UNICIR: UNLOCKING UNIVERSAL EMBEDDINGS FOR TRAINING-FREE COMPOSED IMAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Composed image retrieval (CIR) searches for images that satisfy a textual modification while preserving relevant content from a reference image. Many training-free methods construct queries through generated descriptions, making retrieval depend on the visual details captured in text. Universal multimodal embeddings jointly encode the reference image and modification, providing a strong alternative. However, retrieval can remain visually close to the reference while missing a requested change. We propose `UniCIR`, a training-free method that refines this joint representation through a controlled description pair. A frozen vision–language generator describes the intended target. We then replace one edit-linked value to form a negative description, keeping all other wording fixed. The shared retriever encodes both descriptions and refines the original query by adding the weighted positive and subtracting the weighted negative. This contrast targets the requested distinction while preserving the original joint embedding as an anchor. The resulting vector searches a fixed gallery with all model weights frozen. Our experiments show gains of 6.31, 7.89 and 8.68 percentage points over Qwen-8B on CIRR test R@1, CIRCO test mAP@5 and standard FashionIQ validation average R@10, respectively. Gains extend to four additional retrievers, while matched ablations show improvements over augmentation with positive descriptions.

1 INTRODUCTION

Composed image retrieval (CIR) retrieves a target image using a reference image and a textual modification (Vo et al., 2019). For example, the request in Figure 1 asks for a pink top with light-blue sleeves. The image supplies the source appearance, while the text specifies how the target should differ. Retrieval must apply these changes while retaining the reference context needed to identify the target. We study training-free CIR adaptation with pretrained models kept fixed.

In this context, many CIR methods build a query through a textual interface. Learned zero-shot methods, such as Pic2Word, SEARLE and LinCIR, map the reference image into text tokens and combine them with the edit (Saito et al., 2023; Baldrati et al., 2023a; Gu et al., 2024). Training-free methods such as CIReVL and OSrCIR instead generate a target description for an image–text retriever (Karthik et al., 2024; Tang et al., 2025b). For the garment query, this description might specify the desired colours (e.g., "a pink top with light-blue sleeves" in Figure 1a). Reference information then reaches retrieval through the generated wording or learned token. A generated sentence can specify the requested sleeve colour while leaving other visible reference details unstated, making retrieval depend on what the intermediate representation retains.

Universal multimodal retrievers directly encode the original image and edit in a shared retrieval space (Wei et al., 2024; Jiang et al., 2025; Li et al., 2026a). In the same example, the garment photograph remains an input to the retrieval encoder alongside the colour request. We call this frozen joint-input baseline Universal. In our evaluation, Qwen-8B reaches 53.19% CIRR test R@1 and 50.41% standard FashionIQ validation average R@10, exceeding the compared training-free methods (Tables 1 and 2). BGE-VL-large and WISER score higher on CIRCO. These results make strong frozen universal retrievers an important baseline for training-free CIR.

However, joint access to the inputs can still leave a requested condition unsatisfied. Figure 1b shows Universal retrieving a *pink top* whose sleeves are also *pink*. The overall colour matches, but the

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.



Figure 1: Motivation with the same reference and edit. (a) Caption and projection routes pass a description or learned token to a text encoder. Universal directly encodes the original image and edit. (b) Its recorded retrieval misses the requested sleeve colour. (c) A controlled pair highlights that distinction. The edit is shortened. Descriptions and refinement are schematic, and the target is annotated.

sleeve-colour condition is missed. The query must distinguish such a partial match from a top with light-blue sleeves, while continuing to use the original visual reference.

We propose `UniCIR` to make this distinction explicit before retrieval. Its *Contrastive Intent Constructor* (CIC) uses a frozen vision–language generator to describe the intended target and propose an alternative for one edit-linked value. For example, replacing "light-blue" with "pink" for the sleeves in Figure 1c produces the negative description. Both descriptions refer to the same top, with all other wording fixed. Their contrast distinguishes the requested sleeve colour from a partial match. The shared retriever encodes both descriptions, adding the positive and subtracting the negative from the original joint vector. The original joint representation remains part of the query. One resulting vector searches the fixed gallery, with all model weights unchanged.

We establish strong universal-retrieval baselines and identify concrete remaining errors, such as the unmet sleeve-colour condition in Figure 1. `UniCIR` complements this representation with a controlled description pair. Our matched Qwen-8B evaluations give gains of 6.31, 7.89 and 8.68 percentage points in CIRR test R@1, CIRCO test mAP@5 and standard FashionIQ validation average R@10 (Tables 1 and 2). The gains extend to four additional retrievers. Matched positive controls and component ablations show the benefit of both the paired negative and the original joint query.

2 Related Work

Composed image retrieval. CIR combines a reference image and a textual modification to retrieve a target image (Vo et al., 2019; Delmas et al., 2022). Supervised methods learn this composition from annotated triplets. Zero-shot methods such as Pic2Word, SEARLE and LinCIR learn visual-to-text mappings (Saito et al., 2023; Baldrati et al., 2023a; Gu et al., 2024). Training-free methods use pretrained generators to interpret the query. CIReVL modifies a reference caption, whereas OSrCIR jointly processes the image and modification with reflective chain-of-thought prompting (Karthik et al., 2024; Tang et al., 2025b). WISER combines generated text and image queries through verification and adaptive fusion (Wang et al., 2026a). Paracosm edits the reference image and synthesizes gallery counterparts for matching (Wang et al., 2026b). We retain the original image–edit embedding during retrieval and use generated descriptions to refine it.

Universal embeddings and synthetic supervision. UniIR, VLM2Vec and Qwen3-VL-Embedding support image, text and mixed-modal queries in a shared retrieval space (Wei et al., 2024; Jiang et al., 2025; Li et al., 2026a). Qwen3-VL-Embedding combines large-scale contrastive pretraining and reranker distillation for diverse multimodal retrieval tasks (Li et al., 2026a). MegaPairs trains MMRet-Large (BGE-VL-large) on synthetic reference-image–instruction–target-image triplets (Zhou et al., 2024). Think-Then-Embed trains embedders with generated reasoning, and CORE distills compositional representations from a reranker (Cui et al., 2025; Song et al., 2026). CIRCUS examines the compositional demands of CIR benchmarks (Attimonelli et al., 2026). Using a pretrained retriever, `UniCIR` refines queries at inference without further parameter updates or gallery changes.

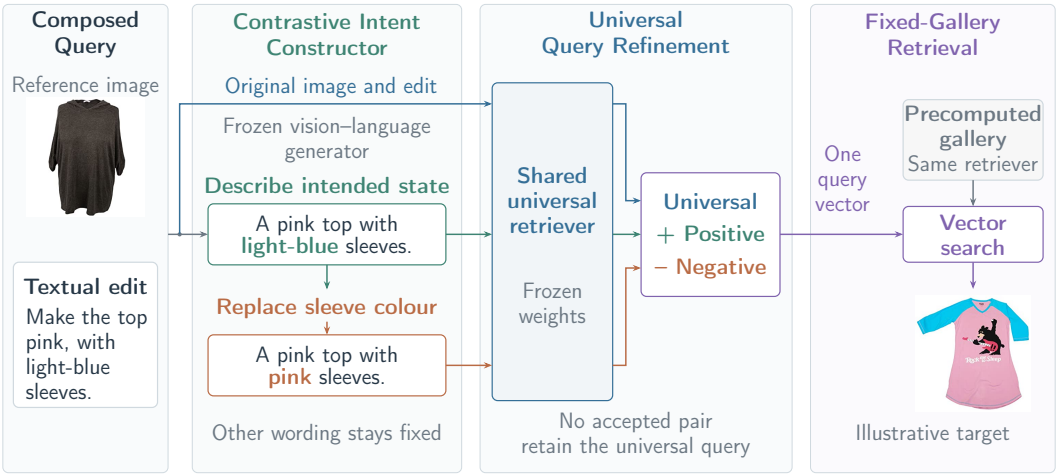


Figure 2: UniCIR from input to retrieval. Universal directly encodes the image and edit. The Contrastive Intent Constructor produces a controlled pair. Universal Query Refinement combines their encodings with the original query for Fixed-Gallery Retrieval. The pair and annotated target illustrate the process.

Intent-aware retrieval and negative evidence. CIR-LVLM, ReCALL, QuRe and TEMA improve query–target alignment through intent-conditioned encoding, corrective supervision, query-relevant hard negatives or entity composition (Sun et al., 2025b; Yang et al., 2026; Kwak et al., 2025; Li et al., 2026b). DualCIR and CoTMR combine target-caption scores with positive and negative attribute or object descriptions (Zhao et al., 2025; Sun et al., 2025a). DEO iteratively moves a query toward positive subqueries and away from negative ones while regularizing it toward the original query (Lee et al., 2026). MulVec uses role-specific probes with global and local candidate representations (Zhang et al., 2026). UniCIR builds on signed refinement through a controlled description pair. CIC changes one edit-linked value while fixing the remaining wording, and the original image–edit embedding anchors the correction. Appendix C.3 relates this update to prior formulations.

3 Methodology

Figure 2 presents our UniCIR framework. We use the original image–edit query to retain reference context and a controlled description pair to specify a requested change. The pair is encoded by the same retriever and combined with the original query to search a fixed gallery.

3.1 Composed Query

Let $I_r \in \mathcal{I}$ be a reference image in the image space $\mathcal{I}$, $T_m \in \mathcal{T}$ a textual modification in the space $\mathcal{T}$, and $\mathcal{D}$ an image gallery. The *universal retriever* Ψ_θ has frozen weights and maps image, text and joint inputs to unit vectors in a shared d-dimensional space. The universal query is $\mathbf{a} = \Psi_\theta(I_r, T_m)$, and each gallery image has embedding $\mathbf{v}_I = \Psi_\theta(I)$. Let $\mathcal{D}_r \subseteq \mathcal{D}$ denote the eligible retrieval gallery for this query. Its benchmark-specific construction is described in Section 4.1. The Universal retrieval result is

$$\widehat{I}_{\mathrm{U}} = \arg\max_{I \in \mathcal{D}_r} \langle \mathbf{a}, \mathbf{v}_I \rangle \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product, which equals cosine similarity for these unit vectors. All input modes use the same frozen checkpoint θ, so the original image–edit query, auxiliary descriptions and gallery images are compared in the same space. Encoding the original image also preserves direct access to visual details omitted from a generated description.

The joint embedding encodes the reference image and requested modification. The description branch adds an explicit contrast for a selected condition, such as sleeve colour, in the same retrieval space.

3.2 CONTRASTIVE INTENT CONSTRUCTOR

We construct the description pair to express a requested condition and a nearby violation. A frozen vision–language generator h_ϕ, with parameters separate from the retriever, first produces the intended-state description $c^+ = h_\phi(I_r, T_m)$. The prompt asks it to combine the requested changes with the relevant reference context.

We form the negative by editing this positive description because independently generating both descriptions can also change the object, pose or background. Fixing the surrounding text confines the explicit difference to the selected condition. Generation uses greedy decoding, with output caps of 256 tokens for the positive description and 64 for the replacement proposal. The positive describes the intended image and relevant context. The proposal only supplies a replacement for one selected value.

The constructor matches the modification with c^+ to select an edit-linked condition $\ell = (o, b, w)$. Here, o identifies the object or part, b its attribute, and w the corresponding value span in c^+. We consider explicitly requested conditions with a unique corresponding value span. Part-level attributes take priority over global attributes, with ties resolved by their order in the modification. A text-only call to the same generator proposes one alternative value w' of the same attribute type. We replace the selected value to obtain

$$\begin{aligned} c^+ &= c_{\mathrm{L}} \oplus w \oplus c_{\mathrm{R}}, \\ c^- &= c_{\mathrm{L}} \oplus w' \oplus c_{\mathrm{R}}, \qquad w' \neq w. \end{aligned} \tag{2}$$

where $\oplus$ denotes string concatenation and c_{L} and c_{R} are the unchanged prefix and suffix. For *A pink top with light-blue sleeves*, the selected attribute is sleeve colour. Replacing *light-blue* with *pink* gives *A pink top with pink sleeves*. The top colour and other wording stay fixed, while the new sleeve colour violates the selected request.

The supported conditions include colour, brightness, local attribute bindings, pattern regularity and restricted counts. We reject unchanged or unsupported replacements and those that fail local compatibility checks. We also compare the prefix and suffix to verify that the other wording stays unchanged. After rejection, we retain the Universal query without sampling another proposal. Only accepted pairs enter query refinement. We evaluate their semantic quality in Appendix D.2. Generation settings and embedding interfaces are provided in Appendix A.1.

3.3 UNIVERSAL QUERY REFINEMENT

To incorporate the local contrast into retrieval, we encode an accepted pair as $\mathbf{p} = \Psi_\theta(c^+)$ and $\mathbf{n} = \Psi_\theta(c^-)$. We add the positive embedding to the universal query and subtract the paired negative, giving

$$\widetilde{\mathbf{q}} = \mathbf{a} + \alpha\mathbf{p} - \beta\mathbf{n}, \qquad \mathbf{q} = \frac{\widetilde{\mathbf{q}}}{\|\widetilde{\mathbf{q}}\|_2}. \tag{3}$$

The coefficient of the original query $\mathbf{a}$ is fixed at one. Nonnegative weights α and β control the positive and negative contributions and are shared across queries. We normalize the nonzero combined vector once. We use $\mathbf{q} = \mathbf{a}$ when a description is unavailable, a unique admissible value span cannot be selected, or the pair fails a construction check. These queries keep the original Universal ranking. The default weights and their sensitivity are described in Section 4.1 and Appendix C.1.

With $\alpha = 0.50, \beta = 0.25$, the unnormalized query can be written as $\mathbf{a} + 0.25\mathbf{p} + 0.25(\mathbf{p} - \mathbf{n})$. The correction combines positive-description evidence with the unnormalized paired difference, alongside the original image–edit query. Text embeddings depend on the full description, so the paired difference contrasts two complete descriptions in the retrieval space. Component ablations appear in Section 4.3, followed by behavioural analysis in Section 4.4.

3.4 FIXED-GALLERY RETRIEVAL

We pre-encode the gallery with the same retriever and keep its index fixed. Eligible images $I \in \mathcal{D}_r$ are ranked by $\langle \mathbf{q}, \mathbf{v}_I \rangle$. For an accepted pair, Equation (3) gives

$$\langle \mathbf{q}, \mathbf{v}_I \rangle = \frac{\langle \mathbf{a}, \mathbf{v}_I \rangle + \alpha\langle \mathbf{p}, \mathbf{v}_I \rangle - \beta\langle \mathbf{n}, \mathbf{v}_I \rangle}{\|\widetilde{\mathbf{q}}\|_2}. \tag{4}$$

The positive denominator is shared by all candidates. We can therefore obtain this ranking with one vector search, using a single precomputed embedding per gallery image. The pair is constructed entirely on the query side. Candidate images and retrieval scores are excluded from condition selection and pair acceptance.

Query construction generates the target description and, when an admissible edit-linked span is found, a replacement proposal. Accepted pairs require two text encodings. These operations depend on the query alone. Retrieval then uses one vector search over the fixed gallery.

To apply the construction to another universal retriever, we reuse the descriptions and weights but encode the pair, joint query and gallery in that retriever's own space. Figure 4 evaluates this transfer. Algorithm 1 in Appendix A.2 summarizes query construction and retrieval.

4 EXPERIMENTS

We evaluate universal embeddings for CIR, same-retriever gains and the contribution of the controlled pair. Further analyses examine edit adherence, retrieval errors and inference cost.

4.1 EXPERIMENTAL SETUP

We evaluate on CIRR and CIRCO test sets and the standard FashionIQ validation set (Liu et al., 2021; Baldrati et al., 2023a; Wu et al., 2021). CIRR/CIRCO test relevance labels are hidden, and their official servers evaluate the submitted rankings. We report the standard benchmark cutoffs, including CIRR subset recall and equally weighted FashionIQ category averages. The primary metrics are CIRR R@1, CIRCO mAP@5 and FashionIQ average R@10. The reference image is eligible on CIRCO and excluded on CIRR and FashionIQ. Full-query evaluation includes queries that fall back to Universal.

We use Qwen3-VL-Embedding-8B as the default retriever and frozen Qwen3-VL-4B-Instruct within CIC. Training-free refers to our CIR adaptation with frozen pretrained models. Qwen3-VL-Embedding is trained through contrastive learning and reranker distillation (Li et al., 2026a). Universal measures this model's retrieval performance before the proposed refinement. The full method uses $\alpha = 0.50, \beta = 0.25$ (Appendix C.1, Figure 10 and Table 19). All query-construction controls fix the retriever and gallery. Independent positive baselines select their weights separately. Component-removal variants keep Full's remaining weights and modify the same queries as Full. All other queries use the original Universal embedding. The second-positive control matches information and token limits (Appendix A). We report query latency for each complete pipeline. Published baselines use their reported configurations (Appendix A.1).

4.2 QUANTITATIVE AND QUALITATIVE RESULTS

Tables 1 and 2 compare `UniCIR` with training-free, trained zero-shot and benchmark-supervised CIR methods, alongside frozen Qwen and GME retrievers (Zhang et al., 2025). Paired Qwen-8B rows share the retriever, gallery and split. Figure 4 extends the same-retriever comparison to five retrievers.

Table 1 shows that `UniCIR` raises CIRCO test mAP@5 from the universal Qwen-8B baseline's 31.26 to 39.15, a gain of 7.89 points. BGE-VL-large has the highest mAP@5 among the compared baselines at 39.20. `UniCIR` reaches 39.15 and leads this comparison at mAP@10/25/50, exceeding BGE-VL-large by 0.53/0.12/0.37 points. All gallery embeddings remain fixed during query refinement. Figure 3(a) illustrates the paired update on a development query requesting three same-colour urinals in front view. The universal query and positive-only control both return oblique views. Both augmented queries use the same positive description, which already specifies *front-facing*. Adding its paired negative changes the returned image from an oblique view to a frontal view. Figure 17 gives the complete top-five rankings.

The universal Qwen-8B baseline reaches 53.19 R@1 on CIRR test (Table 1), exceeding the compared training-free and trained zero-shot methods. `UniCIR` reaches 59.50, improving the same retriever by 6.31 points and the strongest compared baseline, ReCALL (Yang et al., 2026), by 3.98 points.

Table 1: CIRCO and CIRR test. Unstarred baselines use published scores. $*$ denotes our evaluation. † marks external synthetic-triplet pretraining. Bold/underline mark best/second-best scores. Additional cutoffs appear in Table 12.

	CIRCO (test)				CIRR (test)					
	mAP@k				Recall@k			Recall$_{\text{subset}}$@k		
Method	k=5	10	25	50	k=1	5	10	k=1	2	3
Training-free										
OSrCIR	30.47	31.14	35.03	36.59	37.26	67.25	77.33	69.22	85.28	93.55
CoTMR	32.23	32.72	35.60	36.83	36.36	67.52	77.82	71.19	86.34	93.87
WISER	36.53	38.14	40.46	41.26	49.54	77.40	85.76	78.10	89.06	92.68
Trained zero-shot										
LinCIR	19.71	21.01	23.13	24.18	35.25	64.72	76.05	63.35	82.22	91.98
PrediCIR	23.70	24.60	25.40	26.00	37.00	66.10	77.90	—	—	—
CVSI	33.51	34.96	37.86	38.98	41.37	72.58	82.39	71.35	87.54	94.68
BGE-VL-large†	**39.20**	40.20	42.90	44.00	38.00	70.30	81.10	73.20	88.00	94.30
Benchmark-supervised										
CIR-LVLM	—	—	—	—	53.64	83.76	90.60	79.12	92.33	96.67
QuRe	—	—	—	—	52.22	82.53	90.31	78.51	91.28	96.48
ReCALL	—	—	—	—	55.52	84.07	91.83	81.49	93.35	97.64
Universal methods										
GME-7B*	24.68	26.07	29.31	30.35	41.31	73.91	83.93	68.22	85.06	93.07
Qwen-8B*	31.26	29.41	32.06	33.20	53.19	82.35	89.74	79.74	91.14	95.09
UniCIR	39.15	**40.73**	**43.02**	**44.37**	**59.50**	**88.21**	**94.27**	**83.32**	**94.74**	**98.47**



Figure 3: Diagnostics of an earlier clause-pair construction on development queries. Panels show viewpoint (a) and local-structure correction (b). Green borders mark annotated positives. Appendix H gives the configuration and description pairs.

Improvements also occur at R@5 and R@10. Subset R@1/2/3 increase by 3.58/3.60/3.38 points over Qwen-8B (Appendix B).

Figure 3(b) illustrates a local-structure change. The request asks for an opening in the front of a knee pad. Universal and positive-only retrieval return closed-front pads, whereas paired correction retrieves the annotated open-front target. The target moves from rank 4 under positive-only retrieval to rank 1, showing the requested change directly in the returned image.

Table 2 reports standard FashionIQ validation, where the universal Qwen-8B query reaches 50.41 average R@10, ahead of the compared training-free and trained zero-shot methods. UniCIR

Table 2: FashionIQ standard validation with equal-category averages. $*$ denotes our evaluation. Other baselines use published scores. † marks external synthetic-triplet pretraining. Bold/underline mark best/second-best scores. Appendix G details numerical conventions.

	Shirt		Dress		Toptee		Average	
Method	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
Training-free								
OSrCIR	38.65	54.71	33.02	54.78	41.04	61.83	37.57	57.11
CoTMR	38.32	62.24	34.51	57.36	41.90	64.30	38.25	61.32
WISER	45.98	61.58	38.57	58.35	49.21	66.96	44.59	62.30
Trained zero-shot								
LinCIR	46.76	65.11	38.08	60.88	50.48	71.09	45.11	65.69
PrediCIR	48.20	67.40	39.70	62.40	53.70	73.60	47.20	67.80
BGE-VL-large†	37.00	56.10	29.70	50.30	37.00	59.30	34.60	55.20
Benchmark-supervised								
CIR-LVLM	58.59	75.86	50.42	73.57	59.61	78.99	56.21	76.14
QuRe	53.53	72.87	46.80	69.81	57.47	77.77	52.60	73.48
ReCALL	58.49	76.59	51.81	73.48	60.83	79.19	57.04	76.42
Universal methods								
GME-7B*	52.21	69.63	39.71	63.41	51.50	71.29	47.81	68.11
Qwen-8B*	51.77	70.66	45.96	67.77	53.49	73.99	50.41	70.81
UniCIR	**60.99**	**79.05**	**53.45**	**75.26**	**62.83**	**82.00**	**59.09**	**78.77**

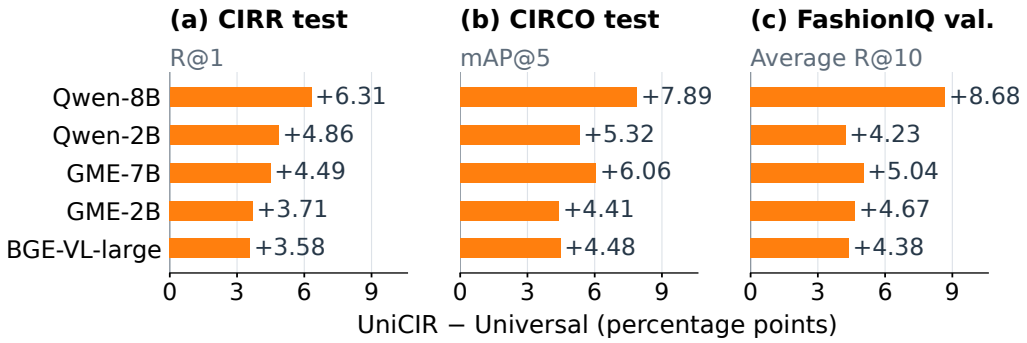


Figure 4: Gains over each retriever's Universal baseline in our matched evaluations. Panels use CIRR test R@1, CIRCO test mAP@5 and standard FashionIQ validation average R@10. Every bar compares the same retriever, split and gallery. Table 11 reports paired scores and their increments.

reaches 59.09, gaining 8.68 points over the same retriever and 2.05 points over ReCALL, the strongest compared baseline. Average R@50 rises from 70.81 to 78.77. Both recall cutoffs improve in each FashionIQ category.

Gains across universal retrievers. To examine transfer beyond Qwen-8B, we reuse the same descriptions and coefficients across five frozen Qwen, GME and BGE retrievers. Each retriever encodes the joint query, descriptions and gallery in its own space. Figure 4 shows gains over the corresponding Universal baseline on all three benchmarks. Qwen-2B improves by 4.86/5.32/4.23 points on CIRR/CIRCO/FashionIQ, and GME-2B gains 3.71/4.41/4.67 points. Both smaller retrievers benefit from the descriptions and weights used with Qwen-8B. Each gain measures the additional contribution of refinement for its fixed retriever. Table 17 also compares refinement with positive augmentation.

4.3 Ablation Study

We test whether the gain comes from the controlled contrast or target-description augmentation. Table 3 fixes the retriever and gallery on CIRR test and standard FashionIQ validation. Positive

Table 3: Ablation study on CIRR test and standard FashionIQ validation. Scores include all queries. FashionIQ averages its three categories equally. † methods modify exactly the queries for which UniCIR accepts a description pair. All other queries use the original Universal embedding.

	CIRR (test)			FashionIQ (val)	
Configuration	R@1	R@5	R@10	Avg. R@10	Avg. R@50
Reference and positive augmentation					
Universal	53.19	82.35	89.74	50.41	70.81
Positive augmentation	57.04	85.93	92.50	56.46	76.29
Two-positive augmentation	57.09	85.97	92.54	56.74	77.14
Component and construction controls					
w/o negative evidence†	56.28	85.22	91.96	56.12	76.34
w/o positive evidence†	50.47	79.82	87.79	50.55	70.94
Negative replaced by a second positive†	57.10	85.98	92.55	56.28	76.47
Shuffled negative†	51.50	80.78	88.53	52.38	72.62
Independent negative	57.05	85.93	92.51	55.71	76.00
w/o universal anchor†	49.89	79.29	87.37	48.78	69.31
UniCIR (Full)	**59.50**	**88.21**	**94.27**	**59.09**	**78.77**

Table 4: Reference evidence on canonical validation. Qwen-8B encodes the same description without or with the reference image.

Metric	Description only	Image + description
CIRR R@1	40.85	55.20
CIRCO mAP@5	27.96	33.41
FashionIQ R@10	48.09	51.39

The reference improves all three metrics. FashionIQ uses standard galleries with equal weighting across categories.

Table 5: Stronger positive controls on CIRCO validation with Qwen-8B. Table 9 gives the complete definitions and scores.

Query construction	mAP@5
Image + description	33.41
Image + edit + description	35.50
Tuned positive weighting	36.03
Reflective positive	37.09
UniCIR	**38.25**

Paired correction improves on the stronger positive descriptions, with a 1.16-point margin over reflective generation.

and two-positive augmentation use every available description and select their weights separately. Full exceeds these baselines by 2.46/2.63 and 2.41/2.35 points on CIRR/FashionIQ, respectively. Removing the negative while keeping Full's positive, weight and corrected queries reduces the scores by 3.22/2.97 points. Replacing it with a second positive also trails Full by 2.40/2.81 points under the same queries, input information and token limits. Shuffling negatives breaks their correspondence with the query and lowers recall. Independently generated negatives also perform worse and can change which queries receive a correction. Removing positive evidence or the universal anchor reduces both primary scores. Additional validation and accepted-pair subset comparisons appear in Tables 7 and 8.

Generator substitution. We replace Qwen3-VL-4B-Instruct with MiniCPM-V-4.0 while fixing the Qwen-8B retriever and gallery. On validation, paired refinement improves over MiniCPM's own positive augmentation by 2.22/4.42/2.42 points on CIRR/CIRCO/FashionIQ (Table 28), showing that the gain extends to a different generator family.

4.4 Analysis

Universal retrieval and reference evidence. We examine the contribution of reference context by fixing Qwen-8B and the generated description, then adding or removing the reference image (Table 4). Including the image improves all three validation metrics. Universal also outperforms description-only retrieval (Table 7). The stronger image-plus-description results motivate the additional positive controls used to evaluate our paired correction.

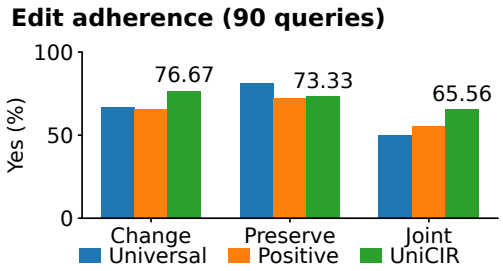


Figure 5: Edit adherence on 90 shared queries. Joint success requires both change and preservation. Rates include uncertain judgments. Table 24 gives the counts.

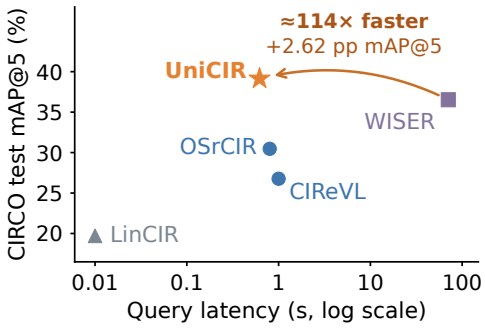


Figure 6: Reported query latency and CIRCO test mAP@5. The arrow indicates UniCIR's improvement over default WISER.

Stronger positive inputs. We further compare with joint image–edit–description encoding, tuned positive weights and reflective positive generation to examine whether stronger positive inputs recover the gain. Table 5 shows the CIRCO results. On canonical validation, Full exceeds validation-tuned positive weighting by 1.05/2.22/0.70 points on CIRR/CIRCO/FashionIQ. Figure 8 and Table 9 give the complete comparisons. Appendix J reports the paired query-level gains and losses on CIRR validation.

Edit adherence. To examine whether retrieval improvements correspond to the requested changes, we assess Universal, positive augmentation and Full on the same 90 queries. We evaluate both the requested change and preservation of relevant reference context, requiring both for joint adherence. The semantic assessments were conducted by two authors and three student volunteers, who judged retrieved images and generated descriptions.

Full improves joint adherence from Universal's 50.00% to 65.56%, compared with 55.56% for positive augmentation (Figure 5). Change adherence rises by 10.00 points, while preservation decreases by 7.78. These rates count positive judgments over all 90 queries, with no and uncertain judgments retained in the denominator. Counts and criterion overlap are reported in Tables 24 and 35.

We separately assess the descriptions in 180 accepted pairs, of which 65.00% satisfy all semantic criteria. Appendix D.2 details these criteria and errors, with broader limitations discussed in Appendix K. The description-pair review assesses the inputs to refinement, while the 90-query assessment evaluates the retrieved images.

Inference cost. The frozen generator adds query-side computation. Figure 6 compares reported latency and accuracy, including generation and retrieval for each pipeline.

`UniCIR` has a reported latency of approximately 0.62 s/query and exceeds default WISER by 2.62 points in CIRCO mAP@5. Default WISER takes about $114\times$ as long. It also improves over OSrCIR by 8.68 points at about $1.29\times$ its speed (Figure 6). LinCIR is faster at 0.01 s/query but has lower accuracy of 19.71. Appendix F.1 compares faster WISER settings and their component costs.

5 Conclusion

In this work, we study universal multimodal embeddings as a foundation for training-free composed image retrieval and introduce `UniCIR` to refine their joint image–edit queries. Its controlled description pair adds contrast for a requested change while retaining the original query. All models and gallery embeddings remain fixed. Experiments on three benchmarks show consistent gains over the corresponding universal baselines across five retrievers, with matched ablations supporting the paired correction. These findings establish strong universal baselines and motivate further research on query construction for composed image retrieval.

AI USE STATEMENT

Generative AI tools assisted with manuscript polishing and code preparation, and provided feedback on research framing, hypotheses, experimental design and result interpretation. They also supported qualitative inspection and annotation conducted by two authors and three student volunteers. The research team conducted the experiments. The authors take responsibility for the final text, claims and artifacts.

ETHICS STATEMENT

This work evaluates retrieval on the existing CIRR, CIRCO and FashionIQ benchmarks. The semantic assessments of 90 queries and 180 description pairs were performed by two authors and three student volunteers. The volunteers gave informed consent. Qualitative figures use benchmark photographs and include failure cases. Generated descriptions may omit or misstate image content, and pretrained models may inherit biases. Applications involving personal or sensitive images require additional privacy and fairness assessment.

REPRODUCIBILITY STATEMENT

Section 3 and Algorithm 1 describe query construction and fixed-gallery retrieval. Evaluation protocols are specified in Section 4.1. Appendix A gives model configurations, generation settings and control definitions. Appendix B reports additional results, and Appendix C provides pair availability and the evaluated parameter grid.

REFERENCES

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? *arXiv preprint arXiv:2605.14787*, 2026.

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 2022a.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4959–4968, June 2022b.

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

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Composed image retrieval using contrastive learning and task-oriented clip-based features. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023b.

Xuanming Cui, Jianpeng Cheng, Hong-you Chen, Satya Narayan Shukla, Abhijeet Awasthi, Xichen Pan, Chaitanya Ahuja, Shlok Kumar Mishra, Yonghuan Yang, Jun Xiao, Qi Guo, Ser-Nam Lim, Aashu Singh, and Xiangjun Fan. Think then embed: Generative context improves multimodal embedding. *arXiv preprint arXiv:2510.05014*, 2025.

Ginger Delmas, Rafael S. Rezende, Gabriela Csurka, and Diane Larlus. ARTEMIS: Attention-based retrieval with text-explicit matching and implicit similarity. In *International Conference on Learning Representations*, 2022.

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

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

Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. In *International Conference on Learning Representations*, 2024.

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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 32063–32077. PMLR, 2025. URL https://proceedings.mlr.press/v267/kwak25a.html.

Taegyeong Lee, Jiwon Park, Seunghyun Hwang, and JooYoung Jang. DEO: Training-free direct embedding optimization for negation-aware retrieval, 2026. URL https://arxiv.org/abs/2603.09185.

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. *arXiv preprint arXiv:2601.04720*, 2026a.

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 *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24421–24442. Association for Computational Linguistics, July 2026b. doi: 10.18653/v1/2026.acl-long.1121. URL https://aclanthology.org/2026.acl-long.1121/.

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

Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2Word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19305–19314, 2023.

Tingyu Song, Mingxin Li, Yanzhao Zhang, Dingkun Long, Chu Liu, Pengjun Xie, Yilun Zhao, and Shu Wu. CORE: Improving compositional reasoning in MLLM embedding via reranker distillation. *arXiv preprint arXiv:2609.04083*, 2026.

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

Zelong Sun, Dong Jing, Guoxing Yang, Nanyi Fei, and Zhiwu Lu. Leveraging large vision-language model as user intent-aware encoder for composed image retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(7):7149–7157, 2025b. doi: 10.1609/aaai.v39i7.32768.

Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Gaopeng Gou, and Qi Wu. Missing target-relevant information prediction with world model for accurate zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24785–24795, June 2025a.

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

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*, pp. 6439–6448, 2019.

Tianyue Wang, Leigang Qu, Tianyu Yang, Xiangzhao Hao, Yifan Xu, Haiyun Guo, and Jinqiao Wang. Wiser: Wider search, deeper thinking, and adaptive fusion for training-free zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16865–16875, June 2026a.

Tong Wang, Yunhan Zhao, and Shu Kong. Generating a paracosm for training-free zero-shot composed image retrieval. In *European Conference on Computer Vision*, 2026b.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. UniIR: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, 2024.

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

Tianyu Yang, ChenWei He, Xiangzhao Hao, Tianyue Wang, Jiarui Guo, Haiyun Guo, Leigang Qu, Jinqiao Wang, and Tat-Seng Chua. ReCALL: Recalibrating capability degradation for mllm-based composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38763–38773, June 2026.

Yongcong Ye, Kai Zhang, Yanghai Zhang, Enhong Chen, Longfei Li, and Jun Zhou. Fine-grained zero-shot composed image retrieval with complementary visual-semantic integration, 2026. URL `https://arxiv.org/abs/2601.14060`.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Bridging modalities: Improving universal multimodal retrieval by multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9274–9285, June 2025.

Zihao Zhang, Dayan Wu, Xinze Liu, Hengjie Zhu, Yiliang Zhu, Ding Wang, Peng Fu, Zheng Lin, and Weiping Wang. MulVec: Fine-grained role-aware matching for training-free zero-shot composed image retrieval. *arXiv preprint arXiv:2608.25305*, 2026.

Jingjiao Zhao, Jiaju Li, Dongze Lian, Liguo Sun, and Pin Lv. DualCIR: Enhancing training-free composed image retrieval via dual-directional descriptions. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 5926–5936, February 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. *arXiv preprint arXiv:2412.14475*, 2024.

A Implementation and Comparison Details

A.1 Inputs, Construction and Availability

The retriever is Qwen3-VL-Embedding-8B. Within CIC, a separate frozen Qwen3-VL-4B-Instruct generates the intended-state description and proposes replacement values. Universal and the generator both receive the original reference image and edit. Target relevance annotations are used only to evaluate the returned rankings. Both auxiliary descriptions use the text-only interface of the retriever, with compatible precomputed gallery vectors. Generation is greedy (`do_sample=False`, `num_beams=1`), with maximum output lengths of 256 tokens for the positive description and 64 for the replacement proposal. The first output describes the intended image, including requested changes and relevant reference context. The second supplies a local alternative to one selected value. The smaller cap reflects this local output format.

We keep the positive description fixed during local replacement. Full uses a pair when it passes the construction checks. The matched controls modify exactly these same queries. All remaining queries use the original Universal embedding and are included in the reported scores. Positive-all uses every available positive description.

Algorithm 1 `UniCIR` query construction and retrieval

Require: Reference I_r, modification T_m, fixed gallery vectors $\{\mathbf{v}_I\}$
Compute $\mathbf{a} \leftarrow \Psi_\theta(I_r, T_m)$ and initialize $\mathbf{q} \leftarrow \mathbf{a}$
Generate intended-state description $c^+ \leftarrow h_\phi(I_r, T_m)$
For an available c^+, identify requested conditions with unique value spans
Prioritize part-level attributes, then use their order in T_m to break ties
if an eligible condition is available **then**
    Select the highest-priority condition and propose one same-type alternative
    Construct c^- by replacing only the selected value span
    if the replacement changes the value and the pair passes the acceptance checks **then**
        Encode $\mathbf{p} \leftarrow \Psi_\theta(c^+)$ and $\mathbf{n} \leftarrow \Psi_\theta(c^-)$
        Compose $\mathbf{q} \leftarrow \mathrm{unit}(\mathbf{a} + 0.50\mathbf{p} - 0.25\mathbf{n})$
    end if
end if
Search $\mathcal{D}_r$ using $\langle \mathbf{q}, \mathbf{v}_I \rangle$

The normalized embedding interface is $\Psi_\theta(x) = f_\theta(x)/\|f_\theta(x)\|_2$ for a nonzero output. All retrieval input modes share θ, and the generator has separate parameters ϕ. The main-text symbols $(I_r, T_m, \mathcal{D})$ correspond to $(r, t, \mathcal{G})$ in the control definitions. Vectors $\mathbf{a}$, $\mathbf{p}$ and $\mathbf{n}$ denote the normalized universal, positive and negative embeddings.

Comparison model configurations. WISER, OSrCIR, CoTMR, LinCIR, PrediCIR and CVSI report ViT-G/14. Benchmark-supervised methods use target-benchmark training annotations. Their configurations are Qwen-VL-Chat for CIR-LVLM, BLIP-2 ViT-L for QuRe and Qwen2.5-VL-7B for ReCALL (Table 14). The complete systems differ in composition, training and inference. Universal and `UniCIR` share the retrieval checkpoint named in each paired row. The published BGE-VL-large results correspond to MMRet-Large in MegaPairs (Zhou et al., 2024), a CLIP-L retriever using image–text feature fusion. It is contrastively trained on external synthetic image–instruction–target triplets and evaluated zero-shot on the CIR benchmarks.

A.2 Query Construction Procedure

Algorithm 1 describes query construction and retrieval. When a description is unavailable or a pair is rejected, the initialized Universal query is used. Rejected proposals are not resampled.

A.3 Control Definitions

Table 6 defines each query direction. Best positive-all (A3) and Double-positive-all (A5-all) correspond to positive and two-positive augmentation in the main text. Table 7 reports the complete CIRR/CIRCO/FashionIQ validation comparisons, including literal-edit and joint-description augmentation.

A3 selects its positive weight independently and uses every available positive description. A5-all is an independent two-positive system. A4 removes the negative while keeping Full's positive text, weight and set of corrected queries. A5 applies to the same queries and replaces the negative with a second positive, using the same input information and token limits.

Negative-only removes the positive term. Shuffled negatives assign each corrected query another query's negative description from the same category. Independent negatives are generated separately from the positives. They can produce usable description pairs for a different set of queries. Literal-edit augmentation uses the original edit without generated text. Joint Direct encodes the reference together with a reference-only observation and the verbatim edit. All query directions are normalized before ranking.

Figure 7 summarizes the anchor and negative-content controls and shows the CIRCO parameter slice. Its active-only anchor panel is evaluated on the common population in Table 8. The other component comparisons use complete validation query sets.

Table 6: Control definitions. Each displayed direction is unit normalized before retrieval. $\mathbf{e}$ denotes the literal edit, $\mathbf{d}$ the joint reference/observation/edit input, and $\mathbf{p}_2$ the second positive description.

ID	Name	Direction
A0	Universal	$\mathbf{a}$
A1	Target description only	$\mathbf{p}$
A2	Universal + literal edit	$\mathbf{a} + \eta\mathbf{e}$
A3	Best positive-all	$\mathbf{a} + \alpha_{\text{pos}}\mathbf{p}$
A4	Positive-pair (same weight)	$\mathbf{a} + \alpha\mathbf{p}$
A5	Double-positive (same queries)	$\mathbf{a} + \eta_1\mathbf{p} + \eta_2\mathbf{p}_2$
A5-all	Double-positive-all	$\mathbf{a} + \eta_1\mathbf{p} + \eta_2\mathbf{p}_2$
A6	Negative-only	$\mathbf{a} - \beta\mathbf{n}$
A7	No universal anchor (active only)	$\alpha\mathbf{p} - \beta\mathbf{n}$
A8	Shuffled negative	$\mathbf{a} + \alpha\mathbf{p} - \beta\mathbf{n}_{\text{shuf}}$
A9	Independent negative	$\mathbf{a} + \alpha\mathbf{p} - \beta\mathbf{n}_{\text{ind}}$
A10	**`UniCIR` (Full)**	$\mathbf{a} + \alpha\mathbf{p} - \beta\mathbf{n}$
A11	Universal + joint Direct	$\mathbf{a} + \eta\mathbf{d}$

Table 7: Complete full-query validation controls. Active-only no-anchor results are separate.

ID	Method	CIRR R@1	CIRCO mAP@5	FIQ R@10
A0	Universal	53.22	31.48	50.41
A1	Target description only	40.85	27.96	48.09
A2	Universal + literal edit	53.67	27.79	50.68
A3	Best positive-all	56.16	33.49	56.46
A4	Positive-pair (same weight)	55.54	32.98	56.12
A5	Double-positive (same queries)	56.49	33.71	56.28
A5-all	Double-positive-all	56.28	35.33	56.74
A6	Negative-only	50.28	26.37	50.55
A8	Shuffled negative	51.76	30.46	52.38
A9	Independent negative	56.23	34.14	55.71
A10	**`UniCIR` (Full)**	57.93	38.25	59.09
A11	Universal + joint Direct	54.17	28.48	51.24

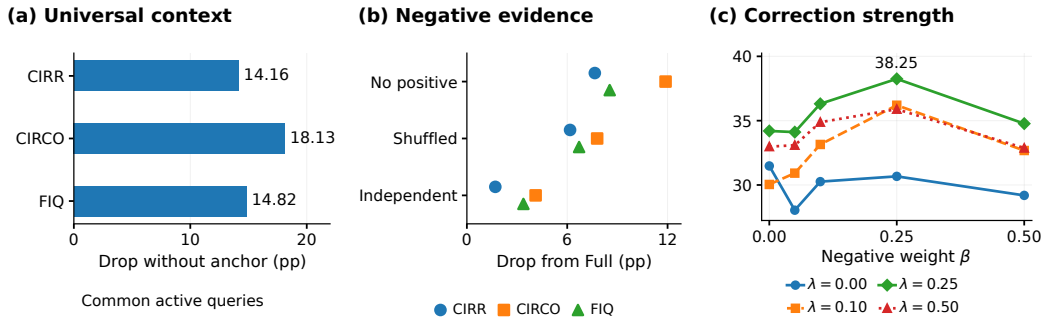


Figure 7: Supplementary component analysis. (a) Anchor removal on the common active subset. (b) Full-query losses relative to `UniCIR` when removing positive evidence, shuffling negatives or generating them independently. (c) CIRCO sensitivity with $\lambda = \alpha - \beta$ fixed on each curve. Markers correspond to evaluated settings. Independent negative generation can change which queries receive a correction.

The active-only comparison evaluates queries for which Full accepts a pair. These subsets contain 2,838 CIRR, 135 CIRCO and 4,184 FashionIQ queries. FashionIQ pools these counts across categories but averages category scores equally. Table 8 compares Full with anchor removal on these subsets. The full-query no-anchor control in Table 3 instead uses Universal on unaccepted queries, reaching 49.89 CIRR test R@1 and 48.78 standard FashionIQ R@10. Its evaluation includes the complete query set and equally weights the FashionIQ categories.

Table 8: Anchor removal on the common accepted-query subset. Δ is Full minus No anchor. FashionIQ averages category scores equally.

Validation metric	N	Full	No anchor	Δ
CIRR R@1	2838	64.02	49.86	14.16
CIRCO mAP@5	135	50.49	32.36	18.13
FIQ Avg. R@10	4184	64.73	49.90	14.82

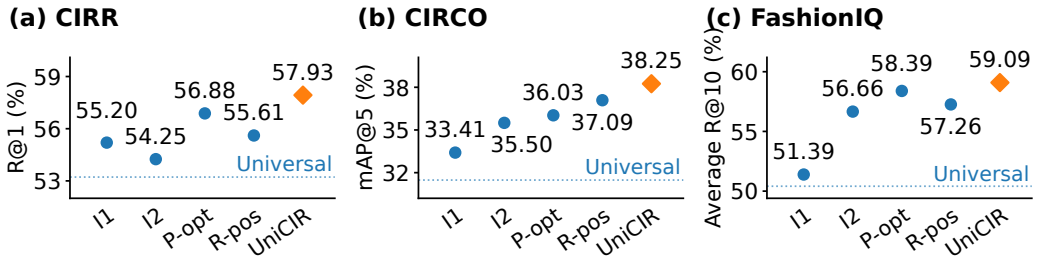


Figure 8: Strong query-construction alternatives on validation. I1 jointly encodes the image and target description. I2 adds the original edit. P-opt tunes positive weighting on all validation queries, modifies the same queries as Full and uses Universal for the rest. R-pos uses a reflective positive. Dotted lines mark Universal. See Table 9 for definitions.

A.4 Additional Input and Positive Controls

Table 9 isolates alternatives to paired refinement. I1 jointly encodes the reference and the same positive description as Full. I2 additionally includes the original edit. Both are standalone joint readouts, and both retain Universal when the positive description is unavailable. Their complete validation denominators match Table 7. These inputs differ from A11, which combines Universal with a separate joint observation–edit representation.

Table 9: Additional input and positive-description controls on full validation. I1 and I2 reuse the same intended-state description as `UniCIR`. P-opt tunes its positive weight separately on each validation dataset while modifying the same queries as Full. R-pos is a reflective-positive adaptation with the same generator. The full-query validation controls are in Table 7.

Query construction	CIRR R@1	CIRCO mAP@5	FIQ R@10
Joint image + target description (I1)	55.20	33.41	51.39
Joint image + edit + description (I2)	54.25	35.50	56.66
Tuned positive-pair (P-opt)	56.88	36.03	58.39
Reflective-positive adaptation (R-pos)	55.61	37.09	57.26
`UniCIR`	**57.93**	**38.25**	**59.09**

P-opt uses the positive from Full on the same pair-available queries and returns Universal elsewhere. The supplementary protocol specifies ten nonnegative candidates,

$$\eta \in \{0.00, 0.05, 0.10, 0.15, 0.20, 0.25, 0.35, 0.50, 0.75, 1.00\}.$$

Selection is by CIRR R@1, CIRCO mAP@5 and equal-category FashionIQ R@10 on the respective validation datasets. FashionIQ uses one weight across all three categories. P-opt reports the best validation aggregate from this expanded search. The fixed-weight no-negative control measures component removal with Full's weight.

R-pos uses the same generator to produce a reflective positive description and independently weights its embedding. This control adapts reflective positive construction to our retriever. Table 9 compares its aggregate performance with Full.

Table 10: Canonical validation. (a) CIRR uses 4,181 queries, 2,297 gallery images and reference exclusion. (b) Fixed Qwen-8B and generated description, with or without the reference image. FashionIQ uses standard galleries and equal-category R@10. Tables 17, 7 and 9 provide full scores.

(a) CIRR validation

Retriever	R@1	R@5	R@10	R@50
Qwen-8B	53.22	82.18	89.62	97.75
Qwen-2B	42.26	74.79	83.43	95.81
GME-7B	41.35	73.91	83.93	96.36
GME-2B	26.31	57.69	70.20	91.58
BGE-VL-large	42.38	73.00	81.73	94.93

(b) Reference-image control

	Description only	Image + description
CIRR R@1	40.85	55.20
CIRCO mAP@5	27.96	33.41
FashionIQ R@10	48.09	51.39

Table 11: Same-retriever gains over Universal. CIRR/CIRCO use test and FashionIQ uses standard validation. Universal and `UniCIR` columns report absolute scores. Δ columns give percentage-point differences of the reported scores. $*$ denotes our matched evaluations. † marks external synthetic-triplet pretraining.

	CIRR test R@1			CIRCO test mAP@5		
Retriever	Universal	`UniCIR`	Δ	Universal	`UniCIR`	Δ
Qwen-8B*	53.19	59.50	+6.31	31.26	39.15	+7.89
Qwen-2B*	42.33	47.19	+4.86	24.96	30.28	+5.32
GME-7B*	41.31	45.80	+4.49	24.68	30.74	+6.06
GME-2B*	26.21	29.92	+3.71	16.35	20.76	+4.41
BGE-VL-large*†	42.41	45.99	+3.58	22.03	26.51	+4.48

	FashionIQ val average R@10			FashionIQ val average R@50		
Retriever	Universal	`UniCIR`	Δ	Universal	`UniCIR`	Δ
Qwen-8B*	50.41	59.09	+8.68	70.81	78.77	+7.96
Qwen-2B*	43.88	48.11	+4.23	63.83	69.75	+5.92
GME-7B*	47.81	52.85	+5.04	68.11	72.59	+4.48
GME-2B*	35.42	40.09	+4.67	55.50	60.88	+5.38
BGE-VL-large*†	34.34	38.72	+4.38	55.40	59.36	+3.96

B Additional Benchmark and Transfer Results

CIRR subset comparison. The subset scores compare ranking within curated CIRR candidates. Qwen-8B reaches 79.74/91.14/95.09 on subset R@1/2/3, while `UniCIR` reaches 83.32/94.74/98.47 (Table 1). Keeping the retriever fixed yields gains of 3.58/3.60/3.38 points. Against ReCALL, the margins are 1.83/1.39/0.83 points. Best positive-all scores 0.80 points higher at subset R@1. `UniCIR` exceeds it by 0.86/1.45 points at subset R@2/3 (Table 16).

Table 11 gives the five matched Universal/`UniCIR` comparisons used in Figure 4. Tables 12 and 13 provide additional cutoffs and category results, alongside published baselines. Starred rows identify our matched evaluations. Published-method CIRR subset scores in Table 1 follow the corresponding reported configurations in the published comparisons (Wang et al., 2026a; Tang et al., 2025b; Sun et al., 2025a; Gu et al., 2024; Ye et al., 2026; Sun et al., 2025b; Kwak et al., 2025; Yang et al., 2026). ViT-G/14 denotes the OpenCLIP retrieval backbone. PrediCIR subset entries remain unreported (Tang et al., 2025a).

Table 14 provides additional published-method comparisons with separate training categories and evaluation splits. The CLIP-based supervised references include Combiner and CLIP4Cir (Baldrati et al., 2022a;b; 2023b). TEMA's FashionIQ results are included, while its CIRR validation score is excluded from the test column. The Pic2Word and SEARLE-G entries use the retrainings reported by the LinCIR authors. Dashes denote unreported scores.

Table 15 gives the category-level positive and two-positive baselines. Table 16 evaluates CIRR subset retrieval, where candidates are restricted to a visually related group. This panel compares `UniCIR` with its Qwen-8B baseline and Best positive-all, and includes GME-7B as a separate reference.

Table 12: CIRR/CIRCO test cutoffs, with CIRR-only supervised methods below. Other `UniCIR` backbones report only their primary metrics in Table 11. $*$ denotes matched evaluations and † external synthetic-triplet pretraining. Published BGE-VL-large is MMRet-Large. Bold/underline mark best/second-best scores. Dashes mark unreported scores.

	CIRR test				CIRCO test			
Method	R@1	R@5	R@10	R@50	mAP@5	mAP@10	mAP@25	mAP@50
Training-free								
OSrCIR	37.26	67.25	77.33	—	30.47	31.14	35.03	36.59
CoTMR	36.36	67.52	77.82	93.99	32.23	32.72	35.60	36.83
WISER	49.54	77.40	85.76	94.17	36.53	38.14	40.46	41.26
Trained zero-shot								
LinCIR	35.25	64.72	76.05	—	19.71	21.01	23.13	24.18
PrediCIR	37.00	66.10	77.90	—	23.70	24.60	25.40	26.00
CVSI	41.37	72.58	82.39	—	33.51	34.96	37.86	38.98
BGE-VL-large†	38.00	70.30	81.10	94.70	**39.20**	<u>40.20</u>	<u>42.90</u>	<u>44.00</u>
Matched retriever comparisons								
Qwen-2B*	42.33	74.84	83.50	95.86	24.96	26.49	29.12	30.14
GME-7B*	41.31	73.91	83.93	96.36	24.68	26.07	29.31	30.35
GME-2B*	26.21	57.62	70.15	91.53	16.35	18.46	20.66	21.69
BGE-VL-large*†	42.41	73.04	81.75	94.95	22.03	23.07	25.21	26.02
Qwen-8B*	53.19	82.35	89.74	—	31.26	29.41	32.06	33.20
`UniCIR`	**59.50**	**88.21**	**94.27**	—	<u>39.15</u>	**40.73**	**43.02**	**44.37**

Benchmark-supervised. CIRR test only				
Method	R@1	R@5	R@10	R@50
CIR-LVLM	53.64	83.76	90.60	97.93
QuRe	52.22	82.53	90.31	<u>98.17</u>
ReCALL	<u>55.52</u>	<u>84.07</u>	<u>91.83</u>	**98.55**

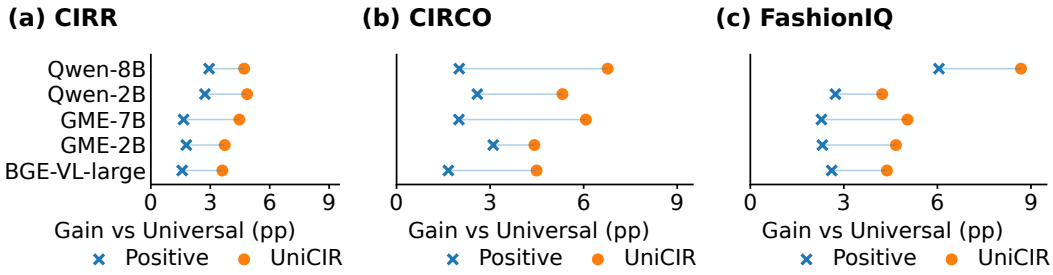


Figure 9: Retriever transfer on validation. Texts and weights are reused, with each retriever encoding the texts in its own space. Crosses and circles compare positive augmentation and Full against its Universal baseline. Absolute scores and FashionIQ R@50 are in Table 17.

Table 17 compares Full with positive augmentation on validation, including FashionIQ R@50. Figure 9 shows the gains. The CIRR/CIRCO splits differ from the test sets in Table 12. We reuse the CIC descriptions and the Qwen-8B refinement configuration across retrievers, each with its compatible gallery. These retrievers differ in pretraining, model size and embedding dimension.

CIRCO Universal uses 31.26 mAP@5 on test and 31.48 on canonical validation. Its test mAP@10/25/50 values are 29.41/32.06/33.20.

C Coverage and Parameter Sensitivity

Table 18 reports 10,105 available positive descriptions and 7,157 accepted pairs among 10,417 validation queries. Accepted pair, unsupported slot, rejected pair and unavailable

Table 13: FashionIQ standard validation with equal-category averages. Other `UniCIR` backbones have reported averages in Table 11 and unreported category scores. $*$ denotes matched evaluations and † external synthetic-triplet pretraining. Published BGE-VL-large is MMRet-Large. Bold/underline mark best/second-best scores. Appendix G details numerical conventions.

	Shirt		Dress		Toptee		Average	
Method	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
Training-free								
OSrCIR	38.65	54.71	33.02	54.78	41.04	61.83	37.57	57.11
CoTMR	38.32	62.24	34.51	57.36	41.90	64.30	38.25	61.32
WISER	45.98	61.58	38.57	58.35	49.21	66.96	44.59	62.30
Trained zero-shot								
LinCIR	46.76	65.11	38.08	60.88	50.48	71.09	45.11	65.69
PrediCIR	48.20	67.40	39.70	62.40	53.70	73.60	47.20	67.80
BGE-VL-large†	37.00	56.10	29.70	50.30	37.00	59.30	34.60	55.20
Benchmark-supervised								
CIR-LVLM	<u>58.59</u>	75.86	50.42	<u>73.57</u>	59.61	78.99	56.21	76.14
QuRe	53.53	72.87	46.80	69.81	57.47	77.77	52.60	73.48
ReCALL	58.49	<u>76.59</u>	<u>51.81</u>	73.48	<u>60.83</u>	<u>79.19</u>	<u>57.04</u>	<u>76.42</u>
TEMA	52.90	73.55	49.66	71.98	56.49	77.07	53.02	74.20
Matched retriever comparisons								
Qwen-2B*	46.07	64.28	39.27	61.23	46.30	65.99	43.88	63.83
GME-7B*	52.21	69.63	39.71	63.41	51.50	71.29	47.81	68.11
GME-2B*	42.15	60.60	26.57	47.10	37.53	58.80	35.42	55.50
BGE-VL-large*†	36.06	56.13	29.90	51.12	37.07	58.95	34.34	55.40
Qwen-8B*	51.77	70.66	45.96	67.77	53.49	73.99	50.41	70.81
`UniCIR`	**60.99**	**79.05**	**53.45**	**75.26**	**62.83**	**82.00**	**59.09**	**78.77**

Table 14: Published results for 15 reference methods. Panel (a) includes all three benchmarks. Panel (b) reports FashionIQ and CIRR results from benchmark-supervised methods, whose CIRCO scores are unreported. Rows give the reported model and are grouped by training regime.

(a) Training-free and trained zero-shot

		FashionIQ val		CIRR test	CIRCO test
Method	Reported model	R@10	R@50	R@1	mAP@5
Training-free					
CIReVL	ViT-G/14	32.19	52.36	34.65	26.77
OSrCIR	ViT-G/14	37.57	57.11	37.26	30.47
CoTMR	ViT-G/14	38.25	61.32	36.36	32.23
WISER	ViT-G/14	44.59	62.30	49.54	36.53
Trained zero-shot					
Pic2Word	ViT-G/14	31.28	51.89	30.41	5.54
LinCIR	ViT-G/14	45.11	65.69	35.25	19.71
SEARLE-G	ViT-G/14	34.81	55.71	34.80	13.20
PrediCIR	ViT-G/14	47.20	67.80	37.00	23.70
CVSI	ViT-G/14	44.97	65.72	41.37	33.51

(b) Benchmark-supervised

		FashionIQ val		CIRR test
Method	Reported model	R@10	R@50	R@1
CLIP4Cir	RN50x4	43.78	66.93	44.82
Combiner + partial FT	RN50x4	38.32	61.74	38.53
CIR-LVLM	Qwen-VL-Chat	56.21	76.14	53.64
QuRe	BLIP-2 ViT-L	52.60	73.48	52.22
ReCALL	Qwen2.5-VL-7B	57.04	76.42	55.52
TEMA	BLIP	53.02	74.20	—

positive are mutually exclusive construction outcomes within each dataset. We assess semantic quality separately in Appendix D.2.

Table 15: FashionIQ positive-description baselines on standard validation.

	Best positive-all		Double-positive-all	
Category	R@10	R@50	R@10	R@50
Shirt	58.44	76.45	58.54	77.58
Dress	51.31	73.18	51.66	73.28
Toptee	59.61	79.25	60.02	80.57
Average	56.46	76.29	56.74	77.14

Table 16: CIRR subset retrieval on the test benchmark. Candidates are restricted to visually related groups. Positive augmentation and UniCIR use Qwen-8B.

Method	$R_{\text{subset}}@1$	$R_{\text{subset}}@2$	$R_{\text{subset}}@3$
GME-7B	68.22	85.06	93.07
Qwen-8B	79.74	91.14	95.09
Best positive-all (Qwen-8B)	84.12	93.88	97.02
`UniCIR`	83.32	94.74	98.47
`UniCIR` − Qwen-8B (pp)	3.58	3.60	3.38
`UniCIR` − Best positive-all (pp)	-0.80	0.86	1.45

Table 17: Retriever transfer on validation using shared descriptions and refinement weights. FashionIQ uses standard galleries and equal-category averages.

Retriever	Val. metric	Universal	Positive	Full	Δ Universal	Δ Positive
Qwen-8B	CIRR R@1	53.22	56.16	57.93	4.71	1.77
Qwen-8B	CIRCO mAP@5	31.48	33.49	38.25	6.77	4.76
Qwen-8B	FIQ std val R@10	50.41	56.46	59.09	8.68	2.63
Qwen-8B	FIQ std val R@50	70.81	76.29	78.77	7.96	2.48
Qwen-2B	CIRR R@1	42.26	44.99	47.12	4.86	2.13
Qwen-2B	CIRCO mAP@5	25.02	27.61	30.34	5.32	2.73
Qwen-2B	FIQ std val R@10	43.88	46.61	48.11	4.23	1.50
Qwen-2B	FIQ std val R@50	63.83	67.12	69.75	5.92	2.63
GME-7B	CIRR R@1	41.35	43.00	45.83	4.47	2.82
GME-7B	CIRCO mAP@5	24.64	26.64	30.71	6.07	4.07
GME-7B	FIQ std val R@10	47.81	50.09	52.85	5.04	2.76
GME-7B	FIQ std val R@50	68.11	70.91	72.59	4.48	1.68
GME-2B	CIRR R@1	26.31	28.10	30.04	3.73	1.94
GME-2B	CIRCO mAP@5	16.29	19.39	20.71	4.42	1.32
GME-2B	FIQ std val R@10	35.42	37.73	40.09	4.67	2.36
GME-2B	FIQ std val R@50	55.50	57.92	60.88	5.38	2.96
BGE-VL-large	CIRR R@1	42.38	43.96	45.99	3.61	2.03
BGE-VL-large	CIRCO mAP@5	22.05	23.71	26.54	4.49	2.83
BGE-VL-large	FIQ std val R@10	34.34	36.95	38.72	4.38	1.77
BGE-VL-large	FIQ std val R@50	55.40	57.83	59.36	3.96	1.53

Table 18: Description and pair availability on validation. The four construction outcomes partition each dataset. No-pair fallback remains in the full-query denominator.

	CIRR	CIRCO	Shirt	Dress	Toptee
Queries	4181	220	2038	2017	1961
Positive available	4021	215	2007	1953	1909
Valid pair	2838	135	1498	1329	1357
Pair coverage (%)	67.88	61.36	73.50	65.89	69.20
No supported slot	747	60	345	424	402
Pair check rejected	436	20	164	200	150
Positive unavailable	160	5	31	64	52

C.1 Correction Settings

The reported configuration uses $\alpha = 0.50, \beta = 0.25$. The coefficient of the universal vector $\mathbf{a}$ is fixed at one, and the two auxiliary coefficients specify their weights relative to it. Multiplying all three coefficients by the same positive constant leaves the normalized query unchanged. We express the sensitivity sweep as $\lambda = \alpha - \beta$ and $\mu = \beta$. Figure 10 shows the 20 settings with $\beta \leq 0.50$. Table 19 lists all 21 evaluated settings. The default attains the highest CIRCO and FashionIQ scores in this grid and ties the best CIRR score. A curve fixes λ while varying μ, which changes both original coefficients. This sweep differs from removing the negative at fixed α.

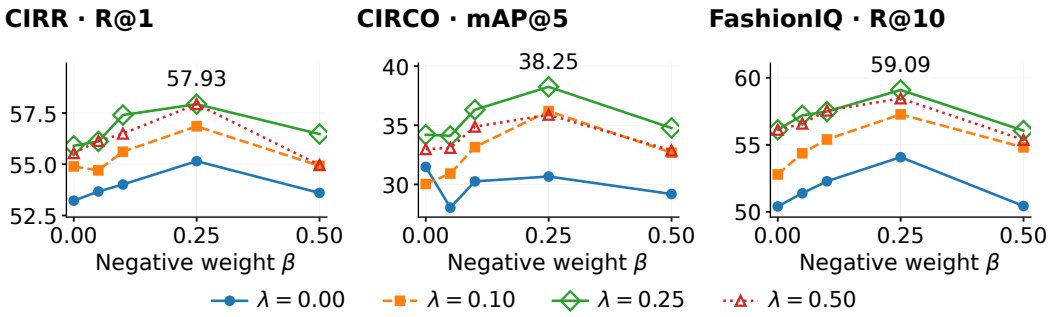


Figure 10: Parameter sensitivity on canonical validation. Curves fix $\lambda = \alpha - \beta$ and vary $\mu = \beta$. Points show the 20 settings with $\beta \leq 0.50$. Table 19 also includes the $\beta = 1.00$ setting. Open markers distinguish coincident values.

Table 19: Full parameter grid (21 settings). $\lambda = \alpha - \beta$ and $\mu = \beta$ re-express the same weighted query.

λ	μ	α	β	CIRR val R@1	CIRCO val mAP@5	FIQ val R@10
0.00	0.00	0.00	0.00	53.22	31.48	50.41
0.00	0.05	0.05	0.05	53.67	28.04	51.39
0.00	0.10	0.10	0.10	54.01	30.26	52.28
0.00	0.25	0.25	0.25	55.15	30.67	54.08
0.00	0.50	0.50	0.50	53.60	29.19	50.44
0.10	0.00	0.10	0.00	54.89	30.04	52.79
0.10	0.05	0.15	0.05	54.70	30.92	54.37
0.10	0.10	0.20	0.10	55.61	33.15	55.40
0.10	0.25	0.35	0.25	56.85	36.20	57.27
0.10	0.50	0.60	0.50	54.92	32.68	54.81
0.25	0.00	0.25	0.00	55.90	34.20	56.12
0.25	0.05	0.30	0.05	56.11	34.11	57.20
0.25	0.10	0.35	0.10	57.40	36.31	57.48
0.25	0.25	0.50	0.25	57.93	38.25	59.09
0.25	0.50	0.75	0.50	56.47	34.77	56.06
0.50	0.00	0.50	0.00	55.54	32.98	56.12
0.50	0.05	0.55	0.05	56.13	33.09	56.58
0.50	0.10	0.60	0.10	56.49	34.88	57.55
0.50	0.25	0.75	0.25	57.93	35.89	58.46
0.50	0.50	1.00	0.50	54.96	32.87	55.37
-0.50	1.00	0.50	1.00	50.97	26.51	51.97

At $\lambda = 0.25$, increasing μ from 0.25 to 0.50 lowers CIRCO mAP@5 from 38.25 to 34.77 and reduces performance on the other two datasets. A positive-only point at $\alpha = 0.25$ obtains 34.20 on CIRCO, compared with 32.98 for the same-weight positive-pair ablation at $\alpha = 0.50$. The latter removes the negative at Full's fixed positive weight, whereas the former uses a different positive weight.

C.2 Normalization and Score Equivalence

When $\mathbf{p} = \mathbf{n}$, the paired difference vanishes and the default query reduces algebraically to $\mathrm{unit}(\mathbf{a} + 0.25\mathbf{p})$. Equation (4) in the main Method expands the composed-query score. The positive denominator is independent of the candidate and preserves the numerator's ranking. For unit input vectors, the reverse triangle inequality gives $\|\widetilde{\mathbf{q}}\|_2 \geq 1 - \alpha - \beta$. At the working point, $\|\widetilde{\mathbf{q}}\|_2 \geq 0.25$, ensuring that normalization is well-defined in exact arithmetic.

C.3 Relation to Signed Query Refinement

DualCIR and CoTMR combine caption and attribute or object scores linearly (Zhao et al., 2025; Sun et al., 2025a). When the text embeddings and weights are fixed, this scoring rule can also use one query vector in the shared gallery space.

For one positive and one negative embedding, the squared-distance objective of DEO (Lee et al., 2026) takes the form

$$L(\mathbf{u}) = \lambda_p \|\mathbf{u} - \mathbf{p}\|^2 - \lambda_n \|\mathbf{u} - \mathbf{n}\|^2 + \lambda_o \|\mathbf{u} - \mathbf{a}\|^2. \tag{5}$$

For $\gamma = \lambda_o + \lambda_p - \lambda_n > 0$, expanding this unconstrained quadratic gives the unique minimizer

$$\mathbf{u}^* = \frac{\lambda_o \mathbf{a} + \lambda_p \mathbf{p} - \lambda_n \mathbf{n}}{\gamma}. \tag{6}$$

With $\lambda_o = 1$, $\lambda_p = \alpha$ and $\lambda_n = \beta$, the normalized direction agrees with Equation (3) whenever $\mathbf{u}^* \neq \mathbf{0}$. DEO uses finite-step Adam optimization. The expression above describes the exact minimum of this special case. In `UniCIR`, the construction of c^+ and c^- supplies the controlled contrast, while $\mathbf{a}$ jointly encodes the original image and modification.

D Gains, Losses and Semantic Assessment

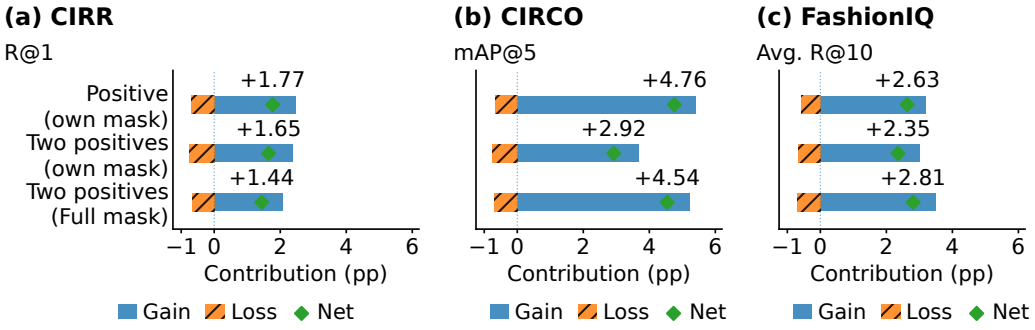


Figure 11: Retrieval gains and losses. Bars show positive and negative utility contributions, and diamonds show reported net gains. All rows use complete validation sets and FashionIQ category weights. The first two use each baseline's own mask, and the last shares Full's mask. Appendix D and Table 20 provide definitions and counts.

Figure 11 visualizes three positive-baseline comparisons from Table 20, which also includes the Universal comparison. Retrieval improvements and new losses are aggregated over the canonical validation queries with available relevance annotations. For query i and a control B, let $d_i = u_{\mathrm{Full}}(i) - u_B(i)$, where u is the R@1 indicator for CIRR, AP@5 for CIRCO, or the R@10 indicator for FashionIQ, each on the $[0, 1]$ scale. The utility contributions are

$$G^+ = 100 \sum_i \omega_i \max(d_i, 0), \qquad G^- = 100 \sum_i \omega_i \max(-d_i, 0).$$

CIRR and CIRCO use $\omega_i = 1/N$. For FashionIQ, a query in category c has weight $\omega_i = 1/(3N_c)$, retaining the equal-category mean. The net change is $G^+ - G^-$. Improved and worsened counts

Table 20: Retrieval repairs and losses on canonical validation. G^+ and G^- measure utility contributions in percentage points. FashionIQ uses equal-category weights. The CIRCO Full–Universal comparison reports a net gain of 6.77 points without its component decomposition.

Metric	Improved	Worsened	G^+	G^-	Net
Full – Universal					
CIRR R@1	240	43	5.74	1.03	4.71
FIQ std val R@10	590	68	9.81	1.13	8.68
Full – Best positive-all					
CIRR R@1	103	29	2.46	0.69	1.77
CIRCO mAP@5	52	18	5.41	0.65	4.76
FIQ std val R@10	192	34	3.20	0.57	2.63
Full – Double-positive (matched mask)					
CIRR R@1	87	27	2.08	0.65	1.44
CIRCO mAP@5	41	19	5.24	0.70	4.54
FIQ std val R@10	211	42	3.51	0.70	2.81
Full – Double-positive-all (independent)					
CIRR R@1	100	31	2.39	0.74	1.65
CIRCO mAP@5	26	9	3.67	0.75	2.92
FIQ std val R@10	181	40	3.01	0.66	2.35

record queries with $d_i > 0$ and $d_i < 0$, respectively. For CIRCO Full–Universal, only the net gain of 6.77 points is reported (Table 20).

Figure 12 expands the edit-type comparison in Figure 14(a). Tables 21, 22 and 23 report group sizes, availability and increments for overlapping edit groups. Queries with unsupported conditions use Universal in both comparisons.

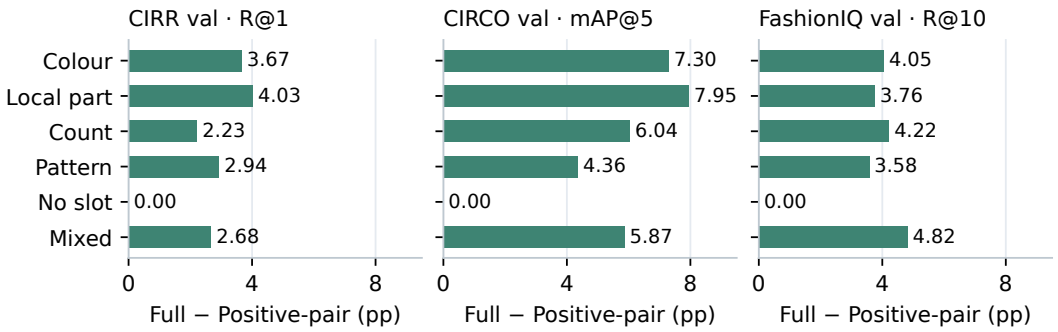


Figure 12: Full minus the same-mask positive control by edit type. The three panels use CIRR R@1, CIRCO AP@5 and category-macro FashionIQ R@10. Groups overlap. Unsupported queries retain Universal in both arms.

Table 21: CIRR validation results by edit type. U denotes Universal. Groups overlap, and queries with unsupported conditions use Universal in both comparisons.

Group	N	Pair %	P-pair−U	Full−U	Full−P-pair
Colour / brightness	1008	81.94	3.08	6.75	3.67
Local part / binding	769	68.92	3.38	7.41	4.03
Object count	224	54.91	2.23	4.46	2.23
Pattern / regularity	306	48.69	2.61	5.56	2.94
Spatial / action (no slot)	498	0.00	0.00	0.00	0.00
Mixed edit	1530	45.95	2.48	5.16	2.68

Table 22: CIRCO validation gain over the matched positive control by edit type. Groups overlap. Unsupported queries use Universal in both arms. Their two Universal-relative increments are both 0.00.

Group	N	Pair %	Full−P-pair
Colour / brightness	51	80.39	7.30
Local part / binding	26	53.85	7.95
Object count	31	61.29	6.04
Pattern / regularity	8	50.00	4.36
Spatial / action (no slot)	42	0.00	0.00
Mixed edit	146	63.70	5.87

Table 23: FashionIQ validation results by edit type, with equal-category weighting. U denotes Universal. Groups overlap, and queries with unsupported conditions use Universal in both comparisons.

Group	N	Pair %	P-pair−U	Full−U	Full−P-pair
Colour / brightness	3892	71.15	5.54	9.59	4.05
Local part / binding	1808	84.68	5.38	9.14	3.76
Object count	240	51.25	2.55	6.77	4.22
Pattern / regularity	1326	70.59	5.43	9.01	3.58
Spatial / action (no slot)	558	0.00	0.00	0.00	0.00
Mixed edit	1305	53.49	4.20	9.02	4.82

D.1 Same-Query Edit Adherence Assessment

The H1 and H2 semantic assessments were performed by two authors and three student volunteers. The H1 assessment evaluates Universal, positive augmentation and Full on the same 90 queries. Each candidate is judged for the requested change and the reference details needed for the target. Joint yes requires both criteria to be yes. Rates include all 90 queries, and joint uncertain counts are reported separately (Table 24). Universal has 60/73/45 yes decisions for change/preservation/joint. Positive augmentation has 59/65/50, and Full has 69/66/59. Joint uncertain counts are 8/9/8.

Table 24: Edit adherence on the same 90 queries for all three methods. Joint success requires both change and preservation. Rates include uncertain judgments in the denominator.

Method	N	Change yes	Preserve yes	Joint yes	Joint uncertain
Universal	90	60	73	45	8
Positive augmentation	90	59	65	50	9
`UniCIR`	90	69	66	59	8

Full improves joint adherence over Universal by $100(59-45)/90 = 15.56$ points, while preservation changes by $100(66-73)/90 = -7.78$ points. Joint improvement therefore accompanies a preservation trade-off. The analysis uses marginal and joint counts within each method.

A separate two-arm pilot yielded 54/90 and 63/90 joint successes. Table 24 presents the three-arm assessment used here. The judgments concern the query and returned image. Benchmark test relevance labels are hidden.

D.2 Semantic Quality of CIC Outputs

We assess whether accepted description pairs express a valid contrast for the current request (Figure 13). This assessment concerns the generated text, separately from the returned-image adherence evaluated above.

H2 assesses 180 accepted pairs separately from H1. The criteria in Table 25 cover the completeness of the positive (A), avoidance of stale source attributes (B), edit linkage of the selected condition (L), validity of the local contrast (D), violation of the requested condition (V),

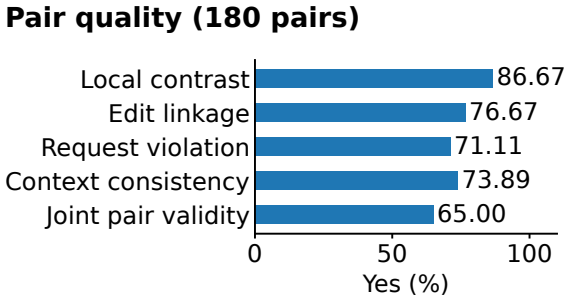


Figure 13: Semantic quality of 180 accepted pairs. Joint validity also requires positive completeness and stale-attribute avoidance. Every criterion uses all 180 pairs. This sample is separate from the 90-query returned-image assessment.

and consistency with unmodified context (C). Context consistency concerns the parts of the description outside the selected condition.

Table 25: Semantic quality of 180 accepted description pairs (H2). Each yes rate uses all 180 pairs. The request-violation event requires both edit linkage and a valid local contrast. Joint validity requires every listed criterion.

Code	Criterion	Yes	Rate (%)
A	Positive contains all explicit changes	151	83.89
B	Positive avoids stale source attributes	145	80.56
L	Selected condition is linked to the edit	138	76.67
D	Replacement makes a valid local contrast	156	86.67
V	Replacement violates the requested condition	128	71.11
C	Negative remains consistent with its context	133	73.89
Q	Joint pair validity	117	65.00

The request-violation judgment requires both edit linkage and a valid local contrast, giving $V \subseteq L \cap D$. Joint validity is $Q = A \cap B \cap L \cap D \cap V \cap C$. Every yes rate uses all 180 pairs. Local contrast is valid for 156 pairs, but only 128 meet the request-violation criterion and 117 satisfy every requirement. Thus, a valid local substitution may still fail to express the requested contrast or preserve the surrounding context.

Of the 156 locally valid contrasts, 28 lack a confirmed requested violation. A further 11 satisfy the violation criterion but fail another joint requirement.

A separate sample of 90 available positive descriptions evaluates change completeness and erroneous retention (Table 26). Its rates use that sample's denominator, independently of the 180-pair assessment.

Table 26: Separate assessment of 90 available positive descriptions. This sample is distinct from the 180 accepted pairs in Table 25.

Criterion	Yes / N	Rate (%)
Contains all explicit requested changes	72/90	80.00
Avoids incorrect retention of source attributes	83/90	92.22

D.3 Choosing the Contrasted Condition

C1 compares the edit-linked condition with another admissible context-linked condition in the same positive text. The retriever, weights and local replacement procedure are held fixed, and both arms are evaluated on their common accepted queries. Table 27 reports edit-linked minus context-linked

differences. Edit-linked selection gains 0.30/0.05/1.53 points on CIRR/CIRCO/FashionIQ. The small CIRR and CIRCO differences leave context-linked selection competitive on these datasets.

Table 27: Edit-linked versus context-linked condition selection on queries accepted by both constructions. Values are edit-linked minus context-linked scores, with the positive description and refinement weights fixed.

Comparison	CIRR R@1	CIRCO mAP@5	FIQ R@10
Edit-linked − context-linked (pp)	+0.30	+0.05	+1.53

Figure 14 places the edit-group gains alongside positive and negative retrieval changes under the full validation protocol.

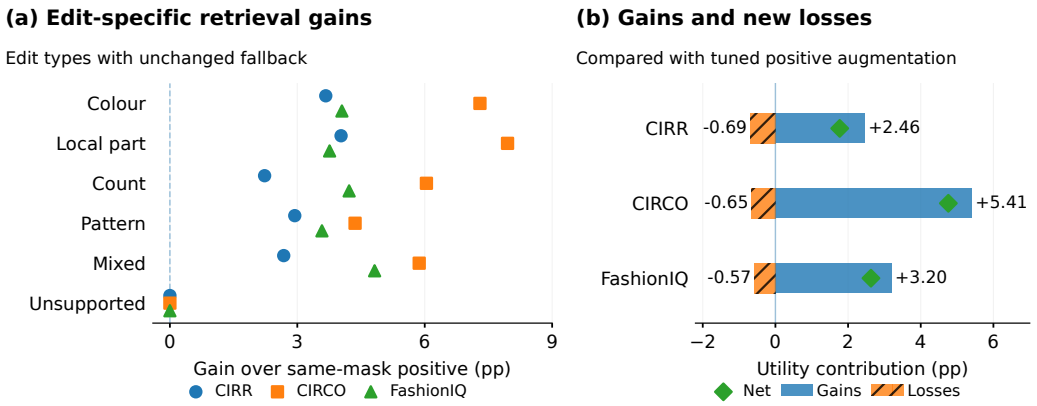


Figure 14: Retrieval changes on full validation. (a) Gain over the same-mask positive control by overlapping edit type. (b) Positive and negative utility contributions relative to independent positive augmentation. These ranking diagnostics complement the direct semantic assessment in Figure 5.

E Generator and Proposal Variants

Edit types and generators. Figure 14(a) tests whether gains extend beyond colour edits using Full's same-mask positive control. Parts, counts, patterns and mixed edits also improve, including 7.95 points for CIRCO local parts and 4.82 for FashionIQ mixed edits. The groups overlap. Figure 15 compares the generator variants detailed in Table 28.

Table 28 changes the generator while fixing the Qwen-8B retriever and gallery. Each generator supplies its own descriptions for positive augmentation and Full. Qwen3-VL-2B-Instruct tests a smaller model from the same family, and MiniCPM-V-4.0 tests another family. The model repositories are `Qwen/Qwen3-VL-2B-Instruct`[1] and `openbmb/MiniCPM-V-4`[2].

Full exceeds the corresponding positive baseline by 0.74/2.95/1.42 points for Qwen3-VL-2B-Instruct and 2.22/4.42/2.42 for MiniCPM-V-4.0 on CIRR/CIRCO/FashionIQ. With Qwen3-VL-2B-Instruct, Full reaches 54.14 on FashionIQ, compared with 59.09 under the default generator. Table 29 reports pair coverage by category.

G3 replaces the generator's alternative-value proposal with a deterministic choice among admissible values. The target description, selected condition, set of corrected queries, retriever and refinement weights are fixed. G3 changes CIRR/CIRCO/FashionIQ scores by $+0.19/-0.59/-0.11$ points relative to Full (Table 30). Rule-based proposals remain competitive when applied to the same queries.

[1] `https://huggingface.co/Qwen/Qwen3-VL-2B-Instruct`
[2] `https://huggingface.co/openbmb/MiniCPM-V-4`

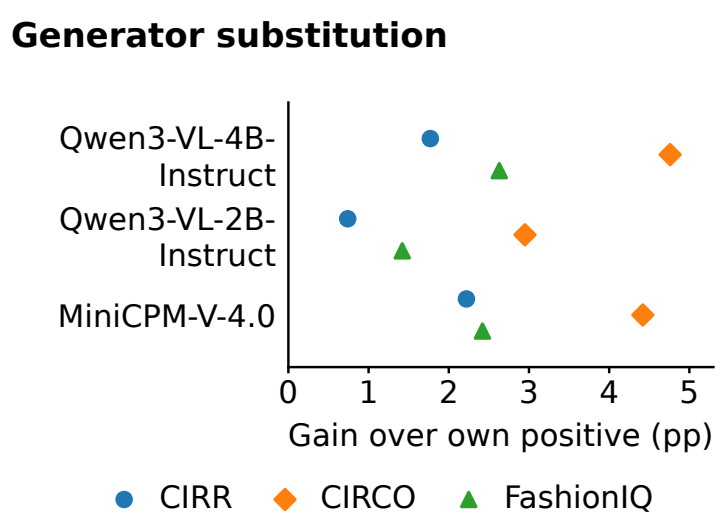


Figure 15: Generator substitution with the Qwen-8B retriever fixed. Markers show Full minus each generator's own positive augmentation. Qwen3-VL-2B-Instruct and MiniCPM-V-4.0 replace the default Qwen3-VL-4B-Instruct. All results use validation splits and the metrics in Section 4.1. Absolute scores are in Table 28.

Table 28: Generator variants on the CIRR, CIRCO, and FashionIQ validation splits, measured by R@1, mAP@5, and standard R@10, respectively. The Qwen3-VL-Embedding-8B retriever and gallery are fixed. Each generator is compared with its own positive augmentation.

Generator	Method	CIRR R@1	CIRCO mAP@5	FIQ R@10
Qwen3-VL-4B-Instruct	Positive	56.16	33.49	56.46
	Full	57.93	38.25	59.09
Qwen3-VL-2B-Instruct	Positive	55.68	30.97	52.72
	Full	56.42	33.92	54.14
MiniCPM-V-4.0	Positive	55.42	32.42	55.20
	Full	57.64	36.84	57.62

Table 29: Accepted-pair coverage (%) on the CIRR, CIRCO, and FashionIQ validation splits under generator substitution, with the Qwen3-VL-Embedding-8B retriever and gallery fixed. Each FashionIQ category is reported separately. Coverage is the percentage of queries with a pair that passes the construction checks.

Generator	CIRR	CIRCO	Shirt	Dress	Toptee
Qwen3-VL-4B-Instruct	67.88	61.36	73.50	65.89	69.20
Qwen3-VL-2B-Instruct	51.47	56.36	70.36	54.73	57.78
MiniCPM-V-4.0	65.46	54.55	53.83	75.26	54.11

Table 30: Replacing the generated proposal with a deterministic admissible-value rule (G3). The target description, selected condition, set of corrected queries, retriever and refinement weights are held fixed for this content comparison. Differences use the displayed aggregate precision.

Proposal strategy	CIRR R@1	CIRCO mAP@5	FIQ R@10
Generated proposal (Full)	57.93	38.25	59.09
Rule-based proposal (G3)	58.12	37.66	58.98
G3 − Full (pp)	+0.19	−0.59	−0.11

The main benchmarks use the generated-proposal configuration.

F Inference-Time Breakdown

F.1 WISER Configurations and Component Costs

To examine the efficiency trade-off beyond WISER's default configuration, Table 31 reports three configurations from an earlier measurement series. The WISER settings share a single-GPU environment and batch size, using ViT-bigG/14 retrieval on CIRCO. Times are rounded averages over repeated runs, with conditional refinement averaged over all profiled queries.

Table 31: Previously measured WISER configurations. Times are seconds per query. Conditional refinement is averaged over all profiled queries. K is the retrieval depth of each branch.

WISER configuration	Editing	Verification	Refinement	Total	mAP@5
SDXL-Turbo, $K = 10$	0.2	8.3	2.9	11.4	29.47
BAGEL, 10 steps, $K = 50$	7.1	24.6	9.4	41.1	32.91
BAGEL, default, $K = 50$	35.7	21.3	13.6	70.6	36.53

The default WISER pipeline spends 35.7 s on generative editing and a further 34.9 s on candidate verification and conditional refinement. Reducing BAGEL to ten diffusion steps lowers latency to 41.1 s and mAP@5 to 32.91. The SDXL-Turbo configuration also reduces retrieval depth from 50 to 10 candidates per branch. It reaches 11.4 s/query and 29.47 mAP@5, with 11.2 s spent on verification and refinement. These settings show how faster editing and a smaller candidate pool change both latency and retrieval accuracy.

`UniCIR` constructs a single query vector before searching its fixed gallery. Its default configuration reaches 39.15 mAP@5 at approximately 0.62 s/query. The latency-oriented WISER configuration takes $18.4\times$ as long as `UniCIR`, whose mAP@5 is 9.68 points higher. The balanced and default WISER configurations take $66.3\times$ and $113.9\times$ as long, respectively. Figure 6 uses the default WISER measurement. Each complete pipeline uses its own models.

F.2 UniCIR Component Costs

Tables 32 and 33 give the stage breakdown of a separate timing run with an end-to-end latency of 1.94111 s/query. Figure 6 uses the updated reported server measurement of approximately 0.61987 s/query.

Table 32: Earlier end-to-end query-time measurements. Full includes description construction, universal and auxiliary encoding, vector composition and gallery search. Serial throughput is 1000 divided by mean milliseconds.

Method	Mean (ms)	p50 (ms)	p95 (ms)	Peak (GiB)	Serial q/s
Best positive-all	1819.28	1047.38	4140.08	30.91	0.55
Double-positive-all	2445.03	2032.21	4582.83	29.55	0.41
`UniCIR` (Full)	1941.11	1133.69	3569.84	27.29	0.52

Table 33: Cost components. Times are in milliseconds. Calls and input counts are per-query means.

Method	Generation	Joint enc.	Aux.	Search	Other	Gen. calls	Emb. inputs
Positive-all	1591.84	88.68	31.08	49.00	58.68	1.05	1.97
Double-positive-all	2235.74	65.49	53.02	15.95	74.83	2.18	2.81
Full	1719.49	81.04	66.10	28.52	45.96	1.93	2.37

The component sums are 1,819.28 ms for Positive-all, 2,445.03 ms for Double-positive-all and 1,941.11 ms for Full. The inverse-mean rates of 0.55/0.41/0.52 query/s describe serial execution. Full averages 1.93 generation calls and 2.37 effective embedding inputs per query.

The Universal encoding/readout measurement has a mean of 74.09 ms, a median of 60.24 ms and a peak memory use of 16.14 GiB. These measurements cover encoding and readout. Figure 6 compares complete pipelines including generation and retrieval.

G Numerical Conventions

Scores are percentages, increments are percentage points, and counts are integers. We use source-reported averages for published baselines. CoTMR reports 61.32 average FashionIQ R@50, while its displayed category values average 61.30. MegaPairs reports BGE-VL-large scores to one decimal place, displayed here with two decimal places. CIRR and CIRCO benchmark scores use test data, and FashionIQ uses standard validation with equal category weights. Appendix J gives conditional query-level tests from the available CIRR validation discordance counts.

Several reported increments differ by 0.01 from subtraction of displayed scores. GME-7B CIRR Full–Universal and Full–Positive are listed as 4.47/2.82, versus 4.48/2.83 from subtraction. Active FashionIQ Full–No-anchor is 14.82, versus 14.83. CIRR Full–matched Double-positive has a net increment of 1.44, versus 1.43 from its displayed components. The CIRR pattern Full–Positive-pair increment is 2.94, versus 2.95 from the Universal-relative increments.

Sleeve-colour example. The original edit is *It is a brighter color in pink and blue. and has light sleeve blue and is pink color.* Figure 1 uses a shortened version with the real reference, Universal top-1 and annotated target. Figure 2 illustrates description construction for the same request. The image IDs are B00CQNY0EG, B00BHIFMDO and B007KBMC58, respectively. The target's identity comes from the dataset annotation.

H Qualitative Comparisons on Development Queries

The outputs in Figure 16 are CIRR/CIRCO development predictions with relevance annotations. They come from the joint-description/clause study at $\alpha = 0.50, \beta = 1.00$, using Qwen3-VL-4B-Instruct and Qwen3-VL-Embedding-8B. Its construction and weights differ from the main benchmark configuration. The examples were selected to show local structure, multi-positive viewpoint retrieval and a count regression.

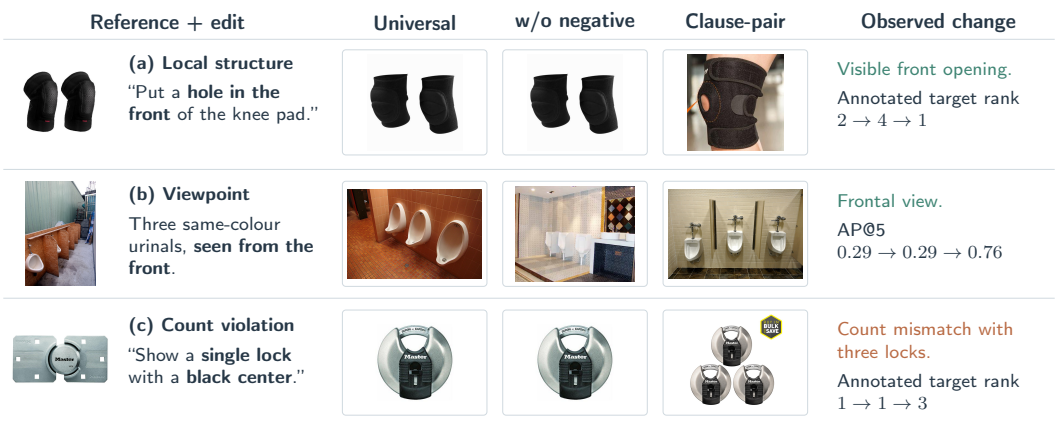


Figure 16: Clause-pair examples for local structure, viewpoint and count, using the configuration specified in this appendix. Two improvements accompany a count regression. CIRCO AP@5 uses all annotated positives.

Local structure. The edit is "Put a hole in the front of the knee pad." The positive description is "Black knee pads with textured front panel." and the negative is "Black knee pads with intact front panel." The annotated target ranks $2 \to 4 \to 1$ under Universal, the no-negative control and the paired correction. Although the correction retrieves a knee pad with a visible opening, the positive omits that requirement.

Viewpoint. The figure abbreviates the request as three same-colour urinals seen from the front. The pair is "Three urinals, same color, front-facing." and "Three urinals, same color, not front-facing." Figure 17 shows the complete top-five comparison. Universal and the no-negative control have annotated positives at ranks $3, 4, 5$, while the paired correction has positives at $1, 2, 3, 5$. AP@5 values are $0.29, 0.29, 0.76$. The designated target remains third in every row.



Figure 17: Complete top-five rankings for the viewpoint diagnostic. Rows show the returned images in rank order. Green borders mark annotated positives for this development query. AP@5 is displayed on the $[0, 1]$ scale. The designated target remains third in every row.

Count. The edit is "Show a single lock with a black center." The positive repeats this request, and the negative is "Show a single lock with a silver center." Universal and the no-negative control return the designated target first. The correction ranks it third and returns an image of three locks. The returned image violates the single-lock requirement.

H.1 Additional Value-Substitution Cases

Figures 18 and 19 show a separate value-substitution construction at $\alpha = 0.50, \beta = 1.00$. These are development queries, and FashionIQ uses the smaller development gallery (q0gal). The benchmark method uses the controlled constructor described in Section 3.

Ground and count conditions. The request asks to remove the people on the left, the lake and some of the birds, leaving three birds on the ground, two brown and orange and one grey. The positive description is "Three birds—two brown-orange, one grey—stand on ground. No people, lake, or other birds remain. Background shows distant fountain and trees, unchanged." The negative replaces the initial "Three" with "Two" while retaining the subsequent enumeration of three birds. This textual inconsistency accompanies a target-rank change from $1 \to 1 \to 6$ under Universal, the positive-only control and the paired correction.

Background condition. The edit reads "is gray colored with dirty humor and has a grey background." The positive reads "Gray-colored t-shirt with dirty humor, featuring a gray background. Graphic shows a man holding a beer, with text 'LET'S GET READY TO STUMBLE!' surrounded by shamrocks. All elements retain gray tone and humorous intent." The negative substitutes "black



Figure 18: Value-substitution examples on development queries. Reference, Universal, positive-only and paired top-1 images are shown alongside the annotated target. Requests are shortened in (a). The background example's rank is censored at > 50. The comparisons use the development construction and galleries specified in this section.

background" for "gray background". The target ranks $16 \to 9 \to >50$ under the same three arms. Both positive controls have the same top-1 and target rank in this example.

Dress colour and other edits. The original edit is "is a black dress with longer sleeves and The desired product is slightly longer and is less textured." The positive describes "A black dress with longer sleeves, slightly longer length, and reduced texture, worn by a model against a plain background." Its paired negative substitutes "white dress" for "black dress". Figure 19 shows the paired top-three outputs and the annotated target. The target ranks $23 \to 19 \to 41$ under Universal, positive-only augmentation and the pair. All three query constructions place the target outside the top ten, giving R@10 = 0.



Figure 19: Top-three outputs for the dress value-substitution example. The rightmost image is the annotated target, which is outside the displayed top three. All images and ranks use the development construction at $\alpha = 0.50, \beta = 1.00$.

I Development-Set Comparisons

The clause-pair study covers 160 development queries with five usable pairs and smaller FashionIQ galleries. Its operating point improves the three-dataset average by 1.53 points over Universal and 0.49

over the same-weight no-negative control. A separate 160-query value-substitution study has 13 usable pairs and loses 1.39 points to Universal and 2.78 to its same-mask positive at the same coefficients. Both studies use development settings and coefficients different from the main benchmarks.

The full-query ablations in Table 3 evaluate the default controlled constructor on CIRR test and standard FashionIQ validation.

J Paired Retrieval Outcomes and Joint Adherence

We analyze retrieval repairs and losses from the reported discordant counts, then examine the overlap of change and preservation judgments.

J.1 Repairs and Losses on CIRR Validation

For CIRR R@1, repairs and losses form the discordant pairs. Conditional on their total d, the null assigns each direction probability $1/2$. Table 34 gives two-sided exact binomial probabilities with Holm adjustment across the four available comparisons from Table 20.

Table 34: Paired CIRR validation outcomes at R@1. Repairs change an incorrect top-1 to a correct result, while losses reverse this change. Win share uses discordant queries. Nominal two-sided binomial probabilities assume independent query outcomes. Holm adjustment covers the four displayed comparisons.

Full versus	Repairs	Losses	Win share (%)	Adjusted p
Universal	240	43	84.81	9.58×10^{-34}
Best positive-all	103	29	78.03	1.94×10^{-10}
Two-positive (matched mask)	87	27	76.32	1.53×10^{-8}
Two-positive (independent)	100	31	76.34	2.32×10^{-9}

Against matched two-positive augmentation, 87 queries improve and 27 worsen. Full repairs 76.32% of these discordant queries, giving a 1.44-point net gain over all 4,181 queries. The three positive-description comparisons have win shares of 76.32–78.03%, alongside 27–31 new errors.

These conditional tests use the CIRR validation discordant counts and assume independent query outcomes. Repeated reference images can introduce dependence. Holm adjustment covers the four comparisons, while model and parameter selection were informed by validation. The resulting probabilities describe query-level directional imbalance under these assumptions.

J.2 Joint Adherence

Within each method, change, preservation and joint-yes counts C, P and J give the disjoint categories J, $C-J$, $P-J$ and $90-C-P+J$. Table 35 partitions the counts from Table 24.

Table 35: Change/preservation judgment overlap on the same 90 queries. Each column specifies which criteria receive a yes. The remaining criteria receive no or uncertain judgments.

Method	Both yes	Change only	Preserve only	Neither yes
Universal	45	15	28	2
Positive	50	9	15	16
Full	59	10	7	14

K Limitations

CIC handles a restricted set of local conditions and falls back to Universal elsewhere. Accepted pairs may omit requested changes or introduce conflicting text. In the edit-adherence study, improved joint

adherence accompanies fewer positive judgments of reference preservation. Handling interacting edits and reducing the frozen generator's query-side computation remain directions for future work.