X^R: Cross-Modal Agents for Composed Image Retrieval

Zhongyu Yang
BCML, Heriot-Watt University
Edinburgh, UK
zy4028@hw.ac.uk

Wei Pang
BCML, Heriot-Watt University
Edinburgh, UK
w.pang@hw.ac.uk

Yingfang Yuan*
BCML, Heriot-Watt University
Edinburgh, UK
y.yuan@hw.ac.uk

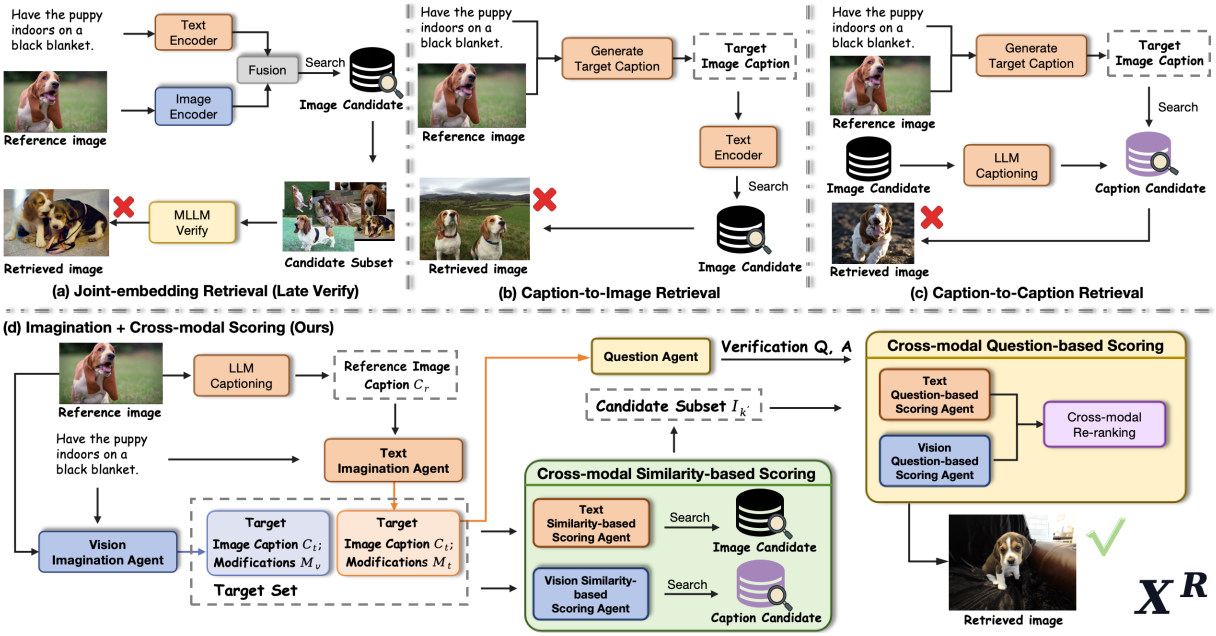


Figure 1: The workflows of existing CIR methods and ours: (a) *Joint embedding–based* methods encode a multimodal query into a shared space, but they often struggle to capture complex text-specified edits. (b) *Caption-to-Image* methods first generate a target caption from the multimodal query prior to retrieval, but they often fail to preserve fine-grained details. (c) *Caption-to-Caption* methods build upon Caption-to-Image but restrict comparison to the text space, thereby discarding visual cues. (d) X^R (ours) introduces an agentic AI framework with cross-modal agents and a progressive retrieval process consisting of an imagination stage followed by coarse-to-fine filtering, enabling robust reasoning that better aligns results with user intent.

Abstract

Retrieval is being redefined by agentic AI, demanding multimodal reasoning beyond conventional similarity-based paradigms. Composed Image Retrieval (CIR) exemplifies this shift as each query combines a reference image with textual modifications, requiring compositional understanding across modalities. While embedding-based CIR methods have achieved progress, they remain narrow in perspective, capturing limited cross-modal cues and lacking semantic reasoning. To address these limitations, we introduce X^R, a training-free multi-agent framework that reframes retrieval as a progressively coordinated reasoning process. It orchestrates three specialized types of agents: *imagination agents* synthesize target representations through cross-modal generation, *similarity agents* perform coarse filtering via hybrid matching, and *question agents* verify factual consistency through targeted reasoning for fine filtering. Through progressive multi-agent coordination, X^R iteratively refines retrieval to meet both semantic and visual query constraints, achieving up to a 38% gain over strong training-free and training-based baselines on FashionIQ, CIRR, and CIRCO, while ablations show each agent is essential. Code is available: https://01yzzyu.github.io/xr.github.io/.

CCS Concepts

• **Information systems** → **Information retrieval**; **Retrieval models and ranking**; **Users and interactive retrieval**.

Keywords

Compose Image Retrieval, Agents, Cross-modality

ACM Reference Format:
Zhongyu Yang, Wei Pang, and Yingfang Yuan. 2026. X^R: Cross-Modal Agents for Composed Image Retrieval. In *Proceedings of the ACM Web Conference 2026 (WWW '26), April 13–17, 2026, Dubai, United Arab Emirates.* , 12 pages. https://doi.org/10.1145/3774904.3792276

*Corresponding author

WWW '26, Dubai, United Arab Emirates

ACM ISBN 979-8-4007-2307-0/2026/04
https://doi.org/10.1145/3774904.3792276

1 Introduction

Composed Image Retrieval (CIR) [4, 7, 17, 30, 42] is a retrieval paradigm where a query is explicitly composed by the user through a reference image and a modification text. CIR queries embody specific intent through the controlled composition of image and text. This not only establishes CIR as a new direction in web information access, where users refine searches by combining images and text, but also links it to broader developments in retrieval-augmented agentic AI. The demand for such interaction is evident in applications such as e-commerce [6, 56] and search engines [46], where navigating massive image repositories requires fine-grained multimodal control. Compared with conventional retrieval [11, 14, 29, 43], CIR is particularly challenging because it requires cross-modal reasoning to integrate heterogeneous signals rather than relying on a single unimodal cue. By pairing a reference image with textual modifications, CIR moves retrieval beyond simple content matching toward retrieving images that preserve reference semantics while faithfully applying the edits.

As illustrated in Figure 1, existing approaches can be broadly grouped into three categories: **(a) Joint embedding-based** methods project the multimodal query into a vector space and formulate CIR as similarity-based matching; **(b) Caption-to-Image** methods first generate a target caption based on the multimodal query, then embed it and compare it with candidate image embeddings in terms of similarity; and **(c) Caption-to-Caption** methods that directly compare candidate captions with the target caption. Despite notable progress, these approaches exhibit persistent limitations. First, joint embeddings struggle to capture fine-grained, edit-specific correspondences due to imperfect cross-modal alignment. Second, a single similarity-based matching approach may fail to capture both textual and visual evidence. Third, the absence of cross-modal verification-based refinement undermines reliability, as each modality provides important information.

These challenges highlight our motivation that effective CIR must fully exploit cross-modal interactions. To address these challenges, we propose X^R, a training-free multi-agent framework that explicitly orchestrates cross-modal reasoning, providing robust retrieval under heterogeneous signals. X^R consists of three sequential modules: imagination, coarse filtering, and fine filtering. In imagination, agents construct a target proxy by generating captions from two cross-modal pairings, namely modification text with the reference image caption and modification text with the reference image, which helps reduce modality gaps and anchor the target semantics. In coarse filtering, similarity-based agents evaluate candidates by producing multi-perspective scores using visual and textual cues, each conditioned on cross-modal captions. Reciprocal Rank Fusion (RRF) then aggregates these scores to form an initial ranked subset that addresses the limitations of single-criterion matching. In fine filtering, question-based agents re-evaluate this subset through cross-modal factual verification by testing candidate images and captions with predicate-style queries, which mimic how humans validate retrieval consistency. Finally, verification scores are integrated with similarity scores through re-ranking to produce the final retrieval set, benefiting from both similarity-based matching and factual verification.

The similarity-based and question-based agents play complementary roles, where the former enables efficient high-level retrieval for broad coverage, while the latter enforces factual verification to refine results for accuracy. This design preserves diverse sources of evidence that single-score pipelines would otherwise overlook. Moreover, the cross-modality employed in both agents within X^R enhances reliability by providing multi-perspective evidence. This is achieved through a combination of implicit coupling and explicit decoupling of modalities, enabling effective integration while maintaining per-modality interpretability. The proposed framework is tailored for edit-sensitive compositionality, capturing fine-grained modifications beyond the capability of unimodal systems.

We evaluate X^R on three CIR benchmarks, CIRR, CIRCO, and FashionIQ, covering diverse retrieval scenarios from controlled reference-based queries to open-domain compositional settings. Across all datasets, X^R consistently improves edit-sensitive retrieval accuracy over strong training-free and training-based baselines, demonstrating both its effectiveness and generality. These results suggest practical value for web systems and applications, including personalized e-commerce search and multimodal recommendation.

In summary, our contributions are as follows:

- We propose X^R, a training-free framework that orchestrates multiple cross-modal agents for CIR.
- We demonstrate the necessity of explicit cross-modality by showing its advantage over unimodal and single-score pipelines, which fail on edit-sensitive reasoning.
- Extensive experiments on CIRR, CIRCO, and FashionIQ show consistent gains over strong baselines, with ablations substantiating each module's role, positioning X^R as a general paradigm for multimodal retrieval.

2 Related Works

Multimodal Agent Systems: The rapid progress in MLLMs [15, 16, 33, 44] has enabled agentic frameworks with emerging abilities in autonomous planning, tool use, and decision-making. Such frameworks show strong potential for decomposing complex reasoning and coordinating across modalities [37, 49, 50, 53, 55], though coordination remains fragile in practice. By iterative reflection and collaborative strategies, they mitigate hallucination and enhance interpretability, outperforming single-pass inference albeit at higher cost [28, 32, 34, 52]. Yet most multimodal agents operate under a closed-world assumption, relying solely on internal inference and often hallucinating unsupported content [10, 20], reflecting a lack of external grounding. In contrast, retrieval has long served as grounding in NLP pipelines, reducing uncertainty and improving adaptability [2, 26, 38]. Recent studies on retrieval-augmented agents, such as Storm [21, 36] and WikiAutoGen [48] highlight the promise of retrieval-augmented agents, yet remain limited: Storm is text-centric, while WikiAutoGen extends to multimodality but in a narrow scope. Overall, these findings underscore retrieval as a key enabler of reasoning, yet a systematic integration into general multimodal agents remains missing.

Composed Image Retrieval: CIR provides a natural testbed for retrieval-enhanced reasoning, where the task is to locate a target image given a reference image and textual modifications [39, 54]. Most existing methods fuse features into a joint embedding and rank

candidates by similarity, achieving coarse alignment but blurring fine-grained changes [7, 23, 42]. Training-based models enhance representations but demand costly supervision and frequent retraining [3, 5, 18, 47]. Training-free approaches avoid task-specific supervision and generalize across domains. However, they rely on static fusion and one-shot pipelines, with little flexibility to refine uncertain retrieval results (e.g., candidate images or captions) [8, 24, 25]. Reasoning-style retrieval has been explored [13, 40, 41], but existing methods remain fixed templates rather than adaptive workflows. In practice, models still fail on fine-grained edits, for example, misinterpreting color changes in FashionIQ or mismatching object replacements in CIRR, underscoring the persistent limits of static similarity matching. This indicates that static similarity is not only brittle to edits but also fundamentally unable to capture compositional semantics.

In short, multimodal agents excel at reasoning but underexploit retrieval, while CIR methods leverage retrieval but lack reasoning, leaving the two largely disconnected. X^{R} bridges this gap by embedding retrieval within an agentic workflow: (1) imagination agents approximate the target image, preserving fine-grained details often missed by embeddings; (2) similarity-based agents score candidates across modalities, reducing the rigidity of one-shot pipelines; (3) question-based agents enforce factual checks, ensuring textual modifications are faithfully satisfied. Unified in a training-free system, these components elevate retrieval into dynamic reasoning, addressing the above limitations and yielding results that are both faithful to user intent and verifiable across modalities.

3 Method

3.1 Preliminaries

Given a multimodal query consisting of a reference image I_r and a modification text T_m, CIR assumes the existence of an ideal target image I_i that represents the desired outcome by preserving the visual characteristics of I_r while incorporating the modifications specified by T_m. The CIR task then aims to retrieve a subset of images $\mathcal{I}^* \subseteq \mathcal{I}$, where $\mathcal{I} = \{I_1, I_2, \ldots, I_N\}$ denotes the candidate image set containing N images. Each $I \in \mathcal{I}^*$ is expected to approximate the ideal target image I_i.

To obtain $\mathcal{I}^*$, CIR typically proceeds in two stages. In the *scoring* stage, each candidate image $I \in \mathcal{I}$ is evaluated against the query (I_r, T_m), which implicitly defines the ideal target image I_i. A matching score is then assigned, $S(I) = p(I \in \mathcal{I}^* \mid I_r, T_m)$, which represents the conditional probability that I belongs to the target set $\mathcal{I}^*$ given the query. This score can also be viewed as a similarity measure between I and the ideal target I_i. In the subsequent *ranking* stage, candidates are ordered by their scores, and the top-k images are selected to form $\mathcal{I}^*$, where $|\mathcal{I}^*| = k$.

3.2 X^{R} Framework

To address CIR, we propose X^{R}, a training-free multi-agent framework that emphasizes the role of cross-modality in improving retrieval accuracy. Unlike existing approaches, our framework is composed of three cross-modal modules: imagination, coarse filtering, and fine filtering. In the coarse filtering stage, similarity-based scoring agents identify an initial subset of candidates that approximate the ideal target image through similarity evaluation. In the fine filtering stage, question-based scoring agents perform factual verification to further filter and refine this subset, producing the final ordered set $\mathcal{I}^*$. These two components complement each other, working together to progressively narrow down candidates and improve ranking. Throughout the process, cross-modal mechanisms offer multi-perspective support that enhances robustness and reliability in retrieval while reducing the risks associated with multimodal misalignment.

The workflow of X^{R} is outlined in Algorithm 1 and depicted in Figure 1. The caption agent $\mathcal{A}_c$ generates a set of candidate captions C by iteratively producing a caption for each $I \in \mathcal{I}$ (Line 1). It also generates the caption C_r for the reference image I_r (Line 2). To construct or imagine the ideal target image I_i, the text imagination agent $\mathcal{A}_t^i$ and the vision imagination agent $\mathcal{A}_v^i$ jointly form the cross-modal imagination module. These two agents generate captions C_t and C_v, respectively, to describe I_i from different modalities. In addition, $\mathcal{A}_t^i$ produces M_t, which specifies the manipulations required to transform C_r with T_m in order to approximate I_i, thereby unifying the information in the text modality. Meanwhile, M_v denotes a set indicating the presence or absence of visual attributes.

In coarse filtering, to evaluate each candidate image I, X^{R} employs a text similarity-based scoring agent $\mathcal{A}_t^s$ (Line 6) and a vision similarity-based scoring agent $\mathcal{A}_v^s$ (Line 7). These agents assess the a-th candidate from different modalities, using C_a and I_a respectively, conditioned on C_t and C_v. Each scoring agent produces two scores, denoted s^t and s^v, through a process termed cross-modal multiperspective scoring, which operates within the embedding space based on similarity. The text-based score s^t and vision-based score s^v are then aggregated across agents in Lines 8 and 9. In Line 10, the score vectors $S^t = \{s_1^t, \ldots, s_N^t\}$ and $S^v = \{s_1^v, \ldots, s_N^v\}$ are processed separately using a Reciprocal Rank Fusion function. The resulting rank scores are combined, after which candidates are ranked and filtered to yield the top-k' results, denoted by $\mathcal{I}^{k'}$.

To enhance the ranking accuracy of the selected top-k' candidates and further refine this subset, X^{R} incorporates cross-modal question-based scoring agents in the fine filtering stage. The question agent $\mathcal{A}^q$ (Line 11) formulates a set of questions Q with corresponding answers A based on the information M_t, M_v, and T_m, focusing on the essential attributes that the ideal target image I_i should contain. Unlike the similarity-based scoring agents, $\mathcal{A}^q$ emphasizes the critical differences between a candidate image I and the ideal target I_i. Two question-based scoring agents, $\mathcal{A}_t^q$ and $\mathcal{A}_v^q$, then use Q and A to re-evaluate candidates in $\mathcal{I}^{k'}$ from the text and vision modalities, applying C_a and I_a respectively (Lines 13–14). Finally, the scores from the question-based scoring agents are combined with the aggregated similarity-based scores through a cross-modal re-ranking function (Line 15). The candidates are then re-ordered based on the resulting scores to form the fine-filtered subset, which constitutes the final output $\mathcal{I}^*$ of size k, where $k < k'$. The following paragraphs describe in detail the roles of imagination, coarse filtering, and fine filtering.

3.3 Imagination

The imagination agents $\mathcal{A}_t^i$ and $\mathcal{A}_v^i$ play a central role in X^{R}. In CIR, accurate retrieval requires a clear representation of the ideal target image I_i that satisfies the multimodal query, since defining

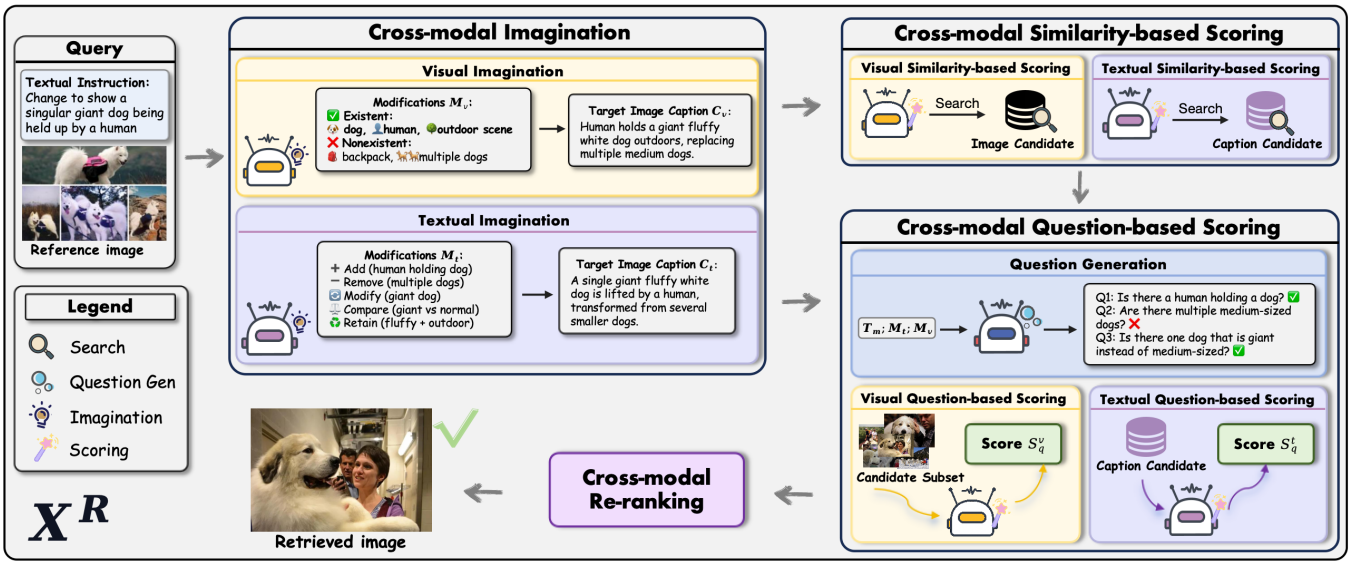


Figure 2: Framework of X^R. The multi-agent system integrates textual and visual imagination with cross-modal similarity and question-based scoring, followed by re-ranking. This multi-stage reasoning process exploits complementary cues from both modalities, effectively handling fine-grained modifications that single-modality approaches often miss.

Algorithm 1 X^R

Input: $\mathcal{I}$ cadidate image set, I_r reference image, T_m modification text
Output: $\mathcal{I}^*$ target image set

initialization
$C = \mathcal{A}_c(\mathcal{I})$ // candidate image caption set
$C_r = \mathcal{A}_c(I_r)$
imagination
$M_t, C_t = \mathcal{A}_t^i(T_m, C_r)$ // text imagination agent
$M_v, C_v = \mathcal{A}_v^i(T_m, I_r)$ // vision imagination agent
coarse filtering
for $a = 1$ **to** N **do** // $I_a \in \mathcal{I}, C_a \in C$
  $s_t^t, s_t^v = \mathcal{A}_t^s(C_t, C_v, C_a)$ // text similarity-based scoring agent
  $s_v^t, s_v^v = \mathcal{A}_v^s(C_t, C_v, I_a)$ // vision similarity-based scoring agent
  $s_a^t = s_t^t + s_v^t$
  $s_a^v = s_v^v + s_t^v$
$\mathcal{I}^{k'} = \text{ranking}(S^t, S^v)$ // reciprocal rank fusion function
fine filtering
$Q, A = \mathcal{A}^q(M_t, M_v, T_m)$ // question agent
for $a = 1$ **to** k' **do**
  $s_q^t = \mathcal{A}_t^q(C_a, Q, A)$ // text question-based scoring agent
  $s_q^v = \mathcal{A}_v^q(I_a, Q, A)$ // vision question-based scoring agent
$S^{k'} \leftarrow (S_q^t + S_q^v) * \text{norm}(\lambda S^t + (1-\lambda) S^v)$ // cross-modal re-ranking
$\mathcal{I}^* = \text{re-ranking}(S^{k'})$
return $\mathcal{I}^*$ // $|\mathcal{I}^*| = k$

I_i provides prior knowledge and evidence to guide the retrieval of similar candidates. This prior knowledge is critical, as incorrect priors can trigger a cascade of errors. To address this, the agents $\mathcal{A}_t^i$ and $\mathcal{A}_v^i$ are designed to imagine and approximate the ideal target image by generating cross-modal captions that capture complementary aspects of I_i. The cross-modality arises from the fact that $\mathcal{A}_v^i$ and $\mathcal{A}_t^i$ take different inputs: $\mathcal{A}_v^i$ uses the reference image I_r and $\mathcal{A}_t^i$ uses the reference caption C_r, and both are conditioned on the modification text T_m. Their outputs are the vision-based caption C_v and the text-based caption C_t, respectively.

The design of cross-modal imagination is motivated by the observation that information from different modalities provides complementary strengths and weaknesses when estimating I_i. Specifically, the pair (T_m, C_r) is more straightforward, which facilitates the extraction of key information for depicting I_i. In contrast, the pair (T_m, I_r) combines textual and visual inputs, where the image contributes fine-grained details that complement the textual description. We argue that combining both perspectives enables the model to adapt flexibly to diverse situations encountered in real-world retrieval tasks. As shown in Figure 2, in cross-modal imagination, C_v and C_t generate captions that are similar but not identical. The former captures fine-grained visual details, such as "outdoors" and "multiple medium dogs", which are grounded in the image. The latter, by contrast, emphasizes semantic transformation, for example "transformed from several smaller dogs", reflecting a more abstract and text-driven perspective rather than visually specific cues.

Additionally, $\mathcal{A}_t^i$ is designed to output M_t, which represents the modifications between T_m and C_r. While T_m is provided by the user or predefined to describe changes in the text modality that are applied to the visual modality, M_t is derived entirely from the text modality and provides more specific and explicit modifications. At the same time, $\mathcal{A}_v^i$ produces M_v, which denotes a set indicating the presence or absence of visual attributes. The example of M_v and M_t can be found in Figure 2. The outputs M_t and M_v will later be used by the question agent, which will be discussed in detail in a subsequent section.

Table 1: Performance comparison on CIRCO and CIRR test set. The best results are in bold, and the second best are underlined.

Backbone	Method	Venue	Training-free	CIRCO				CIRR				CIRR$_{subset}$		
				mAP@5	mAP@10	mAP@25	mAP@50	R@1	R@5	R@10	R@50	R@1	R@2	R@3
CLIP-ViT-B/32	PALAVRA [9]	*ECCV 2022*	✘	4.61	5.32	6.33	6.80	16.62	43.49	58.51	83.95	41.61	65.30	80.95
	SEARLE [4]	*ICCV 2023*	✘	9.35	9.94	11.13	11.84	24.00	53.42	66.82	89.78	54.89	76.60	88.19
	SEARLE-OTI [4]	*ICCV 2023*	✘	7.14	7.38	8.99	9.60	24.27	53.25	66.10	88.84	54.10	75.81	87.33
	iSEARLE [1]	*TPAMI 2025*	✘	10.58	11.24	12.51	13.26	25.23	55.69	68.05	90.82	-	-	-
	iSEARLE-OTI [1]	*TPAMI 2025*	✘	10.31	10.94	12.27	13.01	26.19	55.18	68.05	90.65	-	-	-
	CIReVL [22]	*ICLR 2024*	✔	14.94	15.42	17.00	17.82	23.94	52.51	66.00	86.95	60.17	80.05	90.19
	LDRE [51]	*SIGIR 2024*	✔	17.96	18.32	20.21	21.11	25.69	55.13	69.04	89.90	60.53	80.65	90.70
	ImageScope [31]	*WWW 2025*	✔	22.36	22.19	23.03	23.83	34.36	60.58	71.40	88.41	74.63	87.93	93.83
	X^{R}(Ours)	***Proposed***	✔	**27.51**	**28.33**	**30.28**	**30.95**	**43.06**	**73.86**	**83.15**	**94.36**	**77.54**	**90.27**	**95.21**
CLIP-ViT-L/14	Pic2Word [35]	*CVPR 2023*	✘	8.72	9.51	10.64	11.29	23.90	51.70	65.30	87.80	-	-	-
	SEARLE [4]	*ICCV 2023*	✘	11.68	12.73	14.33	15.12	24.24	52.48	66.29	88.84	53.76	75.01	88.19
	SEARLE-OTI [4]	*ICCV 2023*	✘	10.18	11.03	12.72	13.67	24.87	52.32	66.29	88.58	53.80	74.31	86.94
	iSEARLE [1]	*TPAMI 2025*	✘	12.50	13.61	15.36	16.25	25.28	54.00	66.72	88.80	-	-	-
	iSEARLE-OTI [1]	*TPAMI 2025*	✘	11.31	12.67	14.46	15.34	25.40	54.05	67.47	88.92	-	-	-
	LinCIR [12]	*CVPR 2024*	✘	12.59	13.58	15.00	15.85	25.04	53.25	66.68	-	57.11	77.37	88.89
	FTI4CIR [27]	*SIGIR 2024*	✘	15.05	16.32	18.06	19.05	25.90	55.61	67.66	89.66	55.21	75.88	87.98
	CIReVL [22]	*ICLR 2024*	✔	18.57	19.01	20.89	21.80	24.55	52.31	64.92	86.34	59.54	79.88	89.69
	LDRE [51]	*SIGIR 2024*	✔	23.35	24.03	26.44	27.50	26.53	55.57	67.54	88.50	60.43	80.31	89.90
	ImageScope [31]	*WWW 2025*	✔	25.39	25.82	27.07	27.98	34.99	61.35	71.49	88.84	74.94	88.24	94.02
	X^{R}(Ours)	***Proposed***	✔	**31.38**	**32.88**	**35.46**	**36.50**	**43.13**	**73.59**	**83.09**	**94.05**	**77.98**	**90.68**	**95.06**

3.4 Coarse Filtering

In Lines 5–9, each candidate image $I \in \mathcal{I}$ is evaluated by the text similarity-based scoring agent $\mathcal{A}_t^s$ and the vision similarity-based scoring agent $\mathcal{A}_v^s$. To improve robustness, the scoring process integrates three levels of cross-modality, providing multiperspective information across different stages and producing more reliable results through hybrid cross-modal similarity.

First, the scoring of both agents relies on C_t and C_v produced by $\mathcal{A}_t^i$ and $\mathcal{A}_v^i$, which together approximate the ideal target image I_i. Although both C_t and C_v are textual representations, they are derived from different modalities: the text modality and the vision modality, respectively. Second, each candidate image indexed by $a \in \{1, \ldots, N\}$ is also evaluated with respect to its caption C_a and its visual content I_a by $\mathcal{A}_t^s$ and $\mathcal{A}_v^s$, respectively. Third, each scoring agent produces two cross-modal scores. In Line 6, the scores s_t^t and s_t^v are generated by the text similarity-based scoring agent $\mathcal{A}_t^s$. Specifically, s_t^t measures the similarity between C_t and C_a, while s_t^v measures the similarity between C_v and C_a. Since C_t originates from the text modality and C_v implicitly reflects visual content, these two scores capture cross-modal signals that combine implicitly coupled and explicitly decoupled multimodality. Here, explicitly decoupled multimodality refers to processing that occurs entirely within the text modality, whereas implicit coupling indicates that visual information is embedded within the textual representation. The same procedure is applied by the vision similarity-based scoring agent $\mathcal{A}_v^s$. In Line 7, the scores s_v^t and s_v^v measure the similarity between I_a and C_t, and between I_a and C_v, respectively. Here, the pair (I_a, C_v) belongs to the visual modality, whereas the pair (I_a, C_t) combines the vision and text modalities. These procedures are collectively referred to as cross-modal scoring. Each score is generated by its corresponding agent, which encodes the inputs using an MLLM and computes the cosine similarity between the paired representations.

In Lines 8 and 9, for each candidate I, the scores from the text and vision modalities across agents are summed to obtain s^t and s^v, respectively. Rather than aggregating the outputs of a single agent, the scores are aligned within each modality to ensure consistency in cross-modal scoring. In Line 10, a reciprocal rank fusion function is applied to transform the similarity score vectors $S^t = (s_1^t, \ldots, s_N^t)$ and $S^v = (s_1^v, \ldots, s_N^v)$ into rank values, which are then summed across the text and vision modalities. The ranking function is defined as:

$$\mathrm{RRF}(a) = \frac{1}{z + \mathrm{rank}\left(s_a^t\right)} + \frac{1}{z + \mathrm{rank}\left(s_a^v\right)}, \tag{1}$$

where $\mathrm{rank}(s_a^t)$ and $\mathrm{rank}(s_a^v)$ denote the rank positions of the text-based score s_a^t and the vision-based score s_a^v among all candidates, respectively, and z is a smoothing constant. This formulation ensures that higher-ranked candidates in either modality contribute more to the final score, while still incorporating signals from both modalities. Finally, the top-k' candidates $\mathcal{I}^{k'}$ are selected and passed to the next stage for fine filtering, balancing retrieval accuracy with computational cost. If $k' = N$, then all candidates are selected and proceed to the next step.

Notably, the process of producing and aggregating multiple scores is also inspired by human cognitive mechanisms. When humans search for target images, cross-modal information is interwoven in the mind, collectively forming a unified body of evidence that supports decision-making.

3.5 Fine Filtering

The previously discussed efforts for CIR primarily focus on approximating I_i through similarity-based scoring. However, in real-world scenarios, ensuring retrieval accuracy often requires factual verification with semanic reasoning.

To address this, we introduce the question agent $\mathcal{A}^q$ (Line 11). This agent operates on three types of instructions: M_t, which represents modifications derived from $\mathcal{A}_t^i$; M_v, which represents visual attribute indicators; and T_m, which encodes atomic user-specified instructions. The agent $\mathcal{A}^q$ generates a set of verification questions Q with corresponding answers A by transforming each atomic instruction in T_m into a declarative statement, while using M_t as supporting context and M_v as factual grounding. When T_m alone

Table 2: Performance comparison on FashionIQ validation set. The best results are in bold, and the second best are underlined.

Backbone	Method	Venue	Training-free	Shirt		Dress		Toptee		*Avg.*	
				R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
CLIP-ViT-B/32	PALAVRA [9]	*ECCV 2022*	✘	21.49	37.05	17.25	35.94	20.55	38.76	19.76	37.25
	SEARLE [4]	*ICCV 2023*	✘	24.44	41.61	18.54	39.51	25.70	46.46	22.89	42.53
	SEARLE-OTI [4]	*ICCV 2023*	✘	25.37	41.32	17.85	39.91	24.12	45.79	22.45	42.34
	iSEARLE [1]	*TPAMI 2025*	✘	25.81	43.52	20.92	42.19	26.47	48.70	24.40	44.80
	iSEARLE-OTI [1]	*TPAMI 2025*	✘	27.09	43.42	21.27	42.19	26.82	48.75	25.06	44.79
	CIReVL [22]	*ICLR 2024*	✔	28.36	47.84	25.29	46.36	31.21	53.85	28.29	49.35
	LDRE [51]	*SIGIR 2024*	✔	27.38	46.27	19.97	41.84	27.07	48.78	24.81	45.63
	ImageScope [31]	*WWW 2025*	✔	24.29	37.49	18.00	35.20	24.99	41.41	22.42	38.03
	X$^{\mathbf{R}}$(Ours)	***Proposed***	✔	**36.06**	**54.66**	**30.94**	**52.06**	**42.99**	**64.56**	**36.66**	**57.10**
CLIP-ViT-L/14	Pic2Word [35]	*CVPR 2023*	✘	26.20	43.60	20.00	40.20	27.90	47.40	24.70	43.73
	SEARLE [4]	*ICCV 2023*	✘	26.89	45.58	20.48	43.13	29.32	49.97	25.56	46.23
	SEARLE-OTI [4]	*ICCV 2023*	✘	30.37	47.49	21.57	44.47	30.90	51.76	27.61	47.91
	iSEARLE [1]	*TPAMI 2025*	✘	28.75	47.84	22.51	46.36	31.31	52.68	27.52	48.96
	iSEARLE-OTI [1]	*TPAMI 2025*	✘	31.80	50.20	24.19	45.12	31.72	53.29	29.24	49.54
	LinCIR [12]	*CVPR 2024*	✘	29.10	46.81	20.92	42.44	28.81	50.18	26.28	46.48
	FTI4CIR [27]	*SIGIR 2024*	✘	31.35	50.59	24.49	47.84	32.43	54.21	29.42	50.88
	CIReVL [22]	*ICLR 2024*	✔	29.49	47.40	24.79	44.76	31.36	53.65	28.55	48.57
	LDRE [51]	*SIGIR 2024*	✔	31.04	51.22	22.93	46.76	31.57	53.64	28.51	50.54
	ImageScope [31]	*WWW 2025*	✔	27.82	41.76	20.18	37.48	28.61	44.42	25.54	41.22
	X$^{\mathbf{R}}$(Ours)	***Proposed***	✔	**38.91**	**56.82**	**28.71**	**52.50**	**43.91**	**62.57**	**37.18**	**57.30**

is insufficient to define a clear question, M_t provides additional context. All questions are formulated in a True/False format, with examples provided in Figure 2.

In Lines 13 and 14, $\mathcal{A}_t^q$ and $\mathcal{A}_v^q$ are two question-based scoring agents that use Q to evaluate each candidate image $a \in \{1, \ldots, k'\}$ based on different modality information, namely C_a and I_a. We consider a fact to be true if the candidate is able to pass the verification check under both modalities. If the answer to a question is correct, the agent assigns a score of +1; otherwise, the score is 0. The results are denoted as s_q^t and s_q^v. We consider this design to be effective because it provides discrete, verifiable signals that emphasize factual consistency across modalities.

In Lines 15 and 16, we define the cross-modal re-ranking procedure for the top-k' previously selected candidates. For notational simplicity, we omit the index k' here. First, the question-based scores for each candidate are summed as $S_q^t + S_q^v$, where $s_q^t \in S_q^t$, $s_q^v \in S_q^v$, and $|S_q^t| = |S_q^v| = k'$. This sum is then multiplied by a normalized weighted combination of the similarity-based scores S^t and S^v from Lines 8 and 9, with weights λ and $1 - \lambda$, respectively. It is worth noting that only the similarity-based scores S^t and S^v of the top-k' candidates are used at this stage. The result, denoted $S^{k'}$, represents the refined scores used for re-ranking in Line 16. Finally, the re-ranked set $\mathcal{I}^*$ is returned as the final output.

This design is motivated by the complementary strengths of the two scoring mechanisms. Similarity-based scores (S^t and S^v) capture soft alignment between the candidate images and the multimodal query, but they may overlook fine-grained factual details. In contrast, question-based scores (S_q^t and S_q^v) enforce explicit verification of atomic modifications and provide binary, interpretable signals. By combining the two, the re-ranking step integrates the broad cross-modal coverage of similarity-based scoring with the precision of question-based verification, thereby improving robustness and ensuring that the final retrieved ordered set $\mathcal{I}^*$ more faithfully reflects the intended modifications through semantic reasoning.

In fact, the use of X$^{\mathrm{R}}$ is flexible. So far, we have discussed how similarity-based scoring and question-based scoring collaborate through ranking (selection) and re-ranking (re-selection). Moreover, when $k' = k$, the two scoring processes act as ranking–selection and re-ranking. When $k' = N$, the two processes operate jointly, performing ranking and selection directly. It is important to note that these three configurations correspond to increasing computational cost.

In summary, we propose X$^{\mathrm{R}}$, a training-free cross-modal multi-agent framework for CIR. The framework highlights the benefits of agentic AI, including minimal human intervention and autonomous collaboration among multiple agents that mimic human cognitive processes. By enabling retrieval that is both robust and adaptive across modalities, X$^{\mathrm{R}}$ points toward promising directions for large-scale information access scenarios that are becoming increasingly central in web-driven environments.

4 Experiments

4.1 Experiment Setup

Benchmark. We evaluate X$^{\mathrm{R}}$ on three representative CIR benchmarks (Table 4): CIRR [30], the first natural-image CIR dataset with subset retrieval for fine-grained candidate discrimination; CIRCO [4], a large-scale benchmark with multiple ground truths to reduce false negatives; and FashionIQ [45], a fashion-domain dataset with three categories (dress, shirt, toptee).

Metrics. Following the original protocols, we use Recall@k (R@k) for CIRR and FashionIQ to capture retrieval accuracy, and mean average precision (mAP@k) for CIRCO to account for multiple valid ground truths.

Table 3: Ablation studies on CLIP-ViT-B/32 with InternVL3-8B.

Similarity-based		Question-based		FashionIQ		CIRCO			CIRR			CIRR$_{subset}$		
Textual	Visual	Textual	Visual	R@10	R@50	mAP@5	mAP@10	mAP@25	R@1	R@5	R@10	R@1	R@2	R@3
✘	✘	✘	✘	14.78	29.60	2.65	3.25	4.14	11.71	35.06	48.94	32.77	56.89	74.96
✔	✘	✘	✘	19.36	37.65	11.98	13.40	14.11	18.12	51.16	65.11	59.76	79.88	90.00
✘	✔	✘	✘	32.48	54.55	15.18	16.02	17.54	27.02	61.04	74.05	64.53	83.18	91.93
✔	✔	✘	✘	32.84	55.37	16.73	17.69	19.29	27.33	63.57	76.36	66.39	84.00	92.89
✔	✘	✔	✘	23.93	41.72	17.22	17.7	19.05	30.53	58.96	69.78	68.12	83.88	91.64
✘	✔	✘	✔	36.01	56.57	24.12	24.84	26.53	40.24	71.57	81.75	75.42	89.68	93.77
✔	✔	✔	✘	34.78	55.63	24.87	25.51	27.34	36.60	65.69	76.72	72.72	86.41	93.64
✔	✔	✘	✔	36.62	56.84	26.21	27.01	28.87	41.34	73.06	82.43	76.45	89.95	95.13
✔	✔	✔	✔	**36.66**	**57.10**	**27.51**	**28.33**	**30.28**	**43.06**	**73.86**	**83.15**	**77.54**	**90.27**	**95.21**

Table 4: Benchmark details.

Dataset	Split	Type	# Queries	# Images
CIRR [30]	Test	CIR	4,148	2,316
CIRCO [4]	Test	CIR	800	123,403
FashionIQ-Shirt [45]	Val.	CIR	2,038	6,346
FashionIQ-Dress [45]	Val.	CIR	2,017	3,817
FashionIQ-Toptee [45]	Val.	CIR	1,961	5,373

Baselines. We compare X^R against nine representative CIR baselines. Since X^R is training-free, we primarily focus on zero-shot methods for fair comparison, while also reporting strong training-based models for completeness:

- **Training-based:** PALAVRA [9], Pic2Word [35], SEARLE [4], iSEARLE [1], LinCIR [12], and FTI4CIR [27].
- **Training-free:** CIReVL [22], LDRE [51], and ImageScope [31].

Implementation Details. We use CLIP-ViT-L/14 and CLIP-ViT-B/32 [19] as dual-encoder backbones for similarity agents, and InternVL3-8B [57] for imagination and question-based verification. We set the temperature = 0 and top-p = 1 for deterministic outputs. The fusion weight is $\lambda = 0.15$ to balance text–image similarity. We set $k' = 100$ candidates for fine filtering to allow question-based scoring to take effect. If k' were set equal to k, as in Recall@k, the benefit of re-ranking would be masked since Recall@k is order-insensitive. In contrast, fine filtering both re-ranks and prunes candidates, making $k' = 100$ a balanced choice. All experiments are conducted on a single NVIDIA H800-80G GPU with FP16 precision.

4.2 Main Result.

Table 1 and Table 2 illustrate that X^R consistently outperforms both training-free and training-based CIR methods. On FashionIQ, X^R achieves consistent gains across all three categories. With CLIP-ViT-B/32, it reaches 36.66% R@10 and 57.10% R@50 on average, surpassing CIReVL by over 8 points in R@10. These gains hold across shirts, dresses, and tops, indicating that the method generalizes across diverse attribute-level edits rather than overfitting to a single category. On CIRCO, which introduces large distractor sets and multiple ground truths, X^R attains 30.95% mAP@50, over 7 points higher than the best baseline. This shows that multi-agent reasoning maintains robustness in noisy, large-scale retrieval where static similarity models often collapse. On CIRR, X^R achieves 83.15% R@10 and 95.21% R@3 in the fine-grained subset retrieval task (CIRR$_{\text{subset}}$), surpassing training-free and training-based baselines. These results show that imagination and verification act as complementary safeguards against error propagation, ensuring faithful alignment to user intent in fine-grained scenarios. Overall, X^R demonstrates consistent advantages across domain-specific (FashionIQ), distractor-heavy (CIRCO), and fine-grained (CIRR) benchmarks, pointing to strong cross-benchmark generalization.

4.3 Ablation Studies.

Ablation on key modules. Table 3 presents results from progressively enabling different modules. On FashionIQ, the visual similarity agent alone lifts R@10 from 14.78% to 32.48%. Compared with the textual agent, the two together show clear complementarity: the visual branch captures appearance-level cues but risks semantic drift, while the textual branch enforces semantic alignment but may miss subtle details. Combining both further yields substantial gains, with CIRCO mAP@25 rising from 4.14% to 19.29%, confirming that cross-modal similarity is more reliable than unimodal signals. Adding a textual question-based agent then further boosts CIRR R@10 to 76.72%, showing that factual checks reduce false positives. Finally, with both textual and visual question-based agents, CIRR$_{subset}$ R@3 reaches 95.21%, demonstrating that explicit verification enforces modification faithfulness. In summary, each module contributes individually, but their integration ultimately delivers the strongest performance: similarity agents provide broad alignment, and question-based agents enforce correctness, together validating the multi-agent cross-modal reasoning of X^R.

4.4 Discussion.

RRF z vs. summation. Figure 3(a) compares reciprocal rank fusion (RRF) with varying z against direct score summation. Direct score summation performs worst, showing that naive averaging fails to normalize heterogeneous modalities. RRF instead focuses on ranks, which makes aggregation more robust to noisy candidates. At $z = 60$, it strikes the best balance, raising CIRCO mAP@25 to 30.28% and CIRR R@10 to 83.15%.

Generality of MLLMs. Figure 3(b) compares different multimodal backbones. Medium-scale models such as InternVL3-8B and Qwen2.5VL-7B achieve the best trade-off, combining strong grounding ability with efficiency (e.g., 57.10% R@50 on FashionIQ and 95.21% R@3 on CIRR$_{subset}$). Smaller backbones lack grounding ability, while extremely large ones bring only marginal gains at a much higher cost.

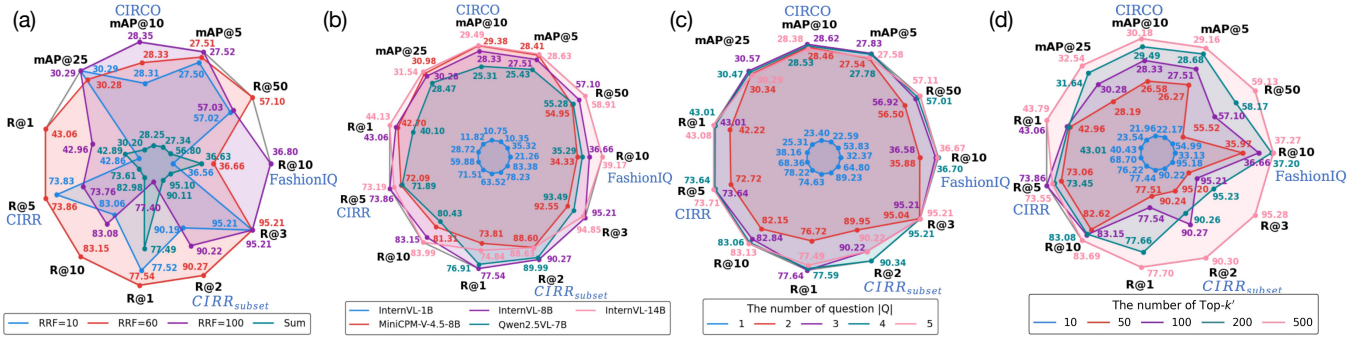


Figure 3: Parameter analysis of X^R. (a) Effect of RRF with different z values. (b) Comparison across multimodal backbones. (c) Impact of the number of verification questions. (d) Influence of candidate pool size k'.

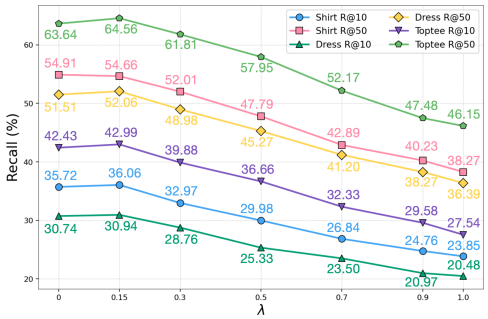
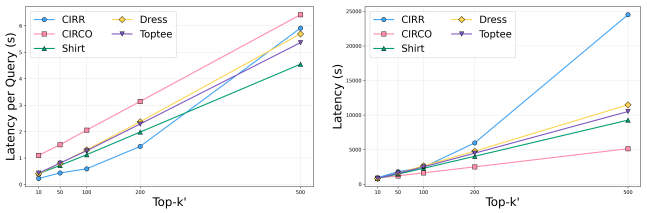


Figure 4: Effect of λ on text–image fusion: best at λ=0.15; extremes degrade by losing cross-modal cues.



(a) Average latency per query. **(b) Total latency.**

Figure 5: Latency of X^R under different top-k': larger pools increase cost nearly linearly, but $k' \approx 100$ balances coverage and overhead.

This suggests that X^R benefits most from medium-scale MLLMs, which balance expressiveness and efficiency and indicate that the framework scales smoothly across backbones.

Number of questions. Figure 3(c) shows that a single verification question is insufficient, while three yield consistent gains across benchmarks (e.g., CIRR R@10 = 82.84%). Using more than three slightly reduces performance, as redundant checks add overhead without new information. These results indicate that a small, diverse set of questions suffices for robust factual alignment, consistent with the principle that each agent contributes distinct value.

Top-k' analysis. Figure 3(d) illustrates the impact of the coarse filtering pool size. Small k' limits recall, while larger pools improve coverage by ensuring candidate diversity. Gains plateau beyond $k' = 100$, with only marginal improvements up to $k' = 500$. This indicates that moderately large pools provide the best efficiency–effectiveness balance.

Effect of λ. Figure 4 examines the balance between textual and visual similarity signals. Relying solely on either modality ($\lambda = 0$ or $\lambda = 1$) significantly degrades performance, as it ignores complementary cues. The best results occur at $\lambda = 0.15$, suggesting that cross-modal fusion is most effective when neither modality dominates, validating the principle of balanced agent collaboration.

Latency analysis. Figure 5 shows average and total latency under different candidate pool sizes (k'). As expected, larger k' increases cost as more candidates enter fine filtering. Per-query latency grows nearly linearly, with CIRCO highest due to its many distractors. Total latency is dominated by CIRR because of its dataset scale, while FashionIQ remains relatively lightweight. Overall, a moderately large k' (around 100) offers the best trade-off, balancing diversity for robust retrieval against computational overhead.

These results highlight that the strength of X^R comes from orchestrating multiple agents rather than relying on any single component. By uniting semantic alignment with factual verification, cross-modal reasoning refines retrieval and overcomes the inherent limits of unimodal pipelines. More broadly, this shows that cross-modal multi-agent reasoning is not only effective for CIR but establishes a general paradigm for multimodal retrieval and reasoning.

5 Conclusion

We presented X^R, a training-free cross-modal multi-agent framework for composed image retrieval. Unlike unimodal pipelines, X^R integrates imagination, coarse filtering, and fine filtering through similarity- and question-based agents, progressively refining results via semantic alignment and factual verification. Experiments on FashionIQ, CIRCO, and CIRR show consistent improvements over both training-free and training-based baselines, particularly in fine-grained and distractor-rich scenarios. Ablation analyses confirm that while each agent contributes independently, their coordination yields more stable and accurate retrieval. Overall, our findings underscore that cross-modal reasoning is not only advantageous but often essential for aligning retrieval with user intent. Looking ahead, we envision X^R as a foundation for retrieval-augmented reasoning, where agentic systems actively interpret, verify, and adapt across modalities to achieve reliable and human-aligned intelligence.

References

[1] Lorenzo Agnolucci, Alberto Baldrati, Marco Bertini, and A. Bimbo. 2024. iSEARLE: Improving Textual Inversion for Zero-Shot Composed Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47 (2024), 10801–10817. https://api.semanticscholar.org/CorpusID:269604752

[2] Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei A. Zaharia, and O. Khattab. 2025. GEPA: Reflective Prompt Evolution Can Outperform Reinforcement Learning. *ArXiv* abs/2507.19457 (2025). https://api.semanticscholar.org/CorpusID:280046245

[3] Yang bai, Xinxing Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong Goh, and Chun-Mei Feng. 2024. Sentence-level Prompts Benefit Composed Image Retrieval. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=m3ch3kJL7q

[4] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and A. Bimbo. 2023. Zero-Shot Composed Image Retrieval with Textual Inversion. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), 15292–15301. https://api.semanticscholar.org/CorpusID:257766776

[5] Tong Bao, Che Liu, Derong Xu, Zhi Zheng, and Tong Xu. 2025. MLLM-I2W: Harnessing Multimodal Large Language Model for Zero-Shot Composed Image Retrieval. In *International Conference on Computational Linguistics*. https://api.semanticscholar.org/CorpusID:275821154

[6] Ben Chen, Linbo Jin, Xinxin Wang, Dehong Gao, Wen Jiang, and Wei Ning. 2023. Unified Vision-Language Representation Modeling for E-Commerce Same-style Products Retrieval. *Companion Proceedings of the ACM Web Conference 2023* (2023). https://api.semanticscholar.org/CorpusID:256808380

[7] Yanbei Chen, Shaogang Gong, and Loris Bazzani. 2020. Image Search With Text Feedback by Visiolinguistic Attention Learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 2998–3008. https://api.semanticscholar.org/CorpusID:219401805

[8] Zhangtao Cheng, Yuhao Ma, Jian Lang, Kunpeng Zhang, Ting Zhong, Yong Wang, and Fan Zhou. 2025. Generative Thinking, Corrective Action: User-Friendly Composed Image Retrieval via Automatic Multi-Agent Collaboration. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2* (2025). https://api.semanticscholar.org/CorpusID:280448098

[9] Niv Cohen, Rinon Gal, Eli A. Meirom, Gal Chechik, and Yuval Atzmon. 2022. "This is my unicorn, Fluffy": Personalizing frozen vision-language representations. *ArXiv* abs/2204.01694 (2022). https://api.semanticscholar.org/CorpusID:247939764

[10] Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, Katsushi Ikeuchi, Hoi Vo, Fei-Fei Li, and Jianfeng Gao. 2024. Agent AI: Surveying the Horizons of Multimodal Interaction. *ArXiv* abs/2401.03568 (2024). https://api.semanticscholar.org/CorpusID:266844635

[11] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva N. Mody, Steven Truitt, and Jonathan Larson. 2024. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *ArXiv* abs/2404.16130 (2024). https://api.semanticscholar.org/CorpusID:269363075

[12] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. 2023. Language-only Efficient Training of Zero-shot Composed Image Retrieval. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 13225–13234. https://api.semanticscholar.org/CorpusID:265609308

[13] Jiawei Gu, Ziting Xian, Yuanzhen Xie, Ye Liu, Enjie Liu, Ruichao Zhong, Mochi Gao, Yunzhi Tan, Bo Hu, and Zang Li. 2025. Toward Structured Knowledge Reasoning: Contrastive Retrieval-Augmented Generation on Experience. In *Annual Meeting of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:279075943

[14] Venkat N. Gudivada and Vijay V. Raghavan. 1995. Content-Based Image Retrieval Systems - Guest Editors' Introduction. *Computer* 28 (1995), 18–22. https://api.semanticscholar.org/CorpusID:206402728

[15] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).

[16] Demis Hassabis. 2023. Introducing Gemini: our largest and most capable AI model. *Google Blog* (2023). Accessed 2025-10-03.

[17] M. Hosseinzadeh and Yang Wang. 2020. Composed Query Image Retrieval Using Locally Bounded Features. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 3593–3602. https://api.semanticscholar.org/CorpusID:219963281

[18] Chuong Huynh, Jinyu Yang, Ashish Tawari, Mubarak Shah, Son Tran, Raffay Hamid, Trishul M. Chilimbi, and Abhinav Shrivastava. 2025. CoLLM: A Large Language Model for Composed Image Retrieval. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), 3994–4004. https://api.semanticscholar.org/CorpusID:277314021

[19] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:231879586

[20] Bowen Jiang, Yangxinyu Xie, Xiaomeng Wang, Weijie J. Su, Camillo Jose Taylor, and Tanwi Mallick. 2024. Multi-Modal and Multi-Agent Systems Meet Rationality: A Survey. *ArXiv* abs/2406.00252 (2024). https://api.semanticscholar.org/CorpusID:278935961

[21] Yucheng Jiang, Yijia Shao, Dekun Ma, Sina J. Semnani, and Monica S. Lam. 2024. Into the Unknown Unknowns: Engaged Human Learning through Participation in Language Model Agent Conversations. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:271963301

[22] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. 2024. Vision-by-Language for Training-Free Compositional Image Retrieval. In *International Conference on Learning Representations*.

[23] Wei Li, Hehe Fan, Yongkang Wong, Yi Yang, and Mohan S. Kankanhalli. 2024. Improving Context Understanding in Multimodal Large Language Models via Multimodal Composition Learning. In *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:272330468

[24] You Li, Fan Ma, and Yi Yang. 2024. Imagine and Seek: Improving Composed Image Retrieval with an Imagined Proxy. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), 3984–3993. https://api.semanticscholar.org/CorpusID:274281272

[25] Zhe Li, Lei Zhang, Kun Zhang, Weidong Chen, Yongdong Zhang, and Zhendong Mao. 2025. Rethinking Pseudo Word Learning in Zero-Shot Composed Image Retrieval: From an Object-Aware Perspective. *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2025). https://api.semanticscholar.org/CorpusID:280069892

[26] Jintao Liang, Gang Su, Huifeng Lin, You Wu, Rui Zhao, and Ziyue Li. 2025. Reasoning RAG via System 1 or System 2: A Survey on Reasoning Agentic Retrieval-Augmented Generation for Industry Challenges. arXiv:2506.10408 [cs.AI] https://arxiv.org/abs/2506.10408

[27] Haoqiang Lin, Haokun Wen, Xuemeng Song, Meng Liu, Yupeng Hu, and Liqiang Nie. 2024. Fine-grained Textual Inversion Network for Zero-Shot Composed Image Retrieval. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2024). https://api.semanticscholar.org/CorpusID:271114410

[28] Liping Liu, Chunhong Zhang, Likang Wu, Chuang Zhao, Zheng Hu, Ming He, and Jianpin Fan. 2025. Instruct-of-Reflection: Enhancing Large Language Models Iterative Reflection Capabilities via Dynamic-Meta Instruction. *ArXiv* abs/2503.00902 (2025). https://api.semanticscholar.org/CorpusID:276741572

[29] Ying Liu, Dengsheng Zhang, Guojun Lu, and Wei-Ying Ma. 2007. A survey of content-based image retrieval with high-level semantics. *Pattern Recognit.* 40 (2007), 262–282. https://api.semanticscholar.org/CorpusID:9160719

[30] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. 2021. Image Retrieval on Real-life Images with Pre-trained Vision-and-Language Models. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), 2105–2114. https://api.semanticscholar.org/CorpusID:236956879

[31] Pengfei Luo, Jingbo Zhou, Tong Xu, Yuan Xia, Linli Xu, and Enhong Chen. 2025. ImageScope: Unifying Language-Guided Image Retrieval via Large Multimodal Model Collective Reasoning. *Proceedings of the ACM on Web Conference 2025* (2025). https://api.semanticscholar.org/CorpusID:276961009

[32] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-Refine: Iterative Refinement with Self-Feedback. *ArXiv* abs/2303.17651 (2023). https://api.semanticscholar.org/CorpusID:257900871

[33] OpenAI. 2025. Introducing GPT-5. Accessed 2025-10-03.

[34] Matthew Renze and Erhan Guven. 2024. The benefits of a concise chain of thought on problem-solving in large language models. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*. IEEE, 476–483.

[35] 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*. 19305–19314.

[36] Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. 2024. Assisting in writing wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 6252–6278.

[37] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems* 36 (2023), 8634–8652.

[38] Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. 2025. Agentic retrieval-augmented generation: A survey on agentic rag. *arXiv preprint arXiv:2501.09136* (2025).

[39] Xuemeng Song, Haoqiang Lin, Haokun Wen, Bohan Hou, Mingzhu Xu, and Liqiang Nie. 2025. A comprehensive survey on composed image retrieval. *ACM Transactions on Information Systems* 44, 1 (2025), 1–54.
[40] Yuanmin Tang, Jue Zhang, Xiaoting Qin, Jing Yu, Gaopeng Gou, Gang Xiong, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Wu. 2025. Reason-before-retrieve: One-stage reflective chain-of-thoughts for training-free zero-shot composed image retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 14400–14410.
[41] Rong-Cheng Tu, Wenhao Sun, Hanzhe You, Yingjie Wang, Jiaxing Huang, Li Shen, and Dacheng Tao. 2025. Multimodal Reasoning Agent for Zero-Shot Composed Image Retrieval. *arXiv preprint arXiv:2505.19952* (2025).
[42] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2019. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6439–6448.
[43] Ji Wan, Dayong Wang, Steven Chu Hong Hoi, Pengcheng Wu, Jianke Zhu, Yongdong Zhang, and Jintao Li. 2014. Deep learning for content-based image retrieval: A comprehensive study. In *Proceedings of the 22nd ACM international conference on Multimedia*. 157–166.
[44] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191* (2024).
[45] 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/CVF Conference on computer vision and pattern recognition*. 11307–11317.
[46] Xiaohui Xie, Yiqun Liu, Maarten De Rijke, Jiyin He, Min Zhang, and Shaoping Ma. 2018. Why people search for images using web search engines. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 655–663.
[47] Eric Xing, Pranavi Kolouju, Robert Pless, Abby Stylianou, and Nathan Jacobs. 2025. ConText-CIR: Learning from Concepts in Text for Composed Image Retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 19638–19648.
[48] Zhongyu Yang, Jun Chen, Dannong Xu, Junjie Fei, Xiaoqian Shen, Liangbing Zhao, Chun-Mei Feng, and Mohamed Elhoseiny. 2025. WikiAutoGen: Towards Multi-Modal Wikipedia-Style Article Generation. *arXiv preprint arXiv:2503.19065* (2025).
[49] Zhongyu Yang, Junhao Song, Siyang Song, Wei Pang, and Yingfang Yuan. 2025. MERMAID: Multi-perspective Self-reflective Agents with Generative Augmentation for Emotion Recognition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 24639–24655. doi:10.18653/v1/2025.emnlp-main.1252
[50] Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Bo Li, Chengwei Qin, Shijian Lu, Xingxuan Li, and Lidong Bing. 2025. LongVT: Incentivizing "Thinking with Long Videos" via Native Tool Calling. arXiv:2511.20785 [cs.CV] https://arxiv.org/abs/2511.20785
[51] Zhenyu Yang, Dizhan Xue, Shengsheng Qian, Weiming Dong, and Changsheng Xu. 2024. Ldre: Llm-based divergent reasoning and ensemble for zero-shot composed image retrieval. In *Proceedings of the 47th International ACM SIGIR conference on research and development in information retrieval*. 80–90.
[52] Zhongyu Yang, Yingfang Yuan, Xuanming Jiang, Baoyi An, and Wei Pang. 2025. InEx: Hallucination Mitigation via Introspection and Cross-Modal Multi-Agent Collaboration. arXiv:2512.02981 [cs.CV] https://arxiv.org/abs/2512.02981
[53] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.
[54] Kun Zhang, Jingyu Li, Zhe Li, Jingjing Zhang, Fan Li, Yandong Liu, Rui Yan, Zihang Jiang, Nan Chen, Lei Zhang, et al. 2025. Composed multi-modal retrieval: A survey of approaches and applications. *arXiv preprint arXiv:2503.01334* (2025).
[55] Kaichen Zhang, Keming Wu, Zuhao Yang, Bo Li, Kairui Hu, Bin Wang, Ziwei Liu, Xingxuan Li, and Lidong Bing. 2025. OpenMMReasoner: Pushing the Frontiers for Multimodal Reasoning with an Open and General Recipe. arXiv:2511.16334 [cs.AI] https://arxiv.org/abs/2511.16334
[56] Xiaoyang Zheng, Zilong Wang, Sen Li, Ke Xu, Tao Zhuang, Qingwen Liu, and Xiaoyi Zeng. 2023. Make: Vision-language pre-training based product retrieval in taobao search. In *Companion Proceedings of the ACM Web Conference 2023*. 356–360.
[57] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479* (2025).

Appendix

A Detailed Experiment Results

In this section, we provide the complete parameter analysis and ablation study results on the FashionIQ dataset. While the main paper reports the representative metrics (R@10 and R@50), here we include the full set of scores across different categories (shirts, dresses, and tops) and a wider range of evaluation metrics. These results further demonstrate the consistent advantages of our cross-modal framework over single-modality baselines.

B Statistical Significance Study

We conduct 10 independent runs with different random seeds and compare our method against baseline models under identical settings. To assess statistical significance, we perform paired one-sided t-tests and Wilcoxon signed-rank tests. The null hypothesis ($\mathbf{H}_0$) states that X^{R} performs equally or worse than the baseline, while the alternative hypothesis ($\mathbf{H}_1$) states that our method performs better. As shown in Table B.1, X^{R} achieves the highest mean score (57.16 ± 0.07), compared to CIReVL (49.06 ± 0.23), ImageScope (37.99 ± 0.25), and the Raw baseline (29.62 ± 0.11). All comparisons yield p-values well below the threshold $\alpha = 0.05$ (e.g., t-test: 3.94×10^{-17} against CIReVL), thereby allowing us to confidently reject $\mathbf{H}_0$ in favor of $\mathbf{H}_1$. These results confirm that X^{R} consistently and significantly outperforms all baselines.

C Experimental Code

. To promote transparency and ensure the reproducibility of our work, we will release all experimental code, datasets, and detailed tutorials necessary for replicating our experiments. Our goal is to make it straightforward for researchers and practitioners to reproduce our results, regardless of their technical background. Additionally, by providing comprehensive documentation and clear guidelines, we aim to facilitate the extension of our method to other models and architectures, enabling the broader research community to explore its potential applications and improvements. We believe that open and reproducible research is essential for advancing the field and fostering collaboration.

D Limitations and Future Work

While X^{R} achieves strong results on CIR benchmarks, several limitations remain. The framework is currently tailored to image–text composition and has not yet been explored in settings involving richer modalities or temporal data. Its reliance on captions and verification questions generated by large models can also introduce subtle biases, which may affect consistency. Moreover, scaling to very large candidate pools requires further optimization of efficiency. Looking ahead, we envision cross-modal reasoning as the key avenue for progress. Extending X^{R} beyond images and text to modalities such as video, audio, or interactive queries would open new opportunities for retrieval systems. Developing more lightweight and adaptive agents, together with diverse verification signals, could further enhance both robustness and scalability. These directions highlight the potential of cross-modal multi-agent reasoning as a general paradigm for future multimodal search.

Table A.1: Ablation studies on CLIP-ViT-B/32 with InternVL3-8B on FashionIQ.

Similarity-based		Question-based		Shirt		Dress		Toptee		*Avg.*	
Textual	Visual	Textual	Visual	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
✘	✘	✘	✘	13.44	26.25	13.83	30.88	17.08	31.67	14.78	29.60
✔	✘	✘	✘	19.48	35.62	16.21	33.96	22.39	43.35	19.36	37.65
✘	✔	✘	✘	32.92	53.19	26.13	49.18	38.40	61.30	32.48	54.55
✔	✘	✔	✘	23.87	39.63	19.84	37.60	28.06	47.93	23.93	41.72
✘	✔	✘	✔	35.62	54.41	29.90	51.66	42.51	63.64	36.01	56.57
✔	✔	✘	✘	33.45	53.54	27.37	50.69	37.68	61.86	32.84	55.37
✔	✔	✔	✘	34.74	53.48	28.90	50.07	40.69	63.34	34.78	55.63
✔	✔	✘	✔	36.16	54.17	31.04	51.96	42.65	64.41	36.62	56.84
✔	✔	✔	✔	36.06	54.66	30.94	52.06	42.99	64.56	36.66	57.10

Table A.2: Number of questions studies on CLIP-ViT-B/32 with InternVL3-8B on FashionIQ.

Question Num	Shirt		Dress		Toptee		*Avg.*	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
1	32.14	51.67	26.38	49.33	38.60	60.53	32.37	53.83
2	34.79	54.51	30.00	51.41	42.84	63.59	35.88	56.50
3	35.82	54.66	31.18	52.01	42.73	64.10	36.58	56.92
4	35.87	54.76	31.18	51.96	43.04	64.30	36.70	57.01
5	36.11	54.91	30.94	52.01	42.94	64.41	36.67	57.11

Table A.3: Generality of MLLMs studies on CLIP-ViT-B/32 with InternVL3-8B on FashionIQ.

MLLMs	Shirt		Dress		Toptee		*Avg.*	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
InternVL3-1B	18.66	29.99	17.90	34.26	27.72	41.72	21.26	35.32
InternVL3-8B	36.06	54.66	30.94	52.06	42.99	64.56	36.66	57.10
InternVL3-14B	38.86	56.82	33.47	54.88	45.18	65.02	39.17	58.91
MiniCPM-V-4.5-8B	33.52	54.40	30.51	50.33	41.84	61.12	35.29	55.28
Qwen2.5VL-7B	31.79	52.89	31.33	50.22	39.86	61.73	34.33	54.95

Table A.4: RRF z vs. summation studies on CLIP-ViT-B/32 with InternVL3-8B on FashionIQ.

Method	Shirt		Dress		Toptee		*Avg.*	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
RRF=10	35.97	55.10	30.94	52.06	42.78	63.90	36.56	57.02
RRF=60	36.06	54.66	30.94	52.06	42.99	64.56	36.66	57.10
RRF=100	36.31	55.05	30.99	52.06	43.09	64.00	36.80	57.03
Sum	35.91	54.35	30.99	52.06	42.99	64.00	36.63	56.80

Table A.5: Top-k' analysis on CLIP-ViT-B/32 with InternVL3-8B on FashionIQ.

Top-k'	Shirt		Dress		Toptee		*Avg.*	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
10	33.37	53.73	26.87	48.98	39.16	62.26	33.13	54.99
50	35.77	53.29	29.77	51.30	42.38	61.96	35.97	55.52
100	36.06	54.66	30.94	52.06	42.99	64.56	36.66	57.10
200	36.26	55.64	31.68	53.54	43.65	65.32	37.20	58.17
500	36.31	56.87	31.63	54.88	43.86	65.63	37.27	59.13

Table A.6: Latency analysis (in seconds) across categories under different Top-k'.

Top-k'	CIRR	CIRCO	FashionIQ$_{Shirt}$	FashionIQ$_{Dress}$	FashionIQ$_{Toptee}$	Avg.
10	945	881	831	829	847	867
50	1808	1203	1467	1611	1585	1535
100	2464	1645	2298	2643	2507	2311
200	5970	2511	4037	4781	4483	4356
500	24506	5132	9262	11476	10509	12177

Table A.7: Average latency per query (in seconds) across categories under different Top-k'.

Top-k	CIRR	CIRCO	FashionIQ$_{Shirt}$	FashionIQ$_{Dress}$	FashionIQ$_{Toptee}$	Avg.
10	0.228	1.101	0.408	0.411	0.432	0.516
50	0.436	1.504	0.720	0.799	0.608	0.653
100	0.594	2.056	1.128	1.310	1.278	1.273
200	1.439	3.139	1.981	2.370	2.286	2.243
500	5.908	6.415	4.545	5.690	5.359	5.983

Table B.1: Statistical comparison on FashionIQ benchmark with average Recall@50. StdDev denotes standard deviation. Paired one-sided t-test and Wilcoxon signed-rank test p-values are reported ($\alpha = 5\%$).

Method	Mean (%)	StdDev	t-test p	Wilcoxon p
Raw	29.62	0.11	3.39×10^{-26}	4.88×10^{-4}
CIReVL	49.06	0.23	3.94×10^{-17}	
ImageScope	37.99	0.25	6.82×10^{-21}	
X^{R} (Ours)	**57.16**	**0.07**	–	

E Ethical Considerations

Reliability and Transparency. X^{R} enhances retrieval reliability by coordinating imagination, similarity, and verification, reducing semantic drift and promoting more trustworthy multimodal systems. Its modular design decomposes decisions into interpretable stages, enabling auditing and analysis of system behavior.

Responsible Data Use. All experiments are conducted on publicly available datasets with proper licenses, ensuring compliance with ethical data standards.

F Case Studies

Beyond aggregate metrics, we present case studies on CIRR, FashionIQ, and CIRCO to illustrate how X^{R} behaves on concrete queries. These examples highlight complementary aspects of the framework:

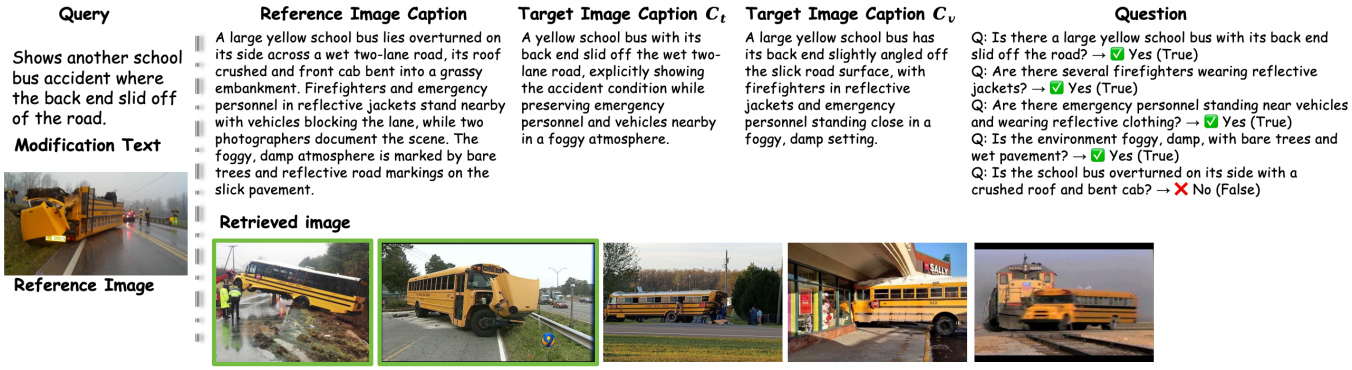


Figure F.1: Case study on CIRR. X^R correctly grounds complex scene edits (e.g., bus orientation, reflective jackets) through factual verification. Target image is marked with the green box.

Figure F.2: Case study on FashionIQ. X^R captures subtle attribute edits (e.g., tone, lettering) and validates them via text-based questioning. Target image is marked with the green box.

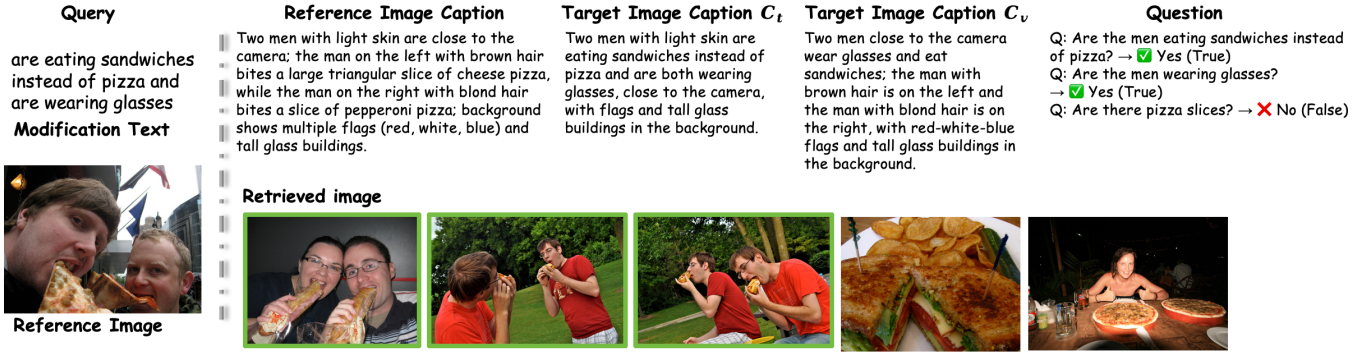


Figure F.3: Case study on CIRCO. X^R remains robust under distractor-heavy settings by verifying entity-level edits (e.g., food type, clothing). Target image is marked with the green box.

on CIRR, it grounds complex scene edits through factual verification; on FashionIQ, it captures subtle attribute modifications such as color or lettering; and on CIRCO, it remains robust under distractor-heavy settings where static similarity often fails. Together, these cases not only showcase the strengths of multi-agent reasoning but also reveal remaining challenges, offering qualitative evidence that complements our quantitative findings.