FOLLOW THE EDIT: COUNTERFACTUAL DIAGNOSTICS AND ON-POLICY RANK DISTILLATION FOR COMPOSED IMAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Composed image retrieval (CIR) retrieves target images based on the reference image and an editing instruction, satisfying the requested changes while preserving relevant visual content. Conventional retrieval metrics measure whether targets appear among the top-ranked results but do not directly assess responses to changes in the user's editing instruction. We introduce CIR-QDiag, a dataset of 10,131 local instruction edits, together with two complementary metrics, Directional Response Rate (DRR) and Calibrated Response Rate (CRR), to assess instruction following in CIR through models' responses to these edits. Our diagnostic reveals that high retrieval recall does not guarantee high scores on both metrics and that instruction sensitivity varies across models. To improve CIR performance by transferring instruction-sensitive relevance judgments from stronger models to smaller retrieval models, we propose On-Policy Rank Distillation (OPRD), which matches a frozen teacher's soft relevance distribution over student-mined candidates. The student minimizes reverse KL on the shared candidates, focusing supervision on its current rankings. This transfers soft ranking information beyond the annotated target, while retaining student-only inference. OPRD trains on original CIR pairs without diagnostic edits. With BGE-MLLM-S2 as the teacher, OPRD achieves higher CRR than the InfoNCE baseline across all four student checkpoints. For CLIP-L, it improves CRR from 2.00% to 6.22% and CIRR Recall@1 from 12.80% to 42.72% relative to the pretrained checkpoint, with additional retrieval gains on PinPoint and FashionIQ. Code: `https://anonymous.4open.science/r/QDiag-OPRD-877B/`.

1 INTRODUCTION

Composed image retrieval (CIR) retrieves target images using a reference image and an editing instruction, enabling applications such as e-commerce search (Vo et al., 2019; Wu et al., 2021). The targets should satisfy the requested changes while preserving relevant reference-image content.

Given a reference image of a long white coat and the user's editing instruction "change the coat to blue and make it shorter," a short red coat or a long blue coat satisfies only part of the request, yet a model relying on overall visual similarity could rank either highly. Recall@k measures whether an annotated target appears among the top-k results, but does not isolate how individual conditions affect its ranking. Even retrieving the correct target does not establish whether both color and length contributed. Replacing "blue" with "red" while keeping the length condition and both images fixed lets us examine the color condition directly: the change in the *target similarity score*, the model's estimated query–target match, reveals its response to the changed requirement.

We introduce CIR-QDiag for such *counterfactual diagnostics*: locally changing one instruction condition and measuring the response for fixed images. Built from PinPoint (Mahadev et al., 2026), it contains 10,131 edits involving objects, attributes, and predicates, including actions and relations. Each edit adds, removes, or substitutes a condition while preserving the remaining instruction. ADD and SUBSTITUTE introduce conditions incompatible with the annotated targets; DELETE removes a constraint and may leave the targets valid. These comparisons complement retrieval evaluation by isolating responses to specific instruction changes.

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.

Directional Response Rate (DRR) measures how often an edit decreases the target similarity score. However, meaning-preserving rewordings can also change that score. We therefore estimate a model-specific threshold from paraphrase variation and define *Calibrated Response Rate (CRR)* as the frequency of decreases exceeding this threshold. For DELETE, a decrease reflects sensitivity to constraint removal rather than rejection of an incompatible target. DRR helps interpret low CRR by distinguishing infrequent score decreases from frequent decreases that do not exceed the calibration threshold.

Our diagnostic shows that high recall can coexist with high DRR but low CRR, indicating that many target similarity score decreases do not exceed the paraphrase-based threshold. In our evaluation, larger embedding models, including BGE-MLLM-S2 (Zhou et al., 2024) and Qwen3-VL-Emb-8B (Li et al., 2026a), achieve higher CRR than smaller pretrained CLIP models (Radford et al., 2021). This gap motivates us to investigate whether their relevance assessments can improve both retrieval and local edit sensitivity in smaller models. We therefore propose *On-Policy Rank Distillation (OPRD)*, which uses a larger model as a frozen teacher to supervise a smaller student's current rankings. When teacher scores distinguish instruction-satisfying candidates from visually similar partial matches, the resulting soft ranking provides graded supervision beyond the target label. Matching these relative preferences may help the student discriminate candidates according to fine-grained instruction conditions. CIR-QDiag evaluates whether this supervision on original queries transfers to local edit responses relative to paraphrase variation.

OPRD minimizes reverse Kullback–Leibler (KL) divergence from the student to the teacher over student-mined candidates including the target. Here, *on-policy* denotes student-driven selection refreshed each epoch; student scores are recomputed at every step. Training uses original CIR pairs, diagnostic edits are reserved for evaluation, and inference uses only the student.

With BGE-MLLM-S2 as the teacher, OPRD achieves higher aggregate CRR than the InfoNCE contrastive baseline (van den Oord et al., 2018) across all four student checkpoints. For CLIP-L, CRR increases from 5.09% to 6.22% and Recall@1 on CIRR (Liu et al., 2021) from 27.63% to 42.72%. CIRR-only training also yields retrieval gains on PinPoint and FashionIQ.

Our contributions are: **(1)** CIR-QDiag with directional and paraphrase-calibrated response metrics; **(2)** OPRD for soft relevance distillation over periodically refreshed, student-mined candidates; and **(3)** evaluations across four student checkpoints and four benchmarks, with ablations of candidate selection, supervision, and target insertion.

2 Related Work

2.1 Composed Image Retrieval and Evaluation

CIR retrieves images that preserve relevant reference-image content while satisfying the editing instruction. Approaches learn image–text composition functions (Vo et al., 2019) or leverage pretrained CLIP features (Baldrati et al., 2022) and large language models (Huynh et al., 2025). Recent work addresses visual identity through instance-level CIR (Psomas et al., 2025), fine-grained instruction semantics through FineCIR (Li et al., 2025), and clause–entity alignment through TEMA (Li et al., 2026b). Gallery information supports query reconstruction via adaptively weighted neighbors in SADN (Wang et al., 2024) and candidate discrimination via relevance-aware hard-negative sampling in QuRe (Kwak et al., 2025).

Evaluation extends beyond aggregate recall: CIRR-D diagnoses fine-grained textual understanding (Sun et al., 2023), and EDIR examines categories of editing instructions (Song et al., 2026). PinPoint includes explicit negative images, multi-image queries, and paraphrase testing (Mahadev et al., 2026); CIRCUS audits whether a single modality suffices to solve queries (Attimonelli et al., 2026). Using controlled image generation, Liu et al. (2026) introduce FISD to address query ambiguity and also evaluate multi-round retrieval. Controlled textual contrasts in ARO test attributes, relations, and word order (Yuksekgonul et al., 2023), while SugarCrepe constructs plausible negative captions with reduced linguistic artifacts (Hsieh et al., 2023). These benchmarks assess image–caption compatibility. CIR-QDiag instead fixes the reference and target images and measures target similarity score responses to single-condition instruction edits relative to paraphrase variation.

2.2 KNOWLEDGE DISTILLATION FOR RETRIEVAL

Knowledge distillation transfers teacher knowledge to more efficient students (Hinton et al., 2015). RankDistil preserves the teacher's top-ranked orderings while selecting hard negatives using student scores (Reddi et al., 2021). In CIR, DistillCIR distills instruction awareness through generated reasoning and multimodal feature alignment (Zhong et al., 2025); TG-CIR and IUDC supervise matching relations over batch candidates (Wen et al., 2023; Ge et al., 2025). EviRank distills relevance scores and hard-pair comparisons for downstream re-ranking (Du et al., 2026).

In text retrieval, RocketQAv2 jointly trains a retriever and re-ranker with listwise relevance distillation (Ren et al., 2021), while AR^2 couples retriever-mined candidates with soft ranker supervision in joint adversarial training (Zhang et al., 2022). Candidate composition matters: Jang et al. (2026) show benefits from stratified sampling across teacher scores over top-K sampling within a fixed pool. GKD studies on-policy distillation with teacher feedback on student-generated language sequences (Agarwal et al., 2024). OPRD builds on candidate-level retrieval distillation: a frozen embedding teacher supervises the student's periodically refreshed image candidates for queries combining visual context and editing instructions. Reverse KL matches relevance distributions over shared candidates, while CIR-QDiag evaluates transfer to local edit responses without diagnostic-edit training.

3 METHOD

3.1 PROBLEM SETUP

A composed query is $x = (\mathcal{R}, m)$, where $\mathcal{R}$ contains one or two reference images and m is the editing instruction. A retriever ψ maps queries and gallery images $y \in \mathcal{G}$ to unit-normalized embeddings and ranks images by their inner product:

$$q_\psi(x) = f_\psi(\mathcal{R}, m), \qquad z_\psi(y) = g_\psi(y), \qquad s_\psi(x, y) = q_\psi(x)^\top z_\psi(y). \tag{1}$$

For an annotated target t, $s_\psi(x, t)$ is the target similarity score. CIR-QDiag compares original and diagnostically edited instructions with images fixed; OPRD trains on original instructions. We denote student and frozen-teacher parameters by θ and ϕ. The teacher and student can use different embedding spaces because OPRD matches their relevance distributions over the same candidate images. No geometric alignment between the two embedding spaces is required.

3.2 COUNTERFACTUAL DIAGNOSTICS WITH CIR-QDIAG

CIR-QDiag measures target similarity score responses to local instruction edits relative to paraphrase variation. We build it from PinPoint (Mahadev et al., 2026) through the three stages in Figure 1.

The publicly released PinPoint data contain 7,635 queries, including human-verified paraphrases with shared retrieval intent and annotations. After removing empty and duplicate candidate IDs, we match reference-image IDs and complete positive- and official negative-image sets to recover paraphrase groups. Keeping groups with two to six instructions yields 1,683 groups and 6,553 instructions. For group g, $\mathcal{I}_g$ and $\mathcal{P}_g$ denote its original instructions and annotated positives. These groups supply editing inputs and calibration paraphrases.

3.2.1 LOCAL INSTRUCTION EDITING

Given an original instruction m_i^+, `gpt-5.6-terra`[1] generates edits without images. We cross ADD, DELETE, and SUBSTITUTE with object identity or presence, attributes, and predicates (actions or relations), producing at most nine proposals per instruction. Each changes one contiguous span while preserving other content, apart from necessary whitespace and terminal-punctuation adjustments. Inapplicable edits may be omitted.

[1] OpenAI model documentation: `https://developers.openai.com/api/docs/models/gpt-5.6-terra`.

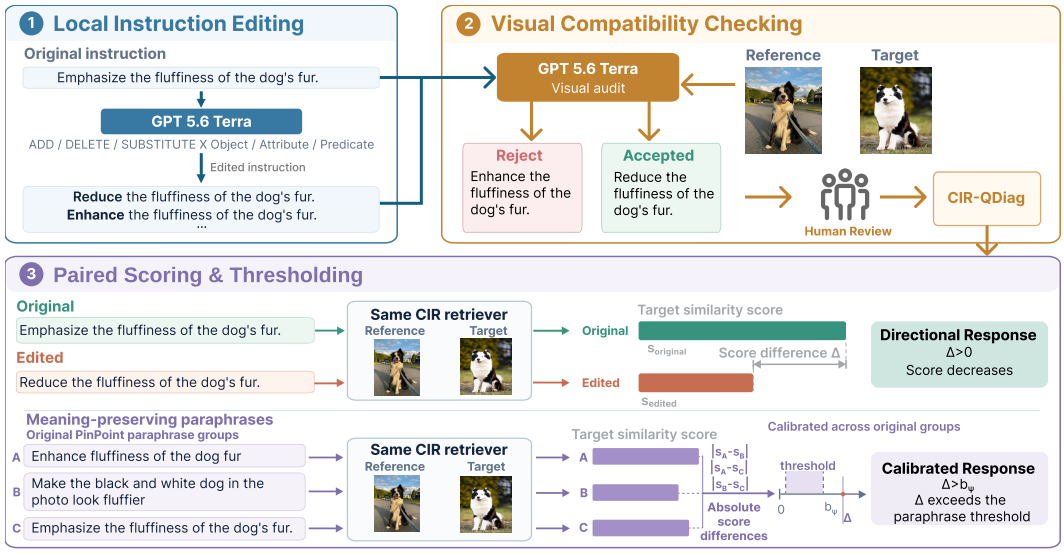

Figure 1: **CIR-QDiag construction and evaluation.** Text-only local editing is followed by visual validation and human review. With reference and target images fixed, paired scoring tests Directional Response ($\Delta^{\psi} > 0$) and Calibrated Response ($\Delta^{\psi} > b_{\psi}$). The threshold b_{ψ} is estimated from original paraphrases across calibration groups using the same retriever. Group-balanced event rates give DRR and CRR. The accepted example illustrates SUBSTITUTE; score bars are schematic.

3.2.2 Visual Compatibility Checking

A separate group-level call to the same model checks proposals against their original instructions, reference images, and annotated positives. It verifies the edited span, category, fluency, content preservation, and operation-specific visual criteria. ADD and SUBSTITUTE must introduce a visually decidable condition false for the positives. DELETE must remove a meaningful constraint while preserving a well-formed instruction; positives may remain compatible. Official negative images are excluded from generation and validation. Model-accepted proposals are divided among five human reviewers for individual review; prompts and review procedures appear in Appendices E and B.1.1.

CIR-QDiag contains 10,131 accepted edits over 930 groups and 3,308 original instructions. Aggregate results include all three operations; full counts and operation-specific analyses appear in Appendices B.1 and B.3.

3.2.3 Paired Scoring & Thresholding

Directional Response. For an original instruction m_i^+, accepted edit $m_{i,n}^-$, and target $t \in \mathcal{P}_g$, we hold the references $\mathcal{R}_g$ fixed and compute

$$\Delta^{\psi}_{g,i,n,t} = s_{\psi}((\mathcal{R}_g, m_i^+), t) - s_{\psi}((\mathcal{R}_g, m_{i,n}^-), t). \tag{2}$$

A Directional Response occurs when $\Delta^{\psi} > 0$. For ADD and SUBSTITUTE, this is the expected decrease after introducing a target-incompatible condition. For DELETE, it measures a decrease after constraint removal, not rejection of an invalid target. Superscripts $+$ and $-$ therefore identify original and edited instructions rather than target validity. Operation-specific results distinguish these interpretations.

Calibrated Response. We estimate wording variation from target similarity score differences between original paraphrases, keeping images fixed:

$$\epsilon^{\psi}_{g,i,j,t} = \left| s_{\psi}((\mathcal{R}_g, m_i^+), t) - s_{\psi}((\mathcal{R}_g, m_j^+), t) \right|, \qquad i < j. \tag{3}$$

The model-specific threshold b_{ψ} is the group-balanced 95th percentile of these absolute differences, assigning equal total weight to each calibration group. Calibrated Response requires the same Δ^{ψ}

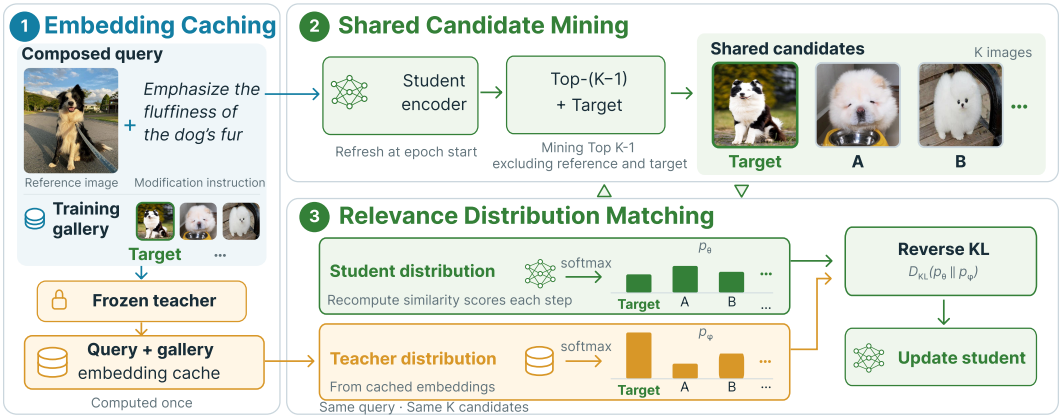

Figure 2: **On-Policy Rank Distillation (OPRD).** The frozen teacher caches training-query and gallery embeddings once. At each epoch's start, the student mines $K-1$ non-reference, non-target images and inserts the annotated target. Candidate IDs remain fixed within the epoch; student scores are recomputed at every optimization step. Reverse KL matches student and teacher distributions over these shared candidates and updates only the student. Probability bars are schematic.

to exceed b_ψ, expressing the response relative to the retriever's paraphrase variation. The response indicators are

$$I^{\psi}_{\text{dir}} = \mathbf{1}[\Delta^{\psi} > 0], \qquad I^{\psi}_{\text{cal}} = \mathbf{1}[\Delta^{\psi} > b_\psi], \tag{4}$$

where $\mathbf{1}[\cdot]$ is the indicator function. Figure 1 illustrates these events with sample indices suppressed; its three paraphrases show one group, whereas b_ψ is estimated across calibration groups.

Group-balanced rates and controls. DRR and CRR are group-balanced percentages of their respective events. Averaging targets within edits, edits within instructions, instructions within groups, and then groups also defines Mean Δ (Appendix B.2). Uncertainty uses a group bootstrap preserving internal structure. We report overall DRR, CRR, Mean Δ, and b_ψ, together with category- and operation-specific response analyses.

Random-image and random-query controls assess general query–image separation by substituting images or queries from other groups (Appendix B.4).

3.3 On-Policy Rank Distillation

OPRD matches a frozen teacher's relevance distribution on student-selected candidates using reverse KL, as summarized in Figure 2.

3.3.1 Embedding Caching

For original CIR training pairs (x, y^+) and the training gallery, the frozen teacher ϕ encodes each query and image once, caching embeddings for reuse throughout training. CIR-QDiag diagnostic edits are never used for training.

3.3.2 Shared Candidate Mining

At the start of epoch r, the student θ_r encodes the training queries and gallery. For each query $x = (\mathcal{R}, m)$, it selects the $K-1$ highest-scoring gallery images excluding references and the annotated target, then inserts the target:

$$\mathcal{C}^K_r(x) = \{y^+\} \cup \operatorname*{Top}_{K-1,\; y \in \mathcal{G} \setminus (\mathcal{R} \cup \{y^+\})} s_{\theta_r}(x, y). \tag{5}$$

Here Top_{K-1} returns image identities. Target insertion ensures that y^+ participates even when ranked poorly. Mining uses no gradients; candidate IDs remain fixed within the epoch while student embeddings and scores are recomputed at every optimization step. Refreshing candidates with the updated student each epoch defines OPRD's on-policy component.

3.3.3 RELEVANCE DISTRIBUTION MATCHING

Student and teacher distributions. For each query, the student scores shared candidates using recomputed embeddings and the teacher uses its cache. Both apply softmax with temperature T:

$$p_\psi(y \mid x, \mathcal{C}_r^K) = \frac{\exp(s_\psi(x, y)/T)}{\sum_{c \in \mathcal{C}_r^K(x)} \exp(s_\psi(x, c)/T)}, \qquad \psi \in \{\theta, \phi\}. \tag{6}$$

Probabilities are aligned by candidate identity, transferring the teacher's allocation of relevance among the target and competing images. Section 4.1 specifies K and T.

Student update. For a minibatch $\mathcal{B}$ of original training queries, we minimize reverse KL from student to teacher:

$$\mathcal{L}_{\mathrm{OPRD}}(\theta) = \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} D_{\mathrm{KL}}\big(p_\theta(\cdot \mid x, \mathcal{C}_r^K(x)) \,\big\|\, p_\phi(\cdot \mid x, \mathcal{C}_r^K(x))\big)\,. \tag{7}$$

Here, $D_{\mathrm{KL}}(p\|q) = \sum_y p(y) \log \frac{p(y)}{q(y)}$ denotes the KL divergence, where p and q are the student and teacher distributions over the same candidate set, respectively. The loss penalizes candidates assigned high probability by the student but low probability by the teacher. Gradients update only student parameters through query and candidate-image embeddings; the teacher and its cached embeddings remain fixed. Inference uses only the student.

Student parameterization. Mining and distillation use the same student encoder. For CLIP students, reference-image and instruction embeddings z_r and z_m are combined by a residual head:

$$q_\theta(x) = \mathrm{norm}(z_r + z_m + h_\theta(z_r, z_m, z_r \odot z_m))\,. \tag{8}$$

Here h_θ models embedding interactions, $\odot$ is element-wise multiplication, and norm is ℓ_2 normalization. OPRD also supports other composed-query encoders; backbone adaptation and initialization are specified in Sec. 4.1.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Data and metrics. All adaptation uses only CIRR training data: 28,225 queries and 16,939 gallery images (Liu et al., 2021). We evaluate retrieval on CIRR validation, PinPoint (Mahadev et al., 2026), and FashionIQ validation (Wu et al., 2021), and local edit sensitivity on CIR-QDiag. The diagnostic reports group-balanced DRR, CRR, Mean Δ, the paraphrase threshold b_ψ, and group-bootstrap confidence intervals (CIs). Bootstrap CIs quantify evaluation uncertainty, not training-seed variation. Retrieval metrics include CIRR gallery/subset Recall@k (R@k), PinPoint mAP and Negative Recall (NegRecall), and FashionIQ recall macro-averaged over dress, shirt, and toptee. NegRecall measures retrieval of annotated negatives (lower is better); noNeg evaluation removes these negatives. Response rates and retrieval metrics are percentages. Protocols and calibration details appear in Appendices A.3 and B.2.

Models and training. We compare Original, InfoNCE, centered mean squared error (CMSE), and OPRD for CLIP-L, CLIP-B, Qwen3-VL-Emb-2B (Li et al., 2026a), and Qwen3-VL-2B-Instruct (Bai et al., 2025). Original denotes the pretrained checkpoint; InfoNCE uses one-hot target labels; CMSE adapts centered score distillation (Dubovikov et al., 2026). BGE-MLLM-S2 (†) supervises the main CMSE/OPRD comparisons; Qwen3-VL-Emb-8B (‡) is an alternative teacher. Superscripts identify the teacher, and both are also evaluated as retrievers. OPRD uses $K = 32$, $T = 0.05$, and epoch-wise candidate refresh. CLIP updates both encoders and its composition head; Qwen uses LoRA with rank and α both 32. Training lasts four epochs on one NVIDIA H100 GPU. Details are in Appendices A.1 and A.2.

4.2 CIR-QDIAG: LOCAL EDIT SENSITIVITY

Responses relative to paraphrase variation. In Table 1, BGE-MLLM-S2 achieves 70.08% DRR but 7.94% CRR; pretrained CLIP-L achieves 64.80% and 2.00%, respectively. DRR counts even

arbitrarily small decreases, whereas CRR identifies those exceeding the paraphrase threshold. Their gap shows that many score decreases remain within paraphrase variation. Random-image and random-query controls measure broad query–image separation rather than local edit responses (Appendix B.4).

Table 1: CIR-QDiag local edit sensitivity across all operations. DRR, CRR, and bootstrap CIs are percentages; bold marks the highest rate per group. Higher rates indicate more frequent responses, not greater semantic correctness.

Model / backbone	Method	Mean Δ	DRR $\uparrow$	CRR $\uparrow$	CRR bootstrap CI	b_ψ
Teacher Models						
BGE-MLLM-S2†		0.017767	**70.08**	7.94	[6.82, 9.14]	0.071220
Qwen3-VL-Emb-8B‡		0.032877	69.56	**10.21**	[9.04, 11.39]	0.112892
Student Models						
CLIP-L	Original	0.005471	64.80	2.00	[1.52, 2.52]	0.042547
	InfoNCE	0.017783	68.00	5.09	[4.24, 5.97]	0.091210
	CMSE†	0.010219	**68.54**	4.62	[3.85, 5.46]	0.053417
	OPRD†	0.014627	67.36	**6.22**	[5.30, 7.16]	0.072281
CLIP-B	Original	0.004775	65.26	2.39	[1.89, 2.99]	0.036016
	InfoNCE	0.017640	67.48	5.15	[4.33, 6.04]	0.092273
	CMSE†	0.009539	**67.76**	5.17	[4.30, 6.13]	0.049115
	OPRD†	0.013778	64.98	**6.10**	[5.28, 7.00]	0.071699
Qwen3-VL-Emb-2B	Original	0.026134	**68.38**	8.69	[7.58, 9.83]	0.103842
	InfoNCE	0.036515	66.94	8.68	[7.57, 9.79]	0.148849
	CMSE†	0.013464	59.10	5.97	[5.22, 6.77]	0.085236
	OPRD†	0.019574	62.09	**10.24**	[9.14, 11.32]	0.084918
Qwen3-VL-2B-Instruct	Original	-0.000241	49.50	0.17	[0.05, 0.30]	0.077324
	InfoNCE	0.030856	64.60	6.91	[5.86, 7.98]	0.157061
	CMSE†	0.011412	62.82	7.32	[6.41, 8.29]	0.059764
	OPRD†	0.016820	**69.04**	**7.35**	[6.32, 8.34]	0.073238

Interpreting CRR. For CLIP-L, OPRD yields higher CRR than InfoNCE (6.22% versus 5.09%) despite lower DRR and Mean Δ; b_ψ also decreases from 0.091210 to 0.072281. The gain concerns the frequency of edit-induced score decreases exceeding the model-specific paraphrase threshold, rather than larger average score decreases.

Transfer from ranking supervision. With BGE-MLLM-S2, OPRD exceeds Original and InfoNCE in all-edit CRR for all four students without diagnostic-edit training. It also exceeds CMSE for the CLIP and Qwen embedding students. For Instruct, OPRD and CMSE are close under BGE supervision (7.35% versus 7.32%), while CMSE leads under the Qwen teacher (10.22% versus 9.98%).

Categories and operations. CRR is generally higher for object edits than for attribute or predicate edits (Appendix B.3). With BGE-MLLM-S2, OPRD increases ADD CRR but lowers SUBSTITUTE CRR relative to InfoNCE for all four students. With Qwen3-VL-Emb-8B, SUBSTITUTE CRR matches or exceeds InfoNCE across the four students. Operation-level outcomes therefore depend on the teacher. DELETE CRR also decreases under BGE supervision; these responses concern constraint removal rather than target incompatibility (Appendix B.3).

4.3 Retrieval Performance and Cross-Dataset Generalization

CIRR. Under BGE-MLLM-S2 supervision, OPRD improves gallery and subset retrieval over InfoNCE and CMSE for both CLIP students (Table 2; Appendix C.1). For CLIP-L, Recall@1 rises from 27.63% with InfoNCE to 42.72%, and subset Recall@1 from 51.88% to 67.66%, indicating better discrimination within groups of similar images. The Qwen Instruct student also improves

Table 2: Retrieval results (%) on CIRR, PinPoint, and FashionIQ. FashionIQ reports macro averages over dress, shirt, and toptee; category-level results are in Appendix C.3. Bold marks the best result within each group.

		CIRR ↑			PinPoint		FashionIQ Macro avg. ↑	
Model / backbone	Method	R@1	R@10	Subset R@1	mAP @10 ↑	NegRecall @10 ↓	R@10	R@50
Teacher Models								
BGE-MLLM-S2†		**49.41**	**89.28**	**75.10**	12.62	**12.19**	34.17	55.13
Qwen3-VL-Emb-8B‡		47.67	86.73	70.94	**15.39**	14.65	**47.44**	**68.29**
Student Models								
CLIP-L	Original	12.80	51.16	34.51	4.62	20.56	20.20	36.45
	InfoNCE	27.63	77.18	51.88	6.63	19.82	26.87	47.55
	CMSE†	35.90	82.18	61.13	8.08	18.80	29.62	50.40
	OPRD†	**42.72**	**87.42**	**67.66**	**8.97**	**17.47**	**31.58**	**52.35**
CLIP-B	Original	12.60	50.01	32.70	3.99	20.51	16.94	31.54
	InfoNCE	25.07	75.65	48.15	5.32	18.55	20.47	38.90
	CMSE†	33.41	81.46	59.15	6.59	17.84	24.81	45.34
	OPRD†	**40.49**	**86.30**	**65.75**	**7.43**	**16.94**	**26.84**	**48.40**
Qwen3-VL-Emb-2B	Original	37.29	80.82	63.48	11.52	16.04	31.01	51.91
	InfoNCE	**53.55**	90.05	**77.33**	12.61	**12.67**	37.37	58.08
	CMSE†	51.54	89.62	75.48	13.98	13.78	38.83	60.17
	OPRD†	53.26	**90.36**	76.42	**14.02**	14.19	**39.92**	**61.64**
Qwen3-VL-2B-Instruct	Original	0.10	1.36	21.14	0.02	**0.20**	0.37	1.35
	InfoNCE	46.42	87.97	73.26	9.62	12.29	27.28	46.53
	CMSE†	48.36	88.38	73.67	11.43	13.41	27.30	48.94
	OPRD†	**52.09**	**89.60**	**76.51**	**12.55**	14.29	**32.41**	**53.36**

over InfoNCE in gallery and subset recall. For the embedding student, OPRD improves Recall@10, but InfoNCE leads at Recall@1 and Recall@5; gains depend on checkpoint and cutoff.

PinPoint. OPRD improves mAP@10 and lowers NegRecall@10 relative to InfoNCE for both CLIP students. For Qwen, higher mAP@10 accompanies higher NegRecall@10: better positive retrieval need not imply better negative suppression. The noNeg analysis is in Appendix C.2. Original Instruct retrieves few positives or negatives, so its low NegRecall alone does not establish effective suppression.

FashionIQ. With CIRR-only training, OPRD improves macro Recall@10 and Recall@50 over InfoNCE for all four students. Category-level results show gains at both cutoffs across dress, shirt, and toptee (Appendix C.3), demonstrating transfer without FashionIQ training data.

4.4 Ablation Studies and Analysis

We use CLIP-L with frozen BGE-MLLM-S2. Candidate-construction comparisons share initialization, objective, candidate count, temperature, and training budget, with ordinary gallery evaluation. Table 3 summarizes results; Appendix D provides full results, trajectories, and a separate hyperparameter sweep. Ablation CRR uses the ADD/SUBSTITUTE subset (Appendix B.1).

Table 3: Effects of candidate selection and target insertion. Bold indicates the best reported result. All retrieval metrics are percentages.

Candidates	Target insertion	CIRR R@1 ↑	FashionIQ R@10 ↑	PinPoint mAP@10 ↑	PinPoint NegRecall@10 ↓
On-policy†	No	42.07	31.39	8.94	17.55
Random†	Yes	36.31	30.86	8.00	19.27
On-policy†	Yes (OPRD)	**42.72**	**31.58**	**8.97**	**17.47**

Candidate selection. With target insertion and per-epoch refresh in both conditions, student-retrieved candidates outperform random candidates: CIRR Recall@1 rises from 36.31% to 42.72%, and ADD/SUBSTITUTE CRR from 5.87% to 10.30% (Table 22).

Graded teacher supervision. Student-mined InfoNCE and OPRD use the same candidate-mining rule but different supervision. OPRD raises CIRR Recall@1 from 33.92% to 42.72% and ADD/SUBSTITUTE CRR from 6.45% to 10.30% (Table 23). These results support graded teacher supervision within the student-driven mining procedure.

Target insertion. Keeping K fixed, insertion raises CIRR Recall@1 from 42.07% to 42.72%. Improvements occur at each checkpoint and are largest after epoch one (Appendix D.3). Cross-dataset effects are smaller and mixed: FashionIQ Recall@50 is slightly higher without insertion.

Teacher choice. BGE-MLLM-S2 yields higher CIRR Recall@1 and lower PinPoint NegRecall@10 for every student; Qwen3-VL-Emb-8B yields higher all-edit CRR for every student and higher FashionIQ Recall@10 for both Qwen students. The preferred teacher for PinPoint mAP@10 depends on the student (Appendix D.4). Teacher choice affects these outcomes differently.

Candidate count and training cost. In the dedicated CLIP-L/BGE-MLLM-S2 sweep (Appendix D.5), reducing K from 32 to 8 at $T = 0.05$ lowers the reported training time from 2.75 to 1.06 hours, while CIRR Recall@1 changes from 42.62% to 43.77%. These results show that smaller candidate sets can retain competitive retrieval performance while reducing training cost in this setting.

5 Conclusion

We introduced CIR-QDiag to measure how local instruction edits affect target similarity scores, using DRR for response direction and CRR for decreases exceeding a model-specific paraphrase threshold. OPRD distills a frozen teacher's soft relevance distribution over periodically refreshed, student-retrieved candidates. Training on original CIR pairs yields retrieval gains and higher aggregate CRR across multiple teacher–student configurations. Candidate-selection and supervision controls support the use of student-retrieved images and graded teacher relevance distributions. Operation-level responses vary with the teacher and edit type, highlighting the complementary information provided by retrieval metrics and calibrated edit diagnostics.

Limitations

Diagnostic coverage and interpretation. CIR-QDiag covers single-condition edits derived from PinPoint; interactions among conditions and broader contextual requirements remain outside its scope. The same model is used for proposal generation and automatic validation, so correlated errors may remain after human review. The diagnostic measures target similarity changes rather than gallery-level rank changes: a lower target score need not imply a lower rank if competing scores also change. Extending coverage to additional domains and multi-condition edits would broaden the diagnostic.

Dependence on teachers and candidate sets. OPRD transfers a teacher's relevance distribution over a finite candidate set. Teacher errors may propagate to the student, and candidates outside the selected set receive no direct supervision at that step. The observed variation across teachers, edit operations, and retrieval metrics motivates further study of teacher reliability and candidate diversity.

Training scalability. Inference uses only the student, but training requires teacher embedding caches, periodic student encoding of queries and gallery images, and retrieval to refresh candidates. These steps add computation and storage overhead. The experiments do not systematically evaluate larger or continually evolving galleries. Incremental indexing and adaptive candidate refresh could reduce these costs.

AI USE STATEMENT

Generative AI models were used in CIR-QDiag construction to generate local instruction edits and perform automated visual compatibility checking, as documented in Section 3.2 and Appendix E. The language of the manuscript was refined and grammatical errors were corrected with the assistance of generative AI tools, without altering the meaning of the original sentences. The authors take responsibility for the final content of this work.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/5be69a584901a26c521c2b51e40a4c20-Paper-Conference.pdf.

Matteo Attimonelli, Alessandro De Bellis, Aryo Pradipta Gema, Rohit Saxena, Monica Sekoyan, Wai-Chung Kwan, Claudio Pomo, Alessandro Suglia, Dietmar Jannach, Tommaso Di Noia, and Pasquale Minervini. Do composed image retrieval benchmarks require multimodal composition?, 2026. URL https://arxiv.org/abs/2605.14787.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21466–21474, June 2022.

Enjun Du, Siyi Liu, Zirong Chen, Xinyu Zuo, Jinwen Luo, Ruiwen Tao, Lisheng Duan, Haijin Liang, Jin Ma, Junfu Pu, and Yongqi Zhang. Evirank: Structured relevance evidence for multimodal image re-ranking, 2026. URL https://arxiv.org/abs/2608.20886.

Kirill Dubovikov, Martin Takac, and Salem Lahlou. Score-only distillation for compact dense retrieval, 2026. URL https://arxiv.org/abs/2607.11465.

Hongfei Ge, Yuanchun Jiang, Jianshan Sun, Kun Yuan, and Yezheng Liu. Llm-enhanced composed image retrieval: An intent uncertainty-aware linguistic-visual dual channel matching model. *ACM Trans. Inf. Syst.*, 43(2), January 2025. ISSN 1046-8188. doi: 10.1145/3699715. URL https://doi.org/10.1145/3699715.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 31096–31116. Curran Associates, Inc., 2023. doi: 10.52202/075280-1355. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/63461de0b4cb760fc498e85b18a7fe81-Paper-Datasets_and_Benchmarks.pdf.

Chuong Huynh, Jinyu Yang, Ashish Tawari, Mubarak Shah, Son Tran, Raffay Hamid, Trishul Chilimbi, and Abhinav Shrivastava. Collm: A large language model for composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3994–4004, June 2025.

Youngjoon Jang, Seongtae Hong, Hyeonseok Moon, and Heuiseok Lim. Beyond hard negatives: The importance of score distribution in knowledge distillation. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '26, pp. 3811–3816, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400725999. doi: 10.1145/3805712.3809843. URL https://doi.org/10.1145/3805712.3809843.

Jaehyun Kwak, Ramahdani Muhammad Izaaz Inhar, Se-Young Yun, and Sung-Ju Lee. QuRe: Query-relevant retrieval through hard negative sampling in composed image retrieval. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 32063–32077. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/kwak25a.html.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking, 2026a. URL https://arxiv.org/abs/2601.04720.

Zixu Li, Zhiheng Fu, Yupeng Hu, Zhiwei Chen, Haokun Wen, and Liqiang Nie. Finecir: Explicit parsing of fine-grained modification semantics for composed image retrieval, 2025. URL https://arxiv.org/abs/2503.21309.

Zixu Li, Yupeng Hu, Zhiheng Fu, Zhiwei Chen, Yongqi Li, and Liqiang Nie. TEMA: Anchor the image, follow the text for multi-modification composed image retrieval. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24421–24442, San Diego, California, United States, July 2026b. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1121. URL https://aclanthology.org/2026.acl-long.1121/.

Yikun Liu, Jiangchao Yao, Weidi Xie, and Yanfeng Wang. A Sanity Check on Composed Image Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 11292–11301, June 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 (ICCV)*, pp. 2125–2134, October 2021.

Rohan Mahadev, Joyce Yuan, Patrick Poirson, David Xue, Hao-Yu Wu, and Dmitry Kislyuk. Pinpoint: Evaluation of composed image retrieval with explicit negatives, multi-image queries, and paraphrase testing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9742–9751, June 2026.

Bill Psomas, George Retsinas, Nikos Efthymiadis, Panagiotis Filntisis, Yannis Avrithis, Petros Maragos, Ondrej Chum, and Giorgos Tolias. Instance-level composed image retrieval. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 68819–68864. Curran Associates, Inc., 2025. doi: 10.52202/085713-2315. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/6391a02f02dc529cfa3f242fb4343c6a-Paper-Conference.pdf.

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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine*

Learning, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Sashank Reddi, Rama Kumar Pasumarthi, Aditya Menon, Ankit Singh Rawat, Felix Yu, Seungyeon Kim, Andreas Veit, and Sanjiv Kumar. Rankdistil: Knowledge distillation for ranking. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2368–2376. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/reddi21a.html.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. RocketQAv2: A joint training method for dense passage retrieval and passage reranking. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2825–2835, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.224. URL https://aclanthology.org/2021.emnlp-main.224/.

Tingyu Song, Yanzhao Zhang, Mingxin Li, Zhuoning Guo, Dingkun Long, Pengjun Xie, Siyue Zhang, Yilun Zhao, and Shu Wu. Rethinking composed image retrieval evaluation: A fine-grained benchmark from image editing. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 46224–46242, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.2144. URL https://aclanthology.org/2026.acl-long.2144/.

Shitong Sun, Jindong Gu, and Shaogang Gong. Benchmarking robustness of text-image composed retrieval. In *R0-FoMo:Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*, 2023. URL https://openreview.net/forum?id=YMutYSbvVe.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL http://arxiv.org/abs/1807.03748.

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

Yifan Wang, Wuliang Huang, Lei Li, and Chun Yuan. Semantic distillation from neighborhood for composed image retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 5575–5583, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3681493. URL https://doi.org/10.1145/3664647.3681493.

Haokun Wen, Xian Zhang, Xuemeng Song, Yinwei Wei, and Liqiang Nie. Target-guided composed image retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, pp. 915–923, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701085. doi: 10.1145/3581783.3611817. URL https://doi.org/10.1145/3581783.3611817.

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 (CVPR)*, pp. 11307–11317, June 2021.

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 *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KRLUvxh8uaX.

Hang Zhang, Yeyun Gong, Yelong Shen, Jiancheng Lv, Nan Duan, and Weizhu Chen. Adversarial retriever-ranker for dense text retrieval. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=MR7XubKUFB.

Wenliang Zhong, Rob Barton, Weizhi An, Feng Jiang, Hehuan Ma, Yuzhi Guo, Abhishek Dan, Shioulin Sam, Karim Bouyarmane, and Junzhou Huang. Zero-shot composed image retrieval via dual-stream instruction-aware distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22221–22231, October 2025.

Junjie Zhou, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, Defu Lian, and Yongping Xiong. Megapairs: Massive data synthesis for universal multimodal retrieval, 2024. URL https://arxiv.org/abs/2412.14475.

A Implementation and Evaluation Details

This appendix specifies the retrieval interfaces, training settings, and benchmark protocols. CIR-QDiag construction and evaluation are detailed in Appendix B; complete retrieval results and OPRD design analyses appear in Appendices C and D, respectively.

A.1 Models and Retrieval Interfaces

The students are CLIP-L, CLIP-B, Qwen3-VL-Emb-2B, and Qwen3-VL-2B-Instruct. We abbreviate Qwen3-VL-Embedding-2B and Qwen3-VL-Embedding-8B as Qwen3-VL-Emb-2B and Qwen3-VL-Emb-8B, respectively.[2] The instruction-tuned checkpoint is `Qwen/Qwen3-VL-2B-Instruct`[3] and is evaluated separately from the embedding checkpoint. Original denotes the checkpoint before CIR adaptation. Query and gallery embeddings are unit-normalized and scored by the inner product in Eq. 1. Each teacher and student computes similarity scores in its own embedding space. CMSE compares these similarity scores, whereas OPRD compares the soft relevance distributions formed from them; both operate on shared candidate identities rather than matching embedding coordinates.

For CLIP students, both encoders and the residual composition head in Eq. 8 are updated. The head's output layer is initialized to zero, so adaptation begins from normalized image–text addition. Both Qwen students are adapted using LoRA. BGE-MLLM-S2 and Qwen3-VL-Emb-8B are frozen during distillation and are also evaluated as standalone retrievers. Their superscripts, † and ‡, identify the teacher used by each distilled configuration.

A.2 Training Configuration and Baseline Implementations

Table 4 summarizes the shared training settings. All student training runs, including the main comparisons and ablation studies, use four epochs on the CIRR training split. At the start of each epoch, OPRD retrieves the student's top-$(K-1)$ non-reference, non-target images and inserts the annotated target. Candidate IDs remain fixed within the epoch, while student similarity scores are recomputed at each optimization step. Target insertion and reference exclusion follow Eq. 5.

Table 4: Shared training settings. Candidate and temperature settings apply to OPRD; LoRA settings apply to the two Qwen students.

Setting	Value
Training data	Original CIRR training split
Training duration	4 epochs
Candidate count	$K = 32$
Distillation temperature	$T = 0.05$
Candidate refresh	Once at the start of each epoch
Teacher updates	Frozen; query and gallery embeddings cached
Qwen LoRA rank / alpha	32 / 32
Hardware	One NVIDIA H100 GPU for training and inference

Baseline definitions. The main-table InfoNCE baselines use contrastive supervision with one-hot target labels. CMSE adapts the centered mean squared error objective of Dubovikov et al. (2026) by matching teacher and student similarity scores after removing the per-query mean score residual. The student-mined InfoNCE control in Appendix D.2 uses student-retrieved candidates refreshed at the start of each epoch and is distinct from the main-table InfoNCE baseline. Random candidates refer only to the reverse-KL candidate-source control in Appendix D.1.

Controlled ablation settings. The candidate-construction controls share the Original CLIP-L (ViT-L/14) initialization and a frozen BGE-MLLM-S2 teacher. They use seed 42 and batch size

[2] Official embedding checkpoints: `https://huggingface.co/Qwen/Qwen3-VL-Embedding-2B` and `https://huggingface.co/Qwen/Qwen3-VL-Embedding-8B`.

[3] `https://huggingface.co/Qwen/Qwen3-VL-2B-Instruct`

32. Their final results use the epoch-four checkpoint on every dataset. For the target-insertion comparison, the composition-head learning rate is 10^{-5} and the backbone learning rate is 5×10^{-7}, with bfloat16 (BF16) precision, 5% learning-rate warmup, and cosine learning-rate decay. These optimizer settings describe that controlled comparison rather than specifying a common configuration for every student.

Training and inference overhead. Inference uses only the trained student and its gallery index. Training additionally requires frozen-teacher embedding caches and student encoding and retrieval to refresh candidate sets. These operations add training cost but do not introduce teacher inference into the retrieval pipeline.

A.3 Data Splits and Retrieval Protocols

All task adaptation uses the original CIRR training split, containing 28,225 queries and 16,939 gallery images. CIRR validation has 4,181 queries, PinPoint has 7,635 evaluation queries, and FashionIQ validation has 6,016 queries. FashionIQ averages dress, shirt, and toptee equally. CIR-QDiag is reserved for evaluation and is not used to select checkpoints, temperatures, or candidate strategies; its construction, subsets, calibration, and aggregation are specified in Appendix B.

CIRR reports gallery and subset recall. Its additional summary is $\mathrm{Avg} = (\mathrm{R@5} + \mathrm{R}_{\mathrm{subset}}\mathrm{@1})/2$. PinPoint reports positive retrieval and retrieval of annotated negatives. Its noNeg evaluation removes these negatives, with the resulting difference defined as $\Delta\mathrm{mAP@10} = \mathrm{mAP@10}_{\mathrm{noNeg}} - \mathrm{mAP@10}$. Precision, recall, mAP, and relative noNeg changes are percentages, and absolute mAP differences are percentage points. Linguistic-sensitivity statistics retain their original scale.

B CIR-QDiag: Construction, Protocol, and Full Results

B.1 Edit Construction and Dataset Statistics

CIR-QDiag follows the generation, model-based visual validation, and human-review procedure in Sec. 3.2. Paraphrase groups are recovered using the same reference images and complete positive and negative annotation sets after removing empty and duplicate IDs. Groups contain two to six original instructions. Proposed diagnostic edits change one contiguous text span. Model-based visual validation checks the edited condition against the reference and annotated positive images.

B.1.1 Human Review Protocol

Following model-based visual validation, five human reviewers checked the model-accepted proposals. They first used a shared reviewer-training set to align their interpretation of the criteria. The remaining proposals were divided among them, with one reviewer assigning a binary ACCEPT/REJECT judgment to each proposal. Human reviewers accepted 95% of the model-accepted proposals. This acceptance rate is not an estimate of inter-annotator agreement.

The human reviewers are provided with the same inputs and review instructions as the model-based visual validation stage: the source and edited instructions, the edit operation and semantic category, the proposed before/after spans, the reference image(s), and the annotated positive images. Accordingly, reviewers check the edited span, semantic category, fluency, preservation of unedited content, and the operation-specific visual criteria. For ADD and SUBSTITUTE, the introduced condition must be visually decidable and false for the supplied positive images. For DELETE, the edit must remove one meaningful constraint while preserving the remaining instruction content and yielding a fluent, strictly weaker query. Proposals that do not satisfy these criteria are rejected and excluded from CIR-QDiag. Official negative images are not used during this review.

The validated diagnostic contains 930 groups, 3,308 original instructions, and 10,131 edits. Table 5 gives the category-by-operation counts. The all-edit diagnostic includes all accepted edits. The ADD/SUBSTITUTE subset excludes DELETE, and its semantic subsets further restrict the edits by category. Compact tables abbreviate ADD/SUBSTITUTE as A/S. All subsets use the paired target similarity score responses and group-balanced aggregation in Sec. 3.2.3.

Table 5: CIR-QDiag edit counts by semantic category and operation.

Semantic category	ADD	DELETE	SUBSTITUTE	Total
Object	1,616	1,090	1,444	4,150
Attribute	693	2,497	1,052	4,242
Predicate	452	809	478	1,739
Total	**2,761**	**4,396**	**2,974**	**10,131**

The ADD/SUBSTITUTE subset contains 5,735 edits: 3,060 object, 1,745 attribute, and 930 predicate edits. DELETE contains 4,396 edits.

B.2 Paraphrase-Based Calibration and Response Rates

Hierarchical aggregation. For a diagnostic subset S, let $\mathcal{H}_S$, $\mathcal{I}_g^S$, and $\mathcal{E}_{g,i}^S$ contain its represented groups, original instructions, and edits. We average targets within each edit, edits within each original instruction, instructions within each group, and finally groups:

$$M_S(h) = \frac{1}{|\mathcal{H}_S|} \sum_{g \in \mathcal{H}_S} \frac{1}{|\mathcal{I}_g^S|} \sum_{i \in \mathcal{I}_g^S} \frac{1}{|\mathcal{E}_{g,i}^S|} \sum_{n \in \mathcal{E}_{g,i}^S} \frac{1}{|\mathcal{P}_g|} \sum_{t \in \mathcal{P}_g} h_{g,i,n,t}. \tag{9}$$

The response rates, expressed as percentages, are

$$\mathrm{DRR}_S^{\psi} = 100\, M_S(I_{\mathrm{dir}}^{\psi}), \qquad \mathrm{CRR}_S^{\psi} = 100\, M_S(I_{\mathrm{cal}}^{\psi}). \tag{10}$$

Mean Δ is $M_S(\Delta^{\psi})$. This hierarchy prevents groups with more instructions, edits, or targets from dominating the diagnostic.

The paraphrase threshold b_{ψ} is estimated separately for each model as the group-balanced 95th percentile of absolute target similarity score differences between meaning-preserving paraphrases, with equal total weight assigned to each calibration group. We construct 95% confidence intervals using 2,000 group-bootstrap replicates, preserving each group's internal structure. The interval bounds are the 2.5th and 97.5th percentiles of the bootstrap estimates. These intervals quantify evaluation uncertainty for each trained checkpoint, not variation across training seeds. Mean Δ, thresholds, and control gaps retain the similarity scale; response rates and their intervals are percentages.

Table 6 compares responses across all edit operations for the student checkpoints and teachers. Table 7 provides bootstrap confidence intervals for the paraphrase threshold, and Table 8 compares results on the ADD/SUBSTITUTE subset. Method orderings can differ between DRR and CRR, so response magnitude should be interpreted together with the model-specific paraphrase threshold.

Table 6: CIR-QDiag results across all edit operations. Bold marks the highest response rate within each teacher or student group. Response rates and intervals are percentages.

Model / method	Mean Δ	DRR ↑	CRR ↑	CRR bootstrap CI	b_{ψ}
Teacher Models					
BGE-MLLM-S2†	0.017767	**70.08**	7.94	[6.82, 9.14]	0.071220
Qwen3-VL-Emb-8B‡	0.032877	69.56	**10.21**	[9.04, 11.39]	0.112892
CLIP-L					
Original	0.005471	64.80	2.00	[1.52, 2.52]	0.042547
InfoNCE	0.017783	68.00	5.09	[4.24, 5.97]	0.091210
CMSE†	0.010219	68.54	4.62	[3.85, 5.46]	0.053417
CMSE‡	0.015716	66.75	5.94	[5.04, 6.89]	0.075822
OPRD†	0.014627	67.36	6.22	[5.30, 7.16]	0.072281
OPRD‡	0.017770	**68.78**	**6.51**	[5.57, 7.47]	0.081524

Continued on the next page

Table 6 continued

Model / method	Mean Δ	DRR ↑	CRR ↑	CRR bootstrap CI	b_ψ
CLIP-B					
Original	0.004775	65.26	2.39	[1.89, 2.99]	0.036016
InfoNCE	0.017640	67.48	5.15	[4.33, 6.04]	0.092273
CMSE†	0.009539	**67.76**	5.17	[4.30, 6.13]	0.049115
CMSE‡	0.014245	66.85	5.83	[4.94, 6.79]	0.071360
OPRD†	0.013778	64.98	6.10	[5.28, 7.00]	0.071699
OPRD‡	0.016911	66.73	**6.38**	[5.49, 7.33]	0.081584
Qwen3-VL-Emb-2B					
Original	0.026134	68.38	8.69	[7.58, 9.83]	0.103842
InfoNCE	0.036515	66.94	8.68	[7.57, 9.79]	0.148849
CMSE†	0.013464	59.10	5.97	[5.22, 6.77]	0.085236
CMSE‡	0.024701	65.96	10.86	[9.61, 12.10]	0.090963
OPRD†	0.019574	62.09	10.24	[9.14, 11.32]	0.084918
OPRD‡	0.032127	**68.83**	**11.88**	[10.62, 13.18]	0.105891
Qwen3-VL-2B-Instruct					
Original	-0.000241	49.50	0.17	[0.05, 0.30]	0.077324
InfoNCE	0.030856	64.60	6.91	[5.86, 7.98]	0.157061
CMSE†	0.011412	62.82	7.32	[6.41, 8.29]	0.059764
CMSE‡	0.021950	65.10	**10.22**	[8.99, 11.46]	0.084456
OPRD†	0.016820	**69.04**	7.35	[6.32, 8.34]	0.073238
OPRD‡	0.025500	65.40	9.98	[8.81, 11.17]	0.098788

Table 7: Bootstrap confidence intervals for all-edit DRR and the paraphrase threshold for Qwen configurations. DRR intervals are percentages; thresholds remain on the similarity scale.

Model / method	DRR CI	b_ψ	b_ψ bootstrap CI
Qwen3-VL-Emb-2B			
InfoNCE	[65.47, 68.48]	0.148849	[0.132754, 0.174525]
CMSE†	[57.39, 60.94]	0.085236	[0.080170, 0.089724]
CMSE‡	[64.39, 67.61]	0.090963	[0.084069, 0.099373]
OPRD†	[60.35, 63.86]	0.084918	[0.078505, 0.091386]
OPRD‡	[67.29, 70.42]	0.105891	[0.096751, 0.115574]
Qwen3-VL-2B-Instruct			
Original	[47.83, 51.19]	0.077324	[0.070726, 0.083208]
InfoNCE	[63.06, 66.17]	0.157061	[0.136622, 0.182705]
CMSE†	[61.18, 64.49]	0.059764	[0.054810, 0.065408]
CMSE‡	[63.48, 66.70]	0.084456	[0.077280, 0.092683]
OPRD†	[67.47, 70.53]	0.073238	[0.068839, 0.078170]
OPRD‡	[63.78, 67.02]	0.098788	[0.090099, 0.107870]

Table 8: CIR-QDiag responses on the ADD/SUBSTITUTE subset. Bold marks the highest response rate within each student checkpoint group. Response rates and intervals are percentages.

Model / method	Mean Δ	DRR	DRR CI	CRR	CRR CI
Qwen3-VL-Emb-2B					
InfoNCE	0.052425	77.54	[75.84, 79.24]	11.03	[9.57, 12.58]
CMSE†	0.035258	84.07	[82.51, 85.68]	10.62	[9.19, 12.06]
CMSE‡	0.041340	82.53	[81.05, 83.99]	15.16	[13.39, 16.99]
OPRD†	0.043424	**85.06**	[83.55, 86.53]	**18.34**	[16.48, 20.20]
OPRD‡	0.049485	81.64	[80.08, 83.09]	16.77	[14.95, 18.68]
Qwen3-VL-2B-Instruct					
Original	-0.000132	47.55	[45.19, 49.94]	0.20	[0.05, 0.39]
InfoNCE	0.039984	70.74	[68.75, 72.85]	7.95	[6.67, 9.33]
CMSE†	0.022042	77.81	[75.99, 79.58]	11.11	[9.62, 12.71]
CMSE‡	0.033876	77.63	[75.85, 79.36]	13.64	[11.94, 15.42]
OPRD†	0.022461	74.20	[72.36, 76.05]	8.95	[7.67, 10.33]
OPRD‡	0.039976	**78.38**	[76.73, 80.06]	**13.78**	[12.17, 15.50]

B.3 Semantic Categories and Edit Operations

Table 9 organizes CRR by semantic category and operation; Table 11 reports Mean Δ, DRR, and CRR for the Qwen configurations, with bootstrap confidence intervals for the two response rates. Each subset is balanced over its represented groups, so the displayed rates cannot be arithmetically averaged to reconstruct the all-edit CRR.

With BGE-MLLM-S2, OPRD increases ADD CRR and lowers SUBSTITUTE CRR relative to InfoNCE for all four students. The Qwen teacher produces a different profile: SUBSTITUTE CRR matches or exceeds InfoNCE for every student. For Qwen3-VL-Emb-2B, SUBSTITUTE CRR is 13.86% with InfoNCE, 8.06% with BGE-supervised OPRD, and 19.00% with Qwen-supervised OPRD.

DELETE relaxes an instruction and need not invalidate the target. Under BGE supervision, OPRD has lower DELETE CRR than InfoNCE for all four students (Table 10); these rates describe score decreases following constraint removal rather than rejection of incompatible targets.

Table 9: CIR-QDiag semantic and operation breakdown. Values are CRR percentages; bold marks the highest rate in each group.

Model / method	Object	Attribute	Predicate	ADD	SUBSTITUTE	DELETE
Teacher Models						
BGE-MLLM-S2†	7.34	**6.56**	4.52	5.38	11.85	6.34
Qwen3-VL-Emb-8B‡	**12.46**	6.29	**4.94**	**9.26**	**16.96**	**6.50**
CLIP-L						
Original	2.99	0.68	0.73	0.13	3.81	1.99
InfoNCE	6.55	2.62	1.49	1.59	10.26	**3.48**
CMSE†	4.66	3.84	1.88	5.34	6.90	1.86
CMSE‡	6.98	3.78	1.64	5.29	**10.85**	1.91
OPRD†	**8.28**	3.63	**3.11**	**12.13**	7.38	0.81
OPRD‡	7.56	**4.33**	2.19	7.64	10.26	1.80

Continued on the next page

Table 9 continued

Model / method	Object	Attribute	Predicate	ADD	SUBSTITUTE	DELETE
CLIP-B						
Original	3.61	1.03	0.99	0.29	4.19	2.75
InfoNCE	6.96	2.36	1.33	2.13	9.40	**3.83**
CMSE†	4.44	**4.70**	2.27	3.92	7.66	3.16
CMSE‡	6.11	4.53	1.62	3.69	**10.44**	3.00
OPRD†	**9.58**	2.77	**2.59**	**12.26**	6.68	0.83
OPRD‡	7.89	3.87	2.51	8.15	9.96	1.93
Qwen3-VL-Emb-2B						
Original	10.85	5.40	4.32	8.01	12.41	**6.26**
InfoNCE	10.70	5.30	4.17	7.48	13.86	5.66
CMSE†	8.71	3.82	2.40	14.78	6.46	0.49
CMSE‡	14.55	6.67	3.87	12.27	17.50	5.34
OPRD†	**17.44**	4.50	4.25	**28.18**	8.06	0.95
OPRD‡	16.26	**7.07**	**4.41**	13.76	**19.00**	6.17
Qwen3-VL-2B-Instruct						
Original	0.13	0.06	0.65	0.15	0.15	0.06
InfoNCE	9.19	3.45	2.87	5.51	10.00	5.49
CMSE†	8.52	5.46	3.35	11.85	8.99	2.42
CMSE‡	12.44	**6.57**	4.32	9.38	**16.56**	**5.98**
OPRD†	7.29	5.98	4.37	7.51	9.21	5.30
OPRD‡	**12.84**	6.18	**4.45**	**12.21**	14.35	5.55

Table 10: CIR-QDiag responses to constraint removal (DELETE). Bold marks the largest response rate in each group; a larger rate does not imply greater semantic correctness.

Model / method	Mean Δ	DRR	CRR
Teacher Models			
BGE-MLLM-S2†	0.013403	**66.21**	6.34
Qwen3-VL-Emb-8B‡	0.015940	56.47	**6.50**
CLIP-L			
Original	0.004506	**62.57**	1.99
InfoNCE	0.009668	59.13	**3.48**
CMSE†	0.002552	57.08	1.86
CMSE‡	0.002233	51.96	1.91
OPRD†	0.000847	54.87	0.81
OPRD‡	0.004938	58.90	1.80
CLIP-B			
Original	0.004480	**63.54**	2.75
InfoNCE	0.009577	58.27	**3.83**
CMSE†	0.003448	56.44	3.16
CMSE‡	0.004408	53.73	3.00
OPRD†	-0.000673	49.09	0.83
OPRD‡	0.003208	54.70	1.93
Qwen3-VL-Emb-2B			
Original	0.015379	**59.41**	**6.26**
InfoNCE	0.019123	55.94	5.66
CMSE†	-0.010523	32.44	0.49
CMSE‡	0.006146	48.89	5.34
OPRD†	-0.006562	38.07	0.95
OPRD‡	0.013146	56.04	6.17

Continued on the next page

Table 10 continued

Model / method	Mean Δ	DRR	CRR
Qwen3-VL-2B-Instruct			
Original	-0.000617	52.85	0.06
InfoNCE	0.021004	58.17	5.49
CMSE†	-0.000969	46.10	2.42
CMSE‡	0.008730	52.90	**5.98**
OPRD†	0.010583	**63.98**	5.30
OPRD‡	0.009522	52.16	5.55

Table 11: Detailed CIR-QDiag responses organized by subset and student checkpoint. Rows labeled ADD/ SUBSTITUTE restrict the corresponding semantic category to those two operations. Response rates and bootstrap intervals are percentages; Mean Δ retains the similarity scale. Bold compares the listed methods within each subset and student checkpoint group.

Method	Mean Δ	DRR	DRR CI	CRR	CRR CI
Object					
Qwen3-VL-Emb-2B					
InfoNCE	0.049503	73.86	[71.92, 75.89]	10.70	[9.16, 12.32]
CMSE†	0.029209	75.01	[72.78, 77.17]	8.71	[7.38, 10.12]
CMSE‡	0.037988	77.51	[75.56, 79.64]	14.55	[12.65, 16.57]
OPRD†	0.038801	**77.90**	[75.88, 80.01]	**17.44**	[15.54, 19.51]
OPRD‡	0.046560	76.76	[74.87, 78.81]	16.26	[14.28, 18.35]
Qwen3-VL-2B-Instruct					
Original	-0.001052	46.35	[43.61, 49.00]	0.13	[0.00, 0.35]
InfoNCE	0.040387	68.23	[66.16, 70.35]	9.19	[7.69, 10.77]
CMSE†	0.017385	72.14	[70.00, 74.44]	8.52	[7.18, 9.93]
CMSE‡	0.030122	72.77	[70.68, 74.93]	12.44	[10.71, 14.28]
OPRD†	0.019990	**74.36**	[72.38, 76.37]	7.29	[6.08, 8.57]
OPRD‡	0.036032	73.16	[71.02, 75.34]	**12.84**	[11.10, 14.64]
Attribute					
Qwen3-VL-Emb-2B					
InfoNCE	0.020927	60.20	[57.92, 62.38]	5.30	[4.11, 6.52]
CMSE†	0.002258	47.58	[45.08, 50.29]	3.82	[2.89, 4.83]
CMSE‡	0.012275	56.89	[54.53, 59.34]	6.67	[5.33, 8.01]
OPRD†	0.004944	49.52	[47.07, 52.12]	4.50	[3.53, 5.52]
OPRD‡	0.018058	**61.66**	[59.36, 63.96]	**7.07**	[5.78, 8.40]
Qwen3-VL-2B-Instruct					
Original	0.000313	52.50	[50.00, 54.97]	0.06	[0.00, 0.15]
InfoNCE	0.018407	60.85	[58.69, 62.98]	3.45	[2.42, 4.52]
CMSE†	0.005790	54.84	[52.43, 57.25]	5.46	[4.25, 6.74]
CMSE‡	0.012409	58.78	[56.48, 61.02]	**6.57**	[5.16, 8.00]
OPRD†	0.012541	**64.52**	[62.22, 66.73]	5.98	[4.66, 7.44]
OPRD‡	0.014477	58.62	[56.31, 60.84]	6.18	[4.86, 7.60]

Continued on the next page

Table 11 continued

Method	Mean Δ	DRR	DRR CI	CRR	CRR CI
Predicate					
Qwen3-VL-Emb-2B					
InfoNCE	0.017875	59.06	[55.88, 62.31]	4.17	[2.91, 5.61]
CMSE†	-0.000852	45.06	[41.27, 49.03]	2.40	[1.38, 3.57]
CMSE‡	0.008007	53.72	[50.08, 57.42]	3.87	[2.63, 5.21]
OPRD†	0.004066	50.03	[46.33, 53.91]	4.25	[3.02, 5.67]
OPRD‡	0.014469	**60.87**	[57.64, 64.15]	**4.41**	[3.08, 5.83]
Qwen3-VL-2B-Instruct					
Original	0.000598	51.24	[47.17, 55.03]	0.65	[0.04, 1.48]
InfoNCE	0.014003	58.22	[54.92, 61.47]	2.87	[1.84, 4.05]
CMSE†	0.004334	56.04	[52.35, 59.68]	3.35	[2.19, 4.55]
CMSE‡	0.010490	58.67	[55.33, 62.14]	4.32	[2.96, 5.85]
OPRD†	0.010333	**63.67**	[60.16, 67.00]	4.37	[3.08, 5.82]
OPRD‡	0.011840	57.31	[53.87, 60.83]	**4.45**	[3.07, 6.03]
ADD					
Qwen3-VL-Emb-2B					
InfoNCE	0.045880	78.61	[76.65, 80.55]	7.48	[6.01, 9.08]
CMSE†	0.047667	93.73	[92.48, 94.97]	14.78	[12.72, 17.03]
CMSE‡	0.040115	86.30	[84.60, 87.99]	12.27	[10.29, 14.40]
OPRD†	0.061995	**94.56**	[93.42, 95.63]	**28.18**	[25.51, 31.04]
OPRD‡	0.047750	84.80	[83.16, 86.53]	13.76	[11.67, 15.94]
Qwen3-VL-2B-Instruct					
Original	-0.001835	42.32	[39.19, 45.49]	0.15	[0.00, 0.40]
InfoNCE	0.033316	69.45	[67.06, 71.89]	5.51	[4.24, 6.95]
CMSE†	0.024620	**80.84**	[78.77, 82.82]	11.85	[10.04, 13.82]
CMSE‡	0.028584	77.68	[75.39, 79.85]	9.38	[7.68, 11.13]
OPRD†	0.018918	70.13	[67.67, 72.46]	7.51	[6.00, 9.10]
OPRD‡	0.039750	80.62	[78.49, 82.62]	**12.21**	[10.32, 14.25]
SUBSTITUTE					
Qwen3-VL-Emb-2B					
InfoNCE	0.057425	76.60	[74.35, 78.87]	13.86	[11.83, 16.15]
CMSE†	0.022502	73.71	[70.99, 76.40]	6.46	[5.07, 7.87]
CMSE‡	0.041285	78.39	[76.07, 80.57]	17.50	[15.05, 19.95]
OPRD†	0.025235	75.12	[72.56, 77.73]	8.06	[6.47, 9.66]
OPRD‡	0.050111	**78.92**	[76.67, 81.01]	**19.00**	[16.42, 21.50]
Qwen3-VL-2B-Instruct					
Original	0.000662	49.96	[47.04, 52.92]	0.15	[0.01, 0.33]
InfoNCE	0.045059	72.35	[69.85, 74.87]	10.00	[8.09, 12.20]
CMSE†	0.017857	73.32	[70.70, 76.09]	8.99	[7.34, 10.79]
CMSE‡	0.037088	**77.41**	[75.03, 79.76]	**16.56**	[14.24, 19.12]
OPRD†	0.024223	76.75	[74.29, 79.18]	9.21	[7.58, 10.93]
OPRD‡	0.039170	76.32	[73.96, 78.68]	14.35	[12.23, 16.62]

Continued on the next page

Table 11 continued

Method	Mean Δ	DRR	DRR CI	CRR	CRR CI
DELETE					
Qwen3-VL-Emb-2B					
InfoNCE	0.019123	55.94	[53.90, 57.98]	5.66	[4.45, 7.00]
CMSE†	-0.010523	32.44	[30.33, 34.57]	0.49	[0.25, 0.77]
CMSE‡	0.006146	48.89	[46.63, 51.21]	5.34	[4.14, 6.69]
OPRD†	-0.006562	38.07	[36.01, 40.13]	0.95	[0.51, 1.50]
OPRD‡	0.013146	**56.04**	[53.82, 58.25]	**6.17**	[4.90, 7.56]
Qwen3-VL-2B-Instruct					
Original	-0.000617	52.85	[50.37, 55.35]	0.06	[0.00, 0.15]
InfoNCE	0.021004	58.17	[56.13, 60.17]	5.49	[4.32, 6.86]
CMSE†	-0.000969	46.10	[44.00, 48.41]	2.42	[1.69, 3.25]
CMSE‡	0.008730	52.90	[50.58, 55.18]	**5.98**	[4.78, 7.35]
OPRD†	0.010583	**63.98**	[61.74, 66.28]	5.30	[4.16, 6.54]
OPRD‡	0.009522	52.16	[49.87, 54.43]	5.55	[4.36, 6.84]
Object (ADD/SUBSTITUTE)					
Qwen3-VL-Emb-2B					
InfoNCE	0.056498	79.58	[77.59, 81.66]	11.32	[9.59, 13.13]
CMSE†	0.041274	89.66	[88.01, 91.34]	11.72	[9.84, 13.54]
CMSE‡	0.046024	86.26	[84.58, 87.93]	16.50	[14.31, 18.80]
OPRD†	0.052863	**90.65**	[89.08, 92.21]	**23.18**	[20.77, 25.70]
OPRD‡	0.055535	84.36	[82.49, 86.22]	18.80	[16.43, 21.18]
Qwen3-VL-2B-Instruct					
Original	-0.000783	45.03	[42.06, 48.13]	0.16	[0.00, 0.40]
InfoNCE	0.043465	71.37	[69.07, 73.77]	9.15	[7.53, 10.89]
CMSE†	0.022685	80.29	[78.06, 82.43]	10.54	[8.80, 12.33]
CMSE‡	0.035720	79.29	[77.10, 81.53]	14.09	[12.09, 16.21]
OPRD†	0.021968	74.98	[72.68, 77.29]	7.94	[6.49, 9.55]
OPRD‡	0.044012	**81.07**	[79.11, 83.18]	**14.89**	[12.91, 16.94]
Attribute (ADD/SUBSTITUTE)					
Qwen3-VL-Emb-2B					
InfoNCE	0.046428	75.17	[72.43, 77.94]	9.80	[7.75, 11.99]
CMSE†	0.031921	80.06	[77.15, 82.91]	11.35	[9.07, 13.87]
CMSE‡	0.038067	79.55	[76.76, 82.17]	**15.05**	[12.47, 17.72]
OPRD†	0.035230	**80.81**	[78.05, 83.57]	13.65	[11.08, 16.37]
OPRD‡	0.043707	79.24	[76.48, 81.81]	14.92	[12.27, 17.71]
Qwen3-VL-2B-Instruct					
Original	0.000793	51.21	[47.47, 55.02]	0.13	[0.00, 0.39]
InfoNCE	0.031596	68.48	[65.19, 71.58]	4.65	[3.20, 6.21]
CMSE†	0.021915	75.09	[71.96, 77.84]	**12.80**	[10.41, 15.45]
CMSE‡	0.031099	**76.35**	[73.39, 79.05]	12.65	[10.28, 15.17]
OPRD†	0.023531	72.56	[69.38, 75.40]	10.32	[8.18, 12.56]
OPRD‡	0.035688	75.80	[72.85, 78.54]	11.73	[9.47, 14.15]

Continued on the next page

Table 11 continued

Method	Mean Δ	DRR	DRR CI	CRR	CRR CI
Predicate (ADD/SUBSTITUTE)					
Qwen3-VL-Emb-2B					
InfoNCE	0.024069	65.98	[62.49, 69.42]	3.86	[2.47, 5.38]
CMSE†	0.013864	66.28	[61.76, 70.73]	3.85	[2.34, 5.56]
CMSE‡	0.016158	68.19	[64.39, 71.92]	3.53	[2.12, 5.16]
OPRD†	0.019872	69.46	[65.32, 73.57]	**7.32**	[5.15, 9.63]
OPRD‡	0.021198	**69.71**	[66.29, 73.04]	4.23	[2.75, 5.93]
Qwen3-VL-2B-Instruct					
Original	-0.000788	47.42	[42.68, 52.17]	0.85	[0.05, 1.98]
InfoNCE	0.018281	64.06	[60.10, 67.68]	2.44	[1.38, 3.73]
CMSE†	0.011458	65.21	[61.22, 69.15]	4.41	[2.86, 6.20]
CMSE‡	0.015627	65.85	[61.91, 69.53]	4.18	[2.66, 5.82]
OPRD†	0.012047	63.78	[59.91, 67.61]	4.06	[2.74, 5.51]
OPRD‡	0.019991	**68.49**	[64.71, 72.09]	**4.42**	[2.86, 6.07]

B.4 Random-Image and Random-Query Controls

For each control, ten images or queries are sampled from other groups with seed 42, and the sampled IDs are shared across models. The random-image control fixes the query and substitutes target images; the random-query control fixes the target and substitutes queries. Table 12 reports control gaps across all edit operations. Each gap is the mean target similarity score minus the corresponding mean control similarity score. Positive control gaps indicate broad query–image separation, while sensitivity to the edited condition is assessed through paired target similarity score changes. The small gaps of the Original Qwen3-VL-2B-Instruct checkpoint indicate weak query–image separation.

Table 13 gives the mean similarity scores used to compute the control gaps across all edit operations. Table 14 reports the subset-level components and control gaps for the Original Qwen3-VL-2B-Instruct checkpoint, including its negative random-query control gaps.

Table 12: Random-image and random-query control gaps across all edit operations. All values retain the similarity scale.

	Random image		Random query	
Model / method	Original query	Edited query	Original query	Edited query
CLIP-L				
Original	0.12652	0.11990	0.12629	0.11921
InfoNCE	0.30359	0.28513	0.30084	0.28086
CMSE†	0.13210	0.12237	0.13105	0.12345
OPRD†	0.15354	0.14077	0.15133	0.14217
Qwen3-VL-Emb-2B				
InfoNCE	0.39475	0.35930	0.39353	0.35683
CMSE†	0.15236	0.13847	0.15065	0.14147
CMSE‡	0.27965	0.25521	0.27911	0.25539
OPRD†	0.18140	0.16456	0.17946	0.16744
OPRD‡	0.32867	0.29756	0.32837	0.29729
Qwen3-VL-2B-Instruct				
Original	0.00220	0.00215	0.00156	0.00120
InfoNCE	0.38768	0.35864	0.38573	0.35319
CMSE†	0.13995	0.12926	0.13881	0.12940
CMSE‡	0.26400	0.24253	0.26406	0.24129
OPRD†	0.17763	0.16281	0.17682	0.16019
OPRD‡	0.28314	0.25897	0.28330	0.25801

Table 13: Mean similarity scores across all edit operations for the Qwen configurations. Superscripts $+$ and $-$ denote original and edited queries; own, rand.img, and rand.qry denote paired targets, random-image controls, and random-query controls.

Model / method	s^{+}_{own}	s^{-}_{own}	$s^{+}_{\text{rand.img}}$	$s^{-}_{\text{rand.img}}$	$s^{+}_{\text{rand.qry}}$	$s^{-}_{\text{rand.qry}}$
Qwen3-VL-Emb-2B						
InfoNCE	0.487458	0.450942	0.092704	0.091645	0.093925	0.094115
CMSE†	0.678141	0.664677	0.525786	0.526203	0.527487	0.523208
CMSE‡	0.516011	0.491310	0.236361	0.236102	0.236901	0.235920
OPRD†	0.690395	0.670822	0.508993	0.506264	0.510941	0.503379
OPRD‡	0.464211	0.432085	0.135544	0.134524	0.135844	0.134795
Qwen3-VL-2B-Instruct						
Original	0.877272	0.877513	0.875072	0.875359	0.875714	0.876311
InfoNCE	0.508428	0.477572	0.120753	0.118933	0.122694	0.124379
CMSE†	0.648113	0.636701	0.508167	0.507443	0.509303	0.507297
CMSE‡	0.524768	0.502818	0.260773	0.260289	0.260711	0.261526
OPRD†	0.502312	0.485493	0.324680	0.322687	0.325494	0.325304
OPRD‡	0.443089	0.417590	0.159949	0.158623	0.159787	0.159579

Table 14: Subset-level similarity scores and control gaps for Qwen3-VL-2B-Instruct Original. Panel (a) uses the score notation of Table 13; panel (b) gives control gaps, defined as mean target similarity scores minus mean control similarity scores. A/S denotes ADD/SUBSTITUTE; rows labeled Object (A/S), Attribute (A/S), and Predicate (A/S) additionally restrict the semantic category. All values retain the similarity scale.

(a) Mean similarity scores

Subset	s^{+}_{own}	s^{-}_{own}	$s^{+}_{\text{rand.img}}$	$s^{-}_{\text{rand.img}}$	$s^{+}_{\text{rand.qry}}$	$s^{-}_{\text{rand.qry}}$
All edits	0.877272	0.877513	0.875072	0.875359	0.875714	0.876311
A/S	0.876990	0.877122	0.875019	0.875204	0.875661	0.875852
Object	0.878442	0.879494	0.876266	0.877406	0.875791	0.876273
Attribute	0.878147	0.877834	0.875804	0.875521	0.875899	0.876668
Predicate	0.878018	0.877421	0.875467	0.874979	0.876393	0.876035
ADD	0.878080	0.879915	0.875909	0.877705	0.875599	0.876102
SUBSTITUTE	0.876384	0.875722	0.874401	0.873906	0.875824	0.876038
DELETE	0.877714	0.878331	0.875389	0.876020	0.875833	0.876802
Object (A/S)	0.878034	0.878817	0.875906	0.876717	0.875954	0.875930
Attribute (A/S)	0.875909	0.875116	0.873685	0.872916	0.875763	0.876207
Predicate (A/S)	0.876403	0.877191	0.874023	0.874905	0.876481	0.875948

(b) Control gaps

	Random image		Random query	
Subset	Original query	Edited query	Original query	Edited query
All edits	0.002200	0.002153	0.001558	0.001202
A/S	0.001971	0.001918	0.001329	0.001270
Object	0.002176	0.002089	0.002651	0.003221
Attribute	0.002343	0.002313	0.002248	0.001165
Predicate	0.002551	0.002441	0.001625	0.001386
ADD	0.002171	0.002210	0.002481	0.003814
SUBSTITUTE	0.001983	0.001816	0.000561	-0.000316
DELETE	0.002325	0.002311	0.001881	0.001529
Object (A/S)	0.002128	0.002101	0.002080	0.002888
Attribute (A/S)	0.002223	0.002200	0.000146	-0.001091
Predicate (A/S)	0.002380	0.002286	-0.000078	0.001243

C Full Retrieval Results

Results follow the benchmark order of the main text: CIRR, PinPoint, and FashionIQ. Within each student checkpoint group, methods under both teachers are listed in the order Original, InfoNCE, CMSE†, CMSE‡, OPRD†, and OPRD‡. Bold marks the best result within each student group or the standalone-teacher group, according to the metric direction. Cross-benchmark teacher comparisons appear in Appendix D.4.

C.1 CIRR

Table 15 combines all gallery and subset cutoffs under both teachers. OPRD's advantage over InfoNCE extends across cutoffs for the CLIP students and the BGE-MLLM-S2-supervised Qwen3-VL-2B-Instruct student; Qwen3-VL-Emb-2B exhibits cutoff-dependent differences. These results complement the source-domain analysis in Sec. 4.3.

Table 15: Full CIRR validation results under both teachers. Recall and Avg are percentages.

	Gallery Recall ↑				Subset Recall ↑			
Model / method	R@1	R@5	R@10	R@50	R@1	R@2	R@3	Avg ↑
Teacher Models								
BGE-MLLM-S2†	**49.41**	**80.32**	**89.28**	**97.56**	**75.10**	**90.10**	**96.10**	**77.71**
Qwen3-VL-Emb-8B‡	47.67	78.21	86.73	97.32	70.94	87.40	94.24	74.58
CLIP-L								
Original	12.80	37.77	51.16	78.40	34.51	58.89	76.63	36.14
InfoNCE	27.63	63.21	77.18	95.96	51.88	75.05	88.38	57.55
CMSE†	35.90	70.34	82.18	96.34	61.13	82.06	92.27	65.74
CMSE‡	28.70	61.59	74.31	95.19	52.86	75.87	87.73	57.22
OPRD†	**42.72**	**76.23**	**87.42**	**97.54**	**67.66**	**86.61**	**94.57**	**71.94**
OPRD‡	36.74	71.87	83.59	97.03	61.66	83.14	92.32	66.77
CLIP-B								
Original	12.60	37.19	50.01	77.45	32.70	57.71	74.93	34.94
InfoNCE	25.07	60.56	75.65	95.65	48.15	72.45	87.47	54.35
CMSE†	33.41	67.76	81.46	96.03	59.15	80.24	91.46	63.45
CMSE‡	26.48	60.70	74.62	94.95	50.92	74.91	87.28	55.81
OPRD†	**40.49**	**74.79**	**86.30**	**97.37**	**65.75**	**85.53**	**93.71**	**70.27**
OPRD‡	33.96	70.51	82.90	96.87	59.58	81.11	91.51	65.04
Qwen3-VL-Emb-2B								
Original	37.29	70.01	80.82	95.46	63.48	83.16	91.82	66.74
InfoNCE	**53.55**	**83.19**	90.05	**97.99**	**77.33**	90.79	95.98	**80.26**
CMSE†	51.54	81.46	89.62	97.92	75.48	90.17	95.77	78.47
CMSE‡	46.66	76.03	85.75	97.11	69.50	87.08	93.57	72.77
OPRD†	53.26	82.95	**90.36**	97.87	76.42	**90.89**	**96.32**	79.68
OPRD‡	48.07	78.28	86.73	97.32	71.32	87.95	94.24	74.80
Qwen3-VL-2B-Instruct								
Original	0.10	0.72	1.36	4.64	21.14	42.12	61.85	10.93
InfoNCE	46.42	78.93	87.97	97.44	73.26	88.54	94.93	76.09
CMSE†	48.36	80.32	88.38	97.46	73.67	89.21	95.43	76.99
CMSE‡	43.82	74.03	83.52	96.56	68.52	85.48	93.06	71.27
OPRD†	**52.09**	**82.47**	**89.60**	**97.82**	**76.51**	**90.07**	**95.84**	**79.49**
OPRD‡	45.73	75.91	85.22	96.80	69.77	86.27	93.88	72.84

C.2 PinPoint

Tables 16–18 separate positive retrieval, annotated-negative retrieval, and the effect of removing those negatives. Better positive ranking need not coincide with lower NegRecall or a smaller

ΔmAP@10, particularly for the Qwen students. The Original Qwen3-VL-2B-Instruct checkpoint's low negative-retrieval rate should be interpreted together with its weak positive retrieval. Table 19 adds precision and linguistic-sensitivity statistics.

Table 16: PinPoint positive retrieval under both teachers. All metrics are percentages.

	mAP ↑				Recall ↑			
Model / method	@1	@5	@10	@50	@1	@5	@10	@50
Teacher Models								
BGE-MLLM-S2†	**20.68**	14.36	12.62	14.08	**3.09**	11.82	17.96	34.58
Qwen3-VL-Emb-8B‡	15.18	**15.75**	**15.39**	**17.97**	2.38	**15.33**	**24.13**	**43.57**
CLIP-L								
Original	1.10	4.47	4.62	5.75	0.16	5.55	9.64	20.84
InfoNCE	2.20	6.42	6.63	8.29	0.32	7.67	12.88	27.31
CMSE†	4.32	8.03	8.08	9.77	0.57	9.10	14.91	30.32
CMSE‡	2.69	7.29	7.56	9.55	0.36	8.51	14.47	30.78
OPRD†	**6.06**	**9.06**	**8.97**	**10.77**	**0.97**	**9.82**	**16.10**	32.20
OPRD‡	3.42	8.04	8.35	10.31	0.53	9.44	15.81	**32.44**
CLIP-B								
Original	1.35	3.98	3.99	5.00	0.18	5.03	8.43	18.94
InfoNCE	1.79	5.17	5.32	6.71	0.27	6.52	10.92	23.65
CMSE†	3.97	6.50	6.59	8.14	0.56	7.61	12.88	26.98
CMSE‡	2.16	5.92	6.17	7.82	0.31	7.11	12.26	26.96
OPRD†	**4.49**	**7.48**	**7.43**	**9.07**	**0.68**	**8.58**	**14.09**	**29.18**
OPRD‡	2.62	6.64	6.83	8.51	0.38	7.87	13.39	28.76
Qwen3-VL-Emb-2B								
Original	9.16	11.54	11.52	13.87	1.30	12.37	19.77	38.10
InfoNCE	17.87	13.95	12.61	14.38	2.72	12.26	18.97	36.10
CMSE†	**19.72**	**15.37**	13.98	15.86	**2.86**	13.26	20.52	38.18
CMSE‡	12.64	14.22	13.99	16.42	1.99	13.92	22.30	41.10
OPRD†	18.15	15.17	14.02	15.91	2.65	13.51	21.05	38.42
OPRD‡	13.20	14.41	**14.14**	**16.64**	2.04	**14.13**	**22.58**	**41.66**
Qwen3-VL-2B-Instruct								
Original	0.04	0.02	0.02	0.02	0.00	0.02	0.05	0.26
InfoNCE	14.17	10.72	9.62	10.94	2.12	9.66	15.01	30.10
CMSE†	17.14	12.75	11.43	12.81	2.46	11.11	16.89	32.73
CMSE‡	11.33	12.26	11.76	13.87	1.72	12.11	19.20	36.71
OPRD†	**17.63**	**13.71**	**12.55**	**14.11**	**2.68**	12.07	18.64	35.07
OPRD‡	9.42	12.02	11.77	13.91	1.44	**12.37**	**19.56**	**37.11**

Table 17: PinPoint annotated-negative retrieval under both teachers. All metrics are percentages.

Model / method	NegRecall@1 ↓	NegRecall@5 ↓	NegRecall@10 ↓	NegRecall@50 ↓
Teacher Models				
BGE-MLLM-S2†	**18.68**	**13.65**	**12.19**	**11.06**
Qwen3-VL-Emb-8B‡	36.18	18.34	14.65	12.94
CLIP-L				
Original	54.46	27.37	20.56	14.64
InfoNCE	50.75	25.71	19.82	15.02
CMSE†	45.30	24.21	18.80	14.68
CMSE‡	50.16	26.54	20.72	16.34
OPRD†	**43.24**	**22.28**	**17.47**	**13.84**
OPRD‡	48.19	24.67	19.13	14.89

Continued on the next page

Table 17 continued

Model / method	NegRecall@1 ↓	NegRecall@5 ↓	NegRecall@10 ↓	NegRecall@50 ↓
CLIP-B				
Original	53.46	27.33	20.51	14.44
InfoNCE	51.26	24.81	18.55	13.80
CMSE†	44.85	23.16	17.84	13.74
CMSE‡	50.39	26.02	19.97	15.15
OPRD†	**43.34**	**21.65**	**16.94**	**13.30**
OPRD‡	49.00	24.16	18.56	14.23
Qwen3-VL-Emb-2B				
Original	44.47	20.72	16.04	13.44
InfoNCE	**28.17**	**15.65**	**12.67**	**11.30**
CMSE†	29.72	16.99	13.78	12.33
CMSE‡	40.54	20.51	16.32	14.18
OPRD†	32.26	17.52	14.19	12.53
OPRD‡	40.18	19.91	15.63	13.47
Qwen3-VL-2B-Instruct				
Original	**0.26**	**0.23**	**0.20**	**0.28**
InfoNCE	29.38	15.63	12.29	10.35
CMSE†	27.83	16.48	13.41	11.71
CMSE‡	39.14	19.51	15.17	12.89
OPRD†	30.62	17.68	14.29	12.12
OPRD‡	43.43	20.99	16.11	13.20

Table 18: PinPoint retrieval after removing annotated negatives. mAP is a percentage; ΔmAP is in percentage points.

	noNeg mAP ↑				
Model / method	@1	@5	@10	@50	ΔmAP@10 ↓
Teacher Models					
BGE-MLLM-S2†	25.17	16.70	14.54	15.73	1.92
Qwen3-VL-Emb-8B‡	**27.50**	**21.34**	**19.54**	**21.45**	4.15
CLIP-L					
Original	13.61	8.86	7.77	8.46	3.15
InfoNCE	16.41	11.79	10.54	11.59	3.90
CMSE†	17.87	13.34	11.95	13.04	3.87
CMSE‡	17.56	13.19	11.93	13.24	4.36
OPRD†	**18.43**	**14.08**	**12.64**	13.86	3.67
OPRD‡	18.05	13.87	**12.64**	**13.96**	4.29
CLIP-B					
Original	12.14	7.72	6.73	7.33	2.73
InfoNCE	13.79	9.39	8.44	9.36	3.12
CMSE†	15.04	10.77	9.80	10.84	3.21
CMSE‡	15.11	10.80	9.72	10.86	3.55
OPRD†	15.90	**11.84**	**10.64**	**11.82**	3.21
OPRD‡	**16.16**	11.65	10.46	11.57	3.63

Continued on the next page

Table 18 continued

Model / method	noNeg mAP ↑ @1	@5	@10	@50	ΔmAP@10 ↓
Qwen3-VL-Emb-2B					
Original	22.92	17.45	15.84	17.49	4.32
InfoNCE	25.21	17.36	15.23	16.55	2.61
CMSE†	**27.91**	19.27	16.98	18.34	2.99
CMSE‡	26.82	20.25	18.35	20.11	4.37
OPRD†	27.87	19.63	17.37	18.69	3.34
OPRD‡	27.35	**20.36**	**18.49**	**20.27**	4.35
Qwen3-VL-2B-Instruct					
Original	0.04	0.03	0.02	0.02	0.00
InfoNCE	20.92	13.49	11.72	12.71	2.10
CMSE†	23.51	15.74	13.78	14.79	2.35
CMSE‡	23.60	17.34	15.42	16.99	3.66
OPRD†	**25.49**	17.44	15.35	16.45	2.80
OPRD‡	23.77	**17.62**	**15.85**	**17.37**	4.08

Table 19: Additional PinPoint statistics for Qwen students. P@k denotes precision among the top-k retrieved images, and std. denotes standard deviation. Precision and relative noNeg mAP changes are percentages; linguistic-sensitivity statistics retain their original scale.

Model / method	P@1	P@5	P@10	P@50	Relative noNeg mAP@10 change	Linguistic range	Linguistic std.
Qwen3-VL-Emb-2B							
InfoNCE	17.87	18.09	14.87	6.33	20.73	0.1387	0.0541
CMSE†	19.72	19.69	16.28	6.79	21.39	0.1477	0.0576
CMSE‡	12.64	19.94	17.25	7.24	31.22	0.1472	0.0582
OPRD†	18.15	19.88	16.57	6.84	23.85	0.1533	0.0595
OPRD‡	13.20	20.15	17.37	7.30	30.77	0.1508	0.0587
Qwen3-VL-2B-Instruct							
Original	0.04	0.05	0.05	0.05	3.41	0.0023	0.0009
InfoNCE	14.17	14.45	12.09	5.29	21.85	0.1221	0.0477
CMSE†	17.14	16.47	13.52	5.79	20.60	0.1254	0.0494
CMSE‡	11.33	17.60	14.95	6.48	31.13	0.1317	0.0515
OPRD†	17.63	17.74	14.70	6.12	22.31	0.1379	0.0539
OPRD‡	9.42	17.56	15.11	6.50	34.64	0.1304	0.0509

A zero absolute change at the displayed precision need not imply an exactly zero change.

C.3 FASHIONIQ

Tables 20 and 21 collect category-level and macro-average recall at all reported cutoffs, with both teachers included in the same student groups. The results extend the main cross-domain comparison to recall at ranks 1 and 5 and show that the preferred teacher can depend on the student checkpoint and cutoff.

Table 20: FashionIQ R@10 and R@50 under both teachers. All recall values are percentages.

	Dress		Shirt		Toptee		Macro avg.	
Model / method	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
Teacher Models								
BGE-MLLM-S2†	27.76	51.12	38.08	57.36	36.67	56.91	34.17	55.13
Qwen3-VL-Emb-8B‡	**43.38**	**65.00**	**48.92**	**68.94**	**50.03**	**70.93**	**47.44**	**68.29**
CLIP-L								
Original	16.46	34.95	21.25	35.28	22.90	39.11	20.20	36.45
InfoNCE	22.91	44.97	28.90	48.82	28.81	48.85	26.87	47.55
CMSE†	24.05	45.51	34.05	52.90	30.75	52.78	29.62	50.40
CMSE‡	24.94	46.80	33.37	52.80	32.69	54.21	30.33	51.27
OPRD†	26.03	47.30	**35.82**	**54.51**	32.89	55.23	31.58	52.35
OPRD‡	**26.08**	**47.79**	34.99	54.02	**33.96**	**55.74**	**31.68**	**52.52**
CLIP-B								
Original	15.02	31.14	16.98	30.13	18.82	33.35	16.94	31.54
InfoNCE	17.85	38.03	21.64	39.16	21.93	39.52	20.47	38.90
CMSE†	19.98	41.60	27.43	47.45	27.03	46.97	24.81	45.34
CMSE‡	18.69	39.56	25.42	44.85	25.85	44.93	23.32	43.11
OPRD†	**21.42**	**44.17**	**30.18**	**50.29**	**28.91**	**50.74**	**26.84**	**48.40**
OPRD‡	19.83	41.65	28.31	48.14	28.25	48.70	25.46	46.16
Qwen3-VL-Emb-2B								
Original	24.49	45.76	34.94	55.25	33.61	54.72	31.01	51.91
InfoNCE	35.35	57.21	37.59	56.13	39.16	60.89	37.37	58.08
CMSE†	34.31	56.57	41.07	61.43	41.10	62.52	38.83	60.17
CMSE‡	38.03	60.34	45.04	65.11	45.64	68.13	42.90	64.53
OPRD†	35.70	59.00	42.30	62.12	41.76	63.79	39.92	61.64
OPRD‡	**39.61**	**61.28**	**46.12**	**66.00**	**47.78**	**68.74**	**44.51**	**65.34**
Qwen3-VL-2B-Instruct								
Original	0.20	1.29	0.39	1.23	0.51	1.53	0.37	1.35
InfoNCE	23.15	44.37	30.32	47.94	28.35	47.27	27.28	46.53
CMSE†	21.77	44.42	31.01	51.72	29.12	50.69	27.30	48.94
CMSE‡	30.09	53.15	39.50	**59.18**	38.30	59.97	35.96	57.43
OPRD†	26.08	48.74	37.10	55.50	34.06	55.84	32.41	53.36
OPRD‡	**31.23**	**54.98**	**40.33**	58.73	**39.21**	**60.99**	**36.93**	**58.24**

Table 21: FashionIQ R@1 and R@5 under both teachers. All recall values are percentages.

	Dress		Shirt		Toptee		Macro avg.	
Model / method	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
Teacher Models								
BGE-MLLM-S2†	8.68	20.13	16.24	31.11	13.72	28.81	12.88	26.68
Qwen3-VL-Emb-8B‡	**14.13**	**33.47**	**18.30**	**40.33**	**19.17**	**41.15**	**17.20**	**38.32**
CLIP-L								
Original	0.55	11.65	0.20	14.28	0.66	16.01	0.47	13.98
InfoNCE	3.62	16.21	2.65	22.18	3.06	21.01	3.11	19.80
CMSE†	**6.89**	17.50	7.61	26.84	7.65	23.76	7.38	22.70
CMSE‡	5.01	17.85	3.53	25.02	4.28	23.87	4.27	22.25
OPRD†	6.84	**18.79**	**9.52**	**27.67**	**8.26**	**25.85**	**8.21**	**24.11**
OPRD‡	6.05	18.54	4.17	26.69	5.35	25.75	5.19	23.66

Continued on the next page

Table 21 continued

Model / method	Dress		Shirt		Toptee		Macro avg.	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
CLIP-B								
Original	0.25	9.92	0.25	11.78	0.31	12.85	0.27	11.51
InfoNCE	2.23	13.19	1.23	15.70	1.73	15.50	1.73	14.80
CMSE†	4.36	13.88	5.54	20.31	5.15	19.28	5.02	17.82
CMSE‡	3.07	12.39	1.42	18.65	2.75	18.46	2.42	16.50
OPRD†	**4.76**	13.78	**6.53**	**23.45**	**5.51**	**21.57**	**5.60**	**19.60**
OPRD‡	3.67	**13.93**	2.26	21.05	3.47	20.55	3.13	18.51
Qwen3-VL-Emb-2B								
Original	4.21	17.65	6.38	26.89	5.81	25.50	5.47	23.35
InfoNCE	11.30	26.87	16.73	30.86	15.20	30.70	14.41	29.48
CMSE†	11.11	26.38	17.47	33.17	15.55	32.28	14.71	30.61
CMSE‡	12.54	28.95	17.91	36.90	16.62	36.16	15.69	34.00
OPRD†	11.35	28.41	**18.89**	35.38	16.88	33.40	15.71	32.40
OPRD‡	**12.94**	**30.29**	18.30	**38.96**	**18.00**	**38.81**	**16.41**	**36.02**
Qwen3-VL-2B-Instruct								
Original	0.00	0.10	0.10	0.20	0.00	0.26	0.03	0.18
InfoNCE	6.30	16.46	10.79	23.21	9.54	21.98	8.88	20.55
CMSE†	5.70	15.07	11.24	24.73	8.87	22.03	8.60	20.61
CMSE‡	**9.27**	22.91	13.79	32.24	**12.49**	29.63	**11.85**	28.26
OPRD†	7.14	19.29	**14.77**	28.75	11.37	26.47	11.09	24.84
OPRD‡	8.92	**23.90**	13.98	**32.68**	12.24	**30.85**	11.72	**29.14**

D OPRD Design Analyses

The candidate and supervision controls use CLIP-L and a frozen BGE-MLLM-S2 teacher, with the settings in Appendix A.2. CIR-QDiag results in these controls use the ADD/SUBSTITUTE subset defined in Appendix B.1, with the same group-balanced aggregation and paraphrase-based calibration. Cross-model results across all edit operations appear in Appendix B.2.

D.1 Student-Driven Candidate Selection

Table 22 compares candidate-source and target-insertion controls against a shared OPRD configuration. For the candidate-source comparison, both conditions retain the target and select the remaining candidates either by student retrieval or uniform sampling from non-target, non-reference gallery images. Student-retrieved candidates improve retrieval performance and CRR on the ADD/SUBSTITUTE subset. Both candidate sets refresh at the start of each epoch, so the comparison measures the effect of candidate source under per-epoch refresh.

Table 22: Candidate construction ablations. Bold indicates the best reported result. Recall, mAP, and response rates are percentages; score differences and thresholds retain their similarity scale.

(a) CIRR

Candidates	Target insertion	Gallery Recall ↑ R@1	R@5	R@10	R@50	Subset Recall ↑ R@1	R@2	R@3
On-policy†	No	42.0713	76.0823	87.3475	97.3930	67.4958	86.4387	94.4511
Random†	Yes	36.3071	71.4183	83.3533	96.9385	61.3011	82.9467	92.8008
On-policy†	Yes (OPRD)	**42.7171**	**76.2258**	**87.4193**	**97.5365**	**67.6632**	**86.6061**	**94.5707**

(b) Cross-dataset transfer

Candidates	Target insertion	FashionIQ ↑ R@1	R@5	R@10	R@50	PinPoint mAP@10 ↑	PinPoint NegRecall@10 ↓
On-policy†	No	8.0909	24.0389	31.3851	**52.4468**	8.94	17.55
Random†	Yes	6.3028	23.2970	30.8637	51.6963	8.00	19.27
On-policy†	Yes (OPRD)	**8.2073**	**24.1062**	**31.5799**	52.3464	**8.97**	**17.47**

(c) CIR-QDiag: ADD/SUBSTITUTE subset

Candidates	Target insertion	Mean Δ	DRR ↑	CRR ↑	b_ψ
On-policy†	No	0.025563	**78.6117**	10.1565	0.070824
Random†	Yes	0.016002	74.0030	5.8667	0.059378
On-policy†	Yes (OPRD)	0.026183	78.4262	**10.3033**	0.072281

D.2 Graded Teacher Supervision

Student-mined InfoNCE and OPRD both retrieve candidates using their own evolving students and refresh them at the start of each epoch. InfoNCE uses one-hot target labels; OPRD matches the frozen teacher's soft relevance distribution on its mined candidates using reverse KL. Table 23 shows better retrieval performance and higher ADD/SUBSTITUTE CRR for OPRD, supporting graded teacher supervision under the student-driven mining procedure. This CLIP-L control uses a separate student-mined InfoNCE run, distinct from the main-table baseline. The comparison evaluates training procedures with the same mining rule, rather than isolating losses on identical candidate sets.

Table 23: Student-mined InfoNCE and OPRD. A/S CRR denotes CRR on the ADD/SUBSTITUTE subset. Bold indicates the best result. All retrieval and response rates are percentages.

Method	CIRR R@1 ↑	FashionIQ R@10 ↑	PinPoint mAP@10 ↑	PinPoint NegRecall@10 ↓	A/S CRR ↑
Student-mined InfoNCE	33.92	29.47	6.51	17.60	6.45
OPRD†	**42.72**	**31.58**	**8.97**	**17.47**	**10.30**

D.3 Target Insertion and Training Trajectory

Without insertion, reverse KL uses the student's top-32 candidates after reference exclusion, including queries whose targets are absent from that set. Target insertion instead combines the target with the 31 highest-ranked non-target, non-reference images, keeping the candidate count fixed (Table 22). Its CIRR benefit is present at each checkpoint and is more pronounced early in training (Table 24). Both variants use the shared training budget in Appendix A.2 and the epoch-four checkpoint for final evaluation on every dataset.

Table 24: Effect of target insertion on CIRR R@1 during training.

Epoch	Without insertion†	With insertion†	Difference
1	37.5269	**39.9904**	+2.4635
2	41.0906	**41.4733**	+0.3827
3	42.1191	**42.5975**	+0.4784
4	42.0713	**42.7171**	+0.6458

D.4 Teacher Choice and Transfer

Table 25 compares both OPRD teachers across the student checkpoints. BGE-MLLM-S2 favors CIRR ranking and PinPoint negative suppression. Qwen3-VL-Emb-8B favors all-edit CRR and FashionIQ transfer for the two Qwen students, while the preferred teacher for PinPoint positive ranking depends on the student checkpoint. Teacher choice therefore depends on the retrieval, transfer, and diagnostic criteria of interest.

Table 25: Effect of teacher choice. CIR-QDiag CRR covers all edit operations. Bold denotes the best in each group. All retrieval and response rates are percentages.

Teacher	CIRR R@1 ↑	FashionIQ R@10 ↑	PinPoint mAP@10 ↑	PinPoint NegRecall@10 ↓	CIR-QDiag CRR ↑
CLIP-L					
BGE-MLLM-S2†	**42.72**	31.58	**8.97**	**17.47**	6.22
Qwen3-VL-Emb-8B‡	36.74	**31.68**	8.35	19.13	**6.51**
CLIP-B					
BGE-MLLM-S2†	**40.49**	**26.84**	**7.43**	**16.94**	6.10
Qwen3-VL-Emb-8B‡	33.96	25.46	6.83	18.56	**6.38**
Qwen3-VL-Emb-2B					
BGE-MLLM-S2†	**53.26**	39.92	14.02	**14.19**	10.24
Qwen3-VL-Emb-8B‡	48.07	**44.51**	**14.14**	15.63	**11.88**
Qwen3-VL-2B-Instruct					
BGE-MLLM-S2†	**52.09**	32.41	**12.55**	**14.29**	7.35
Qwen3-VL-Emb-8B‡	45.73	**36.93**	11.77	16.11	**9.98**

D.5 Candidate Count, Temperature, and Training Cost

Table 26 reports a dedicated CLIP-L/BGE-MLLM-S2 sweep using the common training duration in Appendix A.2. We vary the candidate count at a fixed temperature and compare temperatures at a fixed candidate count. All comparisons use the sweep's shared reference run, which is separate from the main-experiment run. Smaller candidate sets reduce the reported training time while retaining competitive retrieval performance. Training time is reported in hours; separate timings for teacher precomputation, candidate refresh, and optimization are unavailable.

Table 26: Candidate-count and temperature sweep for CLIP-L with BGE-MLLM-S2. Retrieval metrics are percentages, FashionIQ values are macro averages, and training time is in hours. These configurations belong to the same dedicated sweep.

(a) CIRR

		Gallery Recall ↑				Subset Recall ↑		
K	T	R@1	R@5	R@10	R@50	R@1	R@2	R@3
8	0.05	43.7694	77.1586	87.4193	97.3451	68.8591	86.6539	94.6424
16	0.05	43.1476	77.0390	87.0844	97.4647	68.0938	86.8453	94.5228
32	0.05	42.6214	76.2497	87.1083	97.4169	67.3762	86.5343	94.4989
32	1.0	41.9995	75.7713	86.9409	97.3451	67.3762	86.1995	94.4033

(b) Cross-dataset retrieval and training time

		FashionIQ ↑		PinPoint		
K	T	R@10	R@50	mAP@10 ↑	NegRecall@10 ↓	Training hours
8	0.05	32.2379	52.9462	8.97	17.10	1.0563
16	0.05	31.7746	52.8466	8.93	17.14	1.6255
32	0.05	31.5293	52.2481	8.87	17.40	2.7460
32	1.0	31.4182	51.9484	8.90	17.47	2.7378

E PROMPTS FOR CIR-QDIAG DATA CONSTRUCTION

We generate candidate edits using three separate text-only requests per source instruction, corresponding to ADD, DELETE, and SUBSTITUTE. Each request considers object, attribute, and predicate edits, yielding at most nine candidates per source instruction. We then review the candidates jointly at the paraphrase-group level using the reference image(s) and the supplied positive retrieval images. Both stages use `gpt-5.6-terra`.

Below, we present the task instructions, omitting image-count and serialization details for readability. The review criteria apply to the supplied positive images. Accepted ADD and SUBSTITUTE edits form the ADD/SUBSTITUTE subset. DELETE edits yield strictly weaker queries and are included in the all-edit results and also reported separately; they need not contradict the positive images. The prompt excerpts retain their construction-time wording, including "hard-negative", "positive instruction", and fields such as `negative_text`. These denote edit proposals and their source instructions in the prompts; they do not imply that every proposal is incompatible with the targets, particularly for DELETE.

E.1 TEXT-ONLY GENERATION

Each request supplies the original instruction and the designated edit operation. The model returns one proposal or `null` for each semantic category: object, attribute, and predicate. Each proposal includes the edited text and its before/after spans.

ADD.

```
Generate text-only ADD hard-negative instructions. You receive
exactly one positive instruction and no images. Return the
object, attribute, and predicate slots; each slot is one
proposal or null.

For each proposal, insert exactly one contiguous after_span
into the original instruction and change nothing else. Set
before_span=null. Removing that exact after_span from
negative_text must recover the original source_instruction
apart from whitespace and terminal punctuation.

- object: add one named entity or entity-presence constraint.
- attribute: add one property to an entity already named or
  unambiguously referenced.
- predicate: add one action or relation over entities already
  named or referenced.

The result must be fluent, natural, close to the original, and
plausibly difficult. Do not use images, claim visual evidence,
rewrite the sentence, restate existing meaning, or add multiple
constraints. Use null when a safe single insertion is
unavailable. Return only the structured response.
```

DELETE.

```
Generate text-only DELETE hard-negative instructions. You
receive exactly one positive instruction and no images. Return
the object, attribute, and predicate slots; each slot is one
proposal or null.

For each proposal, delete exactly one contiguous before_span
from the original instruction and change nothing else. Set
after_span=null. Deleting that exact span must recover
negative_text apart from whitespace and terminal punctuation.
The result must remain fluent, meaningful, and strictly weaker.

Count semantic constraints rather than raw tokens. A
```

```
before_span may include grammatical glue that belongs only to
the deleted constraint, such as an article, preposition, comma,
or coordinator "and". For example, "brown and white animals"
may become "white animals" by deleting "brown and"; that is one
attribute deletion.

- object: remove one entity constraint while leaving a
  meaningful query.
- attribute: remove one property while retaining its entity.
- predicate: remove one action/relation while retaining its
  arguments.

Do not delete bundled independent constraints and do not repair
or rewrite the remaining sentence. Use null when no safe
deletion exists. Return only the structured response.
```

SUBSTITUTE.

```
Generate text-only SUBSTITUTE hard-negative instructions. You
receive exactly one positive instruction and no images. Return
the object, attribute, and predicate slots; each slot is one
proposal or null.

For each proposal, replace exactly one contiguous before_span
from the original instruction with exactly one contiguous
after_span and change nothing else. Both spans must be explicit
and different. Applying that one replacement must recover
negative_text apart from whitespace and terminal punctuation.

- object: replace one entity identity or category.
- attribute: replace one property value while retaining its
  entity.
- predicate: replace one action or relation while retaining its
  arguments and semantic scope.

The result must be fluent, natural, close to the original, and
plausibly difficult. Do not use images, invent visual evidence,
replace multiple values, switch semantic axes, or rewrite the
sentence. Use null when a safe single replacement is
unavailable. Return only the structured response.
```

E.2 Model-based Visual Validation

The review request supplies the group's original instructions and all generated candidates, including each candidate's source text, edit operation, semantic category, and proposed spans. It also supplies the reference image(s) and positive retrieval images. The model returns an accept/reject decision and a concise reason for each candidate.

```
Validate every supplied hard-negative text for one PinPoint
paraphrase group in exactly one pass.

The supplied images are labeled as REFERENCES or POSITIVE
retrieval results. Official negative images are not included in
this request. Give exactly one verdict for every negative_id;
do not omit, duplicate, invent, or rewrite candidates.

First check every candidate against its own source_text:
- it is fluent, natural, close to source_text, and preserves
  all unedited meaning;
- it performs exactly one stated edit and matches
  semantic_target;
- object means entity identity/presence, attribute means a
  property of an existing entity, and predicate means an
  action/relation with the same arguments and scope.
```

```
- before_span and after_span are untrusted generator claims.
  Verify them against source_text and negative_text yourself.
  Reject as not_single_edit when the text is unchanged, the
  stated span operation cannot reconstruct negative_text, or
  more than one semantic constraint changed.

ADD and SUBSTITUTE:
- use the REFERENCE image(s) to resolve phrases such as "this",
  "same", <img1>, and <img2>;
- accept only when the new fact is clearly false for EVERY
  SUPPLIED POSITIVE image;
- if any supplied POSITIVE supports or remains compatible with
  the new fact, reject as positive_compatible;
- if any required fact cannot be judged reliably from the
  references and all supplied positives, reject as
  not_visually_verifiable;
- difficulty must come from a subtle single edit, not from
  ambiguity or hidden evidence.

DELETE:
- do not require the weaker text to contradict positive images;
- accept when it removes exactly one meaningful constraint,
  preserves everything else, and produces a fluent, meaningful,
  strictly weaker query;
- count semantic constraints, not raw tokens. Removing
  grammatical glue belonging only to the deleted constraint is
  still one edit. For example, "brown and white animals" ->
  "white animals" is one attribute DELETE;
- reject bundled deletions, a wrong semantic axis, changed
  surviving meaning, or broken grammar.

For rejection choose exactly one failure_code from the response
schema and state the decisive reason concisely. For acceptance
use failure_code=null. Return only the structured response.
```