HABIT: Chrono-Synergia Robust Progressive Learning Framework for Composed Image Retrieval

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

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

Abstract

Composed Image Retrieval (CIR) is a flexible image retrieval paradigm that enables users to accurately locate the target image through a multimodal query composed of a reference image and modification text. Although this task has demonstrated promising applications in personalized search and recommendation systems, it encounters a severe challenge in practical scenarios known as the Noise Triplet Correspondence (NTC) problem. This issue primarily arises from the high cost and subjectivity involved in annotating triplet data. To address this problem, we identify two central challenges: the **precise estimation of composed semantic discrepancy** and the **insufficient progressive adaptation to modification discrepancy**. To tackle these challenges, we propose a c**H**rono-synergi**A** ro**B**ust progress**I**ve learning framework for composed image re**T**rieval (**HABIT**), which consists of two core modules. First, the *Mutual Knowledge Estimation* Module quantifies sample cleanliness by calculating the Transition Rate of mutual information between the composed feature and the target image, thereby effectively identifying clean samples that align with the intended modification semantics. Second, the *Dual-consistency Progressive Learning* Module introduces a collaborative mechanism between the historical and current models, simulating human habit formation to retain good habits and calibrate bad habits, ultimately enabling robust learning under the presence of NTC. Extensive experiments conducted on two standard CIR datasets demonstrate that HABIT significantly outperforms most methods under various noise ratios, exhibiting superior robustness and retrieval performance.

1 Introduction

With the rapid growth of image data (Yi et al. 2025; Zhang et al. 2025a, 2023; Lu et al. 2025; Zhou et al. 2024), Composed Image Retrieval (CIR) (Li et al. 2025c; Wen et al. 2023a; Xu et al. 2024) has emerged as a key research focus in information retrieval (Liu et al. 2021a; Yang et al. 2024; Jiang et al. 2024; Pu et al. 2025a; Kong et al. 2025; Sun et al. 2023b; Hu et al. 2021b) and multimodal learning (Huang et al. 2024b; Lu et al. 2024a). As shown in Figure 1(a), CIR differs from unimodal retrieval by supporting multimodal queries consisting of a reference image and modification text, which enables the retrieval of a semantically consistent target image. Although significant progress has been made in the CIR task, real-world applications (Huang et al. 2025d; Li et al. 2025b; Liu et al. 2025b,c; Cao et al. 2025; Wang, Zhang, and Dodgson 2024, 2025; Ou, de Bruijn, and Schulz 2025; Yi-fan 2016) still face substantial challenges. Due to the high cost and inherent subjectivity of triplets annotation, practical datasets are often affected by annotation errors and inaccurate semantic alignment. Moreover, this issue is further exacerbated in large-scale datasets involving large models, owing to the hallucination problem commonly observed in such models. To mitigate this problem, Li et al. (Li et al. 2025a) introduced the Noise Triplet Correspondence (NTC) to enhance the robustness of CIR models.

Figure 1: (a) presents an example of the CIR paradigm. (b) illustrates the commonly observed "unmentioned visual discrepancies" in CIR task, which increase the difficulty of identifying Noise Triplet Correspondence. (c) depicts our proposed Chrono-Synergia Mechanism.

Unlike traditional cross-modal retrieval tasks (Tian et al. 2025b,a; Wei et al. 2019, 2020) such as video-text match-

*Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

ing (Liu et al. 2018a; Hu et al. 2021a; Liu et al. 2018b; Hu et al. 2023b; Liu et al. 2025a), CIR (Wen et al. 2023b; Li et al. 2025c; Chen et al. 2025b,a) is inherently a semantic modification task where the reference image and modification text often contain inconsistent semantics, increasing the complexity of the NTC problem. As a result, existing robust methods for cross-modal matching (Mu, Yang, and Deng 2025; Zha et al. 2025) fail to generalize to CIR. Moreover, during the CIR annotation process (Wu et al. 2021; Guo et al. 2018), annotators receive a reference and target image pair and are asked to describe their differences as modification text. Due to annotation cost constraints, other candidate images are not shown, leading to modification texts that are often brief and omit subtle differences. As shown in Figure 1(b), visual details such as "sky" and "mountains" in the reference image as well as "tree" and "house" in the target image are frequently excluded, resulting in unmentioned visual discrepancies. These discrepancies create persistent semantic gaps between the composed feature and the target image, which vary across samples. However, current robust CIR methods such as TME (Li et al. 2025a) rely heavily on feature similarity to identify noisy correspondences, which may misclassify samples with diverse semantic gaps and ultimately limit model performance and generalization.

However, addressing the above limitations remains challenging due to two key factors. **(1) Precise estimation of composed semantic discrepancy.** To identify unmentioned visual discrepancies and circumvent the limitations of similarity-based noise assessment methods, it is essential to develop an approach capable of accurately quantifying the fine-grained semantic consistency between the composed feature and the target image. Moreover, the semantic fusion of the reference image and modification text in CIR is inherently complex, making it difficult to directly apply conventional robust learning methods. Thus, developing an effective estimation mechanism for composed semantic discrepancy constitutes the first major challenge. **(2) Insufficient progressive adaptation to modification discrepancy.** Due to the intrinsic semantic gap between the reference image and modification text, CIR models often misinterpret this discrepancy during early training stages, leading to clean positive samples being wrongly treated as noisy correspondences. Therefore, the second challenge lies in designing a learning strategy that progressively enhances the model's ability to adapt to diverse triplet compositions and continuously reduces the risk of misdetermination.

To address the aforementioned challenges, we propose a c**H**rono-synergi**A** ro**B**ust progress**I**ve learning framework for composed image re**T**rieval (**HABIT**). HABIT tackles the NTC problem in CIR by modeling mutual knowledge and leveraging progressive learning via dual knowledge collaboration. It comprises two key modules: (1) *Mutual Knowledge Estimation* Module quantifies sample cleanliness by computing the Transition Rate of mutual knowledge between the composed feature and the target image. This facilitates the reliable identification of clean samples that align with the intended modification semantics, improving noise recognition accuracy. (2) *Dual-consistency Progressive Learning* addresses insufficient adaptation to modification discrepancy. As shown in Figure 1(c), we introduce the Chrono-Synergia Mechanism, which simulates human habit formation by integrating predictions from both historical and current models. Historically consistent decisions are retained as good habits, while inconsistent ones are calibrated as bad habits, enabling robust learning in the presence of NTC.

In summary, the contributions of this paper are threefold:

- We conduct an in-depth analysis of the NTC problem in CIR, and for the first time, we identify two key challenges faced by existing methods: the precise estimation of composed semantic discrepancy and the insufficient progressive adaptation to modification discrepancy.
- We propose a novel robust learning framework for CIR, named HABIT, which employs the Transition Rate of mutual knowledge for accurate noise-aware label assignment. In addition, HABIT simulates human habit formation and achieves robust learning under noisy triplet conditions through Dual-consistency Progressive Learning.
- Extensive experiments on two standard CIR benchmarks demonstrate that HABIT outperforms most methods across varying noise levels, significantly improving retrieval performance in noisy settings.

2 Related Work

Conventional Composed Image Retrieval. Composed Image Retrieval (CIR) plays a crucial role in computer vision (Sunmola et al. 2025; Bellavia et al. 2024; Zhang et al. 2025b, 2021) and multimodal learning (Liu et al. 2025d; Wu et al. 2025; Xu et al. 2025), aiming to retrieve the target image based on a reference image and modification text. Existing approaches (Li et al. 2025c; Huang et al. 2025c; Fu et al. 2025; Li et al. 2025d) can generally be categorized into two types. The first category consists of models (Wen et al. 2021) that use traditional architectures (e.g., ResNet, LSTM) to extract image and text features, followed by the multimodal composition. In contrast, the second group (Li et al. 2025c,d) leverages Vision-Language Pretrained (VLP) models such as CLIP (Radford et al. 2021), BLIP-2 (Li et al. 2023), to extract multimodal query features, then apply relatively simple strategies for feature alignment and composition to achieve remarkable performance (Jiang et al. 2024; Xu et al. 2024). Although some methods have investigated false positive samples (Chen et al. 2024), their primary objective is to improve the performance of traditional CIR, making them less effective from the perspective of NTC.

Noisy Correspondence Learning. While robust learning (Pu et al. 2025b; Lu et al. 2024b) has been widely explored in multimodal tasks (Huang et al. 2025a; Feng et al. 2023; Huang et al. 2025b; Li et al. 2025a; Huang et al. 2023; Yifan 2018; Tang et al. 2021), recent work increasingly focuses on the more complex issue of noisy correspondence across modalities. This type of noise, which surpasses simple label errors, often induces overfitting and performance degradation. Existing approaches in visual-language pretraining and cross-modal learning (Huang et al. 2024c; Li et al. 2025a; Liu et al. 2024; Sun et al. 2023a; Feng and Peng 2014; Ting and Listening 2024) mainly consider two-modality noise, overlooking the challenges in real-

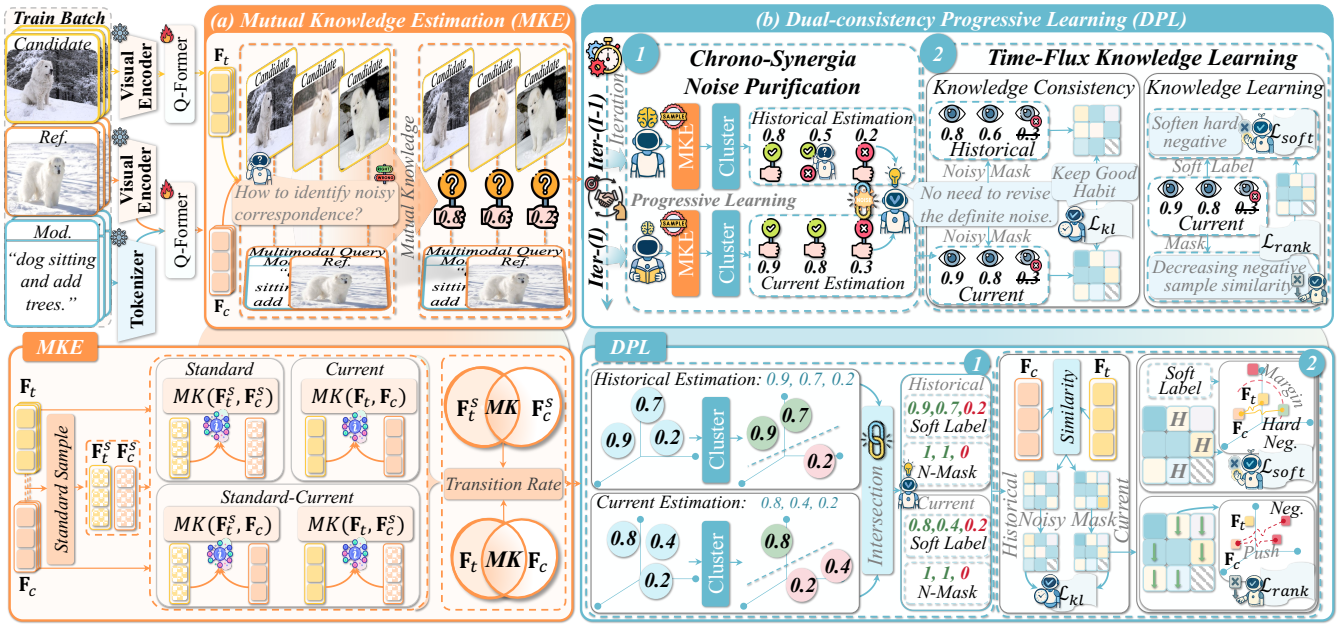

Figure 2: HABIT consists of two modules: (a) Mutual Knowledge Estimation and (b) Dual-consistency Progressive Learning.

world Composed Image Retrieval. To address this, Li et al. (Li et al. 2025a) introduced the Noise Triplet Correspondence (NTC) problem, underscoring semantic ambiguity and annotation errors in large-scale datasets. While prior alignment-based strategies (Li et al. 2025a; Huang et al. 2024a) have enhanced robustness, our HABIT further advances this by leveraging fine-grained mutual knowledge modeling and Dual-consistency Progressive Learning, yielding substantial gains in robustness to NTC.

3 HABIT

As a primary innovation, the proposed HABIT is designed to quantify sample cleanliness via the mutual knowledge transition rate and to achieve robust learning in the *NTC* environment by simulating human habit formation through a dual-consistency progressive learning strategy. As shown in Figure 2, HABIT comprises two core modules: *(a) Mutual Knowledge Estimation (MKE)* and *(b) Dual-consistency Progressive Learning (DPL)*. In this section, we first formulate CIR task with NTC and then elaborate on each module.

3.1 Problem Formulation

The CIR task seeks to retrieve the target image that matches a given multimodal query from an image database. In practice, CIR datasets frequently suffer from annotation errors within their triplets, termed NTC, with such erroneous samples referred to as noisy triplets. Noisy triplets typically fall into two types: (1) partial match: the modification text x_m partially describes the transformation from the reference image x_r to the target image x_t; (2) full mismatch: x_m completely misrepresents the modification. Following TME (Li et al. 2025a), we simulate noisy scenarios by randomly selecting a subset of training triplets according to a noise ratio σ. Given a triplet set with NTC, $\mathcal{T} = \{\langle x_r, x_m, x_t \rangle_n\}_{n=1}^{N}$, where x_r, x_m, and x_t may not be properly aligned, the objective is to learn an embedding function $\mathcal{G}$ that maps the multimodal query (x_r, x_m) close to its corresponding target image x_t in a shared metric space, as $\mathcal{G}(x_r, x_m) \rightarrow \mathcal{G}(x_t)$, where $\mathcal{G}$ is the embedding function to be learned, mapping multimodal queries and target images into one metric space.

3.2 Mutual Knowledge Estimation (MKE)

This module quantifies sample cleanliness by measuring the mutual knowledge transition rate between the composed feature and the target image, enabling effective identification of samples whose semantics align with the modification text and improving noisy correspondence detection. First, the module extracts the composed feature and modality features from the target image. Next, we compute the semantic matching degree between the composed feature and the target feature based on mutual knowledge across modalities, and the mutual knowledge transition rate is defined to capture semantic discrepancies. This metric serves as the criterion for evaluating the semantic noise margin.

Specifically, we employ Q-Former, which is proven effective in various CIR models (Xu et al. 2024; Ventura et al. 2024), to extract cross-modal features from both visual and textual modalities, extracting precise composed and target features, as formulated below,

$$\mathbf{F}_c = \text{Q-Former}(\Phi_{\mathbb{I}}(x_r), \Phi_{\mathbb{T}}(x_m)), \mathbf{F}_t = \text{Q-Former}(\Phi_{\mathbb{I}}(x_t)), \quad (1)$$

where $\mathbf{F}_c, \mathbf{F}_t \in \mathbb{R}^{Q \times D}$ denote the composed feature and the target feature, respectively, Q is the number of learnable queries, D is the embedding dimension, and $\Phi_{\mathbb{I}}$ and $\Phi_{\mathbb{T}}$ denote the visual encoder and the text tokenizer, respectively.

Subsequently, we define the mutual knowledge between the composed feature $\mathbf{F}_c$ and the target feature $\mathbf{F}_t$ for any

sample in the batch, which is used to measure the semantic matching degree between the two features, as formulated as,

$$\mathrm{MK}(\mathbf{F}_c, \mathbf{F}_t) = \sum_{\mathbf{f}_c \in \mathbf{F}_c} \sum_{\mathbf{f}_t \in \mathbf{F}_t} p(\mathbf{f}_c, \mathbf{f}_t) \log \frac{p(\mathbf{f}_c, \mathbf{f}_t)}{p(\mathbf{f}_c)p(\mathbf{f}_t)}, \quad (2)$$

where $p(\mathbf{f}_c, \mathbf{f}_t)$ denotes the joint probability distribution between $\mathbf{f}_c$ and $\mathbf{f}_t$, and $p(\mathbf{f}_c)$ and $p(\mathbf{f}_t)$ denote the marginal probability distributions of $\mathbf{f}_c$ and $\mathbf{f}_t$, respectively.

However, the mutual knowledge value alone does not provide a clear margin between noisy and correctly matched samples. To address this, inspired by (Lyu et al. 2025), in each batch, we select the triplet with the lowest loss as the *Standard Sample* $(\mathbf{F}_c^s, \mathbf{F}_t^s)$, assumed to be the cleanest correspondence. We then calculate the mutual knowledge discrepancy of all other samples relative to this standard sample to define the mutual knowledge transition rate, which quantifies each sample's discrepancy from the standard and serves as the estimation of noisy correspondence, formulated as,

$$\mathrm{TR}(\mathbf{F}_c, \mathbf{F}_t) = \frac{|\,\mathrm{MK}(\mathbf{F}_c^s, \mathbf{F}_t^s) - \mathrm{MK}(\mathbf{F}_c, \mathbf{F}_t)|}{\mathrm{MK}(\mathbf{F}_c^s, \mathbf{F}_t^s)}. \quad (3)$$

Furthermore, since noisy correspondence can originate from either the composed feature (i.e., reference image/-modification text side) or the target feature, we also consider the transition rates between each composed/target feature and the corresponding standard sample, denoted as $\mathrm{TR}(\mathbf{F}_c, \mathbf{F}_t^s)$ and $\mathrm{TR}(\mathbf{F}_t, \mathbf{F}_c^s)$. Using these transition rates, we estimate the cleanliness of each triplet sample as follows,

$$\mathbb{E} = (1 + \mathrm{TR}(\mathbf{F}_c, \mathbf{F}_t) + |\,\mathrm{TR}(\mathbf{F}_c, \mathbf{F}_t^s) - \mathrm{TR}(\mathbf{F}_t, \mathbf{F}_c^s)|)^{-1}. \quad (4)$$

To be specific, when the transition rate between a sample's composed feature and its target feature is small, and when both $\mathrm{TR}(\mathbf{F}_c, \mathbf{F}_t^s)$ and $\mathrm{TR}(\mathbf{F}_t, \mathbf{F}_c^s)$ are close to that of the standard sample, it indicates that the matching relationship between the multimodal query composed feature and the target image is more reliable, i.e. the sample is more likely a clean correspondence.

3.3 Dual-consistency Progressive Learning (DPL)

The *Dual-consistency Progressive Learning (DPL)* module is designed to address the limited progressive adaptivity caused by modification discrepancies, allowing the model to incrementally adapt to the challenges of noisy correspondence detection and enhance robustness. Building on sample cleanliness estimations from the MKE module, DPL incorporates the *Chrono-Synergia Noise Discrimination* and *Time-Flux Knowledge Updating*. These enable HABIT to simulate human habit formation by integrating predictions from both the historical and current models, preserving good habits and calibrating bad habits for robust learning in NTC scenarios. We now explain this module in detail.

Chrono-Synergia Noise Discriminantion Intuitively, the cleanliness estimations of standard samples remain relatively stable during training, whereas those of noisy correspondence samples tend to fluctuate and deviate from clean samples. Based on this, we introduce the *Chrono-Synergia Noise Discrimination*, which aims to chrono-synergize the dynamic changes of cleanliness estimations to identify noisy triplets that fall outside the normal matching range.

Specifically, we denote $\mathbf{e}^{(I)} = [\mathbb{E}_1^{(I)}, \ldots, \mathbb{E}_B^{(I)}]$ as the estimation sequence for all samples in the batch at the I-th iteration, where B is the batch size. For each iteration's estimation sequence, we apply the DBSCAN (Ester et al. 1996) to obtain the set of outlier samples $\mathcal{O}^{(I)}$ for that iteration. Similarly, by applying the same procedure to the historical estimation sequence from the previous iteration $(I{-}1)$, we obtain the corresponding outlier set $\mathcal{O}^{(I-1)}$.

Subsequently, we merge the outlier sets from both the historical and current iterations to compute the noisy mask $\mathbf{m}^{(I)} = [m_1, \ldots, m_b, \ldots, m_B]$ for the current batch at iteration I, where the noisy mask m_b for the b-th triplet sample in the batch, formalized as follows,

$$m_b = \begin{cases} 0, & \text{if } \{b\} \in \mathcal{O}^{(I)} \cap \mathcal{O}^{(I-1)} \\ 1, & \text{otherwise} \end{cases}. \quad (5)$$

The resulting noisy mask can accurately identify samples that are outliers at both time points, synergizing the discrimination capabilities of the historical and current models for noisy triplets, and thereby continuously filtering out those noisy triplets that are stably recognized.

Time-Flux Knowledge Updating Since the noise discrimination mechanism applies strict criteria, some mismatched samples may still be misclassified, leading to model's "bad habits". To further improve robustness against residual noisy correspondence and calibrate these bad habits while keeping good ones, we introduce *Time-Flux Knowledge Updating*, which utilizes the temporal evolution of cleanliness estimations to maintain semantic consistency and clear time-flux margins, thus enhancing training robustness.

Knowledge Consistency. Before calibrating model's bad habits, it is essential to preserve its learned good habits. Thus, to ensure semantic consistency and maintain accurate matching capabilities throughout training, we introduce *Knowledge Consistency*. It constrains the evolution of the model's triplet semantic matching degree across iterations. By comparing the similarity distributions produced at the current iteration I and the previous iteration $(I-1)$, this mechanism prevents the model from losing its "good habits" (the correctly learned determination capabilities).

Specifically, let I denote the current iteration. Define $\mathbf{s}^{(I)} \in \mathbb{R}^{B \times B}$ as the similarity matrix between all multimodal query features $\mathbf{F}_c$ and all target image features $\mathbf{F}_t$ in the current batch, and let $\mathbf{s}^{(I-1)}$ denote the corresponding similarity matrix from the previous iteration on the same batch. Correspondingly, $\mathbf{m}^{(I)}$ and $\mathbf{m}^{(I-1)}$ represent the current and previous noisy masks, respectively. We enforce a consistency constraint on "good habits" by minimizing the Kullback-Leibler (KL) divergence between the similarity distributions of consecutive iterations, as follows,

$$\mathcal{L}_{KL} = \frac{1}{B} \sum_{b=1}^{B} D_{KL}\left(\left(\mathbf{s}_b^{(I)} \cdot \mathbf{m}_b^{(I)}\right) \| \left(\mathbf{s}_b^{(I-1)} \cdot \mathbf{m}_b^{(I-1)}\right)\right), \quad (6)$$

where $\mathbf{s}_b^{(I)}, \mathbf{s}_b^{(I-1)}$ denote the similarity vectors of the b-th multimodal query within the batch at iterations I and $I-1$.

Knowledge Learning. While preserving good habits, we also need to calibrate bad habits formed during previous

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	54.92	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**	55.15	75.02	65.09
	HABIT (Ours)	**49.99**	72.38	**56.62**	**74.68**	**59.51**	78.53	**55.38**	**75.20**	**65.29**
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
	HABIT (Ours)	**49.63**	**71.34**	55.67	**73.19**	**58.14**	**78.32**	**54.48**	**74.28**	**64.38**
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
	HABIT (Ours)	**47.33**	**69.71**	**53.72**	**72.55**	**56.51**	**77.00**	**52.52**	**73.09**	**62.80**
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
	HABIT (Ours)	**42.04**	**65.20**	**50.12**	**69.77**	**52.92**	**73.61**	**48.36**	**69.53**	**58.94**

Table 1: Performance comparison on FashionIQ in terms of R@K (%). The best result under each noise ratio is highlighted in **bold**, while the second-best result is underlined.

training to mitigate the harm caused by noisy correspondence misdetermination. To this end, we introduce the soft estimation margin loss $\mathcal{L}_{soft}$ and the robust contrastive loss $\mathcal{L}_{rank}$, performing robust optimization on negative examples with higher similarity to the current sample, thereby jointly enhancing the model's discriminative capability and robustness to noise interference.

To address model misdetermination, we adopt a margin softening strategy based on cleanliness estimations. Let I denote the current iteration, and $\mathbf{s}^{(I)} \in \mathbb{R}^{B\times B}$ represent the similarity matrix between all composed features $\mathbf{F}_c$ and target images $\mathbf{F}_t$ in the batch. The noise mask for iteration I is denoted as $\mathbf{m}^{(I)}$. We apply this mask to identify noisy correspondences and employ a margin-based hard negative sampling strategy (Huang et al. 2021). The soft estimation margin loss is then defined as follows:

$$\mathcal{L}_{soft} = \frac{1}{B}\sum_{b=1}^{B}\left(\max_{j\neq b}\left[\text{margin}(\mathbf{e}_b^{(I)}) + \mathbf{s}_{bj}^{(I)} - \mathbf{s}_{bb}^{(I)}\right]_+ \cdot \mathbf{m}_b^{(I)}\right), \quad (7)$$

where $[\cdot]+ = \max(0,\cdot)$ denotes the ReLU operation, and $\mathbf{s}bb$ represents the similarity of the diagonal samples. $\text{margin}(\mathbf{e}_b^{(I)})$ is a dynamic margin that depends on the estimation value $\mathbf{e}_b^{(I)}$ of the b-th sample in the batch.

Subsequently, following (Li et al. 2025a; Hu et al. 2023a), we utilize the robust contrastive loss, which actively increases the distance between the multi-modal query and negative samples by reducing their similarity, and prevents interference from noisy correspondence, formulated as,

$$\mathcal{L}_{rank} = \frac{1}{B}\sum_{b=1}^{B} -\log\Big(\Big(1 - \text{Softmax}(\mathbf{s}_b^{(I)}/\tau)\Big)\cdot \mathbf{m}_b^{(I)}\Big), \quad (8)$$

where B is the batch size, τ is the temperature coefficient, $\mathbf{m}^{(I)}$ refers to the noisy mask of the I-th iteration batch, and $\mathbf{s}^{(I)} \in \mathbb{R}^{B\times B}$ is the similarity matrix between all composed features $\mathbf{F}_c$ and target images $\mathbf{F}_t$ in the I-th iteration batch.

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

$$\mathbf{\Theta}^* = \arg\min_{\mathbf{\Theta}} \left(\mathcal{L}_{rank} + \kappa\mathcal{L}_{KL} + \gamma\mathcal{L}_{soft}\right), \quad (9)$$

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

4 Experiments

This section provides an in-depth examination, with $\sigma = 0.2$ for ablation and sensitivity experiments, following TME.

4.1 Experimental Settings

Datasets. Following previous works (Li et al. 2025a; Xu et al. 2024), we selected two widely adopted datasets for the CIR task, the fashion-domain dataset FashionIQ (Wu et al. 2021), and the open-domain dataset CIRR (Liu et al. 2021b).

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**
	HABIT (Ours)	52.71	82.64	**90.63**	98.19	80.99	**92.77**	**97.00**	81.82
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**
	HABIT (Ours)	**51.68**	81.02	**89.24**	**97.81**	78.20	**91.66**	**96.75**	79.61
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
	HABIT (Ours)	**50.32**	**79.63**	**88.34**	**97.06**	**76.84**	**90.60**	**96.27**	**78.87**
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
	HABIT (Ours)	**47.93**	**76.84**	**85.95**	**95.90**	**74.87**	**89.08**	**95.21**	**75.86**

Table 2: 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 underlined, respectively.

Implementation Details. Our HABIT is trained with the learning rate of $5e$–5 with the AdamW optimizer on a V100 GPU. Following previous works (Li et al. 2025a; Xu et al. 2024), we utilize the pre-trained BLIP-2 (Li et al. 2023) as the backbone of HABIT. The learnable query number $Q = 32$. Regarding hyperparameter settings, we utilize the grid search to obtain the final value: $\kappa = 10.0, \gamma = 0.5$. The temperature coefficient $\tau = 0.1$. Following TME (Li et al. 2025a), we introduce noise ratio $\sigma = \{0.0, 0.2, 0.5, 0.8\}$ during training to simulate NTC environment.

Evaluation. We adopt Recall@K (R@K) as the primary metric. For CIRR, we report the performance of R@$\{1, 5, 10, 50\}$ and further provide $\text{Recall}_{\text{sub}}$@$\{1, 2, 3\}$ metrics on its subset. For FashionIQ, we report the R@$\{10, 50\}$ for each category (*Dresses*, *Shirts*, *Tops&Tees*).

4.2 Performance Comparison

To assess the robustness and generalization of HABIT under the NTC scenario, we compare it with selected baselines: ordinary baselines (SSN, CALA, and SPRC) and robust baselines (RCL, RDE, and TME), on the CIRR and FashionIQ datasets across different noise ratios. As shown in Table 1,2, our analysis reveals the following: **1)** HABIT outperforms existing robust methods in addressing NTC, demonstrating clear advantages in complex noise environments. On CIRR, HABIT achieves average improvements over TME of 1.16%, and 1.28% for $\sigma = 0.5$, and 0.8, respectively. On FashionIQ, gains are 0.94%, and 1.31%. Notably, the performance gap with TME widens as noise increases. These gains result from HABIT's effective mining of mutual knowledge between multimodal queries and targets, precise noise estimation, and the integration of dual-consistency progressive learning. **2)** Robust methods consistently surpass ordinary methods, with the performance gap widening as the noise ratio increases. For example, on the CIRR dataset at $\sigma = 0.2$, the SOTA ordinary method SPRC even outperforms some robust models on certain metrics, and its Avg score is only 2.63% below HABIT. However, at $\sigma = 0.8$, SPRC's Avg score drops to 15.12% lower than HABIT, reflecting severe performance decline. This highlights the high sensitivity of traditional methods to noise and their instability in noisy correspondence scenarios, while robust models demonstrate superior noise resistance, emphasizing their necessity.

4.3 Ablation Study

To evaluate the efficacy of each HABIT module, we conduct comprehensive comparisons with variants in two groups:

G1: Ablation on Mutual Knowledge Estimation. This group examines the MKE module: **D#(1): w/o_Sample** randomly selects samples instead of using standard samples Eq.(3) for mutual knowledge calculation. **D#(2): w/o_TR** replaces the Transition Rate Eq.(4) with the difference

in mutual knowledge estimations. **D#(3): w/o_MKE** removes the entire MKE process, setting all estimation values to 1. ***G2: Dual-consistency Progressive Learning.*** This group assesses DPL components: **D#(4): w/o_CS** uses only current iteration estimations, omitting historical values in Eq.(5). **D#(5): w/o_KL** removes the consistency constraint. **D#(6): w/o_History** excludes both the consistency constraint and historical estimations. **D#(7): w/o_mask** bypasses noisy mask computation, using only MKE estimations. **D#(8): w/o_M_Rank**, **D#(9): w/o_M_Soft**, **D#(10): w/o_M_KL** remove the noisy mask in Eq.(8), Eq.(7), and Eq.(6). **D#(11): w/o_$\mathcal{L}_{rank}$** omits the robust contrastive loss. **D#(12): w/o_$\mathcal{L}_{soft}$** omits the soft estimation margin loss. **D#(13): w/o_$\mathcal{L}_{KL}$&$\mathcal{L}_{soft}$** removes both consistency constraint and soft estimation margin loss.

D#	Deriv.	FashionIQ-Avg.		CIRR-Avg.	
		R@10	R@50	R@K	R_{sub}@K
	(a) Mutual Knowledge Estimation (MKE)				
1	w/o Sample	53.77	73.01	79.42	88.12
2	w/o TR	53.40	73.03	79.58	87.71
3	w/o MKE	53.08	73.12	79.32	87.94
	(b) Dual-consistency Progressive Learning (DPL)				
4	w/o CS	53.30	73.22	79.48	88.00
5	w/o KL	53.44	73.92	79.27	87.28
6	w/o History	53.66	73.87	78.32	87.24
7	w/o mask	53.59	73.67	79.46	87.72
8	w/o M_Rank	53.48	73.58	79.67	87.98
9	w/o M_Soft	53.51	73.65	79.46	88.00
10	w/o M_KL	53.73	73.47	79.71	88.03
11	w/o $\mathcal{L}_{rank}$	51.03	71.82	77.92	87.15
12	w/o $\mathcal{L}_{soft}$	53.14	73.34	79.38	88.21
13	w/o $\mathcal{L}_{KL}$&$\mathcal{L}_{soft}$	53.27	73.69	79.51	88.20
HABIT (Ours)		**54.48**	**74.28**	**79.94**	**88.87**

Table 3: Ablation study on FashionIQ and CIRR datasets.

Key findings from Table 3 include: **1)** Removing $\mathcal{L}_{rank}$ **(D#(11))** yields the worst performance, underscoring the essential role of robust contrastive loss in noise suppression and retrieval accuracy. **2)** Omitting any component of the **MKE** module leads to performance degradation. Notably, eliminating standard samples for mutual knowledge **(D#(1))** degrades performance, though less so than removing all MKE estimations **(D#(3))**. Replacing the transition rate with mutual knowledge discrepancy **(D#(2))** results in the largest drop, highlighting the transition rate's importance for capturing intrinsic NTC relationships. **3)** Removing either the historical estimation sequence **(D#(4))** or the "calibrate bad habits" constraint **(D#(5))** reduces performance; removing both **(D#(6))** causes even greater decline. These help retain good habits and noise discrimination, preventing semantic inconsistency over time-flux training. **4)** Excluding the noisy mask **(D#(7))** decreases performance, validating its role in filtering noisy triplets. Further, removing the noisy mask from any loss function **(D#(8)-D#(10))** results in performance drops, reflecting its broad optimization effect via knowledge aggregation. **5)** Excluding $\mathcal{L}_{soft}$ **(D#(11), D#(12))** impairs performance, indicating that the soft estimation margin loss reduces overfitting from partial matches and misclassified noisy samples.

4.4 Case Study

Figure 3 presents the top-5 results from HABIT and the SOTA robust CIR model TME on two CIR datasets. In the CIRR example (Figure 3(a)), HABIT correctly retrieves the top-ranked image featuring both a diver and a sea turtle, fully satisfying the compositional semantics. By contrast, TME fails to interpret the cross-entity relationship, retrieving only manta ray images and missing the "human+different species" requirement, revealing its limited grasp of fine-grained semantic composition. In Figure 3(b), HABIT's top-1 image matches all target attributes (e.g., reddish-brown long dress, warm autumn tones) at top-1, with subsequent results also aligning well with key attributes. In comparison, TME's top-1 result, although color-similar, lacks thin straps. Overall, HABIT reveals clear advantages in handling multi-level semantic compositions, which stem from its meticulous understanding of triplet semantic relationships.

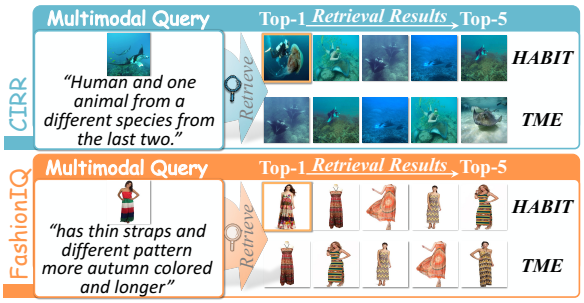

Figure 3: Case Study on CIRR and FashionIQ.

5 Conclusion

In this study, we investigated the NTC problem in the CIR task. To address the challenges of the precise estimation of composed semantic discrepancy and the insufficient progressive adaptation to modification discrepancy, we proposed HABIT, which comprised two key modules. The MKE module, employed the transition rate of variational mutual information to achieve accurate noise-aware label assignment. Furthermore, the DPL module introduced a collaborative mechanism between the historical and current models, progressively enhancing the model's understanding of the complex semantic relations within triplets and continuously reduced the probability of misdetermination. Extensive experiments showed that HABIT outperformed most methods under various noise levels, demonstrating its superiority and robustness.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China, No.:62276155, No.:62576195, and No.:62572282; in part by the China National University Student Innovation & Entrepreneurship Development Program, No.:2025282 and No.:2025283.

References

Bellavia, F.; Zhao, Z.; Morelli, L.; and Remondino, F. 2024. Image Matching Filtering and Refinement by Planes and Beyond. *arXiv preprint arXiv:2411.09484*.

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(9): 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.

Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.; et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, 226–231.

Feng, P.; and Peng, X. 2014. A note on Monge–Kantorovich problem. *Statistics & Probability Letters*, 84: 204–211.

Feng, Y.; Zhu, H.; Peng, D.; Peng, X.; and Hu, P. 2023. ROAD: Robust unsupervised domain adaptation with noisy labels. In *ACM MM*, 7264–7273.

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.

Guo, X.; Wu, H.; Cheng, Y.; Rennie, S.; Tesauro, G.; and Feris, R. S. 2018. Dialog-based Interactive Image Retrieval. In *NeurIPS*, 676–686. MIT Press.

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, 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. 2024a. 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. 2024b. 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.

Huang, Z.; Qian, H.; Cai, Z.; Wang, X.; Xie, L.; and Niu, X. 2025d. An intelligent multilane roadway recognition method based on pseudo-tagging. *CaGIS*, 1–16.

Huang, Z.; Yang, M.; Xiao, X.; Hu, P.; and Peng, X. 2024c. Noise-robust vision-language pre-training with positive-negative learning. *IEEE TPAMI*.

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*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. 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, 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. *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, J.; Shang, F.; Zhu, K.; Liu, H.; Liu, Y.; and Liu, J. 2025b. FedAdamW: A Communication-Efficient Optimizer with Convergence and Generalization Guarantees for Federated Large Models. *arXiv preprint arXiv:2510.27486*.

Liu, J.; Tian, Y.; Shang, F.; Liu, Y.; Liu, H.; Zhou, J.; and Ding, D. 2025c. 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. 2024. 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 *ACM 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. 2025d. Training-Free Multi-Style Fusion Through Reference-Based Adaptive Modulation. arXiv:2509.18602.

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.; Lian, Z.; Zhou, Z.; Zhang, S.; Zhao, C.; and Kong, A. W.-K. 2025. Does FLUX Already Know How to Perform Physically Plausible Image Composition? *arXiv preprint arXiv:2509.21278*.

Lu, S.; Wang, Z.; Li, L.; Liu, Y.; and Kong, A. W.-K. 2024a. Mace: Mass concept erasure in diffusion models. In *CVPR*, 6430–6440.

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

Lyu, S.; Tian, Z.; Ou, Z.; Zhu, Y.; Zhang, X.; Ha, Q.; Luo, H.; and Song, M. 2025. TSVC: Tripartite Learning with Semantic Variation Consistency for Robust Image-Text Retrieval. In *AAAI*, volume 39, 19269–19277.

Mu, C.; Yang, E.; and Deng, C. 2025. Meta-Guided Adaptive Weight Learner for Noisy Correspondence. In *ACM SIGIR*, 968–978.

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.; Peng, D.; Dai, J.; and Ren, Z. 2023a. Stepwise refinement short hashing for image retrieval. In *ACM MM*, 6501–6509.

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

Sunmola, I. O.; Zhao, Z.; Schmidgall, S.; Wang, Y.; Scheikl, P. M.; and Krieger, A. 2025. Surgical Gaussian Surfels: Highly Accurate Real-time Surgical Scene Rendering. *arXiv preprint arXiv:2503.04079*.

Tang, H.; Zhu, J.; Liu, M.; Gao, Z.; and Cheng, Z. 2021. Frame-wise cross-modal matching for video moment retrieval. *IEEE TMM*, 24: 1338–1349.

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.

Ventura, L.; Yang, A.; Schmid, C.; and Varol, G. 2024. CoVR-2: Automatic Data Construction for Composed Video Retrieval. *IEEE TPAMI*.

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, Q.; He, Y.; Wang, J.; Song, X.; Qian, S.; Yuan, X.; Sun, L.; Xin, Y.; Tang, J.; Li, K.; et al. 2025. Score: Story coherence and retrieval enhancement for ai narratives. *arXiv preprint arXiv:2503.23512*.

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.

Zha, Q.; Liu, X.; Cheung, Y.-m.; Peng, S.-J.; Xu, X.; and Wang, N. 2025. UCPM: Uncertainty-Guided Cross-Modal Retrieval with Partially Mismatched Pairs. *IEEE TIP*.

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. 2025a. Spatial Understanding from Videos: Structured Prompts Meet Simulation Data. In *NeurIPS*, 1–16.

Zhang, H.; Xu, H.; Liu, H.; Yu, X.; Zhang, X.; and Wu, C. 2025b. Conditional variational underwater image enhancement with kernel decomposition and adaptive hybrid normalization. *Neurocomputing*, 130845.

Zhang, X.; Liu, H.; Xue, L.; Li, X.; Guo, W.; Yu, S.; Ru, J.; and Xu, H. 2021. Multi-objective collaborative optimization algorithm for heterogeneous cooperative tasks based on conflict resolution. In *ICAUS*, 2548–2557. Springer.

Zhou, Z.; Zhang, J.; Zhang, J.; He, Y.; Wang, B.; Shi, T.; and Khamis, A. 2024. Human-centric reward optimization for reinforcement learning-based automated driving using large language models. *arXiv preprint arXiv:2405.04135*.