INTENT: Invariance and Discrimination-aware Noise Mitigation for Robust Composed Image Retrieval

Zhiwei Chen[1], Yupeng Hu[1]*, Zhiheng Fu[1], Zixu Li[1], Jiale Huang[1], Qinlei Huang[1], Yinwei Wei[1],

[1]School of Software, Shandong University
{zivczw, fuzhiheng8, lizixu.cs}@gmail.com, {huangjiale359, hql}@mail.sdu.edu.cn,
huyupeng@sdu.edu.cn, weiyinwei@hotmail.com

Abstract

Composed Image Retrieval (CIR) is a challenging image retrieval paradigm that enables to retrieve target images based on multimodal queries consisting of reference images and modification texts. Although substantial progress has been made in recent years, existing methods assume that all samples are correctly matched. However, in real-world scenarios, due to high triplet annotation costs, CIR datasets inevitably contain annotation errors, resulting in incorrectly matched triplets. To address this issue, the problem of Noisy Triplet Correspondence (NTC) has attracted growing attention. We argue that noise in CIR can be categorized into two types: **cross-modal correspondence noise** and **modality-inherent noise**. The former arises from mismatches across modalities, whereas the latter originates from intra-modal background interference or visual factors irrelevant to the coarse-grained modification annotations. However, modality-inherent noise is often overlooked, and research on cross-modal correspondence noise remains nascent. To tackle above issues, we propose the **I**nvariance and discrimi**N**a**T**ion-awar**E** **N**oise ne**T**work (**INTENT**), comprising two components: *Visual Invariant Composition* and *Bi-Objective Discriminative Learning*, specifically designed to handle the two-aspect noise. The former applies causal intervention on the visual side via Fast Fourier Transform (FFT) to generate intervened composed features, enforcing visual invariance and enabling the model to ignore modality-inherent noise during composition. The latter adopts collaborative optimization with both positive and negative samples, and constructs a scalable decision boundary that dynamically adjusts decisions based on the loyalty degree, enabling robust correspondence discrimination. Extensive experiments on two widely used benchmark datasets demonstrate the superiority and robustness of INTENT.

1 Introduction

In recent years, the rapid growth of multimedia data (Huang et al. 2025a; Liu et al. 2018b; Tian et al. 2025a; Liu et al. 2025a; Zhang et al. 2025; Liu et al. 2025e) has brought increasing attention to Composed Image Retrieval (CIR) (Xu et al. 2024; Wen et al. 2023a; Yang et al. 2024; Li et al. 2025c; Chen et al. 2025b,a). Unlike traditional single-modal image retrieval, CIR enables more flexible retrieval through multimodal queries, specifically by combining a reference image and a modification text, to accurately retrieve a target image that satisfies specific requirements. This paradigm has shown significant value in various domains, such as information processing (Xu et al. 2025; Tian et al. 2025b; Liu et al. 2021a; Wang, Zhang, and Dodgson 2024; Zhao and Chen 2023; Zhao 2024; Huang et al. 2024c; Ma et al. 2025; Liu et al. 2025b; Wang, Zhang, and Dodgson 2025; Yi-fan 2016; Yifan 2018; Wu et al. 2025) and intelligent systems (Li et al. 2025b, 2023c; Liu et al. 2025d; Huang et al. 2025b; Liu et al. 2025c; Cao et al. 2025; Jiang et al. 2025; Ou, de Bruijn, and Schulz 2025), effectively addressing users' diverse retrieval needs (Liu et al. 2018a; Ting and Listening 2024; He et al. 2025; He, Wang, and Shi 2024).

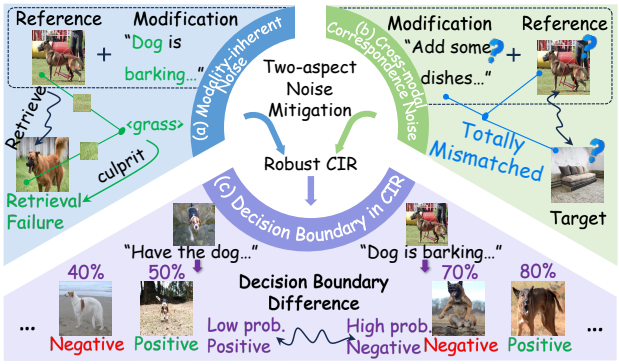

Figure 1: (a) shows typical Modality-inherent Noise in CIR. (b) reveals Cross-modal Correspondence Noise. (c) presents two cases: left successfully retrieves target with low confidence (50%), while right fails with high confidence (70%), illustrating varying decision boundaries in retrievals.

Although existing CIR methods have made significant progress, most of them tend to overlook the issue of Noisy Triplet Correspondence (NTC) in query data, i.e., possible incorrectly matched triplets. Specifically, as illustrated in Figure 1(a) and (b), NTC in CIR is primarily caused by the following two factors: **modality-inherent noise** and **cross-modal correspondence noise**. The former is introduced by intra-modal interference factors, such as complex backgrounds in the image or visual elements irrelevant to

*Corresponding author.

the modification requirements, while the latter arises from incorrect semantic alignment across modalities. As research on NTC in CIR remains at an early stage, achieving robust CIR learning presents two major challenges.

C1: Neglected inherent noise. Existing studies on noise in CIR primarily focus on addressing cross-modal correspondence noise (Li et al. 2025a), with limited analysis or solutions for modality-inherent noise. Such methods typically perform direct fusion of multimodal query features and determine noisy correspondences based on the similarity between composed and the target features. However, they often overlook the fact that modality-inherent noise, such as interference from irrelevant content in the reference image, may distort the composition process and lead to inaccurate feature fusion, thereby introducing uncertainty in noisy correspondence identification. While several recent CIR models attempt to handle modality-inherent noise (Huang et al. 2024a; Li et al. 2025c), they generally assume that cross-modal correspondence is correct and rely on cross-modal interaction to suppress the intra-modal noisy signals. In the context of NTC, where incorrect cross-modal matching may occur, these methods are difficult to apply directly. Therefore, achieving effective noise mitigation under uncertain matching conditions poses the first major challenge in robust CIR learning.

C2: Hard decision boundary. Recent work have focused on noisy dual correspondence (NDC) problem and proposed various strategies such as adversarial training (Dang et al. 2025; Han et al. 2023), sample reweighting (Huang et al. 2021), and self-supervised learning (Tan et al. 2021; Batson and Royer 2019). However, these methods are mainly designed for noise discrimination in paired data and are not readily applicable to the Noisy Triplet Correspondence (NTC) problem. Moreover, current NTC approaches (Li et al. 2025a) typically employ a hard decision boundary. Due to the brevity of modification text in CIR, capturing the semantic gap between reference and target images is challenging. Consequently, similarity scores don't absolutely correspond to noise presence, especially with varying visual complexity across samples. As shown in Figure 1(c), a 50% similarity may suffice for positive labels in some cases (i.e., the left one), while even a 70% score may be negative if the content is simple or there are better-matching candidates in the dataset (i.e., the right one). Thus, accurate noise discrimination in CIR requires contrastive comparison between positive and negative samples, and the design of a scalable decision boundary for robust composed feature alignment, constitutes the second major challenge.

To address these challenges, we propose the Invariance and discrimiNaTion-awarE Noise mitigation neTwork (INTENT), for handling both modality-inherent noise and ambiguous decision boundaries in Composed Image Retrieval (CIR). INTENT comprises two main modules: *(1) The Visual Invariant Composition (VIC) module* applies causal intervention on reference image via Fast Fourier Transform (FFT) to generate counterfactual sample, utilizing it to further enforce visual invariance and enabling the model to ignore modality-inherent noise during composition. *(2) The Bi-Objective Discriminative Learning (BiODL) module* performs collaborative optimization using positive and negative samples, constructing a scalable decision boundary that dynamically adjusts decisions based on the loyalty degree of sample matching, enabling more robust composed feature alignment and improving model's discrimination.

The main contributions of this paper are as follows:

- We are the first to clearly distinguish and investigate two-aspect noise which may lead to Noisy Triplet Correspondence (NTC), focusing on the challenges of neglected inherent noise and hard decision boundaries, and revealing their critical impact on model accuracy and robustness.
- We propose INTENT, a novel robust CIR framework. The VIC module mitigates modality-inherent noise via causal intervention, while the BiODL module performs collaborative optimization over positive and negative samples and constructs a scalable decision boundary for more robust composed feature alignment.
- Extensive experiments on multiple benchmarks show that INTENT significantly outperforms most methods in both accuracy and robustness, confirming the effectiveness of our approach.

2 Related Work

Our work is closely related to Composed Image Retrieval (CIR) with Noisy Correspondence and Causal Intervention.

Composed Image Retrieval with Noisy Correspondence. Composed Image Retrieval (CIR) retrieve target images using reference images combined with modification texts. Current approaches follow two paradigms. Early works (Vo et al. 2019; Wen et al. 2021) used conventional architectures like ResNet and LSTM for separate feature extraction before fusion. Recent advances leverage pre-trained vision-language models such as CLIP (Radford et al. 2021) for joint learning, achieving superior results through streamlined alignment and composition (Jiang et al. 2024; Wen et al. 2023b; Li et al. 2025d; Fu et al. 2025; Huang et al. 2025c). While most CIR research assumes well-aligned triplet annotations, large-scale datasets often contain noisy triplet correspondence (NTC) due to annotation errors or semantic ambiguity. Addressing such robust learning (Lu, Liu, and Kong 2023; Wei et al. 2019; Zhang et al. 2023; Huang et al. 2023; Pu et al. 2025b; Huang et al. 2024b) of noisy correspondences is challenging, as it requires distinguishing reliable supervision from unreliable ones during training. Recent work has begun to tackle this problem through sample selection and realignment strategies, making CIR models more robust to annotation noise (Li et al. 2025a). While some methods (Chen et al. 2024) address false positives, they mainly aim to improve CIR performance rather than addressing robustness under NTC, thus not suited for tackling challenges posed by noisy triplet correspondence.

Causal Intervention in Multimodal Learning. Causal intervention techniques have attracted increasing attention in multimodal learning (Hu et al. 2021b; Kong et al. 2025; Pu et al. 2025a; Sun et al. 2024; Liu et al. 2024a; Wei et al. 2020), especially for information retrieval (Hu et al. 2023b; Liu et al. 2024b; Sun et al. 2023; Hu et al. 2021a; Tang et al. 2024; Lu et al. 2024), visual question answering (Niu

et al. 2021; Wang et al. 2025), and so on. Typical approaches include backdoor adjustment and counterfactual data augmentation, all aimed at mitigating spurious correlations by simulating interventions on input variables. These methods help models focus on causality and improve robustness against confounders or irrelevant variations. Intervention-based methods have been used in image-text scenarios (Li et al. 2023b), perform feature-level interventions, thereby learning representations that are more aligned with true semantic relationships. Despite these advances, the application of causal intervention in complex scenarios (e.g., composed image retrieval) remains underexplored. In this work, we introduce intervention-inspired strategies to CIR, encouraging robust and causally meaningful feature composition.

3 The Proposed INTENT

In this section, we introduce INTENT, as Figure 2 shows.

3.1 CIR from a Causal Perspective

Task definition. Given a set of N multimodal queries $\mathcal{Q} = \{(x_r, y_m)_n\}_{n=1}^{N}$, where x_r, y_m refer to the reference image and modification text, respectively. CIR aims to retrieve the most relevant target image t for each query $\mathcal{Q}$.

Causal analysis. In CIR tasks, since modification texts typically specify changes to partial content in reference images, they naturally share semantic correlations. Therefore, to eliminate modality-inherent noise, we consider leveraging these semantic correlations during multimodal composition to identify noise content. As shown in Figure 3(a), the CIR composition process can be modeled as $I \longrightarrow F \longleftarrow T$. Additionally, the specific content of reference image I influences the interpretation of modification instructions in text T, creating $I \dashrightarrow T$. So if we could model the correlation between I and T, we could explicitly guide the composition. However, this correlation is unobservable due to lack of supervision signals. Thus we explore causal relations during composition to achieve indirect understanding of these correlations. As Figure 3(b) shows, reference image I comprises causal factors C and spurious factors S. Causal factors C represent true semantic attributes the text intends to modify, reflecting genuine causal relations. Spurious factors S represent irrelevant, non-causal visual information (e.g., unmodified objects) that corrupts the composition process, i.e., modality-inherent noise. To eliminate modality-inherent noise, we suppress spurious factors. Since modification texts are user-written and concise, we consider modality-inherent noise primarily exists in reference images.

Two potential paths exist for suppressing spurious factors: 1) precisely decouple image features and remove spurious factors, 2) ignore possible spurious factors during composition. While the first approach is more direct, it's difficult to implement. Existing works in this direction focus on implicit operations (Fu et al. 2025), but uncontrollable implicit processes may cause inaccurate decoupling, harming training. Therefore, we choose the second path, ensuring the final composed features are less affected by modality-inherent noise, implemented by Visual Invariant Composition (VIC).

3.2 Visual Invariant Composition

As shown in Figure 2(a), the Visual Invariant Composition (VIC) performs an intervention described in Section 3.1, forcing the model to focus only on semantics invariant to spurious factors, thus ignoring modality-inherent noise in reference images. This is achieved by enforcing consistency between two composed features derived from the reference image and its counterfactual, intervened version.

Counterfactual Image Generation. To achieve this, we first perform a frequency-domain intervention on the reference image x_r, perturbing its original frequency components while preserving semantic structures. This process yields a counterfactual image $\hat{x}_r$ in which modality-inherent noise is altered, but the core semantic content remains unchanged.

Specifically, as shown in the right part of Figure 2, for each reference image, we randomly sample an irrelevant image x_d, and apply Fast Fourier Transform (FFT), which which has been widely used in the vision domain for manipulating image signals, to both x_r and x_d, obtaining their spectra $\mathbf{F}_r = \mathcal{F}(x_r), \mathbf{F}_d = \mathcal{F}(x_d)$, where $\mathcal{F}(\cdot)$ denotes the FFT function. The spectrum combines an amplitude and a phase spectrum, where amplitude captures low-level statistics (style, texture) while phase contains high-level semantic structures. We intervene solely on amplitude spectra by randomly mixing the cropped central regions with random ratio λ, represented as follows,

$$\hat{\mathbf{A}}_r = \lambda \mathbf{A}_d^{\text{crop}} + (1-\lambda)\mathbf{A}_r^{\text{crop}}, \tag{1}$$

where $\mathbf{A}_r, \mathbf{A}_d$ are amplitude spectra of $\mathbf{F}_r, \mathbf{F}_d$, $\mathbf{A}^{\text{crop}}$ denotes the amplitude spectra of cropped central region. Finally, we reconstruct the counterfactual image via inverse FFT: $\hat{x}_r = \mathcal{F}^{-1}(\hat{\mathbf{A}}_r, \theta_r)$, where θ_r is the phase spectrum of $\mathbf{F}_r$.

In this way, $\hat{x}_r$ preserves key semantics of the original x_r while exhibiting altered noise patterns, providing ideal data pairs for subsequent consistency learning.

Causality-free Composition. After obtaining the factual-counterfactual image pairs, we leverage BLIP-2's Q-Former architecture for multimodal feature composition, combining the reference image x_r and counterfactual image $\hat{x}_r$, with the modification text respectively, formulated as,

$$\begin{cases} \mathbf{F}_c = \text{Q-Former}(\Phi_{\mathbb{X}}(x_r), \Phi_{\mathbb{Y}}(y_m)), \\ \hat{\mathbf{F}}_c = \text{Q-Former}(\Phi_{\mathbb{X}}(\hat{x}_r), \Phi_{\mathbb{Y}}(y_m)), \end{cases} \tag{2}$$

where $\mathbf{F}_c, \hat{\mathbf{F}}_c \in \mathbb{R}^{Q \times D}$ represent the composed feature and intervened composed feature. Q is number of Q-former's learnable queries, and D is the feature dimension. $\Phi_{\mathbb{X}}$ and $\Phi_{\mathbb{Y}}$ denote BLIP-2's frozen visual encoder and tokenizer respectively. Similarly, for target image t, we obtain its feature $\mathbf{F}_t = \text{Q-Former}(\Phi_{\mathbb{X}}(t)) \in \mathbb{R}^{Q \times D}$. Notably, since the intervened composed feature have not been used for training, the model remains causality-free, unable to identify correct causal relations and still affected by spurious factors.

Visual Invariance Learning. In the previous stage, we utilized reference and counterfactual images for composition respectively, obtaining composed features $\mathbf{F}_c$ and intervened composed feature $\hat{\mathbf{F}}_c$ that inherit different modality-inherent noise. Building upon this, we only need to train the model

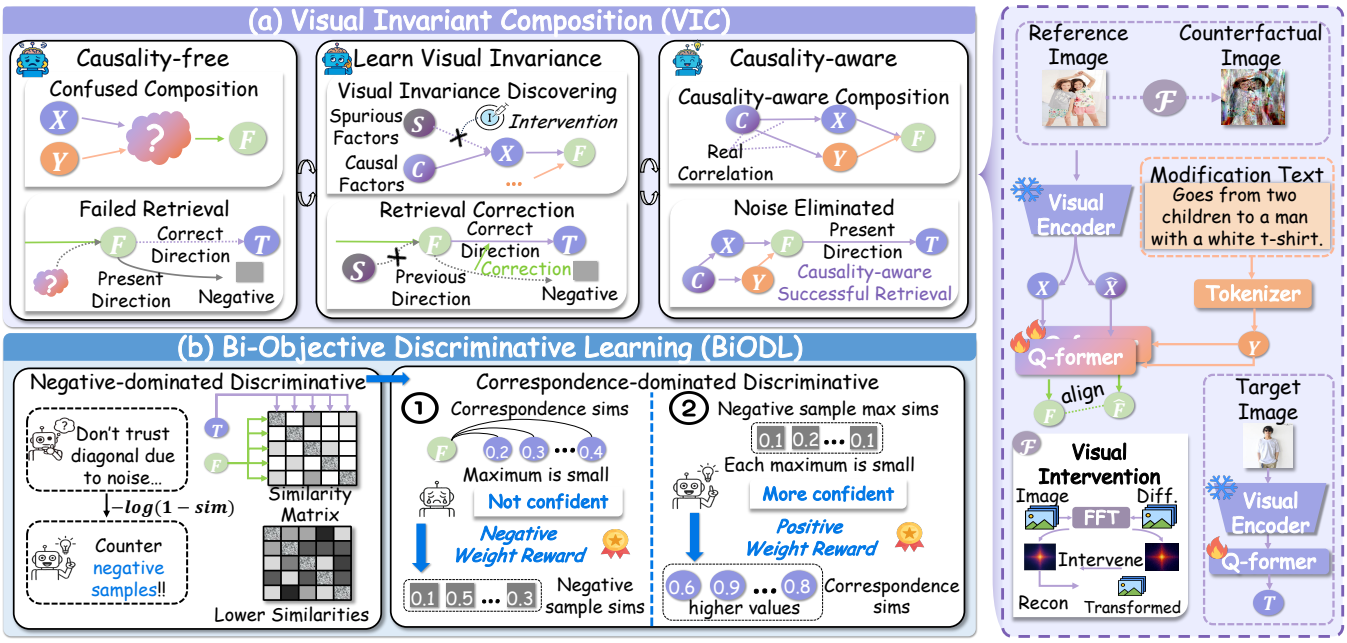

Figure 2: The framework of our proposed INTENT. We designed (a) Visual Invariant Composition and (b) Bi-Objective Discriminative Learning, to mitigate **modality-inherent noise** and **cross-modal correspondence noise**, respectively.

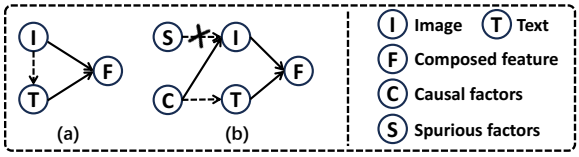

Figure 3: The causal graph of CIR. Solid arrows present the **cause effect**. Dash arrows mean there exist **correlations**.

to identify visual invariants at the feature level, thereby homogenizing noise across different reference images, reducing the model's sensitivity to noise, and indirectly achieving modality-inherent noise mitigation.

To this end, we design visual invariance learning, which forces the model to maintain understanding of key visual information under different modality-inherent noise, obtaining unbiased composed features. Specifically, we define a Causal Consistency Loss based on Centered Kernel Alignment (CKA) (Kornblith et al. 2019), to measure the semantic consistency between $\mathbf{F}_c$ and $\hat{\mathbf{F}}_c$. We first compute the Gram matrices $\mathbf{K}_c = \mathbf{F}_c\mathbf{F}_c^\top \in \mathbb{R}^{Q\times Q}$ and $\mathbf{L}_c = \hat{\mathbf{F}}_c\hat{\mathbf{F}}_c^\top \in \mathbb{R}^{Q\times Q}$. Then we perform centering to eliminate mean shift:

$$\bar{\mathbf{K}}_c = \mathbf{H}\mathbf{K}_c\mathbf{H}, \quad \bar{\mathbf{L}}_c = \mathbf{H}\mathbf{L}_c\mathbf{H}, \tag{3}$$

where $\bar{\mathbf{K}}_c, \bar{\mathbf{L}}_c \in \mathbb{R}^{Q\times Q}$, $\mathbf{H} = \mathbf{I} - \frac{1}{Q}\mathbf{e}\mathbf{e}^\top$, $\mathbf{I}$ is the identity matrix, $\mathbf{e}$ is the unit vector, and Q is the number of Q-former's learnable queries. Finally, we compute the CKA similarity as a causal consistency constraint:

$$\mathcal{L}_{\text{caco}} = \frac{1}{B}\sum_{i=1}^{B}\left(1 - \frac{\langle \bar{\mathbf{K}}_{ci}, \bar{\mathbf{L}}_{ci}\rangle_F}{\|\bar{\mathbf{K}}_{ci}\|_F \|\bar{\mathbf{L}}_{ci}\|_F}\right), \tag{4}$$

where B represents the batch size, $\langle\cdot,\cdot\rangle_F$ denotes Frobenius inner product, $\|\cdot\|_F$ is the Frobenius norm. $\bar{\mathbf{K}}_{ci}$ and $\bar{\mathbf{L}}_{ci}$ represent the i-th Gram matrix of composed features in a batch. By this, the model becomes causality-aware, capable of mining visual invariants while ignoring modality-inherent noise, providing a causal correction for composed features.

3.3 Bi-Objective Discriminative Learning

While mitigating modality-inherent noise, to enhance the model's generalization capability in NTC problem, we propose Bi-Objective Discriminative Learning (BiODL), which jointly optimizes decision boundaries from both negative-dominated and correspondence-dominated perspectives.

Negative-dominated Discriminative Learning. Traditional CIR methods often use InfoNCE (Vo et al. 2019) to pull positive samples closer and push negative samples apart. However, in NTC scenarios, positive sample pairs may contain noisy matches, leading to unreliability. Therefore, inspired by RCL (Hu et al. 2023a), we design a Robust Contrastive Loss that focuses on reducing the negative impact of negative samples, formulated as:

$$\mathcal{L}_{\text{robust}} = -\frac{1}{B}\sum_{i,j\neq i}^{B}\log\left(1 - \frac{\exp\{\mathbf{F}_{ci}\mathbf{F}_{ti}^\top/\tau\}}{\sum_{j=1}^{B}\exp\{\mathbf{F}_{ci}\mathbf{F}_{tj}^\top/\tau\}}\right), \tag{5}$$

where B is batch size, τ is the temperature factor. $\mathbf{F}_{ci}$ and $\mathbf{F}_{tj}$ are i-th query feature and j-th target feature respectively.

The process actively pushes away negative samples, achieving spatial dispersion between positive and negative samples. Notably, although there may be correct correspondences among negative samples, the harm of treating them as negatives for learning is far less than treating noisy correspondences as positives, thus being more robust.

Noise	Methods	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%	SSN (Yang et al. 2024) (AAAI'24)	43.91	77.25	86.48	97.45	71.76	88.63	95.54	74.51
	CALA (Jiang et al. 2024) (SIGIR'24)	49.11	81.21	89.59	98.00	76.27	91.04	96.46	78.74
	SPRC (Xu et al. 2024) (ICLR'24)	51.96	82.12	89.74	97.69	80.65	92.31	96.60	81.39
	RCL (Hu et al. 2023a) (TPAMI'23)	53.16	82.41	90.12	**98.34**	79.57	92.02	96.87	80.99
	RDE (Qin et al. 2024) (CVPR'24)	51.81	82.02	90.60	97.93	78.17	91.90	96.70	80.10
	TME (Li et al. 2025a) (CVPR'25)	**53.42**	82.99	90.24	98.15	**81.04**	**92.58**	**96.94**	**82.01**
	INTENT (Ours)	53.37	**83.16**	**90.73**	98.22	80.24	92.37	96.89	81.70
20%	SSN (Yang et al. 2024) (AAAI'24)	34.02	65.90	75.78	91.33	66.92	85.90	93.45	66.41
	CALA (Jiang et al. 2024) (SIGIR'24)	41.33	72.70	82.84	94.34	71.66	88.15	94.94	72.18
	SPRC (Xu et al. 2024) (ICLR'24)	45.90	75.86	83.52	93.37	78.10	91.40	96.05	76.98
	RCL (Hu et al. 2023a) (TPAMI'23)	50.43	81.11	88.82	96.68	77.52	90.80	95.71	79.31
	RDE (Qin et al. 2024) (CVPR'24)	49.23	78.63	86.80	95.78	76.58	90.31	96.07	77.60
	TME (Li et al. 2025a) (CVPR'25)	**51.35**	81.01	88.53	97.81	**78.46**	91.25	96.39	**79.74**
	INTENT (Ours)	51.25	**81.36**	**90.02**	**98.05**	77.95	**91.40**	**96.46**	79.66
50%	SSN (Yang et al. 2024) (AAAI'24)	25.93	53.71	63.40	82.10	62.10	82.27	91.57	57.90
	CALA (Jiang et al. 2024) (SIGIR'24)	36.10	66.12	77.76	92.10	68.12	85.66	93.59	67.12
	SPRC (Xu et al. 2024) (ICLR'24)	39.93	66.00	73.59	86.48	75.81	89.21	95.37	70.90
	RCL (Hu et al. 2023a) (TPAMI'23)	48.58	77.45	85.93	94.70	75.60	89.28	94.80	76.52
	RDE (Qin et al. 2024) (CVPR'24)	45.98	75.30	83.73	94.48	73.98	88.99	95.13	74.64
	TME (Li et al. 2025a) (CVPR'25)	48.48	78.94	87.28	96.99	76.48	90.07	95.83	77.71
	INTENT (Ours)	**49.78**	**79.64**	**88.99**	**97.37**	**77.18**	**90.41**	**96.00**	**78.41**
80%	SSN (Yang et al. 2024) (AAAI'24)	20.48	43.98	54.27	74.80	56.48	77.20	89.54	50.23
	CALA (Jiang et al. 2024) (SIGIR'24)	31.52	61.49	72.60	89.86	64.34	83.52	92.60	62.92
	SPRC (Xu et al. 2024) (ICLR'24)	29.95	51.25	58.51	73.86	70.22	86.05	93.21	60.74
	RCL (Hu et al. 2023a) (TPAMI'23)	44.94	74.43	82.99	92.31	71.93	86.84	92.96	73.18
	RDE (Qin et al. 2024) (CVPR'24)	42.92	71.30	80.51	92.96	69.64	85.86	93.54	70.47
	TME (Li et al. 2025a) (CVPR'25)	46.31	75.78	84.89	95.83	73.37	88.02	94.89	74.58
	INTENT (Ours)	**47.90**	**78.13**	**87.04**	**96.47**	**73.81**	**89.18**	**95.54**	**75.97**

Table 1: Performance comparison on the CIRR test set in terms of R@K(%) and R_{sub}@K(%). The best and second-best results are highlighted in **bold** and underline, respectively.

Correspondence-dominated Discriminative Learning. While noisy correspondence interference can be reduced at negative-dominated level, the model requires active enhancement of discrimination among reliable correspondence. Thus we propose a scalable decision boundary based on loyalty degree. First, we define the similarity matrix between queries and targets $\mathbf{S} = \left\{\mathbf{s}_{ij} | \mathbf{s}_{ij} = \mathrm{softmax}(\mathrm{s}\left(\mathbf{F}_{ci}^{\top}\mathbf{F}_{tj}\right))\right\}$, where $\mathrm{s}(\cdot,\cdot)$ denotes similarity computation, $\mathbf{F}_{ci}, \mathbf{F}_{tj} \in \mathbb{R}^{1\times D}$ represent the i-th composed feature and the j-th target feature in a mini-batch.

Based on the similarity matrix, we define the maximum positive matching likelihood p^+, and maximum negative matching likelihood p^- for each query to subsequently estimate matching loyalty degree:

$$p_i^+ = \max_j(\mathbf{s}_{ij} \cdot y_{ij}), \quad p_i^- = \max_j(\mathbf{s}_{ij} \cdot (1 - y_{ij})), \tag{6}$$

where y_{ij} denotes the pseudo-label matrix (diagonal elements are 1, others are 0), $\mathbf{s}_{ij}$ denotes the similarity between i-th query and j-th target.

We then perform dynamic decisions based on p^+ and p^-.

Negative Weight Reward. When all queries show low similarity to their positives within a batch, correct correspondences are more likely to exist among negatives. We thus apply Negative Weight Reward to negatives, as follows:

$$\mathbf{N} = \{\mathbf{n}_{ij} | \mathbf{n}_{ij} = (1 - p_i^+) \cdot (1 - y_{ij})\}, \tag{7}$$

where $\mathbf{n}_{ij}$ is the reward from i-th positive to j-th negative.

Positive Weight Reward. Complementarily, when all queries show low similarity to other negatives, positives are more likely to be cleanly labeled. We apply Positive Weight Reward to the positive group as follows:

$$\mathbf{R} = \{\mathbf{r}_{ij} | \mathbf{r}_{ij} = (1 - p_i^-) \cdot y_{ij}\}, \tag{8}$$

where $\mathbf{r}_{ij}$ is the reward from i-th negative to j-th positive.

Finally, we construct a scalable decision boundary via both rewards, obtaining loyalty degree estimation matrix $\mathbf{L}$:

$$\mathbf{L} = (\mathbf{S} + \mathbf{N} + \mathbf{R})/2 = \{\mathbf{l}_{ij} | \mathbf{l}_{ij} = (\mathbf{s}_{ij} + \mathbf{n}_{ij} + \mathbf{r}_{ij})/2\}. \tag{9}$$

Based on $\mathbf{L}$, we further propose the Soft Discriminative Loss, formulated as follows:

$$\mathcal{L}_{\text{sod}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B} y_{ij} \log(\mathbf{l}_{ij}), \tag{10}$$

where B is the batch size, $\mathbf{l}_{ij}$ is the loyalty degree of the i-th query regarding the j-th candidate.

Finally, we obtain the final loss function of INTENT as,

$$\mathbf{\Theta}^* = \arg\min_{\mathbf{\Theta}} \left(\mathcal{L}_{robust} + \mu\mathcal{L}_{sod} + \alpha\mathcal{L}_{caco}\right), \tag{11}$$

where $\mathbf{\Theta}^*$ is the to-be-optimized parameter for INTENT and μ, α are trade-off hyper-parameters.

Noise	Methods	Dress		Shirt		Toptee		Average		
		R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	AVG.
0%	SSN (Yang et al. 2024) (AAAI'24)	34.36	60.78	38.13	61.83	44.26	69.05	38.92	63.89	51.40
	CALA (Jiang et al. 2024) (SIGIR'24)	42.38	66.08	46.76	68.16	50.93	73.42	46.69	69.22	57.96
	SPRC (Xu et al. 2024) (ICLR'24)	49.18	72.43	55.64	73.89	**59.35**	78.58	54.72	74.97	64.85
	RCL (Hu et al. 2023a) (TPAMI'23)	48.79	72.68	55.89	73.90	56.91	77.41	53.86	74.66	64.26
	RDE (Qin et al. 2024) (CVPR'24)	47.84	71.89	54.37	73.55	56.91	77.21	53.04	74.22	63.63
	TME (Li et al. 2025a) (CVPR'25)	49.73	71.69	**56.43**	74.44	59.31	**78.94**	55.16	75.02	65.09
	INTENT (Ours)	**50.32**	**72.10**	56.32	**74.93**	59.28	78.45	**55.31**	**75.16**	**65.24**
20%	SSN (Yang et al. 2024) (AAAI'24)	22.61	45.56	27.87	48.58	31.82	55.28	27.43	49.81	38.62
	CALA (Jiang et al. 2024) (SIGIR'24)	29.05	51.36	35.28	56.23	36.05	58.24	33.46	55.28	44.37
	SPRC (Xu et al. 2024) (ICLR'24)	39.81	62.22	48.58	66.29	50.48	70.58	46.29	66.36	56.33
	RCL (Hu et al. 2023a) (TPAMI'23)	47.05	70.65	53.14	71.74	55.28	75.62	51.82	72.67	62.25
	RDE (Qin et al. 2024) (CVPR'24)	44.62	68.91	50.74	69.09	52.12	73.38	49.16	70.64	59.81
	TME (Li et al. 2025a) (CVPR'25)	49.03	70.35	**55.84**	73.16	57.22	78.23	54.03	73.91	63.97
	INTENT (Ours)	**49.32**	**71.43**	55.32	**73.57**	**58.01**	**78.46**	**54.22**	**74.49**	**64.36**
50%	SSN (Yang et al. 2024) (AAAI'24)	15.27	33.71	23.36	41.61	22.79	42.94	20.47	39.42	29.95
	CALA (Jiang et al. 2024) (SIGIR'24)	20.77	40.95	29.69	46.57	27.03	46.81	24.83	44.78	34.80
	SPRC (Xu et al. 2024) (ICLR'24)	35.94	57.16	42.25	61.63	44.98	64.76	41.06	61.19	51.12
	RCL (Hu et al. 2023a) (TPAMI'23)	43.68	66.44	50.74	69.19	52.63	73.84	49.01	69.82	59.42
	RDE (Qin et al. 2024) (CVPR'24)	41.30	64.75	47.06	66.34	50.13	70.63	46.16	67.24	56.70
	TME (Li et al. 2025a) (CVPR'25)	46.26	68.27	**53.09**	71.88	55.07	**76.59**	51.47	72.25	61.86
	INTENT (Ours)	**47.99**	**71.24**	52.78	**72.48**	**56.79**	76.23	**52.52**	**73.32**	**62.92**
80%	SSN (Yang et al. 2024) (AAAI'24)	11.16	25.24	16.98	30.72	17.03	32.64	15.05	29.53	22.29
	CALA (Jiang et al. 2024) (SIGIR'24)	14.28	30.59	19.73	35.82	19.48	36.10	17.83	34.41	26.00
	SPRC (Xu et al. 2024) (ICLR'24)	28.41	50.77	36.21	54.37	35.90	59.06	33.51	54.03	43.77
	RCL (Hu et al. 2023a) (TPAMI'23)	38.82	60.54	45.44	64.38	47.42	68.38	43.89	64.43	54.16
	RDE (Qin et al. 2024) (CVPR'24)	37.63	59.64	43.62	62.12	46.10	66.50	42.45	62.75	52.60
	TME (Li et al. 2025a) (CVPR'25)	41.45	64.35	47.30	68.20	51.25	73.23	46.67	68.60	57.63
	INTENT (Ours)	**42.07**	**65.58**	**50.38**	**69.41**	**53.09**	**73.91**	**48.51**	**69.63**	**59.07**

Table 2: Performance comparison on the FashionIQ validation set in terms of R@K(%). The best and second-best results are highlighted in **bold** and underline, respectively.

4 Experiments

This section delves into our comprehensive experiments of INTENT and the corresponding analyses. Following the SOTA method TME (Li et al. 2025a), the noise ratio is set to 20% for all ablation and parameter sensitivity experiments.

4.1 Experimental Settings

Datasets. Following previous works, we apply two standard datasets widely used in CIR for evaluation, including a fashion-domain dataset FashionIQ (Wu et al. 2021), and an open-domain dataset CIRR (Liu et al. 2021b).

Implementation Details. We adopt the BLIP-2 (Li et al. 2023a) backbone for INTENT. We set the number Q of learned queries for the Q-former to 32. And the embedding dimension D is set to 256. Through a comprehensive grid search, we set $\mu = 0.2, \alpha = 0.6$ for both datasets. We also adopt the temperature factor τ to 0.07 for Eqn (5). We trained INTENT for 10 epochs using the AdamW optimizer with the initial learning rate of 4e-5, while the batch size is set to 128 and the learning rate for CLIP is 1e-6. All experiments were conducted on a single NVIDIA A40 GPU.

Evaluation. We adopt the widely accepted metric Recall@K (short for R@K) to measure whether the target image is retrieved within top-K candidates. Following existing settings, for CIRR, we report overall recall at K=$1, 5, 10, 50$, and subset performance at K=$1, 2, 3$; for FashionIQ, we report for each of three categories at K=$10, 50$.

4.2 Performance Comparison

To assess the robustness of INTENT under NTC, we conduct experiments on CIRR and FashionIQ, comparing INTENT to both standard CIR models, and robust baselines. As shown in Table 1 and 2, we further observe that: **1) Robust methods outperform ordinary approaches.** According to Table 1 and 2, across all noise ratios, robust methods such as RCL and TME consistently outperform conventional models like CALA and SPRC with this gap growing as noise increases. For example, on CIRR with a noise ratio of 20%, SPRC is only 2.76% behind TME in Avg metric, and even outperforms some robust models on certain metrics. However, at 80% noise, SPRC lags behind TME by 13.84% in Avg. Similar results are observed on FashionIQ. These results highlight the high sensitivity of conventional methods, while robust approaches maintain strong performance under severe noise. **2) INTENT demonstrates superior robustness compared to other robust methods.** As Table 2 shows, on FashionIQ dataset, when noise ratio $\sigma = 20\%$, INTENT surpasses TME by 0.39% on Avg; as σ increases to 50% and 80%, this margin further widens to 1.06% and 1.44%, respectively. Similar trends appear on CIRR. These results highlight INTENT's robustness in NTC scenarios.

A#	Derivative	FashionIQ-Avg		CIRR-Avg	
		R@10	R@50	R@K	R_{sub}@K
Visual Invariant Composition (VIC)					
1	w/o VIC	53.05	72.98	79.36	86.12
2	w/o Intervention	54.01	73.67	80.07	87.52
Bi-Objective Discriminative Learning (BiODL)					
3	w/o PWR	53.37	73.97	79.80	87.28
4	w/o NWR	53.89	74.07	80.03	87.33
5	w/o Both_Raward	52.48	73.20	77.93	85.86
INTENT(Ours)		**54.22**	**74.49**	**80.17**	**88.60**

Table 3: The ablation study for **modules** of INTENT.

B#	Derivative	FashionIQ-Avg		CIRR-Avg	
		R@10	R@50	R@K	R_{sub}@K
Loss Functions					
1	w/o robust	51.14	72.25	78.33	86.06
2	w/o mask	50.46	70.87	74.26	82.88
3	w/o sod	51.06	72.07	77.67	84.65
4	$\mathcal{L}_{\text{caco}}$ w/ MSE	53.88	73.77	79.69	87.71
5	$\mathcal{L}_{\text{caco}}$ w/ L1	53.99	74.15	79.92	87.88
6	$\mathcal{L}_{\text{caco}}$ w/ L2	53.79	73.90	79.79	88.01
INTENT(Ours)		**54.22**	**74.49**	**80.17**	**88.60**

Table 4: The ablation study for **loss functions** of INTENT.

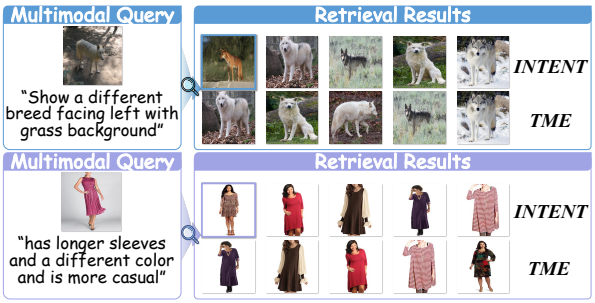

Figure 4: Case Study on (a) CIRR and (b) FashionIQ.

4.3 Ablation Study

To assess the contribution of each component in INTENT, we conduct comprehensive ablation studies on both datasets.

Ablation on modules and their components. As shown in Table 3, **A#1** and **A#2** denote variants where the VIC module is removed and where counterfactual images are replaced by grayscale images, respectively. **A#3** - **#5** correspond to removing $\mathbf{R}$, $\mathbf{N}$, or both from the loyalty degree matrix $\mathbf{L}$ in Eqn (9). Several key observations emerge from the results. 1) Both w/o VIC and w/o Intervention exhibit noticeable performance drops, demonstrating VIC's effectiveness in promoting visual invariance learning and reducing modality-inherent noise impact. The w/o Intervention still retains high performance, indicating the intervention process itself is robust and helps stabilize model training. 2) **A#3**-**#5**, all related to scalable decision boundaries, result in varying performance degradation. Notably, removing both rewards (w/o Both_Reward) leads to the largest drop, highlighting scalable decision boundary's advantage over relying solely on raw similarity scores. Moreover, w/o NWR slightly outperforms w/o PWR, which is expected since with 20% noise, most positives are real matches and thus benefit more from enhanced confidence in positive samples. These results validate the contribution of BiODL in improving discrimination and robustness to correspondence noise.

Ablation on loss functions. As shown in Table 4, **B#1** and **B#3** ablate the Robust Contrastive Loss and Soft Discriminative Loss in BiODL, while **B#2** applies $\mathcal{L}_{\text{robust}}$ without masking positives. **B#4** – **B#6** replace the CKA metric in VIC's Causal Consistency Loss with MSE, L1, or L2, respectively. We have the following observations. 1) Both w/o robust and w/o sod cause notable performance drops, confirming their complementary roles in discriminating noisy and real correspondence. 2) The largest decline occurs with w/o mask. This is expected since removing the mask causes the model to push positive samples apart, severely degrading performance. 3) $\mathcal{L}_{\text{caco}}$ w/ MSE, L1, L2 all lead to a performance decline compared to CKA. This may be due to CKA specifically measures similarities between the relational structures of two features (i.e., centered Gram matrices), rather than merely minimizing point-wise differences.

4.4 Case Study

To intuitively demonstrate the accuracy and robustness of INTENT, we compare its top-5 retrieval results with those of the sub-optimal robust method TME on FashionIQ and CIRR, as shown in Figure 4, where images in colored boxes are target images. INTENT successfully retrieves the target images at top-1 by capturing causal relationships during multimodal query composition and effectively ignoring potential modality-inherent noise. However, TME fails in these cases and does not retrieve the target even within the top-5 results. We attribute this to TME's lack of addressing noise prior to composition, and utlizing hard decision boundary to discriminate correspondences, without addressing modality-inherent noise or employing decision optimization.

5 Conclusion

In this work, we investigate neglected inherent noise and hard decision boundaries in NTC. To address these issues, we propose INTENT, comprising two components: Visual Invariant Composition and Bi-Objective Discriminative Learning. The former uses causal intervention via FFT to generate intervened composed features, promoting visual invariance and allowing the model to ignore modality-inherent noise during composition. The latter employs collaborative optimization with positive/ negative samples, establishing a scalable decision boundary that dynamically adjusts decisions according to the loyalty degree, enabling robust correspondence discrimination. Extensive experiments on two benchmarks reveal superiority and robustness of INTENT.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China, No.:62276155, No.:62576195, and No.:62572282

References

Batson, J.; and Royer, L. 2019. Noise2self: Blind denoising by self-supervision. In *ICML*, 524–533. PMLR.

Cao, F.; Xu, H.; Ru, J.; Li, Z.; Zhang, H.; and Liu, H. 2025. Collision Avoidance of Multi-UUV Systems Based on Deep Reinforcement Learning in Complex Marine Environments. *JMSE*, 13: 1615.

Chen, Y.; Zheng, Z.; Ji, W.; Qu, L.; and Chua, T.-S. 2024. Composed image retrieval with text feedback via multi-grained uncertainty regularization. *ICLR*.

Chen, Z.; Hu, Y.; Li, Z.; Fu, Z.; Song, X.; and Nie, L. 2025a. OFFSET: Segmentation-based Focus Shift Revision for Composed Image Retrieval. In *ACM MM*, 6113–6122. ACM.

Chen, Z.; Hu, Y.; Li, Z.; Fu, Z.; Wen, H.; and Guan, W. 2025b. HUD: Hierarchical Uncertainty-Aware Disambiguation Network for Composed Video Retrieval. In *ACM MM*, 6143–6152. ACM.

Dang, Z.; Luo, M.; Wang, J.; Jia, C.; Han, H.; Wan, H.; Dai, G.; Chang, X.; and Wang, J. 2025. Disentangled noisy correspondence learning. *IEEE TIP*.

Fu, Z.; Li, Z.; Chen, Z.; Wang, C.; Song, X.; Hu, Y.; and Nie, L. 2025. PAIR: Complementarity-guided Disentanglement for Composed Image Retrieval. In *ICASSP*, 1–5. IEEE.

Han, H.; Miao, K.; Zheng, Q.; and Luo, M. 2023. Noisy correspondence learning with meta similarity correction. In *CVPR*, 7517–7526.

He, Y.; Li, S.; Wang, J.; Li, K.; Song, X.; Yuan, X.; Li, K.; Lu, K.; Huo, M.; Tang, J.; et al. 2025. Enhancing low-cost video editing with lightweight adaptors and temporal-aware inversion. *arXiv preprint arXiv:2501.04606*.

He, Y.; Wang, X.; and Shi, T. 2024. Ddpm-moco: Advancing industrial surface defect generation and detection with generative and contrastive learning. In *IJCAI*, 34–49. Springer.

Hu, P.; Huang, Z.; Peng, D.; Wang, X.; and Peng, X. 2023a. Cross-modal retrieval with partially mismatched pairs. *IEEE TPAMI*, 45(8): 9595–9610.

Hu, Y.; Liu, M.; Su, X.; Gao, Z.; and Nie, L. 2021a. Video moment localization via deep cross-modal hashing. *IEEE TIP*, 30: 4667–4677.

Hu, Y.; Nie, L.; Liu, M.; Wang, K.; Wang, Y.; and Hua, X.-S. 2021b. Coarse-to-fine semantic alignment for cross-modal moment localization. *IEEE TIP*, 30: 5933–5943.

Hu, Y.; Wang, K.; Liu, M.; Tang, H.; and Nie, L. 2023b. Semantic collaborative learning for cross-modal moment localization. *ACM TOIS*, 42(2): 1–26.

Huang, F.; Zhang, L.; Fu, X.; and Song, S. 2024a. Dynamic weighted combiner for mixed-modal image retrieval. In *AAAI*, volume 38, 2303–2311.

Huang, J.; Du, L.; Chen, X.; Fu, Q.; Han, S.; and Zhang, D. 2023. Robust mid-pass filtering graph convolutional networks. In *ACM WWW*, 328–338.

Huang, J.; Mo, Y.; Hu, P.; Shi, X.; Yuan, S.; Zhang, Z.; and Zhu, X. 2024b. Exploring the Role of Node Diversity in Directed Graph Representation Learning. In *IJCAI*.

Huang, J.; Mo, Y.; Shi, X.; Feng, L.; and Zhu, X. 2025a. Enhancing the Influence of Labels on Unlabeled Nodes in Graph Convolutional Networks. In *ICML*.

Huang, J.; Shen, J.; Shi, X.; and Zhu, X. 2024c. On Which Nodes Does GCN Fail? Enhancing GCN From the Node Perspective. In *ICML*.

Huang, J.; Xu, J.; Shi, X.; Hu, P.; Feng, L.; and Zhu, X. 2025b. The Final Layer Holds the Key: A Unified and Efficient GNN Calibration Framework. *arXiv preprint arXiv:2505.11335*.

Huang, Q.; Chen, Z.; Li, Z.; Wang, C.; Song, X.; Hu, Y.; and Nie, L. 2025c. MEDIAN: Adaptive Intermediate-grained Aggregation Network for Composed Image Retrieval. In *ICASSP*, 1–5. IEEE.

Huang, Z.; Niu, G.; Liu, X.; Ding, W.; Xiao, X.; Wu, H.; and Peng, X. 2021. Learning with noisy correspondence for cross-modal matching. *NeurIPS*, 34: 29406–29419.

Jiang, W.; Zhang, S.; You, S.; Feng, P.; and Lu, Z. 2025. Traditional Chinese Painting Completion via Hierarchical Optimal Transport. *IEEE Access*.

Jiang, X.; Wang, Y.; Li, M.; Wu, Y.; Hu, B.; and Qian, X. 2024. Cala: Complementary association learning for augmenting comoposed image retrieval. In *ACM SIGIR*, 2177–2187.

Kong, F.; Zhang, J.; Liu, Y.; Zhang, H.; Feng, S.; Yang, X.; Wang, D.; Tian, Y.; Zhang, F.; Zhou, G.; et al. 2025. Modality curation: Building universal embeddings for advanced multimodal information retrieval. *arXiv preprint arXiv:2505.19650*.

Kornblith, S.; Norouzi, M.; Lee, H.; and Hinton, G. 2019. Similarity of neural network representations revisited. In *ICML*, 3519–3529. PMlR.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 19730–19742. PMLR.

Li, S.; He, C.; Liu, X.; Zhou, J. T.; Peng, X.; and Hu, P. 2025a. Learning with Noisy Triplet Correspondence for Composed Image Retrieval. In *CVPR*, 19628–19637.

Li, W.; Su, X.; Song, D.; Wang, L.; Zhang, K.; and Liu, A.-A. 2023b. Towards deconfounded image-text matching with causal inference. In *ACM MM*, 6264–6273.

Li, Y.; Chen, C.; Zhang, Y.; Liu, W.; Lyu, L.; Zheng, X.; Meng, D.; and Wang, J. 2023c. Ultrare: Enhancing receraser for recommendation unlearning via error decomposition. *NeurIPS*, 12611–12625.

Li, Y.; Zhang, Y.; Liu, W.; Feng, X.; Han, Z.; Chen, C.; and Yan, C. 2025b. Multi-Objective Unlearning in Recommender Systems via Preference Guided Pareto Exploration. *IEEE TSC*.

Li, Z.; Chen, Z.; Wen, H.; Fu, Z.; Hu, Y.; and Guan, W. 2025c. ENCODER: Entity Mining and Modification Relation Binding for Composed Image Retrieval. In *AAAI*.

Li, Z.; Fu, Z.; Hu, Y.; Chen, Z.; Wen, H.; and Nie, L. 2025d. FineCIR: Explicit Parsing of Fine-Grained Modification Semantics for Composed Image Retrieval. *https://arxiv.org/abs/2503.21309*.

Liu, F.; Cheng, Z.; Zhu, L.; Gao, Z.; and Nie, L. 2021a. Interest-aware message-passing GCN for recommendation. In *ACM WWW*, 1296–1305.

Liu, F.; Liu, Y.; Chen, H.; Cheng, Z.; Nie, L.; and Kankanhalli, M. 2025a. Understanding Before Recommendation: Semantic Aspect-Aware Review Exploitation via Large Language Models. *ACM TOIS*, 43(2).

Liu, H.; Li, X.; Zhang, X.; Liu, G.; and Lu, M. 2025b. In-Pipe Navigation Development Environment and a Smooth Path Planning Method on Pipeline Surface. In *ICRA*, 128084–128090. IEEE.

Liu, J.; Shang, F.; Zhou, J.; Liu, H.; Liu, Y.; and Liu, J. 2025c. FedMuon: Accelerating Federated Learning with Matrix Orthogonalization. *arXiv preprint arXiv:2510.27403*.

Liu, J.; Tian, Y.; Shang, F.; Liu, Y.; Liu, H.; Zhou, J.; and Ding, D. 2025d. DP-FedPGN: Finding Global Flat Minima for Differentially Private Federated Learning via Penalizing Gradient Norm. *arXiv preprint arXiv:2510.27504*.

Liu, K.; Gong, Y.; Cao, Y.; Ren, Z.; Peng, D.; and Sun, Y. 2024a. Dual semantic fusion hashing for multi-label cross-modal retrieval. In *IJCAI*, 4569–4577.

Liu, M.; Wang, X.; Nie, L.; He, X.; Chen, B.; and Chua, T.-S. 2018a. Attentive moment retrieval in videos. In *SIGIR*, 15–24.

Liu, M.; Wang, X.; Nie, L.; Tian, Q.; Chen, B.; and Chua, T.-S. 2018b. Cross-modal moment localization in videos. In *ACM MM*, 843–851.

Liu, X.; Lu, Y.; Wang, X.; and Wu, X. 2025e. Training-Free Multi-Style Fusion Through Reference-Based Adaptive Modulation. arXiv:2509.18602.

Liu, Y.; Qin, G.; Chen, H.; Cheng, Z.; and Yang, X. 2024b. Causality-inspired invariant representation learning for text-based person retrieval. In *AAAI*, volume 38, 14052–14060.

Liu, Z.; Opazo, C. R.; Teney, D.; and Gould, S. 2021b. Image Retrieval on Real-life Images with Pre-trained Vision-and-Language Models. In *ICCV*, 2105–2114. IEEE.

Lu, S.; Liu, Y.; and Kong, A. W.-K. 2023. Tf-icon: Diffusion-based training-free cross-domain image composition. In *ICCV*, 2294–2305.

Lu, S.; Zhou, Z.; Lu, J.; Zhu, Y.; and Kong, A. W.-K. 2024. Robust watermarking using generative priors against image editing: From benchmarking to advances. *arXiv preprint arXiv:2410.18775*.

Ma, Z.; Luo, Y.; Zhang, Z.; Sun, A.; Yang, Y.; and Liu, H. 2025. Reinforcement Learning Approach for Highway Lane-Changing: PPO-Based Strategy Design.

Niu, Y.; Tang, K.; Zhang, H.; Lu, Z.; Hua, X.-S.; and Wen, J.-R. 2021. Counterfactual vqa: A cause-effect look at language bias. In *CVPR*, 12700–12710.

Ou, Y.; de Bruijn, G.-J.; and Schulz, P. J. 2025. Social Media as an Emotional Barometer: Bidirectional Encoder Representations From Transformers–Long Short-Term Memory Sentiment Analysis on the Evolution of Public Sentiments During Influenza A on Sina Weibo. *JMIR*, 27: e68205.

Pu, R.; Qin, Y.; Song, X.; Peng, D.; Ren, Z.; and Sun, Y. 2025a. SHE: Streaming-media Hashing Retrieval. In *ICML*.

Pu, R.; Sun, Y.; Qin, Y.; Ren, Z.; Song, X.; Zheng, H.; and Peng, D. 2025b. Robust Self-Paced Hashing for Cross-Modal Retrieval with Noisy Labels. In *AAAI*, volume 39, 19969–19977.

Qin, Y.; Chen, Y.; Peng, D.; Peng, X.; Zhou, J. T.; and Hu, P. 2024. Noisy-correspondence learning for text-to-image person re-identification. In *CVPR*, 27197–27206.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

Sun, Y.; Qin, Y.; Peng, D.; Ren, Z.; Yang, C.; and Hu, P. 2024. Dual self-paced hashing for image retrieval. *IEEE TMM*, 26: 9619–9629.

Sun, Y.; Ren, Z.; Hu, P.; Peng, D.; and Wang, X. 2023. Hierarchical consensus hashing for cross-modal retrieval. *IEEE TMM*, 26: 824–836.

Tan, C.; Xia, J.; Wu, L.; and Li, S. Z. 2021. Co-learning: Learning from noisy labels with self-supervision. In *ACM MM*, 1405–1413.

Tang, H.; Hu, Y.; Wang, Y.; Zhang, S.; Xu, M.; Zhu, J.; and Zheng, Q. 2024. Listen as you wish: Fusion of audio and text for cross-modal event detection in smart cities. *Information Fusion*, 110: 102460.

Tian, Y.; Liu, F.; Zhang, J.; Bi, W.; Hu, Y.; and Nie, L. 2025a. Open Multimodal Retrieval-Augmented Factual Image Generation. *arXiv preprint arXiv:2510.22521*.

Tian, Y.; Liu, F.; Zhang, J.; W., V.; Hu, Y.; and Nie, L. 2025b. CoRe-MMRAG: Cross-Source Knowledge Reconciliation for Multimodal RAG. In *ACL*, 32967–32982.

Ting, Y.; and Listening, C. 2024. When Radio Become a Broadcasting Application.

Vo, N.; Jiang, L.; Sun, C.; Murphy, K.; Li, L.; Fei-Fei, L.; and Hays, J. 2019. Composing Text and Image for Image Retrieval - an Empirical Odyssey. In *CVPR*, 6439–6448. IEEE.

Wang, J.; Tang, Y.; Wang, Y.; Yuan, Z.; Wang, H.; He, Y.; and Li, B. 2025. See the Forest and the Trees: A Synergistic Reasoning Framework for Knowledge-Based Visual Question Answering. *arXiv preprint arXiv:2507.17659*.

Wang, Y.; Zhang, F.-L.; and Dodgson, N. A. 2024. Scantd: 360° scanpath prediction based on time-series diffusion. In *ACM MM*, 7764–7773.

Wang, Y.; Zhang, F.-L.; and Dodgson, N. A. 2025. Target Scanpath-Guided 360-Degree Image Enhancement. In *AAAI*, volume 39, 8169–8177.

Wei, Y.; Wang, X.; Nie, L.; He, X.; and Chua, T.-S. 2020. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In *ACM MM*, 3541–3549.

Wei, Y.; Wang, X.; Nie, L.; He, X.; Hong, R.; and Chua, T.-S. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *ACM MM*, 1437–1445.

Wen, H.; Song, X.; Yang, X.; Zhan, Y.; and Nie, L. 2021. Comprehensive Linguistic-Visual Composition Network for Image Retrieval. In *ACM SIGIR*, 1369–1378. ACM.

Wen, H.; Song, X.; Yin, J.; Wu, J.; Guan, W.; and Nie, L. 2023a. Self-Training Boosted Multi-Factor Matching Network for Composed Image Retrieval. *IEEE TPAMI*.

Wen, H.; Zhang, X.; Song, X.; Wei, Y.; and Nie, L. 2023b. Target-guided composed image retrieval. In *ACM MM*, 915–923.

Wu, H.; Gao, Y.; Guo, X.; Al-Halah, Z.; Rennie, S.; Grauman, K.; and Feris, R. 2021. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *CVPR*, 11307–11317.

Wu, Y.; Liu, X.; Zhao, C.; and Wu, X. 2025. Prompt-Guided Dual Latent Steering for Inversion Problems. arXiv:2509.18619.

Xu, M.; Yu, C.; Li, Z.; Tang, H.; Hu, Y.; and Nie, L. 2025. Hdnet: A hybrid domain network with multi-scale high-frequency information enhancement for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*.

Xu, X.; Liu, Y.; Khan, S.; Khan, F.; Zuo, W.; Goh, R. S. M.; Feng, C.-M.; et al. 2024. Sentence-level Prompts Benefit Composed Image Retrieval. In *ICLR*.

Yang, X.; Liu, D.; Zhang, H.; Luo, Y.; Wang, C.; and Zhang, J. 2024. Decomposing Semantic Shifts for Composed Image Retrieval. In *AAAI*, volume 38, 6576–6584.

Yi-fan, O. 2016. Communication and operation of TV WeChat official account. *Journalism and Mass Communication*, 6(12): 730–736.

Yifan, O. 2018. Participating in Chinese Social Question and Answer Communities: A Case Study of Zhihu. com.

Zhang, H.; Liu, M.; Li, Y.; Yan, M.; Gao, Z.; Chang, X.; and Nie, L. 2023. Attribute-guided collaborative learning for partial person re-identification. *IEEE TPAMI*, 45(12): 14144–14160.

Zhang, H.; Liu, M.; Li, Z.; Wen, H.; Guan, W.; Wang, Y.; and Nie, L. 2025. Spatial Understanding from Videos: Structured Prompts Meet Simulation Data. In *NeurIPS*, 1–16.

Zhao, Z. 2024. Balf: Simple and efficient blur aware local feature detector. In *WACV*, 3362–3372.

Zhao, Z.; and Chen, B. M. 2023. Benchmark for Evaluating Initialization of Visual-Inertial Odometry. In *CCC*, 3935–3940. IEEE.