EviRank: Structured Relevance Evidence for Multimodal Image Re-ranking

Enjun Du[1,2,3,†], Siyi Liu[1], Zirong Chen[1,2,4†], Xinyu Zuo[2], Jinwen Luo[2], Ruiwen Tao[2], Lisheng Duan[2], Haijin Liang[2], Jin Ma[2], Junfu Pu[5], Yongqi Zhang[1,*]
[1]The Hong Kong University of Science and Technology (Guangzhou) [2]Yuanbao Team, Tencent, [3]The University of Hong Kong, [4]Tsinghua University, [5]ARC Lab, Tencent
{enjundu, yongqizhang}@hkust-gz.edu.cn

Abstract

Real-world image search queries are multimodal and compositional: "find this shirt in pink" specifies an entity to retain, an attribute to modify, and context to ignore. Yet existing re-rankers either compress such multi-faceted relevance into an opaque embedding or rely on free-form chain-of-thought that easily omits or hallucinates fine-grained constraints. Drawing on rubric- and checklist-based evaluation from NLP, we recast multimodal image re-ranking as a **semantic constraint satisfaction** problem and propose **EviRank**, which parses any query—text-only, image-only, or composed—into a unified *evidence package*: typed criteria across six semantic slots (e.g., entities, attributes, relations), each labelled *required*, *forbidden*, or *ignorable*. Re-ranking then reduces to **evidence-conditioned verification**, combining deterministic rubric scoring and evidence-grounded listwise comparison in a single training-free procedure. The explicit evidence can further serve as structured supervision for optionally distilling a lightweight student. Across five benchmarks spanning text-to-image, image-to-image, and composed image retrieval, EviRank achieves state-of-the-art performance, and the distilled student preserves over 90% of the teacher's capability at substantially lower cost. Code is available at https://github.com/EnjunDu/EviRank.

1 Introduction

Real-world image search queries are rarely single-faceted. A user asking "find *this shirt* in *pink*" is implicitly specifying multiple semantic constraints at once: the system must *retain* the entity "shirt" together with its visual identity (cut, pattern, fit), *modify* a specific attribute (color → pink), and *ignore* incidental factors such as background or lighting in the reference image (Xing et al., 2025; Tang et al., 2025a; Tian et al., 2025). Such queries naturally combine information from multiple modalities—text, image, or both—and constrain the target along several semantic dimensions, such as *entities*, *attributes*, *actions*, *relations*, and discriminative *details*. Whether the query is purely textual, purely visual, or a mixture of the two, the underlying problem is the same: deciding whether each candidate image satisfies the set of fine-grained semantic constraints that the query imposes.

Existing multimodal re-rankers, however, do not explicitly model these constraints. *Representation-centric* approaches—ranging from classical dual-encoder embeddings (Frome et al., 2013) to recent MLLM-based scorers (Liu et al., 2025c; Huynh et al., 2025; Sun et al., 2025b)—collapse all semantic dimensions of a query into a single dense vector. Such an opaque representation cannot tell which evidence should be preserved, modified, or ignored, and offers no handle on fine-grained constraint violations. *Reasoning-centric* approaches (Sun et al., 2025a; Wu et al., 2025) instead ask an MLLM to reason about relevance in free-form chain-of-thought (CoT). While more expressive, this free-form reasoning remains *unstructured* natural language, and exhibits three concrete failure modes. First, it *omits* implicit constraints: when a user says "find this shirt in pink," a CoT trace typically discusses the color change while silently dropping the "preserve the original cut and pattern" constraint. Second, it *hallucinates* evidence, attributing visual properties the model has not actually verified. Third, its semantic coverage is *inconsistent across queries*, so completeness and verifiability of the relevance judgement cannot be guaranteed without a formal structure specifying which dimensions must be examined. These failure modes share a common root: multimodal image re-ranking is fundamentally a **semantic constraint satisfaction** problem, not a generic similarity-matching problem, and methods that do not represent constraints

[†]Work done during an internship at Tencent.
[*]Corresponding author.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 7640–7666
October 24-29, 2026 ©2026 Association for Computational Linguistics

explicitly cannot systematically verify them.

We therefore propose to represent multimodal relevance as a **structured Evidence Frame**. Instead of free-form reasoning, we first *parse the query into typed evidence statements*—REQUIRED, FORBIDDEN, and IGNORABLE—organized along six semantic slots (entities, attributes, actions, relations, scene, and key details; full definitions in Section 3.1). The three labels mirror well-studied semantic relations (entailment, contradiction, and invariance, respectively; Section 3.1). Since the Evidence Frame is modality-agnostic, text-only, image-only and composed queries share a single evidence interface. Re-ranking then reduces to *verifying* which constraints each candidate satisfies or violates—a slot-wise, language-grounded judgement that is transparent and deterministically aggregable.

Building on this representation, **EviRank** re-ranks top-K candidates in two stages—**evidence mining** and **evidence verification**—both training-free. The same slot-wise satisfaction scores and hard-pair distinctions also serve as naturally decomposable supervision for distilling a lightweight student re-ranker. Our core contributions are:

- **A new formulation.** We formulate multimodal image re-ranking as *evidence-conditioned semantic verification*, where text-only, image-only and composed queries are converted into a shared *Evidence Frame* of typed relevance criteria (required / forbidden / ignorable, instantiated as entailment / contradiction / invariance).
- **Evidence mining and evidence verification.** We design a training-free, evidence-conditioned re-ranking procedure that mines slot-wise required, forbidden, and ignorable evidence, and verifies it through a deterministic rubric grounded in the same evidence.
- **Evidence as supervision.** We show that the structured evidence and its slot-wise verification signals form a naturally decomposable supervision source, which can be distilled into a compact student re-ranker.
- **Experiments.** State-of-the-art results across five benchmarks, with analysis showing the structured evidence is robust, interpretable, and contributes beyond free-form augmentation.

2 Related Work

Implicit Relevance Representations for Multimodal Retrieval. A first family encodes queries and candidate images into a joint embedding space and ranks by similarity (Frome et al., 2013; Liu et al., 2024; Levy et al., 2024; Bai et al., 2024); recent MLLM-based re-rankers continue this paradigm by distilling multimodal capability into improved embeddings or learned scoring heads (Liu et al., 2025c; Huynh et al., 2025; Sun et al., 2025b; Lin et al., 2025; Chen et al., 2024), while a complementary line reformulates retrieval as document-identifier generation (Tay et al., 2022; Wang et al., 2022; Li et al., 2024). All of them represent relevance *implicitly*—a scalar similarity, an opaque embedding, or a learned ID distribution—and cannot tell which semantic dimensions of a query a candidate satisfies or violates. Our work is orthogonal: we operate on top of any such retriever and replace its implicit score with explicit, slot-wise evidence verification.

Explicit mechanism	Image Scope	CoT RR	CoT MR	Ours
Fixed, reusable slot taxonomy	○	✗	✗	✓
Typed FORBIDDEN evidence	✗	✗	○	✓
Typed IGNORABLE evidence	✗	✗	✗	✓
One frame for T→I/I→I/CIR	○	✗	✗	✓
Deterministic auditable scoring	✗	✗	✗	✓
Teacher-free distilled student	✗	✗	✗	✓

Table 1: Novelty delineation against the closest reasoning-based re-rankers. ✓: reported as an explicit, fixed and reusable representational mechanism; ○: partially or implicitly present; ✗: not reported as such a mechanism—which does not deny that the underlying MLLM may implicitly possess the capability.

Free-form Reasoning for Image Re-ranking. A second family augments MLLM-based re-ranking with CoT reasoning (Sun et al., 2025a; Wu et al., 2025; Tang et al., 2025b; Luo et al., 2025). More expressive than a single similarity score, the reasoning trace nonetheless remains *unstructured natural language* and exhibits the omission, hallucination, and inconsistent-coverage failures discussed in Section 1. EviRank instead asks the teacher for a typed Evidence Frame (required / forbidden / ignorable over six semantic slots) that can be verified, aggregated, and distilled.

Novelty delineation. Our claim is *representational* rather than about what MLLMs can reason

about. ImageScope decomposes a query into a multi-granularity hierarchy, CoTRR into five reasoning components and CoTMR into predefined verification subtasks, yet in all three the intermediate product is free-form text whose fields are neither typed nor guaranteed to recur across queries. EviRank fixes a six-slot taxonomy and attaches an explicit polarity to every statement, which is what makes deterministic aggregation (Eqs. 2–3), per-slot auditing and reuse as typed supervision possible (Table 1; formulation-level comparison with ImageScope in Appendix N).

Structured Criteria for Semantic Verification. The NLP community has increasingly turned to structured rubric- and checklist-based evaluation, where complex semantic judgements are decomposed into independently checkable criteria (Liu et al., 2023; Cook et al., 2024; Lee et al., 2025; Wei et al., 2025), and rubric criteria are further used as RL reward signals (Zhou et al., 2025; Arora et al., 2025). A closely related line decomposes complex *claims* into verifiable sub-units for fact checking and faithfulness evaluation (Thorne et al., 2018; Min et al., 2023; Golovneva et al., 2023), while a further line makes the intermediate structure itself explicit and inspectable—as a provenance-constrained evidence ledger, as reasoning paths or relational digraphs over a graph (Du et al., 2026, 2025; Zhang et al., 2026, 2023), as a semantics-preserving hierarchy over a text-attributed hypergraph (Pan et al., 2025), or as a structure shared between a neural reasoner and an LLM (Liu et al., 2025a). EviRank imports the same *decompose-then-verify* principle into multimodal image re-ranking: a multimodal query becomes typed evidence statements over six semantic slots, and relevance is established by slot-wise, language-grounded verification rather than by opaque similarity or free-form CoT.

3 Method

We present **EviRank** (Figure 1), which recasts multimodal image re-ranking as *semantic constraint satisfaction*: a query is parsed into a unified package of structured evidence criteria, and candidates are re-ranked by checking which criteria they satisfy or violate. Section 3.1 defines the setting and evidence representation; Sections 3.2 and 3.3 describe evidence mining and evidence-conditioned verification and re-ranking; and Section 3.4 shows how the same evidence serves as decomposable supervision for distilling a lightweight re-ranker.

3.1 Setting and Evidence Representation

Setting. We consider multimodal image re-ranking. A query q may consist of text, an image, or both—without loss of generality, we write $q = (t, I_{\text{ref}})$ where either component may be empty. Given q and a top-K candidate set $\mathcal{C}_K = \{c_1, \ldots, c_K\}$ produced by any off-the-shelf retriever, the goal of re-ranking is to output a permutation π over $\mathcal{C}_K$ that better reflects semantic relevance.

From queries to evidence. Our central observation is that, regardless of how q is expressed, it implicitly defines a set of *semantic constraints* on the target image: some elements must be present, others must not, and many are simply irrelevant. We therefore represent relevance not by a similarity score, but by a typed **evidence package**

$$\mathcal{E}(q) = \big(g,\ \{(R_s, F_s, I_s)\}_{s\in\mathcal{S}}\big), \tag{1}$$

where g is a concise global summary of the query intent and $\mathcal{S}$ is a lightweight default schema of six semantic slots,

$$\mathcal{S} = \left\{\begin{matrix}\textsc{Entities}, \textsc{Attributes}, \textsc{Actions},\\ \textsc{Relations}, \textsc{Scene}, \textsc{KeyDetails}\end{matrix}\right\}.$$

Slots are allowed to remain empty when the query does not express the corresponding factor (e.g., a still-life caption typically populates ENTITIES and ATTRIBUTES but leaves ACTIONS and RELATIONS empty), and the schema can be extended with additional slots for specialized domains without changing the verification procedure. Each slot s contributes three lists of short, checkable sentences: *required* constraints R_s that must hold, *forbidden* constraints F_s that must not hold, and *ignorable* constraints I_s that should not influence the decision (Appendix M draws an analogy between this typing and classical SQL constraints).

Unifying heterogeneous query forms. A key property of this representation is that it is *modality-agnostic*: although users may issue queries in very different forms, the resulting evidence package always lives in the same schema. A purely textual query becomes evidence whose required slots reflect what the user verbalizes and whose ignorable slots absorb under-specified context. A purely visual query is first captioned and then converted to

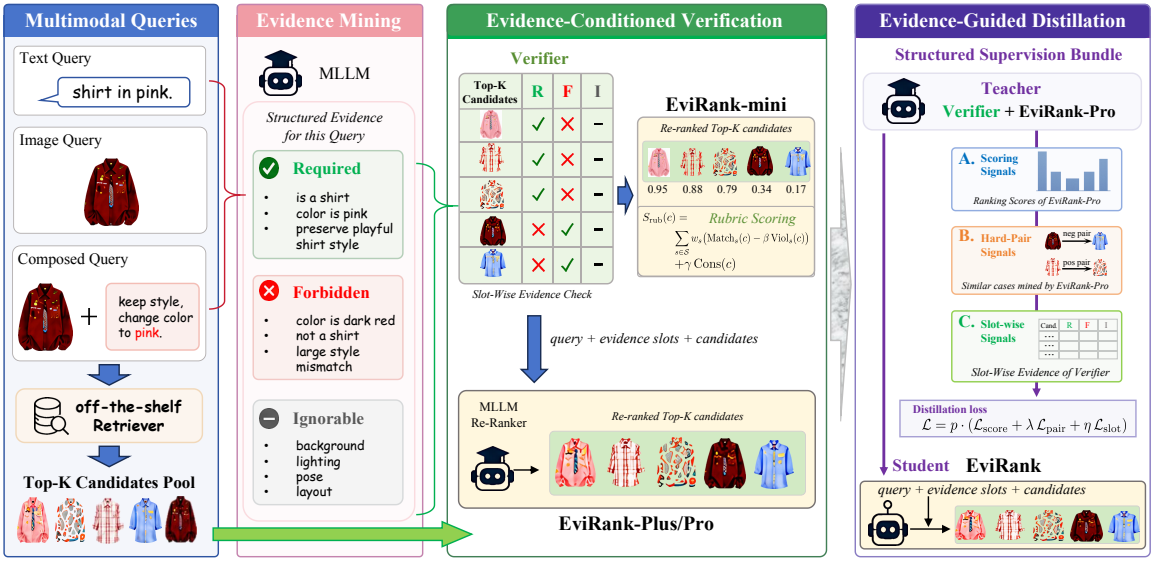

Figure 1: **Overview of EviRank.** A multimodal query (text-only, image-only, or composed) and top-K candidates from any retriever are first parsed by an MLLM teacher into a structured Evidence Frame of typed *required* / *forbidden* / *ignorable* statements over six semantic slots. Candidates are then re-ranked by evidence-conditioned verification (deterministic rubric + MLLM listwise refinement). The same evidence yields a structured supervision bundle that is distilled into a lightweight student; at inference the student consumes only the raw query and candidate image.

evidence in which the reference image's salient content becomes the required constraints. A composed query (image + modification text) is mapped to evidence by treating the reference image as a source of *preserved* required constraints (e.g., shirt cut and pattern) and the modification text as a source of *updated* required constraints together with newly induced forbidden constraints (e.g., original color → pink, so "the shirt is of the original color" becomes forbidden). After this normalization, all subsequent stages of EviRank operate on the same evidence schema, so heterogeneous query forms can be handled through a shared evidence-verification interface rather than paradigm-specific re-ranking pipelines.

3.2 Query-to-Evidence Mining

Building the evidence package $\mathcal{E}(q)$ proceeds in two training-free steps performed by a single multimodal model (referred to as the MLLM teacher throughout this paper).

Query normalization. Different query forms expose different amounts of information. Text queries are explicit but often under-specified; image queries are visually rich but lexically unanchored; composed queries entangle both. We therefore first normalize q into a textual representation that exposes its semantic content in a uniform way. For text queries, we *verbalize* t into $\tilde{t}$ that makes implicit visual details (identity, scene, salient attributes) explicit, while preserving the original intent. For image queries, we generate a concise caption together with key objects. For composed queries, we combine the reference caption with the modification text and explicitly mark which aspects should be *preserved* versus *modified*. This step is training-free; it serves purely as semantic enrichment for downstream evidence extraction, *not* as an additional retrieval pass.

Evidence extraction. Given the normalized query, the MLLM emits the full evidence package $\mathcal{E}(q)$ in a single structured pass. For each slot $s \in \mathcal{S}$, the model produces 2–4 short affirmative sentences for each of R_s, F_s, and I_s. Critically, even forbidden constraints are written as positive statements that describe what an *incorrect* match would look like (e.g., "the shirt is red"), so that every constraint can be compared to a candidate by the same matching operator and yields sign-consistent scores during verification. Ignorable constraints explicitly enumerate non-discriminative variations (e.g., minor shade shifts, background lighting) that should be tolerated rather than penalized. The full prompts used for query normalization and evidence extraction are in Appendix H. The six slots cover the principal semantic axes of

multimodal queries (*entities*, *attributes*, *actions*, *relations*, *scene*, *key details*) and are empirically near-orthogonal (mean pairwise SBERT similarity 0.18; full slot-semantics analysis in Appendix L); the schema is a default rather than a hard requirement (cf. Section 3.1).

3.3 Evidence-Conditioned Verification and Re-ranking

Once $\mathcal{E}(q)$ is mined, re-ranking reduces to *verifying* which constraints each candidate satisfies. We combine a deterministic rubric for stable, interpretable per- candidate scoring with an evidence-grounded listwise step for global, comparative judgement; both operate on the same evidence.

Deterministic rubric scoring. Let $m(\cdot, c) \in [0, 1]$ be a matching operator that measures whether a constraint sentence holds for c. We instantiate m either as embedding-based cosine similarity using a pretrained vision–language encoder, or as a teacher-graded score. The slot satisfaction and violation rates are

$$\begin{aligned} \text{Match}_s(c) &= \frac{1}{|R_s|} \sum_{r \in R_s} m(r, c), \\ \text{Viol}_s(c) &= \frac{1}{|F_s|} \sum_{f \in F_s} m(f, c). \end{aligned} \tag{2}$$

and the rubric score aggregates them with a forbidden penalty and an optional cross-modal consistency term:

$$\begin{aligned} S_{\text{rub}}(c) = &\sum_{s \in \mathcal{S}} w_s \big(\text{Match}_s(c) - \beta\, \text{Viol}_s(c)\big) \\ &+ \gamma\, \text{Cons}(c). \end{aligned} \tag{3}$$

Slot weights w_s are uniform; $\beta = 0.75$ controls the strength of forbidden penalties; and $\text{Cons}(c)$, used only when the query has both text and image components, measures alignment between the expanded query text $\tilde{t}$ and c via a pretrained encoder. Ignorable constraints I_s are not scored directly; instead they mask required constraints whose embedding is too close to any ignore statement, preventing background or stylistic cues from dominating the score.

Evidence-grounded listwise refinement. The rubric is stable and interpretable but evaluates candidates independently and therefore cannot easily resolve close calls between visually similar candidates. We address this with an evidence-grounded listwise step. The teacher is shown the query, the global intent g together with the *required* and *forbidden* constraints of $\mathcal{E}(q)$, and the candidates jointly, and is asked to return a permutation in which required constraints raise rank when clearly satisfied and forbidden constraints heavily penalize salient mismatches. *Ignorable* constraints are not exposed to the listwise prompt; they act as an invariance mask inside the deterministic rubric only, keeping the prompt compact and avoiding asking the teacher to reason over non-discriminative details. In practice, we sort candidates by S_{rub} and invoke the teacher only to verify and locally reorder the top-M (M=5 by default), which substantially reduces cost while preserving accuracy. For larger candidate pools, candidates are processed in batches and merged in a divide-and-conquer fashion (details in Appendix A).

A by-product: structured supervision. Beyond the final ranking, the listwise step naturally produces a JSON-formatted artifact containing per-candidate relevance scores, a self-assessed confidence, and a short list of *hard pairs* (closest competitors with brief distinguishing reasons). These auxiliary signals are not needed for inference but are reused as supervision in the optional distillation stage (Section 3.4).

3.4 Evidence as Supervision: Structured Distillation

Beyond verification, the structured evidence package $\mathcal{E}(q)$ also serves as a *structured supervision signal*. Unlike free-form CoT traces, which a smaller student must imitate verbatim, every component of our pipeline is typed and decomposable, letting us turn the teacher's behaviour into a precisely-targeted supervision bundle for the student.

For each query, the verification stage yields a query-conditioned *supervision bundle* comprising (i) the evidence package $\mathcal{E}(q)$, (ii) per-candidate slot-wise satisfaction and violation rates, (iii) calibrated listwise scores $\mathbf{r} = (r_1, \ldots, r_K)$, (iv) a self-assessed confidence $p \in [0, 1]$, and (v) a set of hard pairs $\mathcal{H} = \{(i, j)\}$ with distinguishing constraints. Every component has a fixed schema and a clear semantic role, which condenses the teacher's behaviour into a compact reusable training signal—in the spirit of condensing data into small training surrogates (Xu et al., 2026)—and makes the supervision transferable to a smaller model.

Student input: evidence as training-time supervision only. We distil this evidence-grounded behaviour into a lightweight student re-ranker. Crucially, at *inference time* the student takes only the raw query (text and/or reference image) and a candidate image as input, and predicts a relevance score $s_i \in [0, 1]$. The evidence package $\mathcal{E}(q)$, slot-wise satisfaction labels, listwise scores, hard pairs, and confidence are used *only as supervision during training* and are never required at deployment. The student therefore runs without any teacher MLLM, evidence generator, or runtime evidence cache.

Training objectives. Three objectives mirror the three structured components of the bundle. Let $\mathbf{s} = (s_1, \ldots, s_K)$ be the student scores for the same list of K candidates. *Score distillation* standardizes the teacher scores per query, $\hat{r}_i = (r_i - \mu_r)/(\sigma_r + \epsilon)$ with μ_r, σ_r the mean and standard deviation of $\mathbf{r}$ and $\epsilon = 10^{-8}$, and matches the two listwise distributions at temperature τ:

$$\mathcal{L}_{\text{score}} = \mathrm{KL}\big(\mathrm{softmax}(\hat{\mathbf{r}}/\tau) \,\|\, \mathrm{softmax}(\mathbf{s}/\tau)\big).$$

Hard-pair supervision imposes a margin δ on every teacher-flagged pair $(i, j) \in \mathcal{H}$ in which i should outrank j, weighted inversely to the teacher's score gap so that closely-scored competitors dominate the gradient:

$$w_{ij} = \frac{1}{|r_i - r_j| + \epsilon},$$
$$\mathcal{L}_{\text{pair}} = \sum_{(i,j)\in\mathcal{H}} w_{ij} \max\big(0,\ \delta - (s_i - s_j)\big).$$

Slot-wise supervision is optional: when slot-wise labels $y_{s,i} = \mathbf{1}[\mathrm{Match}_s(c_i) > \theta]$ are available, a lightweight auxiliary head predicts $\hat{y}_{s,i}$ for each slot $s \in \mathcal{S}$ and is trained by binary cross-entropy,

$$\mathcal{L}_{\text{slot}} = -\frac{1}{K|\mathcal{S}|}\sum_{i,s}\big[y_{s,i} \log \hat{y}_{s,i} + (1 - y_{s,i}) \log(1 - \hat{y}_{s,i})\big],$$

so that the student internalizes *which* constraints make a candidate a good match rather than only its position. The three terms are combined and modulated by the teacher's confidence p, so that uncertain supervision contributes less:

$$\mathcal{L} = p \cdot \big(\mathcal{L}_{\text{score}} + \lambda\, \mathcal{L}_{\text{pair}} + \eta\, \mathcal{L}_{\text{slot}}\big). \quad (4)$$

The auxiliary head is discarded after training, so inference cost is unchanged. Values of $\tau, \delta, \lambda, \eta$ and further implementation details are given in Appendix B.

VLM	Methods	R@1	R@5	R@10	MRR@5
EVA-CLIP-18B	No ReRank	84.0	95.3	96.9	89.7
	ImageScope	84.8	95.6	97.0	90.1
	CoTRR	85.6	96.2	97.3	90.3
	CoTMR	85.9	96.8	97.6	90.9
	EviRank-mini	84.9	96.5	97.5	91.0
	EviRank	86.7	96.9	98.0	91.1
	EviRank-plus	87.2	97.8	98.4	92.2
	EviRank-pro	88.0	97.7	98.7	92.5
CLIP-ViT-B/32	No ReRank	67.1	89.0	90.0	75.8
	AFS	72.4	91.3	91.8	80.1
	Explicit	72.5	93.7	93.9	80.9
	ImageScope	78.8	92.6	95.6	80.2
	CoTRR	84.6	94.7	96.0	80.9
	CoTMR	84.7	94.8	96.3	81.1
	EviRank-mini	85.2	94.0	96.4	82.7
	EviRank	86.2	95.0	96.7	86.8
	EviRank-plus	87.1	96.3	97.1	89.9
	EviRank-pro	87.2	97.2	97.9	90.2
CLIP-ViT-L/14	No ReRank	72.7	91.2	92.4	80.0
	AFS	78.4	94.3	95.2	84.6
	Explicit	74.6	93.3	94.2	82.0
	ImageScope	78.1	91.0	93.3	82.2
	CoTRR	83.8	93.2	94.2	82.8
	CoTMR	84.5	93.9	94.3	83.0
	EviRank-mini	83.6	92.4	93.4	82.3
	EviRank	85.0	94.3	95.4	84.9
	EviRank-plus	86.7	94.6	96.1	85.4
	EviRank-pro	88.7	95.5	96.9	87.1
BLIP-2	No ReRank	86.8	97.5	98.0	91.5
	AFS	89.0	98.2	98.3	93.1
	Explicit	88.6	98.4	98.5	92.8
	ImageScope	88.8	98.2	98.3	92.8
	CoTRR	88.9	98.3	98.4	93.0
	CoTMR	89.2	98.4	98.9	93.4
	EviRank-mini	89.3	98.6	98.9	93.3
	EviRank	91.3	98.6	99.0	93.6
	EviRank-plus	93.5	99.0	99.2	94.2
	EviRank-pro	95.6	99.2	99.4	95.6

Table 2: Retrieval performance (%) with different backbones on Flickr30k.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate **EviRank** across five benchmarks spanning three retrieval paradigms: **T→I** (MS COCO (Lin et al., 2014), Flickr30k (Plummer et al., 2017)), **I→I** (Stanford Online Products (SoP) (Oh Song et al., 2016), CUB-200-2011 (Wah et al., 2011)), and **(T,I)→I** (FashionIQ (Wu et al., 2021)). These collectively cover coarse-grained scene matching, fine-grained product/species discrimination, and cross-modal composed retrieval. Detailed splits are in Appendix C.

Implementation Details. For coarse retrieval (the standard retrieve-then-rerank prerequisite, not part of our contribution), we use EVA-CLIP-18B (Fang et al., 2023), CLIP-ViT-B/32 and L/14 (Radford et al., 2021), BLIP-2 (Li et al., 2023), and DINOv2 (Oquab et al., 2024) (for I→I), retriev-

ing Top-K=20 candidates. For the MLLM teacher, we use Gemini-3-flash (Google DeepMind, 2025a) and Gemini-3-pro (Google DeepMind, 2025b) with temperature T=0.15. The student is Qwen3-VL-2B-Thinking (Qwen Team, 2025), distilled with 20k queries (4K $\times$ 5 datasets). The distillation hyperparameters λ and η are selected via grid search over $[0, 1]$ with step 0.1 on the validation set. We evaluate four variants: **EviRank-mini** (rubric-only, no MLLM at test time), **EviRank** (distilled student), **EviRank-plus** (rubric + Gemini-3-flash), and **EviRank-pro** (rubric + Gemini-3-pro). Under the default K=20 and M=5, rubric scoring runs *locally* over all 20 candidates by reusing the first-stage encoder and costs no MLLM call, so the online budget of **EviRank-plus/-pro** is fixed at *two* calls per query: one evidence-extraction call (the P-CAP/P-B0/P-B1 modules of Appendix H are issued as a single structured request) and one top-M listwise call. All reported results and all efficiency numbers use this same two-call configuration. Full settings are in Appendix D.

Baselines. We compare against AFS (Khaertdinov et al., 2025a,b), ImageScope (Luo et al., 2025), CoTRR (Wu et al., 2025), CoTMR (Sun et al., 2025a), LoCoRE (Xiao et al., 2025), ReMatch (Liu et al., 2025b), VLM2Vec (Jiang et al., 2025), UniME-v2 (Gu et al., 2026), and RetLLM (Su and Wang, 2026).

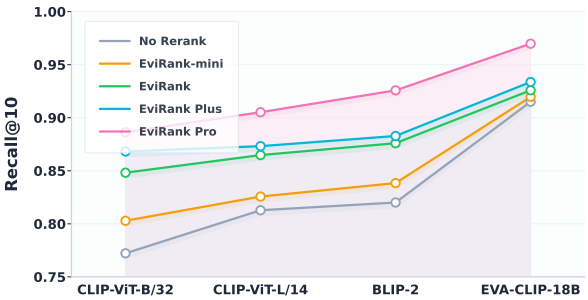

Figure 2: EviRank's Recall@10 on Flickr30k across the four coarse-retrieval backbones of Table 2.

4.2 Main Results

Tables 2–5, Table 13 (Appendix E), and Figure 2 demonstrate that **EviRank** consistently achieves SOTA across paradigms. On Flickr30k with BLIP-2, **EviRank-pro** attains 95.61% R@1, surpassing CoTMR by 6.32 points; even rubric-only **EviRank-mini** outperforms all prior methods without MLLM inference at test time. On COCO with CLIP-ViT-L/14, **EviRank-pro** reaches 69.53% R@1 (+11.22 over CoTRR). For I→I, **EviRank-pro** achieves

VLM	Methods	R@1	R@5	R@10
CLIP-ViT-B/32	No ReRank	39.4	65.5	75.6
	ImageScope	51.2	73.3	80.7
	CoTRR	58.3	77.2	83.5
	EviRank-mini	59.9	77.4	82.4
	EviRank	61.0	81.3	84.9
	EviRank-plus	62.3	83.7	86.6
	EviRank-pro	64.6	84.6	88.3
CLIP-ViT-L/14	No ReRank	46.4	71.1	79.7
	ImageScope	53.7	75.9	83.5
	CoTRR	59.9	79.7	84.7
	EviRank-mini	59.3	79.6	84.4
	EviRank	61.7	81.7	85.9
	EviRank-plus	64.9	84.7	85.9
	EviRank-pro	69.5	87.7	89.4

Table 3: Retrieval performance on COCO.

91.46% (SoP) and 86.89% (CUB-200) R@1, exceeding LoCoRE-base by 7.66 and 8.59 points. On FashionIQ, R@10 improves up to 8.27 points on Shirt over ImageScope. The gap between rubric-only **EviRank-mini** and MLLM-enhanced variants validates that structured evidence and listwise reasoning are *complementary* rather than redundant.

Comparison with recent embedding-based rerankers. Recent T→I progress comes largely from stronger *representations* rather than from reranking, so we also compare against four 2025–2026 systems of that kind (Table 13): **EviRank-pro** reaches 95.6 / 69.5 R@1 on Flickr30k / COCO versus 85.6 / 62.8 for the strongest of them (ReMatch). As these baselines already optimize the embedding space that our first stage consumes, the gap comes from evidence-conditioned verification, not from a better retriever.

	SoP				CUB-200			
Methods	R@1	R@5	R@10	mAP	R@1	R@2	R@4	mAP
No ReRank	80.8	86.8	92.1	65.1	68.9	79.4	87.3	49.8
SSR ReRank	81.2	87.4	91.9	66.7	69.4	79.0	86.1	54.2
α QE	81.1	85.8	90.7	68.1	70.9	78.8	84.7	56.9
RRT	81.9	86.7	92.4	67.2	68.7	85.0	95.6	55.6
LoCoRE-tiny	82.4	87.8	93.1	68.0	71.4	86.8	96.4	58.1
LoCoRE-small	83.3	88.4	92.7	69.4	74.6	89.1	97.3	61.0
LoCoRE-base	83.8	90.3	92.9	71.0	78.3	91.9	98.4	64.8
EviRank-mini	83.7	90.4	92.3	69.8	76.7	91.5	97.9	64.8
EviRank	86.6	90.9	93.3	73.8	82.4	93.9	98.5	66.8
EviRank-plus	90.0	92.3	95.0	77.0	84.6	95.5	99.2	72.5
EviRank-pro	91.5	93.9	96.0	79.6	86.9	97.6	99.4	76.6

Table 4: Retrieval performance (%) on SoP and CUB-200 (DINOv2). mAP refers to mAP@R.

Several older PRF/GRF/PALAVRA/SEARLE-OTI/iSEARLE/CIReVL/LinCIR baseline rows are moved to Appendix F, which reproduces the full tables.

VLM	Methods	Shirt		Dress		Toptee	
		R@10	R@50	R@10	R@50	R@10	R@50
CLIP-ViT-B/32	SEARLE	24.4	41.6	18.5	39.5	25.7	46.4
	iSEARLE-OTI	27.0	43.4	21.2	42.1	26.8	48.7
	CIReVL	28.3	47.8	25.2	46.3	31.2	53.8
	ImageScope	31.6	37.4	26.8	46.3	35.8	55.9
	EviRank-mini	31.7	38.4	27.9	47.4	37.4	56.5
	EviRank	33.8	40.5	32.0	49.7	42.9	59.0
	EviRank-plus	35.7	43.9	34.7	52.3	46.7	62.1
	EviRank-pro	39.7	46.7	36.8	56.7	49.0	64.8
CLIP-ViT-L/14	Pic2Word	26.2	43.6	20.0	40.2	27.9	47.4
	iSEARLE-OTI	31.8	50.2	24.1	45.1	31.7	53.2
	LDRE	31.0	51.2	22.9	46.7	31.5	53.6
	FTI4CIR	31.3	50.5	24.4	47.8	32.4	54.2
	ImageScope	32.8	51.0	26.1	46.1	35.0	55.1
	EviRank-mini	32.9	51.6	27.4	46.7	36.2	54.4
	EviRank	35.0	54.8	30.6	50.0	39.8	57.6
	EviRank-plus	37.6	57.6	34.6	53.3	42.6	59.8
	EviRank-pro	40.1	59.9	38.9	56.8	49.7	64.6

Table 5: Retrieval performance (%) on FashionIQ.

4.3 Ablation Studies

Variants	T→I		I→I	
	Flickr30K	COCO	SoP	CUB-200
EviRank-pro	96.9	89.4	87.6	99.4
w/o B0	94.5 ↓2.4%	88.6 ↓0.8%	82.3 ↓5.9%	96.5 ↓2.9%
w/o Slots	93.5 ↓3.4%	87.9 ↓1.6%	85.4 ↓2.4%	96.3 ↓3.1%
w/o Evidence	93.2 ↓3.8%	84.2 ↓5.7%	84.1 ↓3.9%	95.9 ↓3.5%
w/o Required	94.1 ↓2.8%	86.6 ↓3.0%	84.8 ↓3.0%	96.0 ↓3.4%
w/o Forbidden	94.6 ↓2.2%	87.4 ↓2.1%	85.6 ↓2.2%	97.4 ↓1.9%
w/o Ignore	96.4 ↓0.5%	88.8 ↓0.6%	87.3 ↓0.2%	99.4 ↓0.0%

Table 6: Ablation results on T→I and I→I tasks with CLIP-ViT-L/14.

Variants	(T,I)→I		
	Shirt	Dress	Toptee
EviRank-pro	40.1	38.9	49.7
w/o B0	37.6 ↓6.2%	35.4 ↓8.8%	46.5 ↓6.3%
w/o Slots	37.6 ↓6.0%	35.6 ↓8.2%	44.4 ↓10.5%
w/o Evidence	36.9 ↓7.8%	34.7 ↓10.5%	42.9 ↓13.5%
w/o Required	37.1 ↓7.3%	35.7 ↓8.0%	44.6 ↓10.0%
w/o Forbidden	37.9 ↓5.4%	37.3 ↓3.8%	46.8 ↓5.7%
w/o Ignore	39.2 ↓2.1%	38.4 ↓1.1%	48.6 ↓2.0%

Table 7: Ablation results on composed multimodal retrieval tasks with CLIP-ViT-L/14.

We ablate each component on all five benchmarks (CLIP-ViT-L/14 backbone, R@10 except CUB-200 R@4; Tables 6 and 7). Removing query understanding (w/o B0) most affects I→I (SoP −5.19, CUB −2.92), where caption generation bridges the modality gap. Removing slot decomposition (w/o Slots) degrades performance everywhere, especially on composed retrieval (Toptee −5.26), confirming that explicit slot separation yields more systematic evidence than holistic descriptions, and removing all evidence causes the largest drops (Toptee −6.71). Among constraint types, Required anchors positive matching (COCO −2.73), Forbidden disambiguates fine-grained candidates (CUB −1.96), and Ignore filters non-discriminative variations (Toptee −1.04).

Setting	COCO	SoP	FashionIQ
Rubric-only	59.9 / 82.4	83.7 / 69.8	32.1 / 50.9
Listwise-only	62.5 / 85.6	87.9 / 75.7	37.3 / 55.6
Rubric+Listwise	**64.6 / 88.4**	**91.5 / 79.8**	**42.9 / 60.4**

Table 8: Rubric vs. listwise component analysis. M1: R@1 (T→I/I→I) / R@10 (CIR). M2: R@10 / mAP / R@50 respectively.

4.4 Component Analysis: Rubric vs. Listwise

We further isolate **rubric scoring** versus **listwise comparison** (Table 8). Each alone is strong, but their combination is consistently best, with the largest gain on CIR (+5.6 R@10 on FashionIQ over listwise-only): the rubric handles explicit constraints, while listwise comparison captures implicit cues via joint visual comparison.

4.5 Stability of Structured Evidence Extraction

A core claim of EviRank is that *structuring* relevance estimation yields stable, auditable signals rather than relying on prompt wording. We validate it along two axes: stability of the evidence schema, and gain attribution.

Stability of the evidence schema. We test stability of the structured-slot output along three axes on FashionIQ (mean over 3 categories): (a) *repeated runs* (10× identical-input calls), (b) *prompt perturbation* (paraphrase, shuffle, typos), and (c) *cross-teacher* (Gemini-3-pro, GPT-5.4, Claude-4.6-Opus/Sonnet, Gemini-3-flash). Across all settings, Kendall's $\tau \geq 0.89$, Top-1 agreement $\geq 91\%$, and R@10 std $\leq$ 1.3. We attribute this stability to the fixed schema and ternary R/F/I typing, which collapse the output space into a narrow, JSON-checkable format across teachers. Stability is moreover largely decoupled from teacher strength: replacing Gemini-3-pro with the much cheaper Gemini-3-flash costs 4.6 R@10 (42.9→38.3) yet keeps τ=0.90 and 91.1% Top-1 agreement—a weaker teacher orders candidates worse without destabilizing *what* evidence is extracted. Detailed tables are in Appendix J.

Gains come from structured evidence, not prompt wording. To check that improvements come from the slot-wise R/F/I formulation rather than from prompt phrasing, we run three controlled

Setting	COCO	SoP	FIQ R@10
Query-only listwise	63.0	84.7	36.6
Free-form augmentation	63.5	86.7	38.3
Caption only	64.8	85.3	37.9
Evidence only	68.2	90.8	42.2
Caption + Evidence (Ours)	**69.5**	**91.5**	**42.9**

Table 9: Prompt-engineering controlled experiments (R@1, %).

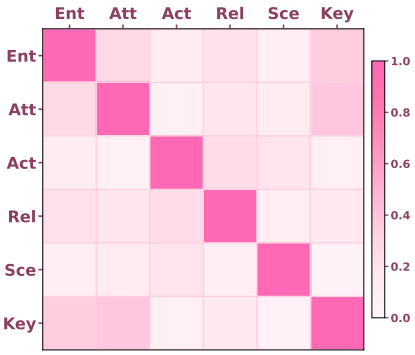

Figure 3: Pairwise SBERT similarity between the six evidence slots over 10k queries (off-diagonal mean 0.18).

experiments (Table 9; full results in Appendix K). Free-form augmentation gives limited gains (+0.5 R@1 on COCO) while structured evidence contributes most of the improvement (+6.5 R@1), and prompt-template rewrites cause negligible variation (R@1 std <0.7).

4.6 Efficiency of Evidence-Conditioned Re-ranking

For deployment, the distilled student inherits the evidence-grounded ranking logic of Section 3.4 without any teacher, evidence generator or runtime evidence cache. On Flickr30k it takes ∼800ms per query while achieving ∼10 R@1 points higher than CLIP coarse retrieval (∼382ms); CLIP is the shared first stage of *all* re-ranking pipelines, so its latency is a fixed cost rather than an alternative. The student is fully batchable: latency scales linearly with K and parallelizes across GPUs. With the teacher in the loop, the default setting costs only two online calls per query (Section 4.1); for larger listwise batches, divide-and-merge (Appendix A) keeps total calls linear in the number of groups and sequential depth logarithmic.

4.7 Case Study and Slot Semantics

Slots are near-orthogonal yet complementary. Figure 3 shows pairwise SBERT similarity between the six slots over 10k queries. The off-diagonal mean of 0.18 confirms they are not redundant restatements; the lowest pairs separate distinct axes (*Scene–KeyDetails*, *Actions–Attributes*), while moderate *Entities/Attributes–KeyDetails* correlations reflect a category-vs-instance hierarchy rather than overlap.

Qualitative case. A bicycle query from SoP illustrates the pipeline: *required* slots anchor the correct category, *forbidden* slots reject the most confusable ones ("mountain bike," "road racing bicycle"), and *ignorable* slots absorb context noise (full case study and R/F/I ablation in Appendices G–L).

5 Conclusion

EviRank shows that multimodal image re-ranking is, at its core, *language-grounded semantic verification*: turning a query into typed required / forbidden / ignorable constraints lets a re-ranker reason about *which* dimensions matter and *how* a candidate satisfies them, instead of committing the decision to an opaque similarity score or an unstructured chain-of-thought. Three findings recur across our five benchmarks: rubric scoring and listwise comparison are *complementary*; *forbidden* constraints, often overlooked in retrieval design, are critical for fine-grained discrimination; and structured evidence is robust across teachers, prompts and decoding randomness. The same formulation generalizes to NLP judgement tasks decomposable into typed atomic units, and distils into a compact student retaining over 90% of the teacher's capability.

Acknowledgements

This work was sponsored by the CCF-Tencent Rhino-Bird Open Research Fund (No. CCF-Tencent RAGR20250119), and was also supported by the Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515010304), the Guangdong Province Project (No. 2024QN11X088), and the Guangzhou Science and Technology Planning Project (No. 2025A03J4491).

Limitations

Our evaluation uses widely adopted public image retrieval benchmarks in English, which are standard for the community but do not capture every deployment condition that real-world systems encounter (e.g., multilingual queries, evolving content distributions, or domain-specific corpora outside common benchmark coverage). We also focus on still-image retrieval; extending the framework to

video, 3D or audio–visual content is an interesting direction. Finally, we report retrieval metrics on offline benchmarks; large-scale human-facing user studies in production search are left as a natural next step.

References

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, et al. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

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 *Proceedings of the International Conference on Learning Representations (ICLR)*.

Zhanpeng Chen, Chengjin Xu, Yiyan Qi, Jian Guo, and Xuhui Jiang. 2024. MLLM is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3096–3120.

Jonathan Cook, Tim Rocktaschel, Jakob Foerster, Dennis Aumiller, and Alex Wang. 2024. TICKing all the boxes: Generated checklists improve LLM evaluation and generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Enjun Du, Siyi Liu, and Yongqi Zhang. 2025. Mixture of length and pruning experts for knowledge graphs reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Enjun Du, Hange Zhou, Chenxu Du, Siyi Liu, Zirong Chen, Ziyu Zheng, and Yongqi Zhang. 2026. LEDGERMIND: Provenance-constrained multimodal agentic reasoning with a structured evidence ledger. *arXiv preprint arXiv:2607.28374*.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xianggang Wang, Tiejun Huang, XInlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19358–19369.

Andrea Frome, Greg S. Corrado, Jonathon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, volume 26, pages 2121–2129.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. ROSCOE: A suite of metrics for scoring step-by-step reasoning. In *The Eleventh International Conference on Learning Representations (ICLR)*.

Google DeepMind. 2025a. Gemini 3 flash: A fast and efficient multimodal language model. https://deepmind.google/technologies/gemini/. Accessed: 2026-02-04.

Google DeepMind. 2025b. Gemini 3 pro: A state-of-the-art multimodal large language model. https://deepmind.google/technologies/gemini/. Accessed: 2026-02-04.

Tiancheng Gu, Kaicheng Yang, Kaichen Zhang, Xiang An, Ziyong Feng, Yueyi Zhang, Weidong Cai, Jiankang Deng, and Lidong Bing. 2026. UniME-V2: MLLM-as-a-judge for universal multimodal embedding learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 21378–21386.

Chuong Huynh, Jinyu Yang, Ashish Tawari, Mubarak Shah, Son Tran, Raffay Hamid, Trishul Chilimbi, and Abhinav Shrivastava. 2025. Collm: A large language model for composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. 2025. VLM2Vec: Training vision-language models for massive multimodal embedding tasks. In *International Conference on Learning Representations*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137.

Bulat Khaertdinov, Mirela Popa, and Nava Tintarev. 2025a. A little more like this: Text-to-image retrieval with vision-language models using relevance feedback. *arXiv preprint arXiv:2511.17255*.

Bulat Khaertdinov, Mirela Popa, and Nava Tintarev. 2025b. A little more like this: Text-to-image retrieval with vision-language models using relevance feedback. *arXiv preprint arXiv:2511.17255*.

Yukyung Lee, Joonghoon Kim, Jaehee Kim, Hyowon Cho, and Pilsung Kang. 2025. CheckEval: Robust evaluation framework using large language model via checklist. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. 2024. Data roaming and quality assessment for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 19730–19742. PMLR.

Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yongkang Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. 2024. From matching to generation: A survey on generative information retrieval. *arXiv preprint arXiv:2404.14851*.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2025. MM-EMBED: Universal multimodal retrieval with multimodal llms. In *Proceedings of the International Conference on Learning Representations*. Poster.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C.Lawrence Zitnick, and Piotr Dollar. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*.

Guangyi Liu, Yongqi Zhang, Yong Li, and Quanming Yao. 2025a. Dual reasoning: A GNN–LLM collaborative framework for knowledge graph question answering. In *Conference on Parsimony and Learning*, volume 280 of *Proceedings of Machine Learning Research*, pages 351–372. PMLR.

Qianying Liu, Xiao Liang, Zhiqiang Zhang, Zhongfei Qing, Fengfan Zhou, Yibo Chen, Xu Tang, Yao Hu, and Paul Henderson. 2025b. ReMatch: Boosting representation through matching for multimodal retrieval. *arXiv preprint arXiv:2511.19278*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2511–2522.

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

Zheyuan Liu, Weixuan Sun, Damien Teney, and Stephen Gould. 2024. Candidate set re-ranking for composed image retrieval with dual multi-modal encoder.

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. In *Proceedings of the ACM on Web Conference 2025*, pages 1666–1682.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. 2016. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, and 7 others. 2024. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*.

Mengting Pan, Fan Li, Chen Chen, Xiaoyang Wang, and Wenjie Zhang. 2025. HiTeC: Hierarchical contrastive learning on text-attributed hypergraph with semantic-aware augmentation. *arXiv preprint arXiv:2508.03104*.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision (IJCV)*.

Qwen Team. 2025. Qwen3-vl-2b-thinking. `https://huggingface.co/Qwen/Qwen3-VL-2B-Thinking`. Alibaba Group.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR.

Dawei Su and Dongsheng Wang. 2026. RetLLM: Training and data-free MLLMs for multimodal information retrieval. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.

Zelong Sun, Dong Jing, and Zhiwu Lu. 2025a. CoTMR: Chain-of-thought multi-scale reasoning for training-free zero-shot composed image retrieval. *arXiv preprint arXiv:2502.20826*.

Zelong Sun, Dong Jing, Guoxing Yang, Nanyi Fei, and Zhiwu Lu. 2025b. Leveraging large vision-language model as user intent-aware encoder for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pages 7149–7157.

Yuanmin Tang, Jue Zhang, Xiaoting Qin, Jing Yu, Gaopeng Guo, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Wu. 2025a. Reason-before-retrieve: One-stage reflective chain-of-thoughts for training-free zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yuanmin Tang, Jue Zhang, Xiaoting Qin, Jing Yu, Gaopeng Guo, Gang Xiong, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Wu. 2025b. Reason-before-retrieve: One-stage reflective chain-of-thoughts for training-free zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14400–14410.

Yi Tay, Vinh Q. Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, Tal Schuster, William W. Cohen, and Donald Metzler. 2022. Transformer memory as a differentiable search index. In *Advances in Neural Information Processing Systems (NeurIPS)*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 809–819.

Likai Tian, Jian Zhao, Zechao Hu, Zhengwei Yang, Hao Li, Lei Jin, Zheng Wang, and Xuelong Li. 2025. Ccin: Compositional conflict identification and neutralization for composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. Caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Yujing Wang, Yingyan Hou, Haonan Wang, Ziming Miao, Shibin Wu, Hao Sun, Qi Chen, Yuqing Xia, Chengmin Chi, Guoshuai Zhao, Zheng Liu, Xing Xie, Hao Sun, Weiwei Deng, Qi Zhang, and Mao Yang. 2022. A neural corpus indexer for document retrieval. In *Advances in Neural Information Processing Systems (NeurIPS)*. Outstanding Paper Award.

Tianjun Wei, Wei Wen, Ruizhi Qiao, Xing Sun, and Jianghong Ma. 2025. RocketEval: Efficient automated LLM evaluation via grading checklist. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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

Shangrong Wu, Yanghong Zhou, Yang Chen, Feng Zhang, and P. Y. Mok. 2025. Chain-of-thought re-ranking for image retrieval tasks. *arXiv preprint arXiv:2509.14746*.

Zilin Xiao, Pavel Suma, Ayush Sachdeva, Hao-Jen Wang, Giorgos Kordopatis-Zilos, Giorgos Tolias, and Vicente Ordonez. 2025. Locore: Image re-ranking with long-context sequence modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Eric Xing, Pranavi Kolojyu, Robert Pless, Abby Stylianou, and Nathan Jacobs. 2025. Context-cir: Learning from concepts in text for composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sijia Xu, Fan Li, Xiaoyang Wang, Zhengyi Yang, and Xuemin Lin. 2026. C^2TC: A training-free framework for efficient tabular data condensation. *arXiv preprint arXiv:2602.21717*.

Yongqi Zhang, Haiquan Qiu, Shuzhi Liu, Enjun Du, and Quanming Yao. 2026. Neural knowledge graph reasoning with relational digraph. *Artificial Intelligence*, 355:104520.

Yongqi Zhang, Zhanke Zhou, Quanming Yao, Xiaowen Chu, and Bo Han. 2023. AdaProp: Learning adaptive propagation for graph neural network based knowledge graph reasoning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3446–3457. ACM.

Yang Zhou, Sunzhu Li, Shunyu Liu, et al. 2025. Breaking the exploration bottleneck: Rubric-scaffolded reinforcement learning for general LLM reasoning. *arXiv preprint arXiv:2508.16949*. Accepted to ICLR 2026.

A Listwise Batching and Rich Signal Package

This appendix complements Section 3.3 with the batching strategy used for evidence-grounded listwise refinement, and the full schema of the by-product supervision bundle that is later reused in distillation (Section 3.4).

Divide-and-merge batching. When the candidate pool is large ($K > 8$), a single listwise call becomes prompt-length-bound and unstable. We therefore partition the candidates into groups of size B (default B=5), rank each group independently conditioned on the same evidence package $\mathcal{E}(q)$, and iteratively merge the resulting sub-lists by pairwise listwise re-ranking. Writing $n = \lceil K/B \rceil$ for the number of groups, the procedure issues n group calls plus $n-1$ pairwise merge calls, i.e. $2n-1 = O(K/B)$ MLLM calls in total. Because all merges at the same level of the binary merge tree are independent, they can be issued concurrently, so the *sequential depth*—and hence wall-clock latency under a parallel client—is $1 + \lceil \log_2 n \rceil$ rounds. The logarithmic factor therefore describes latency, not total call count.

Rich signal package. A single listwise call produces, in addition to the permutation $\pi = (\pi_1, \ldots, \pi_K)$, the following structured artifact for each query–candidate list:

- Absolute relevance scores $\{r_i\}_{i=1}^{K}$ on a bounded scale (0–100), enabling calibrated comparisons across queries.
- A self-assessed confidence $p \in [0, 1]$ reflecting how decisively the candidates can be ordered under the current evidence.
- A set of *hard pairs* $\mathcal{H} = \{(i, j)\}$ consisting of the closest competitors, each accompanied by a short distinguishing reason that references specific constraints in $\mathcal{E}(q)$.
- Optional slot-wise signals: per-slot binary indicators $y_{s,i} = \mathbf{1}[\text{Match}_s(c_i) > \theta]$ and the corresponding rates from the deterministic rubric.

All of these signals share a fixed JSON schema, are produced by the same evidence-conditioned pipeline, and are consumed verbatim by the distillation stage—no additional supervision labelling is required.

B Additional Distillation Details

The three component objectives and the confidence-weighted total loss are defined in Section 3.4 (Eq. 4). This appendix records the motivation behind each term and the implementation details omitted there.

Why teacher scores are standardized. The teacher reports absolute relevance on a bounded 0–100 scale, but its effective dynamic range varies with query difficulty: an unambiguous query may yield $\{95, 90, 20, \ldots\}$ while a hard one yields $\{62, 60, 58, \ldots\}$. Per-query standardization by μ_r, σ_r removes this query-level scale, so $\mathcal{L}_{\text{score}}$ transfers the *shape* of the teacher's preference distribution rather than its absolute calibration. Student scores are not standardized—they are already produced by a trainable head—and both sides share the single temperature τ. The constant $\epsilon = 10^{-8}$ only guards against degenerate lists in which all teacher scores coincide.

Why hard pairs are weighted by the inverse score gap. Hard pairs are exactly the comparisons that a bi-encoder gets wrong, and their difficulty is not uniform: a pair separated by one point of teacher score is far more informative than one separated by thirty. The weight w_{ij} therefore grows as the gap $|r_i - r_j|$ shrinks, concentrating gradient mass on the closest competitors, while the margin δ prevents the loss from being driven to zero by an arbitrarily large score difference. The pair set $\mathcal{H}$ is taken verbatim from the teacher's listwise output; no additional mining is performed.

Slot auxiliary head. The slot head is a small linear projection on top of the student's pooled query–candidate representation, producing one logit per slot $s \in \mathcal{S}$, and is supervised only by the teacher's binary indicators $y_{s,i} = \mathbf{1}[\text{Match}_s(c_i) > \theta]$. It is used as a regularizer that forces the shared representation to encode constraint-level evidence rather than a single scalar preference; the head is dropped after training, so deployed inference cost is identical to a plain scoring student. When slot-wise labels are missing for a query, $\mathcal{L}_{\text{slot}}$ is skipped for that query rather than filled with pseudo-labels.

Hyperparameters and stability. We use τ=2.0, δ=0.5, λ=0.5 and η=0.3, selected on the validation split; the full configuration is listed in Table 11 in Appendix D. The confidence weight $p \in (0, 1)$ acts as a per-query learning-rate scaler, which we found sufficient to keep training stable without additional gradient clipping on low-confidence lists.

C Dataset Details

We evaluate on five benchmarks spanning three retrieval paradigms. Table 10 provides an overview, followed by detailed descriptions of each dataset.

Paradigm	Dataset	Images	Classes	Test Split
T→I	MS COCO	123,287	–	5K
	Flickr30k	31,000	–	1K
I→I	SoP	120,053	22,634	60,502
	CUB-200-2011	11,788	200	5,924
(T,I)→I	FashionIQ	77,684	3	6,016

Table 10: Summary of evaluation datasets.

C.1 Text-to-Image Retrieval (T→I)

MS COCO (Lin et al., 2014) is a large-scale image captioning and retrieval benchmark containing 123,287 images, each annotated with five human-written captions. The dataset covers a diverse range of everyday scenes with multiple objects and complex spatial relationships. We follow the standard Karpathy split (Karpathy and Fei-Fei, 2015) and report results on the 5K test set, where each text query is matched against all 5K images.

Flickr30k (Plummer et al., 2017) consists of 31,000 images collected from the Flickr photo-sharing platform, each accompanied by five descriptive captions. Compared to MS COCO, Flickr30k features more diverse photographic styles and natural language descriptions with richer linguistic variation. We evaluate on the commonly used 1K test split.

C.2 Image-to-Image Retrieval (I→I)

Stanford Online Products (SoP) (Oh Song et al., 2016) is an instance-level product retrieval benchmark containing 120,053 images across 22,634 fine-grained product categories sourced from eBay.com. The dataset is characterized by large intra-class appearance variations (e.g., different viewpoints, lighting conditions, and backgrounds for the same product) and high inter-class similarity among visually related products. The standard split uses 59,551 images from 11,318 categories for training and 60,502 images from 11,316 categories for testing.

CUB-200-2011 (Wah et al., 2011) is a fine-grained visual recognition dataset consisting of 11,788 bird images from 200 species. The dataset exhibits substantial intra-class variations in pose, viewpoint, and appearance, while inter-class differences can be extremely subtle (e.g., distinguishing between closely related warbler species). Following the standard evaluation protocol, we use the first 100 species (5,864 images) for training and the remaining 100 species (5,924 images) for testing.

C.3 Composed Image Retrieval ((T,I)→I)

FashionIQ (Wu et al., 2021) is a benchmark designed for image retrieval based on natural language feedback in the fashion domain. It contains three categories—*Dress*, *Shirt*, and *Toptee*—comprising 77,684 images and 30,134 image–text–image triplets. Each triplet consists of a reference image, a target image, and two natural language descriptions specifying how the target differs from the reference (e.g., "is more colorful and has shorter sleeves"). This setup requires models to jointly reason over visual content and textual modification intent. We report results on the standard validation split following prior work (Luo et al., 2025; Wu et al., 2025).

D Implementation Details

D.1 Stage A: Coarse Retrieval

We employ different backbone encoders depending on the retrieval paradigm. For text-to-image retrieval on Flickr30k and MS COCO, we use EVA-CLIP-18B (Fang et al., 2023), CLIP-ViT-B/32 (Radford et al., 2021), CLIP-ViT-L/14 (Radford et al., 2021), and BLIP-2 (Li et al., 2023). For image-to-image retrieval on SoP and CUB-200-2011, we use DINOv2 (Oquab et al., 2024) following the setup of LoCoRE (Xiao et al., 2025). For composed image retrieval on FashionIQ, we use CLIP-ViT-B/32 and CLIP-ViT-L/14. In all settings, we retrieve top-K=20 candidates per query via cosine similarity. The embedding-based matching operator m_{emb} in the rubric scoring module reuses the same backbone encoder as Stage A.

D.2 Stage B: Teacher Re-ranking

We use Gemini-3-flash (Google DeepMind, 2025a) (for **EviRank-plus**) and Gemini-3-pro (Google DeepMind, 2025b) (for **EviRank-pro**) as the multimodal teacher models. All structured prompting (caption generation, query expansion, evidence extraction, and listwise ranking) is performed with temperature T=0.15 to encourage deterministic outputs. For listwise re-ranking, candidates are processed in batches of 5; when $K > 8$, a divide-and-merge strategy is applied as described in Appendix A. The rubric scoring hyperparameters in

Eq. 3 are configured as follows: slot weights w_s are set uniformly to $1/|\mathcal{S}|$ (i.e., $1/6$), the forbidden penalty β=0.75, and the consistency weight γ=0.1.

D.3 Stage C: Knowledge Distillation

We construct the distillation dataset by uniformly sampling 20k queries across all five datasets (4K per dataset: Flickr30k, MS COCO, SoP, CUB-200-2011, and FashionIQ). For each sampled query, we execute the full Stage B pipeline using Gemini-3-pro as the teacher to produce the complete supervision package, including listwise rankings, absolute relevance scores (0–100), confidence estimates, hard-pair annotations, and per-slot hit/violation rates. Candidate ordering is randomly shuffled before teacher inference to eliminate position bias.

The student model is initialized from Qwen3-VL-2B-Thinking (Qwen Team, 2025) and fine-tuned for 10 epochs using AdamW with cosine learning rate scheduling. Table 11 summarizes the full hyperparameter configuration.

Hyperparameter	Value
Student backbone	Qwen3-VL-2B-Thinking
Training epochs	10
Batch size	32
Learning rate	5×10^{-5}
Optimizer	AdamW
LR schedule	Cosine decay
KL temperature τ	2.0
Hard-pair margin δ	0.5
Pair loss weight λ	0.5
Slot loss weight η	0.3
Distillation queries	20k (4K $\times$ 5 datasets)
Teacher model	Gemini-3-pro
Hardware	8 $\times$ NVIDIA H20 80GB

Table 11: Distillation hyperparameters for Stage C.

D.4 Model Variants

Table 12 summarizes the four **EviRank** variants and their corresponding configurations.

Variant	Rubric	Listwise	Inference Model
EviRank-mini	m_{emb}	✗	CLIP / DINOv2 only
EviRank	m_{emb}	✗	Qwen3-VL-2B (distilled)
EviRank-plus	m_{emb}	✓	Gemini-3-flash
EviRank-pro	m_{emb}	✓	Gemini-3-pro

Table 12: Summary of EviRank variants.

E Comparison with Recent T→I Baselines

Table 13 adds a head-to-head comparison with very recent embedding/CoT-based methods. **EviRank** substantially outperforms all of them on both benchmarks.

Method	Flickr30k	COCO
VLM2Vec	80.0	49.2
RetLLM	82.0	54.1
UniME-v2	85.5	60.9
ReMatch	85.6	62.8
EviRank-pro	**95.6**	**69.5**

Table 13: Comparison with recent SOTA T→I baselines (R@1, %).

F Full Main-Result Tables

For space, the main-text tables (Tables 2–5) omit several older CIR/CIR-style baselines (PRF, GRF, PALAVRA, SEARLE-OTI, iSEARLE, CIReVL, LinCIR, AQE, DQE, etc.). For completeness and reproducibility, we reproduce the full tables below (Tables 14, 15 and 16).

Backbone	Methods	R@1	R@5	R@10	MRR@5
EVA-CLIP-18B	No ReRank	84.08	95.33	96.97	89.75
	ImageScope	84.89	95.68	97.01	90.12
	CoTRR	85.69	96.21	97.34	90.37
	CoTMR	85.93	96.82	97.69	90.96
	EviRank-mini	84.89	96.47	97.52	90.97
	EviRank	86.65	96.93	98.01	91.11
	EviRank-plus	87.19	97.80	98.41	92.20
	EviRank-pro	87.98	97.65	98.73	92.51
CLIP-ViT-B/32	No ReRank	67.10	89.00	90.03	75.80
	PRF	66.90	89.20	90.23	75.70
	GRF	71.60	89.60	90.42	78.90
	AFS	72.40	91.30	91.89	80.10
	Explicit	72.50	93.70	93.96	80.90
	ImageScope	78.84	92.66	95.64	80.23
	CoTRR	84.68	94.72	96.08	80.99
	CoTMR	84.77	94.89	96.39	81.18
	EviRank-mini	85.19	93.97	96.42	82.67
	EviRank	86.18	95.03	96.74	86.78
	EviRank-plus	87.12	96.31	97.09	89.91
	EviRank-pro	87.24	97.17	97.93	90.19
CLIP-ViT-L/14	No ReRank	72.70	91.20	92.45	80.00
	PRF	72.60	91.70	92.63	80.00
	GRF	76.60	92.80	93.12	83.30
	AFS	78.40	94.30	95.20	84.60
	Explicit	74.60	93.30	94.26	82.00
	ImageScope	78.10	91.01	93.38	82.26
	CoTRR	83.81	93.23	94.27	82.89
	CoTMR	84.52	93.97	94.39	83.03
	EviRank-mini	83.62	92.41	93.37	82.29
	EviRank	84.97	94.33	95.39	84.93
	EviRank-plus	86.74	94.63	96.10	85.43
	EviRank-pro	88.69	95.48	96.91	87.06
BLIP-2	No ReRank	86.80	97.50	98.02	91.50
	PRF	86.90	97.60	98.13	91.70
	GRF	89.30	98.50	98.69	93.20
	AFS	89.00	98.20	98.30	93.10
	Explicit	88.60	98.40	98.56	92.80
	ImageScope	88.81	98.27	98.38	92.89
	CoTRR	88.93	98.31	98.42	93.08
	CoTMR	89.29	98.49	98.90	93.42
	EviRank-mini	89.33	98.57	98.89	93.27
	EviRank	91.34	98.62	99.01	93.57
	EviRank-plus	93.52	99.03	99.23	94.21
	EviRank-pro	95.61	99.22	99.41	95.62

Table 14: Full Flickr30k results across all backbones and baselines.

Methods	SoP				CUB-200			
	R@1	R@5	R@10	mAP	R@1	R@2	R@4	mAP
No ReRank	80.80	86.87	92.10	65.10	68.90	79.40	87.30	49.80
SSR ReRank	81.20	87.42	91.90	66.70	69.40	79.00	86.10	54.20
AQE	76.90	73.69	89.30	66.10	66.90	76.80	82.70	58.40
DQE	67.90	72.71	81.50	47.80	67.00	75.50	82.00	54.60
α QE	81.10	85.86	90.70	68.10	70.90	78.80	84.70	56.90
RRT	81.90	86.74	92.40	67.20	68.70	85.00	95.60	55.60
LoCoRE-tiny	82.40	87.80	93.10	68.00	71.40	86.80	96.40	58.10
LoCoRE-small	83.30	88.49	92.70	69.40	74.60	89.10	97.30	61.00
LoCoRE-base	83.80	90.37	92.90	71.00	78.30	91.90	98.40	64.80
EviRank-mini	83.69	90.41	92.34	69.83	76.72	91.49	97.90	64.78
EviRank	86.56	90.88	93.27	73.80	82.38	93.94	98.48	66.79
EviRank-plus	89.99	92.33	94.97	77.00	84.56	95.48	99.23	72.48
EviRank-pro	91.46	93.93	96.02	79.58	86.89	97.62	99.44	76.63

Table 15: Full SoP and CUB-200 results with all I→I baselines.

VLM	Methods	Shirt		Dress		Toptee	
		R@10	R@50	R@10	R@50	R@10	R@50
CLIP-ViT-B/32	PALAVRA	21.49	37.05	17.25	35.94	20.55	38.76
	SEARLE	24.44	41.61	18.54	39.51	25.70	46.46
	SEARLE-OTI	25.37	41.32	17.85	39.91	24.12	45.79
	iSEARLE	25.81	43.52	20.92	42.19	26.47	48.70
	iSEARLE-OTI	27.09	43.42	21.27	42.19	26.82	48.75
	CIReVL	28.36	47.84	25.29	46.36	31.21	53.85
	LDRE	27.38	46.27	19.97	41.84	27.07	48.78
	ImageScope	31.65	37.49	26.82	46.31	35.80	55.94
	EviRank-mini	31.73	38.42	27.93	47.39	37.43	56.52
	EviRank	33.79	40.45	31.96	49.73	42.89	58.98
	EviRank-plus	35.65	43.87	34.73	52.33	46.74	62.13
	EviRank-pro	39.67	46.66	36.80	56.65	48.98	64.79
CLIP-ViT-L/14	Pic2Word	26.20	43.60	20.00	40.20	27.90	47.40
	SEARLE	26.89	45.58	20.48	43.13	29.32	49.97
	SEARLE-OTI	30.37	47.49	21.57	44.47	30.90	51.76
	iSEARLE	28.75	47.84	22.51	46.36	31.31	52.68
	iSEARLE-OTI	31.80	50.20	24.19	45.12	31.72	53.29
	CIReVL	29.49	47.40	24.79	44.76	31.36	53.65
	LDRE	31.04	51.22	22.93	46.76	31.57	53.64
	LinCIR	29.10	46.81	20.92	42.44	28.81	50.18
	FTI4CIR	31.35	50.59	24.49	47.84	32.43	54.21
	ImageScope	32.87	51.07	26.17	46.15	35.03	55.12
	EviRank-mini	32.85	51.58	27.36	46.74	36.18	54.43
	EviRank	34.96	54.83	30.64	49.96	39.83	57.61
	EviRank-plus	37.63	57.62	34.56	53.31	42.57	59.84
	EviRank-pro	40.12	59.89	38.89	56.75	49.68	64.59

Table 16: Full FashionIQ results across all baselines and backbones.

G Complete Case Study: Evidence-Based Bicycle Retrieval

Figure 4: Case study's bicycle case.

This appendix provides comprehensive analysis of **EviRank**'s evidence extraction for the bicycle case study shown in Figure 4. Beyond presenting the extracted evidence, we analyze *why* each constraint emerges and *how* it contributes to retrieval accuracy.

G.1 Query Processing Pipeline

G.1.1 Stage 1: Caption Generation

Generated Caption:

> *This is a women's leisure bicycle featuring a low step-through curved frame design. The overall silhouette is sleek and streamlined, oriented toward daily commuting and light recreational use. The bicycle is primarily finished in a bright purple metallic color with glossy paint, complemented by black saddle, handlebars, and tires. Structural components include spoked wheel assemblies, upright handlebars, a comfort saddle, and basic drivetrain and braking systems. A prominent brand logo is printed on the frame side.*

Analysis: The initial caption establishes three critical semantic anchors: (1) *product category* (women's leisure bicycle), (2) *distinguishing structure* (low step-through frame), and (3) *primary visual attribute* (bright purple metallic). These anchors are essential because bicycle retrieval failures typically occur when the system confuses leisure bikes with mountain bikes or road bikes—categories that share similar components but differ fundamentally in frame geometry and intended use.

G.1.2 Stage 2: Query Expansion

Expanded Description:

> *A leisure/commuter women's bicycle with a classic low step-through curved frame configuration, smooth modern lines, and a style balancing sporty aesthetics with everyday practicality. Primary color is bright purple, accented with white brand lettering and minimal black components. Metal frame and fork with smooth glossy paint finish. Full-size spoked wheel assemblies, stable frame geometry from curved top and down tubes, widened comfort saddle, upright handlebars, and rear single-speed or simple derailleur system. Elongated proportions emphasizing riding stability. Brand logos on frame side, color-matched chain guard and fender areas, functional pedals, chain, braking system, and adjustable seat post—suitable for daily riding and light outdoor use.*

Analysis: The expansion serves two purposes. First, it adds *discriminative details* that were implicit in the image: "elongated proportions," "comfort saddle," and "upright handlebars" collectively define the leisure category's ergonomic signature. Second, it introduces *component-level vocabulary* (chain guard, fender, derailleur) that enables precise matching against product descriptions in the gallery. Without this expansion, the retrieval system would rely solely on global visual similarity, missing the fine-grained attributes that distinguish near-identical products.

G.2 Evidence Extraction with Dimensional Analysis

Global Summary:

> *The image displays a women's leisure/commuter bicycle predominantly finished in bright purple, featuring a low step-through curved frame with elongated and flowing lines. The configuration emphasizes stability and comfort for daily riding, suitable for urban commuting and light outdoor use.*

G.2.1 Entities: Category-Level Disambiguation

Required:

- *The image features a women's leisure/commuter bicycle as the main subject.*
- *This product is a modern-style city bike with a low step-through frame structure and upright riding orientation.*

Forbidden:

- *The image features a mountain bike as the main subject, characterized by robust suspension structure and off-road orientation.*
- *This product is a road racing bicycle, featuring a high top-tube frame and extremely narrow tire configuration.*

Ignore:

- *The image contains simple and clean presentation environment elements.*
- *The frame may contain minor photographic cropping variations.*

Why These Constraints Emerge: Entity-level evidence targets the most consequential retrieval errors—*category confusion*. In embedding space, a purple leisure bike may be closer to a purple mountain bike than to a green leisure bike, because color dominates the learned representation. The required constraints anchor the search to the correct product category, while forbidden constraints explicitly reject the nearest confusable categories.

Functional Role: During re-ranking, candidates satisfying entity requirements receive positive scores, while those matching forbidden descriptions receive strong penalties. This creates a "category gate" that filters out high-similarity but wrong-category results before finer attributes are evaluated.

G.2.2 Attributes: Visual Property Specification

Required:

- *The product's primary color is bright purple, complemented by white brand lettering and minimal black components.*
- *The frame and fork are constructed from metal material, presenting a smooth glossy paint texture.*

Forbidden:

- *The product's primary color is red or orange, complemented by large-area dark camouflage patterns.*
- *The frame material is carbon fiber or plastic, presenting a matte or rough surface texture.*

Ignore:

- *Minor shade variations exist within the same purple color family.*
- *Normal reflection differences exist on local metal surfaces.*

Why These Constraints Emerge: Attribute evidence addresses *intra-category discrimination*. Once the category is established, retrieval must distinguish among dozens of similar leisure bicycles. Color (bright purple vs. red/orange) and material finish (glossy metal vs. matte carbon) are the primary differentiators at this level.

Functional Role: The ignore constraints are particularly important here. Without them, lighting-induced color shifts or reflection variations would cause false negatives—rejecting correct matches due to superficial differences. By explicitly modeling acceptable variation, **EviRank** achieves robustness that rigid attribute matching cannot provide.

G.2.3 Actions: State and Presentation Mode

Required:

- *The bicycle is displayed completely from a side view angle, with the body maintaining an upright balanced state.*

- *The vehicle is in a fully assembled state ready for immediate riding, with all functional components installed.*

Forbidden:

- *The bicycle is displayed in a folded state, with the frame and wheels collapsed together.*
- *The vehicle is in a disassembled or maintenance state, with multiple key components placed separately.*

Ignore:

- *Minor forward or backward offset exists in the shooting angle.*
- *Small variations exist in wheel orientation during display.*

Why These Constraints Emerge: Product images of the same bicycle may appear in vastly different states—assembled, partially disassembled, or folded (for folding bikes). The action dimension ensures that presentation state does not confuse identity matching.

Functional Role: This dimension prevents a subtle failure mode: matching a fully assembled leisure bike to component images or maintenance photos of the same product line. The ignore constraints allow matching across minor viewpoint variations that are common in e-commerce photography.

G.2.4 Relations: Structural Geometry Verification

Required:

- *The front and rear wheels are correctly connected to the frame's front fork and rear dropout through spoked wheel assemblies.*
- *The handlebars, seat post, and frame form an upright comfortable geometric relationship, emphasizing stable riding posture.*

Forbidden:

- *The front and rear wheels exhibit obvious misalignment with the frame, with abnormal structural proportions.*
- *The handlebars and frame are connected through complex linkage structures, forming an unconventional layout.*

Ignore:

- *Minor spacing differences exist between components.*
- *Local structural alignment variations exist within normal tolerance ranges.*

Why These Constraints Emerge: Relational evidence captures the *geometric signature* of a product category. Leisure bicycles have a characteristic relationship between handlebars, seat, and frame that differs from racing bikes (aggressive forward lean) or mountain bikes (suspension-mediated geometry).

Functional Role: This dimension acts as a structural consistency check. Even if color and components match, incorrect geometric relationships indicate a different product. The forbidden constraints specifically target unconventional designs (recumbent bikes, cargo bikes) that might share individual attributes but differ in overall configuration.

G.2.5 Scene: Context Normalization

Required:

- *The product is displayed against a pure white or light-colored background.*
- *The overall scene presents as an e-commerce or product catalog style clean photography environment.*

Forbidden:

- *The product is displayed in a complex outdoor mountain environment.*
- *The overall scene presents as a cluttered indoor space containing numerous other objects.*

Ignore:

- *Minor brightness variations exist in the background.*
- *Subtle shadow differences exist in ambient lighting.*

Why These Constraints Emerge: Scene evidence explicitly models the *presentation context*, which is a major source of embedding noise. The same bicycle photographed in a studio versus on a mountain trail produces dramatically different embeddings, despite identical product identity.

Functional Role: By requiring studio-style presentation and ignoring minor background variations, this dimension normalizes context-induced similarity distortions. The forbidden constraints prevent matching to lifestyle photography where the product may be partially occluded or shown at unusual angles.

G.2.6 Key Details: Discriminative Feature Anchoring

Required:

- *Brand logos are clearly printed on the frame side as an important identifying feature.*
- *The bicycle is equipped with a color-matched chain guard and widened comfort saddle, reflecting daily riding orientation.*

Forbidden:

- *Exaggerated racing numbers and competitive sponsor logos are printed on the frame.*
- *The vehicle is equipped with dual-crown suspension forks and large-diameter disc brake systems as core features.*

Ignore:

- *Minor differences exist in brand font sizing.*
- *Slight positional variations exist in non-critical decals.*

Why These Constraints Emerge: Key details capture the *most discriminative local features*—elements that uniquely identify a specific product or product line. Brand logos and signature components (color-matched chain guard) serve as "visual fingerprints."

Functional Role: This dimension handles the final disambiguation when multiple candidates pass all other checks. The forbidden constraints are particularly valuable: racing numbers and professional-grade suspension are strong negative indicators that immediately disqualify competitive cycling products, even if they share color and basic frame geometry.

G.3 Integrated Analysis: How Evidence Dimensions Interact

The six dimensions form a hierarchical filtering system:

1. **Entities** establishes category boundaries, eliminating cross-category confusion (mountain bikes, road bikes).

2. **Attributes** performs intra-category discrimination based on visual properties (color, material, finish).

3. **Actions** filters presentation state mismatches (assembled vs. disassembled).

4. **Relations** verifies structural consistency with expected geometric configurations.

5. **Scene** normalizes context variations that affect embedding similarity.

6. **Key Details** anchors to unique identifying features for final disambiguation.

The three-way constraint structure (required/forbidden/ignore) operates consistently across all dimensions:

- **Required**: Positive evidence that must be present—defines what the target *is*.

- **Forbidden**: Negative evidence that must be absent—defines what the target *is not*.

- **Ignore**: Variation tolerance—defines what *does not matter* for identity.

This structured approach transforms retrieval from opaque similarity scoring into transparent logical evaluation, where each dimension's contribution can be independently assessed and explained.

H Prompt Templates

This appendix presents the complete prompt templates used across the **EviRank** pipeline. We employ four *logical prompt modules*: P-CAP for initial caption generation (Section H.1), P-B0 for query understanding and expansion (Section H.2), P-B1 for structured evidence slot mining (Section H.3), and P-RANK for listwise re-ranking inference (Section H.4). The four modules are documented separately for clarity, but at inference time they are merged into *two online calls*: P-CAP/P-B0/P-B1 are completed within a single structured evidence-extraction request, and P-RANK issues one listwise call over the rubric top-M candidates (Section 4.1). All prompts are invoked in a training-free manner with no task-specific fine-tuning. Placeholders enclosed in braces (e.g., `{query}`) are filled at runtime with actual query-specific content.

H.1 Prompt P-Cap: Caption Generation

Prompt P-CAP (Figure 5) serves as the entry point of the evidence extraction pipeline. It converts a raw product image into a rich, multi-aspect textual description covering five complementary dimensions: overall appearance, color and material, structural details, design features, and functional elements. This five-way decomposition is intentionally designed to mirror the evidence-slot taxonomy introduced in Section 3.1, ensuring that the generated caption already contains the fine-grained information needed for downstream evidence mining. The prompt is particularly critical for image-to-image retrieval where no user-provided text exists, as the caption becomes the sole textual anchor for all subsequent stages. By requesting 3–5 detailed sentences, we ensure sufficient specificity for discriminating among visually similar products while

PROMPT P-CAP Detailed Caption Generation

You are an expert at generating comprehensive and detailed visual descriptions for product images. Your task is to create a thorough description that captures both the overall appearance and fine-grained details of the product.

Query Image: <attached>

Generate a detailed product description that includes:
1. **Overall Appearance** – Product type/category; general shape and form factor; overall style (modern, vintage, classic, sporty, etc.).
2. **Color & Material** – Primary colors and any color patterns/combinations; material composition (metal, plastic, leather, fabric, wood, etc.); surface texture and finish (matte, glossy, brushed, etc.).
3. **Structural Details** – Key components and parts; construction and assembly details; size proportions and dimensions (relative terms).
4. **Design Features** – Unique design elements or patterns; branding, logos, or text if present; decorative elements or embellishments.
5. **Functional Elements** – Visible functional parts; hardware, fasteners, or connectors; any moving parts or adjustable features.

Please provide a comprehensive description in 3–5 detailed sentences that would allow someone to accurately identify and distinguish this specific product from similar items. Be as specific and descriptive as possible about visual characteristics.

Output: Only the detailed description text, nothing else.

Figure 5: Prompt P-CAP for detailed caption generation.

keeping the output concise enough for efficient processing.

H.2 Prompt P-B0: Query Understanding & Expansion

Prompt P-B0 (Figure 6) enriches the raw caption from the previous stage with background knowledge and semantic context (Section 3.2). The output contains three key fields. `expanded_query` augments the original description with concrete visual details that are implicit in the caption—such as typical use scenarios and related product concepts—thereby improving recall for semantically relevant but lexically dissimilar candidates. `core_intent` distils the single most discriminative identity statement, serving as the primary semantic anchor during evidence-slot extraction. `confidence` is a self-assessed reliability estimate of the expansion itself; it is used only to monitor and, if necessary, discard degenerate expansions, and should not be confused with the listwise confidence p that weights the distillation loss in Stage C (Eq. 4), which is produced by P-RANK. The prompt is conditioned on the caption alone: image paths, filenames and directory names are never exposed to the teacher at any stage, so no category prior can leak from the dataset layout into the evidence package.

H.3 Prompt P-B1: Structured Evidence Slot Mining

Prompt P-B1 (Figure 7) is the core evidence-mining prompt. It decomposes the enriched query into six orthogonal semantic dimensions—ENTITIES, ATTRIBUTES, ACTIONS, RELATIONS, SCENE, and KEYDETAILS—each further organized into *required*, *forbidden*, and *ignore* constraint lists. Two design choices are noteworthy. **(i) Affirmative-only wording.** All constraint sentences, including those in the forbidden list, are phrased as positive statements that *describe what a wrong match looks like* rather than negating the target. For instance, instead of "The bicycle should not be red," the forbidden constraint states "The bicycle is red with black handlebars." This ensures that every constraint can be directly compared with candidate descriptions via embedding cosine similarity (Eqs. 2–3), yielding stable and sign-consistent matching scores. **(ii) Explicit scoring guidance.** The prompt communicates the intended scoring logic (score $= \sum$ required_match $- \sum$ forbidden_match), aligning constraint generation with the downstream deterministic rubric (Eq. 3) and reducing misalignment between extraction and evaluation.

H.4 Prompt P-Rank: Listwise Re-ranking

During inference, the teacher MLLM receives both the query image and all top-K candidate images simultaneously, together with the structured evidence

PROMPT P-B0 Query Understanding & Expansion

You are a query understanding and expansion expert. Given a detailed image caption/description, you need to: (1) Understand the core visual characteristics and product identity; (2) Provide contextual background and semantic expansion; (3) Output in a structured JSON format.

Input Caption: query

Please analyze and output in the following JSON format:

```
{
"original_query":  "<original caption/query>",
"expanded_query":  "<expanded description with background context,
related concepts, and image-specific considerations>",
"core_intent":     "<one sentence describing the main visual identity
and distinguishing features>",
"confidence":      0.9
}
```

Requirements:
- `expanded_query` should add relevant visual and semantic context while preserving all original details.
- `core_intent` should capture the most distinctive visual characteristics.
- `confidence` should be a number between 0.0 and 1.0.
- Rely only on the caption; do not use filenames, directory names, or any other dataset-side metadata.
- Output ONLY the JSON object, no additional text before or after.
- Ensure all strings are properly escaped.

Figure 6: Prompt P-B0 for query understanding and expansion.

text distilled from the preceding stages. The evidence text presented to the re-ranker is assembled at runtime from the outputs of P-B0 and P-B1 following a compact serialization format (described after the prompt box). Figure 8 presents the full re-ranking prompt.

The re-ranking prompt is designed around three principles. **(i) Evidence-conditioned ranking.** The evidence constraints (`{evidence_text}`) are placed prominently before the candidate list, ensuring that the teacher attends to structured required/forbidden constraints as the primary decision basis rather than relying solely on holistic visual impression. This directly implements the "rubric-first stabilization" strategy. **(ii) Multimodal listwise input.** Both the query image and all K candidate images are presented in a single prompt turn, enabling the teacher to perform direct visual comparison across candidates rather than scoring each independently. The images are base64-encoded with controlled resolution (`max_edge`=2048) and JPEG quality (`quality`=85) to balance visual fidelity and token efficiency. **(iii) Structured JSON output.** By requiring a strict JSON schema with a permutation of candidate indices, we eliminate ambiguity in parsing and guarantee a valid ranking. The same call also returns the three supervision fields consumed by Stage C: `relevance_scores` provide the teacher scores $\mathbf{r}$ used in $\mathcal{L}_{\text{score}}$, `confidence` provides the weight p in Eq. 4, and `hard_pairs` provides $\mathcal{H}$ for $\mathcal{L}_{\text{pair}}$; only `ranked_indices` is needed to produce the final ranking at test time. The accompanying `brief_reasoning` field, while not used in the final score, provides interpretable rationales that are valuable for error analysis.

We anonymize candidate identifiers (replacing filenames with opaque indices) before presenting them to the teacher, so that no dataset-side metadata reaches the re-ranker. For large candidate sets ($K > 8$), we adopt a *divide-and-merge* batching strategy: candidates are partitioned into groups of size B (default B=5), each group is independently ranked by the teacher, and the resulting sub-lists are iteratively merged via pairwise re-ranking calls. With $n = \lceil K/B \rceil$ groups this costs $2n - 1$ calls in total and $1 + \lceil \log_2 n \rceil$ sequential rounds when merges at the same tree level are issued in parallel, which keeps inference cost tractable while preserving global ranking quality.

H.5 Evidence Text Assembly

The `{evidence_text}` placeholder in P-RANK is populated at runtime by serializing the outputs of P-B0 and P-B1 into a compact textual format. Figure 9 shows the assembly template.

As shown in Figure 9, the assembled evidence text follows a three-tier hierarchy: *query-level* summaries (core intent, expanded query, global summary), followed by *slot-level* constraint blocks. Within each slot, required and forbidden constraint sentences are serialized as pipe-delimited strings,

PROMPT P-B1 Structured Evidence Slot Mining

You are a visual evidence extraction expert for image-to-image retrieval. Given a detailed image caption, extract structured evidence that would help identify matching images.

Input Caption: query

Your task is to extract evidence into **6 semantic slots**, with each slot containing 2–4 complete **affirmative sentences** (not phrases). Organize into Required / Forbidden / Ignore lists:

• **Required**: Affirmative sentences describing visual elements that MUST be present for a correct match.
Example: "The bicycle is blue with silver handlebars."
• **Forbidden**: Affirmative sentences describing visual elements that would indicate a WRONG match. These describe what incorrect matches would look like.
Example: "The bicycle is red with black handlebars."
• **Ignore**: Affirmative sentences describing similar but non-critical visual elements that can be tolerated.

Scoring logic: score $= \sum$(required_match_scores) $- \sum$(forbidden_match_scores)

Output in the following JSON format:

```
{
"global_summary":  "<2-3 sentences providing an overall visual summary>",
"entities":    {"required":[...], "forbidden":[...], "ignore":[...]},
"attributes":  {"required":[...], "forbidden":[...], "ignore":[...]},
"actions":     {"required":[...], "forbidden":[...], "ignore":[...]},
"relations":   {"required":[...], "forbidden":[...], "ignore":[...]},
"scene":       {"required":[...], "forbidden":[...], "ignore":[...]},
"key_details": {"required":[...], "forbidden":[...], "ignore":[...]}
}
```

Requirements:
• **All** sentences must be affirmative (positive statements), including forbidden items.
• Forbidden sentences describe what wrong matches look like (positive form), NOT using "not" or "should not".
• Each entry should be a complete sentence; each list should have 2–4 sentences.
• Base the evidence strictly on the visual content described in the caption; do not rely on filenames, directories, or any external metadata.
• Sentences should be specific and descriptive for accurate embedding matching.
• Output ONLY valid JSON, no markdown code blocks, no explanations.

Figure 7: Prompt P-B1 for structured evidence slot mining.

with a maximum of six sentences per list to control prompt length. Notably, the *ignore* constraints are deliberately excluded from the re-ranking prompt. This is because their role is to calibrate the deterministic rubric scoring (Eq. 3) by masking non-discriminative variations, whereas the MLLM reranker is expected to internalize such tolerance through its own visual reasoning. Including ignore constraints would increase prompt length without providing actionable ranking guidance. This selective serialization keeps the evidence text compact (typically 200–400 tokens) while preserving all decision-critical information.

Together, the four prompts form a progressive refinement pipeline: P-CAP grounds the image in language, P-B0 enriches the description with semantic context, P-B1 structures the enriched representation into checkable constraints, and P-RANK leverages the full evidence package for listwise visual re-ranking. Each stage builds upon the previous one's output, creating a coherent chain from raw pixels to interpretable, evidence-conditioned ranking decisions.

I Slot Decomposition Example

The brief slot template is given as Table 17.

J Robustness Analysis Details

We complement the main-text robustness and efficiency summaries (Sections 4.5 and 4.6) with full tables on FashionIQ (mean over Shirt/Dress/Toptee).

(a) Repeated-run stability. We invoke the same teacher API 10× with identical inputs and perturb the student's random seed / decoding temperature.

Setting	R@10	Kendall's τ	Top-1 Agr.
Repeated calls	42.93±0.27	0.99	99.6%
Seed perturbation	32.03±1.26	0.93	98.2%
Decoding randomness	32.13±0.89	0.96	98.8%

(b) Prompt perturbation. We perturb the evidence-extraction prompt with three transformations.

PROMPT P-RANK Listwise Re-ranking with Evidence Constraints and Hard-Pair Mining

You are a visual reranker for image-to-image product retrieval. Given a query image, its detailed description, structured evidence constraints, and a list of candidate images, rank the candidate images from best to worst match based on visual similarity to the query image.

The goal is to find images of the SAME product or very similar products as the query image. Focus on:
• Exact or near-exact visual matches, including the same product from different angles.
• Products with identical key visual features, such as shape, color, branding, and distinctive details.
• Products from the same product line or category.

In addition to the final ranking, report (i) an **absolute relevance score** for every candidate, (ii) a **self-assessed confidence** in the ordering, and (iii) **hard-pair cases**: candidate pairs whose relevance scores are very close or visually difficult to distinguish. These scores, the confidence and the hard pairs are used as fine-grained distillation signals for training the student re-ranker.

IMPORTANT OUTPUT RULES:
• Output MUST be a single JSON object.
• JSON schema:

```
{
"ranked_indices": [int, int, ...],
"relevance_scores": [
{"index": int, "score": float}
],
"confidence": float,
"hard_pairs": [
{"pair": [int, int], "reason": "short reason"}
],
"brief_reasoning": "one short paragraph explaining the ranking"
}
```

• `"ranked_indices"` must be a permutation of $0 .. \{N-1\}$, ordered from BEST match to WORST match.
• `"relevance_scores"` must contain one entry per candidate index, with `score` on a 0–100 scale and consistent with `ranked_indices`.
• `"confidence"` must be a single number between 0.0 and 1.0 expressing how decisively the candidates can be ordered under the given evidence.
• `"hard_pairs"` should contain visually ambiguous candidate pairs with close relevance to the query.
• Do NOT output any extra text outside JSON.

```
=== QUERY IMAGE DESCRIPTION ===
{generated_caption}

=== EVIDENCE CONSTRAINTS ===
{evidence_text}

=== CANDIDATE LIST (index -> opaque ID) ===
{candidates_desc}
```

Below you will see:
1. The **QUERY IMAGE** (labeled) — this is what you are trying to match
2. The **CANDIDATE IMAGES** in order $(0, 1, 2, \ldots)$ — rank these by similarity to the query

`[QUERY IMAGE]` ⟨*base64-encoded query image*⟩
`[CANDIDATE IMAGES]` ⟨*base64-encoded candidate images, indexed 0..N-1*⟩

Figure 8: Prompt P-RANK for listwise re-ranking inference and hard-pair mining. The evidence text block is assembled from P-B0 and P-B1 outputs as described in Section H.5. Hard-pair cases are candidate pairs with close relevance scores or subtle visual differences, which are further used as fine-grained distillation signals for training the student re-ranker.

Evidence Text Assembly Template

```
Core intent: {b0.core_intent}
Expanded query: {b0.expanded_query}
Global summary: {b1.global_summary}
Slot constraints:
[entities]
required: {sent_1} | {sent_2} | ...
forbidden: {sent_1} | {sent_2} | ...
[attributes]
required: {sent_1} | {sent_2} | ...
forbidden: {sent_1} | {sent_2} | ...
[actions]
required: ...    forbidden: ...
[relations]
required: ...    forbidden: ...
[scene]
required: ...    forbidden: ...
[key_details]
required: ...    forbidden: ...
```

Figure 9: The runtime assembly format for evidence text injected into P-RANK. Constraint sentences within each slot are concatenated with pipe delimiters (|). Only *required* and *forbidden* constraints are included; *ignore* constraints are omitted from the re-ranking prompt as they serve primarily to calibrate the deterministic rubric score (Eq. 3) rather than guide the MLLM.

Variant	R@10	Kendall's τ	Top-1 Agr.
Paraphrase	42.6	0.97	98.8%
Order-Shuffled	42.4	0.95	99.3%
Minor Typos	41.9	0.95	98.8%

(c) Cross-teacher. We replace the teacher MLLM with five different models.

Teacher	R@10	Kendall's τ	Top-1 Agr.
Gemini-3-pro	42.9	1.00	100%
GPT-5.4	41.3	0.92	92.0%
Claude-4.6-Opus	42.7	0.91	94.2%
Claude-4.6-Sonnet	40.9	0.89	93.3%
Gemini-3-flash	38.3	0.90	91.1%

R/F/I label agreement across runs is $\geq$99% in all settings, confirming the structured-slot output format is highly stable.

K Prompt Sensitivity and Information-Source Ablation

These experiments verify that EviRank's gains stem from structured-evidence slots rather than prompt engineering.

Free-form augmentation vs. structured evidence.

Setting	COCO R@1	SoP R@1	FIQ R@10	FIQ R@50
Query-only listwise	63.02	84.73	36.57	54.43
Free-form augmentation	63.53	86.70	38.29	56.72
Structured evidence (ours)	69.53	91.46	42.90	60.41

Information-source ablation.

Setting	COCO R@1	SoP R@1	FIQ R@10
Original query only	63.02	84.73	36.57
Query + caption only	64.79	85.28	37.93
Query + evidence only	68.17	90.83	42.19
Query + caption + evidence	69.53	91.46	42.90

Prompt template sensitivity.

Template	COCO R@1	SoP R@1	FIQ R@10
A: Original prompt	69.53	91.46	42.90
B: Concise version	69.12	91.24	42.65
C: Rewritten version	69.50	91.38	42.83
D: No explicit penalty wording	68.86	90.72	42.12

The variation across rewrites is well below the gap from removing structured evidence, confirming that improvements come from the slot-wise R/F/I formulation rather than prompt wording.

L Slot Semantics Analysis

This appendix consolidates the semantic-independence analysis and the slot-type functional analysis previously placed in the main text of the original manuscript.

L.1 Semantic Independence Across Slots

We compute pairwise semantic similarity using Sentence-BERT embeddings across extracted constraints from 10k randomly sampled queries. Figure 10 presents the similarity matrix; low off-diagonal values (mean = 0.18) confirm semantic orthogonality. The lowest similarities occur between *Scene–KeyDetails* (0.05) and *Actions–Attributes*

Slot	Required	Forbidden	Ignore
Entities	The image must contain an acoustic guitar with a hollow wooden body and a visible sound hole. The guitar should have a traditional construction with a neck and fretboard clearly visible.	The guitar must not be an electric guitar with solid body and pickups. The instrument should not be a ukulele, bass guitar, banjo, or any other stringed instrument.	The specific brand or manufacturer of the guitar is not relevant for matching. The exact number of strings or whether a pick is being used does not affect the match.
Attributes	The dress must have blue as its primary base color. The fabric must display a floral or flower-based pattern with visible botanical motifs.	The dress should not be a solid single color without any pattern. The pattern must not be geometric stripes, polka dots, or abstract non-floral designs.	The exact shade or tone of blue does not need to match precisely. The specific type of flowers in the pattern and the overall dress length are not critical factors.
Actions	The woman must be actively engaged in playing the guitar with fingers positioned on strings. The posture should indicate ongoing musical performance rather than static positioning.	The woman should not be merely holding the guitar as a prop without playing engagement. The scene must not show her posing statically for a photograph with the guitar as an accessory.	The specific playing technique such as strumming versus fingerpicking is not relevant. The particular song or music being performed does not affect the matching.
Relations	The woman's hands must be positioned on the guitar in a playing configuration, with one hand on the neck and one near the body. The guitar must be held in an active playing position against her body.	The guitar should not be lying flat on a surface or hanging on a wall. The guitar must not be stored in a case or held passively without playing intent.	Whether the woman is sitting or standing while playing is not relevant. The specific hand positioning for left-handed versus right-handed playing does not affect the match.
Scene	The setting must be an outdoor café environment with visible tables, chairs, or typical café furniture. The lighting should indicate an outdoor daytime setting.	The location must not be an indoor concert stage or formal performance venue. The setting should not be a recording studio, private bedroom, or enclosed indoor space.	The number of other patrons or customers visible in the scene is not relevant. The specific café name, branding, or current weather conditions do not affect the match.
KeyDetails	The overall atmosphere must suggest a casual, informal musical performance in a relaxed social setting. The context should indicate spontaneous or low-key entertainment rather than organized performance.	The scene must not depict a formal concert with audience seating arrangements. The context should not suggest a music competition, audition, or professional street busking for monetary collection.	The specific time of day or season when the scene takes place is not relevant. The geographic location or cultural context of the café does not affect the matching.

Table 17: Slot Decomposition: structured evidence slots with required, forbidden, and ignored constraints.

(0.06), reflecting fundamentally different semantic axes: environmental context vs. product identity, and dynamic state vs. static appearance. Moderate correlations between *Entities–KeyDetails* (0.35) and *Attributes–KeyDetails* (0.38) reveal a semantic hierarchy rather than redundancy: *Entities* establishes category-level identity while *KeyDetails* anchors instance-level discrimination; *Attributes* captures global appearance while *KeyDetails* identifies unique local features. This validates our principle that slots should be *semantically orthogonal yet functionally complementary*.

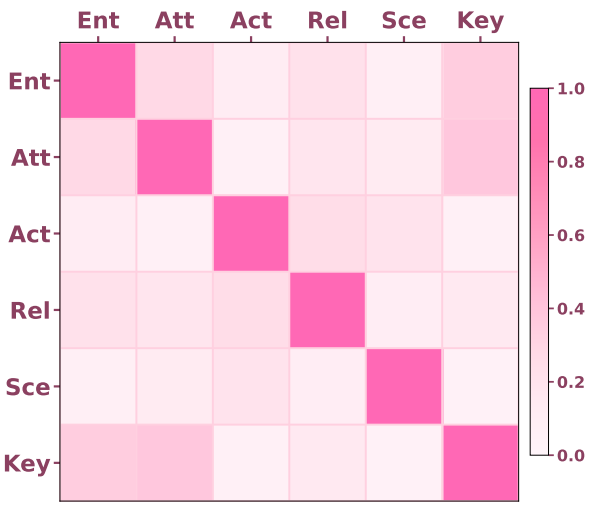

Figure 10: Semantic similarity matrix across evidence slots.

L.2 Functional Roles of Slot Types

To illustrate how different constraint types influence the re-ranking process, we analyze the query: *"A young woman in a blue floral dress playing acoustic guitar at an outdoor café."* The structured evidence slots for this query decompose as

A B C D E

Figure 11: Candidate images for slot-type ablation.

Candidate	Full	w/o R	w/o F	w/o I
(A) Blue floral dress, acoustic, café	1	4	1	1
(B) Blue solid dress, acoustic, café	3	2	1	3
(C) Blue floral dress, electric, stage	5	3	3	5
(D) Blue floral dress, acoustic, studio	4	5	4	2
(E) Green floral dress, acoustic, café	2	1	2	4

Table 18: Ranking shifts under evidence-constraint ablation.

given in Table 17 (Appendix I). Figure 11 and Table 18 show how candidate rankings shift when specific constraint types are ablated. Without required constraints, candidate E (green dress) rises to rank 1 since blue is no longer anchored. Without forbidden constraints, candidates B (solid pattern) and C (electric guitar) gain unjustified ground. Without ignore constraints, candidate D (indoor studio) is over-promoted since setting variations are no longer filtered, while candidate E drops to rank 4 because the green-vs-blue difference is now weighted more heavily without shade variations being explicitly ignorable. This three-way decomposition transforms re-ranking from holistic similarity scoring into explicit multi-dimensional logical evaluation, which additionally provides structured and decomposable supervision signals that facilitate knowledge distillation to smaller models.

M Correspondence to SQL Constraints

Reviewer feedback noted an interesting structural correspondence between our required/forbidden/ignore (R/F/I) labels and classical SQL constraint mechanisms. While SQL constraints are *schema-level* rules attached to a fixed relational model, ours are *query-time* semantic rubrics generated per query. Nonetheless, the mapping provides intuitive grounding:

EviRank	SQL analog	Role
Required	`NOT NULL` / mandatory predicate	must hold
Forbidden	negated `CHECK` constraint	must not hold
Ignore	`NULL`-able / unconstrained	no constraint

This mapping highlights that EviRank effectively performs constraint satisfaction over a query-conditioned rubric, analogous to schema-validated record matching but with the "schema" dynamically generated by the MLLM teacher.

N Detailed Comparison with ImageScope

ImageScope (Luo et al., 2025) is the most closely related multi-stage MLLM-based re-ranker, and it likewise unifies heterogeneous queries before invoking an MLLM. We therefore make the formulation-level differences explicit.

Different unification targets. ImageScope unifies queries into a *single natural-language description* of the user's intent and lets the MLLM perform holistic reasoning over that description. The unified representation is therefore still a piece of free-form text, and the relevance decision is recovered *implicitly* from the MLLM's reasoning trace. EviRank, in contrast, unifies queries into a *typed evidence package* $\mathcal{E}(q)$ consisting of slot-wise required/forbidden/ignorable constraints. The unified representation is itself an evaluation rubric, and relevance is recovered *explicitly* via constraint verification rather than holistic reasoning.

Different failure surfaces. Because ImageScope's representation is a holistic description followed by free-form reasoning, it inherits the three failure modes of CoT-based re-rankers discussed in Section 2 (omission of unstated constraints, hallucinated evidence, and inconsistent semantic coverage across queries). EviRank's typed-evidence representation directly addresses these failure modes: required and forbidden statements force the model to commit to which dimensions matter *before* comparing candidates; ignorable statements explicitly model what should not be penalized; and the slot schema is fixed across queries, making coverage uniform and auditable.

Auditability and supervision. Holistic descriptions are not decomposable: it is hard to ask "which part of the description is satisfied by candidate c?" or to extract per-dimension supervision signals from a single free-form passage. EviRank's evidence package is decomposable by construction, which (i) yields per-slot satisfaction and violation rates that can be inspected and verified independently, and (ii) provides structured supervision (slot-wise indicators, hard pairs, calibrated scores) that can be distilled into a smaller student—neither of which is naturally available from holistic-description pipelines.

Empirical gap. The two formulations are not just stylistically different. Across all five benchmarks (Tables 2–5), EviRank-pro consistently outperforms ImageScope by large margins, and even the rubric-only EviRank-mini—which uses no MLLM at test time—matches or exceeds ImageScope, indicating that the gain originates from the structured-evidence representation itself rather than from additional MLLM compute.