CROSSCHECK: FUSING GLOBAL RELEVANCE AND FINE-GRAINED SEMANTIC EVIDENCE FOR TRAINING-FREE COMPOSED IMAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Composed image retrieval requires overall target compatibility and precise adherence to a requested modification. However, global multimodal relevance can overlook specific attributes and relations, while fine-grained semantic evaluation may miss broader visual compatibility. We present CrossCheck, a training-free framework that combines these complementary sources of evidence. CrossCheck first estimates global relevance using CLIP-based retrieval and image-text matching. It then constructs semantic checks as target-contrast statement pairs and performs contrastive likelihood scoring: a frozen multimodal large language model compares the conditional token likelihoods of the two statements under the same visual context, yielding fine-grained evidence for individual modification and preservation requirements. We aggregate these signals with sensitivity to weakly supported checks and fuse them with global relevance for final ranking. CrossCheck requires no task-specific training and runs entirely locally with public weights. Across CIRCO, CIRR, and Fashion-IQ and three retrieval backbones, CrossCheck consistently improves over the compared methods. On a fixed CIRCO shortlist, global relevance and fine-grained semantic evidence achieve 29.56 and 30.43 mAP@5 individually, while their fusion reaches 41.60, showing that the gain arises from complementary ranking evidence.

1 INTRODUCTION

Composed image retrieval (CIR) searches an image gallery using a reference image and a natural-language modification. The retrieved image should reflect the requested change while retaining query-relevant reference content. Visual similarity alone is insufficient: a candidate can match the overall scene yet miss the detail that motivated the query. For example, given a photograph of a person underwater and the instruction "Add trees to the background," another underwater photograph is insufficient if it contains no trees. Conversely, an image of trees alone does not preserve the context of the request.

CLIP-based retrieval (Radford et al., 2021) provides a strong starting point for this task. Zero-shot CIR (ZS-CIR) methods map the composed query to a shared embedding space through pseudo-word representations, generated target descriptions, or imagined image proxies (Baldrati et al., 2023; Gu et al., 2024; Karthik et al., 2024; Li et al., 2025a). The resulting similarities efficiently identify globally relevant images, but a high similarity score does not establish that each requested attribute or relation is satisfied. This limitation is consistent with compositional evaluations showing that strong image–text retrieval performance can coexist with weaknesses in attribute binding and relational understanding (Yuksekgonul et al., 2023).

Large language models and multimodal large language models (MLLMs) offer a way to interpret these requirements explicitly. Existing methods use them to infer target descriptions, reason about desired content, or verify candidate compatibility (Karthik et al., 2024; Tang et al., 2025; Sun et al., 2025; Wang et al., 2026). However, richer query reasoning can still be followed by CLIP-based candidate scoring: CoTMR and G-MIXER extract object-level or inclusion/exclusion cues but evaluate them through CLIP similarities (Sun et al., 2025; Lim et al., 2026). This motivates complementing global CLIP-based matching with MLLM evaluation of individual semantic requirements.

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

We present **CrossCheck**, a training-free framework that combines CLIP-based global retrieval with MLLM-based semantic checking. A global stage retrieves candidates from three target views and uses image–text matching to form a compact shortlist, retaining its relevance score for the final ranking. An MLLM constructs query-specific semantic checks, each pairing a *target statement* with a matched *contrast statement* about the same attribute or relation. For each candidate, we contrast the statements' conditional token likelihoods under the same visual context. The resulting evidence is aggregated with sensitivity to weakly supported requirements and fused with global relevance. All models remain frozen, and checks are generated once per query without candidate access.

Experiments on CIRCO, CIRR, and Fashion-IQ with three retrieval backbones support this design. CrossCheck improves CIRCO test mAP@5 over WISER by 11.17–15.85% in relative terms. On a fixed CIRCO test shortlist, global relevance and semantic evidence achieve 29.56 and 30.43 mAP@5 individually, while their fusion reaches 41.60. Holding candidate identities fixed attributes this gain to improved ranking rather than increased coverage. Our contributions are threefold:

- We develop a training-free CIR framework that combines global relevance from CLIP-based retrieval and image–text matching with fine-grained semantic evidence from an MLLM.
- We formulate semantic checks as target–contrast statement pairs, scored through conditional likelihood contrasts and aggregated into candidate-level evidence.
- We evaluate CrossCheck on three CIR benchmarks and demonstrate the complementarity of the two evidence sources through fixed-shortlist ablations and per-query analyses.

2 Related Work

2.1 Zero-Shot Composed Image Retrieval

Zero-shot CIR avoids training on annotated reference–modification–target triplets, although it may still learn task-specific mappings from other data. SEARLE maps reference-image features to a pseudo-word that can be combined with the modification in CLIP's text encoder (Baldrati et al., 2023). LinCIR learns a projection through language-only self-supervision, improving scalability across retrieval backbones (Gu et al., 2024). These methods construct a joint query representation whose embedding similarity determines the retrieval order.

Training-free approaches instead compose queries using frozen generative models. CIReVL captions the reference and uses an LLM to rewrite the caption according to the requested change (Karthik et al., 2024). OSrCIR uses reflective chain-of-thought reasoning to produce a target description in one stage (Tang et al., 2025). Image-based composition provides another retrieval view: IP-CIR generates a target-image proxy and combines visual and textual retrieval signals (Li et al., 2025a). WISER combines text-to-image and image-to-image search with verification and iterative refinement (Wang et al., 2026). CrossCheck uses multiple target views to obtain globally relevant candidates and retains this evidence when incorporating semantic checks into the final ranking.

2.2 Reasoning and Reranking for Composed Image Retrieval

Several methods address the semantic detail that a global embedding can overlook. CoTMR reasons at image and object scales, predicts content that should be present or absent, and combines the CLIP similarities (Sun et al., 2025). G-MIXER expands candidate retrieval through geodesic mixup and reranks with CLIP-based inclusion and exclusion cues (Lim et al., 2026). STiTch refines composed captions in embedding space and uses bidirectional transportation to align caption and image distributions (Li et al., 2026). These methods strengthen query representation or cross-modal alignment. Their scoring differs from evaluating statement likelihoods conditioned on a candidate image.

MLLMs can also evaluate retrieved candidates directly. WISER uses the relative probabilities of binary answers to measure compatibility and guide fusion and refinement (Wang et al., 2026). CoTRR decomposes queries into semantic components, generates candidate assessments, and performs listwise chain-of-thought reranking (Wu et al., 2025). Building on candidate evaluation, CrossCheck evaluates each requirement by contrasting target and contrast statement likelihoods under the same visual context. Its final score combines the aggregated likelihood evidence with global relevance, allowing specific checks and broader visual matching to compensate for different ranking errors.

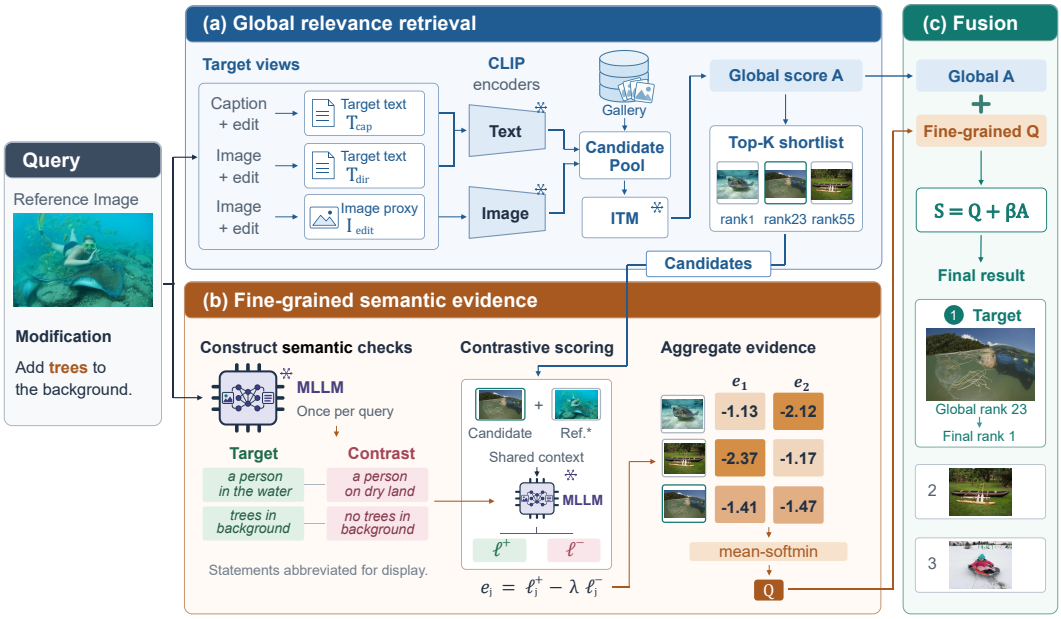


Figure 1: Overview of CrossCheck. Global relevance retrieval provides the shortlist and score $A(I)$. Fine-grained semantic evaluation yields $Q(I)$ through semantic checks. Evidence fusion combines the two signals into the final score $S(I)$.

3 Method

3.1 Framework Overview

CrossCheck combines global relevance retrieval, fine-grained semantic evaluation, and evidence fusion. As illustrated in Figure 1, global relevance retrieval produces a candidate shortlist and a score of target compatibility. Fine-grained semantic evaluation checks query requirements through contrastive likelihoods and aggregates their evidence with sensitivity to weakly supported requirements. Evidence fusion combines these two signals for the final ranking.

The two signals capture different aspects of candidate suitability: global relevance retains holistic matching information that a finite set of checks may omit, while fine-grained evidence distinguishes globally similar candidates by their support for specific requirements. Here, fine-grained refers to semantic specificity, including attributes, counts, relations, and preservation requirements.

Formally, given a reference image I_{ref} and a modification instruction T_{mod}, the task is to rank images in a gallery $\mathcal{D}$ according to how well they satisfy the requested changes while preserving query-relevant reference content. We use I for images and T for text.

All models are frozen. CLIP-style encoders E_{img} and E_{txt} provide unit-normalized retrieval embeddings in a shared space, an image–text matching (ITM) model F_{ITM} scores image–text pairs, and an autoregressive MLLM provides token probabilities P_{MLLM}. The generators G_{caption}, G_{desc}, G_{edit}, and G_{check} produce reference captions, target descriptions, edited images, and semantic checks, respectively. Specific models are given in Section 4.1.

3.2 Global Relevance Retrieval

This stage produces both a candidate shortlist $\mathcal{C}$ and a global relevance score $A(I)$. The score summarizes candidate compatibility with complete target descriptions and an edited-image proxy, and remains part of the final ranking in Section 3.4.

Candidate pooling. To construct the candidate pool, we form three target views from $(I_{\text{ref}}, T_{\text{mod}})$. The captioner first produces $T_{\text{ref}} = G_{\text{caption}}(I_{\text{ref}})$. The caption-based target description is $T_{\text{cap}} = G_{\text{desc}}(T_{\text{ref}}, T_{\text{mod}})$, while the direct target description is $T_{\text{dir}} = G_{\text{desc}}(I_{\text{ref}}, T_{\text{mod}})$. Each uses its

fixed generation prompt. The visual proxy is $I_{\text{edit}} = G_{\text{edit}}(I_{\text{ref}}, T_{\text{mod}})$. Let $\mathcal{V} = \{\text{cap}, \text{dir}, \text{vis}\}$ index these views. Their cosine similarities to a gallery image I are

$$\begin{aligned} s_b(I) &= \langle E_{\text{txt}}(T_b), E_{\text{img}}(I) \rangle , \qquad b \in \{\text{cap}, \text{dir}\}, \\ s_{\text{vis}}(I) &= \langle E_{\text{img}}(I_{\text{edit}}), E_{\text{img}}(I) \rangle . \end{aligned} \tag{1}$$

The top-L images from each view are merged into the candidate pool:

$$\mathcal{C}_v = \text{Top}_L(\mathcal{D}; s_v), \qquad \mathcal{U} = \bigcup_{v \in \mathcal{V}} \mathcal{C}_v. \tag{2}$$

Here $\text{Top}_L(\mathcal{S}; s)$ returns up to L images from $\mathcal{S}$ in descending score order, with ties resolved by a fixed protocol. We next define the evidence used to rank $\mathcal{U}$ before selecting the final shortlist.

Retrieval-rank evidence. To combine view-specific scores through their rankings, define

$$\mathcal{R}_N(I; s, \mathcal{S}) = \begin{cases} \dfrac{N - \text{rank}_{s,\mathcal{S}}(I)}{N}, & I \in \mathcal{S}, \\ 0, & I \notin \mathcal{S}, \end{cases}$$

where $\text{rank}_{s,\mathcal{S}}(I)$ is the zero-based position of I in a nonempty set $\mathcal{S}$ sorted by decreasing s, and $N \geq |\mathcal{S}|$ is the normalization size. Each view contributes

$$R_v(I) = \mathcal{R}_L\big(I; s_v, \mathcal{C}_v\big). \tag{3}$$

Thus, a view contributes positive retrieval-rank evidence only if the image is in its top-L set.

Image–text matching evidence. For each textual branch $b \in \{\text{cap}, \text{dir}\}$, the ITM model scores T_b against images in $\mathcal{U}_b = \mathcal{C}_b \cup \mathcal{C}_{\text{vis}}$, giving $s_b^{\text{ITM}}(I) = F_{\text{ITM}}(I, T_b)$. Let $\mathcal{W}(I) = \{b \in \{\text{cap}, \text{dir}\} : I \in \mathcal{U}_b\}$ be the branches for I. We normalize the branch rankings and average their contributions:

$$\begin{aligned} R_b^{\text{ITM}}(I) &= \mathcal{R}_{|\mathcal{U}_b|}\big(I; s_b^{\text{ITM}}, \mathcal{U}_b\big), \\ \overline{R}^{\text{ITM}}(I) &= \frac{1}{|\mathcal{W}(I)|} \sum_{b \in \mathcal{W}(I)} R_b^{\text{ITM}}(I). \end{aligned} \tag{4}$$

The two branch pools cover $\mathcal{U}$, so every candidate has at least one ITM score.

Global relevance scoring and shortlist selection. Let $o(I)$ be one when $I \in \mathcal{C}_{\text{vis}} \cap (\mathcal{C}_{\text{cap}} \cup \mathcal{C}_{\text{dir}})$, and zero otherwise. This indicator rewards agreement between the visual view and at least one textual view. The global relevance score is

$$A(I) = \max_{v \in \mathcal{V}} R_v(I) + \omega\, o(I) + \overline{R}^{\text{ITM}}(I), \qquad \omega \geq 0. \tag{5}$$

The maximum retains the strongest retrieval-rank signal, the overlap term rewards cross-modal agreement, and the ITM term contributes image–text matching evidence. Given reranking budget K, we select

$$\mathcal{C} = \text{Top}_K(\mathcal{U}; A). \tag{6}$$

The shortlist $\mathcal{C}$ is passed to the likelihood scorer, while $A(I)$ is retained for the final ranking. The budget K controls how many candidates undergo fine-grained evaluation. Appendix A.2 analyzes its coverage–cost trade-off.

3.3 Fine-Grained Semantic Evidence

For shortlisted candidates, this module evaluates semantic checks and summarizes their evidence in $Q(I)$. It comprises check construction, contrastive likelihood scoring, and evidence aggregation.

Modification T_{mod}:
Add trees to the background.



Reference image I_{ref}

Target statement T_j^+	Contrast statement T_j^-
The target image shows trees in the background.	The target image shows no trees in the background.
The target image shows a person in the water.	The target image shows a person on dry land.

Figure 2: Semantic-check illustration for a CIRR query.

Constructing semantic checks. A holistic target description provides a retrieval direction but does not expose individual requirements for candidate evaluation. The MLLM generator constructs a set of semantic checks from $(I_{\text{ref}}, T_{\text{mod}})$:

$$\mathcal{T} = G_{\text{check}}(I_{\text{ref}}, T_{\text{mod}}) = \{(T_j^+, T_j^-)\}_{j=1}^m, \qquad m \geq 1. \tag{7}$$

Each semantic check pairs a target statement T_j^+ describing the intended target state with a contrast statement T_j^- concerning the same subject or attribute. The intended contrast changes one attribute or relation while retaining its subject and comparable wording, probing a semantic requirement rather than a word-level fragment. The generator is instructed to cover requested modifications and query-relevant preservation checks, and the resulting pairs are generated once and reused throughout the shortlist. Concrete prompts are provided in Appendix B.4. Figure 2 shows a concrete example on CIRR.

Contrastive likelihood scoring. For semantic check j, let $\mathsf{X}_j(I)$ denote the scoring context containing candidate image I, a fixed scoring instruction T_{score}, and, when needed for reference-relative comparisons, I_{ref}. Image roles are explicitly labeled, and the context is identical for the two statements in each pair. For a prescribed text sequence T with $|T|$ tokens, the frozen MLLM assigns the length-normalized conditional log-likelihood

$$\ell_{\text{MLLM}}(T; \mathsf{X}) = \frac{1}{|T|} \sum_{u=1}^{|T|} \log P_{\text{MLLM}}(T_u \mid \mathsf{X}, T_{<u}) \,. \tag{8}$$

Here T_u is the token at position u and $T_{<u}$ is its prefix. Only prescribed statement tokens contribute to the average. Context tokens are excluded. Likelihoods are evaluated by teacher forcing rather than by sampling a free-form answer.

For semantic check j, contrastive evidence is

$$e_j(I) = \ell_{\text{MLLM}}\big(T_j^+; \mathsf{X}_j(I)\big) - \lambda\, \ell_{\text{MLLM}}\big(T_j^-; \mathsf{X}_j(I)\big), \qquad 0 \leq \lambda \leq 1. \tag{9}$$

This score combines support for the target state with its advantage over the contrast state:

$$\begin{aligned} e_j(I) = (1-\lambda)\, &\ell_{\text{MLLM}}\big(T_j^+; \mathsf{X}_j(I)\big) \\ &+ \lambda \Big[\ell_{\text{MLLM}}\big(T_j^+; \mathsf{X}_j(I)\big) - \ell_{\text{MLLM}}\big(T_j^-; \mathsf{X}_j(I)\big)\Big]. \end{aligned} \tag{10}$$

Thus, $\lambda = 0$ uses support for the target statement alone, while $\lambda = 1$ uses the full likelihood contrast.

Aggregating fine-grained evidence. A pure mean allows strong evidence on some requirements to compensate for weak evidence on another. A hard minimum avoids this compensation but depends entirely on one potentially noisy check. We interpolate between the mean and the softmin over the same m semantic checks:

$$\begin{aligned} M(I) &= \frac{1}{m} \sum_{j=1}^m e_j(I), \\ B(I) &= -\tau \log \left[\frac{1}{m} \sum_{j=1}^m \exp\left(-\frac{e_j(I)}{\tau} \right) \right], \qquad \tau > 0, \\ Q(I) &= (1-\rho) M(I) + \rho B(I), \qquad 0 \leq \rho \leq 1. \end{aligned} \tag{11}$$

Modification and preservation checks enter this single aggregate. The mean retains overall support, while the softmin emphasizes weakly supported requirements. Smaller τ brings the softmin closer to the minimum check evidence, while ρ controls its contribution relative to the mean. This balances overall support with sensitivity to weak evidence without imposing a hard logical veto. Appendix C.2 gives bounds and limiting properties of the aggregate.

3.4 Evidence Fusion

The fine-grained semantic score $Q(I)$ summarizes support for explicit requirements, while the global relevance score $A(I)$ retains holistic target compatibility. Fine-grained checks can distinguish candidates with similar overall relevance. Global relevance can supplement incomplete or noisy checks with broader matching information. We combine the two signals as

$$S(I) = Q(I) + \beta A(I), \qquad \beta \geq 0, \tag{12}$$

and rank candidates in $\mathcal{C}$ by decreasing $S(I)$. The weight β controls the contribution of global relevance to the fused ranking.

4 Experiments

4.1 Experimental Setup

Datasets and metrics. We evaluate CrossCheck on three composed image retrieval benchmarks. **CIRCO** (Baldrati et al., 2023) contains everyday scenes from COCO and provides multiple target images for a query. **CIRR** (Liu et al., 2021) pairs natural images with descriptions of the changes needed to identify a target, and includes groups of visually similar images for fine-grained evaluation. **Fashion-IQ** (Wu et al., 2021) focuses on clothing retrieval, with natural-language descriptions of differences between garments in three categories: Dress, Shirt, and Tops&Tees.

For CIRCO, we report mean average precision (mAP) at $k \in \{5, 10, 25, 50\}$ on the test set. For CIRR, we report full-gallery Recall@k at $k \in \{1, 5, 10, 50\}$ and subset $\text{Recall}_{\text{sub}}$@$k$ at $k \in \{1, 2, 3\}$ on the test set. The latter evaluates ranking within the predefined groups of similar images. For Fashion-IQ, we report Recall@10 and Recall@50 on the validation set for each category, together with category-averaged results.

Implementation details. We instantiate the retrieval encoders E_{img} and E_{txt} defined in Section 3.1 with OpenCLIP (Cherti et al., 2023), using ViT-B/32, ViT-L/14, and ViT-bigG/14. We denote the last by ViT-G/14 throughout the paper. Following CIReVL (Karthik et al., 2024), we use BLIP-2 (Li et al., 2023) for reference-image captioning (G_{caption}). Its image–text matching variant is also used for F_{ITM}. Following WISER (Wang et al., 2026), we adopt BAGEL (Deng et al., 2025) for target-description generation (G_{desc}) and image editing (G_{edit}). Qwen3.5-9B (Qwen Team, 2026) constructs semantic checks (G_{check}) and provides token probabilities (P_{MLLM}) for contrastive likelihood scoring. Hardware and detailed hyperparameter settings are provided in Appendix B.1.

Baselines. We distinguish methods by additional training and inference-time model access. *Training-dependent methods*, including LinCIR (Gu et al., 2024), MOA (Li et al., 2025c), and HIT (Li et al., 2025b), learn additional representations or mapping modules without annotated CIR triplet supervision. *Training-free methods* use pretrained models without parameter updates for CIR, including CIReVL (Karthik et al., 2024), LDRE (Yang et al., 2024), OSrCIR (Tang et al., 2025), AutoCIR (Cheng et al., 2025), CoTMR (Sun et al., 2025), G-MIXER (Lim et al., 2026), STiTch (Li et al., 2026), and WISER (Wang et al., 2026). Their inference dependencies differ: the reported CIReVL configuration uses GPT-3.5-turbo, OSrCIR and G-MIXER use GPT-4o, and WISER combines a Qwen2.5-VL-7B-Instruct verifier with a GPT-4o refiner, while CoTMR and STiTch use publicly available MLLM weights. We report training requirements and support for fully local inference separately for the compared configurations.

4.2 Zero-Shot CIR Benchmark Comparisons

We compare CrossCheck with state-of-the-art training-free and training-dependent ZS-CIR methods on CIRCO, CIRR, and Fashion-IQ. CIRR subset evaluation are detailed in Appendix B.3.

Table 1: Comparison with state-of-the-art methods on the test sets of CIRCO and CIRR. "TF" denotes training-free. "FL" denotes fully local: all inference stages of the compared configuration can be run locally using publicly available model weights, without external model APIs. Best results within each backbone are in **bold**, and second-best results are underlined.

Backbone	Method	TF	FL	CIRCO				CIRR						
				mAP@k				Recall@k				Recall$_{\text{sub}}$@k		
				$k=5$	$k=10$	$k=25$	$k=50$	$k=1$	$k=5$	$k=10$	$k=50$	$k=1$	$k=2$	$k=3$
ViT-B/32	CIReVL (ICLR'24)	✓	×	14.94	15.42	17.00	17.82	23.94	52.51	66.00	86.95	60.17	80.05	90.19
	LDRE (SIGIR'24)	✓	×	17.96	18.32	20.21	21.11	25.69	55.13	69.04	89.90	60.53	80.65	90.70
	OSrCIR (CVPR'25)	✓	×	18.04	19.17	20.94	21.85	25.42	54.54	68.19	–	62.31	80.86	91.13
	AutoCIR (KDD'25)	✓	×	18.82	19.41	21.38	22.32	30.53	59.42	72.19	91.47	65.11	84.02	92.70
	CoTMR (ICCV'25)	✓	✓	22.23	22.78	24.68	25.74	31.50	60.80	73.04	91.06	66.61	84.50	92.55
	G-MIXER (CVPR'26)	✓	×	21.33	22.00	24.00	24.98	35.18	66.19	76.82	–	69.98	85.78	92.92
	STiTch (CVPR'26)	✓	✓	20.26	21.01	23.01	24.04	25.83	55.25	70.20	–	65.64	83.60	92.80
	WISER (CVPR'26)	✓	×	32.23	33.18	34.82	35.35	49.45	76.55	85.21	93.81	77.30	88.63	92.27
	CrossCheck (Ours)	✓	✓	**37.34**	**37.89**	**39.92**	**40.61**	**50.53**	**77.11**	**85.59**	**94.87**	**78.72**	**89.74**	**94.55**
ViT-L/14	CIReVL (ICLR'24)	✓	×	18.57	19.01	20.89	21.80	24.55	52.31	64.92	86.34	59.54	79.88	89.69
	LinCIR (CVPR'24)	×	✓	12.59	13.58	15.00	15.85	25.04	53.25	66.68	–	57.11	77.37	88.89
	LDRE (SIGIR'24)	✓	×	23.35	24.03	26.44	27.50	26.53	55.57	67.54	88.50	60.43	80.31	89.90
	OSrCIR (CVPR'25)	✓	×	23.87	25.33	27.84	28.97	29.45	57.68	69.86	–	62.12	81.92	91.10
	MOA (SIGIR'25)	×	✓	15.30	17.10	18.50	19.30	27.10	56.50	69.20	90.00	–	–	–
	AutoCIR (KDD'25)	✓	×	24.05	25.14	27.35	28.36	31.81	61.95	73.86	92.07	67.21	84.89	93.13
	CoTMR (ICCV'25)	✓	✓	27.61	28.22	30.61	31.70	35.02	64.75	76.18	92.51	69.39	85.75	93.33
	HIT (ICCV'25)	×	✓	15.50	16.70	18.90	19.90	27.90	57.60	70.50	90.40	–	–	–
	G-MIXER (CVPR'26)	✓	×	28.29	29.04	31.44	32.39	37.42	67.69	78.58	–	71.88	87.04	92.82
	STiTch (CVPR'26)	✓	✓	25.55	26.27	28.81	29.99	28.87	57.97	69.90	–	65.22	84.10	92.37
	WISER (CVPR'26)	✓	×	35.10	36.30	38.46	39.15	49.23	76.72	85.11	94.17	77.81	88.89	92.77
	CrossCheck (Ours)	✓	✓	**39.02**	**40.00**	**42.53**	**43.28**	**50.53**	**77.78**	**85.68**	**95.10**	**79.20**	**90.41**	**94.67**
ViT-G/14	CIReVL (ICLR'24)	✓	×	26.77	27.59	29.96	31.03	34.65	64.29	75.06	91.66	67.95	84.87	93.21
	LinCIR (CVPR'24)	×	✓	19.71	21.01	23.13	24.18	35.25	64.72	76.05	–	63.35	82.22	91.98
	LDRE (SIGIR'24)	✓	×	31.12	32.24	34.95	36.03	36.15	66.39	77.25	93.95	68.82	85.66	93.76
	OSrCIR (CVPR'25)	✓	×	30.47	31.14	35.03	36.59	37.26	67.25	77.33	–	69.22	85.28	93.55
	CoTMR (ICCV'25)	✓	✓	32.23	32.72	35.60	36.83	36.36	67.52	77.82	93.99	71.19	86.34	93.87
	G-MIXER (CVPR'26)	✓	×	31.79	32.54	35.49	36.87	39.18	69.83	79.35	–	72.36	87.25	93.85
	STiTch (CVPR'26)	✓	✓	34.40	35.56	38.07	40.02	39.23	69.95	79.56	–	73.56	89.50	**95.86**
	WISER (CVPR'26)	✓	×	36.53	38.14	40.46	41.26	49.54	77.40	85.76	94.17	78.10	89.06	92.68
	CrossCheck (Ours)	✓	✓	**41.60**	**42.35**	**44.89**	**45.74**	**50.94**	**78.12**	**85.90**	**95.18**	**79.30**	**90.43**	95.08

Consistent gains across retrieval backbones. Among the compared methods, CrossCheck achieves the highest reported CIRCO mAP and CIRR full-gallery Recall at every evaluated cutoff within each backbone group. Relative to WISER, its CIRCO mAP@5 gains range from 11.17% to 15.85% across ViT-B/32, ViT-L/14, and ViT-G/14. With ViT-G/14, mAP@5 increases from 36.53 to 41.60. The improvement also holds for single-target full-gallery retrieval: CIRR Recall@1 increases by 1.08–1.40 percentage points over WISER across the three backbone groups. These results support the effectiveness of the complete framework across the evaluated retrieval backbone choices.

Strong early retrieval across domains. The gains extend to clothing retrieval, where CrossCheck obtains the highest category-averaged Recall@10 among the compared methods in all three backbone groups. With ViT-G/14, it reaches 45.60, compared with 44.59 for WISER and 45.11 for the training-dependent LinCIR. CrossCheck also leads CIRR Recall$_{\text{sub}}$@1 and Recall$_{\text{sub}}$@2 within every backbone group. Together, these results show competitive performance at early retrieval cutoffs in both natural-image and fashion benchmarks, using a training-free pipeline that can run entirely with publicly available model weights.

4.3 Ablation and Further Analysis

We isolate scoring effects on fixed shortlists using the CIRCO and CIRR test sets, then analyze complementarity on their validation sets. Additional experiments, including analyses of candidate coverage and scorer sensitivity, are provided in Appendix A.

Core component ablation. Table 3 shows that global relevance and fine-grained semantic evidence are complementary for full-gallery retrieval. On CIRCO, A and Q individually achieve

Table 2: Comparison with state-of-the-art methods on the validation set of Fashion-IQ. R@k abbreviates Recall@k. "TF" and "FL" follow Table 1. Best results within each backbone are in **bold**, and second-best results are underlined.

Backbone	Method	TF	FL	Shirt		Dress		Tops&Tees		Avg.	
				R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
ViT-B/32	CIReVL (ICLR'24)	✓	×	28.36	47.84	25.29	46.36	31.21	53.85	28.29	49.35
	LDRE (SIGIR'24)	✓	×	27.38	46.27	19.97	41.84	27.07	48.78	24.81	45.63
	OSrCIR (CVPR'25)	✓	×	31.16	51.13	29.35	50.37	36.51	58.71	32.34	53.40
	AutoCIR (KDD'25)	✓	×	32.43	51.67	26.52	46.36	33.96	56.09	30.97	51.37
	CoTMR (ICCV'25)	✓	✓	33.42	53.93	31.09	54.54	38.40	61.14	34.30	56.54
	G-MIXER (CVPR'26)	✓	×	37.24	55.99	36.39	**58.21**	45.23	64.10	39.62	59.43
	STiTch (CVPR'26)	✓	✓	25.22	44.16	18.59	40.16	25.97	47.61	23.26	43.98
	WISER (CVPR'26)	✓	×	41.51	56.92	37.48	56.97	**46.97**	62.32	41.99	58.74
	CrossCheck (Ours)	✓	✓	**43.27**	**59.51**	**38.47**	58.10	45.69	**65.32**	**42.48**	**60.98**
ViT-L/14	CIReVL (ICLR'24)	✓	×	29.49	47.40	24.79	44.76	31.36	53.65	28.55	48.57
	LinCIR (CVPR'24)	×	✓	29.10	46.81	20.92	42.44	28.81	50.18	26.28	46.49
	LDRE (SIGIR'24)	✓	×	31.04	51.22	22.93	46.76	31.57	53.64	28.51	50.54
	OSrCIR (CVPR'25)	✓	×	33.17	52.03	29.70	51.81	36.92	59.27	33.26	54.37
	MOA (SIGIR'25)	×	✓	31.90	50.70	25.20	48.50	33.20	54.80	30.10	51.30
	AutoCIR (KDD'25)	✓	×	34.00	53.43	24.94	45.81	33.10	55.58	30.68	51.60
	CoTMR (ICCV'25)	✓	✓	35.43	54.91	31.18	55.04	38.55	61.33	35.05	57.09
	HIT (ICCV'25)	×	✓	32.40	51.20	25.60	47.10	32.80	54.70	30.30	51.00
	G-MIXER (CVPR'26)	✓	×	40.87	60.35	37.98	**60.93**	**46.91**	**66.14**	41.92	**62.47**
	WISER (CVPR'26)	✓	×	43.13	57.85	38.42	55.92	45.39	61.50	42.17	58.51
	CrossCheck (Ours)	✓	✓	**44.65**	**61.92**	**38.57**	57.61	45.49	64.10	**42.90**	61.21
ViT-G/14	CIReVL (ICLR'24)	✓	×	33.71	51.42	27.07	49.53	35.80	56.14	32.19	52.36
	LinCIR (CVPR'24)	×	✓	46.76	**65.11**	38.08	**60.88**	**50.48**	**71.09**	45.11	**65.69**
	LDRE (SIGIR'24)	✓	×	35.94	58.58	26.11	51.12	35.42	56.67	32.49	55.46
	OSrCIR (CVPR'25)	✓	×	38.65	54.71	33.02	54.78	41.04	61.83	37.57	57.11
	AutoCIR (KDD'25)	✓	×	36.36	55.84	26.18	47.69	37.28	60.38	33.27	54.63
	CoTMR (ICCV'25)	✓	✓	38.32	62.24	34.51	57.36	41.90	64.30	38.25	61.32
	G-MIXER (CVPR'26)	✓	×	39.65	59.61	34.71	58.85	44.77	67.47	39.71	61.98
	STiTch (CVPR'26)	✓	✓	39.48	56.59	35.04	56.74	42.86	64.95	39.12	59.43
	WISER (CVPR'26)	✓	×	45.98	61.58	38.57	58.35	49.21	66.96	44.59	62.30
	CrossCheck (Ours)	✓	✓	**48.13**	64.18	**39.76**	59.10	48.90	67.42	**45.60**	63.57

mAP@5 of 29.56 and 30.43, while their fusion reaches 41.60. On CIRR, fusion improves Recall@1 from 47.49 for Q to 50.94. Candidate identities are identical across variants within each dataset, so these comparisons isolate changes in ranking from changes in candidate coverage. Within Q, adding contrast statements improves CIRCO mAP@5 and CIRR Recall@1, with smaller additional gains from mean–softmin aggregation.

The benefit depends on the evaluation setting. Within CIRR's predefined groups of similar images, Q alone achieves the highest $\text{Recall}_{\text{sub}}$@1/2/3 among the ablated configurations. For example, $\text{Recall}_{\text{sub}}$@1 is 79.93 for Q and 79.30 for the fusion. This pattern is consistent with global matching information benefiting full-gallery retrieval while occasionally perturbing the fine-grained ordering within a restricted set of similar candidates.

Complementary behavior. We analyze CIRCO and CIRR queries with ViT-G/14, using the same candidates for A, Q, and S. For each query q, let n_q be its number of target images and $y_j^X \in \{0,1\}$ indicate relevance at rank j under ranking X. We use $M_5(q;X) = \frac{1}{\min(5,n_q)} \sum_{j=1}^{5} y_j^X \frac{\sum_{\ell=1}^{j} y_\ell^X}{j}$. This is AP@5 on CIRCO and reduces to truncated reciprocal rank (RR@5) on CIRR: $1/r$ if the single target ranks $r \leq 5$, and zero otherwise.

In Figure 3 (left), a query is improved, tied, or degraded according to the difference between $M_5(q;S)$ and its reference value. Best(A,Q) is a label-based per-query selector using $\max\{M_5(q;A), M_5(q;Q)\}$. Fusion exceeds this selector on 30.0% of CIRCO and 9.2% of CIRR

Table 3: Core component ablation on the CIRCO and CIRR test sets. Candidate shortlists are fixed within each dataset. The best result in each metric column is shown in bold.

Component configuration		CIRCO test				CIRR test						
		mAP@k				Recall@k				Recall$_{\text{sub}}$@k		
		5	10	25	50	1	5	10	50	1	2	3
Global relevance only: A		29.56	30.87	33.53	34.60	31.90	60.10	71.35	92.51	61.74	80.55	91.28
Q only	target statements (mean)	29.29	30.92	33.81	34.93	46.77	73.93	81.61	94.31	79.21	90.43	95.33
	+ contrast statements	30.36	31.97	34.93	35.99	47.45	74.10	81.66	94.24	79.81	90.82	95.57
	+ mean–softmin	30.43	32.09	35.03	36.20	47.49	74.15	81.66	94.24	**79.93**	**90.92**	**95.59**
Evidence fusion: $S = Q + \beta A$		**41.60**	**42.35**	**44.89**	**45.74**	**50.94**	**78.12**	**85.90**	**95.18**	79.30	90.43	95.08

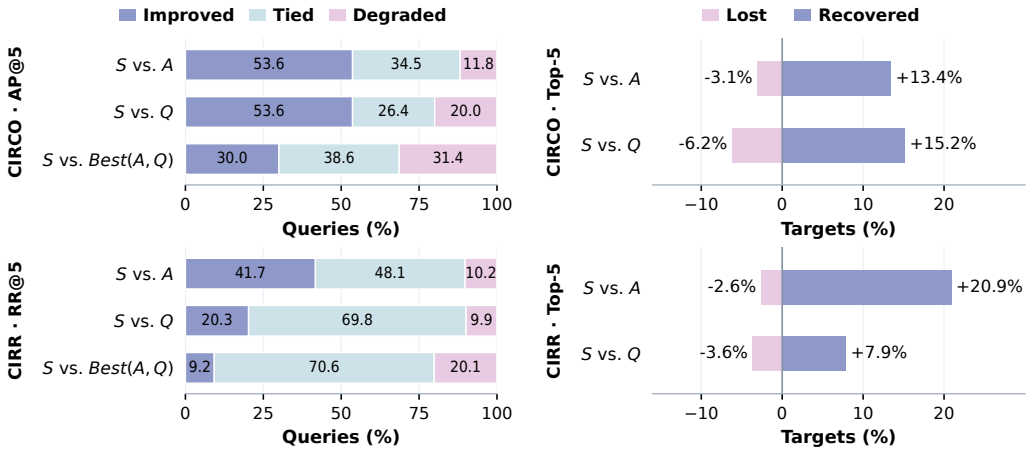


Figure 3: Complementarity diagnostics on CIRCO and CIRR with fixed shortlists. Left: Percentages of queries improved, tied, or degraded by S relative to each reference, measured by AP@5 on CIRCO and RR@5 on CIRR. Right: Target images recovered into (positive) or lost from (negative) the top five after fusion. "Targets (%)" expresses each count as a percentage of all target query–image pairs.

queries, but its mean score remains lower than ideal on both datasets. Relative to either individual signal, improvements are more frequent than degradations on both datasets, supporting the complementary value of the two signals.

The right panel counts target query–image pairs entering (recovered) or leaving (lost) the top five, normalized by all annotated target pairs. Across comparisons, 7.9–20.9% of pairs are recovered and 2.6–6.2% are lost, yielding a net gain in top-five target coverage. Together, these results support the complementarity of global relevance and fine-grained semantic evidence: their fusion helps correct errors in the individual rankings and improves retrieval performance. Exact counts and additional target-recovery statistics are provided in Appendix A.4.

5 Conclusion

We introduced CrossCheck, a training-free framework that couples global relevance with fine-grained semantic evidence for composed image retrieval. Contrastive likelihood scoring makes modification and preservation requirements explicit in candidate scoring, while evidence fusion retains the broader matching information that individual checks may miss. Experiments on CIRCO, CIRR, and Fashion-IQ demonstrate the effectiveness of this combination across three retrieval backbones. Fixed-shortlist analyses show that fusion corrects errors in the individual rankings even when candidate coverage is held fixed. These findings highlight the value of evaluating specific semantic requirements in conjunction with overall target compatibility.

AI USE STATEMENT

Generative AI tools were used to support several aspects of the research and manuscript preparation process, including brainstorming and research ideation, information and literature search, literature summarization and analysis, identification of potential research gaps and relevant prior work, assistance with the creation and modification of software code, and improvements to the clarity, grammar, and readability of the manuscript. All AI-assisted outputs, including literature-related suggestions, code, and textual revisions, were critically reviewed and independently verified by the authors where appropriate. The authors take full responsibility for the methodology, results, claims, and final content of this work.

REPRODUCIBILITY STATEMENT

To support reproducibility, Section 3 defines scoring and fusion procedure, while Section 4.1 reports the datasets, evaluation metrics, model components, and retrieval backbones. Appendix A documents candidate construction, reranking budgets, scorer sensitivity, and the fixed-shortlist analysis protocol. Appendix B provides the hyperparameter settings, evidence calibration procedure, and generation and likelihood-scoring prompts.

REFERENCES

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15338–15347, 2023.

Zhangtao Cheng, Yuhao Ma, Jian Lang, Kunpeng Zhang, Ting Zhong, Yong Wang, and Fan Zhou. Generative thinking, corrective action: User-friendly composed image retrieval via automatic multi-agent collaboration. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 334–344, 2025. doi: 10.1145/3711896.3736982. URL https://doi.org/10.1145/3711896.3736982.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. URL https://arxiv.org/abs/2505.14683.

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

Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. In *International Conference on Learning Representations*, volume 2024, pp. 16926–16941, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/48fd58527b29c5c0ef2cae43065636e6-Paper-Conference.pdf.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 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*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Miaoge Li, Dongsheng Wang, Zening Sun, Jinsen Zhang, Wenhan Luo, and Jingcai Guo. STiTch: Semantic transition and transportation in collaboration for training-free zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12374–12384, 2026.

You Li, Fan Ma, and Yi Yang. Imagine and seek: Improving composed image retrieval with an imagined proxy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3984–3993, 2025a.

Zhe Li, Lei Zhang, Zheren Fu, Kun Zhang, and Zhendong Mao. Hierarchy-aware pseudo word learning with text adaptation for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24319–24329, 2025b.

Zhe Li, Lei Zhang, Kun Zhang, Weidong Chen, Yongdong Zhang, and Zhendong Mao. Rethinking pseudo word learning in zero-shot composed image retrieval: From an object-aware perspective. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 833–843, 2025c. doi: 10.1145/3726302.3730074. URL https://doi.org/10.1145/3726302.3730074.

Jiyoung Lim, Heejae Yang, and Jee-Hyong Lee. G-MIXER: Geodesic mixup-based implicit semantic expansion and explicit semantic re-ranking for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33891–33900, 2026.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2125–2134, 2021.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

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. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Zelong Sun, Dong Jing, and Zhiwu Lu. CoTMR: Chain-of-thought multi-scale reasoning for training-free zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22675–22684, 2025.

Yuanmin Tang, Jue Zhang, Xiaoting Qin, Jing Yu, Gaopeng Gou, Gang Xiong, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Wu. 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*, pp. 14400–14410, 2025.

Tianyue Wang, Leigang Qu, Tianyu Yang, Xiangzhao Hao, Yifan Xu, Haiyun Guo, and Jinqiao Wang. Wiser: Wider search, deeper thinking, and adaptive fusion for training-free zero-shot composed image retrieval. *arXiv preprint arXiv:2602.23029*, 2026.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion IQ: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11307–11317, 2021.

Shangrong Wu, Yanghong Zhou, Yang Chen, Feng Zhang, and P. Y. Mok. Chain-of-thought reranking for image retrieval tasks. *arXiv preprint arXiv:2509.14746*, 2025. URL https://arxiv.org/abs/2509.14746.

Zhenyu Yang, Dizhan Xue, Shengsheng Qian, Weiming Dong, and Changsheng Xu. LDRE: LLM-based divergent reasoning and ensemble for zero-shot composed image retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 80–90, 2024. doi: 10.1145/3626772.3657740. URL https://doi.org/10.1145/3626772.3657740.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KRLUvxh8uaX.

A Additional Experimental Analyses

A.1 Candidate Sources and Coverage

Table 4 compares candidate sources on the CIRCO and CIRR validation sets with the same final shortlist size K, using $L = 100$ per retrieval view. This analysis measures relevant-image coverage before semantic reranking.

Table 4: Global relevance retrieval on the CIRCO and CIRR validation sets with ViT-G/14 and $L = 100$ per view. Recall measures the query-averaged fraction of target images retained. Hit measures the fraction of queries with at least one target image retrieved. On CIRR, Recall equals Hit because each query has one designated target. All values are percentages. The best result in each column is shown in bold.

Candidate source	CIRCO val						CIRR val		
	Recall@K			Hit@K			Recall@K		
	25	50	100	25	50	100	25	50	100
Caption-based text view	54.25	64.69	77.25	80.45	88.18	92.27	80.58	88.73	93.90
Direct text view	47.38	56.01	65.08	68.18	75.91	80.91	82.47	88.64	93.18
Edited-image view	38.83	49.46	61.82	63.18	74.55	85.91	74.98	82.47	88.66
Caption-based text + edited image	57.60	67.86	76.55	81.82	89.09	94.09	83.95	91.39	95.41
Direct text + edited image	53.48	64.70	71.88	80.45	86.36	90.45	86.82	92.85	95.98
All three views, without ITM	59.28	70.35	78.35	84.55	90.00	**95.00**	86.85	93.52	96.56
All three views, with ITM	**65.42**	**73.30**	**81.65**	**89.09**	**91.82**	94.55	**86.99**	**93.64**	**97.18**

At $K = 25$, combining all three views without ITM improves CIRCO Recall from 54.25 for the strongest single view to 59.28, and CIRR Recall from 82.47 to 86.85. Thus, the additional views improve relevant-image coverage under a fixed reranking budget. Adding ITM to the same three-view union further raises Recall to 65.42 on CIRCO and 86.99 on CIRR at $K = 25$. Its contribution is substantially larger on CIRCO in this comparison.

A.2 Reranking Budget: Coverage and Computational Cost

We choose K to balance target coverage against the cost of fine-grained evaluation. An overly large shortlist weakens the filtering role of global retrieval and shifts more candidate discrimination to the expensive MLLM scorer. An overly small shortlist discards relevant images before semantic evaluation: reranking only reorders the admitted candidates and cannot introduce images excluded from the shortlist. For a query with m semantic checks, evaluating both statements in each pair requires $2mK$ statement-likelihood evaluations. Increasing K therefore expands both the set of recoverable targets and the semantic-scoring workload.

Validation protocol. We vary K over prefixes of the same ordering by global relevance in Eq. 6, using all 220 CIRCO and 4,181 CIRR validation queries with ViT-G/14. The retrieval views, their union, and the global relevance scores remain fixed, with $L = 100$ and $\omega = 0.25$ as in Table 8. Each prefix contains exactly K distinct non-reference images. Recall and Hit follow the definitions in Table 4. For CIRCO, we additionally report the fraction of queries whose designated target appears in the shortlist. For CIRR, the single annotated target makes these three coverage measures identical. Group membership is not used to construct the shortlists.

Coverage gained by increasing K. The increase from $K = 50$ to $K = 90$ raises Recall by 7.43 percentage points on CIRCO and 3.08 points on CIRR (Table 4 and Figures 4 and 5). These gains illustrate the relevant images that a smaller shortlist would exclude from subsequent reranking. At $K = 90$, CIRCO Recall is 80.73%, and at least one relevant image is retained for 208 of 220 queries (94.55%). Increasing K to 100 raises Recall to 81.65%, a gain of 0.92 points, while leaving the number of covered queries unchanged. The additional candidates recover more relevant images for queries that are already covered. The 87.62% complete-union ceiling shows the remaining coverage

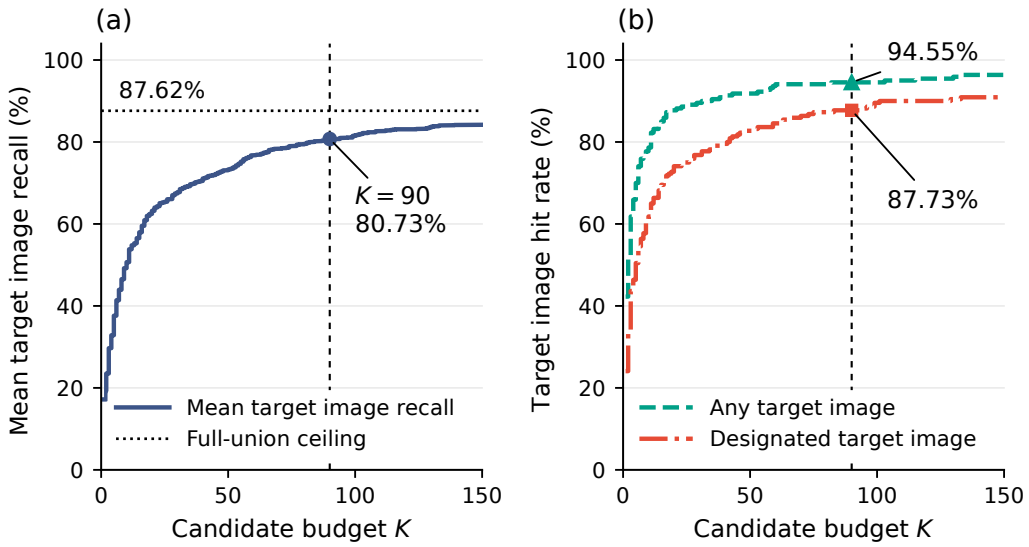


Figure 4: Coverage versus shortlist size on CIRCO validation. (a) Mean per-query relevant-image recall, with the complete-union ceiling at 87.62%. (b) Hit rates for any annotated relevant image and for the designated target. The complete union has a larger, query-dependent candidate count.

available within the retrieved pool. Recovering targets outside that pool requires changes to retrieval rather than a larger K.

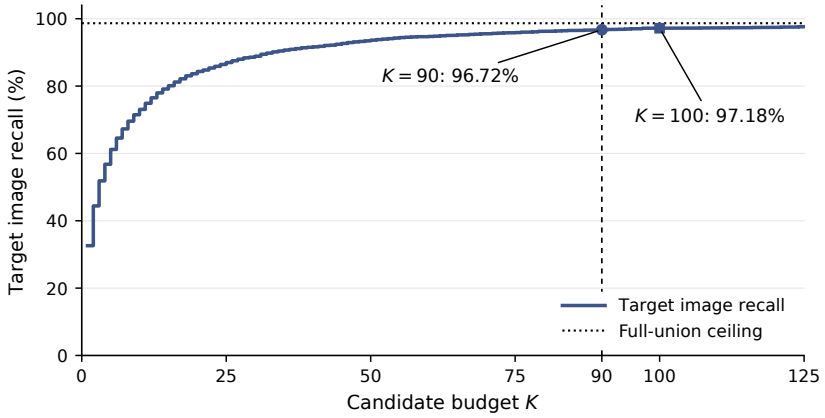


Figure 5: Coverage versus shortlist size on CIRR validation. The selected $K = 90$ retains 4,044 of 4,181 targets (96.72%). Increasing K to 100 retains 4,063 (97.18%). The dotted line denotes complete-union coverage, 4,125/4,181 (98.66%).

Selected operating budgets. On CIRR, increasing K from 90 to 100 recovers 19 additional targets, a gain of approximately 0.45 percentage points. For either dataset, this change requires 11.11% more statement-likelihood evaluations for the same semantic checks. We therefore use $K = 90$ on CIRCO and CIRR as a practical balance: it retains broad target coverage while avoiding the extra semantic-scoring workload for the smaller coverage gains from the next ten candidates. This comparison measures workload by the number of statement evaluations, rather than wall-clock latency. The coverage curves support this operating choice but do not establish an optimum for the final retrieval metrics. For Fashion-IQ, we use $K = 100$ in the category-specific configurations in Table 8. The budget sweeps reported here are limited to CIRCO and CIRR.

A.3 Sensitivity to the MLLM Scorer

We keep the candidate identities, semantic checks, and global relevance scores fixed and vary only the MLLM used for likelihood evaluation. Table 5 reports both semantic evidence alone (Q) and its fusion with the same global relevance signal (A) on the CIRCO test set.

Table 5: MLLM scorer comparison on the CIRCO test set with fixed semantic checks and candidate identities. Q uses semantic evidence alone, and S combines it with the global relevance score A.

Scoring model	Q: mAP@k				S: mAP@k			
	$k=5$	$k=10$	$k=25$	$k=50$	$k=5$	$k=10$	$k=25$	$k=50$
Qwen2.5-VL-7B-Instruct	27.91	29.63	32.60	33.73	39.37	40.18	42.81	43.71
Qwen3.5-9B	30.43	32.09	35.03	36.20	41.60	42.35	44.89	45.74
Qwen3.6-27B	25.87	27.39	30.30	31.47	38.26	39.40	42.18	43.03

All three scorers benefit consistently from evidence fusion. Relative to Q alone, S improves mAP@5 by 11.17–12.39 percentage points, with gains at every reported cutoff. With Qwen2.5-VL-7B-Instruct, the verifier used by WISER, CrossCheck reaches 39.37 mAP@5, exceeding WISER's best reported result of 36.53 in Table 1 by 2.84 points. The improvement therefore persists without upgrading the MLLM scorer. Moreover, the larger Qwen3.6-27B yields lower scores than Qwen3.5-9B for both Q and S, showing that a larger scorer does not necessarily produce better retrieval results. Although absolute performance varies with the scorer, the consistent fusion gains on fixed candidates and checks support the contribution of CrossCheck's scoring and fusion design. These results indicate that CrossCheck works effectively with various off-the-shelf MLLM scorers and that its gains are not solely attributable to a stronger scoring model.

A.4 Exact Complementarity Statistics

Tables 6 and 7 give the exact counts underlying Figure 3, using the per-query metric M_5 and comparison rules from Section 4.3. The diagnostics include all 220 CIRCO and 4,181 CIRR validation queries, with ViT-G/14, Qwen3.5-9B, and the settings in Table 8. The same $K = 90$ candidates are ranked by A, Q, and S for each query, without retuning the parameters for these diagnostics.

Table 6: Exact per-query outcomes at $k = 5$. Counts include every validation query. $\overline{\Delta}$ is the mean change in AP@5 (CIRCO) or RR@5 (CIRR), expressed in percentage points.

Dataset	Reference	Improved	Tied	Degraded	$\overline{\Delta}$
CIRCO	A	118	76	26	$+13.88$
	Q	118	58	44	$+14.59$
	Best(A, Q)	66	85	69	-0.78
CIRR	A	1,744	2,009	428	$+19.69$
	Q	847	2,920	414	$+4.34$
	Best(A, Q)	386	2,953	842	-5.43

Per-query outcomes. The mean AP@5 values on CIRCO are 31.14, 30.42, 45.01, and 45.79 for A, Q, S, and Best(A, Q), respectively. The corresponding mean RR@5 values on CIRR are 43.10, 58.45, 62.79, and 68.22.

Target recovery counts. The counting unit is an annotated query–image pair, giving 916 pairs for CIRCO and 4,181 for CIRR. These totals are the denominators of the target percentages in Figure 3.

We additionally count targets promoted into the fused top five that were absent from both individual top-five rankings. Let G_q be the relevant images for query q and $P_X^5(q) = G_q \cap \mathrm{Top}_5(X, q)$. The count is

$$B_5 = \sum_q \left| P_S^5(q) \setminus \left(P_A^5(q) \cup P_Q^5(q) \right) \right|. \tag{13}$$

Table 7: Relevant query–image pairs recovered or lost at $k = 5$. Net gain equals recovered minus lost. On CIRR, each count also equals the number of affected queries.

Dataset	Reference	Recovered	Lost	Net gain
CIRCO	A	123	28	+95
	Q	139	57	+82
CIRR	A	874	110	+764
	Q	330	152	+178

We obtain $B_5 = 32$ on CIRCO, spanning 31 queries, and $B_5 = 80$ on CIRR. For CIRCO, these newly recovered pairs do not imply that both individual lists lacked any relevant image, since a query can have multiple relevant images. The fixed shortlists contain 696 of the 916 CIRCO pairs and 4,044 of the 4,181 CIRR targets. All validation queries are included, and the AP denominator retains all annotated relevant images.

B Implementation Details

B.1 Hyperparameter Settings

Our experiments are conducted on NVIDIA GeForce RTX 5090 GPUs. Table 8 lists the dataset- and backbone-specific settings for the parameters defined in Section 3.

Table 8: Candidate budgets and semantic scoring settings. B/32, L/14, and G/14 denote ViT-B/32, ViT-L/14, and ViT-G/14, respectively. Dress, Shirt, and Tops&Tees are Fashion-IQ categories.

Dataset	Encoder	L	K	ω	λ	ρ	β
CIRCO	B/32	100	90	0.25	0.43	0.9	1.5
CIRCO	L/14	100	90	0.25	0.43	0.9	1.5
CIRCO	G/14	100	90	0.25	0.45	0.9	1.5
CIRR	B/32	100	90	0.25	0.35	0.05	1
CIRR	L/14	100	90	0.25	0.35	0.05	1
CIRR	G/14	100	90	0.25	0.35	0.05	1
Dress	B/32	100	100	0.5	0.25	0.1	0.8
Dress	L/14	100	100	0.5	0.16	0.1	0.8
Dress	G/14	100	100	0.5	0.1	0.1	0.8
Shirt	B/32	100	100	0.5	0.25	0.25	1
Shirt	L/14	100	100	0.5	0.25	0.25	1.2
Shirt	G/14	100	100	0.5	0.15	0.8	1
Tops&Tees	B/32	100	100	0.5	0.25	0.25	2
Tops&Tees	L/14	100	100	0.5	0.25	0.25	2
Tops&Tees	G/14	100	100	0.5	0.25	0.25	2

Hyperparameter selection and evaluation splits. For CIRCO and CIRR, hyperparameters are selected on the respective validation sets and then fixed for test-set evaluation. For Fashion-IQ, the official test evaluation service was suspended after the official challenge ended. Following prior work and the compared baselines, we therefore report results on the validation set.

B.2 Evidence Calibration

We provide this implementation detail in the appendix to keep the main text concise and focused on the core method. Different semantic checks can produce contrastive evidence with different centers and dynamic ranges. To keep these numerical scales from determining a check's influence on aggregation, we calibrate each check over the fixed shortlist $\mathcal{C}$ from Eq. 6. For the raw evidence $e_j(I)$ in Eq. 9, the implementation computes

$$\begin{aligned} \mu_j &= \mathrm{median}_{I \in \mathcal{C}}\, e_j(I), \\ \sigma_j &= \max\{c_{\mathrm{MAD}}\, \mathrm{median}_{I \in \mathcal{C}}\, |e_j(I) - \mu_j|,\ \epsilon_E\}, \\ z_j(I) &= \frac{e_j(I) - \mu_j}{\sigma_j}. \end{aligned} \tag{14}$$

Here $c_{\mathrm{MAD}} > 0$ is the median absolute deviation (MAD) scale factor, and $\epsilon_E > 0$ is a scale floor that limits amplification when a check has little variation across candidates. Both median operations use the lower median when the shortlist size is even. The implementation evaluates $M(I)$, $B(I)$, and $Q(I)$ in Eq. 11 with $z_j(I)$ in place of $e_j(I)$; the mean–softmin aggregation and the fusion with $A(I)$ in Eq. 12 otherwise retain their stated forms. This calibration uses only model scores within the current query's shortlist, without relevance labels. The calibrated values express relative evidence among candidates, not calibrated probabilities that the corresponding visual requirements are satisfied.

B.3 CIRR Subset Evaluation.

Subset predictions are extracted from a completed full-gallery ranking after shortlist scoring. The implementation proceeds as follows:

1. **Rerank the shortlist.** Sort the images in $\mathcal{C}$ by decreasing fused score $S(I)$ from Eq. 12. This sort is stable: equal scores preserve the incoming candidate order.
2. **Complete the ranking.** Append images in $\mathcal{U} \setminus \mathcal{C}$ in their original global-relevance (admission) order given by $A(I)$. With the reported CIRR budget $K = 90$, these are the retrieved union members outside the top-90 shortlist. Next append images in $\mathcal{D} \setminus \mathcal{U}$ in gallery-index order. Throughout this procedure, exclude I_{ref} and any image already included in the ranking.
3. **Extract the official subset.** Filter the completed ranking to retain only images in the query's official `group_members`, preserving their relative order, and write the first three images to the subset submission file.

Group members outside $\mathcal{C}$ receive no additional MLLM scoring; their positions follow the ranking-completion order above. We do not recompute $A(I)$ or $Q(I)$, or repeat evidence normalization, within the subset. The official `group_members` are used only in the evaluation/export stage after scoring, not to construct the shortlist or to perform a separate group-level reranking.

B.4 Generation and Scoring Prompts

Figure 6 presents a condensed version of the semantic-check generation prompt. Braces denote query inputs, output fields, or omitted few-shot demonstrations. The JSON fields `target` and `contrast` correspond to T_j^+ and T_j^- in Eq. 7. The single-fact and same-slot rules are instructions to the generator, not guarantees for every generated pair. A generated check may combine attributes or fail to isolate a local contrast; some queries may also be difficult to express under these constraints. Such deviations are a limitation of automatic check construction. Figure 10(g) shows one concrete example, discussed in its caption.

Inputs
`{reference picture}`
`{modification text}`

Task
You create visual criteria for composed image retrieval. You receive one reference image and a modification text. Imagine the target image after the requested change, then describe the facts that would distinguish a correct target from an incorrect one.

Rules
1. Each criterion checks exactly one visual fact.
2. Split independent requested changes into separate criteria, and cover every requested change.
3. `target` states the fact that a correct target image should show.
4. `contrast` states one mutually exclusive alternative for the same subject and the same property, count, action, relation, or scene slot.
5. Write the contrast as a specific fact that can be judged from an image.
6. Use the reference image only to understand the subject, pronouns, source state, and relative changes.

Output format
Return JSON only:

```
{"criteria": [
  {"target": "{target statement}",
   "contrast": "{contrast statement}"}
]}
```

`{few-shot examples}`
Modification text: `{modification text}`

Figure 6: The semantic-check construction prompt, with input-dependent content and few-shot examples represented by placeholders. The six rules retain the original template wording.

Likelihood-scoring prompt. The reference image is supplied as Picture 1 and the candidate as Picture 2. The following is a prefix instruction used in the implementation:

> Picture 1 is the reference image and Picture 2 is the candidate image. Continue with exactly one concise target statement that is true of Picture 2. Use Picture 1 only for comparison. Do not explain.

Statement likelihoods are evaluated as described in Section 3.3.

C Supplementary Derivations

C.1 Relationship to Binary Verification

Paired likelihood scoring includes binary answer-token scoring as a special case under the same conditioning context. Let y and n be single tokens representing yes and no, and let $\pi_j^a(I) = P_{\mathrm{MLLM}}(a \mid \mathsf{X}_j(I), T_{\mathrm{pre},j})$ for $a \in \{y, n\}$, where $T_{\mathrm{pre},j}$ is a fixed verification prefix. Renormalizing the two answer probabilities gives

$$c_j(I) = \frac{\pi_j^y(I)}{\pi_j^y(I) + \pi_j^n(I)} = \mathrm{sigmoid}\big(\log \pi_j^y(I) - \log \pi_j^n(I)\big)\,. \tag{15}$$

This two-label normalization does not guarantee calibration to visual requirement satisfaction.

Choose $T_j^+ = T_{\mathrm{pre},j} \,\|\, y$ and $T_j^- = T_{\mathrm{pre},j} \,\|\, n$ with equal scored length $n_j = |T_{\mathrm{pre},j}| + 1$. At $\lambda = 1$, the shared-prefix log-likelihood cancels, yielding

$$e_j^{(1)}(I) = \frac{1}{n_j} \log \frac{\pi_j^y(I)}{\pi_j^n(I)}, \qquad c_j(I) = \mathrm{sigmoid}\Big(n_j e_j^{(1)}(I)\Big)\,. \tag{16}$$

If the prefix belongs entirely to the conditioning context and only the answer token is scored, $n_j = 1$ and the contrast is exactly the yes/no log-odds. The reduction requires identical contexts and tokenized prefixes, equal scored lengths, $\lambda = 1$, and no scored suffix after the answer. Other settings require the full sequence calculation.

C.2 Properties of Mean–Softmin Aggregation

For a fixed candidate, the normalized softmin in Eq. 11 satisfies

$$\min_j e_j \;\le\; -\tau \log\left(\frac{1}{m}\sum_{j=1}^{m} \exp(-e_j/\tau)\right) \;\le\; \frac{1}{m}\sum_{j=1}^{m} e_j. \tag{17}$$

The lower bound follows by bounding each exponential by $\exp(-\min_j e_j/\tau)$. The upper bound follows from Jensen's inequality. As $\tau \to 0^+$, the softmin approaches the minimum. As $\tau \to \infty$, it approaches the mean. The factor $1/m$ ensures that equal inputs return the same value. The aggregate Q is nondecreasing in each check's evidence. Setting $\rho = 0$ recovers the mean, and for $m = 1$ both terms coincide. These properties concern the aggregation function and also hold when its inputs are the calibrated values $z_j(I)$ from Eq. 14, with calibration statistics held fixed. They do not guarantee satisfaction of the underlying semantic requirements.

For stable numerical evaluation, the softmin can be written as

$$B(I) = -\tau\big[\mathrm{LSE}_j(-e_j(I)/\tau) - \log m\big],$$

where LSE denotes log-sum-exp.

D Qualitative Retrieval Examples

We show three validation queries per dataset with ViT-G/14, selecting cases with at least two saved semantic-check pairs. Reference and target images appear on the left. Rows A, Q, and S show the top five candidates from the same fixed shortlist ($K = 90$ for CIRCO/CIRR and $K = 100$ for Fashion-IQ). Green frames mark annotated target images. Unmarked images are not automatically incorrect. The displayed checks were generated using the prompt in Appendix B.4. Here, Q is the complete semantic component of the saved final score. To save space, the figures omit the repeated prefix "The target image shows" where present and standardize "picture" to "image" in the displayed checks, retaining the generated content. These are post-hoc illustrations, not a random sample.

Ranking corrections and limitations. The donut target in (a) improves from ranks 5 and 4 under A and Q to first under S. The close-up portrait in (b) improves from ranks 12 and 2 to first. For CIRR, the bird target in (d) improves from ranks 8 and 3 to first. The smaller dog in (e) improves from rank 10 under both individual scores to rank 3. In (g)–(h), semantic evidence promotes the dress and striped-shirt targets from A ranks 71 and 35 to first. Fusion does not always help: the designated targets in (c), (f), and (i) fall from A ranks 1, 2, and 1 to S ranks 10, 11, and 9. Displayed checks expose the distinctions being examined, but these examples do not identify any single check's causal effect.

(a) CIRCO Target rank: *A* 5 / *Q* 4 / *S* 1

Modification: is chocolate-flavored and is held by a hand

	Target statement T_j^+	Contrast statement T_j^-
1	a chocolate-flavored donut.	a powdered donut.
2	the donut being held by a hand.	the donut resting on a surface.



Figure 7: CIRCO validation query 120: a chocolate-flavored donut held by a hand. Two semantic checks are shown. Rankings under A, Q, and S place the designated target 5th, 4th, and 1st, respectively. Numbers above the images denote ranks 1–5.

(b) CIRCO Target rank: *A* 12 / *Q* 2 / *S* 1

Modification: has only one person wearing the same outfit, the photo is zoomed in

	Target statement T_j^+	Contrast statement T_j^-
1	only one person wearing the same outfit.	multiple people wearing the same outfit.
2	The target image is zoomed in.	The target image is zoomed out.



(c) CIRCO Target rank: *A* 1 / *Q* 33 / *S* 10

Modification: has only one person and is walking away from the camera

	Target statement T_j^+	Contrast statement T_j^-
1	only one person.	two people.
2	the person walking away from the camera.	the person facing the camera.

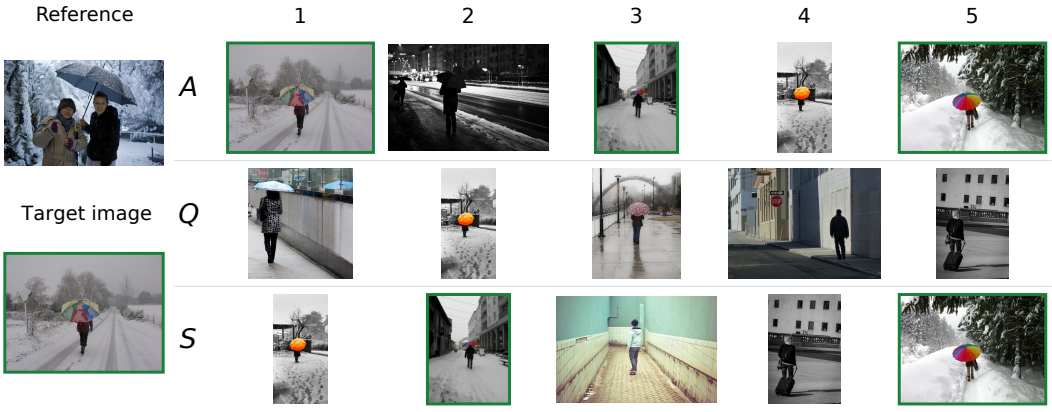
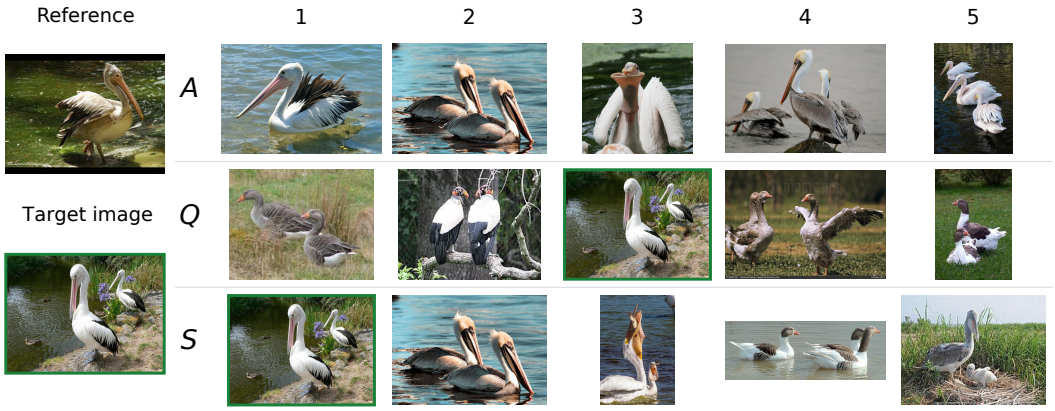


Figure 8: CIRCO validation examples. (b) Query 5 requests one person in the same outfit and a closer view. (c) Query 51 requests one person walking away. Each query has two checks. In (c), fusion demotes the designated target. Rows show rankings under A, Q, and S. Green frames mark annotated target images.

(d) CIRR Target rank: *A* 8 / *Q* 3 / *S* 1

Modification: One more bird and they are further away from the water.

	Target statement T_j^+	Contrast statement T_j^-
1	two birds.	only one bird.
2	The birds are positioned further away from the water.	The birds are positioned close to the water.



(e) CIRR Target rank: *A* 10 / *Q* 10 / *S* 3

Modification: Smaller dog lying on the grass.

	Target statement T_j^+	Contrast statement T_j^-
1	a smaller dog.	a larger dog.
2	the dog lying on the grass.	the dog standing on the grass.

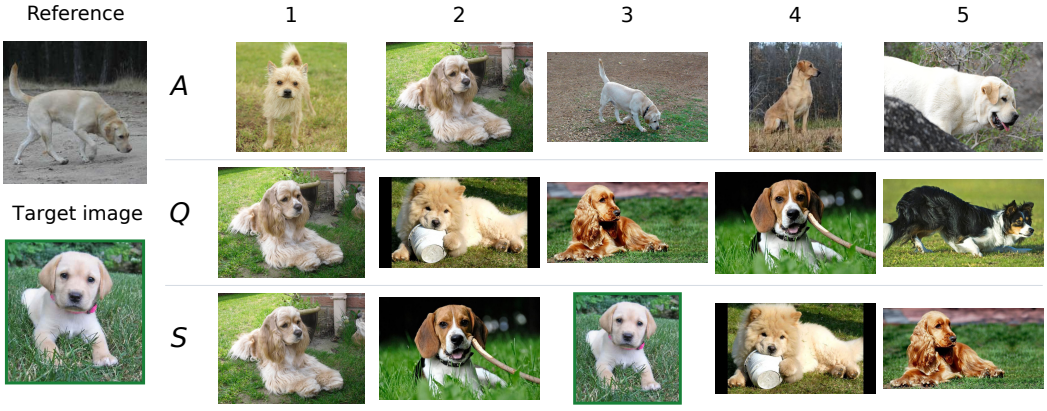
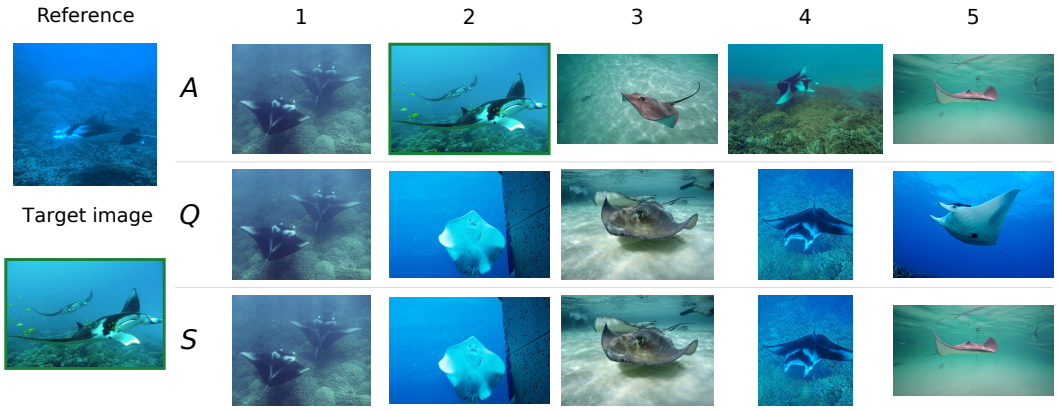


Figure 9: CIRR validation examples. (d) Query 36803 changes bird count and distance from water. (e) Query 37823 changes dog size and posture. Each query has two checks. The examples illustrate a correction and a partial correction. Rows show rankings under A, Q, and S. Green frames mark annotated target images.

(f) CIRR Target rank: *A* 2 / *Q* 12 / *S* 11

Modification: Increase the number of rays to two rays swimming towards the camera

	Target statement T_j^+	Contrast statement T_j^-
1	two rays swimming towards the camera.	only one ray swimming towards the camera.
2	two rays swimming towards the camera.	two rays swimming away from the camera.



(g) Fashion-IQ Dress Target rank: *A* 71 / *Q* 1 / *S* 1

Modification: Is longer and more Asian-inspired and is longer and shiny black

	Target statement T_j^+	Contrast statement T_j^-
1	a dress that is longer than the reference dress.	a dress that is shorter than the reference dress.
2	a dress with an Asian-inspired design.	a dress with a Western-inspired design.
3	a dress that is shiny black.	a dress that is matte purple.



Figure 10: Validation examples. (f) CIRR query 12966 specifies two rays swimming toward the camera (two checks). Fusion demotes the designated target. (g) Fashion-IQ Dress query 99 specifies length, design, and appearance (three checks). Semantic evidence promotes the target to first. The generated pair "a dress that is shiny black" versus "a dress that is matte purple" changes both gloss and color, violating the single-fact, same-slot rules in Appendix B.4. We retain this imperfect output to illustrate that the prompt does not guarantee isolated attribute contrasts; its score cannot separate evidence for gloss from evidence for color. Rows show rankings under A, Q, and S. Green frames mark annotated target images.

(h) Fashion-IQ Shirt Target rank: *A* 35 / *Q* 1 / *S* 1

Modification: is black with white and yellow stripes and it is black and has stripped print

	Target statement T_j^+	Contrast statement T_j^-
1	a black t-shirt with white and yellow stripes.	a grey t-shirt with white and yellow stripes.
2	a t-shirt with a striped print.	a t-shirt with a solid print.



(i) Fashion-IQ Tops&Tees Target rank: *A* 1 / *Q* 73 / *S* 9

Modification: is orange with shorter sleeves and is of orange camo pattern

	Target statement T_j^+	Contrast statement T_j^-
1	a black top with an orange camouflage pattern.	a top with a black camouflage pattern.
2	a top with shorter sleeves.	a top with long sleeves.



Figure 11: Fashion-IQ validation examples. (h) Shirt query 919 changes color and print. (i) Tops&Tees query 9 changes color/pattern and sleeve length. Each query has two checks. The orange camouflage target is a failure case. Rows show rankings under A, Q, and S. Green frames mark annotated target images.