy Optimization (GRPO) [24] with a composite reward function: an outcome reward evaluates the correctness of the final answer, a format reward ensures compliance of structured outputs, and fine-grained verification rewards provide step-level supervision for each reranking and rejection decision, rewarding the model when it correctly identifies a matching candidate or correctly rejects when no match exists. This dense intermediate supervision enables the model to learn when to select and when to reject. Extensive experiments on InfoSeek [6], E-VQA [20], and MMhops [35] demonstrate that MMAgent-R^2 achieves state-of-the-art performance across all benchmarks, with particularly significant gains on E-VQA (+7.2), which has the largest knowledge base and the highest retrieval difficulty, and on MMhops (Bridging +13.2, Comparison +10.4), which involves multi-image inputs and multi-hop reasoning, fully validating the effectiveness of our approach in challenging retrieval scenarios and complex reasoning tasks.

Our main contributions are summarized as follows:

- We propose MMAgent-R^2, an agentic mRAG framework in which **visual reranking** and **active rejection** jointly constitute the internal verification mechanism for external retrieval. The model autonomously searches, verifies, and answers within a multi-turn agent loop, without any additional modules.

- We design a composite reward function with step-level verification rewards that provide fine-grained supervision for each reranking and rejection decision, achieving joint optimization of external retrieval, internal verification, and answer generation via GRPO training.
- Extensive experiments show that MMAgent-R^2 achieves state-of-the-art performance on InfoSeek, E-VQA, and MMhops, with particularly significant improvements in challenging retrieval scenarios and complex reasoning tasks.

2 Related Work

2.1 Knowledge-based VQA

Knowledge-based Visual Question Answering (KB-VQA) requires models to recognize visual entities in images and leverage external knowledge bases to answer related questions. Early benchmarks such as OK-VQA [19] and A-OKVQA [23] focused on commonsense reasoning. Subsequent research shifted towards entity-centric settings with increasingly larger knowledge bases and finer entity granularity [6, 14]. Encyclopedic-VQA [20] further introduced two-hop reasoning that requires chaining multiple retrieval steps, and MMhops [35] extended this to multi-image inputs with multi-hop reasoning chains.

Multimodal Retrieval-Augmented Generation (mRAG) has emerged as the dominant framework for these tasks. Early approaches converted visual inputs into textual representations for text-level retrieval [17]. Subsequent works shifted to image-level retrieval via visual encoders, including cross-modal retrieval that fuses image-to-image and image-to-text similarity for entity matching [13], hierarchical retrieval that first recalls candidates through image similarity and then locates textual evidence [5], cropped-region retrieval that reduces background interference to improve recall accuracy [12, 22], and structured retrieval via multimodal knowledge graphs that organize unstructured knowledge into graph representations for more precise retrieval [34]. Despite these advances, visual retrievers primarily rely on global feature matching, remaining prone to recalling visually similar yet factually mismatched candidates.

2.2 Post-processing in mRAG

After initial retrieval, how to effectively post-process candidates to reduce noise remains a key challenge for mRAG. Existing methods primarily operate at the textual evidence level. Representative approaches include leveraging VLMs to compute multimodal query-passage similarity for paragraph-level reranking [31], employing reflection mechanisms that allow the model to autonomously assess paragraph relevance and decide whether to trigger supplementary retrieval [36], question-guided filtering that injects question semantics into candidate encoding for cross-article chunk-level selection [32], and multi-stage filtering pipelines that combine multimodal LLMs for hierarchical cross-filtering [12] or train independent critic models to eliminate irrelevant texts [8]. Some works also leverage the

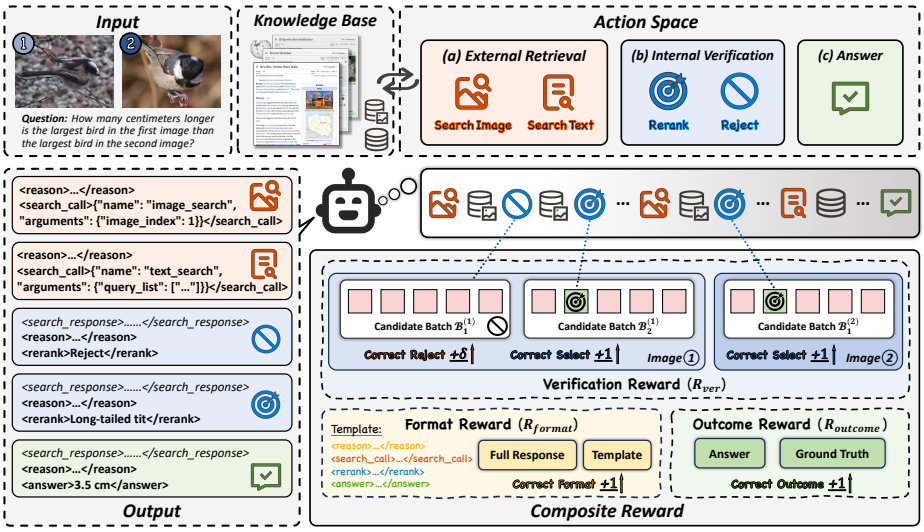

Fig. 2: The overall framework of MMAgent-R^2. Given multi-image inputs and a question, the VLM agent generates an interleaved "Reason-then-Act" trajectory by selecting and executing an operation at each step from an action space comprising External Retrieval, Internal Verification, and Answer Generation. The framework is optimized via RL with a composite reward function, where the verification reward provides dense step-level supervision for each reranking and rejection decision.

model's own knowledge boundaries to dynamically generate semantic tags, filtering retrieved content based on its consistency with the model's existing knowledge [18]. Although these methods have significantly advanced textual denoising, their filtering signals remain confined to the textual semantic space, struggling to distinguish fine-grained visual differences between highly similar entities. In contrast, MMAgent-R^2 introduces visual reranking and active rejection for candidate images within a unified agentic framework, elevating post-processing from the textual to the visual level.

3 Methodology

3.1 Problem Formulation

For the Knowledge-based Visual Question Answering (KB-VQA) task, given N query images $\{I_q^1, \ldots, I_q^N\}$ $(N \geq 1)$ and a question Q, the model has access to an external knowledge base $\mathcal{KB}$ to extract relevant knowledge and generate an accurate answer y. Specifically, the knowledge base $\mathcal{KB}$ consists of a massive collection of structured web pages (*e.g.*, Wikipedia). We formalize each entry as a triplet $(E, \mathcal{I}_e, \mathcal{T}_e)$, where E represents an entity, $\mathcal{I}_e$ is the set of reference images associated with the entity, and $\mathcal{T}_e$ is the set of detailed textual paragraphs describing the entity. In practice, the images and texts across all entries in the

knowledge base constitute the visual knowledge base $\mathcal{KB}_v = \bigcup_e \mathcal{I}_e$ and the textual knowledge base $\mathcal{KB}_t = \bigcup_e \mathcal{T}_e$ used for retrieval.

3.2 MMAgent-R^2

As illustrated in Fig. 2, we formulate the reasoning process as a multi-turn agentic decision-making process. We define the VLM as an agent driven by a policy π_θ and construct an action space $\mathcal{A}$ comprising "external retrieval", "internal verification", and "answer generation". Specifically, the agent can invoke retrieval tools to recall candidate images from $\mathcal{KB}_v$ or extract relevant text from $\mathcal{KB}_t$. Concurrently, based on its perception and reasoning capabilities, it can perform precise "visual reranking" or "active rejection" on the recalled candidate images. Conditioned on the interaction history $\mathcal{H}_t$, the policy π_θ autonomously generates a sequence of thoughts and actions $a_t \in \mathcal{A}$ within a maximum of $T_{\max}$ turns. Through iterative exploration and verification, the agent ultimately executes the answer generation action to output the answer y.

Table 1: Summary of the Agent Action Space $\mathcal{A}$. The policy π_θ can execute External Retrieval actions to gather knowledge, perform Internal Verification actions to validate retrieved candidates, or trigger Answer Generation to conclude the task.

Category	Action	Model Output & Environment Feedback
External Retrieval	`SearchImage`(j)	The model specifies query image index j. The environment returns a batch of Top-K visually similar candidates from $\mathcal{KB}_v$; consecutive calls yield non-overlapping batches.
	`SearchText`($\{q_i\}_{i=1}^m$)	The model generates a list of m text queries. For each query q_i, the environment retrieves and concatenates the Top-K most relevant paragraphs from $\mathcal{KB}_t$.
Internal Verification	`Rerank`(e_k)	The model outputs entity name e_k from the current batch via visual comparison. The environment records this entity for subsequent retrieval and reasoning.
	`Reject`()	The model outputs the Reject signal, determining no match in the current batch. The environment triggers `SearchImage` to provide the next batch (up to $\text{Rej}_{\max}$ times).
Answer Generation	`Answer`(y)	The model synthesizes gathered evidence and generates the final answer y. The environment terminates the reasoning loop.

3.2.1 Action Space As detailed in Table 1, we structure the agent's action space $\mathcal{A}$ into three categories: external retrieval for gathering multimodal knowledge, internal verification driven by the model's intrinsic reasoning, and answer generation for concluding the task.

External Retrieval. These actions enable the agent to query the external knowledge base to gather the multimodal evidence required for problem-solving.

– **Image Retrieval (`SearchImage`(j)):** This action encodes the query image I_q^j and searches a pre-built feature index for similar candidate entries. Given the query index j, the environment returns a batch $\mathcal{B} = \{(I_k, e_k)\}_{k=1}^K$ comprising K candidate entries, each containing an image and its entity name. Crucially, this action supports dynamic candidate expansion: consecutive calls

for the same query image yield non-overlapping subsequent batches, enabling the model to break through the limitations of the initial Top-K results and continuously explore the candidate space.
- **Text Retrieval (`SearchText`($\{q_i\}_{i=1}^m$)):** The model leverages this action to extract relevant descriptive evidence from the textual knowledge base $\mathcal{KB}_t$. To enhance information acquisition efficiency, this action supports the parallel submission of multiple queries $\{q_1, \ldots, q_m\}$. The environment retrieves paragraphs relevant to each query and returns the concatenated results. This mechanism allows the model to autonomously aggregate multi-dimensional textual facts within a single interaction turn, providing comprehensive evidentiary support for the final answer generation.

Internal Verification. MMAgent-R^2 leverages the fine-grained perception and reasoning capabilities of the VLM to conduct deep internal verification on the externally retrieved candidate batches, thereby precisely identifying the target entity among highly similar candidate images. Upon receiving the current candidate batch $\mathcal{B}_t^{(j)}$, the policy π_θ performs a direct visual comparison between the query image and all candidate images. Conditioned on the cumulative interaction history $\mathcal{H}_t$, the model generates a decision action:

$$a_t = \pi_\theta\Big(I_q^j,\ \mathcal{B}_t^{(j)},\ \mathcal{H}_t\Big) \in \{\texttt{Rerank}(e_1), \ldots, \texttt{Rerank}(e_K),\ \texttt{Reject}()\} \tag{1}$$

- **Visual Reranking (`Rerank`(e_k)):** When the model verifies through visual feature comparison that entity e_k in the candidate batch accurately matches the query image, it executes this reranking action and outputs e_k. This entity name is then appended to the context history, serving as a definitive premise for subsequent reasoning.
- **Active Rejection (`Reject`()):** When the model determines that no candidate image in current batch matches the query image, it executes this action. This action triggers dynamic candidate expansion, and the environment provides the next batch of candidate images. To balance search depth and reasoning efficiency, we set a maximum rejection count $\mathrm{Rej}_{\max}$ for each query image.

By integrating "reranking" and "rejection" as the agent's internal verification actions, MMAgent-R^2 achieves strict verification of retrieval results: "reranking" precisely identifies the target among similar retrieval results, while "rejection" dynamically expands the search boundary when initial retrievals fail. Working in synergy, they enable the model to autonomously explore complex retrieval spaces, ensuring that final reasoning and answer generation remain grounded in definitive visual evidence.

Answer Generation. Based on the accumulated interaction history $\mathcal{H}_t$, the agent executes this terminal action to synthesize the gathered evidence and generate the final answer y, thereby concluding the reasoning process.

3.2.2 Agentic Reasoning Loop MMAgent-R^2 operates through a multi-turn "Reasoning-Action-Feedback" paradigm. Each iteration of the loop involves the following logic:

- **Reasoning and Action:** The agent performs autonomous reasoning based on the current interaction history $\mathcal{H}_{t-1}$ and selects the next action. Instead of following a fixed sequence of steps, it flexibly decides, based on the reasoning progress, whether to initiate external searches to acquire new clues or to perform internal verification (`Rerank`/`Reject`) to examine the reliability of external information.
- **Environment Feedback:** Upon executing the selected action a_t, the environment returns a corresponding observation o_t (such as image batches or textual facts). This result, along with the action command, is synchronously appended to the conversation history, forming a continuously updated multimodal evidence chain to support subsequent decisions.
- **Termination and Boundaries:** The loop terminates when the agent provides the final answer `Answer` or reaches the maximum turn limit $T_{\max}$. Furthermore, the system monitors the exploration depth for each individual query image; once the maximum rejection limit $\text{Rej}_{\max}$ is reached, the environment provides a boundary prompt to guide the agent in achieving a balance between search scope and reasoning efficiency.

This framework overcomes the limitations of static pipelines and supports highly autonomous reasoning paths. The agent can flexibly schedule retrieval, reranking, and rejection actions based on the context, or even answer directly using its parametric knowledge. Through reinforcement learning, the model can learn to autonomously optimize interaction strategies within complex reasoning paths in an end-to-end manner.

3.3 Agentic RL Training

We train MMAgent-R^2 using Group Relative Policy Optimization (GRPO). For each training sample consisting of query images $\{I_q^j\}_{j=1}^N$, question Q, ground-truth answer y^*, and the corresponding ground-truth entities $\{e^{*,j}\}_{j=1}^N$, we sample G complete multi-turn rollouts from the current policy π_θ and employ a carefully designed reward function to guide the model toward learning the optimal decision-making strategy.

3.3.1 Composite Reward Function To supervise the agent's complex behaviors across multiple dimensions, we define a composite reward function:

$$R = R_{\text{outcome}} + R_{\text{format}} + R_{\text{ver}} \tag{2}$$

- **Outcome Reward (R_{outcome}):** Evaluates the correctness of the final output y. It yields 1 if the answer matches the ground truth y^*, and 0 otherwise.
- **Format Reward (R_{format}):** Verifies whether the output of each turn strictly adheres to the predefined action syntax. This reward ensures stable structured interaction between the agent and the environment.

– **Verification Reward (R_{ver}):** Serving as the signal to guide the agent in learning internal verification behaviors, we provide step-level supervision for every verification decision (`Rerank` or `Reject`) across all query images:

$$R_{\text{ver}} = \sum_{j=1}^{N} \sum_{t=1}^{T_j} \rho\Big(a_t^{(j)},\ \mathcal{B}_t^{(j)},\ e^{*,j}\Big) \tag{3}$$

where $e^{*,j}$ is the ground-truth entity for the j-th image, T_j denotes the number of verification steps, and $a_t^{(j)}$ is the action taken for the candidate batch $\mathcal{B}_t^{(j)}$. The per-step reward ρ is defined as:

$$\rho(a, \mathcal{B}, e^*) = \begin{cases} \delta & \text{if } a = \texttt{Reject} \text{ and } e^* \notin \mathcal{B} \\ 1 & \text{if } a = \texttt{Rerank}(e^*) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Here, δ represents the reward weight for correct rejections. This dense supervision mechanism not only complements the sparse final answer signal but also explicitly encourages the model to choose "rejection" over "guessing" when faced with uncertain candidate batches.

3.3.2 Environment Response Masking During multi-turn rollout training, we mask all tokens corresponding to environment responses (*e.g.*, retrieved images, textual paragraphs). This operation ensures that policy gradient updates are applied exclusively to the "Reasoning" and "Action" tokens generated by the model itself, ensuring the model remains strictly accountable for its own autonomous decision-making.

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets and Evaluation Metrics We train and evaluate our model on three knowledge-based Visual Question Answering (VQA) benchmarks.
InfoSeek [6] contains approximately 1.3M image-question-answer triplets, covering 11K Wikipedia entities and 2.7K entity categories, with answers classified into string, time, and numerical types. It defines two evaluation subsets: Unseen Question and Unseen Entity. For evaluation, string and time types use VQA accuracy (exact match after normalization, with a ± 1 year tolerance for time); the numerical type uses Relaxed Accuracy (a prediction is correct if it falls within a 10% tolerance range or achieves an IoU $\geq 50\%$ for range answers). The overall accuracy is the harmonic mean of these two subsets. Following previous setups [31, 36], we adopt a knowledge base comprising 100K Wikipedia pages.
Encyclopedic-VQA (E-VQA) [20] consists of 221K independent question-answer pairs, each equipped with up to 5 images, yielding approximately 1M

VQA triplets. It covers 16.7K fine-grained categories sourced from iNaturalist [27] and Google Landmarks [29], featuring single-hop and two-hop reasoning questions. Accuracy is measured using the BEM [4] (BERT Matching) score, where a prediction is correct if its score is ≥ 0.5. We utilize the provided knowledge base containing 2M Wikipedia pages.

MMhops [35] is a benchmark specifically designed for multimodal multi-hop reasoning, containing 31.1K samples divided into Bridging (single-image chained reasoning) and Comparison (multi-image comparative reasoning) tasks. All samples require 3 to 4 reasoning steps to deeply integrate visual and textual knowledge. It adopts the same evaluation protocol as InfoSeek [6]. Following the benchmark's default setting, we employ a Wikipedia knowledge base of 100K pages.

4.1.2 Implementation Details We implement MMAgent-R^2 on two base models: Qwen2.5-VL-7B-Instruct [3] and Qwen3-VL-8B-Instruct [2]. For image retrieval, we encode the images and titles of Wikipedia pages into combined features using EVA-CLIP-8B [26] and build a candidate index with Faiss-GPU [11]. Given a query image, we encode it with the same model and retrieve the top-K most similar candidates by cosine similarity, returning them as a single batch with a default size of K=5. Each query image is allowed up to $\text{Rej}_{\max}$=2 rejections by default, with each rejection triggering retrieval of the next candidate batch. For text retrieval, we build a vector index with E5 [28] over Wikipedia articles segmented into 800-character chunks, returning the top-3 most relevant passages per query. We sample 65K, 67K, and 17K training instances from the training splits of InfoSeek, E-VQA, and MMhops respectively, rather than using the full datasets. Training is conducted with the GRPO algorithm [24] on the verl framework [25], using a global batch size of 1024 for 2 epochs, with G=8 rollout trajectories per instance. The learning rate is set to 1×10^{-6} and the maximum number of interaction turns is capped at $T_{\max}$=8. The default reward weight for correct rejection is δ=0.2. Following DAPO [33], we omit the KL divergence penalty to improve training efficiency.

4.2 Main Results

4.2.1 Comparison with SOTAs We compare MMAgent-R^2 against three categories of baselines: (1) **Direct Answer**, where VLMs rely solely on parametric knowledge, to assess the gains from external retrieval; (2) **Multimodal RAG**, covering diverse retrieval-augmented pipelines, to validate the advantage of visual verification over text-level post-processing; and (3) **Agentic mRAG**, where models invoke tools through multi-turn reasoning, to demonstrate the necessity of actively verifying retrieval results rather than directly trusting them. We report results for two variants: MMAgent-R^2-7B (initialized from Qwen2.5-VL-7B [3]) and MMAgent-R^2-8B (initialized from Qwen3-VL-8B [2]). As shown in Tables 2 and 3, both variants consistently surpass all compared methods across all three benchmarks.

Notable improvements on E-VQA. MMAgent-R^2-8B achieves 55.9% Single-Hop and 54.2% overall accuracy, outperforming the runners-up QKVQA [32]

Table 2: VQA accuracy on E-VQA [20] and InfoSeek [6]. †Results obtained with a different knowledge base and thus not directly comparable.

Method	Model	Retriever	E-VQA		InfoSeek		
			Single-Hop	All	Unseen-Q	Unseen-E	All
Direct Answer							
InstructBLIP [9]	Flan-T5$_{XL}$	-	11.9	12.0	8.9	7.4	8.1
BLIP-2 [15]	Flan-T5$_{XL}$	-	12.6	12.4	12.7	12.3	12.5
GPT-4V [1]	-	-	26.9	28.1	15.0	14.3	14.6
Qwen2.5-VL-7B [3]	Qwen2.5-VL-7B	-	19.0	18.8	19.7	19.4	19.6
Qwen3-VL-8B [2]	Qwen3-VL-8B	-	20.3	20.5	20.8	18.5	19.6
Multimodal RAG							
DPR$_{v+t}$† [13]	Multi-passage BERT	CLIP ViT-B/32	29.1	-	-	-	12.4
RORA-VLM† [21]	LLaVA-v1.5-7B	CLIP ViT-L/14	-	20.3	25.1	27.3	-
CoMEM† [30]	Qwen2.5-VL-7B	Custom VLM	-	-	32.8	28.5	-
Wiki-LLaVA [5]	LLaVA-v1.5-7B	CLIP ViT-L/14	17.7	20.3	30.1	27.8	28.9
mKG-RAG [34]	LLaVA-MORE-8B	Custom VLM	38.4	36.3	41.4	39.6	40.5
EchoSight [31]	Mistral-7B/LLaMA3-8B	EVA-CLIP-8B	41.8	-	-	-	31.3
mR^{2}AG [36]	LLaVA-v1.5-7B	CLIP ViT-L/14	-	-	40.6	39.8	40.2
MMKB-RAG [18]	Qwen2-7B	EVA-CLIP-8B	39.7	35.9	36.4	36.3	36.4
VLM-PRF [12]	InternVL3-8B	EVA-CLIP-8B	40.1	39.2	43.5	42.1	42.5
QKVQA [32]	LLaMA-3.1-8B	EVA-CLIP-8B	49.5	-	45.0	43.5	44.2
QKVQA [32]	Qwen2.5-VL-7B	EVA-CLIP-8B	53.7	-	38.2	37.6	37.9
ReAG [8]	Qwen2.5-VL-7B	EVA-CLIP-8B	44.9	47.0	48.3	46.2	47.2
Agentic mRAG							
MMhops-R1 [35]	Qwen2.5-VL-7B	CLIP ViT-L/14	-	-	33.8	32.6	33.2
MMAgent-R^2	Qwen2.5-VL-7B	EVA-CLIP-8B	55.4 ↑1.7	53.6 ↑6.6	**50.7** ↑2.4	**49.7** ↑3.5	**50.2** ↑3.0
MMAgent-R^2	Qwen3-VL-8B	EVA-CLIP-8B	**55.9** ↑2.2	**54.2** ↑7.2	49.0 ↑0.7	49.2 ↑3.0	49.1 ↑1.9

(+2.2) and ReAG [8] (+7.2), respectively. The E-VQA knowledge base contains ∼2M Wikipedia pages with Recall@1 of only 15.4% (Table 4), demonstrating that visual reranking becomes critical when the initial retrieval is unreliable.

Consistent superiority on InfoSeek. MMAgent-R^2-7B achieves 50.2% overall accuracy, surpassing the previous best ReAG [8] (+3.0). Notably, ReAG relies on multi-stage training with SFT cold-start, an external critic module, and RL, whereas MMAgent-R^2 attains superior performance through RL training alone. Furthermore, MMhops-R1 [35], a concurrent agentic method without visual verification, reaches only 33.2% on InfoSeek, confirming that visual verification is indispensable for fine-grained entity identification.

Pronounced advantages on multi-hop reasoning. As shown in Table 3, MMAgent-R^2-8B achieves 67.2% on Bridging (+13.2) and 39.8% on Comparison (+10.4), substantially surpassing the closed-source Gemini-2.5-Pro [7] (54.0% / 29.4%). These results validate the strong generalization of visual reranking and active rejection to multi-image, multi-hop reasoning.

4.2.2 Improvement on Visual Entity Identification Entity identification accuracy (Ident.) denotes the proportion of samples for which the model correctly identifies the ground-truth entity. For MMhops, which involves multiple query images per sample, a sample is correct only when all query images are simultaneously matched. As shown in Table 4, MMAgent-R^2 substantially outperforms the visual retriever's Recall@1 across all benchmarks. Notably on E-VQA, where

Table 3: VQA accuracy on MMhops [35]. MMAgent-R^2-7B and MMAgent-R^2-8B are initialized from Qwen2.5-VL-7B and Qwen3-VL-8B, respectively.

Method	Bridging	Comparison
Gemini-2.5-flash [7]	46.6	23.2
Gemini-2.5-pro [7]	54.0	29.4
EchoSight [31]	13.6	4.81
OmniSearch [16]	42.7	17.0
MMhops-R1 [35]	51.4	22.0
MMAgent-R^2-7B	65.9 ↑11.9	34.9 ↑5.5
MMAgent-R^2-8B	**67.2** ↑13.2	**39.8** ↑10.4

Table 4: Entity identification accuracy (Ident.%) of MMAgent-R^2 against retriever recall. R@15 represents candidate pool coverage, *i.e.*, the upper bound achievable by reranking and rejection.

	InfoSeek E-VQA		MMhops	
			Bri.	Comp.
R@1	53.7	15.4	50.1	32.1
R@15 (ceil.)	80.2	37.8	78.4	70.4
MMAgent-R^2-7B	**63.6**	28.7	54.8	35.5
MMAgent-R^2-8B	62.0	**29.5**	**55.7**	**36.8**

the retriever's Recall@1 is only 15.4%, MMAgent-R^2-8B raises Ident. to 29.5% (+14.1), indicating that the benefit of visual verification is most pronounced in challenging retrieval scenarios. On InfoSeek, MMAgent-R^2-7B raises Ident. from 53.7% to 63.6% (+9.9). On MMhops, MMAgent-R^2-8B achieves 55.7% on Bridging (+5.6) and 36.8% on Comparison (+4.7). These consistent improvements confirm that visual comparison significantly enhances entity identification beyond the retriever's global feature matching. Meanwhile, the remaining gap relative to the upper bound (R@15) suggests that discrimination among a large number of visually similar entities remains challenging.

4.3 Qualitative Results

Fig. 3 presents qualitative examples on E-VQA [20] and InfoSeek [6]. In the E-VQA case, the moth candidates share similar coloration and wing shapes, differing only in subtle markings; MMAgent-R^2 correctly identifies the target entity through visual comparison, whereas the variant without reranking selects the incorrect top-1 candidate, leading to a wrong answer. In the InfoSeek case, the building candidates exhibit similar spire structures, and the correct entity is absent from the first batch; MMAgent-R^2 actively rejects this batch and successfully identifies the target in the subsequent retrieval. Fig. 4 illustrates a comparison question from MMhops [35], where the model must simultaneously identify entities in both query images and retrieve relevant information for reasoning. These cases demonstrate that visual reranking and rejection effectively prevent retrieval errors from propagating to the final answer.

4.4 Ablation Studies

We train MMAgent-R^2-8B ablation variants on 26K samples from InfoSeek to verify the contribution of each key component.

Effect of Rerank and Reject. As shown in Table 5, removing both mechanisms reduces the model to answering based on the retriever's top-1 result (Ident. 53.7%, VQA accuracy 40.2%). Adding visual reranking improves Ident.

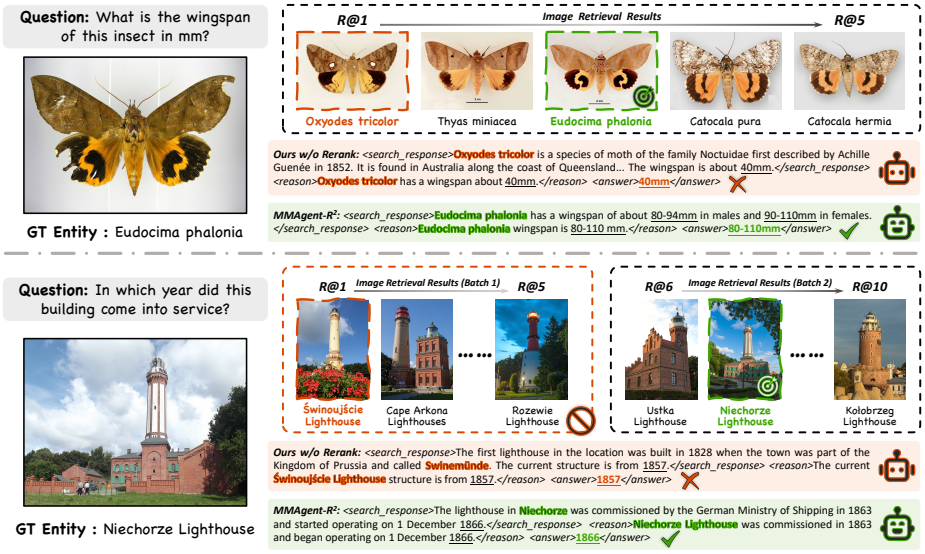

Fig. 3: Qualitative examples of MMAgent-R^2. Top: E-VQA. Bottom: InfoSeek.

Table 5: Effect of reranking and rejection. Ident.: entity identification accuracy. U-Q / U-E: Unseen-Question / Unseen-Entity.

Variant	Ident.	U-Q	U-E	All
w/o Rerank, w/o Reject	53.7	40.5	39.9	40.2
w/o Reject	57.5	44.4	44.3	44.3
w/ Rerank + Reject	58.6	45.9	45.2	45.5

Table 6: Effect of candidate batch size. Avg. Rej.: average rejection count per sample.

K	Ident.	Avg. Rej.	U-Q	U-E	All
3	58.0	0.63	45.1	44.3	44.7
5	58.6	0.59	45.9	45.2	45.5
7	58.7	0.51	46.0	45.4	45.7

and VQA accuracy to 57.5% and 44.3% (+3.8/+4.1), confirming visual comparison as the primary driver. Further incorporating rejection reaches 58.6% and 45.5%, showing that rejection provides additional gains by expanding candidate coverage. The two mechanisms thus improve entity identification from complementary dimensions: precision and candidate coverage.

Effect of Candidate Batch Size K. Table 6 shows the effect of the candidate batch size K, *i.e.*, the number of candidates presented per reranking step. At K=3, the limited per-step coverage forces more frequent rejections (Avg. Rej. 0.63), resulting in lower VQA accuracy (44.7%). At K=7, broader coverage leads to fewer rejections (0.51) and marginally improves accuracy (45.7%), but at higher token cost with diminishing returns. K=5 strikes the best balance between accuracy and efficiency.

Effect of Maximum Rejection Count. As shown in Table 7, performance improves steadily as the rejection limit increases from 0 to 2 (VQA accuracy from 44.3% to 45.5%), but declines at 3 (44.8%), as redundant candidates lengthen the context and degrade reasoning quality. Notably, average rejections remain

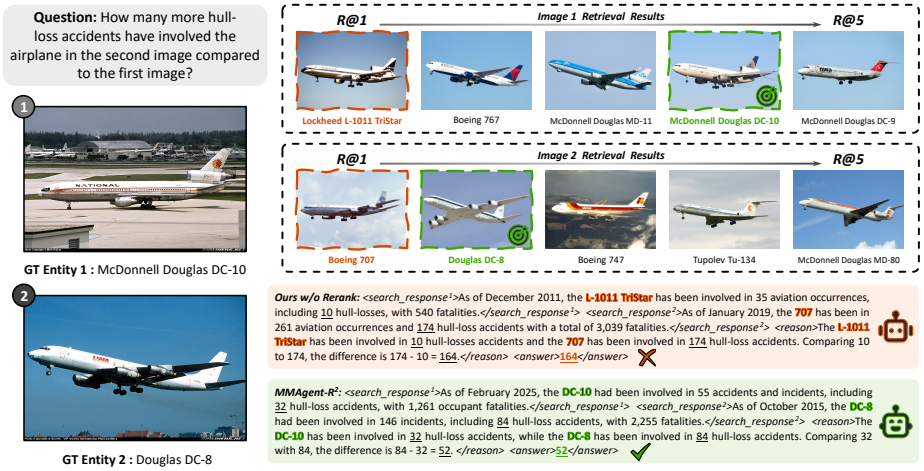

Fig. 4: Qualitative example of MMAgent-R^2 on MMhops comparison question.

Table 7: Effect of maximum rejection count.

$\text{Rej}_{\max}$	Ident.	Avg. Rej.	U-Q	U-E	All
0	57.5	0	44.4	44.3	44.3
1	58.3	0.46	45.2	44.9	45.0
2	58.6	0.59	45.9	45.2	45.5
3	58.1	0.74	45.1	44.5	44.8

Table 8: Effect of rejection reward weight δ in Eq. (4).

δ	Ident.	Avg. Rej.	U-Q	U-E	All
0.1	57.9	0.42	45.2	44.6	44.9
0.2	58.6	0.59	45.9	45.2	45.5
0.3	58.4	0.68	45.5	44.9	45.2

well below the allowed maximum (*e.g.*, only 0.59 at $\text{Rej}_{\max}$=2), indicating that the model learns to reject on demand rather than expanding indiscriminately.

Effect of Rejection Reward Weight δ. As shown in Table 8, performance peaks at δ=0.2 (VQA accuracy 45.5%): a weaker incentive (δ=0.1) under-utilizes the rejection mechanism, yielding lower accuracy (44.9%), while a stronger one (δ=0.3) triggers unnecessary rejections that lengthen the reasoning context and slightly degrade accuracy to 45.2%. A moderate δ thus best balances encouraging necessary and suppressing excessive rejections.

5 Conclusion

We propose MMAgent-R^2, which equips agentic mRAG with visual-level verification over retrieved candidates by introducing visual reranking and active rejection. Combined with a composite reward function incorporating step-level verification rewards and GRPO optimization, MMAgent-R^2 achieves state-of-the-art performance on multiple KB-VQA benchmarks, with particularly notable gains in challenging retrieval scenarios and multi-image multi-hop reasoning tasks. Our results suggest that empowering models with active verification and error correction capabilities is a key direction for advancing knowledge-based VQA.

Limitations. The performance of MMAgent-R^2 is still bounded by the retriever's recall ceiling, and injecting numerous candidate images into the reasoning context incurs additional token overhead. Future work may explore compressing candidate image information while preserving perceptual precision, as well as joint optimization with the retriever.

Acknowledgements

This research was supported by multiple funding sources, including the New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0123401), the National Natural Science Foundation of China (U24A20331, 62302501, 62192782, 62532015, 62536002, U2441241), Beijing Natural Science Foundation (L251005, L223003, L243015, L252032), and Beijing Major Science and Technology Project (Z251100008425008).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
4. Bulian, J., Buck, C., Gajewski, W., Börschinger, B., Schuster, T.: Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 291–305 (2022)
5. Caffagni, D., Cocchi, F., Moratelli, N., Sarto, S., Cornia, M., Baraldi, L., Cucchiara, R.: Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1818–1826 (2024)
6. Chen, Y., Hu, H., Luan, Y., Sun, H., Changpinyo, S., Ritter, A., Chang, M.W.: 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. pp. 14948–14968 (2023)
7. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
8. Compagnoni, A., Morini, M., Sarto, S., Cocchi, F., Caffagni, D., Cornia, M., Baraldi, L., Cucchiara, R.: Reag: Reasoning-augmented generation for knowledge-based visual question answering. arXiv preprint arXiv:2511.22715 (2025)

9. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems **36**, 49250–49267 (2023)
10. Deng, L., Sun, Y., Chen, S., Yang, N., Wang, Y., Song, R.: Muka: Multimodal knowledge augmented visual information-seeking. In: Proceedings of the 31st International Conference on Computational Linguistics. pp. 9675–9686 (2025)
11. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library. IEEE Transactions on Big Data (2025)
12. Hong, Y., Gu, J., Yang, Q., Fan, L., Wu, Y., Wang, Y., Ding, K., Xiang, S., Ye, J.: Knowledge-based visual question answer with multimodal processing, retrieval and filtering. arXiv preprint arXiv:2510.14605 (2025)
13. Lerner, P., Ferret, O., Guinaudeau, C.: Cross-modal retrieval for knowledge-based visual question answering. In: European Conference on Information Retrieval. pp. 421–438. Springer (2024)
14. Lerner, P., Ferret, O., Guinaudeau, C., Le Borgne, H., Besançon, R., Moreno, J.G., Lovón Melgarejo, J.: 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. pp. 3108–3120 (2022)
15. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
16. Li, Y., Li, Y., Wang, X., Jiang, Y., Zhang, Z., Zheng, X., Wang, H., Zheng, H.T., Yu, P.S., Huang, F., et al.: Benchmarking multimodal retrieval augmented generation with dynamic vqa dataset and self-adaptive planning agent. arXiv preprint arXiv:2411.02937 (2024)
17. Lin, W., Byrne, B.: Retrieval augmented visual question answering with outside knowledge. In: Proceedings of the 2022 conference on empirical methods in natural language processing. pp. 11238–11254 (2022)
18. Ling, Z., Guo, Z., Huang, Y., An, Y., Xiao, S., Lan, J., Zhu, X., Zheng, B.: Mmkb-rag: A multi-modal knowledge-based retrieval-augmented generation framework. arXiv preprint arXiv:2504.10074 (2025)
19. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. pp. 3195–3204 (2019)
20. Mensink, T., Uijlings, J., Castrejon, L., Goel, A., Cadar, F., Zhou, H., Sha, F., Araujo, A., Ferrari, V.: Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3113–3124 (2023)
21. Qi, J., Xu, Z., Shao, R., Chen, Y., Di, J., Cheng, Y., Wang, Q., Huang, L.: Rora-vlm: Robust retrieval-augmented vision language models. arXiv preprint arXiv:2410.08876 (2024)
22. Qiu, J., Madotto, A., Lin, Z., Crook, P.A., Xu, Y.E., Damavandi, B., Dong, X.L., Faloutsos, C., Li, L., Moon, S.: Snapntell: Enhancing entity-centric visual question answering with retrieval augmented multimodal llm. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 247–266 (2024)
23. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., Mottaghi, R.: A-okvqa: A benchmark for visual question answering using world knowledge. In: European conference on computer vision. pp. 146–162. Springer (2022)

24. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
25. Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H., Wu, C.: Hybridflow: A flexible and efficient rlhf framework. arXiv preprint arXiv: 2409.19256 (2024)
26. Sun, Q., Fang, Y., Wu, L., Wang, X., Cao, Y.: Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389 (2023)
27. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8769–8778 (2018)
28. Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., Wei, F.: Text embeddings by weakly-supervised contrastive pre-training. arXiv preprint arXiv:2212.03533 (2022)
29. Weyand, T., Araujo, A., Cao, B., Sim, J.: Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2575–2584 (2020)
30. Wu, W., Song, Z., Zhou, K., Shao, Y., Hu, Z., Huang, B.: Towards general continuous memory for vision-language models. arXiv preprint arXiv:2505.17670 (2025)
31. Yan, Y., Xie, W.: Echosight: Advancing visual-language models with wiki knowledge. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 1538–1551 (2024)
32. Ye, W., Su, Y., Chen, Y., Gao, L., Li, J., Li, R., Zhang, R.: Qkvqa: Question-focused filtering for knowledge-based vqa. arXiv e-prints pp. arXiv–2601 (2026)
33. Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., et al.: Dapo: An open-source llm reinforcement learning system at scale. arXiv preprint arXiv:2503.14476 (2025)
34. Yuan, X., Ning, L., Fan, W., Li, Q.: mkg-rag: Multimodal knowledge graph-enhanced rag for visual question answering. arXiv preprint arXiv:2508.05318 (2025)
35. Zhang, T., Zhang, Z., Ma, Z., Chen, Y., Li, B., Yuan, C., Wang, G., Rao, F., Shan, Y., Hu, W.: Mmhops-r1: Multimodal multi-hop reasoning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 28391–28399 (2026)
36. Zhang, T., Zhang, Z., Ma, Z., Chen, Y., Qi, Z., Yuan, C., Li, B., Pu, J., Zhao, Y., Xie, Z., et al.: mr^{2}ag: Multimodal retrieval-reflection-augmented generation for knowledge-based vqa. arXiv preprint arXiv:2411.15041 (2024)