POINTWISE VISION-LANGUAGE RERANKING WITH REINFORCEMENT FINE-TUNING FOR COMPOSED IMAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Composed Image Retrieval (CIR) requires retrieving a target image given a reference image and a modification text describing desired attribute changes. Existing approaches predominantly rely on embedding-based methods that fuse visual and textual features into a joint representation for retrieval. However, these methods struggle with fine-grained semantic understanding of the modification, especially when the desired change is subtle or compositional. We propose a pointwise vision-language model (VLM) reranker trained with reinforcement fine-tuning that operates on top of an off-the-shelf coarse retriever. Specifically, our method frames multi-candidate ranking as a policy optimization problem: the VLM assigns a relevance score to each candidate image through next-token prediction of "Yes"/"No" tokens, and we train the model via a GRPO-style objective whose stage0-aware reward is computed over rankings that fuse the learned scores with the coarse retriever's prior. Additionally, at inference the learned pointwise scores are fused with the stage0 ranking through a rank-log transformation, enabling the model to both exploit its own semantic understanding and preserve the retriever's recall. Because the protocol consumes only the retriever's top-K candidate lists, the same recipe attaches to different coarse retrievers without re-encoding their galleries; we instantiate it with DQU-CIR and MCoT-MVS and train a separate adapter for each. On FashionIQ, Shoes, and CIRR, three benchmark datasets for composed image retrieval, PVR-RFT consistently improves the stage0 retriever it is paired with, raising the controlled top-256 R@10 on all FashionIQ categories and the official CIRR test-server metrics over their matched stage0 baselines, and comprehensive ablation experiments validate the contribution of each component. Our source code and complete training/evaluation configuration are available in an anonymous repository at https://anonymous.4open.science/r/PVRRFT-code.

1 INTRODUCTION

Composed Image Retrieval (CIR) is a multimodal retrieval task where a user provides a reference image and a natural language modification to retrieve a target image that matches the modification intent Vo et al. (2019); Zhu et al. (2023). Unlike traditional text-to-image retrieval, CIR requires the model to jointly reason about visual content and semantic transformations—for example, "a shirt like this, but with longer sleeves and a floral pattern." This task has broad applications in e-commerce, fashion search, and interactive image editing.

The dominant paradigm for CIR is embedding-based retrieval: a model encodes both the reference image and modification text into a single query embedding, then retrieves the target via nearest neighbor search in a shared embedding space Baldrati et al. (2022; 2023). These methods are efficient and scalable but face inherent limitations. First, the fusion of visual and textual modalities into a single vector inevitably loses fine-grained semantic detail. Second, the embedding space is constrained by the training objective—typically triplet loss or contrastive loss—which does not directly optimize for ranking quality at higher cutoff depths (e.g., R@10 or R@50).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Recent advances in vision-language models (VLMs) such as Qwen-VL Bai et al. (2023; 2025) and LLaVA Liu et al. (2023) offer a new opportunity: they can directly compare a reference and a candidate image with natural language through multi-turn visual reasoning. This opens up a *pointwise reranking* paradigm, where a VLM scores individual candidate images and reorders the coarse retrieval results. The reranking design space spans three natural choices. Pointwise scoring independently evaluates each candidate and is computationally light, but it has no explicit comparison signal. Pairwise and listwise scoring can model candidate interactions at higher inference cost and are harder to integrate with large galleries. Generative ranking emits a full order as text, leveraging VLM reasoning but introducing autoregressive decoding and calibration risks. We argue that a practically usable reranker should preserve stage0 recall, remain cheap at inference, and be trainable with a ranking-aware objective.

However, naively applying a VLM as a pointwise scorer faces two key challenges. First, the VLM's "Yes"/"No" probability over closure tokens yields a ranking that is often *worse* than the coarse retriever on R@10–R@50, because the model lacks comparative context and tends to overfit to surface-level visual similarity without understanding the modification intent. Second, the VLM's judgments are independent per candidate, ignoring the ranking structure of the candidate list.

Coarse retrievers and rerankers have complementary advantages: embedding retrievers give efficient full-gallery recall but are weak at fine-grained ranking, while a VLM reranker reasons about modification semantics yet cannot recover targets absent from the candidate pool. PVR-RFT therefore treats the coarse retriever as the recall ceiling, consumes only its top-K lists, and is retriever-agnostic—we instantiate it with DQU-CIR and MCoT-MVS and report both configurations in parallel rather than claiming dependence on either embedding family.

In this paper, we propose a **Pointwise VLM Reranker with Reinforcement Fine-tuning** (PVR-RFT) that addresses both challenges. To demonstrate the effectiveness of the proposed method, we conduct experiments on FashionIQ, Shoes, and CIRR under a controlled top-256 candidate-pool protocol, and report both stage0 configurations in parallel. Our key contributions are:

1. **Pointwise policy as ranking.** We formulate VLM pointwise scoring as a ranking policy: the model outputs "Yes"/"No" logits for each candidate, and we treat the softmax over these logits as a policy distribution, enabling GRPO-style RL with Gumbel-perturbed Plackett-Luce rollouts and ranking-oriented rewards.
2. **Stage0-aware reward.** We train with a stage0-aware objective: the reward is the reciprocal rank of the target under a fused ordering that combines the VLM's pointwise scores with the coarse retriever's prior, encouraging the model to preserve and improve upon the stage0 ordering.
3. **Rank-log fusion at inference.** We introduce a simple yet effective fusion mechanism: VLM log-odds scores are combined with a neg-log transform of the stage0 rank, using a single tunable hyperparameter. This fusion is computationally free and consistently improves results. Across FashionIQ, Shoes, and CIRR, the same recipe improves whatever stage0 retriever it is given, with separately trained adapters for DQU-CIR and MCoT-MVS reported in parallel.

2 Related Work

2.1 Composed Image Retrieval

Early CIR methods trained dedicated compositors that fuse reference image features with text-based transformation features Vo et al. (2019). More recent work leverages vision-language pre-training: CLIP-based methods Baldrati et al. (2022; 2023) map the composed query into the CLIP text space and retrieve via cosine similarity. DQU-CIR Wen et al. (2024) introduces a dual-query unified framework that generates both the target image's visual features and its textual description, achieving state-of-the-art performance on FashionIQ. MegaPairs/BGE-VL Zhou et al. (2025) explores VLM-derived embeddings for retrieval. More recent methods add multi-granular visual signals: COMBINER Li et al. (2026) and MCoT-MVS Ge et al. (2026) use segmentation, multi-level attention, or multimodal chain-of-thought cues to suppress irrelevant visual regions. ReCALL Yang et al. (2026) and FiRE Hou et al. (2025) further investigate MLLM-based generative ranking, where the

model either directly emits ranking text or is fine-tuned with retrieval-oriented objectives. Our work builds on these advances but deliberately stays in pointwise logit space: it preserves the recall and stability of embedding-based stage0 while adding a trainable VLM refinement layer.

2.2 Vision-Language Models for Ranking

The emergence of large VLMs has spurred recent work on generative ranking for multimodal retrieval. FiRE Hou et al. (2025) and related generative ranking methods propose using VLMs to directly generate ranked lists through text generation. However, text-based ranking is difficult to calibrate and susceptible to hallucination. Embed-RL Jiang et al. (2026) takes a different approach by fine-tuning VLMs to produce better embeddings through reinforcement learning. Multi-stage retrieval has been extensively studied in text retrieval Nogueira et al. (2019) and is gaining traction in vision-language tasks. Our method takes a middle ground: we use VLM pointwise logits for scoring (avoiding text generation's instability) while employing RL training to optimize the ranking objective.

2.3 Reinforcement Fine-tuning for Ranking

Reinforcement learning has been applied to improve the ranking quality of language models. RLHF Ouyang et al. (2022) uses human preferences to train reward models, while GRPO Shao et al. (2024); Guo et al. (2025) directly optimizes from group-relative advantages. MMRAG-RFT Zhao et al. (2026) applies two-stage RL fine-tuning to multimodal retrieval-augmented generation. Our work adapts the GRPO framework to pointwise ranking: we treat the softmax over VLM pointwise scores as the policy, sample rankings via Gumbel-perturbed orderings, and compute advantages from a stage0-aware reward.

3 Method

3.1 Overview

Figure 1 provides an overview of our proposed PVR-RFT framework. Given a reference image and modification text, a stage0 coarse retriever retrieves top-K candidates from the image gallery. We report DQU-CIR and MCoT-MVS as two parallel stage0 configurations; for each, we train a separate PVR-RFT adapter on its top-K rankings and apply the same fusion protocol. Our stage1 VLM reranker independently scores each candidate via Yes/No token probabilities, trained with GRPO-style RL using a stage0-aware reward. At inference, VLM scores are fused with stage0 ranks to produce the final ranking.

3.2 Problem Formulation

Given a reference image I_r, a modification text T_m, and a candidate pool $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$ retrieved by a stage0 coarse retriever, the goal of CIR reranking is to produce an ordering π over $\mathcal{C}$ such that the target image c^* (if present) ranks as highly as possible.

In our two-stage framework, the stage0 retriever first retrieves K candidate images via embedding similarity. Our stage1 VLM reranker then scores each candidate individually, producing a list of relevance scores $\{s(c_i)\}_{i=1}^{K}$. The final ranking is obtained by sorting candidates by their fused score:

$$\tilde{s}(c_i) = s_{\text{VLM}}(c_i) + \alpha \cdot \phi(r_i^{\text{stage0}}), \tag{1}$$

where $r_i^{\text{stage0}} \in \{1, \ldots, K\}$ is the integer stage0 rank of candidate i (1 for the top candidate), $\phi(\cdot)$ is a rank transformation ($\phi(r) = -\log r$ for rank-log fusion), and α controls the fusion weight. Training queries whose target is absent from the stage0 pool use a "gold-added" candidate carrying a sentinel stage0 rank beyond the negative-sampling range (rank 101), which yields the weakest stage0 prior under $\phi(r) = -\log r$.

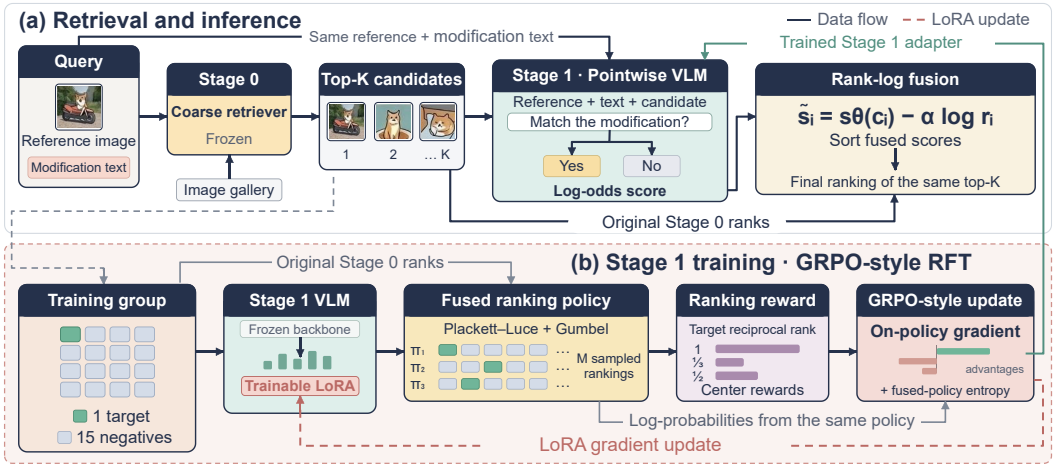

Figure 1: Overview of PVR-RFT. **(a) Retrieval and inference**: a frozen stage0 retriever (DQU-CIR or MCoT-MVS) retrieves top-K candidates from the gallery given a reference image and modification text; a Qwen3-VL-8B pointwise scorer evaluates each candidate with a Yes/No prompt, and its log-odds scores are combined with the stage0 ranks via rank-log fusion to produce the final ordering. **(b) Stage 1 training**: a GRPO-style policy over fused rankings samples candidate orderings with Gumbel noise, receives the target's reciprocal rank as reward, and updates only the LoRA parameters with on-policy policy gradients.

3.3 Pointwise VLM Scorer

Our VLM scorer takes as input the reference image, a candidate image, and the modification text, and outputs a relevance score. We use a simple pointwise prompt: "Reference image: [ref]. Candidate image: [cand]. Does the candidate match the modification: '[mod]'? Answer Yes or No." The VLM (Qwen3-VL-8B-Instruct) processes this multimodal input and produces next-token logits for the closure tokens. We extract the "Yes" and "No" token identities from the tokenizer and compute the pointwise score as the log-odds:

$$s_{\text{VLM}}(c) = \log \sum_{y \in \mathcal{Y}_{\text{Yes}}} P(y \mid I_r, c, T_m) - \log \sum_{n \in \mathcal{Y}_{\text{No}}} P(n \mid I_r, c, T_m). \tag{2}$$

We aggregate valid "Yes" token variants and "No" token variants; $\mathcal{Y}_{\text{Yes}}$ and $\mathcal{Y}_{\text{No}}$ denote the sets of valid token variants, and summing over them yields a calibrated log-odds score.

3.4 Reinforcement Fine-tuning

We train the policy π_θ that maps a group of N candidates to a distribution over rankings. Given N candidates (1 positive + $N-1$ negatives), the policy scores each candidate independently through Eq. 2 and adds the stage0 prior, yielding the training-time fused score

$$\tilde{s}_\theta(c_i) = s_\theta(c_i) + \alpha_{\text{train}}\, \phi(r_i^{\text{stage0}}), \tag{3}$$

which reuses the inference-time fusion of Eq. 1 with $\alpha = \alpha_{\text{train}}$. The policy distribution over rankings is the Plackett-Luce model Plackett (1975); Luce (2005) over these fused scores:

$$P_\theta(\pi \mid \mathcal{C}) = \prod_{j=1}^{N} \frac{\exp(\tilde{s}_\theta(c_{\pi_j})/\tau)}{\sum_{k=j}^{N} \exp(\tilde{s}_\theta(c_{\pi_k})/\tau)}, \tag{4}$$

where τ is the policy temperature. Although the prior term in Eq. 3 does not depend on θ, it enters the normalizers of Eq. 4 and therefore reshapes the ranking distribution rather than acting as a mere additive offset.

GRPO-style training. We follow the Group Relative Policy Optimization (GRPO) framework Shao et al. (2024); Guo et al. (2025). For each group of N candidates, we draw M ranking samples from

the policy P_θ exactly via the Gumbel-max trick Kool et al. (2019): adding independent Gumbel noise, $\tilde{s}(c_i) = \tilde{s}_\theta(c_i) + \epsilon_i$ with $\epsilon_i \sim \text{Gumbel}(0, \tau)$, and sorting by descending $\tilde{s}$ yields a sample $\hat{\pi}^{(m)} \sim P_\theta$. Because the rollouts and the log-probabilities in the objective below come from the same distribution of Eq. 4, the resulting policy gradient is on-policy.

Stage0-aware reward. We define the reward for a sampled ranking $\hat{\pi}$ as the reciprocal rank of the target in that ordering:

$$R(\hat{\pi}) = \frac{1}{\text{Rank}(c^* \mid \hat{\pi})}, \tag{5}$$

where $\text{Rank}(c^* \mid \hat{\pi})$ is the target's rank in the sampled ordering. The Gumbel perturbation introduces the only stochasticity, and the reward of each rollout is evaluated at the target rank in the sampled order of that rollout. Because the sampled rankings are drawn from P_θ (Eq. 4), the reward ties optimization directly to both the learned scorer and the frozen stage0 prior: improving the reward requires the policy to rank the target ahead of where the stage0 prior alone would place it.

Policy gradient. Given M sampled rankings, the advantage for rollout m is the group-centered reward $A_m = R(\hat{\pi}^{(m)}) - \bar{R}$; advantage normalization by the reward standard deviation is disabled in all reported runs. The policy gradient loss is $\mathcal{L}_{\text{RL}} = -\frac{1}{M}\sum_{m=1}^{M} A_m \cdot \log P_\theta(\hat{\pi}^{(m)} \mid \mathcal{C})$, and the full objective adds an entropy regularizer over the softmax of the fused scores, $\mathcal{L} = \mathcal{L}_{\text{RL}} - \beta_{\text{ent}} H$, with $\beta_{\text{ent}} = 0.001$.

We freeze the vision encoder and apply LoRA Hu et al. (2022) ($r = 64$, $\alpha = 128$) to all linear layers in the language model and LM head.

3.5 Training Data Construction and Negative Sampling

PVR-RFT is trained on the stage0 ranking rather than on the full gallery. For each training query, the target candidate is first located in the stage0 top-256 list. When the target is present, it is used as the positive; otherwise, the ground-truth image is appended as a "gold-added" candidate with a sentinel stage0 rank (101) beyond the top-100 range from which negatives are drawn. This ensures that the reranker still observes useful positive supervision even when the coarse retriever misses the target, and because the appended candidate carries the weakest stage0 prior and the lowest coarse score, it is ranked last under the stage0 ordering during reward computation. On FashionIQ, 5,445/5,985 Dress, 5,201/5,988 Shirt, and 5,602/6,027 Toptee training queries have their target inside the stage0 pool; the remaining 540/787/425 queries use the appended gold-added candidate.

For each positive, we sample fifteen negatives from the stage0 top-100 list. We divide ranks 1–10, 11–30, 31–60, and 61–100 into four buckets and allocate nearly equal quotas across buckets. The remaining negatives are sampled from the tail to maintain diversity. This construction deliberately exposes the model to both easy distractors and visually similar hard negatives, preventing the policy from overfitting to the coarse retriever's top candidates.

At training time, the stage0 rank is not treated as an independent label. It is normalized and used in the reward as a prior. Because the policy is trained only to produce pointwise logits, a separately trained adapter can be applied to a new stage0 retriever at inference without re-encoding the gallery or modifying the stage0 index; each adapter is trained on the stage0 rankings it will rerank.

3.6 Rank-log Fusion at Inference

At inference, the VLM produces pointwise scores for all K candidates. We fuse with stage0 ranks via Eq. 1 using $\phi(r) = -\log r$. The logarithmic transform amplifies small rank differences at top positions while compressing differences in the tail. The fusion weight α is selected per (stage0, category) pair on the validation split: for DQU-CIR we use $\alpha = 4$ on Dress and $\alpha = 8$ on Shirt and Toptee, and for MCoT-MVS we use $\alpha = 4$ on all three categories. The official CIRR test-server submissions use the same rank-log fusion with $\alpha = 2$.

Table 1: Comparison on FashionIQ VAL-Split (R@10/R@50). Published baselines use their original full-gallery protocol; our rows rerank a controlled candidate pool of 256 images per query from the named stage0 retriever, whose stage0-only reference on Dress is shown as the dotted lines in Figure 2. Bold marks the best result in each column.

Method	Dress		Shirt		Toptee	
	R@10	R@50	R@10	R@50	R@10	R@50
TIRG Vo et al. (2019)	14.87	34.66	18.26	37.89	19.08	39.62
CLIP4Cir Baldrati et al. (2022)	33.81	59.40	39.99	60.45	41.41	65.37
DQU-CIR Wen et al. (2024)	57.63	78.56	62.14	80.38	66.15	85.73
ENCODER Li et al. (2025)	51.51	76.95	54.86	74.93	62.01	80.88
MCoT-MVS Ge et al. (2026)	58.45	78.92	63.24	81.15	68.02	85.97
COMBINER Li et al. (2026)	57.96	79.08	63.64	81.84	68.18	85.87
Ours + DQU-CIR	62.87	82.25	66.44	82.09	71.29	**87.76**
Ours + MCoT-MVS	**63.11**	**82.60**	**68.55**	**83.42**	**72.87**	87.51

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. *FashionIQ* Wu et al. (2021) is a fashion-oriented CIR benchmark with three categories—Dress, Shirt, and Toptee—and natural-language modifications describing fine-grained attribute changes. Its VAL-Split contains 5,985/5,988/6,027 training queries and 2,017/2,038/1,961 validation queries for the three categories. *CIRR* Liu et al. (2021) contains 28,225 training pairs and 4,181 validation queries over open-domain, real-life images; the gallery contains 2,297 validation images. *Shoes* is a single-domain fashion benchmark containing 10,000 training pairs and 4,658 evaluation gallery images, with 1,761 validation queries in the DQU-CIR evaluation protocol Wen et al. (2024). For all datasets we follow DQU-CIR's evaluation setup and report Recall@K (R@K) for $K \in \{1, 5, 10, 50, 100\}$.

Implementation Details. We use Qwen3-VL-8B-Instruct Bai et al. (2025) as the backbone VLM. The stage0 retriever retrieves the top $K = 256$ candidates per query. All PVR-RFT rows in the main tables use this controlled top-256 candidate pool and rank-log fusion (Eq. 1) at inference, with the fusion weight selected on the validation split; fusion-weight sweeps explicitly change α. We train with 16 candidates per instance (1 positive + 15 negatives), with negatives sampled from the top-100 stage0 candidates using four equal-sized buckets and 3 random negatives from the tail. Training uses GRPO-style RL: $M = 16$ rollouts, policy temperature $\tau = 2.0$, and stage0 prior weight $\alpha_{\text{train}} = 2.0$. We use LoRA ($r = 64$, $\alpha = 128$, dropout 0.05), train for 1 epoch per category with per-device batch size 1, gradient accumulation 4 (effective batch 16), learning rate 10^{-6} under a cosine schedule with 3% warmup, and use DeepSpeed ZeRO-2 on 4×A100-80GB GPUs. Training takes approximately 45 minutes per category. Inference with 2 HF workers on 2×A100 takes approximately 4 hours per category. All reported results are from a single training run per configuration. To quantify sampling uncertainty without retraining, we resample validation queries with a paired bootstrap (2,000 resamples, fixed seed) and report 95

4.2 MAIN RESULTS

Table 1 compares PVR-RFT with representative CIR systems on FashionIQ VAL-Split. We evaluate two frozen stage0 retrievers, DQU-CIR and MCoT-MVS, and train a separate PVR-RFT adapter for each. Under the same top-256 pools, the stage0 retrievers alone reach R@10/R@50 of 56.72/79.08, 62.56/80.52, and 67.11/85.77 (DQU-CIR) and 58.50/79.62, 63.89/81.70, and 69.20/86.23 (MCoT-MVS) on Dress, Shirt, and Toptee; both configurations improve over these baselines. Published rows use their original full-gallery evaluation, while our rows use the controlled top-256 candidate pools described in Section 4. Table 2 reports CIRR results, and Table 3 reports Shoes results.

On the CIRR test server, PVR-RFT with DQU-CIR stage0 attains official metrics of R@1=63.23, R@5=87.30, R@10=92.68, and R@50=98.51, with subset R@1=84.55, R@2=94.48, and

Table 2: Detailed CIRR comparison on the official test split. Published methods report the metrics from their papers; our rows are official test-server submissions that rerank a top-256 candidate pool from the named stage0 retriever. "Subset" restricts ranking to the six images in the query image set.

Method	R@1	R@5	R@10	R@50	Rsub@1	Rsub@2	Rsub@3
CLIP4Cir Baldrati et al. (2022)	33.59	65.35	77.35	95.21	62.39	81.81	92.02
DQU-CIR Wen et al. (2024)	46.22	78.17	87.64	97.81	70.92	87.69	94.68
COMBINER Li et al. (2026)	52.60	82.51	90.12	98.17	81.33	93.06	97.23
FiRE Hou et al. (2025)	43.33	74.02	83.51	95.83	73.01	88.38	94.94
ReCALL Yang et al. (2026)	55.52	84.07	91.83	98.55	81.49	93.35	97.64
MCoT-MVS Ge et al. (2026)	54.99	84.84	91.59	98.55	79.16	92.19	96.89
Ours + DQU-CIR	63.23	87.30	92.68	98.51	84.55	94.48	97.83
Ours + MCoT-MVS	**65.40**	**89.64**	**94.55**	**99.04**	**85.47**	**95.45**	**98.43**

Table 3: Comparison on Shoes (R@1/R@10/R@50). Published baselines use their original full-gallery protocol; the PVR-RFT row uses a top-256 candidate pool from DQU-CIR and rank-log fusion.

Method	R@1	R@10	R@50
TIRG Vo et al. (2019)	12.60	45.45	69.39
VAL Chen et al. (2020)	17.18	51.52	75.83
CLVC-Net Wen et al. (2021)	17.64	54.39	79.47
ARTEMIS Delmas et al. (2022)	18.72	53.11	79.31
TG-CIR Wen et al. (2023)	25.89	63.20	85.07
DQU-CIR Wen et al. (2024)	31.47	69.19	88.52
ENCODER Li et al. (2025)	26.97	65.59	86.48
COMBINER Li et al. (2026)	32.37	70.81	88.81
Ours + DQU-CIR	**36.40**	**74.39**	**91.08**

R@3=97.83. The separately trained MCoT-MVS adapter, evaluated over an independently generated MCoT-MVS top-256 candidate pool, attains R@1=65.40, R@5=89.64, R@10=94.55, and R@50=99.04, with subset R@1=85.47, R@2=95.45, and R@3=98.43. Furthermore, the same candidate pool also yields the official raw MCoT-MVS baseline of R@1=54.99, R@5=84.84, R@10=91.59, R@50=98.55, and subset R@1=79.16, R@2=92.19, R@3=96.89. Both configurations improve substantially over their respective raw stage0 baselines (DQU-CIR: R@1 46.22→63.23; MCoT-MVS: R@1 54.99→65.40). We regard this retriever-agnostic adaptability—the same training and fusion recipe improving whatever stage0 retriever it is given—as the main takeaway from the CIRR experiments, rather than fine-grained ranking against published full-gallery systems.

Pointwise reranking leaves the coarse retriever frozen and only adapts the lightweight VLM reranker. Rather than treating DQU-CIR as canonical and MCoT-MVS as an auxiliary configuration, we report both configurations in parallel because their relative ranking depends on the evaluation depth. For MCoT-MVS, we reproduce it under the same FashionIQ VAL-Split protocol, generate its top-256 candidate lists, and train a separate PVR-RFT adapter with the identical rank-log fusion procedure.

Specifically, across Dress, Shirt, and Toptee, PVR-RFT improves DQU-CIR top-256 R@10 by +6.15, +3.88, and +4.18 absolute points, and MCoT-MVS top-256 R@10 by +4.61, +4.66, and +3.67 points. The corresponding R@50 gains are +3.17/+1.57/+1.99 for DQU-CIR and +2.98/+1.72/+1.28 for MCoT-MVS. In this controlled setup, MCoT-MVS provides stronger stage0 recall at both R@10 and R@50, and the MCoT-MVS-based reranker yields the best deep-rank results on Dress, while the DQU-CIR-based reranker remains competitive. In both cases, PVR-RFT improves the stage0 ranking, especially at R@10. This indicates that the VLM reranker learns a complementary signal that is not specific to DQU-CIR's embedding space.

The comparison also reveals a subtle tradeoff. Because stage0 recall is not perfect, some targets are absent from the top-256 candidate pool. A stronger stage0 retriever is therefore not only a better initialization for fusion but also raises the upper bound of the reranker. We observe that the

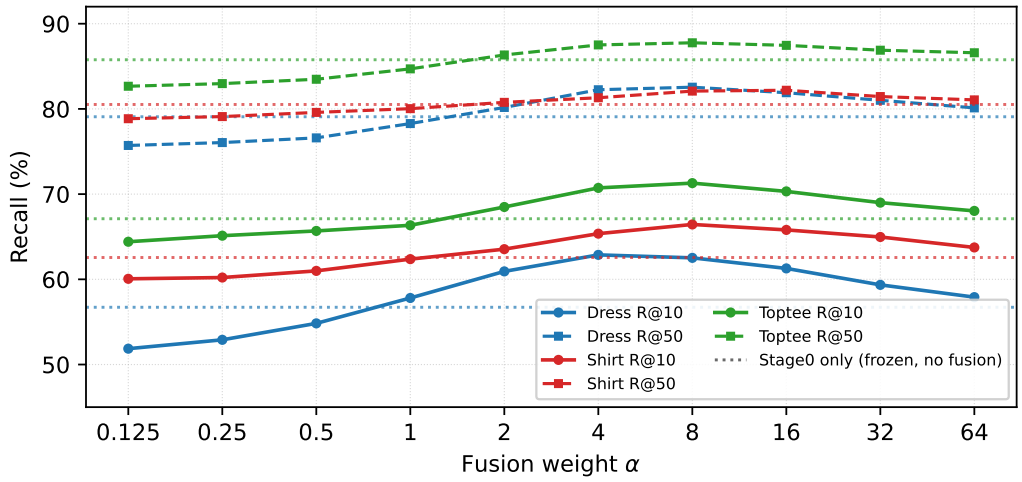

Figure 2: Effect of fusion weight α on FashionIQ Dress, Shirt, and Toptee (DQU-CIR stage0). All three categories are measured at the same ten weights $\alpha \in \{0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64\}$. Dotted lines give the frozen stage0 retriever alone (no fusion); they are flat across α because the stage0 ranking does not depend on the fusion weight.

best stage0/reranker combination depends on the evaluation depth: MCoT-MVS plus PVR-RFT is strongest on Dress and Shirt at every cutoff and at Toptee R@10, whereas DQU-CIR plus PVR-RFT retains a slight advantage at Toptee R@50 (87.76 vs. 87.51). We therefore report both configurations rather than treating a single coarse retriever as canonical.

4.3 Ablation Studies

Effect of fusion weight α. Figure 2 shows R@10 and R@50 as a function of $\alpha \in \{0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64\}$. The grid excludes $\alpha = 0$; we report stage1-alone separately as the no-fusion reference, and the smallest swept weight $\alpha = 0.125$ already exceeds stage1-alone R@10 on Dress (51.86 vs. 51.02). As α increases, R@10 rises, peaking at $\alpha = 4$ (Dress, 62.87), $\alpha = 8$ (Shirt, 66.44), and $\alpha = 8$ (Toptee, 71.29), before declining toward the stage0 baseline as the VLM signal is suppressed. We note that Shirt R@50 peaks later, at $\alpha = 16$ (82.19 vs. 82.09 at the reported $\alpha = 8$); the reported weights follow the R@10-optimal choice.

Module ablation. The two components of PVR-RFT are already visible in Figure 2: the dotted lines are the frozen stage0 retriever, the $\alpha \to 0$ limit is the stage1 VLM scorer alone, and the curves are the fused ranking. On Dress, pointwise scoring alone reaches R@1=23.90, R@5=41.60, R@10=51.02, R@50=75.41, and R@100=84.04 (pool target coverage 93.31%)—weaker than stage0 at every cutoff, showing that independent VLM logits do not automatically refine retrieval. After rank-log fusion ($\alpha = 4$), it improves stage0 R@1 from 24.44 to 28.85, R@5 from 46.80 to 53.15, R@10 from 56.72 to 62.87, R@50 from 79.08 to 82.25, and R@100 from 86.27 to 88.40. A full-pool pairwise comparison is infeasible (32,640 pairs per query, roughly 65 million comparisons for the 2,017 Dress queries), so Table 4 compares the two forms on a pool-matched top-16 set, where pointwise wins and costs $\mathcal{O}(K)$ forwards instead of $\mathcal{O}(K^2)$.

Reranking-form comparison. Table 4 compares pointwise and pairwise stage1 scoring on a pool-matched top-16 candidate set, where target coverage is R@16=66.68%. Pointwise is trained with the proposed GRPO-style objective on 512px inputs; pairwise is trained with a Bradley–Terry classification loss on candidate pairs with 256px inputs, so the comparison reflects each form's default configuration rather than a fully matched training protocol. Candidate ranks for fusion come from the stage0 top-256 export and the top-16 subset preserves the original stage0 ordering; rows are computed on the query subset shared by the top-16 export and both scorers rather than all 2,017 validation queries, which is why the stage0 row reads slightly higher than the full-pool stage0 numbers in the main table. With the final strong adapter, pointwise consistently outperforms pairwise before

Table 4: Stage1 reranking-form comparison on FashionIQ Dress over the pool-matched top-16 candidate set (target coverage R@16=66.68%). The last column reports the number of VLM forward passes per query.

Model	R@1	R@5	R@10	VLM forwards
Stage0 top-16	24.44	47.30	57.46	0
Pointwise	26.23	49.83	60.88	$\mathcal{O}(K)$
Pointwise + rank-log	**28.85**	**53.05**	**62.07**	$\mathcal{O}(K)$
Pairwise	19.29	41.94	57.21	$\mathcal{O}(K^2)$
Pairwise + rank-log	26.97	49.88	60.09	$\mathcal{O}(K^2)$

Table 5: Comparison of rank fusion transforms on FashionIQ Dress. Best result per transform is in bold.

	rank-inv				rank-log			
α	R@1	R@10	R@50	R@100	R@1	R@10	R@50	R@100
1	24.74	51.56	75.66	84.13	27.32	57.81	78.28	86.27
2	25.24	52.80	75.71	84.38	28.46	60.93	80.17	87.41
4	26.18	54.34	76.20	84.83	**28.85**	**62.87**	82.25	88.40
8	28.51	56.62	77.00	85.42	28.26	62.52	**82.55**	**88.45**
16	28.71	58.80	78.14	85.87	26.62	61.28	81.90	88.30
32	26.13	60.93	79.23	86.47	24.74	59.35	81.01	87.61
64	24.49	62.02	80.47	87.26	24.44	57.91	80.12	87.11

and after fusion: at the top-16 pool, pointwise reaches R@1=28.85 vs. 26.97, R@5=53.05 vs. 49.88, R@10=62.07 vs. 60.09 after rank-log fusion, and is also stronger without fusion (R@1=26.23 vs. 19.29, R@5=49.83 vs. 41.94, R@10=60.88 vs. 57.21). Pointwise is also more scalable at the operating point used in our main results: scoring all candidate pairs in a top-K pool requires $\mathcal{O}(K^2)$ forwards, e.g., 32,640 pairs per query at $K = 256$, whereas pointwise needs only $\mathcal{O}(K)$ forwards. This scaling gap is exactly why PVR-RFT is evaluated at top-256 rather than restricted to a small pairwise pool.

Fusion method comparison. Table 5 compares two rank transformations: rank-inv uses $\phi(r) = 1/r$, which decays sharply after the first few positions, while rank-log uses $\phi(r) = -\log r$, which preserves more signal from mid-ranked candidates. We sweep both transforms over the same α values on Dress, extending the grid to $\alpha = 64$: rank-inv keeps improving but stays below rank-log at every cutoff (R@10 62.02 vs. 62.87), and rank-log peaks earlier at $\alpha = 4$ with the strongest R@10/R@50/R@100, indicating that logarithmic compression keeps tail rank outliers from dominating while preserving the coarse-retriever prior.

5 Conclusion

We presented PVR-RFT, a pointwise VLM reranker for CIR trained with reinforcement fine-tuning and a stage0-aware reward. Rank-log fusion combines the complementary strengths of a coarse retriever and learned pointwise scores, and because the protocol consumes only top-K candidate lists, the same recipe attaches to different retrievers: under controlled top-256 pools PVR-RFT consistently improves DQU-CIR on FashionIQ, Shoes, and CIRR, and MCoT-MVS on FashionIQ and CIRR. Future work includes listwise architectures that compare candidates jointly and reducing inference cost through cascaded or cached scoring.

AI USE STATEMENT

AI-assisted tools were used for code development, experiment analysis, and editorial assistance. All AI-assisted content was reviewed by the authors, who take responsibility for the final paper and artifacts.

ETHICS STATEMENT

This work uses only publicly available image-retrieval benchmarks (FashionIQ, CIRR, and Shoes) that contain publicly released consumer and fashion images. No human subjects were involved, and no personal or sensitive data was collected. All models and baselines were trained and evaluated under each dataset's intended research use. The proposed system targets content-based image retrieval research; it does not introduce risks beyond those of standard retrieval systems.

REPRODUCIBILITY STATEMENT

All experiments rely on publicly available datasets and pre-trained models. The complete source code, training scripts, and evaluation configurations used in this work are available in the anonymous repository at https://anonymous.4open.science/r/PVRRFT-code, including the training-group construction with gold-added candidates for queries whose target falls outside the stage0 pool. The official CIRR test-server submissions apply rank-log fusion with weight $\alpha = 2$ over the submitted orderings. We note that the pairwise baseline in the reranking-form comparison is trained with a Bradley–Terry objective rather than RL. The candidate-pool construction, reward design, and evaluation protocol (including full-gallery and pool-matched settings) are described in Sections 3 and 4.

REFERENCES

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization. *Text Reading, and Beyond*, 2(1):1, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21434–21442. IEEE, 2022.

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15292–15301. IEEE, 2023.

Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2998–3008. IEEE, 2020.

Ginger Delmas, Rafael Sampaio de Rezende, Gabriela Csurka, and Diane Larlus. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

Xuri Ge, Chunhao Wang, Xindi Wang, Zheyun Qin, Zhumin Chen, and Xin Xin. Mcot-mvs: Multi-level vision selection by multi-modal chain-of-thought reasoning for composed image retrieval. In *Proceedings of the ACM Web Conference 2026*, pp. 2105–2113, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Bohan Hou, Haoqiang Lin, Xuemeng Song, Haokun Wen, Meng Liu, Yupeng Hu, and Xiangyu Zhao. Fire: Enhancing mllms with fine-grained context learning for complex image retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 803–812, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

Haonan Jiang, Yuji Wang, Yongjie Zhu, Xin Lu, Wenyu Qin, Meng Wang, Pengfei Wan, and Yansong Tang. Embed-rl: Reinforcement learning for reasoning-driven multimodal embeddings. In *European Conference on Computer Vision*, pp. 39–56. Springer, 2026.

Wouter Kool, Herke Van Hoof, and Max Welling. Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement. In *International conference on machine learning*, pp. 3499–3508. PMLR, 2019.

Zixu Li, Zhiwei Chen, Haokun Wen, Zhiheng Fu, Yupeng Hu, and Weili Guan. Encoder: Entity mining and modification relation binding for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5101–5109, 2025.

Zixu Li, Yupeng Hu, Zhiwei Chen, Haokun Wen, Xuemeng Song, and Liqiang Nie. Combiner: Composed image retrieval guided by attribute-based neighbor relations. *IEEE Transactions on Image Processing*, 2026.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2105–2114. IEEE, 2021.

R Duncan Luce. *Individual choice behavior: A theoretical analysis*. Courier Corporation, 2005.

Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. Multi-stage document ranking with bert. *arXiv preprint arXiv:1910.14424*, 2019.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202, 1975.

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

Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6432–6441. IEEE, 2019.

Haokun Wen, Xuemeng Song, Xin Yang, Yibing Zhan, and Liqiang Nie. Comprehensive linguistic-visual composition network for image retrieval. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pp. 1369–1378, 2021.

Haokun Wen, Xian Zhang, Xuemeng Song, Yinwei Wei, and Liqiang Nie. Target-guided composed image retrieval. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 915–923, 2023.

Haokun Wen, Xuemeng Song, Xiaolin Chen, Yinwei Wei, Liqiang Nie, and Tat-Seng Chua. Simple but effective raw-data level multimodal fusion for composed image retrieval. In *Proceedings of the 47th International ACM SIGIR conference on research and development in information retrieval*, pp. 229–239, 2024.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11302–11312. IEEE, 2021.

Tianyu Yang, ChenWei He, Xiangzhao Hao, Tianyue Wang, Jiarui Guo, Haiyun Guo, Leigang Qu, Jinqiao Wang, and Tat-Seng Chua. Recall: Recalibrating capability degradation for mllm-based composed image retrieval. *arXiv preprint arXiv:2602.01639*, 2026.

Shengwei Zhao, Jingwen Yao, Sitong Wei, Linhai Xu, Yuying Liu, Dong Zhang, Zhiqiang Tian, and Shaoyi Du. Mmrag-rft: Two-stage reinforcement fine-tuning for explainable multi-modal retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34968–34975, 2026.

Junjie Zhou, Yongping Xiong, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, and Defu Lian. Megapairs: Massive data synthesis for universal multimodal retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19076–19095, 2025.

Hongguang Zhu, Yunchao Wei, Yao Zhao, Chunjie Zhang, and Shujuan Huang. Amc: Adaptive multi-expert collaborative network for text-guided image retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(6):1–22, 2023.