TRUST: SEMANTIC RELATION-AWARE ROBUST LEARNING FRAMEWORK FOR COMPOSED IMAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Composed Image Retrieval enables flexible image search using a reference image and a modification text, yet suffers from Noisy Triplet Correspondence caused by mismatched triplet components. Existing methods typically discriminate noisy samples via query-target-level signals, while overlooking the intrinsic semantic relations among the reference, modification, and target. Specifically, noisy triplets may still appear reliable when query-target similarity is dominated by unchanged visual content or when the composed query obscures invalid component-target relations. To address these limitations, we propose TRUST: a Semantic Relation-Aware Robust Learning framework, which comprises two coherent and synergistic processes. Transformation-Aware Reliability Estimation first measures continuous triplet reliability on the unit hypersphere by jointly modeling Reference-Anchored Transformation Agreement and Target-oriented Query Decoupling. Then, Reliability-Guided Asymmetric Optimization uses the estimated reliability to weight bidirectional ranking, while shaping the spherical representation space through transformation alignment and target uniformity. We further introduce eXtreme Noisy Triplet Correspondence as a more challenging setting. Experiments on Fashion-IQ and CIRR show that TRUST achieves state-of-the-art performance, with substantial advantages at high noise rates.

1 INTRODUCTION

Composed Image Retrieval (CIR) is one of the most challenging problems in multimodal machine learning (Chen et al., 2020; Liu et al., 2024). A distinctive characteristic of CIR lies in its complex triplet retrieval paradigm, which involves intrinsic semantic correspondences among the reference image, modification text, and target image. Over the years, tremendous efforts have been devoted to CIR, with most existing methods relying on the assumption of ideal and clean training triplets whose components are correctly matched. However, CIR datasets often contain noisy triplets caused by component mismatching, which disrupts the semantic correspondence among triplet components and introduces unreliable supervision. To address this issue, recent studies have investigated the Noisy Triplet Correspondence (NTC) problem (Li et al., 2025a; 2026b; Chen et al., 2026; Fu et al., 2026; Li et al., 2026a). A central strategy in existing NTC methods is to identify noisy triplets through pointwise noise discrimination based on query-target similarity.

Despite these efforts, the intrinsic semantic relations within CIR triplets remain underexplored and are more challenging to capture. Unlike traditional cross-modal data such as text-image pairs (Lee et al., 2018; Yang et al., 2026), which mainly rely on point-to-point matching, CIR triplets involve dynamic semantic relations among their components. In particular, the semantic transformation from the reference to the composed query should be consistent with the actual visual difference from the reference to the target. Such relations provide valuable cues for determining whether the requested modification is faithfully reflected in the target and whether the triplet forms a valid correspondence. In this work, we study CIR from the perspective of intrinsic semantic relations, aiming to move beyond pointwise matching for more reliable noise discrimination for NTC.

Figure 1 illustrates why such a perspective is necessary. As shown in Figure 1(a), when the requested modification is subtle, query-target similarity can be dominated by unchanged visual content even

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.

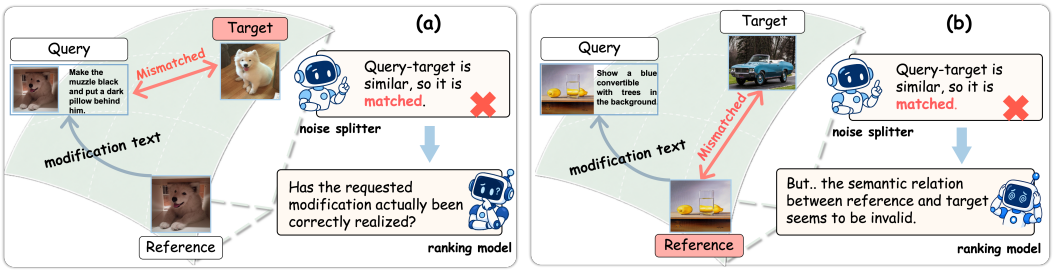

Figure 1: (a) illustrates that query-target similarity can be dominated by unchanged visual content, even when the requested modification is not correctly realized. (b) illustrates that query composition can obscure invalid component-target relations.

if the modification is not correctly realized. In this case, an invalid target may still exhibit high similarity to the composed query because most of its visual content remains similar to the reference image. Moreover, Figure 1(b) shows that composing the reference image and modification text into a unified query inevitably introduces information loss, which may obscure invalid component-target relations during fusion. Consequently, pointwise noise discrimination may become unreliable when visually analogous content dominates similarity estimation or when component-target relations are insufficiently modeled. This limitation can substantially hinder existing robust CIR methods that depend on accurate noise recognition.

Key Challenge: *How can we leverage intra-triplet semantic relations to achieve robust noise discrimination beyond pointwise matching?*

To address this challenge, we propose TRUST, a Seman**T**ic **R**elation-Aware Rob**UST** Learning framework for CIR that explicitly characterizes the intrinsic semantic relations within triplets. TRUST operates on a unit hypersphere via two primary mechanisms. First, *Transformation-Aware Reliability Estimation (TARE)* identifies noisy triplets by jointly evaluating reference-anchored transformation agreement and target-oriented query decoupling. Second, *Reliability-Guided Asymmetric Optimization (RGAO)* learns a well-structured representation space by adjusting each triplet's contribution according to its estimated reliability through reliability-guided ranking and spherical regularization. The motivation behind TRUST is that such intrinsic semantic relations naturally exist in CIR triplets and provide richer supervision for noise discrimination. In particular, comparing the reference-to-query semantic transformation with the actual reference-to-target visual difference offers an intuitive measure of whether the requested modification is faithfully realized. In addition, independently modeling component-target relations helps preserve relational information that may otherwise be obscured during query composition. To summarize, our contributions are as follows:

- **Novel perspective:** We study CIR noise discrimination from the perspective of intrinsic semantic relations. This perspective goes beyond conventional pointwise similarity matching by explicitly exploiting the interdependent semantic structure of CIR triplets.
- **Novel methodology:** We propose TRUST, a robust learning framework for noisy CIR. TRUST incorporates TARE to estimate triplet reliability through transformation agreement and query decoupling, together with RGAO to perform reliability-guided ranking and spherical regularization.
- **Superior performance:** Extensive qualitative and quantitative experiments on Fashion-IQ and CIRR demonstrate the effectiveness of TRUST. We further introduce a more challenging eXtreme Noisy Triplet Correspondence (XNTC) setting. Under both settings, TRUST exhibits slower performance degradation and achieves substantial improvements under high-noise conditions.

2 Related Work

Noisy Correspondence Learning. NCR (Huang et al., 2021) first formalized unreliable cross-modal associations. RCL (Hu et al., 2023) and CRCL (Qin et al., 2023) suppress erroneous gradients through complementary negatives and correspondence refinement. Recently, ReCon (Zha et al.,

2025) further aligns cross-modal and intra-modal relations, while MGCS (Wang et al., 2026) corrects modality-gap-induced directional drift before noise separation. Nevertheless, these methods primarily address noisy correspondence under a point-to-point matching paradigm (Qin et al., 2022; Zha et al., 2024; Shi et al., 2024; Wang et al., 2025; Han et al., 2025).

Composed Image Retrieval. Given a reference image and a text modification, CIR retrieves a target image that realizes the requested change (Yan et al., 2024; Huynh et al., 2025). TIRG (Vo et al., 2019) learned explicit fusion, while CLIP4CIR (Baldrati et al., 2022) leveraged vision-language pretraining. Pic2Word (Saito et al., 2023) and SPRC (Bai et al., 2024) recast the reference as a pseudo-word or dynamic sentence prompt (Baldrati et al., 2023; Suo et al., 2024). Recent methods further refine relation modeling (Wen et al., 2023; Jang et al., 2024; Shi et al., 2025): CCIN (Tian et al., 2025) neutralizes reference-instruction attribute conflicts, whereas ENCODER (Li et al., 2025b) mines visual entities and modification actions and binds their relations. However, most CIR methods assume clean training triplets, overlooking component mismatching noise that disrupts intrinsic semantic relations and introduces unreliable supervision.

Noisy Triplet Correspondence. TME (Li et al., 2025a) first studied robust CIR under NTC. HABIT (Li et al., 2026b) stabilizes mutual-knowledge cleanliness, INTENT (Chen et al., 2026) separates correspondence noise from modality-inherent interference, Air-Know (Fu et al., 2026) decouples an external multimodal-LLM arbiter from the learner, and ConeSep (Li et al., 2026a) combines a cone-space boundary with targeted unlearning. Despite these advances, these methods still rely on pointwise noise discrimination, overlooking the intra-triplet semantic relations within triplets. In contrast, TRUST jointly models reference-anchored transformation agreement and target-oriented query decoupling to estimate triplet reliability and guide robust retrieval optimization.

3 METHODOLOGY

3.1 PROBLEM FORMULATION

Composed image retrieval (CIR) uses a multimodal query, comprising a reference image R and modification text M, to retrieve a target image T that realizes the requested change. The training set can be represented as $\mathcal{D} = \{(R_i, M_i, T_i)\}_{i=1}^{N}$. We consider the Noisy Triplet Correspondence (NTC) setting, where the training set is corrupted by randomly shuffling reference, modification, or target fields, resulting in a noisy training set $\widetilde{\mathcal{D}}$. It can be viewed as $\widetilde{\mathcal{D}} = \mathcal{D}_{\text{clean}} \cup \mathcal{D}_{\text{noise}}$, where the ground-truth clean/noisy labels are unavailable during training. The goal is to train a robust model from $\widetilde{\mathcal{D}}$ that ranks the correct targets highly for test queries while preventing unreliable triplets from distorting the learned representation space. To address this problem, we propose TRUST, which consists of two key modules, TARE and RGAO. The overall framework is illustrated in Figure 2.

3.2 TRANSFORMATION-AWARE RELIABILITY ESTIMATION (TARE)

To faithfully estimate triplet reliability by examining the intra-triplet semantic relations, we introduce Transformation-Aware Reliability Estimation. To assess whether the requested modification is truly realized and whether query composition obscures failures in the intrinsic semantic relations, TARE explores relations from two perspectives, namely *Reference-Anchored Transformation Agreement* and *Target-oriented Query Decoupling*, yielding transformation reliability and composed reliability, respectively. These two reliabilities are then integrated to obtain the final triplet reliability.

Reference-anchored transformation agreement. Query-target (QT) proximity alone may fail to verify that M_i describes the change from R_i to T_i, because for subtle modifications, an unchanged visual subject may dominate their similarity. We instead measure the agreement between the reference-to-query semantic transformation and the actual reference-to-target visual transformation. We model the normalized embeddings on the unit hypersphere $\mathbb{S}^{d-1} = \{z \in \mathbb{R}^d : \|z\|_2 = 1\}$ and represent both transformations in the reference-anchored tangent space. For $z, z^{'} \in \mathbb{S}^{d-1}$, we use the numerically stabilized form of the Riemannian logarithm map at z (Lou et al., 2020):

$$\mathrm{Log}_z(z^{'}) = \frac{\arccos(z^\top z^{'})}{\max\Big(\sqrt{1-(z^\top z^{'})^2}, \epsilon\Big)}\big(z^{'} - (z^\top z^{'})z\big). \tag{1}$$

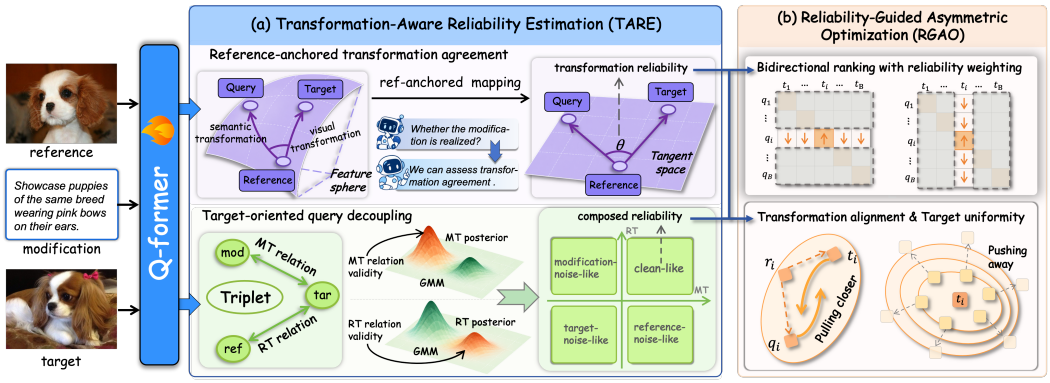

Figure 2: TRUST comprises (a) Transformation-Aware Reliability Estimation (TARE) and (b) Reliability-Guided Asymmetric Optimization (RGAO). TARE estimates transformation and composed reliability through Reference-Anchored Transformation Agreement and Target-oriented Query Decoupling, respectively, and combines them into triplet reliability. RGAO uses triplet reliability to guide bidirectional ranking and spherical regularization.

Let q_i and m_i denote the ℓ_2-normalized embeddings of the composed query (R_i, M_i) and the text-only modification M_i, respectively. Let r_i and t_i be the ℓ_2-normalized reference and target tokens. To characterize the transformations within a triplet, using r_i as the shared tangent origin, we define $\Delta_i^{\text{mod}} = \text{Log}_{r_i}(q_i)$ and $\Delta_i^{\text{vis}} = \text{Log}_{r_i}(t_i)$, which respectively represent the modification-guided transformation and the actual visual transformation. We measure the transformation disagreement:

$$g_i = 1 - \frac{\langle \Delta_i^{\text{mod}}, \Delta_i^{\text{vis}} \rangle}{\|\Delta_i^{\text{mod}}\|_2 \|\Delta_i^{\text{vis}}\|_2}. \tag{2}$$

A smaller g_i indicates stronger transformation agreement. We fit a two-component Gaussian mixture model (GMM) (Dempster et al., 1977) to the disagreement scores. Let $\mathcal{G}^g_{\text{low}}$ denote the posterior probability associated with the lower component. The transformation reliability p_i^{trans} is then defined as $p_i^{\text{trans}} = \mathcal{G}^g_{\text{low}}(g_i)$.

Transformation agreement alone may misclassify an unusually distant target aligned with the same tangent direction. We therefore apply a veto using the modification-target (MT) evidence $e_i^{MT} = s(m_i, t_i)$ and the QT evidence $e_i^{QT} = s(q_i, t_i)$, where $s(\cdot, \cdot)$ denotes cosine similarity. We define the reference-target (RT) displacement magnitude as $d_i^{RT} = \|\Delta_i^{\text{vis}}\|_2$, and let μ^{RT} and σ^{RT} denote the mean and standard deviation of $\{d_n^{RT}\}_{n=1}^N$, respectively. The corrected transformation reliability is

$$\widehat{p}_i^{\text{trans}} = \begin{cases} 0, & d_i^{RT} > \mu^{RT} + \sigma^{RT} \text{ and } e_i^{QT} < e_i^{MT}, \\ p_i^{\text{trans}}, & \text{otherwise.} \end{cases} \tag{3}$$

This veto is triggered only when the observed reference-to-target displacement exceeds the dataset mean by more than one standard deviation, and the paired target receives a higher score from the text-only query than from the composed query, thereby preventing legitimate large modifications from being rejected based solely on displacement magnitude.

Remark. The reference can be viewed as the starting point of the transformation, and the target as its endpoint. The query is a modification-guided prediction of the endpoint from the starting point. For a matched triplet, the prediction should lie in the same direction as the target with respect to the reference. The transformation agreement is designed to capture this intuition.

Target-oriented query decoupling. Query composition may obscure mismatches in the reference-target or modification-target relations due to information loss. To alleviate this issue, we model these two relations separately and then jointly evaluate the validity. Two independent GMMs map the corresponding evidence to soft component posteriors: $p_i^{MT} = \mathcal{G}^{MT}_{\text{high}}(e_i^{MT})$ and $p_i^{RT} = \mathcal{G}^{RT}_{\text{low}}(d_i^{RT})$. These two cues define four soft states. Let $\boldsymbol{\pi}_i = (\pi_i^C, \pi_i^R, \pi_i^M, \pi_i^T)^\top$ denote the

corresponding state weights. We derive the joint distribution from the two marginal posteriors:

$$\begin{bmatrix} \pi_i^C & \pi_i^R \\ \pi_i^M & \pi_i^T \end{bmatrix} = \begin{bmatrix} p_i^{MT} p_i^{RT} & p_i^{MT}(1-p_i^{RT}) \\ (1-p_i^{MT})p_i^{RT} & (1-p_i^{MT})(1-p_i^{RT}) \end{bmatrix}. \tag{4}$$

The four entries correspond to clean-like, reference-noise-like, modification-noise-like, and target-noise-like relation patterns, respectively. Since the clean-like state indicates that both relations remain valid, we define its state weight as the composed reliability $p_i^{\text{comp}} = \pi_i^C$.

We combine the transformation and composed reliabilities to obtain the final triplet reliability as $p_i = (1-\rho) \cdot \widehat{p}_i^{\,\text{trans}} + \rho \cdot p_i^{\text{comp}}$, which reflects the reliability of the supervision provided by each triplet and is used to guide subsequent training.

3.3 Reliability-Guided Asymmetric Optimization (RGAO)

TARE estimates the reliability of each triplet, but a further question remains: *how can such reliability guide robust training and prevent noisy supervision from distorting the learned retrieval space*? To this end, we introduce Reliability-Guided Asymmetric Optimization (RGAO), which assigns different training weights to triplets according to their reliability and combines reliability-guided ranking with spherical regularization. The latter comprises transformation alignment and target uniformity, which jointly shape a well-structured spherical representation space.

Bidirectional ranking with reliability weighting. For a minibatch of size B, let $\mathbf{S} \in \mathbb{R}^{B\times B}$ denote the score matrix between queries and targets, where $S_{ij} = s(q_i, t_j)$ is computed following a MaxSim paradigm(Yao et al., 2022). The same triplet reliability p_i weights both directional objectives, applied to query anchor i in the forward direction and target anchor i in the reverse direction. In both directions, the paired diagonal entry is the sole positive, while all remaining entries in the batch serve as negatives. For each sample i, we define the one-direction margin ranking as

$$\ell_i(\mathbf{S}) = \left[m_{\text{rank}} - S_{ii} + \tau_{\text{rank}} \log \sum\nolimits_{j\neq i} \exp\left(\frac{S_{ij}}{\tau_{\text{rank}}}\right) \right]_+ . \tag{5}$$

As τ_{rank} decreases, the log-sum-exp term(Qin et al., 2024) places greater emphasis on hard negatives while retaining gradients from other competitive candidates. Applying the loss to $\mathbf{S}$ and $\mathbf{S}^\top$, the reliability-weighted bidirectional ranking objective is

$$\mathcal{L}_{\text{rank}} = \frac{\sum_{i=1}^{B} p_i \left(\ell_i(\mathbf{S}) + \ell_i(\mathbf{S}^\top)\right)}{\sum_{i=1}^{B} p_i}. \tag{6}$$

Transformation alignment. The ranking objective encourages the paired target to score higher than in-batch negatives, but satisfying this relative constraint does not ensure close alignment between the composed query and its paired target. Following the MaxSim paradigm, we use the highest-scoring token t_i^* from the paired target and align it with the composed query q_i in the shared tangent space anchored at the reference: $d_i^{\text{align}} = \left\| \text{Log}_{r_i}(q_i) - \text{Log}_{r_i}(t_i^*) \right\|_2$ This regularizer penalizes discrepancies between the transformations. The reliability-weighted transformation alignment loss is (Girshick, 2015) :

$$H(d) = \begin{cases} \alpha \cdot d^2, & d < 1, \\ d - \beta, & d \geq 1, \end{cases} \qquad \mathcal{L}_{\text{align}} = \frac{\sum_i p_i H(d_i^{\text{align}})}{\sum_i p_i}. \tag{7}$$

Target uniformity. Alignment alone may contract target representations. To preserve an informative representation space, target representations should be uniformly distributed over the hypersphere rather than concentrated in a narrow region. We therefore introduce a target uniformity regularizer to promote such a distribution:

$$\mathcal{L}_{\text{uniform}} = \log \left[\frac{1}{\binom{B}{2}} \sum\nolimits_{1\leq i<j\leq B} \exp\left(-2\|t_i^* - t_j^*\|_2^2\right) \right]. \tag{8}$$

Together, these two regularizers complement each other by balancing transformation alignment and target uniformity, thereby maintaining a discriminative spherical representation space that facilitates TARE assessment in the next epoch.

3.4 MODEL TRAINING AND INFERENCE

Warmup. TARE is unreliable with initialized representations, so we first warm-start the shared retrieval model for E_{w} epochs using Robust Contrastive Learning (RCL) (Hu et al., 2023). RCL converts each row into probabilities

$$P_{ij}^{\mathrm{RCL}} = \frac{\exp(S_{ij}/\tau_{\mathrm{model}})}{\sum_{l=1}^{B} \exp(S_{il}/\tau_{\mathrm{model}})}, \tag{9}$$

where τ_{model} is a temperature parameter that controls the sharpness of the distribution, with larger values yielding softer probabilities and smaller values producing more concentrated assignments. The warmup objective is :

$$\mathcal{L}_{\mathrm{RCL}} = -\frac{1}{B} \sum\nolimits_{i=1}^{B} \sum\nolimits_{j \neq i} \log\bigl(1 - P_{ij}^{\mathrm{RCL}}\bigr)\,. \tag{10}$$

By suppressing off-diagonal probabilities rather than explicitly attracting the annotated target, RCL provides a noise-tolerant initialization for subsequent reliability estimation.

Optimization. For each minibatch, let λ_{a} and λ_{u} denote alignment and uniformity weights, and their weighted absolute magnitudes define the auxiliary budget for spherical regularization: $l_{reg} = \lambda_{\mathrm{a}}|\mathcal{L}_{\mathrm{align}}|+\lambda_{\mathrm{u}}|\mathcal{L}_{\mathrm{uniform}}|$. To adapt this budget to the ranking-loss, we set $\zeta_b = \min(1, |\mathcal{L}_{\mathrm{rank}}|/l_{reg})$, i.e., the scaled auxiliary contribution never exceeds the magnitude of $\mathcal{L}_{rank}$. The training objective:

$$\mathcal{L} = \begin{cases} \mathcal{L}_{\mathrm{RCL}}, & 0 \leq e < E_{\mathrm{w}}, \\ \mathcal{L}_{\mathrm{rank}} + \zeta_b(\lambda_{\mathrm{a}}\mathcal{L}_{\mathrm{align}} + \lambda_{\mathrm{u}}\mathcal{L}_{\mathrm{uniform}})\,, & e \geq E_{\mathrm{w}}. \end{cases} \tag{11}$$

The rank loss optimizes query-target discrimination, the alignment loss promotes consistency between the reference-to-query and reference-to-target transformations, and the target uniformity loss encourages representations to be uniformly distributed over the hypersphere.

Inference. At test time, the encoder produces the composed query and gallery image embeddings, whose similarities are computed with MaxSim to retrieve the highest-scoring image.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Implementation details. We instantiate the retrieval model with a pretrained BLIP-2 backbone (Li et al., 2023). We optimize the model with AdamW (Loshchilov & Hutter, 2019), using a batch size of 128, a learning rate of 1×10^{-5}. Training lasts for at most 30 epochs. We optimize the rank loss with temperature $\tau_{\mathrm{rank}} = 0.015$ and margin $m_{\mathrm{rank}} = 0.1$. The alignment and uniformity weights are set to $\lambda_{\mathrm{a}} = 0.1$ and $\lambda_{\mathrm{u}} = 0.01$, respectively, and the reliability fusion coefficient is set to $\rho = 0.4$. We set $\alpha = \beta = \frac{1}{2}$ in the transformation alignment loss and seed 42 for both noise generation and model training. All experiments are conducted on NVIDIA GeForce RTX 3090 GPUs.

Datasets and metrics. We evaluate on Fashion-IQ (Wu et al., 2021) and CIRR (Liu et al., 2021), which assess fine-grained fashion modification and open-domain compositional reasoning, respectively. For Fashion-IQ, we report validation Recall@10 and Recall@50 across three categories and their overall mean. For CIRR, we follow the official test protocol and report full-gallery Recall@K for $K \in 1, 5, 10, 50$ and subset Recall@K for $K \in 1, 2, 3$ within six-image subsets.

4.2 PERFORMANCE COMPARISON

To evaluate the robustness of TRUST under NTC, we compare it with the standard CIR baseline SPRC (Bai et al., 2024) and five robust methods: TME (Li et al., 2025a), HABIT (Li et al., 2026b), INTENT (Chen et al., 2026), Air-Know (Fu et al., 2026), and ConeSep (Li et al., 2026a). Tables 1 and 2 report the results on CIRR and Fashion-IQ at training-noise rates of 0%, 20%, 50%, and 80%. We draw two main observations.

Table 1: Comparison on the CIRR test set under different training-noise rates.

Noise	Methods	R@K				$\mathrm{R_{sub}}@K$			Avg.($\mathrm{R@5}$, $\mathrm{R_{sub}@1}$)
		$K=1$	$K=5$	$K=10$	$K=50$	$K=1$	$K=2$	$K=3$	
0%	SPRC (2024)	51.96	82.12	89.74	97.69	80.65	92.31	96.60	81.39
	TME (2025a)	53.42	82.99	90.24	98.15	81.04	92.58	96.94	82.01
	HABIT (2026b)	52.71	82.64	90.63	98.19	80.99	**92.77**	97.00	81.82
	INTENT (2026)	53.37	83.16	90.73	98.22	80.24	92.37	96.89	81.70
	Air-Know (2026)	51.97	82.25	90.42	98.25	79.21	91.94	96.87	80.73
	ConeSep (2026a)	53.06	83.45	90.60	98.34	**81.23**	92.62	**97.24**	**82.34**
	Ours	**54.12**	**83.71**	**91.08**	**98.51**	79.42	91.76	96.89	81.57
20%	SPRC (2024)	45.90	75.86	83.52	93.37	78.10	91.40	96.05	76.98
	TME (2025a)	51.35	81.01	88.53	97.81	78.46	91.25	96.39	79.74
	HABIT (2026b)	51.68	81.02	89.24	97.81	78.20	91.66	96.75	79.61
	INTENT (2026)	51.25	81.36	90.02	98.05	77.95	91.40	96.46	79.66
	Air-Know (2026)	51.37	81.73	89.78	98.07	79.06	91.66	96.54	80.40
	ConeSep (2026a)	52.29	82.19	89.98	98.19	78.66	**91.76**	96.75	80.43
	Ours	**53.18**	**83.18**	**90.60**	**98.29**	**79.33**	91.64	**97.04**	**81.25**
50%	SPRC (2024)	39.93	66.00	73.59	86.48	75.81	89.21	95.37	70.90
	TME (2025a)	48.48	78.94	87.28	96.99	76.48	90.07	95.83	77.71
	HABIT (2026b)	50.32	79.63	88.34	97.06	76.84	90.60	96.27	78.24
	INTENT (2026)	49.78	79.64	88.99	97.37	77.18	90.41	96.00	78.41
	Air-Know (2026)	50.82	80.42	88.37	97.83	77.43	90.68	96.24	78.93
	ConeSep (2026a)	51.28	80.46	88.46	97.49	77.04	91.04	96.60	78.75
	Ours	**53.59**	**82.27**	**90.00**	**98.12**	**79.01**	**91.61**	**96.96**	**80.64**
80%	SPRC (2024)	29.95	51.25	58.51	73.86	70.22	86.05	93.21	60.74
	TME (2025a)	46.31	75.78	84.89	95.83	73.37	88.02	94.89	74.58
	HABIT (2026b)	47.93	76.84	85.95	95.90	74.87	89.08	95.21	75.86
	INTENT (2026)	47.90	78.13	87.04	96.47	73.81	89.18	95.54	75.97
	Air-Know (2026)	48.84	78.50	87.10	96.47	75.30	89.28	95.42	76.90
	ConeSep (2026a)	48.00	78.53	87.01	97.40	74.22	89.52	95.42	76.38
	Ours	**51.76**	**80.75**	**89.16**	**97.42**	**79.30**	**92.22**	**96.75**	**80.02**

1) TRUST's advantage grows as noise increases. TRUST achieves the best aggregate score on both datasets at all nonzero noise rates, with its gains over the strongest baseline widening from 0.82 to 3.12 points on CIRR and from 0.09 to 2.04 points on Fashion-IQ as the noise rate increases from 20% to 80%. In particular, its strong R@1 performance demonstrates that TRUST better captures the changes specified by the modification text, enabling precise retrieval of the intended target rather than merely identifying broadly similar images. At 0% noise, TRUST remains competitive but does not achieve the best performance, as reliability weighting may slightly downweight informative samples when noisy correspondences are absent.

2) TRUST degrades substantially more slowly as noise increases. From 0% to 80% noise, TRUST's aggregate score decreases by only 1.55 points on CIRR, compared with up to 7.43 points for competing robust methods; a similar advantage is observed on Fashion-IQ. This slower degradation stems from the complementary roles of TARE and RGAO: TARE estimates triplet reliability by exploring the intra-triplet semantic relations, while RGAO uses the reliability to downweight unreliable ranking supervision and maintain the well-structured spherical embedding space. They enable TRUST to remain robust even when most training triplets are mismatched.

To simulate a more realistic and challenging noise scenario, we further introduce Extreme Noisy Triplet Correspondence (XNTC) based on the standard NTC setting. Under high noise rates, XNTC randomly shuffles all components of each noisy triplet. We conduct additional experiments under the XNTC setting, with the results reported in Table 3 and 4. The significant advantage of TRUST demonstrates that it can maintain robust training and prevent wrong supervision from distorting the learned retrieval space under even extreme noise conditions.

4.3 Ablation studies and more analysis

We conduct ablation studies on Fashion-IQ and CIRR, as shown in Table 5. **Effectiveness of TARE.** Removing either transformation reliability $\widehat{p}_i^{\mathrm{trans}}$ (D#1) or compatibility reliability p_i^{comp} (D#2) consistently degrades performance, highlighting their respective roles in verifying the requested modification and detecting relation mismatches obscured by query composition. Removing the complete TARE module (D#3) yields the largest drop in this group, confirming that the two cues complement each other in triplet reliability estimation. **Effectiveness of RGAO.** Removing the transformation alignment loss (D#4) degrades performance, highlighting the importance of reference-anchored transformation alignment. Removing the target uniformity loss (D#5) also reduces performance, supporting its role in preventing representation contraction. Removing both regularizers (D#6) fur-

Table 2: Comparison on the Fashion-IQ validation set under different training-noise rates.

Noise	Methods	Dress		Shirt		Toptee		Average		
		R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	AVG.
0%	SPRC (2024)	49.18	72.43	55.64	73.89	59.35	78.58	54.72	74.97	64.85
	TME (2025a)	49.73	71.69	56.43	74.44	59.31	78.94	55.15	75.02	65.09
	HABIT (2026b)	49.99	72.38	56.62	74.68	**59.51**	78.53	55.38	75.20	65.29
	INTENT (2026)	50.32	72.10	56.32	74.93	59.28	78.45	55.31	75.16	65.24
	Air-Know (2026)	51.03	73.08	56.76	75.05	59.47	79.02	**55.75**	75.72	**65.73**
	ConeSep (2026a)	50.96	73.02	**56.98**	**75.22**	58.80	**79.40**	55.58	**75.88**	**65.73**
	Ours	**51.07**	**73.43**	56.43	75.02	59.36	78.12	55.62	75.52	65.57
20%	SPRC (2024)	39.81	62.22	48.58	66.29	50.48	70.58	46.29	66.36	56.33
	TME (2025a)	49.03	70.35	55.84	73.16	57.22	78.23	54.03	73.91	63.97
	HABIT (2026b)	49.63	71.34	55.67	73.19	58.14	78.32	54.48	74.28	64.38
	INTENT (2026)	49.32	71.43	55.32	73.57	58.01	78.46	54.22	74.49	64.36
	Air-Know (2026)	50.22	**73.48**	56.38	74.73	58.95	78.94	55.18	75.71	65.45
	ConeSep (2026a)	**50.23**	72.19	56.28	74.45	58.29	78.38	54.93	75.01	64.97
	Ours	49.83	73.08	**56.43**	**75.07**	**59.77**	**79.09**	**55.34**	**75.75**	**65.54**
50%	SPRC (2024)	35.94	57.16	42.25	61.63	44.98	64.76	41.06	61.19	51.12
	TME (2025a)	46.26	68.27	53.09	71.88	55.07	76.59	51.47	72.25	61.86
	HABIT (2026b)	47.33	69.71	53.72	72.55	56.51	77.00	52.52	73.09	62.80
	INTENT (2026)	47.99	71.24	52.78	72.48	56.79	76.23	52.52	73.32	62.92
	Air-Know (2026)	47.90	71.09	53.95	73.09	57.71	77.94	53.19	74.04	63.61
	ConeSep (2026a)	48.39	71.34	54.56	73.25	57.22	77.87	53.39	74.15	63.77
	Ours	**48.79**	**71.79**	**55.79**	**73.85**	**58.69**	**78.89**	**54.42**	**74.84**	**64.63**
80%	SPRC (2024)	28.41	50.77	36.21	54.37	35.90	59.06	33.51	54.73	44.12
	TME (2025a)	41.45	64.35	47.30	68.20	51.25	73.23	46.67	68.60	57.63
	HABIT (2026b)	42.04	65.20	50.12	69.77	52.92	73.61	48.36	69.53	58.94
	INTENT (2026)	42.07	65.58	50.38	69.41	53.09	73.91	48.51	69.63	59.07
	Air-Know (2026)	42.52	66.03	51.31	70.26	53.86	74.41	49.23	70.23	59.73
	ConeSep (2026a)	43.33	66.01	51.23	70.35	53.19	74.55	49.25	70.30	59.77
	Ours	**45.36**	**68.52**	**53.34**	**71.79**	**55.38**	**76.49**	**51.36**	**72.27**	**61.81**

Table 3: Comparison on the Fashion-IQ validation set under XNTC setting.

Noise	Methods	Dress		Shirt		Toptee		Average		
		R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	AVG.
50%	SPRC (2024)	46.11	68.62	52.80	71.79	55.38	76.39	51.43	72.26	61.85
	TME (2025a)	47.35	68.72	53.19	72.67	54.82	77.15	51.79	72.85	62.32
	HABIT (2026b)	47.74	70.35	**55.15**	71.98	55.79	75.73	52.89	72.69	62.79
	INTENT (2026)	46.55	69.01	54.07	73.06	55.33	75.78	51.98	72.62	62.30
	ConeSep (2026a)	**49.53**	70.40	54.86	73.16	57.01	76.24	**53.80**	73.27	63.53
	Ours	48.73	**71.19**	54.71	**74.09**	**57.92**	**77.51**	53.79	**74.26**	**64.02**
80%	SPRC (2024)	41.70	64.45	49.51	68.55	50.89	72.26	47.37	68.42	57.89
	TME (2025a)	42.24	63.61	49.90	69.09	51.91	72.67	48.02	68.45	58.24
	HABIT (2026b)	39.66	61.03	48.18	65.06	49.06	68.43	45.63	64.84	55.24
	INTENT (2026)	0.89	2.52	0.39	1.32	0.35	1.63	0.54	1.82	1.19
	ConeSep (2026a)	44.86	66.53	50.44	68.98	52.67	71.08	49.32	68.86	59.09
	Ours	**45.46**	**67.67**	**52.60**	**71.34**	**55.43**	**75.06**	**51.16**	**71.36**	**61.26**

ther confirms their joint contribution. The degradation without the rank loss (D#7) confirms that reliability-weighted bidirectional ranking provides essential retrieval supervision.

Table 6 evaluates the effectiveness and transferability of TARE and RGAO on Fashion-IQ by integrating them individually into TME. Both modules consistently yield clear performance improvements, demonstrating their ability to enhance the performance of existing frameworks. Similar experiments on the CIRR dataset are reported in Table 8.

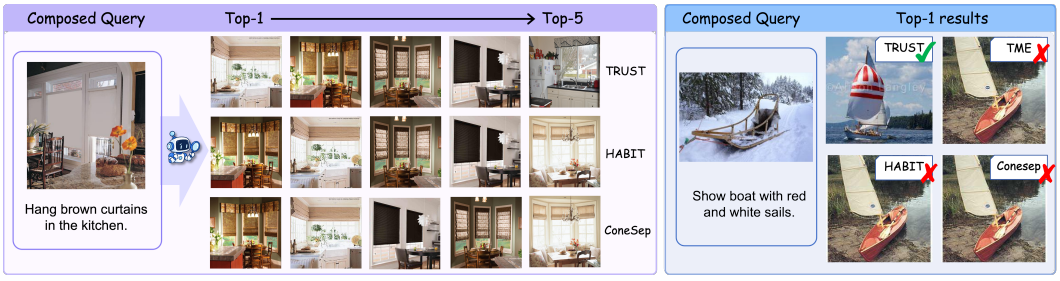

Figure 3: Qualitative Retrieval Comparisons

Table 4: CIRR comparison under XNTC.

Noise	Methods	Avg. R@K	Avg. R_{sub}@K	Avg. (R, R_{sub})
50%	TME (2025a)	79.32	88.15	79.26
	HABIT (2026b)	78.69	87.29	78.21
	INTENT (2026)	78.44	86.65	77.42
	ConeSep (2026a)	78.90	87.79	78.89
	Ours	**80.44**	**88.62**	**79.46**
80%	TME (2025a)	77.01	87.31	77.04
	HABIT (2026b)	75.11	84.93	74.61
	INTENT (2026)	72.76	83.72	71.86
	ConeSep (2026a)	74.58	84.96	73.98
	Ours	**78.28**	**87.86**	**77.33**

Table 5: Ablation study on two datasets.

D#	Deriv.	FashionIQ-Avg R@10	FashionIQ-Avg R@50	CIRR-Avg R@K	CIRR-Avg R_{sub}@K
	(a) Transformation-Aware Reliability Estimation (TARE)				
1	w/o $\widehat{p}_i^{\text{trans}}$	55.06	75.20	80.77	88.95
2	w/o p_i^{comp}	55.13	75.17	81.16	89.08
3	w/o TARE	54.98	75.03	80.69	88.76
	(b) Reliability-Guided Asymmetric Optimization (RGAO)				
4	w/o $\mathcal{L}_{\text{align}}$	54.51	74.87	80.92	89.10
5	w/o $\mathcal{L}_{\text{uniform}}$	55.22	75.34	81.14	89.13
6	w/o $\mathcal{L}_{\text{align}} + \mathcal{L}_{\text{uniform}}$	54.82	75.04	80.77	88.95
7	w/o $\mathcal{L}_{\text{rank}}$	31.52	50.72	50.10	61.22
	TRUST (full)	**55.34**	**75.75**	**81.31**	**89.33**

Table 6: Integration of TARE or RGAO into TME on FashionIQ.

Noise rate	Method	Dress R@10	Dress R@50	Shirt R@10	Shirt R@50	Toptee R@10	Toptee R@50	Average R@10	Average R@50	Average Avg.
20%	TME	49.03	70.35	55.84	73.16	57.22	78.23	54.03	73.91	63.97
	TME+TARE	48.29	71.00	53.93	72.87	57.47	77.77	53.23	73.88	63.55
	TME+RGAO	**49.88**	72.38	55.79	74.34	58.39	77.87	54.68	74.86	64.77
	TRUST	49.83	**73.08**	**56.43**	**75.07**	**59.77**	**79.09**	**55.34**	**75.75**	**65.54**
50%	TME	46.26	68.27	53.09	71.88	55.07	76.59	51.47	72.25	61.86
	TME+TARE	46.41	68.77	52.50	72.03	55.84	76.44	51.58	72.41	62.00
	TME+RGAO	47.60	70.30	54.56	73.55	58.34	77.46	53.50	73.77	63.64
	TRUST	**48.79**	**71.79**	**55.79**	**73.85**	**58.69**	**78.89**	**54.42**	**74.84**	**64.63**

Sensitivity to the routing coefficient ρ. The hyperparameter ρ balances the transformation reliability $\hat{p}_i^{\text{trans}}$ and the compatibility reliability p_i^{comp}. As shown in Figure 4, we vary ρ from 0.2 to 0.7 at a noise rate of 20%. As ρ increases, the performance first increases and then decreases, reaching its peak at $\rho = 0.4$. An excessively low ρ underutilizes the component-target relations, making it harder to detect mismatches obscured by query composition; an excessively high ρ weakens the role of reference-anchored transformation agreement, making it harder to directly assess whether the target actually realizes the intended modification.

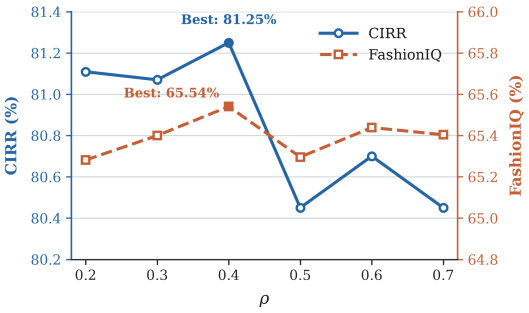

Figure 4: Sensitivity analysis of ρ.

Figure 3 presents qualitative retrieval comparisons. For the kitchen query, TRUST ranks an image with brown window coverings first. HABIT and ConeSep place the same image second, favoring a dining scene that is less consistent with the requested kitchen setting. For the boat query, TRUST retrieves a boat with red-and-white sails, whereas TME, HABIT, and ConeSep return a boat with a red hull and a plain sail. This contrast highlights the importance of associating the requested colors with the sails, rather than merely retrieving a boat containing those colors. Together, these examples illustrate that TRUST follows modification instructions more faithfully while better preserving the relevant scene context.

5 Conclusion

In this paper, we revisited Noisy Triplet Correspondence in Composed Image Retrieval from a relational perspective. We showed that query-target-level noise discrimination can be unreliable because it overlooks the intrinsic semantic relations within triplets. Based on this insight, we proposed TRUST, which models the intra-triplet semantic relations on the unit hypersphere to estimate triplet reliability and guide robust retrieval learning. Extensive results on Fashion-IQ and CIRR demonstrated that TRUST achieved state-of-the-art performance under both NTC and XNTC settings.

AI USE STATEMENT

We used GPT-5.6-series models to assist with literature search and organization, checking mathematical derivations, experimental design, organization of experimental data, visualization and analysis of experimental results, and checking LaTeX formatting. We did not use generative AI to conceive or brainstorm the research ideas. We did not use generative AI to draft the manuscript text. We manually reviewed the AI-assisted work, cross-checked the literature, verified the mathematical derivations, and checked computational results by running code or comparing them with experimental logs. We take full responsibility for the manuscript and associated research artifacts, including the accuracy, originality, and integrity of content developed with AI assistance.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide the TRUST implementation at `https://anonymous.4open.science/r/TRUST-7297/`. The repository includes environment dependencies, instructions for preparing datasets and pretrained weights, and scripts for feature precomputation, training, and evaluation. Section 3 describes the method, and Section 4.1 reports the experimental settings and hyperparameters.

REFERENCES

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 *International Conference on Learning Representations*, 2024.

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, 2023.

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.

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.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39 (1):1–38, 1977. doi: 10.1111/j.2517-6161.1977.tb01600.x.

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.

Ross Girshick. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1440–1448, 2015.

Haochen Han, Alex Jinpeng Wang, Peijun Ye, and Fangming Liu. Unlearning the noisy correspondence makes CLIP more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4518–4528, 2025.

Peng Hu, Zhenyu Huang, Dezhong Peng, Xu Wang, and Xi Peng. Cross-modal retrieval with partially mismatched pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 9595–9610, 2023.

Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29406–29419, 2021.

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

Young Kyun Jang, Donghyun Kim, Zihang Meng, Dat Huynh, and Ser-Nam Lim. Visual delta generator with large multi-modal models for semi-supervised composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16805–16814, 2024.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision*, pp. 201–216, 2018.

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*, 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19628–19637, 2025a.

Zixu Li, Zhiwei Chen, Haokun Wen, Zhiheng Fu, Yupeng Hu, and Weili Guan. ENCODER: Entity mining and modification relation binding for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5101–5109, 2025b.

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.

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. 5753–5762, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Aaron Lou, Derek Lim, Isay Katsman, Leo Huang, Qingxuan Jiang, Ser Nam Lim, and Christopher M. De Sa. Neural manifold ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17548–17558. Curran Associates, Inc., 2020.

Yang Qin, Dezhong Peng, Xi Peng, Xu Wang, and Peng Hu. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 4948–4956, 2022. doi: 10.1145/3503161.3547922.

Yang Qin, Yuan Sun, Dezhong Peng, Joey Tianyi Zhou, Xi Peng, and Peng Hu. Cross-modal active complementary learning with self-refining correspondence. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. Noisy-correspondence learning for text-to-image person re-identification. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27187–27196. IEEE, 2024.

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):149:1–149:26, 2024. doi: 10.1145/3662732.

Jiangming Shi, Xiangbo Yin, Yeyun Chen, Yachao Zhang, Zhizhong Zhang, Yuan Xie, and Yanyun Qu. Multi-schema proximity network for composed image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19999–20008, 2025.

Yucheng Suo, Fan Ma, Linchao Zhu, and Yi Yang. Knowledge-enhanced dual-stream zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26951–26962, 2024.

Likai Tian, Jian Zhao, Zechao Hu, Zhengwei Yang, Hao Li, Lei Jin, Zheng Wang, and Xuelong Li. CCIN: Compositional conflict identification and neutralization for composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3974–3983, 2025.

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.

Wuyuqing Wang, Zeyuan Gu, and Erkun Yang. Noisy correspondence learning with modality gap direction correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10163–10171, 2026.

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, 2025. doi: 10.1609/aaai.v39i20.35439.

Haokun Wen, Xian Zhang, Xuemeng Song, Yinwei Wei, and Liqiang Nie. Target-guided composed image retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 915–923, 2023. doi: 10.1145/3581783.3611817.

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

Cairong Yan, Meng Ma, Yanting Zhang, and Yongquan Wan. Dual-path multimodal optimal transport for composed image retrieval. In *Proceedings of the Asian Conference on Computer Vision*, pp. 1741–1755, 2024.

Haotian Yang, Cheng Peng, Haobo Wang, Zhen Zhu, Xiu Tang, Gongsheng Yuan, Zhongle Xie, Sai Wu, Weiqiang Wang, and Yu Cheng. Beyond contrastive: Generative reranking for multilingual text-based person retrieval. In *Proceedings of the 34th ACM International Conference on Multimedia*, November 2026. doi: 10.1145/3767308.3835886. To appear.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations*, 2022.

Quanxing Zha, Xin Liu, Yiu-ming Cheung, Xing Xu, Nannan Wang, and Jianjia Cao. UGNCL: Uncertainty-guided noisy correspondence learning for efficient cross-modal matching. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 852–861, 2024. doi: 10.1145/3626772.3657806.

Quanxing Zha, Xin Liu, Shu-Juan Peng, Yiu-ming Cheung, Xing Xu, and Nannan Wang. ReCon: Enhancing true correspondence discrimination through relation consistency for robust noisy correspondence learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29680–29689, 2025.

Supplementary Material

This appendix provides mathematical analyses, implementation details, and additional experimental results for TRUST. It is organized as follows:

- **Appendix A:** Mathematical Analysis
  - **Appendix A.1:** Geometry of the Spherical Logarithmic Map and Transformation Agreement
  - **Appendix A.2:** Analysis of Reference-Anchored Alignment
  - **Appendix A.3:** Gradient Analysis of RCL and Reliability-Weighted Ranking
- **Appendix B:** Datasets and Training Algorithm
  - **Appendix B.1:** Datasets and evaluation protocols
  - **Appendix B.2:** Training algorithm
- **Appendix C:** Additional Quantitative Analysis
  - **Appendix C.1:** Efficiency
  - **Appendix C.2:** Cross-method integration of TARE and RGAO
  - **Appendix C.3:** Noise-type and extreme-corruption analysis
- **Appendix D:** Analysis of Transformation-Aware Reliability Estimation
  - **Appendix D.1:** Evaluation of Reliability Estimation
  - **Appendix D.2:** Four-State Analysis
  - **Appendix D.3:** Analysis of Geometric Reliability Cues
  - **Appendix D.4:** Evolution of Reliability Estimates
- **Appendix E:** More Qualitative Results
  - **Appendix E.1:** Semantic Alignment in Top-1 Retrieval
  - **Appendix E.2:** Fine-Grained Ranking in Top-5 Retrieval
  - **Appendix E.3:** Analysis of Retrieval Failures

A Mathematical Analysis

A.1 Geometry of the Spherical Logarithmic Map and Transformation Agreement

All embeddings used by TRUST are ℓ_2-normalized and therefore lie on $\mathbb{S}^{d-1}$. Consider two non-antipodal points $x, y \in \mathbb{S}^{d-1}$ and let

$$c = x^\top y = \cos\theta, \qquad \theta = \arccos(c) \in [0, \pi). \tag{1}$$

For $0 < \theta < \pi$, define the unit tangent direction from x to y as

$$u_{x\to y} = \frac{y - cx}{\sqrt{1-c^2}} = \frac{y - \cos\theta\, x}{\sin\theta}. \tag{2}$$

This vector belongs to the tangent space $T_x\mathbb{S}^{d-1} = \{v \in \mathbb{R}^d : x^\top v = 0\}$ because $x^\top(y - cx) = 0$. Moreover, $\|y - cx\|_2^2 = 1 - c^2 = \sin^2\theta$, so $\|u_{x\to y}\|_2 = 1$. Ignoring only the numerical clamp in Eq. 1, the logarithmic map can therefore be written as

$$\mathrm{Log}_x(y) = \theta u_{x\to y}, \qquad \big\|\mathrm{Log}_x(y)\big\|_2 = \theta. \tag{3}$$

Thus, the logarithmic map separates a spherical displacement into an *intrinsic magnitude* θ, equal to the geodesic distance from x to y, and an *initial direction* $u_{x\to y}$ in the tangent space at x.

For a CIR triplet, both transformations share the reference r_i as their tangent origin. Write

$$\Delta_i^{\mathrm{mod}} = \theta_i^{\mathrm{mod}} u_i^{\mathrm{mod}}, \qquad \Delta_i^{\mathrm{vis}} = \theta_i^{\mathrm{vis}} u_i^{\mathrm{vis}}, \tag{4}$$

where $\theta_i^{\mathrm{mod}} = \arccos(r_i^\top q_i)$ and $\theta_i^{\mathrm{vis}} = \arccos(r_i^\top t_i)$. If ϕ_i denotes the angle between the two unit tangent directions, then the transformation disagreement in Eq. 2 becomes

$$g_i = 1 - (u_i^{\mathrm{mod}})^\top u_i^{\mathrm{vis}} = 1 - \cos\phi_i \in [0, 2]. \tag{5}$$

Thus, g_i is not an endpoint similarity between q_i and t_i: it compares their reference-conditioned directions after removing their components along r_i. In particular, $g_i = 0$ means that the modification-guided and visual transformations initially follow the same direction from the reference; it does not require their endpoints or displacement magnitudes to coincide. Magnitude remains available to TARE through $d_i^{RT} = \theta_i^{\text{vis}}$, which is used by both the RT posterior and the long-displacement veto in Eq. 3.

A.2 Analysis of Reference-Anchored Alignment

The alignment regularizer measures how closely the composed query and its paired target realize the same transformation from the reference. Suppressing the sample index, let r, q, and t^* denote the normalized reference, composed-query, and MaxSim-selected paired-target token. Define the two reference-anchored tangent vectors and their lengths by

$$v_q = \operatorname{Log}_r(q), \qquad v_t = \operatorname{Log}_r(t^*), \qquad a = \|v_q\|_2, \qquad b = \|v_t\|_2. \tag{6}$$

When $a, b > 0$, let $u_q = v_q/a$, $u_t = v_t/b$, and $\kappa = u_q^\top u_t \in [-1, 1]$. Expanding the squared norm gives

$$\begin{aligned} \left(d^{\text{align}}\right)^2 &= \|v_q - v_t\|_2^2 \\ &= a^2 + b^2 - 2ab\kappa \\ &= (a-b)^2 + 2ab(1-\kappa). \end{aligned} \tag{7}$$

The first term measures disagreement in transformation magnitude; the second measures disagreement in direction within the shared tangent space at r. Thus, $d^{\text{align}} = 0$ exactly when $v_q = v_t$. For nonzero vectors, this requires both $a = b$ and $\kappa = 1$. If either vector is zero, the original form $\|v_q - v_t\|_2$ remains well-defined without introducing a direction for the zero vector.

This construction is not equivalent to subtracting Euclidean displacements. Indeed, if $\delta_q = q - r$ and $\delta_t = t^* - r$, then

$$\|\delta_q - \delta_t\|_2 = \|q - t^*\|_2. \tag{8}$$

The reference cancels, so the result depends only on the two endpoints. In contrast, v_q and v_t are obtained by mapping both endpoints into the reference-specific tangent space $T_r\mathbb{S}^{d-1}$; changing r generally changes both tangent vectors and hence d^{align}.

For non-antipodal endpoints, the inverse relationship between the spherical exponential and logarithmic maps gives

$$q = \cos(a)r + \sin(a)u_q, \qquad t^* = \cos(b)r + \sin(b)u_t. \tag{9}$$

Using $u_q^\top u_t = \kappa$, their squared Euclidean endpoint distance is

$$\|q - t^*\|_2^2 = 4\sin^2\left(\frac{a-b}{2}\right) + 2\sin(a)\sin(b)(1-\kappa). \tag{10}$$

Unlike Eq. 7, this endpoint distance is expressed through bounded chordal factors $\sin(a), \sin(b)$, which compress large angular displacements; d^{align} instead compares the intrinsic lengths and directions of the two reference-anchored transformations directly. The robust penalty $H(\cdot)$ and the reliability weights p_i then determine how this geometric discrepancy contributes to $\mathcal{L}_{\text{align}}$.

A.3 Gradient Analysis of RCL and Reliability-Weighted Ranking

RCL warmup. Before TARE can produce meaningful reliability scores, the representation must be sufficiently structured, so TRUST warms up with RCL. Define the row logits $z_{ij} = S_{ij}/\tau_{\text{model}}$ and abbreviate P_{ij}^{RCL} as P_{ij}. Omitting the constant factor $1/B$, the per-anchor objective in Eq. 10 is

$$\mathcal{R}_i = -\sum_{j \neq i} \log(1 - P_{ij}). \tag{11}$$

The diagonal gradient shows how this objective behaves. Using $\partial P_{ij}/\partial z_{ik} = P_{ij}(\mathbf{1}\{j = k\} - P_{ik})$ and $A_i = \sum_{j \neq i} P_{ij}/(1 - P_{ij})$,

$$\frac{\partial \mathcal{R}_i}{\partial z_{ii}} = -P_{ii}A_i. \tag{12}$$

Since each softmax denominator contains both $e^{z_{ii}}$ and $e^{z_{ij}}$, $P_{ii}+P_{ij} \le 1$, hence $P_{ii} \le 1-P_{ij} \le 1$ for $j \neq i$ and

$$P_{ii}(1-P_{ii}) \le \left|\frac{\partial \mathcal{R}_i}{\partial z_{ii}}\right| \le 1-P_{ii}. \tag{13}$$

Positive cross-entropy has $|\partial(-\log P_{ii})/\partial z_{ii}| = 1-P_{ii}$, so RCL's attraction toward the annotated diagonal is at most that of cross-entropy, and can be strictly weaker. RCL thus suppresses off-diagonal assignments rather than placing a direct $-\log P_{ii}$ term on a possibly corrupted pair.

Reliability-weighted bidirectional ranking. After warmup, the ranking loss uses the paired diagonal as the positive. Let

$$v_i = m_{\text{rank}} - S_{ii} + \tau_{\text{rank}} \log \sum_{j \neq i} \exp(S_{ij}/\tau_{\text{rank}}) \tag{14}$$

be the hinge argument in Eq. 5, and let $\omega_{ij} = \exp(S_{ij}/\tau_{\text{rank}}) / \sum_{k \neq i} \exp(S_{ik}/\tau_{\text{rank}})$ for $j \neq i$. At differentiable points,

$$\frac{\partial \ell_i}{\partial S_{ii}} = -\mathbf{1}\{v_i > 0\}, \qquad \frac{\partial \ell_i}{\partial S_{ij}} = \mathbf{1}\{v_i > 0\}\, \omega_{ij}. \tag{15}$$

The same mechanism applies to $\ell_i(\mathbf{S}^\top)$ in the reverse direction. Let $h_i = \ell_i(\mathbf{S}) + \ell_i(\mathbf{S}^\top)$ and $P_B = \sum_{j=1}^{B} p_j > 0$. The reliability pass is detached before minibatch optimization, so the weights $w_i = p_i/P_B$ are constant in the current gradient and

$$\nabla_\Theta \mathcal{L}_{\text{rank}} = \sum_{i=1}^{B} w_i \nabla_\Theta h_i, \qquad \sum_{i=1}^{B} w_i = 1. \tag{16}$$

Reliability weighting thus redistributes anchor-level gradient mass instead of uniformly scaling the objective.

To isolate noisy anchors, partition the batch into clean and noisy index sets C and N, and let

$$\eta_N = \sum_{i \in N} w_i = \frac{\sum_{i \in N} p_i}{\sum_{j \in C \cup N} p_j} \tag{17}$$

be their normalized reliability mass. If $\|\nabla_\Theta h_i\|_2 \le G$ for all noisy anchors,

$$\left\| \sum_{i \in N} w_i \nabla_\Theta h_i \right\|_2 \le G\eta_N. \tag{18}$$

Assigning small reliability to a noisy triplet thus limits the gradient it generates as the paired anchor. This bound is restricted to the direct anchor pathway; noisy samples can still influence other anchors through the in-batch negative pool.

B Datasets and Training Algorithm

B.1 Datasets and evaluation protocols

FashionIQ. FashionIQ (Wu et al., 2021) is a fashion-domain CIR benchmark containing 77,684 web-crawled product images and 30,134 annotated triplets across three clothing categories: Dresses, Shirts, and TopTees. Its modification texts describe fine-grained attribute changes, such as color, pattern, sleeve, neckline, and fit. Each query pairs a reference garment with two short descriptions; we apply the public spelling-correction dictionary and join the descriptions with "and", following recent robust CIR systems (Li et al., 2026b; Fu et al., 2026; Li et al., 2026a). The training split contains 18,000 triplets. We report Recall@10 and Recall@50 for each category and their six-metric mean.

CIRR. CIRR (Liu et al., 2021) is a large-scale open-domain benchmark constructed from natural images in NLVR2. It contains 21,552 unique images and 36,554 annotated triplets, including 28,225 training triplets. Compared with FashionIQ, CIRR emphasizes complex multi-object scenes and more diverse compositional changes, including spatial relations, counting, and object addition or removal. To reduce false negatives caused by sparse annotations, it also provides a six-image candidate subset for each query. We report full-gallery Recall@K for $K \in \{1, 5, 10, 50\}$ and subset Recall@K for $K \in \{1, 2, 3\}$.

B.2 Training algorithm

Algorithm 1 summarizes the two-stage training procedure of TRUST. The first E_{w} epochs warm up the retrieval model with RCL, whose noise-tolerant objective is analyzed in Appendix A.3. In each post-warmup epoch, TARE encodes the noisy training set, fits three two-component GMMs, applies the long-displacement veto, and fuses the two relation cues into the triplet reliability p_i. These reliabilities are detached and held fixed during the RGAO minibatch updates of that epoch, and are re-estimated at the start of the next epoch. The algorithm returns the trained retrieval model; TARE is discarded at inference.

Algorithm 1 Training procedure of TRUST.

Input: Noisy training triplets $\widetilde{\mathcal{D}} = \{(R_i, M_i, T_i)\}_{i=1}^{N}$
Parameters: Total epochs E, warmup epochs E_{w}, batch size B, reliability fusion coefficient ρ, temperatures $(\tau_{\mathrm{model}}, \tau_{\mathrm{rank}})$, and ranking margin m_{rank}
Output: Trained retrieval model f_Θ

Initialize the BLIP-2 retrieval model f_Θ
for $e = 0, \ldots, E-1$ **do**
  if $e < E_{\mathrm{w}}$ **then**
    for each minibatch $\mathcal{B} \subset \widetilde{\mathcal{D}}$ **do**
      Update Θ using $\mathcal{L}_{\mathrm{RCL}}$ on $\mathcal{B}$
    end for
  else
    Encode $\widetilde{\mathcal{D}}$ and compute g_i, e_i^{MT}, and d_i^{RT} for every sample i
    Fit three two-component GMMs and infer p_i^{trans}, p_i^{MT}, and p_i^{RT}
    Apply the veto in equation 3 and compute p_i
    Store StopGrad(p_i) for the current epoch
    for each minibatch $\mathcal{B} \subset \widetilde{\mathcal{D}}$ **do**
      Compute $\mathcal{L}_{\mathrm{rank}}$, $\mathcal{L}_{\mathrm{align}}$, and $\mathcal{L}_{\mathrm{uniform}}$
      Compute l_{reg} and set $\zeta_b \leftarrow \min(1, |\mathcal{L}_{\mathrm{rank}}|/l_{\mathrm{reg}})$
      Update Θ using equation 11
    end for
  end if
end for
return the trained retrieval model f_Θ

C Additional Quantitative Analysis

C.1 Efficiency

To assess practical efficiency together with retrieval robustness, we profile the reproduced robust methods on a single NVIDIA RTX A6000 under the same measurement protocol. Table 7 reports FLOPs and peak GPU memory as measures of computational complexity and training-time resource demand, together with inference latency per query and training time per optimization step. Training time is reported as the mean and standard deviation. The final two columns reproduce the aggregate results from the main comparison under the standard NTC setting with a noise rate of 20%: FashionIQ-Avg is the mean of the six category-recall values, and CIRR-Avg is $(R@5 + R_{\mathrm{sub}}@1)/2$.

The profiled methods have closely matched computational loads: TRUST, HABIT, INTENT, and ConeSep each require 539.49G FLOPs, while TME is slightly higher at 544.96G FLOPs. Their

Table 7: Computational efficiency and retrieval performance on a single NVIDIA RTX A6000.

Method	FLOPs (G)	Peak memory (MiB)	Inference time (s/query)	Training time (s/step)	FashionIQ-Avg	CIRR-Avg
TME (Li et al., 2025a)	544.96	27,946	0.008555	2.708 ± 0.005	63.97	79.74
HABIT (Li et al., 2026b)	539.49	15,925	0.008489	3.150 ± 0.032	64.38	79.61
INTENT (Chen et al., 2026)	539.49	**15,914**	0.008485	$\mathbf{2.382 \pm 0.003}$	64.36	79.66
ConeSep (Li et al., 2026a)	539.49	23,340	0.008501	5.151 ± 0.020	64.97	80.43
TRUST (Ours)	539.49	19,484	**0.008454**	2.480 ± 0.001	**65.54**	**81.25**

inference latencies are likewise tightly clustered below 0.0086s/query, with TRUST recording the lowest value of 0.008454s/query. This is consistent with our design: TARE and the spherical regularizers in RGAO are used only during training and introduce no additional inference branch.

TRUST uses 19,484MiB of peak memory, 30.3% less than TME and 16.5% less than ConeSep, although INTENT and HABIT require less memory. Its training time is 2.480 ± 0.001s/step, second only to INTENT and respectively 8.4%, 21.3%, and 51.9% lower than TME, HABIT, and ConeSep. At the same time, TRUST achieves the strongest retrieval scores among the profiled methods, reaching 65.54 on FashionIQ and 81.25 on CIRR. These scores exceed the strongest profiled baseline, ConeSep, by 0.57 and 0.82 points, respectively. Overall, TRUST improves retrieval robustness while retaining near-identical inference cost and competitive training efficiency on the shared hardware platform.

C.2 Cross-method integration of TARE and RGAO

To test whether TARE and RGAO help beyond their native pairing, we build two hybrids on the TME recipe: TME+TARE swaps only TME's selector for TARE, and TME+RGAO swaps only TME's objective for RGAO. TME and TRUST rows are copied from the corresponding main-text comparison tables so that all variants follow the same evaluation protocol.

Table 8: Integration of TARE or RGAO into TME on CIRR under NTC. Best results at each noise rate are in **bold**.

Noise rate	Method	R@K				R_{sub}@K			Avg.(R@5, R_{sub}@1)
		$K=1$	$K=5$	$K=10$	$K=50$	$K=1$	$K=2$	$K=3$	
20%	TME	51.35	81.01	88.53	97.81	78.46	91.25	96.39	79.74
	TME+TARE	51.78	81.76	89.78	97.88	**80.39**	**92.43**	96.80	81.07
	TME+RGAO	52.36	82.41	90.10	98.22	78.51	91.76	96.60	80.46
	TRUST	**53.18**	**83.18**	**90.60**	**98.29**	79.33	91.64	**97.04**	**81.25**
50%	TME	48.48	78.94	87.28	96.99	76.48	90.07	95.83	77.71
	TME+TARE	50.87	81.18	89.04	97.52	**80.19**	**92.43**	96.70	**80.69**
	TME+RGAO	**54.07**	82.24	89.74	97.98	78.89	91.93	96.58	80.57
	TRUST	53.59	**82.27**	**90.00**	**98.12**	79.01	91.61	**96.96**	80.64

The cross-method results show that TARE and RGAO exhibit complementary transfer behavior when integrated into TME (Tables 8 and 6). On CIRR, TME+TARE improves the aggregate by roughly 1.3-3.0 points, with clear gains in subset recall, while TME+RGAO adds 0.7-2.9 points and reaches the best R@1 at a noise rate of 50%. On FashionIQ, the gains are primarily driven by TME+RGAO ($+0.8$/$+1.8$), whereas TME+TARE changes Avg. by -0.4/$+0.1$. These results suggest that TARE transfers most clearly to relation-sensitive subset retrieval, while RGAO provides more consistent gains across datasets.

Combining the two modules nevertheless yields the strongest overall cross-dataset profile. On FashionIQ, TRUST outperforms the best hybrid by 0.77 and 0.99 points at noise rates of 20% and 50%. On CIRR, it is best at a noise rate of 20%, and at a noise rate of 50% its aggregate is within 0.05 points of TME+TARE while leading R@5, R@10, R@50, and R_{sub}@3. These results indicate that the two modules play complementary roles in the complete framework: TARE supplies relation-aware reliability, whereas RGAO uses it to shape ranking and representation geometry.

C.3 NOISE-TYPE AND EXTREME-CORRUPTION ANALYSIS

Tables 9 and 10 compare TRUST under reference-only, modification-only, and target-only corruption, as well as NTC and XNTC, at noise rates of 50% and 80%. We evaluate on the CIRR test set and Fashion-IQ validation set.

Table 9: Performance of TRUST under different noise types on CIRR.

Noise rate	Noise type	R@K				$\mathrm{R_{sub}}$@K			Avg.($\mathrm{R@5}$, $\mathrm{R_{sub}@1}$)
		$K=1$	$K=5$	$K=10$	$K=50$	$K=1$	$K=2$	$K=3$	
50%	Reference only	52.24	81.54	89.33	97.49	80.36	92.29	96.77	80.95
	Modification only	49.13	80.02	88.55	97.78	74.43	89.66	95.93	77.23
	Target only	51.88	82.43	90.46	98.10	78.24	91.66	96.68	80.34
	XNTC	51.80	81.54	90.41	98.00	77.37	91.38	97.11	79.46
	NTC	53.59	82.27	90.00	98.12	79.01	91.61	96.96	80.64
80%	Reference only	50.60	79.98	87.86	97.16	80.60	92.82	96.82	80.29
	Modification only	40.34	71.49	81.93	96.07	64.29	83.33	92.46	67.89
	Target only	48.82	80.05	88.24	97.13	77.35	91.16	96.15	78.70
	XNTC	48.48	79.06	88.29	97.29	75.59	90.96	97.04	77.33
	NTC	51.76	80.75	89.16	97.42	79.30	92.22	96.75	80.02

Table 10: Performance of TRUST under different noise types on FashionIQ.

Noise rate	Noise type	Dress		Shirt		Toptee		Average		
		R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	Avg.
50%	Reference only	48.93	71.94	54.96	73.80	58.59	78.63	54.16	74.79	64.48
	Modification only	47.60	68.86	53.29	71.10	53.85	74.40	51.58	71.45	61.52
	Target only	48.64	70.90	54.56	73.80	56.81	77.46	53.34	74.05	63.69
	XNTC	48.73	71.19	54.71	74.09	57.92	77.51	53.79	74.26	64.02
	NTC	48.79	71.79	55.79	73.85	58.69	78.89	54.42	74.84	64.63
80%	Reference only	43.28	66.63	50.93	70.71	53.09	75.22	49.10	70.85	59.98
	Modification only	39.86	62.42	42.15	60.75	44.26	65.37	42.09	62.85	52.47
	Target only	45.61	68.62	52.01	70.61	54.46	74.25	50.70	71.16	60.93
	XNTC	45.46	67.67	52.60	71.34	55.43	75.06	51.16	71.36	61.26
	NTC	45.36	68.52	53.34	71.79	55.38	76.49	51.36	72.27	61.81

Modification-only corruption yields the lowest aggregate score on both datasets at both noise rates. As the noise rate increases from 50% to 80%, the score drops by 9.34 points on CIRR and 9.05 points on Fashion-IQ, compared with 0.66-4.50 points under reference- or target-only corruption. In this setting, the original reference-target pair is preserved, but the text no longer describes the change between the two images. The retained visual correspondence may obscure the mismatch between the modification text and the actual reference-to-target visual transformation. Determining whether the requested modification is actually realized is therefore central to noise discrimination under NTC, even when the reference and target remain visually compatible.

XNTC produces lower aggregate scores than standard NTC in all four comparisons, with gaps of 2.69 points on CIRR and 0.55 points on Fashion-IQ at a noise rate of 80%. Despite shuffling more fields, XNTC consistently yields higher scores than modification-only corruption, with gains of 9.44 and 8.79 points, respectively, at this noise rate. Thus, the difficulty of NTC cannot be explained by the noise rate alone; it also depends on which semantic relations are corrupted. These results support using the intrinsic semantic relations within triplets as the basis for noise discrimination.

D ANALYSIS OF TRANSFORMATION-AWARE RELIABILITY ESTIMATION

D.1 EVALUATION OF RELIABILITY ESTIMATION

Retrieval loss measures how well a triplet satisfies the current retrieval objective, but a low loss does not necessarily imply correct correspondence. As discussed in Section 3.2, unchanged visual content can dominate query-target similarity, while query composition can obscure invalid component-target relations. TARE therefore assesses whether the requested transformation agrees with the observed visual change and whether the individual query components are compatible with the target. Here, we examine whether these semantic relations provide more reliable clean/noisy discrimination than loss-based estimation.

We compare TARE in TRUST with a loss-based GMM (Loss GMM) on CIRR and FashionIQ at noise rates of 20%, 50%, and 80%. All evaluations in this subsection use the checkpoint obtained

Table 11: Clean/noisy discrimination of TARE and Loss GMM (%). Best results at each noise rate are in **bold**.

Noise rate	Estimator	CIRR				FashionIQ			
		AUROC ↑	AUPR ↑	Purity ↑	Recall ↑	AUROC ↑	AUPR ↑	Purity ↑	Recall ↑
20%	Loss GMM	82.26	92.12	91.64	**92.14**	84.38	93.29	94.05	**81.41**
	TARE	**92.38**	**97.85**	**94.68**	74.95	**90.81**	**97.28**	**96.44**	72.12
50%	Loss GMM	82.14	73.90	74.14	**89.41**	84.57	80.46	79.95	**78.98**
	TARE	**92.82**	**93.37**	**81.35**	78.70	**88.83**	**88.89**	**83.17**	72.56
80%	Loss GMM	82.83	43.50	**43.63**	**87.49**	84.03	48.10	45.36	**82.87**
	TARE	**89.53**	**77.84**	41.77	75.16	**85.34**	**66.49**	**50.08**	65.49

Table 12: False-positive rates and probabilistic prediction errors. Best results at each noise rate are in **bold**.

Noise rate	Estimator	CIRR			FashionIQ		
		FPR@95TPR (%) ↓	Brier ↓	NLL ↓	FPR@95TPR (%) ↓	Brier ↓	NLL ↓
20%	Loss GMM	36.04	**0.117**	0.665	100.00	**0.178**	1.803
	TARE	**25.70**	0.168	**0.587**	**41.50**	0.214	**0.802**
50%	Loss GMM	36.25	0.185	0.872	100.00	0.191	1.532
	TARE	**27.68**	**0.119**	**0.405**	**47.04**	**0.158**	**0.547**
80%	Loss GMM	**32.63**	0.233	1.079	**42.56**	0.220	1.478
	TARE	36.23	**0.119**	**0.366**	55.75	**0.121**	**0.390**

at the end of warmup. Synthetic corruption labels are used only for evaluation, with clean triplets as the positive class. We assess both the discrimination and probabilistic prediction quality of the reliability scores, and examine the composition of the predicted clean sets. AUROC and AUPR evaluate discrimination across thresholds, while purity and clean recall are measured at a threshold of 0.5.

Table 11 shows that TARE achieves higher AUROC and AUPR in all six settings. These metrics assess the overall separation of clean and noisy triplets without relying on a single selection threshold. The AUPR gap is especially pronounced at a noise rate of 80%, reaching 34.34 percentage points on CIRR and 18.39 points on FashionIQ. In this setting, clean triplets constitute only one-fifth of the training set, making it more difficult to identify them without including noisy ones. The higher AUPR indicates better overall precision-recall performance. These results support the motivation for examining the semantic relations within a triplet, whose validity may not be fully reflected by its retrieval loss.

Discrimination metrics evaluate the ordering of samples, whereas reliability estimation also concerns the values assigned to them. Two estimators can produce the same ordering while assigning very different probabilities. We therefore use Brier score and negative log-likelihood (NLL) to evaluate agreement between the predicted probabilities and the clean/noisy labels. Table 12 shows lower NLL for TARE in all six settings and lower Brier scores in four. At a noise rate of 80%, NLL decreases from 1.079 to 0.366 on CIRR and from 1.478 to 0.390 on FashionIQ. Since NLL strongly penalizes confident incorrect predictions, these reductions provide evidence of improved probabilistic prediction quality beyond the gains in sample ranking.

We also examine noise rejection when most clean triplets must be retained. FPR@95TPR measures the fraction of noisy triplets classified as clean at a fixed clean recall of 95%. TARE reduces this error in four of the six settings. At a noise rate of 50%, it decreases from 36.25% to 27.68% on CIRR and from 100% to 47.04% on FashionIQ. Thus, in these settings, TARE rejects more noisy triplets while recovering the same proportion of clean ones, complementing the overall discrimination results.

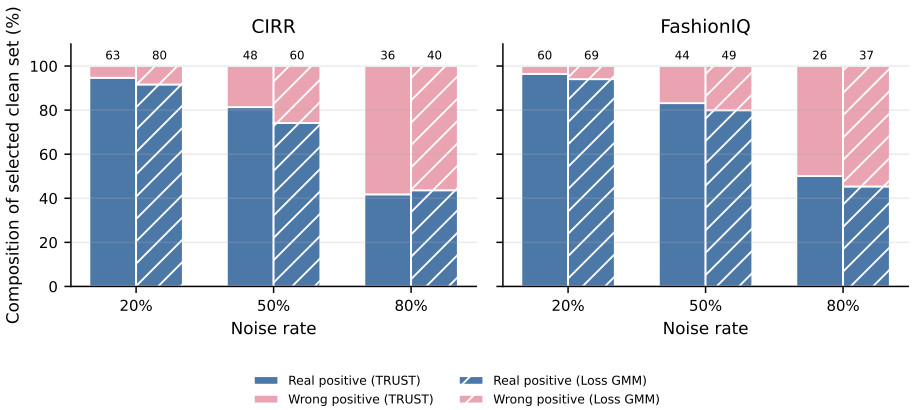

Figure 5: Composition of predicted clean sets. Blue and pink show the clean and noisy fractions; numbers above bars indicate coverage.

Figure 5 illustrates how these differences appear in the predicted clean sets. At noise rates of 20% and 50%, TARE yields higher purity than Loss GMM on both datasets. At a noise rate of 50%, purity increases from 74.14% to 81.35% on CIRR and from 79.95% to 83.17% on FashionIQ. The smaller noisy fraction means that a lower proportion of the triplets identified as reliable provide mismatched targets. This connects the discrimination results to the quality of the supervision identified by the estimator.

The consistent discrimination gains and lower NLL, together with the purity improvements at noise rates of 20% and 50%, support TARE's use of intrinsic semantic relations to assess triplet reliability.

D.2 Four-State Analysis

The four-state formulation assigns a semantic interpretation to each joint validity pattern of the modification-target (MT) and reference-target (RT) relations. We examine whether these interpretations correspond to the relation patterns induced by different corruption sources. In Eq. equation 4, C favors both relations being valid, R only MT being valid, M only RT being valid, and T neither being valid. Figure 6 compares the predicted states with the clean and corruption labels on CIRR and FashionIQ at a noise rate of 50%.

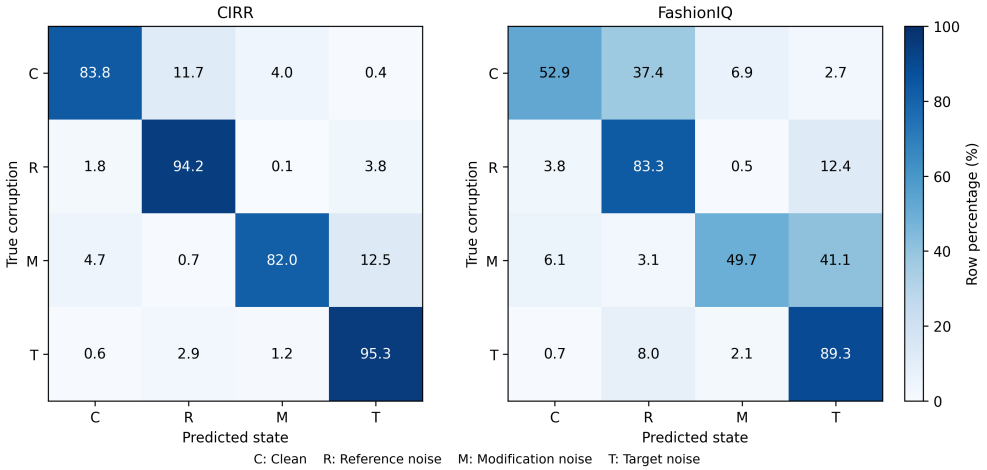

Figure 6: Four-state relation analysis on CIRR and Fashion-IQ at 50% noise. Rows indicate ground-truth corruption labels; columns indicate predicted clean-like, reference-noise-like, modification-noise-like, and target-noise-like states. Entries are row percentages.

The diagonal entry is the largest in every row on both datasets, with per-class recall ranging from 82.0% to 95.3% on CIRR. The distinction between reference and modification corruption is particularly informative: both preserve part of the original correspondence, with reference shuffling retaining the MT pair and modification shuffling retaining the RT pair. The predicted states reflect this distinction, with $R \to M$ and $M \to R$ confusion rates of only 0.1% and 0.7% on CIRR, and 0.5% and 3.1% on FashionIQ. Together with the diagonal concentrations, these results support the intended interpretation of the two states: separately modeling MT and RT provides evidence about which component-target relation is inconsistent, even when the other original pair is preserved.

The main off-diagonal patterns further reveal where relation assessment is ambiguous. On FashionIQ, the largest confusions are $C \to R$ (37.4%) and $M \to T$ (41.1%). Both preserve the validity status of the MT relation while changing the RT relation from valid to invalid, locating the principal attribution ambiguity in the reference-target relation. Thus, the four-state representation both captures distinct corruption patterns and makes its attribution errors interpretable in terms of specific semantic relations. These observations support jointly assessing MT and RT validity as the basis for composed reliability in TARE.

D.3 Analysis of Geometric Reliability Cues

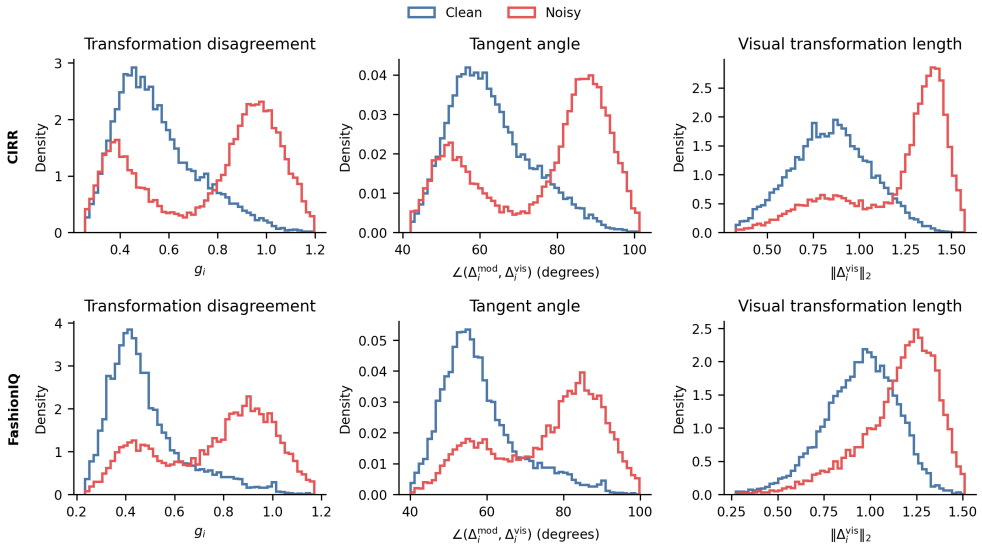

Figure 7: Distributions of transformation-agreement features for clean and noisy triplets under the NTC setting with a noise rate of 50%.

To examine whether reference-anchored geometry provides informative cues for triplet reliability, Figure 7 compares the distributions of clean and noisy triplets on CIRR and FashionIQ under the NTC setting with a noise rate of 50%. We visualize transformation disagreement g_i, its corresponding tangent angle, and the reference-to-target displacement magnitude $d_i^{RT} = \|\Delta_i^{\text{vis}}\|_2$.

On both datasets, clean triplets concentrate at smaller disagreement values, whereas the dominant peak of noisy triplets shifts toward larger values. The tangent-angle plots provide an intuitive interpretation of this separation: the modification-guided and observed visual transformations tend to follow more consistent directions for clean triplets. These observations support using reference-anchored transformation agreement as evidence for reliability estimation and associating the lower-disagreement GMM component with higher transformation reliability. Nevertheless, some noisy triplets also exhibit small disagreement, indicating that directional agreement alone does not establish a valid correspondence.

The displacement distributions reveal another aspect of the relational geometry. Noisy triplets tend to have larger reference-to-target displacement magnitudes, while clean triplets concentrate at shorter distances. This pattern is consistent with using the lower-distance component to estimate RT rela-

tion validity. The overlap between the distributions also shows that a large displacement does not necessarily indicate corruption, since valid modifications can involve substantial visual changes. Together, these results support transformation agreement and RT relation assessment as complementary evidence for triplet reliability estimation, while highlighting the uncertainty in these assessments.

D.4 Evolution of Reliability Estimates

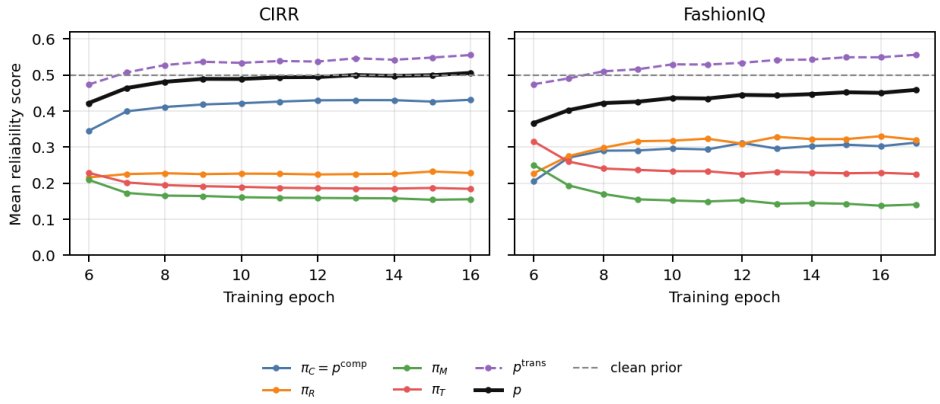

Figure 8: Evolution of relation states and reliability during training under the NTC setting with a noise rate of 50%.

Figure 8 examines the evolution of reliability estimates on CIRR and FashionIQ under the NTC setting with a noise rate of 50%. We evaluate the entire training set and report the mean scores for transformation reliability, composed reliability, and final triplet reliability, together with the mean weights of the four relation states. Since TARE re-estimates reliability from the updated representations at each epoch, these curves reveal how its supervisory signals evolve alongside retrieval learning.

Both datasets exhibit a consistent pattern of initial adjustment followed by gradual stabilization. Transformation reliability and clean-like state weights increase during the early post-warmup epochs, while their subsequent changes become modest. The final reliability follows a similarly smooth trajectory, approaching approximately 0.50 on CIRR and 0.46 on FashionIQ toward the end of the observed training period. This consistent stabilization across two distinct domains indicates that TARE can maintain stable aggregate reliability estimates as the representation space evolves, providing a stable basis for reliability-guided optimization.

The two reliability cues also exhibit distinct estimation profiles. Transformation reliability reaches approximately 0.55 on both datasets, whereas composed reliability remains lower, particularly on FashionIQ. Their fusion balances these different assessments rather than relying exclusively on either relation cue. Meanwhile, the increasing clean-like weight and decreasing modification-noise-like and target-noise-like weights indicate a redistribution of probability mass among the four relation states during early training, followed by relatively stable state proportions. These observations illustrate how TARE preserves differentiated assessments of relational patterns within a stable estimation process, while the mean final reliability moves closer to the true proportion of clean triplets as training progresses.

E More Qualitative Results

We provide complementary qualitative comparisons of the highest-ranked retrieval and the Top-5 candidate list, examining how the methods respond to semantic modifications and rank images under multiple visual constraints. We also analyze retrieval failures to identify limitations in satisfying negation, category changes, and multiple simultaneous requirements.

E.1 Semantic Alignment in Top-1 Retrieval

Figure 9 compares the Top-1 retrieval results of TRUST with TME, HABIT, and ConeSep on four CIRR queries. Each panel shows the composed query on the left and the retrieved images on the right; green check marks and red crosses indicate correct and incorrect retrievals, respectively. The examples involve coordinated attribute changes, category replacement, and spatial modifications, requiring the retrieved image to reflect the requested transformation of the reference. TRUST returns the correct result in all four illustrated cases, while the baselines exhibit partial semantic matches or retrieve unrelated content.

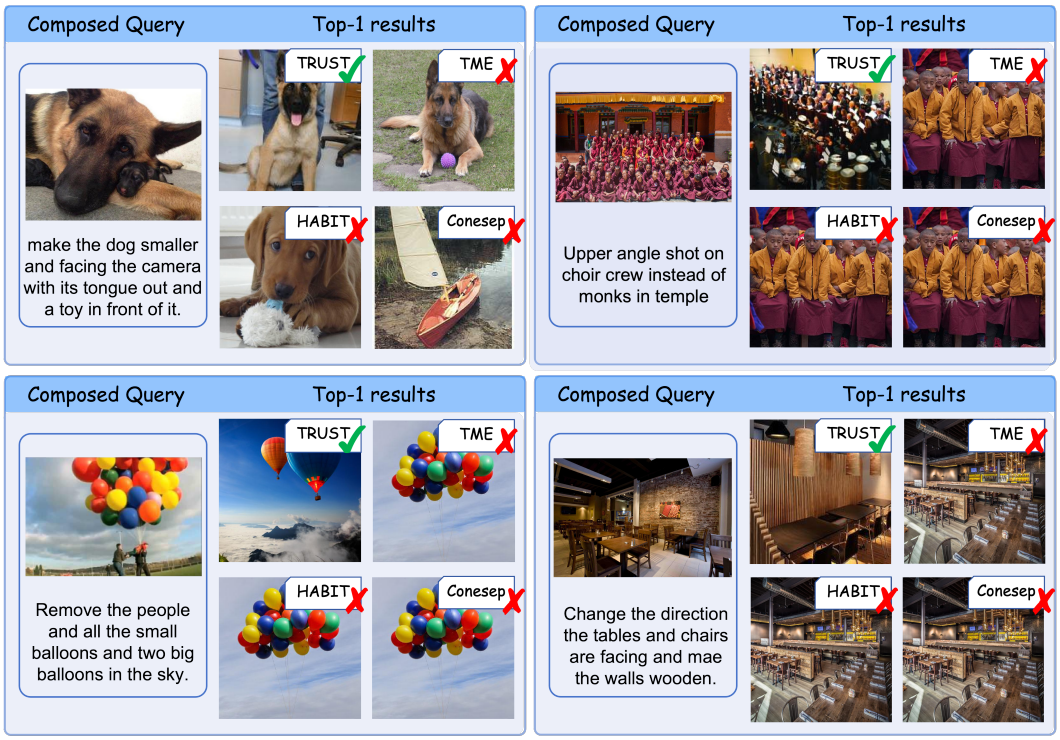

Figure 9: Top-1 retrieval results of TRUST, TME, HABIT, and ConeSep on CIRR.

The choir and balloon examples reveal a shared difficulty in following changes that substantially alter the reference content. For the choir query, all three baselines retain monks as the main subject, missing the requested group replacement. TRUST retrieves a choir viewed from an elevated angle, capturing both the new subject and the specified viewpoint. For the balloon query, the baseline results omit the people but retain a cluster of small balloons. TRUST instead retrieves two large hot-air balloons in the sky. These comparisons illustrate that satisfying one part of a modification can still leave its central semantic change unresolved.

The dog and indoor examples further examine the joint handling of multiple constraints. In the dog example, TME captures the presence of a toy but misses the requested facial expression, HABIT returns a puppy holding a toy, and ConeSep retrieves an unrelated boat. TRUST retrieves a dog facing the camera with its tongue out. In the indoor example, the baselines return a restaurant with a different furniture arrangement but without the requested wooden wall surface, while TRUST retrieves tables and seating alongside a wood-paneled wall. Together, these cases illustrate TRUST's ability to prioritize the semantic changes linking the reference and target. This behavior is consistent with its relation-aware training objective: TARE assesses transformation agreement and jointly evaluates the validity of the reference-target and modification-target relations, while RGAO uses the resulting reliability to guide ranking and spherical regularization.

E.2 Fine-Grained Ranking in Top-5 Retrieval

Figure 10 extends the comparison to the Top-5 results of TRUST, HABIT, and ConeSep on CIRR and FashionIQ. The top two examples are from CIRR, and the bottom example is from FashionIQ. Within each example, the three result rows correspond to TRUST, HABIT, and ConeSep from top to bottom, with candidates ranked from left to right. These examples examine how the methods order semantically related candidates when relevance depends on object number, spatial relations, or a combination of clothing attributes. Comparing shared candidates across the three lists makes the differences in ranking particularly clear.



Figure 10: Top-5 retrieval results of TRUST, HABIT, and ConeSep on CIRR and FashionIQ.

For the first CIRR query, TRUST places the image of two puppies lying together at rank 1; HABIT and ConeSep place the same image at ranks 4 and 2, respectively. Both baselines instead rank an

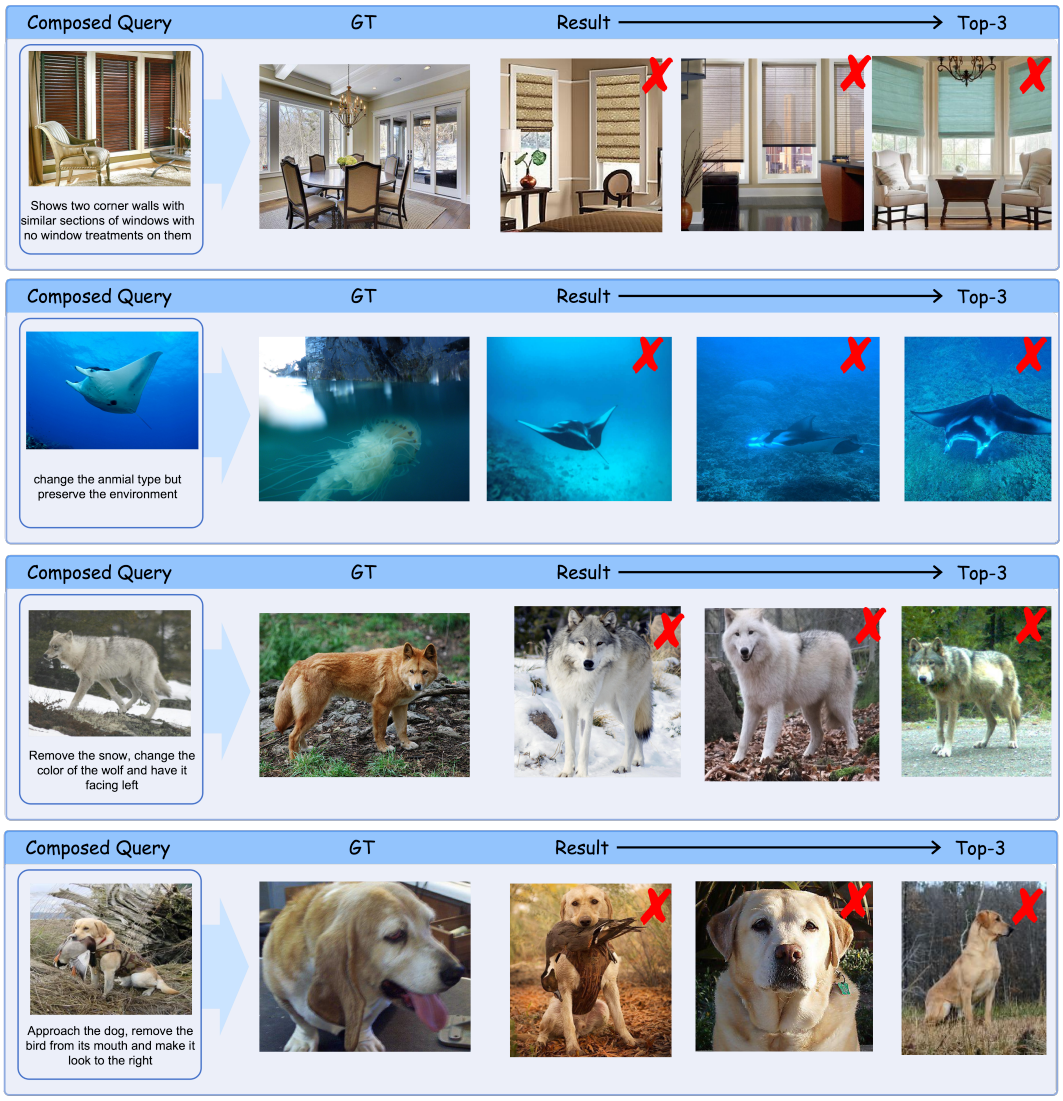

Figure 11: Failure cases of TRUST on CIRR, showing the composed query, ground-truth target, and Top-3 retrieval results.

adult dog with a puppy first, retaining an adult animal despite the requested change to two puppies. This comparison highlights the importance of resolving age and number jointly within the same object category. In the second CIRR example, TRUST ranks first a room in which a rug lies beneath a coffee table in front of a couch. Both baselines rank this image second and prefer a different room with a table over an animal-hide rug. TRUST's first result presents the requested table-rug-couch arrangement more clearly, illustrating the value of considering relations among objects when ordering otherwise relevant scenes.

The FashionIQ example probes finer distinctions among clothing attributes. The query asks for a shorter dress in lighter brown or tan with long sleeves. TRUST's first result captures these broad changes, while HABIT's first result is light-colored but sleeveless and long. The rankings also expose a remaining challenge: the same taupe dress with a dark hem appears at rank 1 for ConeSep and rank 5 for both TRUST and HABIT. Moreover, TRUST's list includes candidates that satisfy only a subset of the requested attributes, such as a sleeveless brown dress. Thus, the example illustrates both sensitivity to the main modification and residual difficulty in consistently satisfying all fine-grained constraints across the candidate list.

Taken together, the Top-5 comparisons show how TRUST can promote images that capture the requested object and spatial relations, while also revealing scope for improving the ordering of closely related fashion items. They complement the Top-1 results by examining the semantic distinctions within the retrieved candidate set.

E.3 Analysis of Retrieval Failures

Figure 11 presents four failure cases of TRUST on CIRR. Each row shows the composed query on the left, the annotated ground-truth target (GT) in the middle, and the Top-3 retrieved images on the right, ordered from rank 1 to rank 3 from left to right. Red crosses indicate retrieved images that do not match the annotated target, which is absent from all four displayed candidate lists. Although the retrieved candidates remain visually related to the query, they reveal unresolved difficulties in excluding unwanted attributes, changing the object category, and jointly satisfying several modification requirements.

The first two cases illustrate how reference appearance can remain dominant when the instruction requires removing or replacing a salient feature. The window query explicitly requests windows with no treatments, yet all three retrieved images retain blinds or shades. The candidates match the broad indoor scene but overlook the exclusion requirement. In the underwater example, the instruction asks to change the animal type while preserving the environment. TRUST retrieves three images of rays, retaining both the underwater setting and the reference animal category. The annotated target instead depicts a jellyfish. This example shows that preserving the requested context does not ensure that the intended category change is captured.

The wolf and dog examples expose a related difficulty in satisfying multiple constraints simultaneously. For the wolf query, the first result retains snow, while the other two show largely snow-free ground but retain gray or white fur. The requested changes are therefore only partially reflected in the candidate list. For the dog query, the first result still shows a bird in the dog's mouth. The second provides a closer view without a bird but shows a largely frontal face, while the third shows a dog looking to the right without providing the requested close-up. Individual candidates capture different parts of the instruction, but none combines them as in the annotated target.

These failures indicate that relation-aware reliability learning does not eliminate all errors in semantic composition. TARE and RGAO improve the quality of training supervision and the learned retrieval geometry, but do not explicitly enforce each textual constraint at retrieval time. The cases motivate further investigation of explicit modeling of negation and joint verification of multiple modification requirements, particularly when unchanged reference content provides a strong competing similarity cue.