Boosting Noisy Correspondence Discrimination via Dynamic Neighborhood Semantic Verification

Yu Wang[1]*, **Fengxia Han**[1] **and Jianyu Wang**[1]

[1]School of Computer Science and Technology, Tongji University
csyuwang@tongji.edu.cn

Abstract

Noisy correspondence, characterized by mismatches in cross-modal data pairs, presents a significant challenge for real-world applications. Current approaches primarily rely on direct cross-modal pairwise similarity metrics, which suffer from two critical limitations: noise sensitivity, where direct similarity calculations are easily corrupted by noisy or ambiguous instances, and contextual blindness, where isolated pairwise comparisons fail to exploit the rich semantic context embedded in neighboring instances. To address this issue, we propose to improve noise correspondence discrimination through a well-designed **D**ynamic **N**eighborhood **S**emantic association verification paradigm, namely ***DNS***. Specifically, we hypothesize that the matching degree of current samples can be quantified through the interrelationships among their respective semantic neighbors. For this reason, we develop a novel semantic drift distance and local relation proximity based on dynamic neighborhood association. Furthermore, beyond implicit approaches to semantic gap modeling in cross-modal data, we introduce an explicit decomposition framework that disentangles the gap into the semantic orientation and scalar magnitude. Through the strategic integration of these proposed mechanisms, ***DNS*** achieves substantial enhancement in noisy correspondence discrimination, yielding remarkable performance gains. Extensive experiments on three widely-used benchmark datasets, including Flickr30K, MS-COCO, and Conceptual Captions, demonstrate the superiority of ***DNS*** over state-of-the-art methods.

Introduction

Cross-modal retrieval, which aims to retrieve semantically relevant content across different modalities, has garnered significant research attention (Shen et al. 2025; Kim et al. 2023; Fu et al. 2023; Pan, Wu, and Zhang 2023; Wang, Zhao, and Chen 2024). Most existing methods depend heavily on high-quality annotations, yet the labor-intensive data collection process and unreliable non-expert annotations often lead to the inadvertent introduction of semantically irrelevant samples, termed noisy correspondence (NC). This challenge frequently results in degraded retrieval performance.

To mitigate the impact of NC, early approaches (Huang et al. 2021; Yang et al. 2023; Han et al. 2023) leverage the *memory effect* (Han et al. 2018), wherein deep neural networks (DNNs) tend to prioritize learning simple, low-level patterns during the initial training phase. By utilizing this characteristic, they estimate soft correspondence scores to quantify the degree of mismatch between cross-modal samples. This enables adaptive instance weighting during training, down-weighting noisy pairs while emphasizing cleaner local correspondences, thereby guiding the model to focus on robust feature alignment. Some studies (Pan, Wu, and Zhang 2023; Feng et al. 2023; Ma et al. 2024) have further proposed fine-grained partitioning schemes to actively exclude noisy pairs. To further mitigate the adverse impact of mismatches, some works (Han et al. 2024; Duan et al. 2024) have leveraged pseudo samples of these mismatches to uncover meaningful correspondences. Additionally, recent works (Zha et al. 2025; Zhao et al. 2024) have also explored the intrinsic structure and relationships among data. However, these methods determine text-image alignment through direct similarity computation between paired features, suffering from two critical limitations: 1) noise sensitivity, where noisy or ambiguous instances easily corrupt similarity calculations, and 2) contextual blindness, where isolated pairwise comparisons overlook the rich semantic context embedded in neighboring instances.

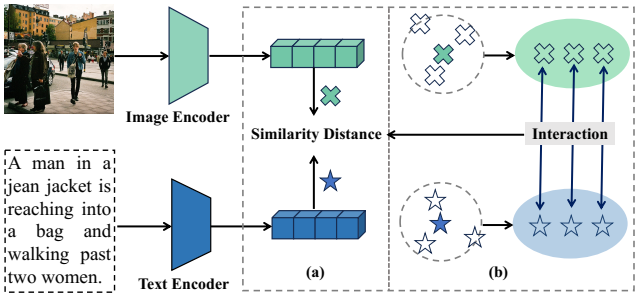

Figure 1: Illustration of our motivation. Existing methods rely on direct cross-modal pairwise similarity metrics, as shown in (a), which are prone to noise sensitivity (direct similarity calculations are easily corrupted by noise) and contextual blindness (isolated comparisons ignore the semantic context in neighboring instances). Thus, we propose leveraging neighboring instances to provide complementary signals for disambiguating noisy pairs, as shown in (b).

*Corresponding author.

In this paper, we argue that the matching degree of current pairs can be quantified through the interrelationships among their respective semantic neighbors. To this end, we propose to improve noise correspondence discrimination through a novel **D**ynamic **N**eighborhood **S**emantic association verification paradigm, namely ***DNS***. In this paradigm, the semantic neighbors of samples are expected to provide complementary signals for disambiguating noisy pairs. Specifically, leveraging the neighbors of data from each modality, we introduce two novel mechanisms: a semantic drift distance to quantify cross-modal semantic gap, and a local relation proximity to capture structure similarity. The core strength of neighborhood-based association verification lies in its capacity to leverage local contextual information for robust cross-modal alignment verification. Besides, to better measure noisy correspondence between pairs, we further propose a novel local relation proximity strategy to evaluate the semantic distance of different modalities by explicitly disentangling the gap into the semantic orientation and scalar magnitude.

Overall, the main contributions of our work are: (1) We hypothesize that the matching degree of pairs can be quantified through interrelationships among their semantic neighbors, and thus propose a novel **D**ynamic **N**eighborhood **S**emantic association verification paradigm to enhance noise correspondence discrimination. (2) We propose a novel semantic drift distance and local relation proximity mechanism to leverage local contextual information for robust cross-modal alignment. (3) Beyond implicit approaches to semantic gap modeling in cross-modal data, we develop an explicit decomposition framework that disentangles the gap into the semantic orientation and scalar magnitude. (4) Our method achieves state-of-the-art performance on three popular benchmarks, particularly under high noisy rates.

Related Work

Cross-Modal Retrieval

Cross-modal retrieval (CMR) (Shen et al. 2025; Kim et al. 2023; Fu et al. 2023; Pan, Wu, and Zhang 2023) concentrates on utilizing information from one modality to retrieve the most relevant content in other modalities. The core of CMR is to bridge semantic gaps by aligning heterogeneous modalities within a unified semantic space. This mapping framework enforces that semantically correlated instances exhibit higher feature similarity, while simultaneously amplifying dissimilarity between unrelated pairs. The mainstream paradigm of CMR consists of two categories: coarse-grained representation alignment (Li et al. 2022; Chen et al. 2021) and fine-grained representation alignment (Diao et al. 2021; Fu et al. 2023; Pan, Wu, and Zhang 2023; Lee et al. 2018; Pham et al. 2024; Wang and Chen 2025). The former prioritizes computational efficiency through a holistic alignment strategy that systematically correlates cross-modal features extracted by dedicated modality-specific encoders. The latter focuses on fine-grained semantics through dynamic learning-to-reason frameworks, enabling the discovery of latent cross-modal fragment alignments. Unfortunately, the promising performance of all these methods is constrained by their reliance on paired training data, which inherently overlooks the prevalence of NC arising from human-annotated or web-crawled datasets. Such NC inherently introduces spurious cross-modal associations during training, which compromise feature distribution alignment and ultimately result in degraded performance.

Noisy Correspondence Learning

Noisy correspondence (NC) learning encompasses robustness-enhancing methodologies specifically engineered to suppress performance degradation caused by mismatched cross-modal pairs. This concept is first introduced in (Huang et al. 2021). To address this problem, early attempts (Huang et al. 2021; Han et al. 2023; Yang et al. 2023) utilize DivideMix (Li, Socher, and Hoi 2020) to isolate clean correspondence pairs from noisy counterparts, based on the memorization effect of DNNs. Following these developments, subsequent studies have proposed sophisticated criteria for refined data partitioning, including prediction inconsistency metrics (Feng et al. 2023; Ma et al. 2024) and uncertainty quantification strategies (Zha et al. 2024). Some methods also further refine alignment through re-pairing of mismatched instance pairs (Han et al. 2024) and the integration of pseudo-caption (Duan et al. 2024). In addition, some endeavors have increasingly adopted robust loss functions that adaptively attenuate the gradient contributions of mismatched samples through dynamic weighting mechanisms, *e.g.*, evidential loss (Qin et al. 2022) and complementary loss (Qin et al. 2023; Hu et al. 2023). Recent advancements leverage intrinsic geometric constraints (Zhao et al. 2024; Yang et al. 2024) and relational invariance principles (Zha et al. 2025) to generate probabilistic correspondence labels within data-driven alignment frameworks.

Methodology

Preliminary

Consider a multi-modal dataset $\mathcal{D} = \{V_i, L_i, M_i\}$, where each sample comprises an image-text pair $\{V_i, L_i\}$ annotated with corresponding label y_i indicating whether the pair is positively correlated ($M_i = 1$) or not ($M_i = 0$). Visual and textual inputs are first independently encoded into a shared latent space to obtain their representations I and Q, using modality-specific encoders for images and texts. Then, the semantic correspondence between these heterogeneous embeddings is quantified through a similarity metric $S(I, Q)$, which measures their alignment in the joint embedding space. However, there exists a non-negligible portion of mismatched pairs (*i.e.*, NCs) in datasets, resulting in notable performance degradation. The goal of our method is to address the challenge of NCs.

Ideally, we expect the matched pairs to have higher similarity while mismatched pairs should have lower similarity. This is typically achieved by minimizing the InfoNCE loss (Oord, Li, and Vinyals 2018). The success of this process relies on well-matched pairs. However, a non-negligible portion of mismatched pairs (*i.e.*, NCs) exist within each

training batch, which can significantly undermine the learning process of cross-modal representation alignment. Leveraging the memorization effect inherent to deep neural networks, where models first capture clean structural patterns before overfitting to noisy instances, we mitigate the adverse effects of NC by learning an accurate correspondence indicator y^i. The purified cross-modal loss can be denoted as follows:

$$\mathcal{L}_{base}(I_i, Q_i) = -\frac{1}{2N}\sum_{i=1}^{N} y^i \frac{exp(< I_i, Q_i > /\tau)}{\sum_{j=1}^{N} exp(< I_i, Q_j > /\tau)} - \frac{1}{2N}\sum_{j=1}^{N} y^j \frac{exp(< I_i, Q_i > /\tau)}{\sum_{j=1}^{N} exp(< I_j, Q_i > /\tau)}, \tag{1}$$

where N is the batch size and τ represents the temperature.

Overall Framework

We present the proposed DNS framework to learn correspondence y^i for noise purification, as shown in Figure 2. DNS first extracts image and text representations through separate pre-trained encoders. Based on the well-established cross-modal structure in the early stage, we adopt a cross-modal bidirectional contrastive metric as the basic correspondence indicator:

$$y^i_{CM} = \frac{1}{2}[\frac{exp(< I_i, Q_i > /\tau)}{\sum_{j=1}^{N} exp(< I_i, Q_j > /\tau)} + \frac{exp(< I_i, Q_i > /\tau)}{\sum_{j=1}^{N} exp(< I_j, Q_i > /\tau)}], \tag{2}$$

where τ is the temperature coefficient. This metric quantifies the relative similarity between the current data pairs I_i and Q_i against all negative samples within the batch context. For clean positive pairs, the numerator term, representing the similarity between the current I_i and Q_i, dominates the denominator summation due to semantic alignment, yielding an indicator value approaching 1. Conversely, noisy pairs exhibit lower similarity across the batch, causing the value to approach 0.

To further mitigate the adverse effects of NCs, we propose a novel dynamic neighborhood semantic verification framework to enhance discriminative capacity for noisy correspondences. This framework consists of three core components: local relation proximity (LRP), semantic drift distance (SDD), and semantic gap decomposition (SGD). Specifically, we first select the top-K semantically most similar text samples for the current text Q_i within a batch, denoted as $Q_i^{nbr} = \{Q_{i,0}^{nbr}, Q_{i,1}^{nbr}, ..., Q_{i,K-1}^{nbr}\}$. Similarly, we identify the top-K most similar neighboring images for the current image, represented as $I_i^{nbr} = \{I_{i,0}^{nbr}, I_{i,1}^{nbr}, ..., I_{i,K-1}^{nbr}\}$. Based on these semantic neighbors, we explore local relation proximity and semantic drift distance for robust semantic consistency, yielding the correspondence indicator y^i_{LRP} and y^i_{SDD} respectively. Besides, we investigate the semantic gap between different modalities by explicitly decomposing the gap into the semantic orientation and scalar magnitude, yielding the correspondence indicator y^i_{SGD}. Then, we get a contrastive indicator $y^i_C = (y^i_{CM} + y^i_{LRP})/2.$, and a gap indicator $y^i_G = y^i_{SDD} \times y^i_{SGD}$. To address the challenges arising from different aspects and accurately identify all samples with noisy correspondence, we define the final correspondence indicator for each sample as the minimum of y^i_C and y^i_G, *i.e.*,

$$y^i = min\{y^i_C, y^i_G\}. \tag{3}$$

Next, we will elaborate on the proposed three components.

Local Relation Proximity

Since the pair $\{I_i, Q_i\}$ and their respective top-K neighbors are semantically relevant, we investigate their local relation proximity to prevent the impact of noisy correspondences. Specifically, we establish the correlation relationships between the current text and the video features corresponding to its semantic neighbors, as well as between the current video and the text features corresponding to its semantic neighbors. Taking the image feature I_i as an example, we first identify its top-K semantic neighbors $\{I_{i,0}^{nbr}, I_{i,1}^{nbr}, ..., I_{i,K-1}^{nbr}\}$ in the feature space, and utilize their corresponding visual features $\{Q_{i,0}^{nbr}, Q_{i,1}^{nbr}, ..., Q_{i,K-1}^{nbr}\}$ as proxies to calculate the similarity with I_i, getting their similarity scores S_i^I. Similarly, we can get similarity scores S_i^Q that describe the local semantic relation for texts. In this manner, S_i^I and S_i^Q leverage broader semantic context embedded in neighboring instances to reflect structured semantic cues, which are more robust in expressing themselves. Last, we measure the affinity between normalized S_i^I and S_i^Q as the correspondence indicator:

$$y^i_{LRP} = norm(S_i^Q) \cdot norm((S_i^I)^T), \tag{4}$$

where the symbol $norm$ denotes ℓ_2 normalization operation.

Semantic Drift Distance

Unlike previous methods using isolated pairwise comparisons, here we utilize the broader semantic context embedded in neighboring instances, which could provide complementary signals for disambiguating noisy pairs. Specifically, given the current pair $\{I_i, Q_i\}$ and their respective semantic neighbors Q_i^{nbr} and I_i^{nbr}, we define a bidirectional semantic drift distance y^i_{SDD} as correspondence indicator for each pair as:

$$\begin{aligned} d^{t\to v}_{i,i} &= ||Q_i - I_i||_2^2 - \frac{1}{K}\sum_{k=1}^{K} ||Q_i - I_{i,k}^{nbr}||_2^2, \\ d^{v\to t}_{i,i} &= ||I_i - Q_i||_2^2 - \frac{1}{K}\sum_{k=1}^{K} ||I_i - Q_{i,k}^{nbr}||_2^2, \\ y^{i,i}_{SDD} &= CLAMP(\frac{1}{2}((1 - d^{t\to v}_{i,i}) + (1 - d^{v\to t}_{i,i}))), \end{aligned} \tag{5}$$

where $CLAMP$ denotes a truncation operation that ensures the value of $y^{i,i}_{SDD}$ is between 0 and 1. This distance leverages neighboring samples that are semantically related to the current pair to measure whether the given pair exhibits semantic drift. When $d_{t\to v}$ and $d_{v\to t}$ are smaller, the target

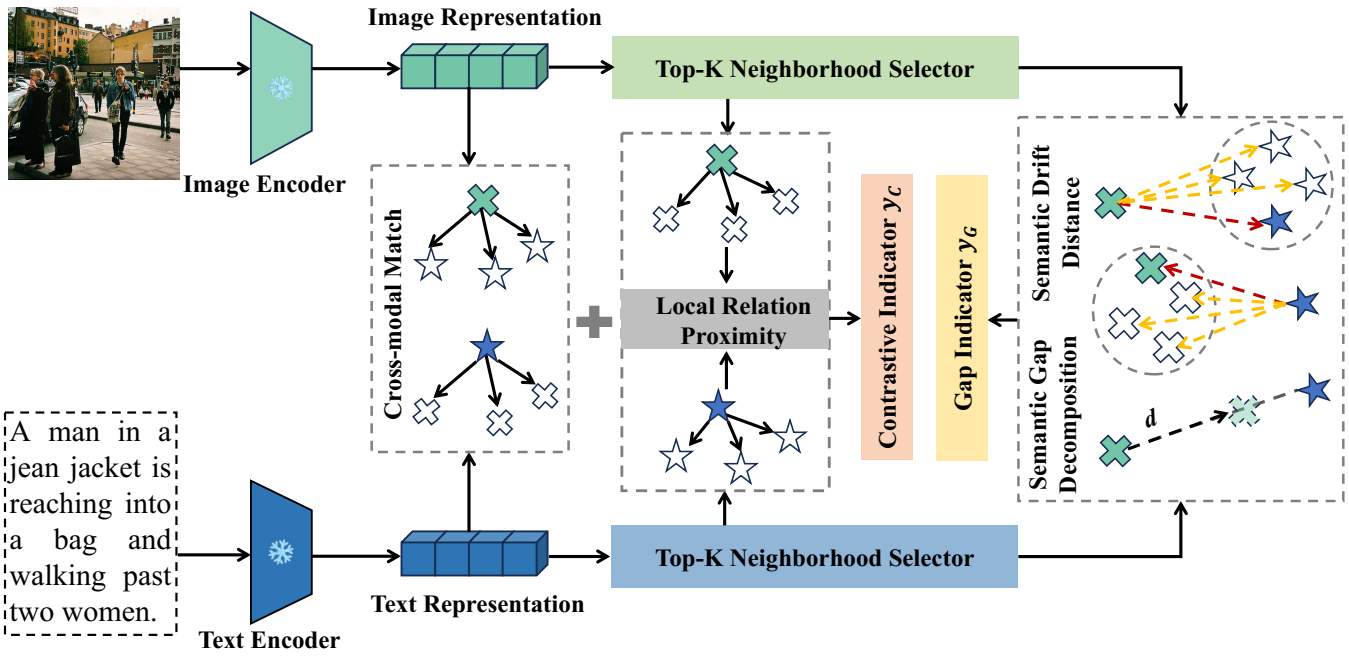

Figure 2: An Overview of DNS. DNS identifies noisy samples via a dynamic neighborhood semantic verification paradigm, which consists of three core components: local relation proximity, semantic drift distance, and semantic gap decomposition.

pair is deemed semantically aligned. Conversely, the pair is identified as exhibiting semantic misalignment. Besides, we also apply contrastive loss to encourage smaller distances between matched samples:

$$\mathcal{L}_{sdd} = -\frac{1}{2N}\sum_{i=1}^{N} y^i \frac{y_{SDD}^{i,i}/\tau}{\sum_{j=1}^{N} exp(y_{SDD}^{i,j}/\tau)}. \tag{6}$$

Semantic Gap Decomposition

Previous methods have discriminated against noisy correspondences by employing implicit strategies, such as similarity-based metrics, to filter out erroneous alignments. They depend heavily on prior knowledge enriched in the pre-trained backbones. In this section, we propose to model explicitly the semantic gap R_i between different modalities, where the gap is disentangled into semantic orientation and scalar magnitude. Since visual features typically contain richer information than textual features, we shift textual features by a distance d_i along the text-to-vision direction $\frac{I_i - Q_i}{||I_i - Q_i||_2^2}$, enabling the adjusted textual features to better align with visual features. Specifically, we define the semantic gap R_i between I_i and Q_i as:

$$\begin{aligned} R_i &= d_i \times \frac{I_i - Q_i}{||I_i - Q_i||_2^2}, \\ d_i &= \exp^{fc(norm(I_i)\cdot norm(Q_i^T))}. \end{aligned} \tag{7}$$

Here, d_i is a learnable scalar parameterized by the similarity between I_i and Q_i, which dynamically regulates the magnitude of the gap. fc denotes a linear transformation using the fully-connected layer. In this way, we can get the adjusted textual features by $Q_i^{adj} = Q_i + R_i$. Meanwhile, we define the noisy correspondence indicator y_{sgd}^i as the degree of the gap between text and image:

$$y_{sgd}^i = 1 - \frac{|d_i| - d_{min}}{d_{max} - d_{min}}, \tag{8}$$

where $|\cdot|$ is the magnitude of d_i, and d_{min} and d_{max} are the minimum and maximum magnitudes in the training batch. Notably, to prevent the learnable parameter d_i from degenerating into trivial solutions, we impose constraints on the adjusted features using contrastive loss:

$$\mathcal{L}_{sgd}(I_i, Q_i^{adj}) = -\frac{1}{2N}\sum_{i=1}^{N} y^i \frac{exp(< I_i, Q_i^{adj} >)}{\sum_{j=1}^{N} exp(< I_i, Q_j^{adj} >)}. \tag{9}$$

Training Details

We optimize the network using the overall loss function as:

$$\mathcal{L}_{all} = \mathcal{L}_{base} + \lambda_1 \mathcal{L}_{sdd} + \lambda_2 \mathcal{L}_{sgd}, \tag{10}$$

where λ_1 and λ_2 are used to balance different terms. Besides, we employ the temporal ensembling strategy, as adopted in (Liu et al. 2020; Zhao et al. 2024), to iteratively refine the estimated correspondence indicators y_C and y_G. Specifically, both indicators are dynamically updated through an exponential moving average mechanism, which integrates values from the current epoch t with those from the previous epoch $t-1$. The detailed process is:

$$\begin{aligned} y_C^i(t) &= \beta_1 y_C^i(t) + (1-\beta_1) y_C^i(t-1), \\ y_G^i(t) &= \beta_2 y_G^i(t) + (1-\beta_2) y_G^i(t-1), \end{aligned} \tag{11}$$

where β_1 and β_2 are the momentum coefficient. This strategy not only stabilizes the training process but also mitigates noise in single-epoch estimates, enabling smoother convergence of the correspondence indicators.

Noise Ratio	Methods	Flickr30K							MS-COCO 1K						
		Image to Text			Text to Image				Image to Text			Text to Image			
		R@1	R@5	R@10	R@1	R@5	R@10	rSum	R@1	R@5	R@10	R@1	R@5	R@10	rSum
20%	NCR (NeurIPS′21)	73.5	93.2	96.6	56.9	82.4	88.5	491.1	76.6	95.6	98.2	60.8	88.8	95.0	515.0
	DECL (ACM MM′22)	77.5	93.8	97.0	56.1	81.8	88.5	494.7	77.5	95.9	98.4	61.7	89.3	95.4	518.2
	MSCN (CVPR′23)	77.4	94.9	97.6	59.6	83.2	89.2	501.9	78.1	97.2	98.8	64.3	90.4	95.8	524.6
	BiCro (CVPR′23)	78.1	94.4	97.5	60.4	84.4	89.9	504.7	78.8	96.1	98.6	63.7	90.3	95.7	523.2
	RCL (TPAMI′23)	75.9	94.5	97.3	57.9	82.6	88.6	496.8	78.9	96.0	98.4	62.8	89.9	95.4	521.4
	CRCL (NeurIPS′23)	78.9	94.8	**97.9**	58.7	83.0	89.2	502.5	77.8	96.1	98.5	63.4	90.3	95.9	522.0
	SREM (AAAI′24)	79.5	94.2	**97.9**	61.2	84.8	90.2	507.8	78.5	96.8	98.8	63.8	90.4	95.8	524.1
	PC^2 (ACM MM′24)	78.7	94.9	96.9	59.8	83.9	89.6	503.8	77.8	95.7	98.4	62.8	89.7	95.3	519.7
	L2RM (CVPR′24)	77.9	95.2	97.8	59.8	83.6	89.5	503.8	80.2	96.3	98.5	64.2	90.1	95.4	524.7
	ESC (CVPR′24)	79.0	94.8	97.5	59.1	83.8	89.1	503.3	79.2	97.0	99.1	64.8	90.7	96.0	526.8
	GSC (CVPR′24)	78.3	94.6	97.8	60.1	84.5	90.5	505.8	79.5	96.4	98.9	64.4	90.6	95.9	525.7
	ReCon (CVPR′25)	80.3	95.3	97.8	61.6	85.5	91.3	511.8	80.9	96.6	98.8	**65.2**	91.0	96.0	528.6
	DNS (Ours)	**81.4**	**96.3**	97.8	**62.0**	**86.1**	**91.8**	**515.4**	**81.4**	**97.6**	**99.6**	64.4	**91.3**	**96.5**	**530.8**
40%	NCR (NeurIPS′21)	75.3	92.1	95.2	56.2	80.6	87.4	486.8	76.5	95.0	98.2	60.7	88.5	95.0	513.9
	DECL (ACM MM′22)	72.7	92.3	95.4	53.4	79.4	86.4	479.6	75.6	95.5	98.3	59.5	88.3	94.8	512.0
	MSCN (CVPR′23)	74.4	94.4	96.9	57.2	81.7	87.6	492.2	74.8	94.9	98.0	60.3	88.5	94.4	510.9
	BiCro (CVPR′23)	74.6	92.7	96.2	55.5	81.1	87.4	487.5	77.0	95.9	98.3	61.8	89.2	94.9	517.1
	RCL (TPAMI′23)	72.7	92.7	96.1	54.8	80.0	87.1	483.4	77.0	95.5	98.3	61.2	88.5	94.8	515.3
	CRCL (NeurIPS′23)	74.1	92.6	96.9	55.5	80.9	87.6	487.6	76.6	95.6	98.5	62.3	89.7	95.4	518.1
	SREM (AAAI′24)	76.5	93.9	96.3	57.5	82.7	88.5	495.4	77.2	96.0	98.5	62.1	89.3	95.3	518.4
	PC^2 (ACM MM′24)	75.8	93.5	96.9	57.5	81.9	88.2	493.8	77.4	95.8	98.4	62.1	89.4	95.1	518.2
	L2RM (CVPR′24)	75.8	93.2	96.9	56.3	81.0	87.3	490.5	77.5	95.8	98.4	62.0	89.1	94.9	517.7
	ESC (CVPR′24)	76.1	93.1	96.4	56.0	80.8	87.2	489.6	78.6	96.6	99.0	63.2	90.6	95.9	523.9
	GSC (CVPR′24)	76.5	94.1	97.6	57.5	82.7	88.9	497.3	78.2	95.9	98.2	62.5	89.7	95.4	519.9
	ReCon (CVPR′25)	79.4	94.3	97.6	59.9	83.9	90.1	505.2	79.9	96.2	98.6	63.5	90.5	95.9	524.5
	DNS (Ours)	**80.0**	**94.9**	**98.1**	**60.3**	**84.7**	**91.2**	**509.2**	**80.2**	**96.6**	**99.1**	**63.9**	**90.7**	**96.1**	**526.6**
60%	NCR (NeurIPS′21)	68.7	89.9	95.5	52.0	77.6	84.9	468.6	72.7	94.0	97.6	57.9	87.0	94.1	503.3
	DECL (ACM MM′22)	65.2	88.4	94.0	46.8	74.0	82.2	450.6	73.0	94.2	97.9	57.0	86.6	93.8	502.5
	MSCN (CVPR′23)	70.4	91.0	94.9	53.4	77.8	84.1	471.6	74.4	95.1	97.9	59.2	87.1	92.8	506.5
	BiCro (CVPR′23)	67.6	90.8	94.4	51.2	77.6	84.7	466.3	73.9	94.4	97.8	58.3	87.2	93.9	505.5
	RCL (TPAMI′23)	67.7	89.1	93.6	48.0	74.9	83.3	456.6	74.0	94.3	97.5	57.6	86.4	93.5	503.3
	CRCL (NeurIPS′23)	70.4	90.4	94.9	52.6	78.1	85.1	471.5	75.2	94.9	98.0	60.1	88.5	94.8	511.5
	SREM (AAAI′24)	71.0	92.1	96.1	54.0	80.1	87.0	480.3	74.5	94.5	97.9	58.7	87.5	93.9	506.9
	PC^2 (ACM MM′24)	70.8	90.3	94.4	53.1	79.0	85.9	473.5	74.2	94.4	97.8	58.9	87.5	93.8	506.6
	L2RM (CVPR′24)	70.0	90.8	95.4	51.3	76.4	83.7	467.6	75.4	94.7	97.9	59.2	87.4	93.8	508.4
	ESC (CVPR′24)	72.6	90.9	94.6	53.0	78.6	85.3	475.0	77.2	95.1	98.1	61.1	88.6	94.9	515.0
	GSC (CVPR′24)	70.8	91.1	95.9	53.6	79.8	86.8	478.0	75.6	95.1	98.0	60.0	88.3	94.6	511.7
	ReCon (CVPR′25)	74.3	93.6	96.6	55.7	81.6	88.1	489.9	77.2	95.9	98.4	61.8	89.3	95.2	517.8
	DNS (Ours)	**76.4**	**94.8**	**97.7**	**57.6**	**82.4**	**89.1**	**497.9**	**79.9**	**96.6**	**98.5**	**61.6**	**89.7**	**95.8**	**522.2**

Table 1: Retrieval performance comparison on Flickr30K and MS-COCO 1K datasets under 20%, 40% and 60% noise ratio respectively. The best and the second best results are respectively marked by bold and underline.

Experiments

Datasets and Evaluation Protocol

Datasets. We evaluate the effectiveness of our method on three prevalent benchmarks. Specifically, **Flickr30K** (Young et al. 2014) contains 31000 Flickr-sourced images, each annotated with five crowd-sourced captions. Following standard protocol, we use 1000 pairs for validation, 1000 for testing, and the remainder for training. **MS-COCO** (Lin et al. 2014) contains 123287 images, each annotated with five captions. We utilize 113287 training pairs, 5000 validation pairs, and 5000 test pairs. We report results using both 5-fold cross-validation (1000 test pairs per fold) and full evaluation on the 5000-pair test set. **Conceptual Captions (CC)** (Sharma et al. 2018) is a large-scale web-crawled corpus containing image-text pairs with inherent noise, where approximately 3%-20% instances exhibit mismatched or weakly-aligned semantics. We employ its subset named CC152K, which comprises 150000 pairs for training, 1000 pairs for validation, and 1000 pairs for testing.

Evaluation Protocol. Retrieval performance is evaluated using Recall@K (R@K), defined as the fraction of ground-truth items retrieved in the top-K positions. Our experiments include bidirectional retrieval (image-to-text and text-to-image) and report R@1, R@5, R@10, and their sum (rSum) for comprehensive performance assessment.

Implementation Details

For fair comparisons, all experiments are conducted using the same backbone SGR (Diao et al. 2021). We train the model using Adam optimizer with a batch size of 128, initi-

Methods	Image to Text			Text to Image			
	R@1	R@5	R@10	R@1	R@5	R@10	rSum
NCR	39.5	64.5	73.5	40.3	64.6	73.2	355.6
DECL	39.0	66.1	75.5	40.7	66.3	76.7	364.3
MSCN	40.1	65.7	76.6	40.6	67.4	76.3	366.7
BiCro	40.8	67.2	76.1	42.1	67.6	76.4	370.2
RCL	41.7	66.0	73.6	41.6	66.4	75.1	364.4
CRCL	41.8	67.4	76.5	41.6	68.0	**78.4**	373.7
SREM	40.9	67.5	77.1	41.5	68.2	77.0	372.2
PC^2	39.3	66.4	75.4	39.8	66.4	76.8	364.1
L2RM	43.0	67.5	75.7	42.8	68.0	77.2	374.2
ESC	42.8	67.3	76.9	44.8	68.2	75.9	375.9
GSC	42.1	68.4	77.7	42.2	67.6	77.1	375.1
ReCon	<u>43.1</u>	<u>68.7</u>	<u>78.1</u>	<u>44.9</u>	<u>68.3</u>	77.4	<u>380.5</u>
DNS	**43.7**	**69.4**	**78.9**	**45.2**	**68.8**	<u>78.1</u>	**384.1**

Table 2: Comparisons on CC52K. The best and the second best results are respectively marked by bold and underline.

Noise	Methods	Image to Text			Text to Image			
		R@1	R@5	R@10	R@1	R@5	R@10	rSum
0% Zero-shot	CLIP-32	58.4	81.5	88.1	37.8	62.4	72.2	400.4
	CLIP-14	50.2	74.6	83.6	30.4	56.0	66.8	361.6
	ReCon	61.6	86.7	92.7	44.4	73.1	83.1	441.6
	DNS	**62.1**	**87.4**	**93.3**	**44.8**	**73.7**	**83.9**	**445.2**
20% Finetune	CLIP-32	21.4	49.6	63.3	14.8	37.6	49.6	236.3
	CLIP-14	36.1	61.3	72.5	22.6	43.2	53.7	289.4
	ReCon	61.1	85.7	92.2	43.5	72.4	82.7	437.6
	DNS	**61.7**	**86.5**	**92.6**	**43.9**	**73.0**	**83.4**	**441.1**
50% Finetune	CLIP-32	10.9	27.8	38.3	7.8	19.5	26.8	131.1
	ReCon	58.1	85.1	91.9	41.5	70.7	81.0	428.3
	DNS	**59.0**	**86.3**	**92.6**	**42.4**	**71.9**	**82.1**	**434.3**

Table 3: Comparisons with CLIP on MS-COCO 5K. The best results are marked in bold.

ating with a learning rate of 2×10^{-4}, which is reduced by a factor of 0.2 every 15 epochs. For hyperparameter configuration, we adopt the following default settings: temperature scaling factor $\tau = 0.07$ controls the sharpness of distribution calibration, with neighborhood size $K = 3$ determining the scope of context aggregation. Momentum coefficients β_1 and β_2 are both set to 0.7 to stabilize training dynamics, while the regularization weights λ_1 and λ_2 are set to 0.05 to balance loss components. These hyper-parameters are further analyzed in the section of Ablation Study. We conduct all experiments on an NVIDIA Tesla A100 GPU.

Comparison with the State-of-the-Arts

We performed a comprehensive evaluation of our proposed DNS against twelve state-of-the-art methods across three benchmark datasets, demonstrating its superior effectiveness. These baselines include NCR (Huang et al. 2021), DECL (Qin et al. 2022), MSCN (Han et al. 2023), BiCro (Yang et al. 2023), RCL (Hu et al. 2023), CRCL (Qin et al. 2023), SREM (Dang et al. 2024), PC^2 (Duan et al. 2024), L2RM (Han et al. 2024), ESC (Yang et al. 2024), GSC (Zhao et al. 2024) and ReCon (Zha et al. 2025). To evaluate model robustness, we simulated noisy correspondences at three different noise levels (20%, 40%, and 60%) by randomly permuting captions in Flickr30K and MS-COCO, as in (Huang et al. 2021). To validate the applicability of our approach in real-world noisy scenarios, we perform a comprehensive evaluation on the CC152K benchmark. Besides, we compare it with large-scale pre-trained vision-language models.

Results on Simulated NC. To assess the robustness of various methods under different noise intensities, we systematically introduce simulated noise at rates of 20%, 40%, and 60% on carefully curated benchmark datasets, as previous methods (Huang et al. 2021). This controlled experimentation enables quantitative evaluation of algorithm performance across a spectrum of noise conditions. We summarize the experimental results in Table 1, which demonstrate that our proposed DNS consistently outperforms other state-of-the-art baselines across the majority of evaluation metrics. Notably, our DNS achieves greater performance gains under higher levels of noise, such as 4.4 absolute points for the rSum compared with ReCon. This suggests that DNS can effectively handle NCs. We observe that the performance gain becomes more pronounced as noise levels increase. Specifically, under high-noise conditions, DNS achieves a substantial absolute improvement of 4.4 points in the rSum metric compared to the ReCon baseline. This observation highlights DNS's capability in effectively addressing NCs.

Results on Real-World NC. To further validate the effectiveness of DNS in addressing noisy correspondence within practical real-world scenarios, we present comprehensive quantitative results evaluated on the CC152K benchmark. The experimental results are summarized in Table 2. We observe that the proposed DNS achieves a substantial performance advantage over baseline methods, demonstrating a notable absolute improvement of 3.6 points in the overall score compared to the second-best performer, ReCon (380.5 vs. 384.1). The superiority is consistently observed across different sub-metrics, indicating in handling complex correspondence patterns in real-world applications.

Comparison to pre-trained model. To further highlight the superiority of DNS, we perform a comparative analysis against the popular large-scale pre-trained vision-language model, namely CLIP (Radford et al. 2021). CLIP is a powerful baseline trained on 400 million web-crawled image-text pairs with a large number of real NCs. In line with the previous studies (Huang et al. 2021; Zha et al. 2025), we adopt three baseline models from the CLIP family—specifically CLIP-14 and CLIP-32, and ReCon, to systematically investigate two experimental paradigms: zero-shot learning (0% NCs) and fine-tuning (20% and 50% NCs), on the MS-COCO 5K dataset. From the Table 3, we observe that there exists an apparent performance decline in CLIP models for all different noise rates. This phenomenon can be primarily attributed to the model's inherent limitations in handling noisy image-text correspondences. Specifically, CLIP lacks robust mechanisms to filter or mitigate the impact of mis-

LRP	SGD	SDD	Image to Text R@1	R@5	R@10	Text to Image R@1	R@5	R@10	rSum
✘	✘	✘	69.2	90.4	94.3	51.8	77.7	84.9	468.3
✔	✘	✘	72.7	91.1	95.4	53.6	78.5	86.1	477.4
✔	✔	✘	74.8	92.2	96.3	55.8	79.3	87.7	486.1
✔	✘	✔	75.2	93.0	96.8	56.2	80.4	88.5	490.1
✔	✔	✔	**76.4**	**94.8**	**97.7**	**57.6**	**82.4**	**89.1**	**497.9**

Table 4: Ablation studies on Flick30K with 60% noise with different components. The best results are marked in bold.

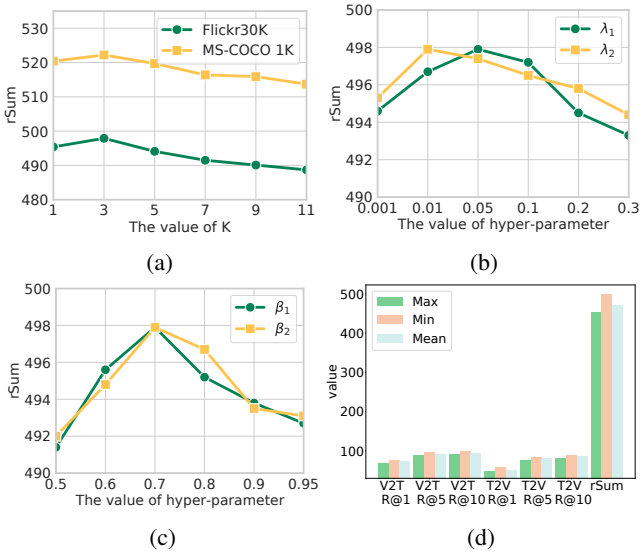

Figure 3: Performance under different hyper-parameters and optimization strategies with 60% NCs.

matched image-text pairs. Such noisy correspondences introduce ambiguities during cross-modal alignment, thereby degrading the model's ability to learn discriminative representations. In contrast, our DNS exhibits remarkable performance across all experimental settings. Notably, when evaluated under challenging conditions with 50% noisy correspondences, DNS demonstrates remarkable superiority and robustness, and its performance even surpasses CLIP's zero-shot results achieved under ideal noise-free conditions.

Ablation Study

Impacts of components. We conduct a component-wise ablation study on the MS-COCO 1K dataset with 60% NCs to dissect the individual contributions of each module within the DNS framework, with quantitative results summarized in Table 4. Empirical observations reveal that the full DNS framework achieves optimal performance when all components are synergistically integrated. Notably, the incremental integration of each module yields apparent performance gains, with ablation experiments demonstrating: 1) each proposed component independently enhances model capability, and 2) their collaboration generates a compounding effect that substantially elevates the algorithm's robustness against

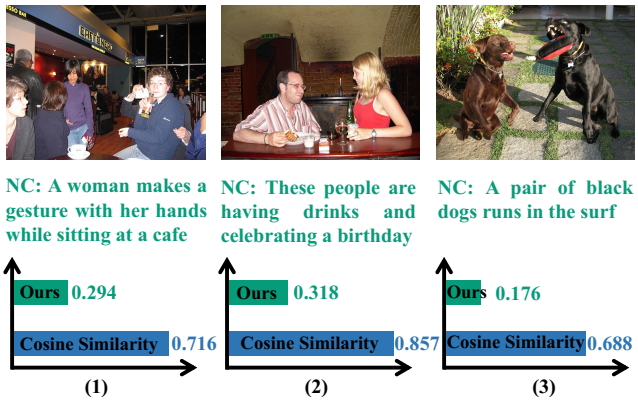

Figure 4: Visualization of predicted mismatched pairs.

noisy correspondence challenges.

Impacts of hyper-parameters. We explore different hyper-parameter settings and optimization strategies in this section. The detailed results are presented in Figure 3. Specifically, K serves as a critical hyper-parameter in neighborhood-based verification. To determine its optimal value, we conducted a systematic investigation on Flickr30K and MS-COCO 1K with 60% NCs. Through empirical analysis, we observe that $K = 3$ yields the best performance. This can be attributed to the fact that an excessively small K fails to capture sufficient contextual information from neighboring instances, while an overly large K introduces irrelevant or noisy information that degrades performance. Also, we also explore different values of the hyper-parameters λ_1 and λ_2 in Figure 3 (b), and β_1 and β_2 in Figure 3 (c) on Flickr30K dataset with 60% NCs.

Besides, we further investigate various optimization strategies for the final correspondence indicator y^i in Eq. 3. Here, we systematically study three distinct aggregation strategies: maximization, minimization, and averaging, with results visualized in Figure 3 (d). As expected, the minimization strategy, which mitigates challenges arising from different aspects, outperforms other strategies, while the maximization strategy demonstrates the weakest performance.

Visualization. Figure 4 visualizes some predicted mismatched pairs of our DNS and the baseline that directly uses cosine similarity of different modalities. While the examples demonstrate significant local correspondence, DNS exhibits superior discriminative capability by assigning lower affinity scores to true mismatches.

Conclusion

This paper introduces a dynamic neighborhood semantic verification paradigm for noisy correspondence learning. Existing methods, often relying on direct cross-modal pairwise similarity, are limited by noise sensitivity and contextual blindness. We hypothesize that pairwise matching can be quantified through semantic neighbor relationships, and propose three core components, *i.e.*, local relation proximity, semantic drift distance, and semantic gap decomposition, to enhance noisy correspondence discrimination. Experimental results on multiple benchmarks validate its superiority.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62406226, in part sponsored by Shanghai Sailing Program under Grant 24YF2748700, in part by New-Generation Information Technology under the Shanghai Key Technology R&D Program under Grant 25511103500, and in part by the National Key Research and Development Project under Grant 2023YFC3806000, and in part sponsored by Tongji University Independent Original Cultivation Project under Grant 22120240326.

References

Chen, J.; Hu, H.; Wu, H.; Jiang, Y.; and Wang, C. 2021. Learning the best pooling strategy for visual semantic embedding. In *CVPR*, 15789–15798.

Dang, Z.; Luo, M.; Jia, C.; Dai, G.; Chang, X.; and Wang, J. 2024. Noisy correspondence learning with self-reinforcing errors mitigation. In *AAAI*, 1463–1471.

Diao, H.; Zhang, Y.; Ma, L.; and Lu, H. 2021. Similarity reasoning and filtration for image-text matching. In *AAAI*, 1218–1226.

Duan, Y.; Gu, Z.; Ying, Z.; Qi, L.; Meng, C.; and Shi, Y. 2024. PC2: Pseudo-Classification Based Pseudo-Captioning for Noisy Correspondence Learning in Cross-Modal Retrieval. In *ACM MM*, 9397–9406.

Feng, Z.; Zeng, Z.; Guo, C.; Li, Z.; and Hu, L. 2023. Learning from noisy correspondence with tri-partition for cross-modal matching. *IEEE TMM*, 26: 3884–3896.

Fu, Z.; Mao, Z.; Song, Y.; and Zhang, Y. 2023. Learning semantic relationship among instances for image-text matching. In *CVPR*, 15159–15168.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 8536–8546.

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

Han, H.; Zheng, Q.; Dai, G.; Luo, M.; and Wang, J. 2024. Learning to rematch mismatched pairs for robust cross-modal retrieval. In *CVPR*, 26679–26688.

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

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

Kim, J. M.; Koepke, A.; Schmid, C.; and Akata, Z. 2023. Exposing and mitigating spurious correlations for cross-modal retrieval. In *CVPR*, 2585–2595.

Lee, K.-H.; Chen, X.; Hua, G.; Hu, H.; and He, X. 2018. Stacked cross attention for image-text matching. In *ECCV*, 201–216.

Li, J.; Socher, R.; and Hoi, S. C. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. In *ICLR*.

Li, K.; Zhang, Y.; Li, K.; Li, Y.; and Fu, Y. 2022. Image-text embedding learning via visual and textual semantic reasoning. *IEEE TPAMI*, 45(1): 641–656.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*, 740–755.

Liu, S.; Niles-Weed, J.; Razavian, N.; and Fernandez-Granda, C. 2020. Early-learning regularization prevents memorization of noisy labels. In *NeurIPS*, 20331–20342.

Ma, X.; Yang, M.; Li, Y.; Hu, P.; Lv, J.; and Peng, X. 2024. Cross-modal retrieval with noisy correspondence via consistency refining and mining. *IEEE TIP*, 33: 2587–2598.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Pan, Z.; Wu, F.; and Zhang, B. 2023. Fine-grained image-text matching by cross-modal hard aligning network. In *CVPR*, 19275–19284.

Pham, K.; Huynh, C.; Lim, S.-N.; and Shrivastava, A. 2024. Composing object relations and attributes for image-text matching. In *CVPR*, 14354–14363.

Qin, Y.; Peng, D.; Peng, X.; Wang, X.; and Hu, P. 2022. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *ACM MM*, 4948–4956.

Qin, Y.; Sun, Y.; Peng, D.; Zhou, J. T.; Peng, X.; and Hu, P. 2023. Cross-modal active complementary learning with self-refining correspondence. In *NeurIPS*, 24829–24840.

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.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2556–2565.

Shen, L.; Gong, G.; Hao, T.; He, T.; Zhang, Y.; Liu, P.; Zhao, S.; Han, J.; and Ding, G. 2025. DiscoVLA: Discrepancy Reduction in Vision, Language, and Alignment for Parameter-Efficient Video-Text Retrieval. In *CVPR*, 19702–19712.

Wang, Y.; and Chen, S. 2025. Learning Event Completeness for Weakly Supervised Video Anomaly Detection. In *ICML*.

Wang, Y.; Zhao, S.; and Chen, S. 2024. SQL-Net: Semantic query learning for point-supervised temporal action localization. *IEEE TMM*, 27: 84–94.

Yang, S.; Xu, Z.; Wang, K.; You, Y.; Yao, H.; Liu, T.; and Xu, M. 2023. Bicro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency. In *CVPR*, 19883–19892.

Yang, Y.; Wang, L.; Yang, E.; and Deng, C. 2024. Robust noisy correspondence learning with equivariant similarity consistency. In *CVPR*, 17700–17709.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2: 67–78.

Zha, Q.; Liu, X.; Cheung, Y.-m.; Xu, X.; Wang, N.; and Cao, J. 2024. Ugncl: Uncertainty-guided noisy correspondence learning for efficient cross-modal matching. In *ACM SIGIR*, 852–861.

Zha, Q.; Liu, X.; Peng, S.-J.; Cheung, Y.-m.; Xu, X.; and Wang, N. 2025. ReCon: Enhancing True Correspondence Discrimination through Relation Consistency for Robust Noisy Correspondence Learning. In *CVPR*, 29680–29689.

Zhao, Z.; Chen, M.; Dai, T.; Yao, J.; Han, B.; Zhang, Y.; and Wang, Y. 2024. Mitigating noisy correspondence by geometrical structure consistency learning. In *CVPR*, 27381–27390.