FLOW-CIR: GENERATIVE FEATURE RECTIFICATION FOR NOISE-ROBUST COMPOSED IMAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Composed image retrieval aims to retrieve a target image according to a reference image and a textual modification. Existing methods usually assume that training triplets are correctly matched, while real-world annotations may contain noisy triplet correspondence caused by subjective modification texts, ambiguous visual changes, and imperfect target selection. Such noise is especially challenging because a suspicious triplet is not always completely wrong; it may still preserve partial semantic cues that are useful for learning fine-grained visual transformations. Most noise-robust CIR methods mainly identify, suppress, or down-weight unreliable triplets, which reduces harmful supervision but may also discard informative signals from hard samples. In this paper, we propose Flow-CIR, a feature-level generative rectification framework for noise-robust composed image retrieval. Instead of simply forgetting suspicious triplets, Flow-CIR learns a conditional semantic vector field from reliable triplets and uses it to generate pseudo-target features for suspicious samples. To avoid over-correction, the rectified target feature is softly mixed with the observed target representation and is applied only after a warm-up stage. Since rectification is performed in the compact feature space, Flow-CIR introduces no additional inference cost. Experiments on FashionIQ and CIRR under different noise ratios demonstrate the robustness and effectiveness of the proposed framework.

1 INTRODUCTION

Composed image retrieval (CIR) aims to retrieve a target image according to a multimodal query composed of a reference image and a modification text. Compared with conventional image retrieval, CIR allows users to express fine-grained search intent by specifying how the reference image should be changed. This task has been widely studied on benchmarks such as FashionIQ (Wu et al., 2021) and CIRR (Liu et al., 2021), and is useful for personalized search, e-commerce recommendation, and interactive multimodal retrieval.

The fine-grained nature of CIR makes triplet supervision particularly important. In practice, constructing high-quality reference-text-target triplets is expensive and subjective. Existing analyses have also revealed limitations in current CIR datasets, including limited scale, modality redundancy, and incomplete or ambiguous supervision (Levy et al., 2024). As a result, the modification text may be partially inconsistent with the observed target image. As illustrated in Figure 1, a noisy triplet can still be partially correct: the target image may satisfy some modification cues, such as color, while violating others, such as sleeve style.

Recent studies have begun to address noisy triplet correspondence in CIR. Related annotation ambiguity was studied by Css-Net (Zhang et al., 2024b), while noisy triplet correspondence was explicitly formulated by TME (Li et al., 2025) and further explored by ConeSep, HABIT, INTENT, and RankVR (Li et al., 2026a;b; Chen et al., 2026; Huang et al., 2026). These methods mainly identify, re-weight, suppress, or separate unreliable triplets. Although such strategies reduce harmful supervision, suspicious triplets are not always useless; they may still contain valuable partial semantics. Purely suppressing them may discard useful training signals, especially when hard positive samples and noisy samples are difficult to distinguish.

In this paper, we rethink noisy triplet learning from a more constructive perspective. Instead of asking whether a triplet should be trusted or forgotten, we ask whether a noisy target representation

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.

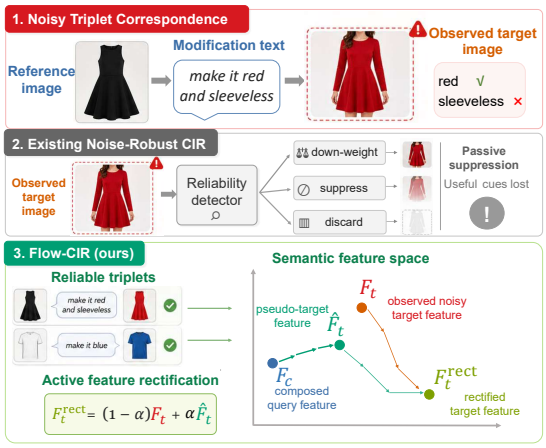


Figure 1: Motivation of Flow-CIR. Unlike existing methods that passively suppress suspicious triplets, Flow-CIR actively rectifies noisy target features using semantic transport learned from reliable triplets.

can be actively rectified toward a semantically cleaner target. Motivated by flow matching and rectified flow (Lipman et al., 2023; Liu et al., 2023), we propose Flow-CIR, a feature-level generative rectification framework for noise-robust CIR. Flow-CIR learns a conditional semantic vector field from reliable triplets and uses it to generate pseudo-target features for suspicious triplets, shifting the learning paradigm from passive noise suppression to active noisy supervision rectification.

Our main contributions are summarized as follows:

- We propose Flow-CIR, a noise-robust composed image retrieval framework that reformulates noisy supervision learning as feature-level generative rectification rather than passive noise suppression.
- We introduce a reliability-aware training strategy that partitions reliable and suspicious triplets with a ranking-based fidelity strategy, and learns a conditional vector field from reliable triplets to provide semantic correction directions for suspicious samples.
- We provide a theoretical clarification of why Flow-CIR is different from simple feature interpolation, and analyze its training and inference cost. Experiments on FashionIQ and CIRR under multiple noise ratios validate the robustness of Flow-CIR.

2 Related Work

2.1 Composed Image Retrieval

Composed image retrieval retrieves a target image from a multimodal query consisting of a reference image and a textual modification. Early methods learn task-specific composition modules to fuse visual and textual cues, such as TIRG (Vo et al., 2019), VAL (Chen et al., 2020), ComposeAE (Anwaar et al., 2021), DCNet (Kim et al., 2021), and ARTEMIS (Delmas et al., 2022). With the development of large-scale vision-language pre-training, CLIP-, BLIP-, and BLIP-2-based methods have been widely adopted for stronger cross-modal alignment and compositional representation learning (Radford et al., 2021; Li et al., 2022; 2023; Baldrati et al., 2022; 2023b). Recent studies further improve CIR through target-guided learning, prompt learning, bidirectional training, candidate re-ranking, and uncertainty modeling (Wen et al., 2023; Bai et al., 2024; Liu et al., 2024a;b; Chen et al., 2024; Tang et al., 2026). Zero-shot CIR methods reduce the dependence on labeled triplets through textual inversion, pseudo-word learning, large-scale instruction synthesis, or LLM-enhanced reasoning (Saito et al., 2023; Baldrati et al., 2023a; Agnolucci et al., 2024; Zhang et al., 2024a; Ge et al., 2025; Wang et al., 2025a). These methods mainly study generalization under limited or absent triplet supervision, whereas our work focuses on robustness under noisy supervised triplets. Nevertheless, both settings highlight the importance of learning transferable compositional

representations. Most existing supervised CIR methods still assume that the annotated triplets are reliable, making them sensitive to noisy correspondence.

2.2 Noise-Robust Composed Image Retrieval

Recent noise-robust CIR methods focus on noisy triplet correspondence, where the reference image, modification text, and target image may be partially mismatched. Css-Net addresses annotation ambiguity through consensus learning (Zhang et al., 2024b). TME explicitly formulates noisy triplet correspondence and enhances noisy textual supervision (Li et al., 2025). ConeSep, HABIT, INTENT, RankVR, and Air-Know further explore robust learning by estimating sample reliability, separating noisy triplets, or mitigating unreliable supervision (Li et al., 2026a;b; Chen et al., 2026; Huang et al., 2026; Fu et al., 2026). Beyond CIR, noisy correspondence has also been studied in broader cross-modal matching, including meta similarity correction, bidirectional consistency, distribution modeling, and asymmetric similarity learning (Han et al., 2023; Yang et al., 2023; Shi et al., 2024; Wang et al., 2025b). General noisy-label learning methods also show the importance of sample selection and semi-supervised treatment under corrupted supervision (Li et al., 2020). Different from these identification- or suppression-based paradigms, Flow-CIR treats suspicious triplets as imperfect but rectifiable supervision, aiming to preserve useful partial semantics while correcting unreliable target features.

2.3 Flow Matching and Feature Rectification

Flow matching provides an efficient generative modeling framework by directly learning a vector field that transports samples between distributions (Lipman et al., 2023). Rectified Flow further formulates generation as learning simple transport trajectories between source and target distributions through ordinary differential equations (Liu et al., 2023). These methods suggest that modeling the direction from one representation state to another can provide a flexible way to construct semantically meaningful transformations.

Inspired by this idea, we introduce flow-based rectification into noise-robust composed image retrieval. Different from pixel-level generation, our method performs lightweight feature-level rectification in the retrieval embedding space. Specifically, reliable triplets are used to learn a conditional vector field from composed query features to target features. For suspicious triplets, the learned vector field generates pseudo target features, which are softly mixed with the original target representations. In this way, Flow-CIR does not simply discard suspicious supervision, but attempts to correct its semantic direction while keeping the inference pipeline unchanged.

3 Method

3.1 Problem Formulation

Composed image retrieval (CIR) aims to retrieve a target image according to a multimodal query consisting of a reference image and a textual modification. Let $\mathcal{D} = \{(I_r^{(i)}, T_m^{(i)}, I_t^{(i)})\}_{i=1}^{N}$ denote the training set, where I_r is the reference image, T_m is the modification text, and I_t is the observed target image. Under noisy triplet correspondence, I_t may satisfy only part of the modification semantics and therefore may not represent the correct semantic destination of the composed query.

We first extract the reference-image feature, modification-text feature, and observed target feature as:

$$F_r = E_I(I_r), F_m = E_T(T_m), F_t = \mathcal{N}(E_I(I_t)), \tag{1}$$

where E_I and E_T denote the image and text encoders, respectively, and $\mathcal{N}(z) = z/\|z\|_2$ denotes ℓ_2-normalization. The composed query feature is obtained using a learnable composition module C_ϕ:

$$F_c = \mathcal{N}(C_\phi(F_r, F_m)). \tag{2}$$

Accordingly, F_c represents the retrieval intent jointly specified by the reference image and modification text, whereas F_t represents the supervision provided by the observed target image. When the triplet is corrupted, directly aligning F_c with F_t may pull the composed query toward an incorrect

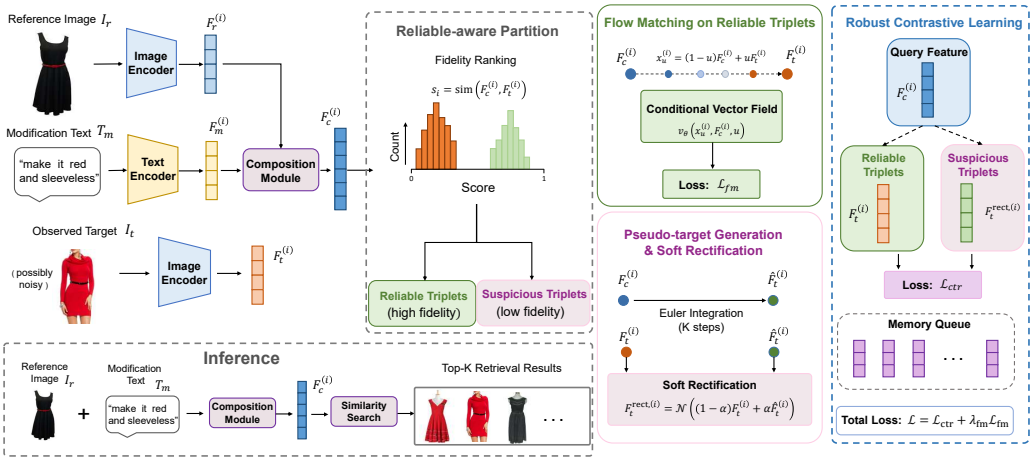


Figure 2: Overview of the proposed Flow-CIR framework. Given a reference image, a modification text, and an observed target image, Flow-CIR first composes the multimodal query feature and estimates triplet reliability. Reliable triplets are used to learn a conditional flow field, while suspicious triplets are softly rectified with generated pseudo-target features for robust contrastive learning. During inference, Flow-CIR follows the standard CIR retrieval pipeline without feature rectification.

semantic region. Flow-CIR therefore aims to estimate triplet reliability and rectify suspicious target supervision in the feature space.

We adopt a BLIP-2 Q-Former-based backbone as the image-text composition module C_ϕ. Instead of designing an additional complex fusion architecture, we keep the composition backbone standard so that the robustness improvement can be attributed mainly to the proposed feature-level rectification mechanism. This design also makes Flow-CIR complementary to stronger fusion modules: the reliability-aware partition and flow rectification operate on retrieval features and can be combined with different CIR composition backbones. Since reliability estimation depends on the composed query feature F_c, we use a warm-up stage before activating pseudo-target rectification, which reduces the risk of unstable early fusion features causing incorrect correction.

3.2 Overview of Flow-CIR

Figure 2 presents Flow-CIR. The model composes the query feature F_c, encodes the observed target as F_t, and partitions triplets by calibrated fidelity. Reliable triplets supervise a conditional semantic field; for suspicious triplets, the field generates pseudo-targets that are softly mixed with observed targets for contrastive learning. These branches are training-only, so inference follows the standard CIR retrieval pipeline.

3.3 Reliability-Aware Triplet Partition

A low query–target similarity does not necessarily indicate an incorrect triplet, since hard positive samples may also receive low scores when the modification is subtle or the representation is not yet sufficiently discriminative. We therefore use the matching score only as a reliability cue rather than an absolute noise label.

For the i-th training triplet, the reliability score is computed as

$$s_i = \mathrm{sim}\left(F_c^{(i)}, F_t^{(i)}\right), \tag{3}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity. Since the features are normalized, Eq. (3) is equivalent to the inner product $(F_c^{(i)})^\top F_t^{(i)}$.

Since a low matching score may also correspond to a hard positive sample, we do not use s_i as an absolute noise label. Instead, we treat it as a relative reliability cue and convert it into a calibrated

fidelity score:

$$q_i = g(s_i; b), \tag{4}$$

where b denotes a batch-level background matching boundary estimated from randomly mismatched image-text pairs. Specifically, we first compute

$$\bar{s}_i = \frac{\max(s_i - b, 0)}{\max_j \max(s_j - b, 0) + \epsilon}, \tag{5}$$

and then obtain the calibrated fidelity score as

$$q_i = -\bar{s}_i^2(\bar{s}_i - 1)(\bar{s}_i + 1). \tag{6}$$

This calibration suppresses scores below the background boundary and normalizes the remaining scores within the mini-batch.

Given a suspicious ratio ρ, the number of suspicious samples in a mini-batch of size B is defined as

$$K_{\text{sus}} = \lfloor \rho B \rfloor. \tag{7}$$

We then select the samples with the lowest fidelity scores as suspicious triplets:

$$\mathcal{T}_{\text{sus}} = \text{BottomK}\left(\{q_i\}_{i=1}^{B}, K_{\text{sus}}\right), \quad \mathcal{T}_{\text{rel}} = \{1, \ldots, B\} \setminus \mathcal{T}_{\text{sus}}. \tag{8}$$

Here, $\mathcal{T}_{\text{rel}}$ and $\mathcal{T}_{\text{sus}}$ denote the reliable and suspicious subsets, respectively. This ranking-based partition avoids relying on a fixed global similarity threshold and adapts to the score distribution of each mini-batch. More importantly, suspicious samples are not discarded; they are retained and softly rectified in the feature space.

3.4 Conditional Flow Matching on Reliable Triplets

For reliable triplets, the observed target feature is assumed to provide sufficiently trustworthy supervision. We formulate the semantic transformation from F_c to F_t as a conditional flow-matching problem. To distinguish the continuous flow time from the target-feature subscript, we use $u \in [0, 1]$ as the flow-time variable.

For each $i \in \mathcal{T}_{\text{rel}}$, we sample $u \sim \mathcal{U}(0, 1)$ and construct the linear probability path

$$x_u^{(i)} = (1 - u)F_c^{(i)} + uF_t^{(i)}. \tag{9}$$

The corresponding target velocity is

$$v^{\star(i)} = \frac{dx_u^{(i)}}{du} = F_t^{(i)} - F_c^{(i)}. \tag{10}$$

We use a conditional vector field $v_\theta(x_u, F_c, u)$ to predict the semantic transport direction at each intermediate state. The flow-matching objective is

$$\mathcal{L}_{\text{fm}} = \frac{1}{|\mathcal{T}_{\text{rel}}|} \sum_{i \in \mathcal{T}_{\text{rel}}} \mathbb{E}_{u \sim \mathcal{U}(0,1)} \left[\left\| v_\theta\left(x_u^{(i)}, F_c^{(i)}, u\right) - v^{\star(i)} \right\|_2^2 \right]. \tag{11}$$

The straight path in Eq. (9) is used only to provide supervision for learning the vector field. FlowCIR is therefore different from applying a fixed linear interpolation to each suspicious sample. The shared conditional field is learned from many reliable triplets and captures transferable, distribution-level semantic transport patterns in the retrieval embedding space.

3.5 Pseudo-Target Generation and Soft Rectification

For each suspicious triplet $i \in \mathcal{T}_{\text{sus}}$, the observed target feature $F_t^{(i)}$ is not fully trusted. We initialize the flow trajectory from the composed query feature:

$$x^{(i)}(0) = F_c^{(i)}, \qquad \frac{dx^{(i)}(u)}{du} = v_\theta\left(x^{(i)}(u), F_c^{(i)}, u\right). \tag{12}$$

The trajectory is numerically integrated from $u = 0$ to $u = 1$ using a K-step Euler solver:

$$x_{k+1}^{(i)} = x_k^{(i)} + \Delta u \, v_\theta \left(x_k^{(i)}, F_c^{(i)}, u_k \right), \tag{13}$$

where

$$x_0^{(i)} = F_c^{(i)}, \qquad \Delta u = \frac{1}{K}, \qquad u_k = k\Delta u. \tag{14}$$

After K integration steps, the generated pseudo-target feature is

$$\hat{F}_t^{(i)} = \mathcal{N}\left(x_K^{(i)} \right). \tag{15}$$

Although $\hat{F}_t$ provides a semantic correction direction learned from reliable triplets, completely replacing the observed target may remove valid visual information that remains in a partially mismatched sample. We therefore adopt a soft rectification strategy:

$$F_t^{\mathrm{rect},(i)} = \mathcal{N}\left((1-\alpha) F_t^{(i)} + \alpha \hat{F}_t^{(i)} \right), \qquad i \in \mathcal{T}_{\mathrm{sus}}, \tag{16}$$

where $\alpha \in [0, 1]$ is a fixed rectification coefficient. We use a small fixed α for stability: it preserves the observed target as the main supervision while limiting errors from early pseudo-target trajectories. Confidence-adaptive rectification remains a promising extension.

For unified notation, the target representation used by the retrieval loss is defined as

$$\widetilde{F}_t^{(i)} = \begin{cases} F_t^{(i)}, & i \in \mathcal{T}_{\mathrm{rel}}, \\ F_t^{\mathrm{rect},(i)}, & i \in \mathcal{T}_{\mathrm{sus}}. \end{cases} \tag{17}$$

Thus, reliable samples retain their original target features, whereas only suspicious samples are rectified.

To stabilize reliability estimation and feature rectification, we adopt a delayed activation strategy. During the first two epochs, the model learns basic compositional alignment using the observed target features without pseudo-target rectification. After this warm-up stage, the flow-matching loss is optimized on reliable triplets, and pseudo-target mixing is activated from epoch e_s. In our implementation, e_s is controlled by the flow-start epoch. This strategy reduces the risk of misclassifying hard positive samples and prevents inaccurate early flows from over-correcting target supervision.

3.6 Robust Contrastive Optimization

Using the unified targets in Eq. (17), the query-to-target contrastive objective is

$$\mathcal{L}_{q\to t} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp\left(\mathrm{sim}\left(F_c^{(i)}, \widetilde{F}_t^{(i)} \right) / \tau \right)}{\sum_{j=1}^{B} \exp\left(\mathrm{sim}\left(F_c^{(i)}, \widetilde{F}_t^{(j)} \right) / \tau \right)}, \tag{18}$$

where B is the mini-batch size and τ is the temperature parameter. The target-to-query objective $\mathcal{L}_{t\to q}$ is defined symmetrically. The robust contrastive objective is

$$\mathcal{L}_{\mathrm{ctr}} = \frac{1}{2} \left(\mathcal{L}_{q\to t} + \mathcal{L}_{t\to q} \right). \tag{19}$$

When the optional memory queue is enabled, detached features from previous mini-batches are appended to the negative candidate set. This enlarges the cross-batch negative pool without changing the positive correspondence in the current mini-batch. The memory queue is used only during training and does not affect inference.

The final objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{ctr}} + \lambda_{fm} \mathcal{L}_{\mathrm{fm}}, \tag{20}$$

where λ_{fm} controls the contribution of the flow-matching objective.

Table 1: Main results on FashionIQ under different noise ratios. "Mean" denotes the average of Avg R@10 and Avg R@50. The best and second-best results are highlighted in bold and underlined, respectively. We compare with CALA, SPRC, TME, HABIT, INTENT, and RankVR (Jiang et al., 2024; Bai et al., 2024; Li et al., 2025; 2026b; Chen et al., 2026; Huang et al., 2026).

Noise	Method	Dress		Shirt		Toptee		Avg.		Mean
		R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	
0.0	CALA	42.38	66.08	46.76	68.16	50.93	73.42	46.69	69.22	57.96
	SPRC	49.18	72.43	55.64	73.89	59.35	78.58	54.92	74.97	64.85
	TME	49.73	71.69	56.43	74.44	59.31	78.94	55.15	75.02	65.09
	HABIT	49.99	72.38	**56.62**	74.68	59.51	78.53	55.38	75.20	65.29
	INTENT	50.32	72.10	56.32	**74.93**	59.28	78.45	55.31	75.16	65.24
	RankVR	50.26	72.42	55.82	74.02	57.95	77.91	54.68	74.75	64.73
	Flow-CIR	**50.98**	**73.84**	56.58	74.90	**59.55**	**79.37**	**55.69**	**76.04**	**65.86**
0.2	CALA	29.05	51.36	35.28	56.23	36.05	58.24	33.46	55.28	44.37
	SPRC	39.81	62.22	48.58	66.29	50.48	70.58	46.29	66.36	56.33
	TME	49.03	70.35	55.84	73.16	57.22	78.23	54.03	73.91	63.97
	HABIT	49.63	71.34	55.67	73.19	58.14	78.32	54.48	74.28	64.38
	INTENT	49.32	71.43	55.32	73.57	58.01	78.46	54.22	74.49	64.36
	RankVR	49.81	71.48	55.97	73.22	57.29	77.94	54.36	74.21	64.29
	Flow-CIR	**50.56**	**72.66**	**56.22**	**73.92**	**59.24**	**78.64**	**55.34**	**75.07**	**65.20**
0.5	CALA	20.77	40.95	29.69	46.57	27.03	46.81	24.83	44.78	34.80
	SPRC	35.94	57.16	42.25	61.63	44.98	64.76	41.06	61.19	51.12
	TME	46.26	68.27	53.09	71.88	55.07	76.59	51.47	72.25	61.86
	HABIT	47.33	69.71	53.72	72.55	56.51	77.00	52.52	73.09	62.80
	INTENT	47.99	71.24	52.78	72.48	56.79	76.23	52.52	73.32	62.92
	RankVR	46.54	69.29	53.50	**72.84**	55.93	**77.56**	51.99	73.23	62.61
	Flow-CIR	**49.12**	**71.80**	**54.89**	72.53	**58.60**	77.52	**54.20**	**73.95**	**64.07**
0.8	CALA	14.28	30.59	19.73	35.82	19.48	36.10	17.83	34.41	26.00
	SPRC	28.41	50.77	36.21	54.37	35.90	59.06	33.51	54.03	43.77
	TME	41.45	64.35	47.30	68.20	51.25	73.23	46.67	68.60	57.63
	HABIT	42.04	65.20	50.12	69.77	52.92	73.61	48.36	69.53	58.94
	INTENT	42.07	65.58	50.38	69.41	53.09	73.91	48.51	69.63	59.07
	RankVR	42.02	64.23	48.85	68.24	51.48	73.17	47.45	68.55	58.00
	Flow-CIR	**45.85**	**67.36**	**50.57**	**70.22**	**53.38**	**75.12**	**49.93**	**70.90**	**60.42**

Computational Complexity. Flow-CIR adds only training cost, approximately $O((1-\rho)BC_v + \rho BKC_v)$, where C_v is one field evaluation. With $d = 256$ and $K = 3$, rectification remains lightweight; the field and memory queue are removed at inference.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate Flow-CIR on two standard composed image retrieval benchmarks, FashionIQ (Wu et al., 2021) and CIRR (Liu et al., 2021). FashionIQ contains three fashion categories, including Dress, Shirt, and Toptee, where each query consists of a reference image and a modification text. CIRR is a more general-domain benchmark and provides both global retrieval and subset retrieval protocols.

Noisy triplet setting. Following the noisy triplet correspondence protocols used in previous studies (Li et al., 2025; 2026a;b), we randomly corrupt a proportion of training triplets according to the noise ratio σ. We evaluate four noise ratios: $\sigma \in \{0.0, 0.2, 0.5, 0.8\}$. A larger σ indicates more severe mismatch. This reproducible synthetic protocol approximates, but does not fully capture,

Table 2: Main results on CIRR under different noise ratios. Avg. denotes the average of R@5 and R_{sub}@1. The best and second-best results are highlighted in bold and underlined, respectively. We compare with CALA, SPRC, TME, HABIT, INTENT, and RankVR (Jiang et al., 2024; Bai et al., 2024; Li et al., 2025; 2026b; Chen et al., 2026; Huang et al., 2026).

Noise	Method	R@K				R_{sub}@K			Avg. (R@5, R_{sub}@1)
		K=1	K=5	K=10	K=50	K=1	K=2	K=3	
0.0	CALA	49.11	81.21	89.59	98.00	76.27	91.04	96.46	78.74
	SPRC	51.96	82.12	89.74	97.69	80.65	92.31	96.60	81.39
	TME	**53.42**	82.99	90.24	98.15	**81.04**	92.58	96.94	**82.01**
	HABIT	52.71	82.64	90.63	98.19	80.99	**92.77**	**97.00**	81.82
	INTENT	53.37	83.16	**90.73**	**98.22**	80.24	92.37	96.89	81.70
	RankVR	51.94	81.89	90.41	97.41	80.12	92.40	95.65	81.01
	Flow-CIR	52.04	**83.23**	90.41	98.09	80.70	92.68	96.08	81.97
0.2	CALA	41.33	72.70	82.84	94.34	71.66	88.15	94.94	72.18
	SPRC	45.90	75.86	83.52	93.37	78.10	91.40	96.05	76.98
	TME	51.35	81.01	88.53	97.81	78.46	91.25	96.39	79.74
	HABIT	51.68	81.02	89.24	97.81	78.20	91.66	96.75	79.61
	INTENT	51.25	81.36	90.02	**98.05**	77.95	91.40	96.46	79.66
	RankVR	51.72	81.72	89.05	98.02	**78.77**	**91.89**	96.48	80.25
	Flow-CIR	**52.33**	**82.28**	**90.19**	97.75	78.54	91.70	**96.84**	**80.41**
0.5	CALA	36.10	66.12	77.76	92.10	68.12	85.66	93.59	67.12
	SPRC	39.93	66.00	73.59	86.48	75.81	89.21	95.37	70.90
	TME	48.48	78.94	87.28	96.99	76.48	90.07	95.83	77.71
	HABIT	**50.32**	79.63	88.34	97.06	76.84	**90.60**	**96.27**	**78.87**
	INTENT	49.78	79.64	88.99	97.37	**77.18**	90.41	96.00	78.41
	RankVR	48.94	78.58	87.88	96.88	76.51	90.45	95.82	77.55
	Flow-CIR	49.77	**80.24**	**89.09**	**97.44**	76.91	89.19	95.24	78.58
0.8	CALA	31.52	61.49	72.60	89.86	64.34	83.52	92.60	62.92
	SPRC	29.95	51.25	58.51	73.86	70.22	86.05	93.21	60.74
	TME	46.31	75.78	84.89	95.83	73.37	88.02	94.89	74.58
	HABIT	47.93	76.84	85.95	95.90	74.87	89.08	95.21	75.86
	INTENT	47.90	78.13	87.04	96.47	73.81	89.18	**95.54**	75.97
	RankVR	47.63	76.04	85.77	95.88	**74.90**	89.45	95.36	75.47
	Flow-CIR	**50.20**	**79.98**	**88.35**	**97.20**	74.60	**89.72**	95.24	**77.29**

non-uniform real annotation ambiguity; we therefore complement it with qualitative retrieval and feature analyses.

Evaluation metrics. For FashionIQ, we report Recall@10 and Recall@50 on Dress, Shirt, and Toptee, together with their average scores. For CIRR, we report global retrieval metrics R@1, R@5, R@10, and R@50, as well as subset retrieval metrics R_{sub}@1, R_{sub}@2, and R_{sub}@3.

Implementation details. We use a BLIP-2 first-stage Q-Former backbone with a frozen vision encoder, 32 query tokens, maximum text length 32, 224×224 inputs, and 256-dimensional retrieval features. The field is a 1024-hidden-dimension MLP with sinusoidal time embeddings. AdamW uses betas $(0.9, 0.98)$, epsilon 10^{-7}, weight decay 0.05, mixed precision, learning rate 10^{-5}, and batch size 64. We set $\tau = 0.07$, $K = 3$, $\lambda_{fm} = 0.10$, $\alpha = 0.20$, $e_s = 4$, and a FIFO queue of 128 detached features. Both datasets are trained for 10 epochs with two warm-up epochs; results average two independent runs. The rectifier and queue are removed at inference.

4.2 Main Results

Table 1 and Table 2 report the main results on FashionIQ and CIRR under different noise ratios. Compared with existing CIR and noise-robust CIR methods, Flow-CIR achieves competitive or superior performance across different noise levels. In particular, under high-noise settings, Flow-

Table 3: Ablation study of Flow-CIR on FashionIQ and CIRR under the noise ratio of 0.5.

Method	FashionIQ, noise=0.5			CIRR, noise=0.5		
	Avg. R@10	Avg. R@50	Mean	R@5	R_{sub} @1	Avg.
Base Model	51.86	71.34	61.60	78.50	74.21	76.36
Base + Noise-aware Selection	52.03	72.12	62.08	79.54	74.91	77.23
Base + Selection + Flow Loss	52.36	72.14	62.25	79.51	75.15	77.33
Base + Selection + Flow Rectification	53.97	73.20	63.59	80.11	76.38	78.25
Full Flow-CIR (+ Memory Queue)	**54.20**	**73.95**	**64.07**	**80.24**	**76.91**	**78.58**

CIR maintains stronger average performance than most baselines, suggesting that actively rectifying noisy target supervision in the feature space is a promising alternative to simply suppressing or down-weighting suspicious triplets. The results suggest that actively rectifying noisy target supervision in the feature space is a promising alternative to simply suppressing or down-weighting noisy triplets.

4.3 Ablation Study

Table 3 evaluates each component at noise ratio 0.5. Selection improves the base model from 61.60 to 62.08 on FashionIQ and from 76.36 to 77.23 on CIRR; flow loss adds a smaller gain, while rectification reaches 63.59 and 78.25. The memory queue yields the best scores, 64.07 and 78.58, confirming the complementarity of selection, rectification, and cross-batch negatives.

4.4 Limitations

This work evaluates model robustness on noisy supervised CIR with synthetic noise injected into reference-text-target triplets. Our gains reflect better robustness under noisy triplet supervision, not zero-shot retrieval performance. While the learned conditional vector field may transfer semantic correction across attribute-object pairs, we make no claims of compositional zero-shot generalization without dedicated tests. Future work will study whether feature-level flow rectification generalizes to unseen attribute-object combinations and real-world noisy CIR datasets.

Our synthetic noise setup ensures cross-noise-level fair comparison yet fails to capture real annotation ambiguities from subjective text, local visual uncertainty or partially matched targets. Flow-CIR adopts a fixed rectification coefficient α for stable training, though suspicious samples demand adaptive correction magnitudes. We leave confidence-aware rectification and tests on naturally noisy CIR benchmarks for future research.

5 Conclusion

In this paper, we proposed Flow-CIR, a feature-level generative rectification framework for noise-robust composed image retrieval. Different from existing methods that mainly identify, suppress, or down-weight suspicious triplets, Flow-CIR reformulates noisy supervision learning as an active rectification problem. By learning a conditional semantic vector field from reliable triplets, the model can generate pseudo-target features for suspicious samples and softly mix them with the observed target representations. This design allows Flow-CIR to exploit useful cues from imperfect triplets while reducing the negative effect of noisy correspondence. Experiments on FashionIQ and CIRR under different noise ratios demonstrate the robustness and effectiveness of the proposed framework. We hope this work provides a new perspective for learning from noisy multimodal supervision in composed image retrieval.

AI USE STATEMENT

Generative AI tools assisted methodological discussion, code implementation, experiment execution and auditing, statistical analysis, language polishing, manuscript organization, and LaTeX formatting. Numerical results were computed by executable pipelines from the recorded inputs rather than synthesized from manuscript tables. Automated checks do not substitute for independent author verification. The authors take responsibility for the final manuscript, including its claims, implementation, data use, and results.

REPRODUCIBILITY STATEMENT

Sections 3 and 4.1 specify the model, objectives, noise-generation protocol, datasets, evaluation metrics, optimization settings, and number of independent runs. Appendix A reports performance across noise ratios, Appendix B gives parameter sensitivity, and Appendix C provides qualitative retrieval and feature analyses. These details, together with the cited public benchmarks and pretrained backbone, are intended to support reproduction of the reported experiments.

ETHICS STATEMENT

This work uses the publicly available FashionIQ and CIRR benchmarks and does not collect new personal data or involve human-subject experiments. The method may inherit biases from pretrained vision-language models and benchmark data, and robustness to synthetic correspondence noise does not establish fairness or reliability in deployment. Dataset images remain subject to their original terms and licenses; downstream users should evaluate bias, privacy, and failure modes in their application context.

REFERENCES

Lorenzo Agnolucci, Alberto Baldrati, Marco Bertini, and Alberto Del Bimbo. isearle: Improving textual inversion for zero-shot composed image retrieval. *arXiv preprint arXiv:2405.02951*, 2024.

Muhammad Umer Anwaar, Egor Labintcev, and Martin Kleinsteuber. Compositional learning of image-text query for image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1140–1149, 2021.

Yang Bai, Xinxing Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong Goh, and Chun-Mei Feng. Sentence-level prompts benefit composed image retrieval. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=m3ch3kJL7q.

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*, pp. 21466–21474, 2022.

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*, 20(3):1–24, 2023b.

Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3001–3011, 2020.

Yiyang Chen, Zhedong Zheng, Wei Ji, Leigang Qu, and Tat-Seng Chua. Composed image retrieval with text feedback via multi-grained uncertainty regularization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Yb5KvPkKQg.

Zhiwei Chen, Yupeng Hu, Zhiheng Fu, Zixu Li, Jiale Huang, Qinlei Huang, and Yinwei Wei. Intent: Invariance and discrimination-aware noise mitigation for robust composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20463–20471, 2026.

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

Zhiheng Fu, Yupeng Hu, Qianyun Yang, Shiqi Zhang, Zhiwei Chen, and Zixu Li. Air-know: Arbiter-calibrated knowledge-internalizing robust network for composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2658–2670, 2026.

Hongfei Ge, Yuanchun Jiang, Jianshan Sun, Kun Yuan, and Yezheng Liu. LLM-enhanced composed image retrieval: An intent uncertainty-aware linguistic-visual dual channel matching model. *ACM Transactions on Information Systems*, 43(2):1–30, 2025. doi: 10.1145/3699715.

Haochen Han, Kaiyao Miao, Qinghua Zheng, and Minnan Luo. Noisy correspondence learning with meta similarity correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7517–7526, 2023.

Jiale Huang, Zixu Li, Zhiheng Fu, Zhiwei Chen, Qinlei Huang, and Yupeng Hu. Rankvr: Low-rank structure perception and value recalibration for robust composed image retrieval. In *Proceedings of the 2026 International Conference on Multimedia Retrieval*, pp. 269–278, 2026.

Xintong Jiang, Yaxiong Wang, Mengjian Li, Yujiao Wu, Bingwen Hu, and Xueming Qian. Cala: Complementary association learning for augmenting comoposed image retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2177–2187, 2024.

Jongseok Kim, Youngjae Yu, Hoeseong Kim, and Gunhee Kim. Dual compositional learning in interactive image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 1771–1779, 2021.

Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Data roaming and quality assessment for composed image retrieval. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 2991–2999, 2024.

Junnan Li, Richard Socher, and Steven C. H. Hoi. DivideMix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Shuxian Li, Changhao He, Xiting Liu, Joey Tianyi Zhou, Xi Peng, and Peng Hu. Learning with noisy triplet correspondence for composed image retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19628–19637, 2025.

Zixu Li, Yupeng Hu, Zhiwei Chen, Mingyu Zhang, Zhiheng Fu, and Liqiang Nie. Conesep: Cone-based robust noise-unlearning compositional network for composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16897–16909, 2026a.

Zixu Li, Yupeng Hu, Zhiwei Chen, Shiqi Zhang, Qinlei Huang, Zhiheng Fu, and Yinwei Wei. Habit: Chrono-synergia robust progressive learning framework for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6762–6770, 2026b.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023.

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

Zheyuan Liu, Weixuan Sun, Yicong Hong, Damien Teney, and Stephen Gould. Bi-directional training for composed image retrieval via text prompt learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5741–5750, 2024a.

Zheyuan Liu, Weixuan Sun, Damien Teney, and Stephen Gould. Candidate set re-ranking for composed image retrieval with dual multi-modal encoder. *Transactions on Machine Learning Research*, 2024b. URL https://openreview.net/forum?id=fJAwemcvpL.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 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.

Haitao Shi, Meng Liu, Xiaoxuan Mu, Xuemeng Song, Yupeng Hu, and Liqiang Nie. Breaking through the noisy correspondence: A robust model for image-text matching. *ACM Transactions on Information Systems*, 42(6):1–26, 2024. doi: 10.1145/3662732.

Haomiao Tang, Jinpeng Wang, Minyi Zhao, Guanghao Meng, Ruisheng Luo, Long Chen, and Shu-Tao Xia. Heterogeneous uncertainty-guided composed image retrieval with fine-grained probabilistic learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9386–9394, 2026.

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.

Peng Wang, Zining Chen, Zhicheng Zhao, and Fei Su. Prompting vision-language fusion for zero-shot composed image retrieval. In *Proceedings of the 16th Asian Conference on Machine Learning*, volume 260 of *Proceedings of Machine Learning Research*, pp. 671–686. PMLR, 2025a.

Yunbo Wang, Yujie Wu, Zhien Dai, Can Tian, Jun Long, and Jianhai Chen. Noisy correspondence rectification via asymmetric similarity learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21384–21392, 2025b.

Haokun Wen, Xian Zhang, Xuemeng Song, Yinwei Wei, and Liqiang Nie. Target-guided composed image retrieval. In *Proceedings of the ACM International Conference on Multimedia*, pp. 915–923. ACM, 2023.

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.

Shuo Yang, Zhaopan Xu, Kai Wang, Yang You, Hongxun Yao, Tongliang Liu, and Min Xu. BiCro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19883–19892, 2023.

Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. Magiclens: Self-supervised image retrieval with open-ended instructions. *arXiv preprint arXiv:2403.19651*, 2024a.

Xu Zhang, Zhedong Zheng, Linchao Zhu, and Yi Yang. Collaborative group: Composed image retrieval via consensus learning from noisy annotations. *Knowledge-Based Systems*, 300:112135, 2024b.

A Robustness under Different Noise Ratios

To further analyze robustness, we plot the average retrieval performance under different noise ratios in Figure 3. For clarity, the curves include representative strong baselines. As the noise ratio increases, all methods show clear performance degradation on both FashionIQ and CIRR. Flow-CIR generally shows a slower degradation trend and achieves strong average performance under high-noise settings, especially when the noise ratio reaches 0.8. This indicates that feature-level generative rectification helps the model better exploit suspicious triplets and alleviate the negative effect of noisy supervision.

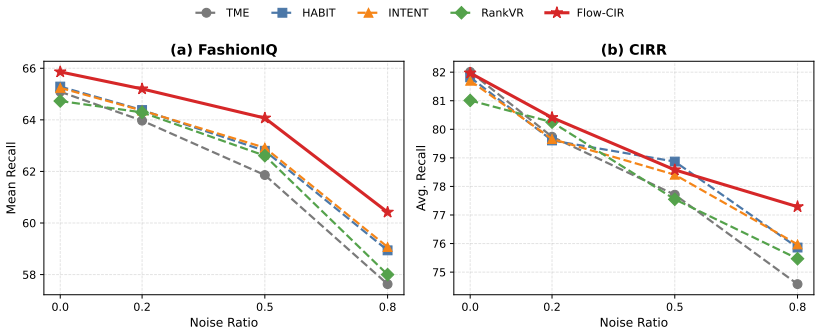


Figure 3: Robustness curves under different noise ratios on FashionIQ and CIRR. The x-axis denotes the noise ratio, and the y-axis denotes the average retrieval performance.

B Parameter Sensitivity

Table 4 analyzes the sensitivity of λ_{fm}, α, and the memory queue size under the noise ratio of 0.5. The best results are obtained around $\lambda_{fm} = 0.10$, $\alpha = 0.20$, and a queue size of 128 on both datasets. Smaller values provide insufficient rectification, while overly large values may introduce excessive correction or stale negatives. Overall, Flow-CIR remains stable within a reasonable parameter range.

Table 4: Parameter sensitivity analysis of Flow-CIR on FashionIQ and CIRR under the noise ratio of 0.5. Each parameter is varied independently while keeping the others fixed.

Parameter	Value	FashionIQ Mean	CIRR Avg.
λ_{fm}	0.01	63.65	77.62
	0.05	63.91	78.14
	0.10	**64.07**	**78.58**
	0.20	63.78	77.74
α	0.10	63.82	77.79
	0.20	**64.07**	**78.58**
	0.30	63.74	78.18
Queue size	0	63.76	76.71
	128	**64.07**	**78.58**
	256	64.03	77.98

C Visualization and Analysis

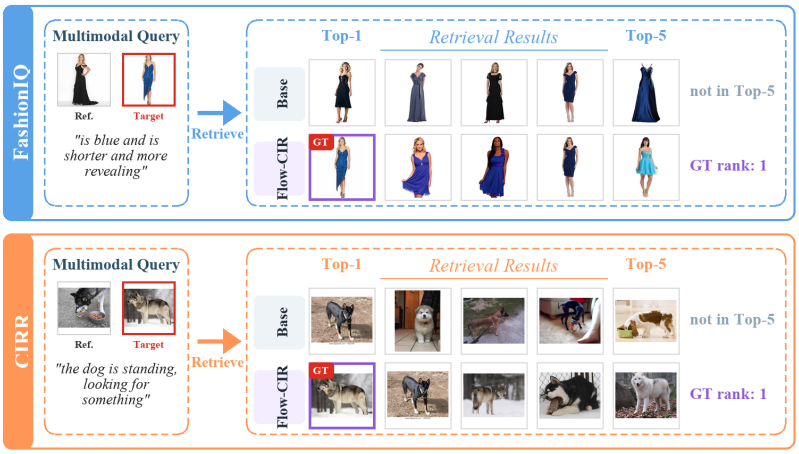
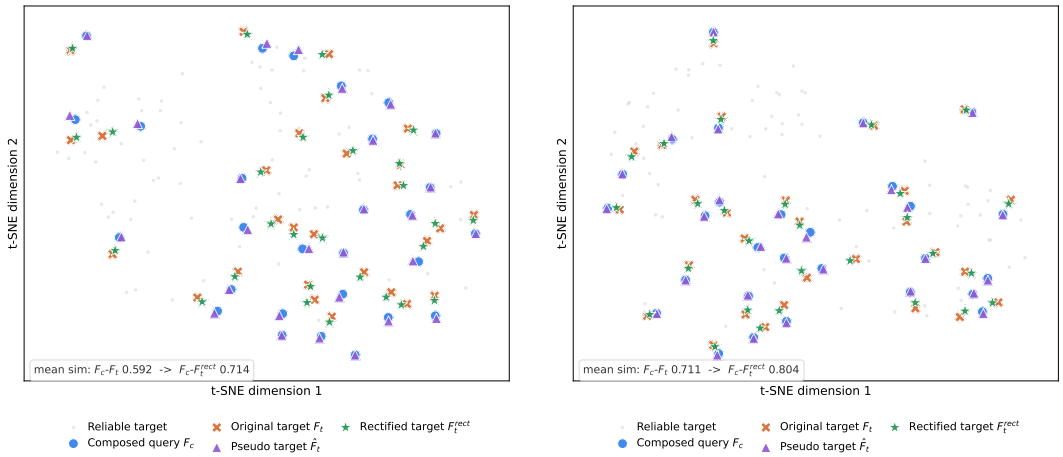


Figure 4: Qualitative retrieval examples on FashionIQ and CIRR. Given the same multimodal query, the base model retrieves visually related but semantically inconsistent images, while Flow-CIR retrieves the ground-truth target within the top-ranked results.



(a) FashionIQ (b) CIRR

Figure 5: Feature-level rectification visualization on FashionIQ and CIRR. The rectified target features move closer to the composed query features, indicating that Flow-CIR can adjust suspicious targets toward more query-consistent semantic directions.

To intuitively analyze the effect of Flow-CIR, we present qualitative retrieval examples on FashionIQ and CIRR in Figure 4. Given the same multimodal query, the base model retrieves visually related but semantically inconsistent images, and fails to include the ground-truth target in the top-5 results. In contrast, Flow-CIR successfully retrieves the correct target at rank 1 on both datasets. These examples show that feature-level rectification helps correct the retrieval direction and improves the semantic consistency between the modified query and the target image.

Figure 5 further visualizes the feature-level rectification process with t-SNE. On both FashionIQ and CIRR, the rectified target features F_t^{rect} are generally moved closer to the composed query features F_c than the original target features F_t. The average query-target similarity also increases after rectification, which qualitatively supports that Flow-CIR adjusts suspicious targets toward more query-consistent semantic directions.