# MMhops-R1: Multimodal Multi-hop Reasoning

**Tao Zhang**[1, 2, 3, 4], **Ziqi Zhang**[1, 3], **Zongyang Ma**[1, 3], **Yuxin Chen**[4], **Bing Li**[1, 3, 5*], **Chunfeng Yuan**[1, 3], **Guangting Wang**[4], **Fengyun Rao**[4], **Ying Shan**[4], **Weiming Hu**[1, 2, 3, 6]

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

## Abstract

The ability to perform multi-modal multi-hop reasoning by iteratively integrating information across various modalities and external knowledge is critical for addressing complex real-world challenges. However, existing Multi-modal Large Language Models (MLLMs) are predominantly limited to single-step reasoning, as existing benchmarks lack the complexity needed to evaluate and drive multi-hop abilities. To bridge this gap, we introduce **MMhops**, a novel, large-scale benchmark designed to systematically evaluate and foster multi-modal multi-hop reasoning. MMhops dataset comprises two challenging task formats, **Bridging** and **Comparison**, which necessitate that models dynamically construct complex reasoning chains by integrating external knowledge. To tackle the challenges posed by MMhops, we propose **MMhops-R1**, a novel multi-modal Retrieval-Augmented Generation (mRAG) framework for dynamic reasoning. Our framework utilizes reinforcement learning to optimize the model for autonomously planning reasoning paths, formulating targeted queries, and synthesizing multi-level information. Comprehensive experiments demonstrate that MMhops-R1 significantly outperforms strong baselines on MMhops, highlighting that dynamic planning and multi-modal knowledge integration are crucial for complex reasoning. Moreover, MMhops-R1 demonstrates strong generalization to tasks requiring fixed-hop reasoning, underscoring the robustness of our dynamic planning approach.

**Code** — https://github.com/taoszhang/MMhops-R1

## Introduction

With continuous advancement in reasoning capabilities, Large Language Models (LLMs) like OpenAI's o1 (Jaech et al. 2024), DeepSeek-R1 (Guo et al. 2025), and Kimi-k1.5 (Team et al. 2025) demonstrate strong performance in complex problem-solving by extending chain-of-thought reasoning during inference. Multimodal large language models (MLLMs), by inheriting the reasoning abilities or adopting similar training paradigms, achieve significant progress in integrating visual understanding and language reasoning (Xu et al. 2024; Peng et al. 2025; Zhang et al. 2025; Zheng et al. 2025). However, current multimodal reasoning research primarily focuses on stimulating intrinsic model capabilities, such as spatial reasoning (Zhou et al. 2025), object detection (Chen et al. 2025a; Liu et al. 2025), and mathematical reasoning (Meng et al. 2025; Leng et al. 2025). By contrast, complex real-world problems typically require the integration of multimodal reasoning with external knowledge retrieval through multi-turn interactions, enabling multimodal multi-hop reasoning. For example, Figure 1($B_1$) requires the model to extract information from the image, retrieve relevant external knowledge, and perform multi-step reasoning to reach the answer. Figure 1($B_2$) requires identifying entities across multiple images, retrieving corresponding knowledge, and conducting quantitative reasoning.


Figure 1: Comparison of reasoning types. (A) Previous KB-VQA: Single-step visual recognition followed by knowledge retrieval. ($B_1$) Bridging reasoning: Multi-step sequential inference on a single image. ($B_2$) Comparison reasoning: Cross-image entity identification and comparative analysis.

Despite progress, existing Visual Question Answering

*Corresponding author.

(VQA) datasets remain insufficient for multimodal multi-hop reasoning. Current datasets are limited in both visual and textual reasoning depth: standard VQA datasets typically require only single-step visual understanding (Goyal et al. 2017; Hudson and Manning 2019; Singh et al. 2019), while several knowledge-based VQA datasets (Marino et al. 2019; Schwenk et al. 2022; Lerner et al. 2022; Chen et al. 2023) introduce external knowledge retrieval to increase complexity. However, as shown in Figure 1(A), models usually use one step of visual recognition and one step of text retrieval to answer, without constructing complex reasoning chains. E-VQA (Mensink et al. 2023) extends questions to two-hop reasoning, but this extension remains restricted to the textual domain and features a fixed reasoning path length, lacking multimodal integration and diverse reasoning types. These limitations make existing datasets inadequate for effectively supporting model training and evaluation in complex multimodal multi-hop reasoning tasks.

Based on these challenges, we propose **MMhops**, a novel large-scale Multimodal Multi-hop reasoning dataset that systematically increases reasoning depth in both visual and textual dimensions. MMhops features two types of reasoning tasks: **Bridging** reasoning and **Comparison** reasoning. Bridging reasoning starts from a single image and requires the model to perform multi-step chain reasoning, with each step building on the previous one, supporting reasoning depths from two hops and beyond. Comparative reasoning is based on multiple images, requiring the model to identify multiple visual entities and compare their shared attributes, involving cross-image information integration and comparative analysis. Both task types demand deep reasoning abilities in visual understanding and textual inference, enabling the model to decompose complex questions and dynamically construct answers through multi-round interactions, thus providing a comprehensive evaluation of multimodal reasoning and knowledge integration capabilities.

To address the challenges of multimodal multi-hop reasoning, we propose **MMhops-R1**, the first framework to leverage reinforcement learning (RL) for multimodal multi-hop reasoning. MMhops-R1 adopts a dynamic interaction strategy that overcomes the limitations of fixed processes in conventional multimodal Retrieval-Augmented Generation (mRAG) frameworks. Specifically, the model supports three core actions: **1)** selecting an input image and invoking the image retriever; **2)** submitting a text query to the text retriever; and **3)** generating answer based on the current information. With a tailored reward mechanism, MMhops-R1 can autonomously select reasoning strategies, dynamically adjust reasoning depth according to question complexity, and adaptively plan the reasoning path.

We evaluate MMhops-R1 on the proposed MMhops benchmark against four categories of strong baselines: open-source MLLMs, multi-hop RAG, multimodal RAG, and proprietary MLLMs. Results demonstrate the profound effectiveness of our proposed RL-driven framework for dynamic mRAG and underscores two critical requirements for complex multi-modal reasoning: the ability to integrate multimodal external knowledge and to dynamically interact with a retrieval system. Furthermore, MMhops-R1 shows strong generalization, achieving robust performance on the single-hop questions from INFOSEEK and the two-hop questions from E-VQA. These findings validate our contributions and highlight the potential of our approach to drive future research in multi-modal multi-hop reasoning.

Our contributions are summarized as follows:

- We introduce **MMhops**, the first large-scale benchmark for multimodal multi-hop reasoning, requiring the synthesis of diverse visual and textual information across various reasoning depths.
- We propose **MMhops-R1**, a novel mRAG framework that leverages reinforcement learning to optimize the model, enabling it to dynamically interact with multiple retrievers and adaptively plan the reasoning path.
- We set a new state-of-the-art on multimodal multi-hop reasoning tasks, demonstrating the superiority of our dynamic mRAG framework over existing methods.

## Related Work

**Knowledge-Based VQA.** To advance VQA beyond perception towards more complex reasoning, the task of KB-VQA was introduced, which requires models to incorporate external knowledge. However, prominent KB-VQA datasets like OK-VQA (Marino et al. 2019) and A-OKVQA (Schwenk et al. 2022) were largely confined to commonsense knowledge or simple facts. Subsequent efforts, including ViQuAE (Lerner et al. 2022) and INFOSEEK (Chen et al. 2023), expanded the knowledge domain to large-scale corpora such as Wikipedia. Nonetheless, these datasets predominantly feature questions solvable via a two-step process: identifying a visual entity and executing a single query against a knowledge base. While the recent E-VQA (Mensink et al. 2023) dataset introduced textual two-hop reasoning, its reasoning chains are confined to the textual modality and a fixed length. In contrast, the MMhops dataset is the first to systematically require multi-hop reasoning across both visual and textual modalities, featuring diverse, variable-length reasoning paths, demanding a more profound integration of multi-modal information.

**Multimodal and Multi-hop RAG.** Early mRAG frameworks (Caffagni et al. 2024; Yan and Xie 2024; Zhang et al. 2024) typically employ a static, single-step pipeline: they first retrieve relevant documents based on the initial query and then feed them to the generator. A key limitation of these approaches is their reliance on a static, pre-defined process, which lacks the flexibility to adapt to queries of varying complexity. While recent work like OmniSearch (Comanici et al. 2025) introduces a planning agent, it relies on manually engineered prompts or supervised fine-tuning, which does not equip the model with the intrinsic capability to learn complex reasoning policies autonomously. In parallel, multi-hop RAG has emerged in the unimodal text domain to address similar challenges. To move beyond fixed reasoning chains, methods such as Search-R1 (Jin et al. 2025) and ReSearch (Chen et al. 2025b) leverage reinforcement learning (RL) with algorithms like GRPO (Shao et al. 2024) and PPO (Schulman et al. 2017) to train an agent that learns a dynamic retrieval policy. However, these powerful RL-based

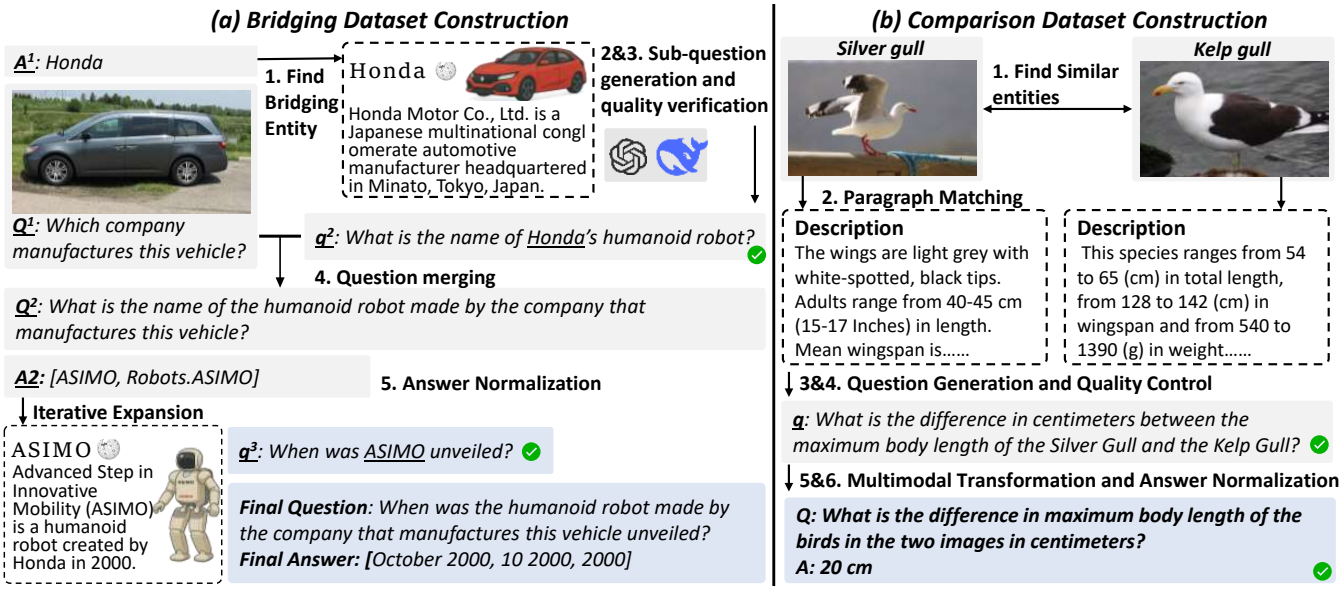


Figure 2: The multi-stage construction process for the MMhops dataset.

paradigms have thus far been confined to the textual modality. Our work, MMhops-R1, bridges this divide by extending this RL-based paradigm to the mRAG domain, training an agent to strategically orchestrate retrieval and reasoning across both visual and textual knowledge sources.

## MMhops Dataset

In this section, we present MMhops, a large-scale multimodal multi-hop reasoning dataset. MMhops requires models to: (1) interact with diverse external knowledge sources for targeted retrieval; (2) perform multi-step reasoning by dynamically integrating and updating knowledge across multiple retrieval and reasoning steps; (3) align and combine information from multiple images for cross-image and cross-modal reasoning. Through a well-designed data construction and evaluation framework, MMhops serves as an important resource for multimodal reasoning research.

### MMhops Construction

The MMhops dataset is built based on the Wikipedia Knowledge Base (KB), with an automated data annotation and quality filtering process designed using powerful language models like GPT-4o (Hurst et al. 2024). The dataset includes two core reasoning types: Bridging Questions and Comparison Questions, covering various reasoning depths and different numbers of image inputs.

**Bridging Dataset** Bridging Questions begin with the visual information from a single image and progressively link relevant entities and knowledge through multi-step chain reasoning. We generate them iteratively by starting with single-hop questions and progressively increasing the reasoning depth, a process depicted in Figure 2 (a).

**Initialization Phase:**

0. **Data Collection:** Gather existing single-hop knowledge-based datasets $(V, Q^1, A^1)$ as the foundation for constructing multi-hop reasoning chains.

**Iterative Expansion Phase:**

1. **Bridging Entity Identification:** From the current $n$ $(\geq 1)$ hop dataset $(V, Q^n, A^n)$, select samples where the answer corresponds to a Wikipedia entity $(V_i, Q_i^n, A_i^n)$, excluding vague entity types such as numbers or years. The answer $A_i^n$ is designated as the bridging entity for subsequent reasoning chains.
2. **Sub-question Generation:** Using the Wikipedia page of the bridging entity $A_i^n$, prompt a large language model to generate a knowledge-based question $q_i^{n+1}$, ensuring that the entity $A_i^n$ is explicitly mentioned in the question and labeling the answer $A_i^{n+1}$.
3. **Question Quality Control:** Verify that the sub-question $q_i^{n+1}$ meets the criteria of an independent single-hop question, meaning that removing the entity $A_i^n$ from $q_i^{n+1}$ should render the question unanswerable.
4. **Question Merging:** Merge the sub-question $q_i^{n+1}$ with the current question $Q_i^n$, replacing the reference to the bridging entity $A_i^n$ in $q_i^{n+1}$ with the current question $Q_i^n$, resulting in the complete $(n+1)$ hop question $Q_i^{n+1}$.
5. **Answer Normalization:** Categorize the answer $A_i^{n+1}$ into three types: numerical values, time-related entities (e.g., years, dates), and strings. Construct standardized answer sets for each category.

Through this iterative process, we systematically develop multi-level reasoning question sets, ranging from two-hop to three-hop.

**Comparison Dataset** This question type evaluates cross-image reasoning, as exemplified in Figure 2 (b). Generating

| Dataset | Scale | Visual Reasoning | Text Reasoning | Total Reasoning | Multi-image | Knowledge Source |
|---|---|---|---|---|---|---|
| OK-VQA (Marino et al. 2019) | 14K | 1 | 1 | 2 | ✗ | Factoid |
| A-OKVQA (Schwenk et al. 2022) | 24.9K | 1 | 1 | 2 | ✗ | Common sense/Factoid |
| ViQuAE (Lerner et al. 2022) | 3.7K | 1 | 1 | 2 | ✗ | Wikipedia |
| INFOSEEK (Chen et al. 2023) | 1.35M | 1 | 1 | 2 | ✗ | Wikipedia |
| E-VQA (Mensink et al. 2023) | 1M | 1 | 1-2 | 2-3 | ✗ | Wikipedia |
| **MMhops** | **31.1K** | **1-2** | **2-3** | **3-4** | ✓ | **Wikipedia** |

Table 1: Comparison with Existing Knowledge-based VQA Datasets.

| Statistic Dimension | Value | Percentage |
|---|---|---|
| **Dataset Scale** | | |
| Total VQA Samples | 31,117 | 100.0% |
| Bridging VQA Samples | 26,437 | 85.0% |
| Comparison VQA Samples | 4,680 | 15.0% |
| Number of Questions Involved | 20,483 | - |
| Number of Entities Involved | 8,832 | - |
| Number of Images Involved | 28,256 | - |
| **Reasoning Complexity** | | |
| Requiring External Knowledge | 31,117 | 100.0% |
| 3 steps | 22,016 | 70.8% |
| 4 steps | 9,101 | 29.2% |
| **Content Characteristics** | | |
| Average Question Length (words) | 17.3 | - |
| Average Answer Length (words) | 1.6 | - |
| **Answer Type Distribution** | | |
| Entity-type Answers | 5,923 | 19.0% |
| Temporal Answers | 5,016 | 16.1% |
| Numerical Answers | 20,178 | 64.9% |

Table 2: Statistics of MMhops Dataset

these questions requires identifying entities across multiple images and utilizing external knowledge to formulate a comparative query. The construction process is as follows:

1. **Entity Collection:** Collect a large number of visual entities from the Wikipedia knowledge base. Use the embedding model NV-Embed-v2 (Lee et al. 2024) to perform semantic similarity matching based on the entity names and summary, selecting entities with high relevance. Use LLMs to perform semantic deduplication and remove pairs of entities that refer to the same concept.
2. **Paragraph Matching:** For the selected similar entity pairs, use a rule-based method to extract paragraphs with the same title that contain quantifiable numerical information, providing consistent background knowledge for subsequent comparative analysis.
3. **Question Generation:** Based on the entity pairs and their background knowledge paragraphs, prompt the LLMs to focus on quantifiable attributes and generate questions that compare the attributes of the two entities.
4. **Quality Control:** Perform automated validation to ensure that the questions are clearly stated and the answers are quantifiable and verifiable.
5. **Multimodal Transformation:** Replace the entity names in the questions with corresponding images to construct a multimodal reasoning scenario that forces the model to reason based on visual content.
6. **Answer Normalization:** Standardize numerical answers, ensuring that clear units are included in the question to support accurate evaluation.

## MMhops Analysis

MMhops is the first large-scale dataset designed for multimodal multi-hop reasoning. As detailed in Table 2, the dataset comprises 31,117 samples, which include 20,483 unique questions, 8,832 distinct entities, and 28,256 images. A key feature of MMhops is its focus on complex reasoning chains; all samples require more than two reasoning hops that span both visual and textual modalities. Specifically, 70.8% of samples require three reasoning steps and 29.2% require four. Furthermore, all samples necessitate the integration of external knowledge. Linguistically, the average question length is 17.3 words, with concise answers averaging 1.6 words. Most answers are numerical, facilitating precise evaluation of the model's reasoning ability.

As detailed in Table 1, existing Knowledge-based VQA (KVQA) datasets are largely confined to shallow reasoning, typically involving a single visual step and 1–2 textual reasoning hops. Consequently, they are insufficient for evaluating complex, multi-step reasoning abilities. MMhops dataset incorporates multi-image inputs, which necessitates 1–2 steps of cross-image relational reasoning. Furthermore, we extend the textual reasoning depth to 2–3 hops via a scalable, iterative pipeline. Collectively, these enhancements result in a total reasoning depth of 3–4 steps, establishing MMhops as a more challenging and practical benchmark to drive progress in advanced multimodal reasoning.

## Dataset Splits

We split the MMhops dataset into training, validation, and test sets with a 7:1:2 ratio using stratified sampling based on reasoning depth and question type.

# Methodology

## Problem Formulation

We consider the task of answering a question $Q$ based on a collection of images $\mathcal{I} = \{I_1, \ldots, I_n\}$. The policy model $\pi_\theta$ can leverage a set of external retrievers $\mathcal{R} = \{R_I, R_T\}$. $R_I$ is an image retriever that, given a query image, returns the information about the most similar image. $R_T$ is a text

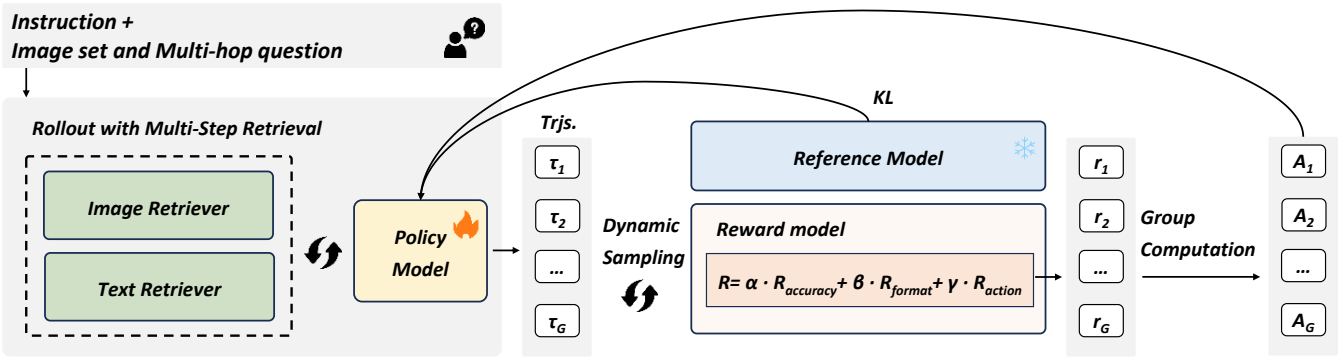


Figure 3: Overview of the training pipeline for MMhops-R1.

retriever that, given a text query, returns the top-$k$ most relevant passages. The model's action space is defined as $\mathcal{A} = \{a_t, a_{is}, a_{ts}, a_a\}$, where $a_t$ represents thinking and reasoning based on all current inputs, $a_{is}$ and $a_{ts}$ invoke the retrievers $R_I$ and $R_T$, respectively, and $a_a$ terminates the process by generating the final response. The training pipeline for our policy model is illustrated in Figure 3.

## Rollout with Multi-Step Retrieval

At each time step $t$, the policy model $\pi_\theta$ first performs thinking based on the current state $s_t$ (containing historical interaction information), then selects the next action from the action set $\{a_{is}, a_{ts}, a_a\}$. The specific action execution mechanisms are as follows:

- **Image Retrieval** ($a_{is}$): The policy issues a query to an image retriever $R_I$ by generating target image indices, formatted within `<image_search>` and `</image_search>` tags. The retriever returns the corresponding image information as observation $o_{t+1}$.
- **Text Retrieval** ($a_{ts}$): The policy generates a textual query, formatted within `<text_search>` and `</text_search>` tags, for a text retriever $R_T$, which returns the top-$k$ relevant passages as observation $o_{t+1}$.
- **Answer** ($a_a$): The policy generates the final answer, formatted within `<answer>` and `</answer>` tags, based on the information gathered throughout the trajectory. This is a terminal action that concludes the episode.

If the policy generates an action with a malformed syntax or one outside $\mathcal{A}$, the environment provides a fixed penalty signal as the observation $o_{t+1}$ to encourage valid action generation. The rollout process terminates when the policy executes the answer action $a_a$ or a maximum of $T$ steps is reached. This interaction generates a trajectory $\tau$, defined as a sequence of states, actions, and observations:

$$\tau = \{(s_0, a_0, o_1), (s_1, a_1, o_2), \ldots, (s_T, a_T)\}. \quad (1)$$

## Reward Modeling

To guide the model's generation, we design a composite reward function for MMhops-R1. This function comprises three components designed to promote correctness, structural clarity, and effective tool use. For a given trajectory $\tau$, the total reward is a weighted sum of these components.

1. **Outcome Reward ($R_{\text{outcome}}$).** This binary reward evaluates the correctness of the final answer. It is defined as $R_{\text{outcome}}(\tau) = 1$ if the model's answer matches the ground truth, and 0 otherwise.
2. **Format Reward ($R_{\text{format}}$).** This binary reward encourages adherence to the previously defined structured format. A trajectory receives $R_{\text{format}}(\tau) = 1$ if all generated thoughts and actions are correctly formatted with their respective tags, and 0 otherwise.
3. **Action Reward ($R_{\text{action}}$).** This rewards effective tool use. A key aspect of our design is that this reward is gated by the overall success of the trajectory. It is only granted if the model both produces the correct final answer and adheres to the required format. This encourages the model to learn tool-use policies that directly contribute to successful outcomes. The reward is defined as:

$$R_{\text{action}}(\tau) = R_{\text{outcome}}(\tau) \cdot R_{\text{format}}(\tau) \cdot R_{\text{tool}}(\tau) \quad (2)$$

where $R_{\text{tool}}(\tau)$ is a separate reward, defined as the number of syntactically correct tool invocations.

The total reward for a trajectory $\tau$ is a weighted sum of these components:

$$R(\tau) = \alpha \cdot R_{\text{outcome}}(\tau) + \beta \cdot R_{\text{format}}(\tau) + \gamma \cdot R_{\text{action}}(\tau) \quad (3)$$

where $\alpha, \beta,$ and $\gamma$ are non-negative hyperparameters that balance the contribution of each component.

## Objective Function

To optimize our policy $\pi_\theta$ using a composite reward signal, we adapt the objective function from DAPO (Yu et al. 2025). This objective is coupled with a dynamic sampling strategy that filters generated response groups. Specifically, we enforce that each group of $G$ responses must contain at least one factually correct sample, as stipulated by the constraint in our objective. This design enables the policy to optimize for procedural correctness by creating non-zero advantages ($\hat{A}_{i,t}$) from process-based rewards, such as format adherence ($R_{\text{format}}$), even when the final outcome is already correct.

| Method | Base Model | Retriever | Bridging | | | | Comparison |
|---|---|---|---|---|---|---|---|
| | | | String | Numerical | Time | Overall | |
| **Closed-sourced model** | | | | | | | |
| GPT-4o-mini (Hurst et al. 2024) | – | – | 29.67 | 20.54 | 26.08 | 23.80 | 7.05 |
| GPT-4o (Hurst et al. 2024) | – | – | 41.66 | 33.60 | 39.28 | 36.62 | 8.76 |
| Gemini-2.5-flash (Comanici et al. 2025) | – | – | 51.08 | 43.51 | 50.10 | 46.58 | 23.18 |
| Gemini-2.5-pro (Comanici et al. 2025) | – | – | **58.80** | **50.83** | **57.32** | **53.98** | **29.39** |
| **Direct Answer** | | | | | | | |
| Zero-shot | Qwen2.5-vl-7B-Instruct | – | 24.21 | 15.24 | 26.60 | 19.53 | 6.20 |
| Zero-shot | Qwen2.5-vl-72B-Instruct | – | 37.51 | 32.11 | 37.32 | 34.39 | 7.59 |
| **Multi-hop RAG (Text-only)** | | | | | | | |
| Search-r1 (Jin et al. 2025) | Qwen2.5-7b-Instruct | Caption, Text | 14.45 | 23.05 | 17.85 | 19.98 | 6.62 |
| Self-Ask (Press et al. 2022) | GPT-4o | Caption, Text | 27.59 | 31.41 | 31.13 | 30.42 | 18.27 |
| **Multimodal RAG** | | | | | | | |
| Vanilla mRAG | Qwen2.5-vl-7B-Instruct | Text | 14.37 | 14.65 | 14.95 | 14.63 | 3.95 |
| Vanilla mRAG | Qwen2.5-vl-7B-Instruct | Image, Text | 26.52 | 25.68 | 28.97 | 26.49 | 9.72 |
| EchoSight (Yan and Xie 2024) | LLaMA3 | Image, Text | 19.14 | 11.83 | 11.86 | 13.63 | 4.81 |
| OmniSearch (Li et al. 2024) | GPT-4o | Image, Text | 31.02 | 49.77 | 36.5 | 42.65 | 17.02 |
| **MMhops-R1 (Ours)** | **Qwen2.5-vl-7B-Instruct** | **Image, Text** | **44.66** | **55.33** | **47.94** | **51.35** | **22.01** |

Table 3: Main results on MMhops.

Our full optimization objective is formulated as:

$$J(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q;R)} \left[ \frac{1}{\sum_{i=1}^{G}|o_i|} \sum_{i=1}^{G}\sum_{t=1}^{|o_i|} \min\Big( r_{i,t}(\theta)\hat{A}_{i,t}, \text{clip}\left(r_{i,t}(\theta), 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}\right)\hat{A}_{i,t} \Big) \right] \quad (4)$$

$$\text{s.t.} \quad 0 < \left|\{o_i \mid \text{is\_equivalent}(a, o_i)\}\right|.$$

where

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_i, \leq t; R)}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_i, \leq t; R)}, \text{and}$$

$$\hat{A}_{i,t} = \frac{R_i - \text{mean}(\{R_i\}_{i=1}^{G})}{\text{std}(\{R_i\}_{i=1}^{G})}$$

Where $R$ represents the retriever, and the model samples while interacting with multiple retrievers $R = \{R_I, R_T\}$.

## Loss Masking for External Observations

The observation $o_{t+1}$ at each timestep $t$ includes tokens from external sources, such as results from the image ($R_I$) and text ($R_T$) retrievers and environmental feedback on invalid actions. Since these tokens are not generated by the policy model, we mask them from the loss computation during policy optimization. This ensures that the optimization objective is confined to the model's own generated reasoning and action tokens, improving training stability.

# Experiments

## Experimental Settings

**Implementation Details** We optimize our policy using the Verl framework (Sheng et al. 2024), employing Qwen2.5-VL-7B-Instruct (Bai et al. 2025) as the backbone model. The model is trained for a single epoch on the MMhops dataset with a constant learning rate of $1 \times 10^{-6}$. During policy optimization, we use a batch size of 256 and a group size of 8. Our knowledge base for retrieval comprises 100K Wikipedia articles, each accompanied by an image. For image retrieval, we utilize the CLIP-ViT-L/14@336px model (Radford et al. 2021). For text retrieval, we employ the E5 model (Wang et al. 2022) to fetch the top-3 most relevant passages for each query. The maximum number of interaction turns with the knowledge base is set to 4 during both training and inference. The hyperparameters $\alpha$, $\beta$, and $\gamma$ for the reward function are set to 1.0, 1.0, and 0.25, respectively.

**Evaluation Metrics** We adopt the evaluation protocol from INFOSEEK (Chen et al. 2023), categorizing answers into three types: STRING, TIME, and NUMERICAL. For STRING answers, we report Exact Match (EM) accuracy. For TIME answers, we employ EM with a tolerance of $\pm 1$ year. For NUMERICAL answers, a prediction is deemed correct if it falls within a $\pm 0.1$ margin of the ground truth or achieves an Intersection-over-Union (IoU) of at least 50%. The overall score is the weighted average of the accuracies for each type. We report performance on the *Test* set, with a breakdown for bridging and comparison questions.

## Comparison with SOTAs

To enable a comprehensive comparison with existing approaches, we evaluate four categories of models: advanced open-source general-purpose multimodal large models with direct answer generation, single-modal text-only multi-hop RAG methods (which convert images into descriptions and combine them with the question as input), multimodal RAG methods, and benchmark closed-source MLLMs. For fairness, all compared methods share the same image and text retriever as ours. Detailed results are presented in Table 3,

| Model | INFOSEEK | | |
|---|---|---|---|
| | Unseen Q | Unseen E | Overall |
| CLIP-PaLM (Chen et al. 2023) | 22.7 | 18.5 | 20.4 |
| CLIP-FiD (Chen et al. 2023) | 23.3 | 19.1 | 20.9 |
| Wiki-LLaVA (Caffagni et al. 2024) | 30.1 | 27.8 | 28.9 |
| EchoSight (Yan and Xie 2024) | – | – | 31.3 |
| **MMhops-R1** | **33.8** | **32.6** | **33.2** |

Table 4: Comparison on INFOSEEK. Q: Question, E: Entity.

| Model | PaLI | PaLM | GPT-3 | **MMhops-R1** |
|---|---|---|---|---|
| Two hop | 14.7 | 22.8 | 18.7 | **23.3** |

Table 5: Comparison on E-VQA.

with key findings summarized as follows:

**1. Rich domain knowledge and strong reasoning capabilities are essential for solving multimodal multi-hop problems.** However, general-purpose open-source MLLMs are relatively weak in both aspects, posing challenges for them to generalize to this task. Even the 72B Qwen2.5-VL model falls short by 16.96% and 14.42% in overall accuracy on bridging and comparison questions, respectively, compared to our 7B-based model.

**2. Incorporating visual information is fundamental to effective multimodal reasoning.** Text-only multi-hop RAG methods are unable to access critical visual information, making it difficult to perform appropriate knowledge retrieval and accurate reasoning for multimodal multi-hop problems. Specifically, the state-of-the-art method Self-Ask, which significantly boosts base model performance (*e.g.*, GPT-4o) on textual multi-hop tasks, even shows an overall performance drop on comparison questions in MMhops compared to GPT-4o alone (30.42% vs. 36.62%).

**3. Accurate multi-turn reasoning and retrieval interactions are critical to successfully solving multimodal multi-hop problems.** Existing multimodal RAG methods, such as those designed for KB-VQA, are tailored to single-hop tasks and lack the ability to properly decompose multimodal multi-hop questions into sequential reasoning and retrieval steps, thereby limiting their answer accuracy. Even with the support of GPT-4o's OmniSearch, the overall accuracy on bridging and comparison questions remains 9.7% and 4.99% lower than ours.

**4. Closed-sourced commercial MLLMs remain the performance ceiling but still fall short of real-world applicability.** Gemini-2.5-Pro, which has likely undergone reasoning-specific optimization and large-scale pretraining, outperforms our method but answers only about half of the bridging questions correctly, with lower accuracy on comparison questions. This underscores that multimodal multi-hop RAG remains largely unexplored.

## Cross-dataset Generalization Verification

To verify the generalizability of the proposed method, we evaluate it on two widely used knowledge-based VQA datasets: INFOSEEK (Chen et al. 2023) and E-VQA (Mensink et al. 2023). Results on INFOSEEK show the effectiveness of MMhops-R1 on multimodal single-hop questions, while its performance on two-hop questions in E-VQA confirms its generalization ability to multi-hop reasoning.

| **Method** | **Bridging** | | | | **Comparison** |
|---|---|---|---|---|---|
| | **String** | **Numerical** | **Time** | **Overall** | |
| **MMhops-R1** | **44.66** | **55.33** | **47.94** | **51.35** | **22.01** |
| w/o $R_{\text{action}}$ | 39.74 | 51.19 | 46.8 | 47.57 | 20.62 |
| w/o $R_{\text{format}}$ | 43.12 | 53.64 | 47.73 | 49.97 | 14.42 |
| w/o $R_{\text{format}}$,$R_{\text{action}}$ | 40.43 | 42.68 | 40.62 | 41.75 | 13.03 |

Table 6: Effect of $R_{\text{outcome}}$, $R_{\text{format}}$ and $R_{\text{action}}$.

| **Method** | **Bridging** | | | | **Comparison** |
|---|---|---|---|---|---|
| | **String** | **Numerical** | **Time** | **Overall** | |
| 5 | 44.20 | **56.13** | 49.18 | **51.92** | 20.09 |
| 4 | **44.66** | 55.33 | **47.94** | **51.35** | **22.01** |
| 3 | 40.05 | 51.79 | 46.70 | 47.97 | 13.78 |
| 2 | 30.75 | 46.98 | 30.13 | 39.93 | 9.83 |

Table 7: Effect of the maximum retriever interaction count.

## Ablation Studies

**Effect of $R_{\text{outcome}}$,$R_{\text{format}}$ and $R_{\text{action}}$.** As shown in Table 6: (1) Removing either the retrieval reward $R_{\text{action}}$ or the format reward $R_{\text{format}}$ leads to a notable performance drop, particularly on comparison questions; (2) Removing both $R_{\text{action}}$ and $R_{\text{format}}$ results in an even greater decline. These findings indicate that encouraging appropriate retrieval, enforcing correct feedback formats, and imposing strong constraints on answer precision all contribute positively to model performance.

**Effect of Number of Interaction Rounds.** To demonstrate that the MMhops dataset indeed requires multi-step reasoning and RAG interaction for problem solving, we report model performance in Table 7 under maximum rounds constrained to 2, 3, 4, and 5. As the number of rounds increases from 2 to 4, overall performance consistently improves, while further increasing to 5 yields no significant gains but introduces more computational overhead. Therefore, four-step reasoning is most suitable for MMhops.

## Conclusion

In this work, we introduce the first large-scale multimodal multi-hop reasoning dataset MMhops to evaluate models' capabilities in multi-turn interactive reasoning and external knowledge utilization, and extensive experiments show that existing MLLMs struggle on MMhops. To address this, we further propose a novel reinforcement learning-based framework MMhops-R1 for multimodal reasoning and RAG interaction. Results demonstrate that MMhops-R1 substantially outperforms existing methods by effectively leveraging reasoning and retrieval capabilities. The code, dataset, and model weights will be open-sourced to encourage future research on the multimodal multi-hop reasoning task.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. U24A20331, No. 62302501), the Beijing Natural Science Foundation (No. L251005, No. L243015) and the Key Research and Development Program of Xinjiang Uyghur Autonomous Region (No. 2023B01005).

## References

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

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

Chen, L.; Li, L.; Zhao, H.; and Song, Y. 2025a. Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than $3.

Chen, M.; Li, T.; Sun, H.; Zhou, Y.; Zhu, C.; Wang, H.; Pan, J. Z.; Zhang, W.; Chen, H.; Yang, F.; et al. 2025b. Learning to reason with search for llms via reinforcement learning. *arXiv preprint arXiv:2503.19470*.

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

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. 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*, 6904–6913.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6700–6709.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Jin, B.; Zeng, H.; Yue, Z.; Yoon, J.; Arik, S.; Wang, D.; Zamani, H.; and Han, J. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Lee, C.; Roy, R.; Xu, M.; Raiman, J.; Shoeybi, M.; Catanzaro, B.; and Ping, W. 2024. NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models. *arXiv preprint arXiv:2405.17428*.

Leng, S.; Wang, J.; Li, J.; Zhang, H.; Hu, Z.; Zhang, B.; Zhang, H.; Jiang, Y.; Li, X.; Zhao, D.; et al. 2025. Mmr1: Advancing the frontiers of multimodal reasoning.

Lerner, P.; Ferret, O.; Guinaudeau, C.; Le Borgne, H.; Besançon, R.; Moreno, J. G.; and Lovón Melgarejo, J. 2022. ViQuAE, a dataset for knowledge-based visual question answering about named entities. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3108–3120.

Li, Y.; Li, Y.; Wang, X.; Jiang, Y.; Zhang, Z.; Zheng, X.; Wang, H.; Zheng, H.-T.; Yu, P. S.; Huang, F.; et al. 2024. Benchmarking multimodal retrieval augmented generation with dynamic vqa dataset and self-adaptive planning agent. *arXiv preprint arXiv:2411.02937*.

Liu, Z.; Sun, Z.; Zang, Y.; Dong, X.; Cao, Y.; Duan, H.; Lin, D.; and Wang, J. 2025. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*.

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

Meng, F.; Du, L.; Liu, Z.; Zhou, Z.; Lu, Q.; Fu, D.; Shi, B.; Wang, W.; He, J.; Zhang, K.; et al. 2025. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *CoRR*.

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

Peng, Y.; Wang, P.; Wang, X.; Wei, Y.; Pei, J.; Qiu, W.; Jian, A.; Hao, Y.; Pan, J.; Xie, T.; et al. 2025. Skywork r1v: Pioneering multimodal reasoning with chain-of-thought. *arXiv preprint arXiv:2504.05599*.

Press, O.; Zhang, M.; Min, S.; Schmidt, L.; Smith, N. A.; and Lewis, M. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

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

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

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

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Sheng, G.; Zhang, C.; Ye, Z.; Wu, X.; Zhang, W.; Zhang, R.; Peng, Y.; Lin, H.; and Wu, C. 2024. HybridFlow: A Flexible and Efficient RLHF Framework. *arXiv preprint arXiv: 2409.19256*.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8317–8326.

Team, K.; Du, A.; Gao, B.; Xing, B.; Jiang, C.; Chen, C.; Li, C.; Xiao, C.; Du, C.; Liao, C.; et al. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Wang, L.; Yang, N.; Huang, X.; Jiao, B.; Yang, L.; Jiang, D.; Majumder, R.; and Wei, F. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Xu, G.; Jin, P.; Wu, Z.; Li, H.; Song, Y.; Sun, L.; and Yuan, L. 2024. Llava-cot: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*.

Yan, Y.; and Xie, W. 2024. EchoSight: Advancing visual-language models with Wiki knowledge. *arXiv preprint arXiv:2407.12735*.

Yu, Q.; Zhang, Z.; Zhu, R.; Yuan, Y.; Zuo, X.; Yue, Y.; Dai, W.; Fan, T.; Liu, G.; Liu, L.; et al. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Zhang, J.; Huang, J.; Yao, H.; Liu, S.; Zhang, X.; Lu, S.; and Tao, D. 2025. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*.

Zhang, T.; Zhang, Z.; Ma, Z.; Chen, Y.; Qi, Z.; Yuan, C.; Li, B.; Pu, J.; Zhao, Y.; Xie, Z.; et al. 2024. mR$^2$AG: Multimodal Retrieval-Reflection-Augmented Generation for Knowledge-Based VQA. *arXiv preprint arXiv:2411.15041*.

Zheng, Z.; Yang, M.; Hong, J.; Zhao, C.; Xu, G.; Yang, L.; Shen, C.; and Yu, X. 2025. DeepEyes: Incentivizing" Thinking with Images" via Reinforcement Learning. *arXiv preprint arXiv:2505.14362*.

Zhou, H.; Li, X.; Wang, R.; Cheng, M.; Zhou, T.; and Hsieh, C.-J. 2025. R1-Zero's" Aha Moment" in Visual Reasoning on a 2B Non-SFT Model. *arXiv preprint arXiv:2503.05132*.