V-Retrver: Evidence-Driven Agentic Reasoning for Universal Multimodal Retrieval

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) have recently been applied to universal multimodal retrieval, where Chain-of-Thought (CoT) reasoning improves candidate reranking. However, existing approaches remain largely language-driven, relying on static visual encodings and lacking the ability to actively verify fine-grained visual evidence, which often leads to speculative reasoning in visually ambiguous cases. We propose V-Retrver, an evidence-driven retrieval framework that reformulates multimodal retrieval as an agentic reasoning process grounded in visual inspection. V-Retrver enables an MLLM to selectively acquire visual evidence during reasoning via external visual tools, performing a multimodal interleaved reasoning process that alternates between hypothesis generation and targeted visual verification. To train such an evidence-gathering retrieval agent, we adopt a curriculum-based learning strategy combining supervised reasoning activation, rejection-based refinement, and reinforcement learning with an evidence-aligned objective. Experiments across multiple multimodal retrieval benchmarks demonstrate consistent improvements in retrieval accuracy (with 23.0% improvements on average), perception-driven reasoning reliability, and generalization.

1 Introduction

The rapid development of Multimodal Large Language Models (MLLMs) has substantially advanced universal multimodal retrieval (Chen et al., 2024a; Lin et al., 2024a; Wang et al., 2024b; Zhu et al., 2025d; Sun et al.), enabling a single model to support diverse retrieval scenarios such as text-to-image, image-to-text, and interleaved multimodal queries. Recent works further demonstrate that incorporating Chain-of-Thought (CoT) reasoning can improve retrieval performance by enhancing interpretability and candidate discrimination (Zhu et al., 2025d; Xu et al., 2025b; Narayan et al., 2025). However, despite these advances, existing CoT-based retrieval systems remain fundamentally language-driven, even when retrieval decisions critically depend on visual evidence.

This limitation becomes particularly pronounced in visually ambiguous retrieval scenarios, where candidate images share similar semantic content but differ in fine-grained visual attributes such as object appearance, style, or local context. Most current MLLM-based retrieval methods (Liu et al., 2025; Chen et al., 2024a; Lin et al., 2024a) compress visual inputs into fixed embeddings or textual descriptions, forcing the reasoning process to rely on language alone to infer visual differences. Consequently, the model often produces speculative or hallucinated reasoning when the required evidence lies in the visual modality. Even recent reasoning-enhanced retrieval frameworks, such as Retrv-R1 (Zhu et al., 2025d) and MM-R5 (Xu et al., 2025b), improve textual reasoning depth but still rely on single-pass visual encoding, lacking the ability to actively verify visual hypotheses during reasoning.

To overcome this gap, we propose **V-Retrver**, an evidence-driven retrieval framework that reformulates multimodal retrieval as an agentic reasoning process grounded in visual inspection. Instead of treating visual representations as static inputs, V-Retrver enables an MLLM to selectively acquire visual evidence during reasoning by invoking external visual tools. Through a multimodal interleaved Chain-of-Thought process, the model alternates between hypothesis generation and targeted visual verification, allowing it to dynamically resolve visual ambiguities and progressively refine ranking decisions, as illustrated in Fig. 1.

Training such an evidence-gathering retrieval agent requires not only strong reasoning ability but also effective alignment between retrieval performance and visual tool usage. We therefore adopt

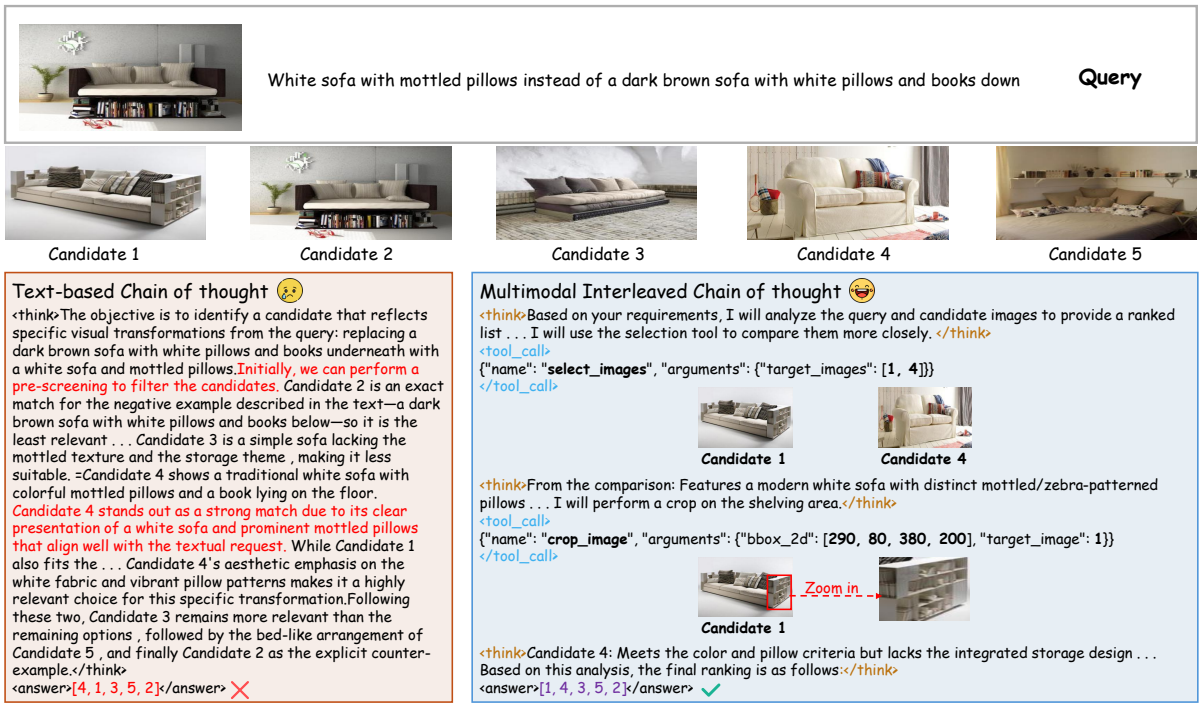


Figure 1: Comparison between text-based CoT (left) and multimodal interleaved CoT (right) for multimodal retrieval. Text-based CoT relies on language-driven inference over static visual representations, often failing to resolve fine-grained differences. In contrast, V-Retrver performs multimodal interleaved CoT reasoning by invoking visual tools to inspect candidate images, enabling grounded reasoning and more reliable ranking decisions.

a curriculum-based training strategy consisting of three stages. First, a cold-start supervised stage initializes the model with basic reasoning capabilities and operation formatting using synthesized high-quality CoT data. Second, rejection sampling fine-tuning consolidates high-quality reasoning trajectories and improves structural compliance. Finally, we introduce **Evidence-Aligned Policy Optimization (EAPO)**, instantiated via Group Relative Policy Optimization (GRPO) (Guo et al., 2025), which reinforces correct ranking decisions while encouraging informative visual verification and discouraging redundant tool usage.

Extensive experiments on the universal multimodal retrieval benchmark M-BEIR, as well as multiple out-of-domain datasets, demonstrate that V-Retrver consistently outperforms strong baselines across diverse retrieval settings. The results show that V-Retrver achieves higher retrieval accuracy, more reliable perception-grounded reasoning, and stronger generalization ability, validating the effectiveness of interleaved visual reasoning for multimodal retrieval. In summary, our contributions are three-fold:

- We propose V-Retrver, an evidence-driven agentic retrieval framework that enables MLLMs to actively acquire visual evidence during multimodal reasoning.
- We introduce a curriculum-based training strategy with an evidence-aligned reinforcement learning objective that jointly improves reasoning quality, ranking accuracy, and efficient visual tool usage.
- Extensive experiments across multiple benchmarks demonstrate that V-Retrver consistently outperforms existing methods and generalizes well to diverse multimodal retrieval scenarios.

2 Related Work

Multi-modal Large Language Models. In recent years, the rapid advancement of MLLMs has driven the deep integration of visual perception and language reasoning, leading to the emergence of high-performing open-source models, notably the LLaVA (Liu et al., 2024; Guo et al., 2024; Zhang et al., 2025b; Lin et al., 2023a; Li et al., 2023a), Qwen-VL (Bai et al., 2023; Wang et al., 2024a; Yang et al., 2024), and InternVL (Chen et al., 2024c; Gao et al., 2024; Lu et al., 2025) series. In parallel, large-scale models such as Flamingo (Alayrac et al., 2022), mPLUG-Owl (Ye et al., 2023, 2024b,a), and GPT-4V (Yang et al., 2023) pursue a holistic vision-language modeling

paradigm, incorporating advanced mechanisms including mixture-of-experts (Shu et al., 2024; Li et al., 2025b) and image generation (Xie et al., 2024; Xu et al., 2025a). However, these models generally lack Chain-of-Thought reasoning and test-time scalability (Muennighoff et al., 2025; Zhang et al., 2025a; Chen et al., 2024b), and still largely decouple visual perception from language reasoning.

Multimodal Retrieval. Recent advances in deep learning (Zhu et al., 2021, 2024, 2025a,c,b; Ji et al., 2024) have driven progress across retrieval tasks, including text–image retrieval (Pham et al., 2024; Fu et al., 2024; Zhang et al., 2020; Chun et al., 2021; Kim et al., 2023b,a), composed image retrieval (Baldrati et al., 2022; Saito et al., 2023; Gu et al., 2024; Suo et al., 2024; Baldrati et al., 2023), multimodal document retrieval (Chen et al., 2023; Hu et al., 2023; Liu et al., 2023), and instruction-based retrieval (Wu et al., 2021; Zhang et al., 2024a; Asai et al., 2023). Among these, CLIP (Radford et al., 2021) has demonstrated strong effectiveness and scalability in multimodal retrieval (Baldrati et al., 2022; Wei et al., 2024; Sain et al., 2023; Pei et al., 2023; Jin et al., 2024). For instance, Kim et al. (2023a) improve CLIP via prompt tuning, enabling enhanced generalization across diverse retrieval settings. More recently, MLLMs have been introduced to advance retrieval (Liu et al., 2025; Jiang et al., 2024; Lin et al., 2024a; Zhou et al., 2024): some use MLLM embeddings for similarity-based retrieval (Zhou et al., 2024; Lan et al., 2025; Lin et al., 2024a; Zhang et al., 2024b; Jian et al., 2025; Gu et al., 2025), while others employ MLLMs as reranking agents (Liu et al., 2025; Li et al., 2025a). Retrv-R1 (Zhu et al., 2025d) further equips the model with text reasoning via reinforcement learning. In contrast, V-Retrver introduces an evidence-driven framework that adaptively invokes visual tools during reasoning, enabling a more flexible and effective process with significant retrieval improvements.

3 Method

3.1 Problem Formulation

We study the problem of *universal multimodal retrieval*. Given a query q of arbitrary modality (text, image, or interleaved multimodal input) and a candidate pool $\Omega = \{c_n\}_{n=1}^{N}$, the objective is to identify the most relevant candidate $\hat{c} \in \Omega$. Conventional multimodal retrieval approaches typically formulate this problem as static similarity matching or language-only reranking over fixed visual representations. Such formulations implicitly assume that all necessary visual evidence has been fully encoded into embeddings or textual descriptions *prior* to reasoning. However, this assumption breaks down in fine-grained or visually ambiguous retrieval scenarios, where subtle local details determine relevance and cannot be reliably inferred from compressed representations alone.

To address this limitation, we reformulate multimodal retrieval as an *evidence-grounded reasoning problem*. Under this formulation, retrieval is no longer a single-pass inference process, but an iterative decision-making procedure in which the model is required to actively acquire and verify visual evidence during ranking. Specifically, the retrieval process consists of three tightly coupled steps: (i) generating hypotheses about candidate relevance based on available information, (ii) selectively inspecting visual evidence to resolve uncertainty, and (iii) refining the ranking decision based on verified observations. This perspective naturally gives rise to an *agentic reranking* paradigm, where a retrieval model is endowed with the ability to reason, inspect, and revise its decisions, rather than passively scoring candidates using fixed representations.

3.2 Overview of V-Retrver

Building on the above formulation, we propose **V-Retrver**, an evidence-driven reasoning framework for universal multimodal retrieval, As illustrated in Fig. 2. V-Retrver follows a coarse-to-fine retrieval pipeline that decouples efficient candidate proposal from computationally intensive evidence-based reasoning. In the first stage, an embedding model ϕ encodes the query q and each candidate c_n into a shared representation space, retrieving the top-K candidates based on similarity. We adopt the same method as LamRA (Liu et al., 2025) for constructing the embedding model ϕ. This stage serves as an efficient candidate proposal mechanism and substantially reduces the search space:

$$\mathcal{C} = \{c_k\}_{k=1}^{K}, \quad K \ll N.$$

In the second stage, V-Retrver employs a reasoning agent θ to perform fine-grained reranking over the reduced candidate set $\mathcal{C}$. Crucially, θ is not a conventional reranker that operates over static features. Instead, it is designed as an *agentic evidence-gathering model* that can iteratively reason, invoke

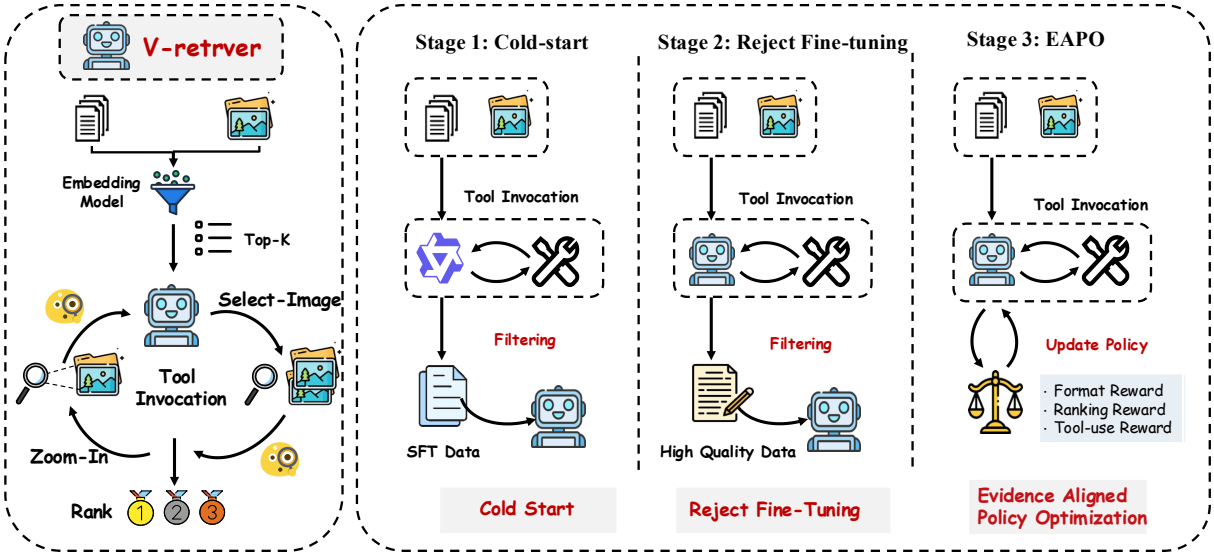


Figure 2: Overview of the V-Retrver framework. The left panel illustrates the inference pipeline, featuring a coarse-to-fine process with embedding-based retrieval and agentic reranking. The right panel details the three training stages we proposed, including Cold Start, Rejection sampling Fine-Tuning, and EAPO.

visual inspection tools, and revise its ranking decisions based on newly acquired visual observations. The final prediction is produced as:

$$\hat{c} = \theta(q, \mathcal{C}).$$

The remainder of this section details the core mechanisms that enable evidence-driven reasoning in V-Retrver, including multimodal interleaved reasoning, visual tools, and a curriculum-based training strategy.

3.3 Multimodal Interleaved Evidence Reasoning

We introduce **Multimodal Interleaved Evidence Reasoning (MIER)**, a reasoning paradigm that tightly interleaves textual hypothesis generation with targeted visual evidence acquisition. Unlike language-only Chain-of-Thought reasoning, MIER allows intermediate reasoning steps to be explicitly grounded in visual observations obtained on demand. Formally, given an initial textual query T_0 and a candidate image set I_0, the reasoning agent iteratively produces outputs:

$$O_k = f_{\text{MLLM}}\big(\{T_i, C_i, V_i\}_{i=0}^{k}\big),$$

where T_i denotes a textual reasoning step, C_i denotes a tool invocation request, and V_i represents the visual evidence returned by the tool. A parser then determines whether to extract the next reasoning step and tool request (T_{k+1}, C_{k+1}), or to terminate the process and output a final ranking.

If a tool is invoked, the corresponding visual tool is executed and returns new visual evidence V_{k+1}, which is appended to the reasoning context. This process yields a multimodal reasoning trajectory:

$$\tau = \{T_1, C_1, V_1, T_2, C_2, V_2, \ldots, T_n, A_n\},$$

where A_n denotes the final ranked list of candidates. By explicitly grounding intermediate reasoning steps in dynamically acquired visual evidence, MIER mitigates speculative inference and hallucination, enabling more reliable ranking decisions in visually ambiguous cases.

3.4 Visual Tools

To support MIER, we equip the reasoning agent with a set of **Visual Tools**, which serve as external perceptual interfaces for selective visual inspection. These tools allow the model to control *what* to observe and *where* to focus during reasoning. Specifically, we implement two tools:

(1) SELECT-IMAGE, which enables the agent to select a subset of candidate images for closer inspection when multiple candidates exhibit high semantic similarity.

(2) ZOOM-IN, which performs localized zoom-in operations on specified regions of an image, allowing fine-grained analysis of discriminative visual attributes such as objects, textures, or spatial configurations.

These tools facilitate *selective perception* during retrieval. Rather than encoding all visual information upfront, the agent dynamically expands its

visual receptive field only when necessary, closely mirroring human retrieval behavior in which ambiguous candidates are resolved by "looking again" at critical details.

3.5 Training V-Retrver via Curriculum-Based Agentic Learning

Training V-Retrver requires transforming a general-purpose MLLM into an agent capable of stable, evidence-driven reasoning and strategic tool usage. To this end, we design a three-stage curriculum that progressively builds reasoning structure, reliability, and decision-making optimality.

Stage I: Reasoning Activation via Supervised Fine-Tuning. We begin with a cold-start supervised fine-tuning stage to activate basic reasoning and tool-use behaviors. Since existing retrieval datasets lack annotated reasoning trajectories, we synthesize multimodal Chain-of-Thought data using Qwen2.5-VL-72B-Instruct. These trajectories include structured reasoning steps and valid tool invocation patterns. After applying rule-based filtering to remove logically inconsistent or malformed samples, the base model is fine-tuned using standard SFT loss. This stage establishes foundational reasoning syntax and tool awareness, but does not yet guarantee robustness or optimal tool-use strategies.

Stage II: Rejection Fine-Tuning for Reasoning Reliability. Although Stage I activates tool-use behavior, the resulting policy exhibits high variance and produces a large fraction of low-quality trajectories. To improve reasoning reliability, we perform Rejection Sampling Fine-Tuning (RSFT). For each training instance, we sample multiple reasoning trajectories and retain only those that strictly satisfy formatting constraints and yield correct retrieval rankings. Fine-tuning on this filtered dataset significantly improves logical consistency and format compliance, providing a stable initialization for reinforcement learning.

Stage III: Evidence-Aligned Policy Optimization. While the previous stages activate structured reasoning and improve trajectory reliability, they do not explicitly optimize *how* visual evidence should be acquired during retrieval. In practice, the model may either underutilize visual inspection or invoke tools redundantly without contributing to better ranking decisions. To address this limitation, we introduce **Evidence-Aligned Policy Optimization (EAPO)**, a reinforcement learning objective that explicitly aligns retrieval performance with effective and economical visual verification behavior.

EAPO formulates multimodal retrieval as a trajectory-level decision-making problem, where each reasoning trajectory o_i is evaluated based on both ranking quality and evidence utilization. Specifically, we define a composite reward:

$$R_i = \alpha r_{\text{format}}(o_i) + \beta r_{\text{rank}}(o_i) + r_{\text{tool}}(o_i), \quad (1)$$

where the three components respectively encourage structural correctness, accurate ranking, and informative visual inspection. Below, we detail each reward term.

Format Compliance Reward. The format compliance reward r_{format} ensures that the model adheres to the required reasoning and output protocols, which is essential for stable policy optimization with structured multimodal outputs. Let Ω_{tag} denote the set of trajectories whose outputs are correctly enclosed by predefined `<think>` and `<answer>` tags, and let Ω_{list} denote the set of trajectories whose final answers strictly follow the required integer ranking list format. We define:

$$r_{\text{format}}(o_i) = \frac{1}{2}\,\mathbb{I}_{\{o_i \in \Omega_{\text{tag}}\}} + \frac{1}{2}\,\mathbb{I}_{\{o_i \in \Omega_{\text{list}}\}}, \quad (2)$$

where $\mathbb{I}_{\{\cdot\}}$ is the indicator function. This term primarily serves as a stabilizing signal, preventing malformed trajectories from dominating policy updates.

Contrastive Ranking Reward. A key challenge in training retrieval agents with reinforcement learning is the sparsity of ranking supervision. Binary correctness signals provide insufficient gradient information for list-wise optimization, while standard positional penalties ignore the relative ordering between the correct candidate and confusable hard negatives.

To address this, we propose a **Contrastive NDCG** (CNDCG) reward that provides dense and discriminative feedback by explicitly modeling the relationship between the correct candidate j^* and hard negatives $\mathcal{H}$, defined as the top-m non-correct candidates with the highest embedding similarity to the query (formal definition in Appendix H). We assign graded relevance scores to each position k in the predicted ranking $\hat{\pi}$:

$$\text{rel}_{\hat{\pi}(k)} = \begin{cases} 2 & \hat{\pi}(k) = j^* \\ 1 & \hat{\pi}(k) \in \mathcal{H} \text{ and ranked after } j^* \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Crucially, a hard negative receives $\mathrm{rel} = 1$ *only* when ranked below j^*; ranking above j^* yields zero credit, explicitly penalizing the most harmful errors. The Discounted Cumulative Gain is computed over the top-K_r positions as:

$$\mathrm{DCG}(\hat{\pi}) = \sum_{k=1}^{K_r} \frac{2^{\mathrm{rel}_{\hat{\pi}(k)}} - 1}{\log_2(k+1)}, \tag{4}$$

and the contrastive ranking reward is defined as:

$$r_{\mathrm{rank}}(o_i) = \frac{\mathrm{DCG}(\hat{\pi})}{\mathrm{IDCG}}, \tag{5}$$

where IDCG is the ideal-ranking DCG with j^* placed first (see Appendix H). This formulation encourages the agent to (i) rank j^* as high as possible, (ii) suppress hard negatives below j^* for dense gradient feedback, and (iii) avoid placing any hard negative above j^*, which incurs zero reward.

Tool-Use Reward. The tool-use reward r_{tool} directly governs the agent's evidence acquisition behavior, encouraging visual inspection only when it contributes to correct decisions while discouraging redundant or excessive tool usage. Let N_{tool} denote the number of valid visual tool invocations in trajectory o_i, and let k be the rank position of the correct candidate. We define:

$$\begin{aligned} r_{\mathrm{tool}}(o_i) = \eta \cdot \mathbb{I}_{\{k=1\}} \cdot \mathbb{I}_{\{N_{\mathrm{tool}}>0\}} \\ - \rho \cdot \max(0, N_{\mathrm{tool}} - \tau), \end{aligned} \tag{6}$$

where η incentivizes successful evidence-based verification, ρ penalizes excessive tool invocations, and τ specifies a tolerance threshold. This design explicitly encodes the principle that *effective* tool usage, rather than frequent usage, should be rewarded.

Policy Optimization. We instantiate EAPO using Group Relative Policy Optimization (GRPO) (Guo et al., 2025). Given a group of G trajectories sampled for the same query, we compute normalized advantages:

$$A_i = \frac{R_i - \mathrm{mean}(R)}{\mathrm{std}(R)}. \tag{7}$$

Let $r_i(\theta) = \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\mathrm{old}}}(o_i|q)}$, the final optimization objective is:

$$\begin{aligned} \mathcal{J}_{\mathrm{EAPO}}(\theta) = \mathbb{E}_q \left[\frac{1}{G} \sum_{i=1}^{G} \mathrm{clip}(r_i(\theta),\, 1 \pm \epsilon)\, A_i \right] \\ - \lambda\, \mathbb{E}_q[\mathrm{KL}(\pi_\theta \,\|\, \pi_{\mathrm{ref}})] \end{aligned} \tag{8}$$

Through EAPO, the model learns not only *what* to rank, but also *how* and *when* to acquire visual evidence in order to support reliable and efficient retrieval decisions.

4 Experiments

4.1 Experimental Setup

Datasets and Metrics. We train V-Retrver on M-BEIR (Wei et al., 2024), which encompasses eight distinct retrieval tasks across 10 datasets comprising 1.1M training samples. We evaluate on the M-BEIR test set and five previously unseen datasets: CIRCO (Baldrati et al., 2023), GeneCIS (Vaze et al., 2023), Visual Storytelling (Huang et al., 2016), Visual Dialog (Das et al., 2017), and Multi-round FashionIQ (Yuan and Lam, 2021). We primarily use Recall@K as the evaluation metric and additionally report MAP@5 for CIRCO. Dataset statistics and evaluation settings are in Appendix I.

Baselines. We compare V-Retrver against four categories: (1) foundational VLMs (CLIP, SigLIP, BLIP, BLIP2); (2) fine-tuned universal retrievers (UniIR (Wei et al., 2024), MM-Embed (Lin et al., 2024a), LamRA (Liu et al., 2025), U-MARVEL (Li et al., 2025a)); (3) reasoning-enhanced models (Vision-R1 (Huang et al., 2025), VLM-R1 (Shen et al., 2025), Retrv-R1 (Zhu et al., 2025d)); and (4) agentic tool-use models (DeepEyes (Zheng et al., 2025)). Detailed baseline descriptions and evaluation protocols are in Appendix J.

Reranking Setup. Following the coarse-to-fine paradigm, the embedding model ϕ first retrieves the top-K candidates, over which the reasoning agent θ performs listwise reranking in a single inference pass. We set K=50 for M-BEIR evaluation (reranking from local pool) and K=10 for unseen datasets.

Implementation Details. Our model is initialized from Qwen2.5-VL-7B-Instruct (Bai et al., 2025). The vision encoder remains frozen throughout all training stages while the language model is fine-tuned. Complete training configurations and hyperparameter settings are in Appendix F.

4.2 Main Results

Performance on M-BEIR. As presented in Table 1, V-Retrver-7B establishes a new state-of-the-art on M-BEIR with an average Recall of 69.7%, surpassing U-MARVEL-7B by +4.9% (64.8%) and

Table 1: **Comparison with other methods on M-BEIR test set.** R@K refers to the Recall@K metric. q^t, q^i, c^t and c^i denote the text query, image query, text candidates and image candidates, respectively. Abbreviations: VN=VisualNews, F200K=Fashion200K, InfoS=InfoSeek, FIQ=FashionIQ. Best results in **bold**.

Models	$q^t \to c^i$			$q^t \to c^t$	$q^t \to (c^i, c^t)$		$q^i \to c^t$			$q^i \to c^i$	$(q^i, q^t) \to c^t$		$(q^i, q^t) \to c^i$		$(q^i, q^t) \to (c^i, c^t)$		Avg.
	VN	COCO	F200K	WebQA	EDIS	WebQA	VN	COCO	F200K	NIGHTS	OVEN	InfoS	FIQ	CIRR	OVEN	InfoS	
	R@5	R@5	R@10	R@5	R@5	R@5	R@5	R@5	R@10	R@5	R@5	R@5	R@10	R@5	R@5	R@5	
CLIP-L (Radford et al., 2021)	43.3	61.1	6.6	36.2	43.3	45.1	41.3	79.0	7.7	26.1	24.2	20.5	7.0	13.2	38.8	26.4	32.5
SigLIP (Zhai et al., 2023)	30.1	75.7	36.5	39.8	27.0	43.5	30.8	88.2	34.2	28.9	29.7	25.1	14.4	22.7	41.7	27.4	37.2
BLIP (Li et al., 2022)	16.4	74.4	15.9	44.9	26.8	20.3	17.2	83.2	19.9	27.4	16.1	10.2	2.3	10.6	27.4	16.6	26.8
BLIP2 (Li et al., 2023b)	16.7	63.8	14.0	38.6	26.9	24.5	15.0	80.0	14.2	25.4	12.2	5.5	4.4	11.8	27.3	15.8	24.8
UniIR-BLIP$_{FF}$ (Wei et al., 2024)	23.4	79.7	26.1	80.0	50.9	79.8	22.8	89.9	28.9	33.0	41.0	22.4	29.2	52.2	55.8	33.0	46.8
UniIR-CLIP$_{SF}$ (Wei et al., 2024)	42.6	81.1	18.0	84.7	59.4	78.7	43.1	92.3	18.3	32.0	45.5	27.9	24.4	44.6	67.6	48.9	50.6
Qwen2.5-VL-3B (Bai et al., 2025)	36.0	67.8	16.1	69.5	45.2	61.7	23.3	82.3	12.0	20.9	36.7	22.3	24.3	53.5	56.4	49.8	42.4
Qwen2.5-VL-7B (Bai et al., 2025)	40.2	71.9	20.3	71.9	49.4	64.5	29.3	84.6	19.4	25.5	42.4	32.1	25.0	55.1	60.8	54.9	46.7
Vision-R1-7B (Huang et al., 2025)	41.9	75.0	22.0	70.6	51.3	69.1	35.4	85.1	22.4	25.9	48.8	44.0	29.2	57.7	66.2	59.0	50.2
VLM-R1-7B (Shen et al., 2025)	40.5	77.2	22.5	72.3	50.0	67.9	36.2	86.3	20.9	26.4	48.8	37.5	29.9	57.4	64.0	62.3	50.0
DeepEyes-7B (Zheng et al., 2025)	43.5	73.2	23.5	75.4	55.2	70.1	38.6	86.5	24.1	28.3	50.4	46.2	31.5	59.2	68.4	63.2	52.3
MM-Embed-7B (Lin et al., 2024a)	41.0	71.3	17.1	95.9	68.8	85.0	41.3	90.1	18.4	32.4	42.1	42.3	25.7	50.0	64.1	57.7	52.7
LamRA-7B (Liu et al., 2025)	48.0	85.2	32.9	96.7	75.8	87.7	48.6	92.3	36.1	33.5	59.2	64.1	37.8	63.3	79.2	78.3	63.7
U-MARVEL-7B (Li et al., 2025a)	49.4	85.6	34.2	**98.5**	81.4	89.4	50.5	88.4	37.7	34.7	63.7	62.9	38.2	63.2	80.8	78.9	64.8
Retrv-R1-3B (Zhu et al., 2025d)	46.6	83.0	34.4	96.7	77.9	88.2	48.0	92.7	34.5	34.7	64.3	70.0	45.2	67.9	83.3	80.0	65.5
Retrv-R1-7B (Zhu et al., 2025d)	50.5	86.7	39.3	97.4	82.1	89.5	51.4	94.1	**39.8**	39.5	69.0	**75.6**	49.5	72.3	86.6	83.7	69.2
V-Retrver-7B	**51.8**	**87.5**	**40.3**	96.9	**82.9**	**90.2**	**52.2**	**94.8**	37.8	**39.8**	**69.8**	73.2	**51.2**	**73.5**	**87.8**	**85.0**	**69.7**

Table 2: **Experimental results on unseen datasets.** q^{dialog} and $(q^i \oplus q^t)$ refer to dialog queries and multi-interleaved image-text queries, respectively.

Models	$(q^i, q^t) \to c^i$		$q^{\text{dialog}} \to c^i$	$(q^i \oplus q^t) \to c^i$	
	CIRCO	GeneCIS	VisD	VIST	MT-FIQ
	MAP@5	R@1	R@1	R@1	R@5
CLIP-L (Radford et al., 2021)	4.0	13.3	23.7	0.6	17.7
UniIR-CLIP (Wei et al., 2024)	12.5	16.8	26.8	0.6	39.4
E5-V (Jiang et al., 2024)	24.8	18.5	54.6	10.0	19.2
MagicLens-L (Zhang et al., 2024a)	29.6	16.3	28.0	3.3	22.6
MM-Embed-7B (Lin et al., 2024a)	35.5	22.9	64.7	25.7	59.0
LamRA-7B (Liu et al., 2025)	42.8	24.8	70.9	28.6	63.9
V-Retrver-7B	**48.2**	**30.7**	**75.1**	**31.2**	**68.3**

Table 3: **Experimental results on held-out tasks.** * indicates training on remaining tasks only, w/o exposure to the three held-out tasks.

Models	$q^i \to c^i$	$(q^i, q^t) \to c^t$		$(q^i, q^t) \to (c^i, c^t)$		Avg.
	NIGHTS	OVEN	InfoS	OVEN	InfoS	
	R@5	R@5	R@5	R@5	R@5	
Supervised						
UniIR-BLIP$_{FF}$ (Wei et al., 2024)	33.0	41.0	22.4	55.8	33.0	37.0
UniIR-CLIP$_{SF}$ (Wei et al., 2024)	32.0	45.5	27.9	67.6	48.9	44.4
Zero-shot						
Qwen2.5-VL-7B (Bai et al., 2025)	20.3	38.5	40.4	53.6	44.9	39.5
Vision-R1-7B (Huang et al., 2025)	22.9	39.8	42.9	57.4	46.5	41.9
LamRA-7B* (Liu et al., 2025)	29.2	46.9	54.2	65.1	59.1	50.9
V-Retrver-7B*	**36.2**	**57.8**	**65.9**	**75.3**	**70.3**	**61.1**

Retrv-R1-7B by +0.5% (69.2%). Gains are especially pronounced on fine-grained visual tasks: V-Retrver achieves 51.2% on FIQ and 73.5% on CIRR, outperforming U-MARVEL-7B (38.2% and 63.2%) by large margins. While Retrv-R1-7B surpasses V-Retrver on certain text-centric subtasks (e.g., F200K $q^i \to c^t$, InfoS $(q^i, q^t) \to c^t$), V-Retrver achieves consistent advantages on visually ambiguous tasks that require fine-grained inspection.

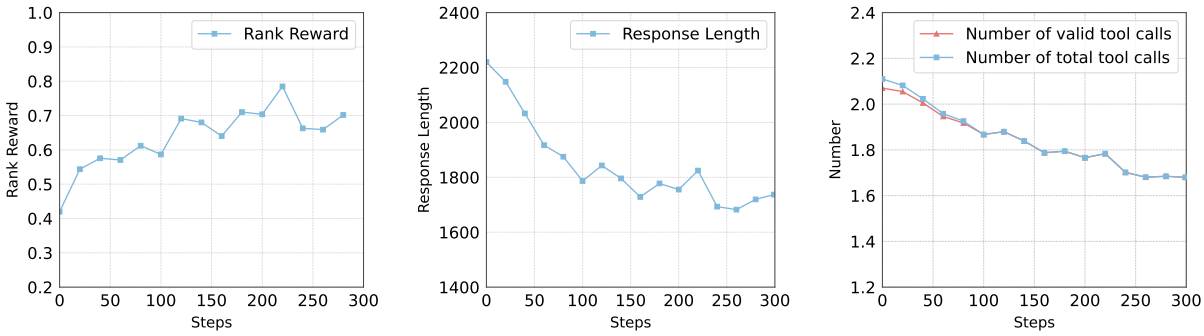
Generalization to Unseen Datasets. Table 2 confirms robustness on datasets not seen during training. V-Retrver consistently outperforms specialized models and generalist MLLMs, achieving MAP@5 of 48.2 on CIRCO (vs. MM-Embed 35.5 and LamRA 42.8) and R@1 of 30.7 on GeneCIS (vs. 24.8 for LamRA).

Robustness on Held-out Tasks. Even without exposure to three modality-combination types during training, V-Retrver* achieves 61.1% average Recall (Table 3), outperforming LamRA-7B* (50.9%) by +10.2%. This confirms that MIER effectively decouples evidence-gathering reasoning from specific input modalities.

4.3 Ablation Study

Impact of Training Stages. Table 5 ablates each training stage. Zero-shot tool invocation without any alignment collapses to 45.8% (below the 47.2% baseline), confirming alignment is critical. SFT alone reaches 59.4%; adding RL without RSFT gives 66.3%; SFT+RSFT without RL yields 60.9%. The full three-stage pipeline achieves 67.2%, demonstrating that each stage addresses specific shortcomings of the previous one.

Effectiveness of Visual Tool. Table 4 compares V-Retrver against a text-only RL variant trained identically but without visual tools. The text-only variant achieves 61.8% vs. V-Retrver's 67.2%, confirming that active zoom-in and image selection resolve fine-grained ambiguities that compressed static representations cannot capture.The baseline frequently struggles to distinguish between hard negatives that share identical global semantics but differ only in minute visual details.



(a) Rank Reward (b) Response Length (c) Tool Calls

Table 4: **Ablation: visual tool-use mechanism.** Multimodal interleaved CoT (with Visual Tool) vs. text-only baseline under identical RL training.

Variants	$q^t \to c^i$ COCO R@5	$q^i \to c^t$ F200K R@10	$(q^i, q^t) \to c^i$ CIRR R@5	$(q^i, q^t) \to c^t$ OVEN R@5	Avg.
Qwen2.5-VL-7B (Bai et al., 2025)	71.9	19.4	55.1	42.4	47.2
RL w/o tool	84.1	33.2	66.5	63.2	61.8
V-Retrver-7B	**87.5**	**37.8**	**73.5**	**69.8**	**67.2**

Table 5: **Ablation: training stages.** Impact of Cold Start (SFT), Rejection Sampling Fine-Tuning (RSFT), and Reinforcement Learning (RL).

Training Stage	$q^t \to c^i$ COCO R@5	$q^i \to c^t$ F200K R@10	$(q^i, q^t) \to c^i$ CIRR R@5	$(q^i, q^t) \to c^t$ OVEN R@5	Avg.
Qwen2.5-VL-7B (Bai et al., 2025)	71.9	19.4	55.1	42.4	47.2
w/o SFT & RSFT & RL	71.5	18.1	53.4	40.2	45.8
w/o RSFT & RL	83.2	31.6	63.7	59.0	59.4
w/o RSFT	87.2	37.3	72.4	68.3	66.3
w/o RL	83.9	32.8	65.3	61.5	60.9
V-Retrver-7B	**87.5**	**37.8**	**73.5**	**69.8**	**67.2**

Figure 4: RL Training curves showing rank reward improvement, response length stabilization, and convergence of valid vs. total tool calls. Tables below show ablation results for visual tool mechanism (left) and training stages (right).

4.4 Analysis

RL Training Dynamics. Fig. 4 shows that rank reward improves steadily under EAPO. Initially, valid tool calls are slightly fewer than total invocations, reflecting occasional formatting errors from early SFT/RSFT training. As RL progresses, the two curves converge, confirming that policy optimization eliminates erroneous tool actions. Concurrently, response length and tool-call frequency decrease and stabilize, demonstrating that the model learns to invoke tools only when necessary, with an average of 1.6 to 1.7 calls per query.

Efficiency Analysis. V-Retrver employs a listwise ranking approach, preserving uncompressed visual representations while remaining competitive in latency. Against the pointwise baseline LamRA-Rank that invokes the LLM once per candidate (50 calls for K=50, ITR 4.41×), V-Retrver is substantially faster. Compared to Retrv-R1, which achieves lower ITR by compressing visual tokens, V-Retrver accepts a modest latency overhead in exchange for fine-grained visual fidelity and higher retrieval accuracy. Full inference time ratio (ITR) and GPU memory ratio (GMR) comparisons are in Appendix A.

5 Conclusion

In this paper, we presented V-Retrver, an evidence-driven MLLM framework tailored for universal multimodal retrieval. V-Retrver adopts multimodal interleaved Chain-of-Thought (CoT) reasoning, enabling the model to dynamically inspect and verify candidate images through visual tool invocation, thereby achieving more fine-grained ranking of candidate result lists. We adopt a three-stage training pipeline to multimodal interleaved CoT reasoning abilities. Extensive experimental results demonstrate that V-Retrver achieves significant improvements in both model effectiveness and task generalization. We regard V-Retrver to be an important step toward effectively introducing agentic MLLMs to enhance downstream multimodal tasks, laying a solid foundation for building general agentic MLLMs with advanced reasoning capabilities.In future work, we plan to extend V-Retrver by exploring fine-grained token-level advantage estimation. Integrating nuanced generalized advantage estimation across the entire multi-turn reasoning sequence could effectively resolve credit assignment challenges, ultimately driving further optimizations in real-time inference efficiency and broader tool-use scalability.

Limitations

Despite its strong performance, V-Retrver still has several limitations. First, the multi-turn tool invocation process introduces a credit assignment challenge: since EAPO evaluates trajectories at the sequence level, it is difficult to attribute the contribution of each individual tool call to the final ranking outcome, particularly in long reasoning chains. Fine-grained step-level credit assignment remains an open problem for agentic retrieval training. Second, the inference efficiency of V-Retrver remains a concern for large-scale deployment. The multi-turn agentic reasoning with interleaved tool invocations and uncompressed visual representations incurs non-negligible latency and memory overhead, limiting applicability in latency-sensitive scenarios with large candidate pools. Future work will explore step-level credit assignment mechanisms, lightweight adaptive inference strategies such as early-exit and dynamic tool scheduling, and extensions to broader downstream tasks such as multimodal recommendation and retrieval-augmented generation.

Ethics Statement

This work focuses on universal multimodal retrieval and does not involve the collection or annotation of human subjects data. All datasets used in our experiments are publicly available and were obtained in accordance with their respective licenses. The proposed framework, V-Retrver, is intended to improve the accuracy and reliability of multimodal retrieval systems. We do not foresee direct negative societal consequences arising from this research, though we acknowledge that any retrieval system may inherit biases present in its training data.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. 2023. Task-aware retrieval with instructions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3650–3675.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. 2023. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the International Conference on Computer Vision*.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2022. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21466–21474.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. Can pre-trained vision and language models answer visual information-seeking questions? In *Proceedings of the Conference on Empirical Methods in Natural Language Processinng*.

Zhanpeng Chen, Chengjin Xu, Yiyan Qi, and Jian Guo. 2024a. Mllm is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training. *arXiv preprint arXiv:2407.21439*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. 2021. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8415–8424.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zheren Fu, Lei Zhang, Hou Xia, and Zhendong Mao. 2024. Linguistic-aware patch slimming framework

for fine-grained cross-modal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26307–26316.

Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, and 1 others. 2024. Mini-internvl: a flexible-transfer pocket multi-modal model with 5% parameters and 90% performance. *Visual Intelligence*, 2(1):1–17.

Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. 2024. Language-only training of zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13225–13234.

Tiancheng Gu, Kaicheng Yang, Ziyong Feng, Xingjun Wang, Yanzhao Zhang, Dingkun Long, Yingda Chen, Weidong Cai, and Jiankang Deng. 2025. Breaking the modality barrier: Universal embedding learning with multimodal llms. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 2860–2869.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. 2024. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In *European Conference on Computer Vision*, pages 390–406. Springer.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. 2023. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proceedings of the International Conference on Computer Vision*.

Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, and 1 others. 2016. Visual storytelling. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.

Deyi Ji, Feng Zhao, Lanyun Zhu, Wenwei Jin, Hongtao Lu, and Jieping Ye. 2024. Discrete latent perspective learning for segmentation and detection. *arXiv preprint arXiv:2406.10475*.

Weijian Jian, Yajun Zhang, Dawei Liang, Chunyu Xie, Yixiao He, Dawei Leng, and Yuhui Yin. 2025. Rzenembed: Towards comprehensive multimodal retrieval. *arXiv preprint arXiv:2510.27350*.

Dongfu Jiang, Yi Lu, Zhuofeng Li, Zhiheng Lyu, Ping Nie, Haozhe Wang, Alex Su, Hui Chen, Kai Zou, Chao Du, and 1 others. 2025. Verltool: Towards holistic agentic reinforcement learning with tool use. *arXiv preprint arXiv:2509.01055*.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. 2024. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*.

Huilong Jin, Yingxue Zhang, Lei Shi, Shuang Zhang, Feifei Kou, Jiapeng Yang, Chuangying Zhu, and Jia Luo. 2024. An end-to-end graph attention network hashing for cross-modal retrieval. *Advances in Neural Information Processing Systems*, 37:2106–2126.

Dongwon Kim, Namyup Kim, and Suha Kwak. 2023a. Improving cross-modal retrieval with set of diverse embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23422–23431.

Jae Myung Kim, A Koepke, Cordelia Schmid, and Zeynep Akata. 2023b. Exposing and mitigating spurious correlations for cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2585–2595.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS Symposium on Operating Systems Principles*.

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. 2025. Llave: Large language and vision embedding models with hardness-weighted contrastive learning. *arXiv preprint arXiv:2503.04812*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning*.

Xiaojie Li, Chu Li, Shi-Zhe Chen, and Xi Chen. 2025a. U-marvel: Unveiling key factors for universal multimodal retrieval via embedding learning with mllms. *arXiv preprint arXiv:2507.14902*.

Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. 2025b. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023a. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2024a. Mm-embed: Universal multimodal retrieval with multimodal llms. *arXiv preprint arXiv:2411.02571*.

Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. 2023b. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. In *Advances in Neural Information Processing Systems*.

Weizhe Lin, Jingbiao Mei, Jinghong Chen, and Bill Byrne. 2024b. PreFLMR: Scaling up fine-grained late-interaction multi-modal retrievers. In *Association for Computational Linguistics*.

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, and 1 others. 2024. Llava-plus: Learning to use tools for creating multimodal agents. In *European Conference on Computer Vision*, pages 126–142. Springer.

Yikun Liu, Pingan Chen, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. 2025. Lamra: Large multimodal model as your advanced retrieval assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, and Ge Yu. 2023. Universal vision-language dense retrieval: Learning a unified representation space for multi-modal retrieval. In *The Eleventh International Conference on Learning Representations*.

Dongchen Lu, Yuyao Sun, Zilu Zhang, Leping Huang, Jianliang Zeng, Mao Shu, and Huo Cao. 2025. Internvl-x: Advancing and accelerating internvl series with efficient visual token compression. *arXiv preprint arXiv:2503.21307*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. 2023. Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. In *Proceedings of the International Conference on Computer Vision*.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

Kartik Narayan, Yang Xu, Tian Cao, Kavya Nerella, Vishal M Patel, Navid Shiee, Peter Grasch, Chao Jia, Yinfei Yang, and Zhe Gan. 2025. Deepmmsearch-r1: Empowering multimodal llms in multimodal web search. *arXiv preprint arXiv:2510.12801*.

Renjing Pei, Jianzhuang Liu, Weimian Li, Bin Shao, Songcen Xu, Peng Dai, Juwei Lu, and Youliang Yan. 2023. Clipping: Distilling clip-based models with a student base for video-language retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18983–18992.

Khoi Pham, Chuong Huynh, Ser-Nam Lim, and Abhinav Shrivastava. 2024. Composing object relations and attributes for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14354–14363.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*.

Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. 2023. Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2765–2775.

Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. 2023. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19305–19314.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, and 1 others. 2025. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*.

Fangxun Shu, Yue Liao, Le Zhuo, Chenning Xu, Lei Zhang, Guanghao Zhang, Haonan Shi, Long Chen, Tao Zhong, Wanggui He, and 1 others. 2024. Llava-mod: Making llava tiny via moe knowledge distillation. *arXiv preprint arXiv:2408.15881*.

Nan Sun, Jing Tang, Lei Sun, Rui Chen, Yuxing Lu, Xiangxiang Chu, and Hefei Ling. Reflection from retrieval: Mllm-guided iterative reasoning for zero-shot composed image retrieval.

Yucheng Suo, Fan Ma, Linchao Zhu, and Yi Yang. 2024. Knowledge-enhanced dual-stream zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26951–26962.

Sagar Vaze, Nicolas Carion, and Ishan Misra. 2023. Genecis: A benchmark for general conditional image similarity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yabing Wang, Le Wang, Qiang Zhou, Zhibin Wang, Hao Li, Gang Hua, and Wei Tang. 2024b. Multimodal llm enhanced cross-lingual cross-modal retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8296–8305.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2024. Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, pages 387–404. Springer.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2021. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. 2024. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*.

Chenkai Xu, Xu Wang, Zhenyi Liao, Yishun Li, Tianqi Hou, and Zhijie Deng. 2025a. Show-o turbo: Towards accelerated unified multimodal understanding and generation. *arXiv preprint arXiv:2502.05415*.

Mingjun Xu, Jinhan Dong, Jue Hou, Zehui Wang, Sihang Li, Zhifeng Gao, Renxin Zhong, and Hengxing Cai. 2025b. MM-R5: Multimodal reasoning-enhanced reranker via reinforcement learning for document retrieval. *arXiv preprint arXiv:2506.12364*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024a. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, and 1 others. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024b. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Yifei Yuan and Wai Lam. 2021. Conversational fashion image retrieval via multiturn natural language feedback. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the International Conference on Computer Vision*.

Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. 2024a. Magiclens: self-supervised image retrieval with open-ended instructions. In *Proceedings of the 41st International Conference on Machine Learning*, pages 59403–59420.

Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z Li. 2020. Context-aware attention network for image-text retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3536–3545.

Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Zhihan Guo, Yufei Wang, Irwin King, Xue Liu, and Chen Ma. 2025a. What, how, where, and how well? a survey on test-time scaling in large language models. *arXiv preprint arXiv:2503.24235*.

Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. 2025b. Llava-mini: Efficient image and video large multimodal models with one vision token. *arXiv preprint arXiv:2501.03895*.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2024b. Gme: Improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. 2025. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*.

Junjie Zhou, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, Defu Lian, and Yongping Xiong. 2024. Megapairs: Massive data synthesis for universal multimodal retrieval. *arXiv preprint arXiv:2412.14475*.

Lanyun Zhu, Tianrun Chen, Deyi Ji, Peng Xu, Jieping Ye, and Jun Liu. 2025a. Llafs++: Few-shot image segmentation with large language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Lanyun Zhu, Tianrun Chen, Deyi Ji, Jieping Ye, and Jun Liu. 2024. Llafs: When large language models meet few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3065–3075.

Lanyun Zhu, Tianrun Chen, Deyi Ji, Jieping Ye, and Jun Liu. 2025b. Not every patch is needed: Towards a more efficient and effective backbone for video-based person re-identification. *IEEE Transactions on Image Processing*.

Lanyun Zhu, Tianrun Chen, Jianxiong Yin, Simon See, De Wen Soh, and Jun Liu. 2025c. Replay master: Automatic sample selection and effective memory utilization for continual semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Haiyang Wu, and Shiqi Wang. 2025d. Retrv-r1: A reasoning-driven mllm framework for universal and efficient multimodal retrieval. *arXiv preprint arXiv:2510.02745*.

Lanyun Zhu, Deyi Ji, Shiping Zhu, Weihao Gan, Wei Wu, and Junjie Yan. 2021. Learning statistical texture for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12537–12546.

A Efficiency Analysis

We evaluate V-Retrver's inference efficiency against strong baselines using two metrics: **ITR** (Inference Time Ratio, relative to V-Retrver) and **GMR** (GPU Memory Ratio, relative to V-Retrver). Results under two candidate pool sizes, K=20 and K=50, are reported in Table 6, together with the representative CIRR R@5 metric to cover compositional retrieval settings.

Comparison with pointwise methods. LamRA-Rank-Pointwise invokes the LLM *independently for each candidate*: processing K candidates requires K separate MLLM calls, each containing only a single candidate. This pointwise paradigm leads to an ITR of $4.41\times$ at K=50, as inference latency scales linearly with the candidate pool size. Moreover, because each call processes only one candidate in isolation, the model lacks a global receptive field and cannot directly contrast similar candidates, which is a critical weakness for distinguishing visually similar hard negatives.

Comparison with token-compression methods. Retrv-R1 achieves lower ITR by heavily compressing visual tokens, allowing all 50 candidates to fit within a single inference pass. However, this compression inevitably discards fine-grained visual details. V-Retrver accepts a modest latency overhead (ITR $1.0\times$ vs. Retrv-R1's $0.22\times$ at K=50) to preserve uncompressed visual representations, enabling precise tool-based inspection and yielding $+1.2\%$ on CIRR at K=50.

Tool-call efficiency under EAPO. As shown in Fig. 4(c), the model stabilizes at **1.6 to 1.7 average tool calls per query**, having learned under EAPO to invoke visual inspection only when it genuinely resolves ambiguity. This autonomously acquired efficiency makes V-Retrver well-suited for precision-critical applications (e.g., medical imaging, trademark detection) where retrieval accuracy outweighs minor latency overhead.

B Reward Component Ablation

To validate each component of the EAPO reward (Eq. 1), we train three variants by removing one term at a time, keeping all other settings identical. Results are presented in Table 7.

Removing r_{format} leads to a 1.7% drop in average recall, as malformed trajectories pollute policy updates. Replacing r_{rank} with a binary correctness signal causes the largest performance drop (2.4%), confirming the critical value of dense gradient feedback from CNDCG. Finally, removing r_{tool} reduces performance by 1.1%, as the model loses the incentive to perform successful evidence-based verification and has no mechanism to suppress redundant calls.

C Tool Tolerance Threshold Ablation

We ablate the tolerance threshold τ in the tool-use reward (Eq. 6) against a dynamic variant that scales τ by the semantic similarity variance of the top-K candidates. Results are summarized in Table 8.

Stricter τ=0 reduces accuracy by discouraging necessary tool calls. More lenient τ=2 or dynamic τ yield marginal accuracy gains ($\leq 0.3\%$) while inflating overhead (+28% to +46% more calls relative to τ=1). Fixed τ=1 achieves the optimal accuracy and efficiency balance for scalable retrieval.

D Exploration of RAG Applications

To further validate the practical utility of our framework, we extend our evaluation to Retrieval-Augmented Generation (RAG) scenarios. Following the experimental setup of LamRA (Liu et al., 2025), we evaluate our method on three Knowledge-based Visual Question Answering (KVQA) benchmarks. Specifically, we train the retrieval and VQA tasks simultaneously, allowing the model to align the agentic visual reasoning process with downstream generation needs. As detailed in Table 9, V-Retrver achieves superior performance in both retrieval precision and VQA accuracy, demonstrating that our Multimodal Interleaved Evidence Reasoning significantly enhances MLLM capabilities in RAG settings.

E Visual Toolset Extensibility

The current V-Retrver toolset (SELECT-IMAGE and ZOOM-IN) was intentionally kept minimal to validate the MIER framework and ensure stable EAPO policy optimization. To demonstrate the extensibility of our framework to richer perceptual operations, we integrate an additional OCR tool (EXTRACT-TEXT) into V-Retrver-7B and evaluate under identical settings. EXTRACT-TEXT allows the agent to invoke optical character recognition on a specified region of a candidate image, enabling it to acquire exact textual evidence (e.g., labels, captions, product names) that standard visual encodings may fail to capture reliably. As shown in

Table 6: Efficiency comparison. ITR and GMR are relative to V-Retrver ($= 1$); $\downarrow$ is better. All models use a *single* MLLM inference call, except LamRA-Rank-Pointwise which invokes the LLM independently for each candidate.

	$(q_i, q_t) \rightarrow c_i$ (CIRR)					
	$K = 20$			$K = 50$		
Methods	R@5	ITR $\downarrow$	GMR $\downarrow$	R@5	ITR $\downarrow$	GMR $\downarrow$
Qwen2.5-VL-7B	52.6	0.60	0.71	55.1	0.62	0.70
Vision-R1-7B	53.7	0.94	0.92	57.7	0.95	0.93
VLM-R1-7B	53.1	0.96	0.95	57.4	0.96	0.96
Retrv-R1-7B	69.9	0.25	0.31	72.3	0.22	0.28
LamRA-Rank-Pointwise-7B	65.4	4.17	0.21	67.9	4.41	0.11
V-Retrver-7B	**71.2**	1.00	1.00	**73.5**	1.00	1.00

Table 7: Ablation of EAPO reward components. Removing any term degrades performance; r_{rank} has the largest individual impact.

Variants	COCO R@5	F200K R@10	CIRR R@5	OVEN R@5	Avg.
V-Retrver-7B (full)	**87.5**	**37.8**	**73.5**	**69.8**	**67.2**
w/o r_{format}	86.2	36.1	71.5	68.2	65.5
w/o r_{rank}	85.5	35.3	70.4	68.0	64.8
w/o r_{tool}	86.8	37.0	72.1	68.5	66.1

Table 8: Ablation of tool tolerance threshold τ. Fixed τ=1 achieves the best accuracy and efficiency trade-off.

Variants	COCO R@5	F200K R@10	CIRR R@5	OVEN R@5	Avg.	Avg. Calls
Fixed τ=0	87.1	37.2	73.2	69.5	66.8	1.12
Fixed τ=1 (ours)	**87.5**	37.8	73.5	69.8	**67.2**	1.68
Fixed τ=2	87.8	**38.2**	**73.9**	**70.2**	67.5	2.45
Dynamic τ	87.8	37.8	73.6	70.0	67.3	2.15

Table 9: **Comparison of RAG capabilities on KVQA tasks.**

Method	OKVQA (Marino et al., 2019)	Infoseek (Chen et al., 2023)	E-VQA (Mensink et al., 2023)
	Retrieval (PR@5)		
PreFLMR (Lin et al., 2024b)	70.9	62.1	73.7
LamRA-7B (Liu et al., 2025)	89.0	73.4	75.0
V-Retrver-7B	**90.9**	**78.3**	**78.1**
	VQA (ACC)		
RA-VQAv2 w/ PreFLMR (Lin et al., 2023b)	61.9	32.1	54.5
LamRA-7B (Liu et al., 2025)	64.3	28.8	56.2
V-Retrver-7B	**65.7**	**31.9**	**58.0**

Table 10, adding EXTRACT-TEXT yields a consistent $+0.2\%$ average performance gain. The improvement is most pronounced on the knowledge-intensive OVEN dataset ($+0.4\%$ on $(q^i, q^t) \rightarrow c^t$), where exact text matching in candidate images provides discriminative evidence that visual features alone cannot supply. These results confirm that the MIER framework and EAPO training pipeline generalize naturally to new tool types: the agent successfully learns *when* to invoke EXTRACT-TEXT (i.e., when candidate images contain informative text) without requiring task-specific re-engineering of the reward or trajectory format. Expanding the toolset further (for example with object detection or spatial relationship queries) is a promising direction for future work.

Table 10: Effect of adding an OCR tool (EXTRACT-TEXT) to V-Retrver-7B under identical training and evaluation settings.

Variants	$q^t \rightarrow c^i$ COCO R@5	$q^i \rightarrow c^t$ F200K R@10	$(q^i, q^t) \rightarrow c^i$ CIRR R@5	$(q^i, q^t) \rightarrow c^t$ OVEN R@5	Avg.
V-Retrver-7B	87.5	37.8	73.5	69.8	67.2
V-Retrver-7B + EXTRACT-TEXT	**87.8**	**37.9**	**73.7**	**70.2**	**67.4**

F Implementation Details

Model Initialization. V-Retrver is initialized from Qwen2.5-VL-7B-Instruct (Bai et al., 2025). Throughout all training stages, the vision encoder is frozen while the language model is fine-tuned. The embedding model ϕ is constructed following the same method as LamRA (Liu et al., 2025).

Stage I: Supervised Fine-Tuning. We synthesize multimodal Chain-of-Thought data using Qwen2.5-VL-72B-Instruct, applying rule-based filtering to remove logically inconsistent or malformed samples. Training is conducted using the LLaMA-Factory (Zheng et al., 2024) framework on 8 A800 GPUs, with a batch size of 64, a learning rate of 1×10^{-5}, and 2 epochs.

Stage II: Rejection Sampling Fine-Tuning. For each training instance, multiple reasoning trajectories are sampled, and only those satisfying both format validity and correct retrieval ranking are retained. Training uses the same framework and hardware as Stage I.

Stage III: Evidence-Aligned Policy Optimization. RL training is based on the verl-tool (Jiang et al., 2025) framework, which extends verl (Sheng et al., 2024) and vLLM (Kwon et al., 2023) to support multimodal tool-augmented multi-turn training. The model is trained for 1 epoch with a learning rate of 1×10^{-6} and 8 rollouts per query.

Reward Hyperparameters. Table 11 summarizes all reward-related hyperparameters used in EAPO.

Table 11: Hyperparameter configurations for the EAPO training stage.

Symbol	Description	Value
K_r	CNDCG reward cutoff depth	5
m	Number of hard negatives for CNDCG	5
α	Format reward weight (Eq. 1)	0.2
β	Ranking reward weight (Eq. 1)	0.8
η	Tool success incentive (Eq. 5)	0.2
ρ	Tool excess penalty (Eq. 5)	0.1
τ	Tool tolerance threshold (Eq. 5)	1
λ	KL penalty coefficient (Eq. 7)	0

G Prompt Template

G.1 System Prompt

Fig. 5 illustrates the system prompt used for both training and inference.

G.2 User Prompt

Fig. 6 illustrates the user prompt used for both training and inference.

G.3 Annotation Prompt

Fig. 7 illustrates the annotation prompt. Specifically, for the CoT annotation process, this prompt is inserted into the user prompt to guide generation.

H Formal Definition of Contrastive NDCG Reward

Hard Negative Set. Given the candidate set $\mathcal{C}$, the hard negative set $\mathcal{H}$ consists of the top-m non-correct candidates ranked by embedding similarity to the query:

$$\mathcal{H} = \underset{j \neq j^*,\, |\mathcal{H}|=m}{\text{top-}m} \; \text{sim}(\phi(c_j),\ \phi(q)) \,. \qquad (9)$$

The remaining candidates form the easy negative set $\mathcal{E} = \mathcal{C} \setminus (\mathcal{H} \cup \{j^*\})$.

The key distinction from vanilla NDCG lies in the contrastive relevance assignment. This contrastive property provides targeted penalization precisely for the failure mode most detrimental to retrieval quality: a visually similar hard negative outranking the correct candidate.

System Prompt for Training and Inference

You are a helpful assistant for multimodal retrieval tasks.
Tools

You may call at most two tool invocations in total. In any single turn, you may emit only one <tool_call></tool_call> block.
You are provided with function signatures within <tools></tools> XML tags.

<tools>
{"type": "function", "function": {"name": "crop_image", "description": "Zoom in on a query image or a candidate image based on the bounding box coordinates.", "parameters": {"type": "object", "properties": {"bbox_2d": {"type": "array", "description": "Bounding box of the area to zoom in, using pixel coordinates. Minimum is 0 and maximum is the image width/height.", "items": {"type": "number"}}, "target_image": {"type":"number", "description": "Index of the image to crop. Use 0 for the query image, or an integer from 1 to the number of candidate images."}}, "required": ["bbox_2d", "target_image"]}}
{"type": "function", "function": {"name": "select_images", "description": "Select 1 to 4 query or candidate images for detailed analysis from the query/candidate list. You can select 1, 2, 3, or 4 images, but never more than 4.", "parameters": {"type": "object", "properties": {"target_images": {"type": "array", "description": "List containing 1 to 4 image indices to select for detailed analysis. Use index 0 to select the query. Use indices 1, 2, 3, ... up to the number of candidates to select candidate. You can mix query and candidate images, e.g., [0] for query only, [10] for candidate 10 only, [0, 10, 11] for query and candidates 10, 11, [2, 3, 10] for candidates 2, 3, 10, etc. Maximum 4 images total.", "items": {"type": "integer", "description": "Image index: 0 for the query image, or 1 to the number of candidates for candidate images."}, "minItems": 1, "maxItems": 4}}, "required": ["target_images"]}}}
</tools>

For each function call, return exactly one json object with function name and arguments within <tool_call></tool_call> XML tag:
<tool_call>
{"name": <function-name>, "arguments": <args-json-object>}
</tool_call>

Figure 5: System prompt template for training and inference.

User Prompt for Training and Inference

Query: <question> Candidates: (1) <candidate$_1$>(2) <candidate$_2$>.....(K) <candidate$_K$>

Guidelines: You are given a query (which can be text, image, or both) and multiple candidate items (which can be text, image, or both). Your task is to rank all candidates from most relevant to least relevant based on how well they match the query.

You can use the following tools to assist with analysis:
1. `crop_image`: Zoom in on specific areas of the query image or a candidate image for detailed analysis.
2. `select_images`: Select 1 to 4 query images or candidate images from the provided list for focused comparison.

Analyze the visual and textual features step by step. Provide your final answer in one of the following exact formats:
If you DO need to use tools:
<think>Your detailed reasoning process here, explaining your analysis steps and tool usage.</think><tool_call>{"name": <function-name>, "arguments": <args-json-object>}</tool_call><think>Continue your reasoning based on tool outputshere.</think><answer>[ranked_list]</answer>
If you do NOT need to use tools:
<think>Your detailed reasoning process here, explaining your analysis steps.</think><answer>[ranked_list]</answer>

Example:
<think>Your reasoning process here</think><tool_call>{"name": "select_images", "arguments": {"target_images": [0, 1, 2]}}</tool_call><think>Continue your reasoning based on tool outputs here</think><answer>[14, 13, 20, 11, 16, 19, 6, 7, 2, 17, 3, 9, 1, 18, 10, 5, 15, 12, 4, 8]</answer>

Figure 6: User prompt template for training and inference.

Annotation Prompt

Analyze the visual and textual features step by step. Your reasoning process should generally follow these stages:

First, analyze the key features and requirements of the query (both visual and textual).
Then perform a quick pre-screening of all candidates to filter out clearly irrelevant ones. Next, identify a subset of the most promising candidates (e.g., the top 5 to 8 potential matches) that appear to be the strongest contenders based on this initial scan.
For these promising candidates, provide a more detailed analysis of their distinguishing features, explaining why each one might or might not match the query (e.g., taxonomy, visual characteristics, context, or any other relevant attributes).
If there are important, ambiguous, or missing visual details that are critical for making a precise decision, decide whether to call tools. You may, for example, use `crop_image` to zoom into specific regions when finer visual inspection is needed, or use `select_images` to focus on a small subset of the most relevant images.
After each tool call, carefully summarize the visual information obtained from the tool outputs (e.g., what you observe in the selected or cropped images) and integrate this information into your reasoning. Use these visual observations to refine your comparison among the promising candidates and to resolve any ambiguities.
Based on the combined textual and visual analysis, determine which candidate is the single best match to the query. Then use your reasoning to order all remaining candidates by relevance, from more relevant to less relevant, until you obtain a complete ranking of all candidates.
Finally, construct the complete ranking list containing ALL candidate indices from 1 to N. Place your determined best match first, followed by the remaining candidates ordered by relevance, and then provide your final answer in the required format.

Figure 7: Annotation prompt template.

I Dataset Details

We present the statistics of all evaluation benchmarks in Table 12 and the full breakdown of the M-BEIR training and evaluation splits in Table 13.

J Detailed Experiment Settings

We establish three distinct experiment settings to comprehensively evaluate V-Retrver.

Setting I: Supervised Evaluation on M-BEIR. To validate the versatility of our method, we train V-Retrver on all 8 tasks in M-BEIR and evaluate on the corresponding test sets. Baselines include: (1) foundational VLMs: Qwen2.5-VL-3B/7B, CLIP-L, SigLIP, BLIP, BLIP2; (2) fine-tuned universal retrievers: UniIR-BLIP$_{\text{FF}}$, UniIR-CLIP$_{\text{SF}}$; (3) reasoning-enhanced and universal retrieval models: Vision-R1 (Huang et al., 2025), VLM-R1 (Shen et al., 2025), MM-Embed (Lin et al., 2024a), LamRA (Liu et al., 2025), U-MARVEL (Li et al., 2025a), Retrv-R1 (Zhu et al., 2025d); and (4) agentic tool-use models: DeepEyes.

Setting II: Zero-shot Evaluation on Unseen Datasets. To evaluate generalization on previously unseen retrieval datasets, we perform zero-shot experiments on 5 datasets not encountered during training: CIRCO, GeneCIS, Visual Storytelling, Visual Dialog, and Multi-round FashionIQ. Baselines include: E5-V (Jiang et al., 2024), MagicLens (Zhang et al., 2024a), MM-Embed (Lin et al., 2024a), LamRA (Liu et al., 2025).

Setting III: Held-out Task Evaluation. To investigate generalization to unseen retrieval tasks, we intentionally exclude three task types during training: image-to-image retrieval ($q^i \rightarrow c^i$), text-image-to-text retrieval ($(q^i, q^t) \rightarrow c^t$), and text-image-to-text-image retrieval ($(q^i, q^t) \rightarrow (c^i, c^t)$). Training is conducted on the remaining five tasks, and evaluation is performed on the three held-out tasks. In this zero-shot setting, we compare against Qwen2.5-VL-7B, Vision-R1, and LamRA* (also trained on the five non-held-out tasks for a fair comparison).

The unseen datasets used in Setting II are summarized in Table 14.

K Algorithms and Detailed Analysis

We present the formal algorithms for the inference and training processes of V-Retrver. The inference process, formulated as a coarse-to-fine pipeline with sliding window agentic reasoning, is detailed in Algorithm 1. The three-stage curriculum learning strategy, designed to progressively align the model with evidence-driven retrieval objectives, is presented in Algorithm 2.

L Qualitative Examples

To provide an intuitive illustration of our approach and further demonstrate the effectiveness of V-Retrver, we present qualitative results in Figs. 8 through 12. These examples illustrate how V-Retrver performs accurate retrieval through fine-grained and structured reasoning, highlighting the strong effectiveness of the proposed method.

Table 12: Summary of evaluation benchmarks, categorized into Supervised and Zero-shot settings. # Queries is the number of test queries; # Candidates is the number of test candidates per query.

Benchmark	**# Queries**	**# Candidates**	**Setting**
M-BEIR (Wei et al., 2024)	190K	5.6M	Supervised
CIRCO (Baldrati et al., 2023)	800	120K	Zero-shot
GeneCIS (Vaze et al., 2023)	8K	10–15	Zero-shot
Visual Storytelling (Huang et al., 2016)	5K	8K	Zero-shot
Visual Dialog (Das et al., 2017)	2K	2K	Zero-shot
Multi-round FashionIQ (Yuan and Lam, 2021)	2.4K	6.2K	Zero-shot

Table 13: Full breakdown of the M-BEIR benchmark by task, dataset, and split.

Task	**Dataset**	**Domain**	**# Train**	**# Dev**	**# Test**	**# Pool**
$q^t \to c^i$	VisualNews	News	99K	20K	20K	542K
	MSCOCO	Misc.	100K	24.8K	24.8K	5K
	Fashion200K	Fashion	15K	1.7K	1.7K	201K
$q^t \to c^t$	WebQA	Wiki	16K	1.7K	2.4K	544K
$q^t \to (c^i, c^t)$	EDIS	News	26K	3.2K	3.2K	1M
	WebQA	Wiki	17K	1.7K	2.5K	403K
$q^i \to c^t$	VisualNews	News	100K	20K	20K	537K
	MSCOCO	Misc.	113K	5K	5K	25K
	Fashion200K	Fashion	15K	4.8K	4.8K	61K
$q^i \to c^i$	NIGHTS	Misc.	16K	2K	2K	40K
$(q^i, q^t) \to c^t$	OVEN	Wiki	150K	50K	50K	676K
	InfoSeek	Wiki	141K	11K	11K	611K
$(q^i, q^t) \to c^i$	FashionIQ	Fashion	16K	2K	6K	74K
	CIRR	Misc.	26K	2K	4K	21K
$(q^i, q^t) \to (c^i, c^t)$	OVEN	Wiki	157K	14.7K	14.7K	335K
	InfoSeek	Wiki	143K	17.6K	17.6K	481K
Total	10 datasets	4 domains	**1.1M**	**182K**	**190K**	**5.6M**

Table 14: Summary of unseen datasets used in zero-shot evaluation (Setting II).

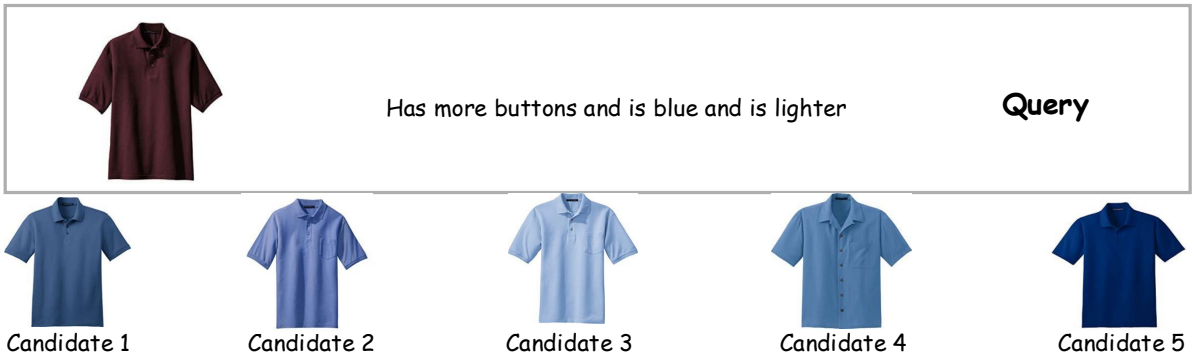
Dataset	**Image Source**	**Task**	**Query Format**
CIRCO	MSCOCO (unlabeled)	$(q^i, q^t) \to c^i$	`<image><rel. caption>`
GeneCIS	MSCOCO	$(q^i, q^t) \to c^i$	`<image><rel. caption>`
Visual Dialog	MSCOCO	$q^{\text{dialog}} \to c^i$	`<Q1><A1>...<Qj>`
Vis. Story.	Flickr	$(q^i \oplus q^t) \to c^i$	`<text><image>...<text>`
MT-FashionIQ	FashionIQ	$(q^i \oplus q^t) \to c^i$	`<image><cap>...<cap>`

Algorithm 1 V-Retrver Inference Pipeline

Input: Query q, Candidate Pool $\Omega = \{c_n\}_{n=1}^{N}$, Embedding Model Φ, Reasoning Agent π_θ, Top-K size K, Window size W, Stride S
Output: Ranked Candidate List $\hat{L}$
{Stage 1: Coarse Retrieval}
Compute $s_n = \cos(\Phi(q), \Phi(c_n))$ for all $c_n \in \Omega$
$\mathcal{C}_{\text{top}} \leftarrow$ Top-K$(\Omega, \{s_n\})$
{Stage 2: Agentic Reranking}
Initialize $\mathcal{L}_{\text{global}} \leftarrow \emptyset$
Split $\mathcal{C}_{\text{top}}$ into windows $\{w_1, \ldots, w_m\}$
for each window w_j **do**
  $H_0 \leftarrow (q,\ w_j,\ \text{Instruction}), \quad t \leftarrow 0$
  while True **do**
    Generate output: $o_t \sim \pi_\theta(H_t)$
    if o_t contains `<tool_call>` **then**
      Parse a_t from o_t
      $v_{\text{obs}} \leftarrow f_{\text{tool}}(a_t, w_j)$
      $H_{t+1} \leftarrow H_t \oplus o_t \oplus v_{\text{obs}}$
    else if o_t contains `<answer>` **then**
      Parse $\hat{r}_j$ from o_t
      Update $\mathcal{L}_{\text{global}}$ with $\hat{r}_j$; **break**
    end if
    $t \leftarrow t + 1$
  end while
end for
$\hat{L} \leftarrow$ AggregateRanks$(\mathcal{L}_{\text{global}})$

Algorithm 2 Curriculum-Based Agentic Training

Input: Pretrained MLLM θ_{init}, Dataset $\mathcal{D}$, Synth Model M_{syn}
Output: Optimized Policy π_{θ^*}
{Stage 1: Reasoning Activation (SFT)}
$\mathcal{D}_{\text{sft}} \leftarrow \{(q, c, \tau_{\text{cot}})\}$ via M_{syn} on $\mathcal{D}$
Filter $\mathcal{D}_{\text{sft}}$ for format compliance
$\theta_{\text{sft}} \leftarrow$ Minimize $\mathcal{L}_{\text{SFT}}(\theta_{\text{init}}, \mathcal{D}_{\text{sft}})$
{Stage 2: Reliability Refinement (RSFT)}
Initialize $\mathcal{D}_{\text{rsft}} \leftarrow \emptyset$
for each $(q, c) \in \mathcal{D}$ **do**
  Sample $\{\tau_1, \ldots, \tau_k\} \sim \pi_{\theta_{\text{sft}}}(q, c)$
  if IsFormatValid$(\tau_i) \wedge$ IsRankCorrect(τ_i) **then**
    Add τ_i to $\mathcal{D}_{\text{rsft}}$
  end if
end for
$\theta_{\text{rsft}} \leftarrow$ Minimize $\mathcal{L}_{\text{SFT}}(\theta_{\text{sft}}, \mathcal{D}_{\text{rsft}})$
{Stage 3: EAPO}
$\theta \leftarrow \theta_{\text{rsft}}, \quad \pi_{\text{ref}} \leftarrow \theta_{\text{rsft}}$
while not converged **do**
  Sample $B_q \sim \mathcal{D}$
  for each $q \in B_q$ **do**
    Sample $\{o_1, \ldots, o_G\} \sim \pi_\theta(q)$
    $R(o_i) = \alpha r_{\text{fmt}} + \beta r_{\text{rank}} + r_{\text{tool}}$
  end for
  Compute A_i via group normalization
  Update $\theta \leftarrow$ Optimize $\mathcal{J}_{\text{EAPO}}(\theta)$
end while
return π_θ



<think>The query specifies a shirt that is blue, lighter in color, and has more buttons. Let's analyze the candidates based on these criteria:1. Candidate 1: This shirt is blue but appears to be a darker shade. It has a collar but no visible buttons.2. Candidate 2: This shirt is blue and lighter in color. It has a pocket but no visible buttons.3. Candidate 3: This shirt is light blue and has a pocket. It also has a visible button on the collar.4. Candidate 4: This shirt is blue and lighter in color. It has multiple visible buttons and a pocket.5. Candidate 5: This shirt is blue but appears to be a darker shade. It has a collar but no visible buttons.Based on the initial analysis, Candidate 4 seems to match the query as it is blue, lighter in color, and has multiple buttons. Candidate 3 is also a strong contender as it is light blue and has a button, but it seems to have fewer buttons compared to Candidate 4. Candidate 2 is light blue but lacks buttons. Candidate 1 and Candidate 5 are darker blue and lack buttons.To confirm the button count and color shade, I will use the `select_images` tool to focus on Candidates 3 and 4</think>
<tool_call>{"name": "select_images", "arguments": {"target_images": [3, 4]}} </tool_call>

Candidate 3 Candidate 4

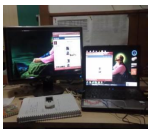
<think>The selected images confirm the following:- Candidate 3 (Image 1): This shirt is light blue and has a single button on the collar. It matches the query's requirement of being blue and lighter in color but does not have multiple buttons.- Candidate 4 (Image 2): This shirt is blue and lighter in color. It has multiple buttons, which aligns with the query's requirement of "more buttons." Based on the detailed analysis, Candidate 4 is the best match as it meets all the query requirements. Candidate 3 is the second-best match as it is light blue and has a button, but it has fewer buttons. The remaining candidates are less relevant. The final ranking is: 4, 3, 2, 1, 5 </think>
<answer>[4, 3, 2, 1, 5]</answer>

Figure 8: A qualitative example of the retrieval result generated by V-Retrver.

Find a smaller monitor and an additional laptop

Query

Candidate 1



Candidate 2

Candidate 3

Candidate 4

Candidate 5

<think>The query image shows a setup with a smaller monitor and an additional laptop. Let's analyze the candidates:1. Candidate 1: This is the same image as the query, so it matches perfectly. 2. Candidate 2: This setup has a smaller monitor and a laptop, similar to the query. The arrangement and the presence of both devices match the query. 3. Candidate 3: This setup has three laptops, which does not match the query's description of a smaller monitor and a laptop. 4. Candidate 4: This setup has a laptop and a monitor, but the monitor is larger compared to the smaller monitor in the query. 5. Candidate 5: This setup has a smaller monitor and a laptop, similar to the query. The arrangement and the presence of both devices match the query. Based on the analysis, the most promising candidates are 1, 2, and 5. Candidates 3 and 4 are less relevant. To further refine the ranking, let's select images 2 and 5 for detailed analysis.</think>
<tool_call>{"name": "select_images", "arguments": {"target_images": [2, 5]}} </tool_call>

Candidate 2

Candidate 5

<think>After analyzing the selected images:Image 2 (Candidate 2): This setup has a smaller monitor and a laptop, similar to the query. The arrangement and the presence of both devices match the query. The smaller monitor is clearly visible, and the laptop is positioned next to it. Image 5 (Candidate 5): This setup also has a smaller monitor and a laptop, similar to the query. The arrangement and the presence of both devices match the query. The smaller monitor is clearly visible, and the laptop is positioned next to it. Both images 2 and 5 are very similar to the query. However, image 2 has a more similar arrangement and size ratio of the smaller monitor and laptop compared to the query. Based on the detailed analysis, the final ranking is: 1. Candidate 2, 2. Candidate 5,3. Candidate 1 (same as the query image), 4. Candidate 4, 5. Candidate 3</think>
<answer>[2, 5, 1, 4, 3]</answer>

Figure 9: A qualitative example of the retrieval result generated by V-Retrver.

Green metallic knitted dress **Query**

Candidate 1 Candidate 2 Candidate 3 Candidate 4 Candidate 5

<think>The query specifies a "Green metallic knitted dress." Let's analyze the candidates based on the key features: color (green), material (metallic and knitted), and style (dress).1. Candidate 1: This dress is green and knitted, but it appears to be a solid matte color with no visible metallic finish.2. Candidate 2: This dress is metallic and knitted, but the color is dark/olive, not the requested green.3. Candidate 3: This dress is green and has a knitted texture. However, looking at the full-body image, I am uncertain if the fabric actually has a metallic sheen or if it is just the lighting on standard yarn.4. Candidate 4: This dress is green and knitted, but looks like a standard textile without metallic properties.5. Candidate 5: This dress is metallic and knitted, but the color is incorrect (not green).Because I am not sure if Candidate 3 actually possesses the required metallic luster, I need to use the crop tool to inspect the fabric details more closely to confirm.</think>
<tool_call>{"name": "crop_image", "arguments": {"bbox_2d": [0, 0, 180, 250], "target_image": 3}}</tool_call>

Candidate 3 Cropped image

<think>The selected images confirm the following:- Candidate 3 (Image 1): This shirt is light blue and has a single button on the collar. It matches the query's requirement of being blue and lighter in color but does not have multiple buttons.- Candidate 4 (Image 2): This shirt is blue and lighter in color. It has multiple buttons, which aligns with the query's requirement of "more buttons." Based on the detailed analysis, Candidate 4 is the best match as it meets all the query requirements. Candidate 3 is the second-best match as it is light blue and has a button, but it has fewer buttons. The remaining candidates are less relevant. The final ranking is: 4, 3, 2, 1, 5 </think>
<answer>[4, 3, 2, 1, 5]</answer>

Figure 10: A qualitative example of the retrieval result generated by V-Retrver.

What is this plant named after? **Query**

Candidate 1

Lilium humboldtii. Lilium humboldtii, or Humboldt's lily, is a species of lily native to the US state of California and the Mexican state of Baja California. It is named after naturalist and explorer Alexander von Humboldt. It is native to the South High Cascade Range, High Sierra Nevada, south Outer South Coast Ranges, and the Santa Monica Mountains and others in Southern California, growing at elevations from 600 (m) to 1200 (m).Description. "Lilium humboldtii" grows up to 6 (ft) tall, with flowers that are maroon-spotted, golden-orange with dark red splotches, with orange to brown stamens. The plant flowers in June, with flowers.

Candidate 2

Lilium humboldtii. Roezl and Leichtlin, gave it the name "Lilium bloomerianum". For some time afterward, the name was still applied to the southern California "Lilium humboldtii" subsp. "ocellatum". Cultivation."Lilium humboldtii" is sold as a garden bulb. It prefers dry summer dormancy, with no water after blooming, good drainage, and part shade. It was one of the parents, along with "Lilium pardalinum", that produced the Bellingham hybrid lilies, which eventually resulted in the popular 'Shuksan' and 'Star of Oregon' lilies.Harlow, Nora, Jakob, Kristin, and Raiche, Roger (2003) "Wild Lilies, Irises, and Grasses". University of California Press. USDA Plants Profile: "Lilium humboldtii" Treatment from the.

Candidate 3

Lilium humboldtii. growing in a pyramidal inflorescence. The flowers are on stout stems, which are sometimes brown-purple. The subrhizomatous bulb is large, with yellowish-white scales, and grows very deep in the soil. The leaves grow in whorls, and are undulate, shiny, and oblanceolate. It is summer-deciduous, dying back after flowering in mid to late summer.Subspecies "Lilium humboldtii" subsp. "humboldtii" - central California"Lilium humboldtii" subsp. "ocellatum" - southern California, Baja CaliforniaBoth subspecies are on the California Native Plant Society "Inventory of Rare and Endangered Plants of California" and described as "fairly endangered in California".Albert Kellogg, unaware that the plant had already been named by.

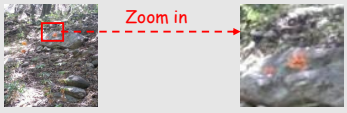
Candidate 4

Lilium humboldtii. Jepson Manual-"Lilium humboldtii" CalFlora Taxon Report: "Lilium humboldtii " "Lilium humboldtii"-CalPhotos galleryiucnredlist.,

Candidate 5

Lilium parryi. especially "Hyles lineata" and "Sphinx perelegans".Threats to this species include grazing, recreation, natural flooding and human alterations in water regimes, and horticultural collecting of the bulbs and flowers."Lilium parryi" was named for Charles Christopher Parry (28 August 1823 2013 20 February 1890), a British-American botanist and mountaineer.Idyllwild, California, hosts the Lemon Lily Festival, which celebrates this species.Calflora Database: "Lilium parryi" (lemon lily)Jepson Manual Treatment of "Lilium parryi"USDA Plants Profile for "Lilium parryi" (lemon lily)UC Photos gallery: "Lilium parryi.

<think>The query image shows a natural setting with orange flowers that appear to be lilies. The key features to consider are the flower color, shape, and the environment in which they are growing. The candidates provided are descriptions of different lily species, with the most relevant being those that match the visual characteristics of the flowers in the imagePre-screening:- Candidates (1) to (4) are about "Lilium humboldtii," which is described as having golden-orange flowers with dark red splotches, fitting the visual characteristics of the flowers in the image.- Candidate (5) is about "Lilium parryi," which is described as having lemon-colored flowers, which does not match the orange color of the flowers in the image. Detailed Analysis of Promising Candidates:- Candidate (1): Describes "Lilium humboldtii" with golden-orange flowers and dark red splotches, matching the visual characteristics of the flowers in the image. It is named after Alexander von Humboldt.- Candidate (2): Also describes "Lilium humboldtii" and provides additional information about its cultivation and history. It confirms the name "Lilium humboldtii" and its association with Humboldt.- Candidate (3): Provides more detailed botanical information about "Lilium humboldtii," including its subspecies and conservation status. It also confirms the name and association with Humboldt.- Candidate (4): Mentions "Lilium humboldtii" and provides references for further information. It confirms the name and association with Humboldt.Tool Usage:The image is clear enough to identify the flower color and shape, but a closer look at the flower details might help confirm the presence of dark red splotches mentioned in the descriptions. Therefore, I will use the `crop_image` tool to zoom in on one of the flowers for a more detailed analysis..</think>
<tool_call>{"name": "crop_image", "arguments": {"bbox_2d": [125, 88, 145, 108], {"target_image": 0}} </tool_call>



Query image Cropped image

<think>The cropped image shows a close-up of one of the orange flowers in the query image. The flower appears to have dark red splotches, which aligns with the description of "Lilium humboldtii" in the candidates. This visual confirmation strengthens the match between the query image and the descriptions of "Lilium humboldtii." Final Ranking:1. Candidate (1): Best match as it directly describes "Lilium humboldtii" with golden-orange flowers and dark red splotches, matching the visual characteristics of the flowers in the image.2. Candidate (2): Also a strong match, providing additional information about the cultivation and history of "Lilium humboldtii."3. Candidate (4): Mentions "Lilium humboldtii" and provides references for further information. 4. Candidate (3): Provides detailed botanical information about "Lilium humboldtii," including its subspecies and conservation status. 5. Candidate (5): Least relevant as it describes "Lilium parryi" with lemon-colored flowers, which do not match the orange flowers in the image. </think>
<answer>[1, 2, 4, 3, 5]</answer>

Figure 11: A qualitative example of the retrieval result generated by V-Retrver.

Query

Candidate 1

Candidate 2

Candidate 3

Candidate 4

Candidate 5

<think>The query image shows a barn with a curved roof, wooden siding, and a closed double door. The barn is set against a grassy field with a colorful sky in the background. Let's analyze the candidates:1. Candidate 1: This barn has a triangular roof, not a curved one, and the doors are different. It is not a match.2. Candidate 2: This barn has a curved roof and a similar door style, but the surroundings and the barn's color are different. It is a potential match.3. Candidate 3: This barn has a triangular roof and an open door, which is not similar to the query. It is not a match.4. Candidate 4: This barn has a triangular roof and a different door style. It is not a match.5. Candidate 5: This barn has a curved roof and a similar door style, but the color and surroundings are different. It is a potential match.Candidates 2 and 5 are the most promising. Let's use the `select_images` tool to focus on these two for a detailed comparison./think>
<tool_call>{"name": "select_images", "arguments": {"target_images": [2, 5]}}</tool_call></tool_call>

Candidate 2

Candidate 5

<think>After analyzing the selected images:- Candidate 2: This barn has a curved roof and a similar door style to the query image. The surroundings include trees, which are not present in the query image, but the barn's structure is very similar.- Candidate 5: This barn also has a curved roof and a similar door style. The surroundings are more similar to the query image, with a grassy field and a clear sky.Both candidates are very similar to the query image, but Candidate 5 has surroundings that are more similar to the query image. Therefore, Candidate 5 is the best match.Now, let's rank all candidates:1. Candidate 5: Best match due to similar structure and surroundings.2. Candidate 2: Second best match due to similar structure but different surroundings.3. Candidate 1: Not a match due to different roof shape.4. Candidate 3: Not a match due to different roof shape and open door.5. Candidate 4: Not a match due to different door style.</think>
<answer>[5, 2, 1, 3, 4]</answer>

Figure 12: A qualitative example of the retrieval result generated by V-Retrver.