Region-R1: Reinforcing Query-Side Region Cropping for Multi-Modal Re-Ranking

Anonymous ACL submission

Abstract

Multi-modal retrieval-augmented generation (MM-RAG) depends on re-ranking to surface the most relevant evidence for image-question queries. We propose **Region-R1**, which treats query-side region cropping as a decision problem during re-ranking, allowing the system to retain the full image or focus on a question-relevant region before scoring candidates with a fixed vision-language encoder. The policy is trained with reinforcement learning using designed rewards. Across two challenging benchmarks, Region-R1 delivers consistent gains in top-heavy ranking, improving *top-1* retrieval over prior re-rankers and increasing conditional Recall@1 by up to 20% in relative terms. These results show that query-side adaptation is a simple way to strengthen MM-RAG re-ranking.

1 Introduction

Multi-modal retrieval-augmented generation (MM-RAG) has become an increasingly important paradigm for grounding vision language models (VLMs) with both textual and visual information (Hu et al., 2023b; Lin and Byrne, 2022; Yasunaga et al., 2023; Chen et al., 2022). In MM-RAG systems, queries may include images in addition to natural language, and retrieved evidence may consist of heterogeneous modalities such as images, captions, diagrams, or structured visual content. As a result, the effectiveness of MM-RAG depends not only on language understanding, but also on how visual information is represented.

In practical MM-RAG pipelines, an upstream retriever first produces a candidate set, after which a re-ranking model refines their ordering with respect to a given query. Recent work has primarily focused on improving retrievers (Lin et al., 2023; Lin and Byrne, 2022; Zhou et al., 2024) or designing more expressive multi-modal re-rankers (Yu et al., 2024; Yan and Xie, 2024; Yang et al., 2025). By contrast, comparatively little attention has been paid to how the *multi-modal query representation itself* is constructed and controlled during re-ranking. In particular, re-rankers typically rely on a single global embedding of the query image, implicitly assuming that all regions of the image are relevant to the user's question.

This assumption is often violated in realistic MM-RAG scenarios. Query images frequently contain distractive objects, background regions, or contextual elements that are irrelevant to the question being asked. When such irrelevant regions dominate the global visual representation, they can distort visual similarity estimates and degrade re-ranking performance, even when the candidate set is fixed. This issue is specific to multi-modal settings and does not arise in purely textual RAG.

A natural response is to perform region cropping on the query image in a question-conditioned manner. Modern vision-language models exhibit strong localization capabilities, suggesting that they can identify regions relevant to a given question. Indeed, our preliminary analysis shows that, under a fixed candidate pool and a scoring model, replacing the full query image with an appropriately selected region can substantially improve multi-modal re-ranking performance. However, naive or heuristic cropping strategies can also remove useful visual context and lead to worse rankings. This raises a key challenge for MM-RAG systems: *how can query-side visual information be selectively modified in a way that reliably improves multi-modal re-ranking performance?*

In this work, we formulate query-side region cropping as a decision problem optimized directly for multi-modal re-ranking objectives, named **Region-R1**. We focus on the re-ranking stage because retrieval is a coarse-grained, high-recall step that must run efficiently at corpus scale, whereas re-ranking performs fine-grained discrimination over a small top-K set. Applying region cropping during retrieval would require policy execution during large-scale negative mining and may even necessitate re-indexing, substantially increasing training and system cost. In addition, it also risks discarding

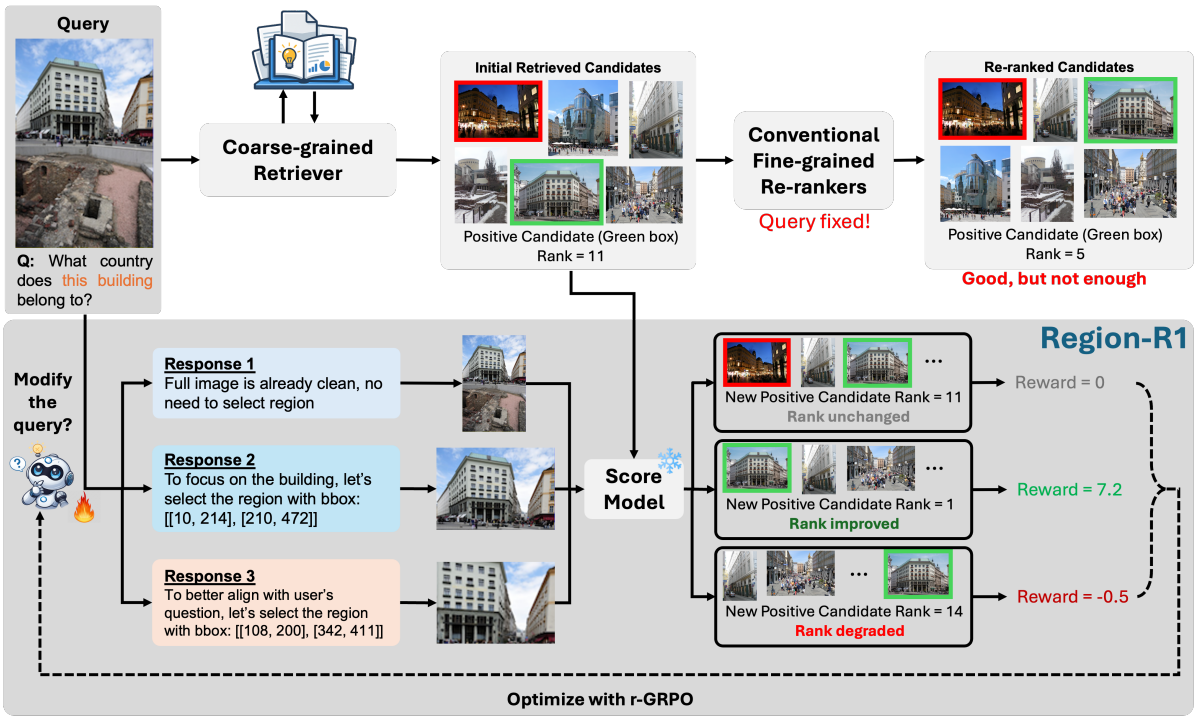

Figure 1: **Overview of query-side region cropping approach.** Conventional re-rankers treat the query as fixed and only re-order candidates. Our method instead automatically adapt the query by selecting an informative region or keeping the original query before scoring candidates.

information too early and hurting recall. Accordingly, we apply region cropping only at re-ranking, where computation is bounded and the selected region helps disambiguate among already plausible candidates.

Given a query image and user question, a policy determines whether to retain the full image or to select a region, with the explicit goal of improving the ranking of a fixed multi-modal candidate set. Our model learns the region cropping policy using reinforcement learning, with rewards defined as improvements over baseline measured by standard information retrieval metrics such as Mean Reciprocal Rank (Craswell, 2016) and Normalized Discounted Cumulative Gain (Wang et al., 2013).

We evaluate the proposed approach on two challenging datasets, E-VQA (Mensink et al., 2023) and InfoSeek (Chen et al., 2023). Experimental results show that the learned policy selectively applies region cropping when beneficial. Most notably, Region-R1 improves *top-1 recall* over prior re-rankers in our setting. Ours CondRecall@1 improves 20% on E-VQA and 8% on InfoSeek. These gains translate into more reliable downstream grounding because MM-RAG systems often condition generation on the top-ranked evidence.

In summary, this paper makes the following contributions:

- We introduce *query-side region cropping* for multi-modal re-ranking, and formulate it as a *decision problem* to improve ranking.
- We propose **Region-R1**, a *candidate-agnostic* reinforcement learning framework that learns a region cropping policy optimized directly for multi-modal re-ranking.
- Our experiment shows Region-R1 consistently improves performance on two challenging datasets, boosting CondRecall@1 by 20% on E-VQA and 8% on InfoSeek. We further provide diagnostic analyses of when region cropping is applied and ablations isolating the effect of each reward component.

2 Related Works

Knowledge-Based VQA. Knowledge-based Visual Question Answering (KB-VQA) requires models to answer questions by retrieving external knowledge beyond the query input (Marino et al., 2019; Schwenk et al., 2022). Early methods leveraged structured knowledge graphs (Narasimhan et al., 2018; Zhu et al., 2020; Gardères et al., 2020; Li et al., 2020; Wu et al., 2022), but often struggled with the open-ended and diverse nature of real-world questions. Recent work has

largely shifted to Retrieval-Augmented Generation (RAG), retrieving evidence from unstructured corpora (e.g., Wikipedia). Benchmarks such as InfoSeek (Chen et al., 2023) and Encyclopedic-VQA (E-VQA) (Mensink et al., 2023) formalize this paradigm and emphasize fine-grained entity linking and multi-hop reasoning. State-of-the-art systems commonly follow a "Retriever-Reranker-Generator" pipeline. Wiki-LLaVA (Caffagni et al., 2024) applies hierarchical retrieval, while EchoSight (Yan and Xie, 2024) and OMGM (Yang et al., 2025) use Q-Former-based re-ranking after initial retrieval.

More recently, studies optimize the interaction between VLMs and retrieval. ReflectiVA (Cocchi et al., 2025) and MMKB-RAG (Ling et al., 2025) add self-reflective signals and consistency checks to decide when retrieval is needed, and ReAG (Compagnoni et al., 2025) and Iterative-RAG (Choi et al., 2025) investigate iterative retrieval to refine evidence chains. A persistent challenge is the semantic gap between visual queries and textual documents. To mitigate it, several works adapt query expansion from text retrieval (Mo et al., 2023), using generated captions or entity-aware rewriting to better guide search (Adjali et al., 2024; Zhu et al., 2024). However, most still treat the visual input as a fixed global representation, underutilizing discriminative visual query reformulation for filtering distractors.

Multi-Modal Re-Ranking. Re-ranking is a critical refinement stage in MM-RAG pipeline, designed to re-order candidate evidence to maximize the rank of true positives. Recent approaches fall into two primary categories. *(1) Interaction-Centric Methods.* These models rely on deep cross-modal alignment or generative reasoning to score candidates. EchoSight (Yan and Xie, 2024), OMGM (Yang et al., 2025), and ReflectiVA (Cocchi et al., 2025) orchestrate fine-grained evidence fusion and self-reflective tokens to weigh relevance dynamically. RAMQA (Bai et al., 2025b) and Mario (Ramezan et al., 2025) employ Large Multimodal Models as generative list-wise re-rankers, synthesizing text and visual history to predict the optimal permutation of documents. *(2) Representation-Centric Methods.* Alternatively, other works optimize the embedding space itself. EchoSight (Yan and Xie, 2024), OMGM (Yang et al., 2025), MM-Embed (Lin et al., 2024), and DocReRank (Wasserman et al., 2025) force retrievers to distinguish subtle visual disagreements with hard negative samples.

Despite these advancements, both paradigms typically treat the visual query as a fixed, global tensor. Our work addresses this by introducing *discriminative region cropping*, effectively reformulating the visual query for the re-ranking stage.

3 Methodology

3.1 Problem Definition

We consider the re-ranking stage of a MM-RAG pipeline. Each query is represented by an image-question pair $x = (I_q, q)$, where I_q denotes the query image and q denotes the associated question. Given a query, an upstream retriever produces a candidate set $\mathcal{C} = \{c_j\}_{j=1}^{N}$, where each candidate $c_j = (I_j, t_j)$ consists of an image I_j and textual description t_j. The objective of re-ranking is to produce an ordering π over $\mathcal{C}$ that reflects relevance to the query. Among candidates $\mathcal{C}$, the relevance labels $y \in \mathbb{R}^N$ are given. $y_j = 1$ for the positive candidate, and $y_j = 0$ otherwise.

Throughout this work, the candidate set is assumed to be fixed, and only the query representation is modified during re-ranking.

3.2 Query-Side Region Cropping

Let $d \in \{\mathsf{REGION}, \mathsf{FULL}\}$ denote a discrete decision variable indicating whether a region is selected from the query image. When $d = \mathsf{REGION}$, a continuous bounding box $b = (x_1, y_1, x_2, y_2)$ is predicted. A deterministic cropping operator $g(\cdot)$ that extracts the region specified by b.

The transformed query image $\tilde{I}_q$ is defined as

$$\tilde{I}_q = \begin{cases} g(I_q, b), & d = \mathsf{REGION}, \\ I_q, & d = \mathsf{FULL}. \end{cases} \tag{1}$$

The decision d and bounding box b are generated by a vision-language model conditioned on the query image and question.

3.3 Scoring Model

Images and text are embedded into a shared representation space using a pre-trained vision-language encoder. Let $f_I(\cdot)$ and $f_T(\cdot)$ denote the image and text encoders, respectively.

The query embedding is computed as

$$\mathbf{v}_q(d, b) = \frac{f_I(\tilde{I}_q)}{\left\| f_I(\tilde{I}_q) \right\|}. \tag{2}$$

For each candidate $c_j \in \mathcal{C}$, the candidate embedding is defined as

$$\mathbf{v}_j = \frac{f_I(I_j) + [t_j \neq \emptyset]\ f_T(t_j)}{\left\| f_I(I_j) + [t_j \neq \emptyset]\ f_T(t_j) \right\|} \tag{3}$$

Method	E-VQA						InfoSeek					
	MRR	NDCG	R@1	R@5	R@10	R@20	MRR	NDCG	R@1	R@5	R@10	R@20
EVA-CLIP	0.224	0.289	14.2	33.4	47.4	50.8	0.553	0.614	46.3	71.2	77.3	81.7
ReflectiVA	-	-	15.6	36.1	-	49.8	-	-	56.1	77.6	-	**86.4**
Wiki-LLaVA	-	-	3.3	-	9.9	13.2	-	-	36.9	-	66.1	77.9
mR^{2}AG	-	-	-	-	-	-	-	-	38.0	-	65.0	71.0
Random	0.060	0.160	1.2	5.9	20.7	50.8	0.193	0.323	10.6	27.1	40.7	81.7
Center	0.143	0.220	7.5	20.1	31.5	50.8	0.468	0.545	38.8	56.8	66.6	81.7
Qwen2.5-3B	0.231	0.293	15.3	34.1	44.3	50.8	0.582	0.618	54.1	68.8	74.0	81.7
Qwen2.5-7B	0.240	0.300	16.1	35.2	45.6	50.8	0.598	0.638	54.8	70.5	75.1	81.7
EchoSight	0.402	0.423	36.5	47.9	48.8	48.8	0.586	0.631	53.2	74.0	77.4	77.9
OMGM	**0.473**	**0.500**	42.8	**55.7**	**58.1**	**58.7**	0.681	0.717	64.0	**80.8**	**83.6**	84.8
Ours	**0.473**	0.480	**44.7**	48.2	49.9	50.8	**0.706**	**0.732**	**66.5**	78.2	80.2	81.7

Table 1: Re-ranking results on E-VQA and InfoSeek with top-20 candidates. We report MRR, NDCG, and Recall@$\{1, 5, 10, 20\}$. Best in bold, second-best underlined.

Methods	E-VQA			InfoSeek		
	CondR@1	CondR@5	CondR@10	CondR@1	CondR@5	CondR@10
EVA-CLIP (Sun et al., 2023)	0.28	0.66	0.93	0.57	0.87	0.95
ReflectiVA (Cocchi et al., 2025)	0.31	0.72	-	0.65	0.90	-
Wiki-LLaVA (Caffagni et al., 2024)	0.25	-	0.75	0.47	-	0.85
mR^{2}AG (Zhang et al., 2024)	-	-	-	0.54	-	0.92
Random	0.02	0.12	0.41	0.13	0.33	0.49
Center	0.15	0.40	0.62	0.47	0.70	0.82
Qwen2.5-3B (Bai et al., 2025a)	0.30	0.67	0.87	0.69	0.84	0.91
Qwen2.5-7B (Bai et al., 2025a)	0.32	0.69	0.90	0.70	0.86	0.92
EchoSight (Yan and Xie, 2024)	0.75	**0.98**	**1.00**	0.68	0.95	**0.99**
OMGM (Yang et al., 2025)	0.73	0.95	0.99	0.75	0.95	**0.99**
Ours	**0.90**	0.95	0.98	**0.81**	**0.96**	0.98

Table 2: **Conditional recall** on E-VQA and InfoSeek. CondR@K denotes CondRecall@K (Eq. 10). Best in bold, second-best underlined.

Candidate relevance scores are computed using cosine similarity

$$s_j(d, b) = \cos\big(\mathbf{v}_q(d, b),\ \mathbf{v}_j\big). \tag{4}$$

Sorting $\{s_j(d, b)\}_{j=1}^{N}$ in descending order yields a ranking $\pi(d, b)$.

3.4 Policy Learning and Reward Design

We fine-tune the vision-language model as a stochastic policy $\pi_\theta(a \mid x)$ that outputs a region cropping action $a = (d, b)$ conditioned on the query $x = (I_q, q)$. The discrete decision is $d \in \{\mathsf{REGION}, \mathsf{FULL}\}$; when $d = \mathsf{REGION}$, the model additionally predicts a bounding box $b = (x_1, y_1, x_2, y_2)$. Given a, we compute the ranking $\pi(d, b)$ using the fixed scoring model in Sec. 3.3.

Reward from re-ranking improvement. To explicitly optimize re-ranking quality, we define rewards using *improvements* over the full-image baseline. Let $\pi_{\mathsf{full}} = \pi(\mathsf{FULL}, \cdot)$ be the baseline ranking and $\pi_a = \pi(d, b)$ be the ranking induced by action a. We compute MRR and NDCG over the candidate pool, and define the corresponding deltas:

$$\begin{aligned} \Delta\mathrm{MRR} &= \mathrm{MRR}(\pi_a) - \mathrm{MRR}(\pi_{\mathsf{full}}), \\ \Delta\mathrm{NDCG} &= \mathrm{NDCG}(\pi_a) - \mathrm{NDCG}(\pi_{\mathsf{full}}). \end{aligned} \tag{5}$$

To further stabilize training, we include two auxiliary improvement signals. Let $\mathrm{rank}(\pi) = \min\{r : y_{\pi(r)} = 1\}$ denote the rank of the positive candidate. Let $\mathrm{pos} = \max_{j:y_j=1} s_j$ and $\mathrm{neg} = \max_{j:y_j=0} s_j$ be the scoring of positive and negative candidates with their maximum score, and

define $\mathrm{margin}(\pi) = \mathrm{pos} - \mathrm{neg}$. We then define:

$$\Delta\mathrm{Rank} = \log \frac{\mathrm{rank}(\pi_{\mathsf{full}}) + 1}{\mathrm{rank}(\pi_a) + 1)}, \qquad (6)$$
$$\Delta\mathrm{Margin} = \mathrm{margin}(\pi_a) - \mathrm{margin}(\pi_{\mathsf{full}}),$$

where the margin improvement term to provide a more direct training signal for *discriminability* in the similarity space. Ranking metrics such as MRR and NDCG depend on the induced ordering and can be relatively insensitive when many candidates receive similar scores or when small score changes do not alter the rank. In contrast, maximizing $\Delta\mathrm{Margin}$ explicitly encourages the policy to increase the similarity between the query representation and the positive candidate while reducing the similarity to the most competitive negative candidates.

Eventually, when $d = \mathsf{REGION}$, the reward is a weighted combination of these improvements:

$$\begin{aligned} r(x, a) = {} & w_1\, \Delta\mathrm{MRR} + w_2\, \Delta\mathrm{NDCG} \\ & + w_3\, \Delta\mathrm{Rank} + w_4\, \Delta\mathrm{Margin} \\ & - \ \eta(b), \end{aligned} \qquad (7)$$

where $\sum_{i=1}^{4} w_i = 1$ are scalar weights. $\eta(b)$ penalizes malformed boxes (e.g., $x_2 \le x_1$ or $y_2 \le y_1$).

When $d = \mathsf{FULL}$, we reward only if the baseline already ranks positive candidate at the rank 1, indicating the model makes the correct decision.

$$r(x, a) = \mathbb{1}[\mathrm{rank}(\pi_{\mathsf{full}}) = 1]\,, \qquad d = \mathsf{FULL}. \qquad (8)$$

Optimization with r-GRPO. We optimize $\pi_\theta(a \mid x)$ using a region-aware variant of GRPO (Shao et al., 2024), which we denote **r-GRPO**. For each query x, we sample a group of N actions $\{a^{(n)}\}_{n=1}^{N} \sim \pi_\theta(\cdot \mid x)$ and compute rewards $r^{(n)} = r(x, a^{(n)})$. We form normalized advantages within the group

$$A^{(n)} = \frac{r^{(n)} - \mu_r}{\sigma_r + \epsilon} \qquad (9)$$

where μ_r and σ_r are the mean and standard deviation of $\{r^{(n)}\}_{n=1}^{N}$.

A challenge in our setting is that the action space is structured as $a = (d, b)$, where d is discrete while b is continuous and only meaningful when $d = \mathsf{REGION}$. Naively applying GRPO can suffer from high variance and unstable updates when the sampled group is dominated by one decision.

To address this, r-GRPO uses decision-balanced group sampling. For each query x, we sample a group of N actions $\{a^{(n)}\}_{n=1}^{N}$ from $\pi_\theta(\cdot \mid x)$, while *ensuring both decisions appear in the group.* This strategy prevents the more frequent decision from setting the baseline for the less frequent one, yielding lower-variance advantages. By balancing decisions in the sampled group and normalizing advantages conditional on d, r-GRPO stabilizes training for structured region-cropping actions while retaining the original GRPO update form.

4 Experiments

4.1 Experimental Settings

Datasets. We conduct experiments on two challenging KB-VQA benchmarks, InfoSeek (Chen et al., 2023) and E-VQA (Mensink et al., 2023), which are commonly used to evaluate MM-RAG systems. Each dataset provides image-question pairs and a knowledge source containing multimodal evidence, which are entity-centric Wikipedia content with associated images. We use Qwen-2.5-VL-3B (Bai et al., 2025a) as the base model for our experiments.

Training. During training, we construct candidate pools on-the-fly for each query by retrieving the top-K candidates from the knowledge base using a pre-trained EVA-CLIP model. Each candidate contains the evidence image and its associated text. Following standard re-ranking practice, we retain only training queries for which at least one ground-truth relevant candidate appears in the retrieved top-K pool, as re-ranking cannot recover missing positives. We set $K = 20$ and retrieve the top-20 most relevant entities, balancing computational efficiency with retrieval effectiveness.

Testing and Evaluation. Following prior work (Yan and Xie, 2024; Yang et al., 2025), we evaluate on the original test splits of InfoSeek and E-VQA. For each test query, we retrieve a top-K candidates from the knowledge base. Re-ranking is then performed over this candidate pool.

We measure re-ranking performance using standard information retrieval metrics, including Mean Reciprocal Rank (MRR), Normalized Discounted Cumulative Gain (NDCG), and Recall@K. Recall@K reflects whether relevant candidate is ranked within the top-K results and is particularly relevant for MM-RAG that condition downstream generation on a small number of retrieved candidates. Since Recall@K can be influenced by retrieval coverage, which may not directly reflect the re-ranking performance, we additionally report *conditional* Recall@K to better isolate re-ranking quality. Let $\mathrm{hit@}K(i) \in \{0, 1\}$ indicate whether

Query (Full) *+ question*	**EchoSight** *rank #1*	**OMGM** *rank #1*	**Ours**	**GT Positive**
 ***Q:*** *What is the closest upper taxonomy of this insect?*	 Archigenes neophron, the tailed Judy, is a small butterfly found in India… Wrong	 *Eurytela hiarbas*, commonly known as the pied piper, is a butterfly of the family… Wrong	 Focuses on *the insect*	 *Lyssa zampa*, the tropical swallowtail moth or Laos brown butterfly… Correct
 ***Q:*** *Where is this animal native to?*	 The California toad (Anaxyrus boreas halophilus) is a subspecies of the western toad, along with the boreal toad… Wrong	 Couch's spadefoot toad or Couch's spadefoot (Scaphiopus couchii) is a species of North American… Wrong	 Focuses on *the animal*	 Scaphiopus holbrookii, commonly known as the eastern spadefoot, is a species of American spadefoot toad… Correct

Table 3: Qualitative examples. Left-to-right: original query, top-1 from EchoSight (Yan and Xie, 2024) and OMGM (Yang et al., 2025) *after re-ranking*, our query with region cropping, and the ground-truth positive candidate. Our region cropping flips the rank-1 prediction to the correct positive by removing distractors and emphasizing question-relevant region.

query i has positive candidate ranked within the top-K positions after re-ranking.

$$\begin{aligned}\text{CondRecall@}K &= \sum_{i \in \mathcal{Q}} \frac{\text{hit@}K(i)}{\text{hit@20}(i)} \\ &= \mathbb{E}[\text{hit@}K \mid \text{hit@20} = 1]\,. \end{aligned} \tag{10}$$

where $\mathcal{Q}$ denotes the test query set. Unless otherwise specified, same in training, we set $K = 20$ and report Recall@$\{1, 5, 10\}$ together with CondRecall@$\{1, 5, 10\}$ and Recall@20. An ideal re-ranker that always promotes positive candidate to the top-K whenever one is present in the top-20 achieves $\text{CondRecall@}K = 1$ for all $K \leq 20$.

4.2 Baselines

We compare our query-side region cropping approach against baselines that reflect common design choices in multi-modal re-ranking.

Heuristic Region Cropping. We evaluate two non-learned region cropping strategies. *Center Region* selects a fixed central region of the query image. *Random Region* selects a region at a random location, with selected area matched to the distribution produced by the learned policy.

Zero-shot VLM Region Cropping. A vision-language model is prompted to predict a question-conditioned region of interest for the query image. No fine-tuning or reinforcement learning is applied. In this study, we experiment with Qwen2.5-VL, with both 3B and 7B models.

Re-ranker Methods. We additionally compare against prior studies that optimizes the re-ranker trained with hard-negative samples, including EchoSight (Yan and Xie, 2024) and OMGM (Yang et al., 2025), which improve multi-modal re-ranking by modifying the re-ranking model or query-candidate interaction mechanism rather than controlling the query representation.

4.3 Main Results

Tables 1 and 2 summarize retrieval/re-ranking performance across different approaches. On InfoSeek, our learned *query-side region cropping* consistently improves ranking quality over the *No Region Cropping* baseline, with clear gains in top-heavy metrics. Notably, the improvements concentrate on bringing the correct evidence to the very top of the list. Our method increases the frequency of ranking a positive candidate at rank #1, indicating more effective ordering within the same candidate

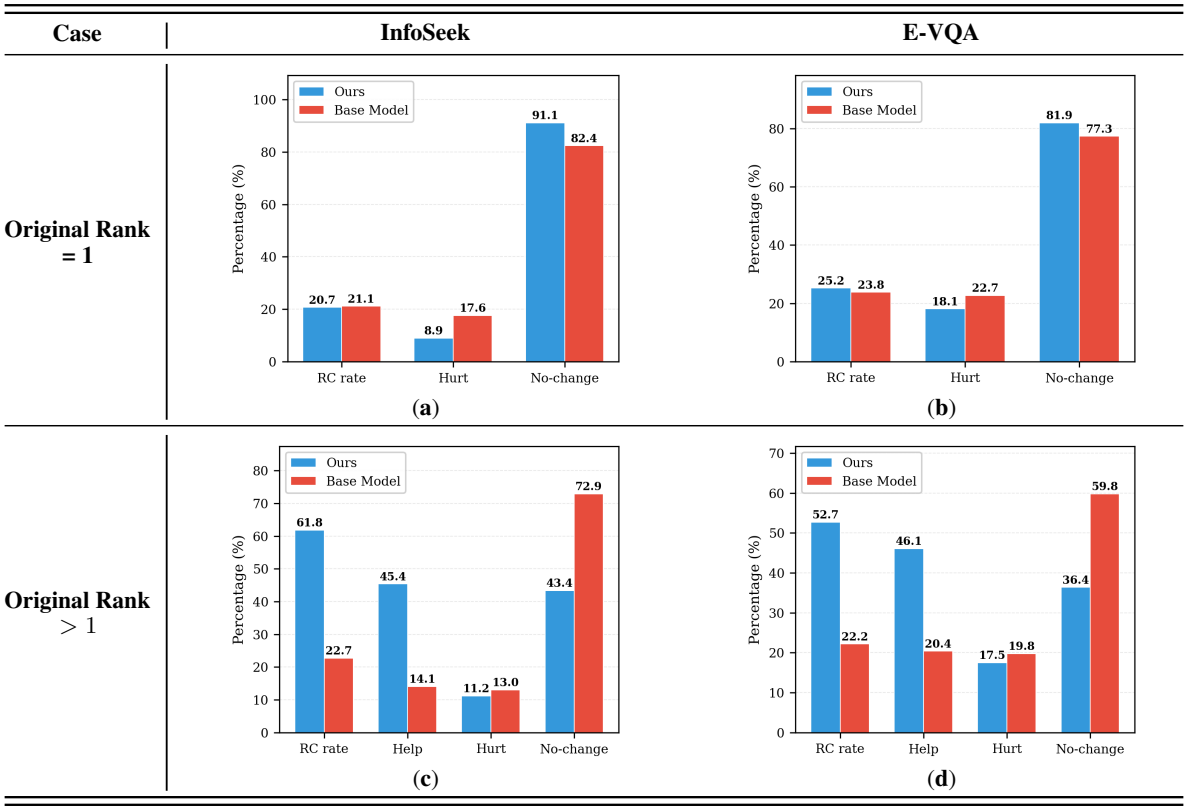

Table 4: **Quantitative results of our region cropping (RC) behavior across two datasets.** (a,b) show cases where the original query already ranks the positive at #1 on InfoSeek and E-VQA, respectively. (c,d) show harder cases where the original rank is > 1. *RC rate* indicates how often does the model choose the crop a region.

set.

Heuristic crops are brittle and can easily discard the evidence needed for matching. Random cropping performs poorly and a center crop only partially mitigates this effect. In contrast, zero-shot VLM-based region cropping provides a stronger baseline than heuristics, but still falls short of the learned policy, A key reason is that the zero-shot VLM *rarely commits to cropping*. Table 4 shows that the base model has a low RC rate, leaving the query unchanged for roughly 80% of test examples, effectively behaving like the no-cropping baseline, suggesting that *directly optimizing the cropping decisions for the re-ranking objective* is necessary.

Compared with prior retrieval/re-ranking methods, our approach achieves the strongest overall top-rank behavior with best MRR/NDCG and the highest Recall@1, while remaining competitive at larger cutoffs. This trend is further supported by conditional recall in Table 2. Our method attains the best CondR@1/CondR@5, confirming that when relevant evidence exists in the candidate pool, region cropping helps the model reliably surface it in the top positions.

Table 3 presents qualitative comparisons on different queries where the retrieved pool already contains the correct evidence, yet competing methods still place a distractor at rank #1. In these cases, full image is dominated by irrelevant but visually salient regions, such as hands or leaves, causing the re-ranker to over-score mismatched candidates. By selecting a question-conditioned region before scoring, our method suppresses these distractors and shifts similarity toward the correct visual evidence to rank #1.

4.4 Analysis

Does the model perform region cropping (RC) correctly? Region cropping is not universally beneficial. Suppressing distractive regions can sharpen cross-modal similarity, but can also remove contextual cues that are necessary for disambiguation. Therefore, a well-behaved RS policy should (i) avoid altering the query when the full-image representation is already sufficient, and (ii) perform RS when the full-image representation is likely contaminated by distractors. To demonstrate whether our learned policy exhibits this capability, we analyze the RS behavior for two case: *Original Rank = 1* and *Original Rank > 1*. Table 4 presents how often the model chooses region cropping (*RC rate*) and the resulting effect on the ranking, where

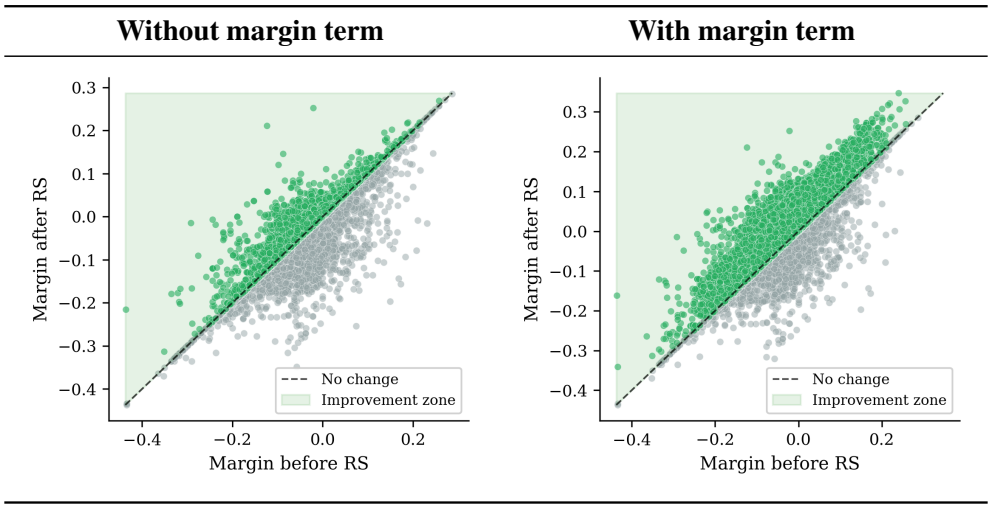

Figure 2: Ablation on the margin term in the reward design.

Help, *Hurt*, and *No-change* means the percentage of rank improves, degrades, and unchanged through test samples, correspondingly.

When the positive candidate is already at rank #1, the model is expected to select region safely. Table 4(a) and (b) show our learned policy does not simply disable RC, but it is less destructive than base model. Harmful outcomes decrease while *No-change* remains dominant.

When the positive candidate is not at rank #1, the model is expected to crop a region and suppress distractors. Table 4(c) and (d) show our model performs RC substantially more often than the base model, yielding a higher *Help* rate without increasing *Hurt*. In contrast, the base model remains biased toward *No-change*, which limits its ability to correct hard cases. Overall, Table 4 suggests that our learned policy acquires the capability of *when to select region*.

Ablation study on reward function To isolate the contribution of components in the reward function, we train the policy under progressively richer reward variants while keeping the scoring model, candidate pool construction, and all hyperparameters fixed. Table 5 reports MRR on test sets of InfoSeek and E-VQA. From the table, ranking-metric deltas yield only marginal improvements, whereas introducing the margin term produces a huge jump.

To further understand this effect, we analyze how RC changes the score separation between the true positive and the most competitive negative, i.e. the margin defined in Eq. 7. Figure 2 visualizes the margin before RC versus after RC. Without the margin term (left), most samples lie close to the no-change diagonal, with a substantial portion falling below it. In contrast, adding the margin term (Right) produces a clear shift. A much larger fraction of samples moves into the improvement zone above the diagonal, meaning that RC increasingly improves the margin rather than merely perturbing ranks. This supports the intuition that ranking-metric deltas provide sparse supervision, where small score changes may not flip an ordering. whereas ΔMargin directly encourages the policy to pull the positive closer while pushing the hardest negative away. This study explains the substantial performance jump observed when the margin term is included.

Reward Component	InfoSeek	E-VQA
ΔMRR only	0.611	0.408
+ ΔNDCG	0.613 (↑)	0.425 (↑)
+ ΔRank	0.613 (-)	0.426 (↑)
+ ΔMargin (Full)	**0.706**	**0.473**

Table 5: **Ablation on reward design.** Each row adds one component to the region-cropping reward (Eq. 7).

5 Conclusion

Region-R1 demonstrates that re-ranking can be improved substantially by *thinking the query image representation*, rather than only building heavier re-rankers. Our approach consistently outperforms prior multi-modal re-rankers across two challenging datasets, with particularly strong gains in retrieving the correct evidence at rank #1. Future directions include adapting richer query, such as selecting multiple regions or learning soft, question-conditioned attention over the image. Another promising direction is to extend query-side adaptation from re-ranking to earlier retrieval stages under efficiency constraints, enabling end-to-end improvements while preserving scalibility.

6 Limitations

Region-R1 operates strictly at the re-ranking stage and therefore cannot recover evidence that is absent from the retriever's top-K candidate pool, so end performance remains bounded by retriever recall. In addition, our model is trained to optimize a fixed similarity-based objective, which may bias the learned region cropping behavior toward the scorer's inductive biases and reduce transferability to other re-rankers or objectives. Furthermore, we evaluate on a limited set of benchmarks and a fixed re-ranking regime. While inference remains bounded by top-K, the method introduces additional decision/cropping overhead and RL-style training can be sensitive to reward design and hyperparameters.

References

Omar Adjali, Olivier Ferret, Sahar Ghannay, and Hervé Le Borgne. 2024. Multi-level information retrieval augmented generation for knowledge-based visual question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16499–16513. Association for Computational Linguistics.

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

Yang Bai, Christan Grant, and Daisy Zhe Wang. 2025b. Ramqa: A unified framework for retrieval-augmented multi-modal question answering. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1061–1076.

Andrea Burns, Krishna Srinivasan, Joshua Ainslie, Geoff Brown, Bryan A Plummer, Kate Saenko, Jianmo Ni, and Mandy Guo. 2023. Wikiweb2m: A page-level multimodal wikipedia dataset. *arXiv preprint arXiv:2305.05432*.

Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1818–1826.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. 2022. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14948–14968, Singapore. Association for Computational Linguistics.

Changin Choi, Wonseok Lee, Jungmin Ko, and Wonjong Rhee. 2025. Multimodal iterative rag for knowledge-intensive visual question answering. *arXiv preprint arXiv:2509.00798*.

Federico Cocchi, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2025. Augmenting Multimodal LLMs with Self-Reflective Tokens for Knowledge-based Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Alberto Compagnoni, Marco Morini, Sara Sarto, Federico Cocchi, Davide Caffagni, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2025. Reag: Reasoning-augmented generation for knowledge-based visual question answering. *arXiv preprint arXiv:2511.22715*.

Nick Craswell. 2016. Mean reciprocal rank. In *Encyclopedia of database systems*, pages 1–1. Springer.

François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. 2020. Conceptbert: Concept-aware representation for visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 489–498.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. 2023a. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. *arXiv preprint arXiv:2302.11154*.

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. 2023b. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23369–23379.

Guohao Li, Xin Wang, and Wenwu Zhu. 2020. Boosting visual question answering with context-aware knowledge aggregation. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1227–1235.

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

Weizhe Lin and Bill Byrne. 2022. Retrieval augmented visual question answering with outside knowledge. *arXiv preprint arXiv:2210.03809*.

Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. 2023. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems*, 36:22820–22840.

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

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/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3090–3101.

Fengran Mo, Kelong Mao, Yutao Zhu, Yihong Wu, Kaiyu Huang, and Jian-Yun Nie. 2023. ConvGQR: Generative query reformulation for conversational search. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4998–5012, Toronto, Canada. Association for Computational Linguistics.

Medhini Narasimhan, Svetlana Lazebnik, and Alexander Schwing. 2018. Out of the box: Reasoning with graph convolution nets for factual visual question answering. *Advances in neural information processing systems*, 31.

Kimia Ramezan, Alireza Amiri Bavandpour, Yifei Yuan, Clemencia Siro, and Mohammad Aliannejadi. 2025. Multi-turn multi-modal question clarification for enhanced conversational understanding. *arXiv preprint arXiv:2502.11442*.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

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

Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. 2021. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12884–12893.

Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. 2013. A theoretical analysis of ndcg type ranking measures. In *Conference on learning theory*, pages 25–54. PMLR.

Navve Wasserman, Oliver Heinimann, Yuval Golbari, Tal Zimbalist, Eli Schwartz, and Michal Irani. 2025. Docrerank: Single-page hard negative query generation for training multi-modal rag rerankers. *arXiv preprint arXiv:2505.22584*.

Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. 2020. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584.

Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. 2022. Multi-modal answer validation for knowledge-based vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2712–2721.

Yibin Yan and Weidi Xie. 2024. EchoSight: Advancing visual-language models with Wiki knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1538–1551, Miami, Florida, USA. Association for Computational Linguistics.

Wei Yang, Jingjing Fu, Rui Wang, Jinyu Wang, Lei Song, and Jiang Bian. 2025. OMGM: Orchestrate multiple granularities and modalities for efficient multimodal retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24545–24563, Vienna, Austria. Association for Computational Linguistics.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Retrieval-augmented multimodal language modeling. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. *Advances in Neural Information Processing Systems*, 37:121156–121184.

Tao Zhang, Ziqi Zhang, Zongyang Ma, Yuxin Chen, Zhongang Qi, Chunfeng Yuan, Bing Li, Junfu Pu, Yuxuan Zhao, Zehua Xie, and 1 others. 2024. mr2ag: Multimodal retrieval-reflection-augmented generation for knowledge-based vqa. *arXiv preprint arXiv:2411.15041*.

Tianshuo Zhou, Sen Mei, Xinze Li, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, Yu Gu, and Ge Yu. 2024. Marvel: unlocking the multi-modal capability of dense retrieval via visual module plugin. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14608–14624.

Hongyi Zhu, Jia-Hong Huang, Stevan Rudinac, and Evangelos Kanoulas. 2024. Enhancing interactive image retrieval with query rewriting using large language models and vision language models. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 978–987.

Zihao Zhu, Jing Yu, Yujing Wang, Yajing Sun, Yue Hu, and Qi Wu. 2020. Mucko: Multi-layer cross-modal knowledge reasoning for fact-based visual question answering. *arXiv preprint arXiv:2006.09073*.

A Details of Dataset

Encyclopedic VQA (Mensink et al., 2023) E-VQA is an encyclopedic visual question answering benchmark built around fine-grained entity recognition and evidence-grounded answering. It contains approximately 221K question-answer pairs spanning 16.7K distinct entities, where each entity is associated with up to five images. The entity inventory and images are sourced from iNaturalist 2021 (Van Horn et al., 2021) and Google Landmarks Dataset V2 (Weyand et al., 2020). To support evidence-based reasoning, E-VQA additionally provides a controlled knowledge base derived from WikiWeb2M (Burns et al., 2023), consisting of roughly 2M Wikipedia articles with images, from which supporting evidence for each answer can be retrieved. Questions are categorized by reasoning complexity into single-hop and two-hop variants. The dataset is organized into training/validation/test splits with approximately 1M, 13K, and 5.8K samples, respectively. Following common practice for comparable evaluation, we restrict both training and testing to the single-hop subset.

InfoSeek (Chen et al., 2023) comprises 1.3M image-question-answer triplets grounded in visual entities, covering roughly 11K entities from OVEN (Hu et al., 2023a). It includes 8.9K human-authored visual information-seeking questions alongside about 1.3M automatically generated questions. The official split contains approximately 934K/73K/348K samples for train/validation/test, respectively. Because the test split does not provide ground-truth answers, we report results on the validation set. Notably, both validation and test emphasize generalization by featuring unseen entities and queries not observed during training. InfoSeek is accompanied by a large knowledge base of about 6M Wikipedia entities. To align with prior work (Yan and Xie, 2024; Yang et al., 2025), we adopt the same evaluation setting with a 100K-entity subset that still includes the 6,741 entities referenced by the training and validation questions. Following this standard protocol, we evaluate on 71,335 validation samples.

B Implementation Details

We instantiate the region cropping policy from `Qwen2.5-VL-3B-Instruct`. Given a query image, the policy predicts either Full (keeping the original query representation) or REGION, which outputs a single 2D bounding box that specifies a region used to form an alternative query view for subsequent re-ranking.

For scoring and training feedback, we use `BAAI/EVA-CLIP-8B` (Sun et al., 2023) to compute cosine-similarity scores between the query image and candidate evidence images, inducing the ranking used for both optimization and evaluation. When candidate-side text is available, we fuse it on the candidate side by combining candidate image and text embeddings before normalization. We train the policy with group-based RL using $N{=}8$ sampled outputs per query, learning rate $5e{-}5$ with a cosine schedule and 50 warmup steps, batch size 4 per device, and 2 epochs; we cap the maximum prompt length at 8192 tokens and generate up to 256 new tokens. We adopt parameter-efficient tuning with LoRA applied to all linear layers of the policy backbone ($r{=}8$, $\alpha{=}32$, dropout 0.1), updating only the LoRA parameters. We trained the model with 4 A6000 GPUs with 48GB VRAM. In addition, we provide the system prompt that we used to train the model in Table 6.

C Licenses

The datasets we used, InfoSeek and E-VQA, are licensed under Apache License 2.0 and CC BY 4.0, respectively. The scoring model, EVA-CLIP-8B, is licensed under MIT License. The prior re-ranker models from EchoSight and OMGM were released without an accompanying license. Qwen-2.5-VL series models are licensed under Apache License 2.0.

Table 6: System Prompt

System Prompt

```
You are an intelligent region cropping assistant for re-ranking tasks.

Given an image and the user's question, your task is to decide whether finding a
    specific region would help remove the redundant information and rank more
    relevant information to a higher rank.

# Instructions
1. Carefully analyze the image in the context of the user's question.
2. Based on the user's question, decide whether region selection would improve re-
    ranking accuracy among many candidates, and push the most relevant candidate to
    the rank 1. You can decide:
   - If you think a specific region is most relevant to answering the question,
       select it to focus on it.
   - If you think the full image is already optimal for the question, output "FULL".
3. Output your decision with "FULL" or "REGION", if your decision is "REGION", you
    have to specify a region.

# Tools
<tools>
{"type":"function","function":{"name":"image_zoom_in_tool","description":"Zoom in on
     a specific region of an image.","parameters":{"type":"object","properties":{"
    bbox_2d":{"type":"array","items":{"type":"number"},"minItems":4,"maxItems":4,"
    description":"Bounding box as [x1, y1, x2, y2], where (x1, y1) is top-left and (
    x2, y2) is bottom-right."},"label":{"type":"string"}},"required":["bbox_2d"]}}}
</tools>

# How to call a tool
Return a json object with function name and arguments within <tool_call></tool_call>
     XML tags:
<tool_call>
{"name": <function-name>, "arguments": <args-json-object>}
</tool_call>

# Output format
{"Decision": "FULL" or "REGION", "Tool": <Call the tool with parameters if the
    decision is "REGION">}


Again, if you call the tool, you MUST follow the format exactly as specified.

Otherwise, I will be unable to parse your response.
```