Re-Weighting Cross-Modal Pairs via Rank Consistency for Noise-Robust Retrieval

Weiran Pan[1,2], **Wei Wei**[1,2*]
[1]School of Computer Science and Technology, Huazhong University of Science and Technology,
[2]Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL).
panwr789@gmail.com, weiw@hust.edu.cn

Abstract

Image-text alignment relies on large-scale, well-aligned image-text pairs, yet web-crawled datasets inevitably introduce noisy correspondence. Existing methods typically re-weight training pairs by estimating semantic matching degrees from the model's own similarity predictions, which are inherently unreliable under noise and often overestimate partially aligned pairs. To overcome this limitation, we introduce a novel semantic matching degree estimation method based on ranking consistency. Our basic idea is that for a well-aligned image-text pair (I, T), the image I and text T should occupy semantically consistent positions in the embedding space. Thus, using I and T as queries to retrieve other texts (images) should yield two highly consistent rankings. Therefore, we reweight the training paris by measuring the Normalized Discounted Cumulative Gain (NDCG) between its cross-modal and intra-modal retrieval results. This signal is derived from the global data manifold structure rather than point-wise predictions, making it more robust to noise. Experiments on multiple cross-modal retrieval benchmarks validate the effectiveness of our method. Code is available at https://github.com/Aliinton/RCR.

1 Introduction

Cross-modal alignment learns a unified space where semantically related data lie close together [Girdhar *et al.*, 2023], yet its success heavily relies on large-scale, correctly aligned multi-modal datasets that are expensive to curate. Crawling multi-modal pairs from the Internet [Jia *et al.*, 2021] scales better but inevitably introduces mismatched or partially relevant pairs (*e.g.,* an image unrelated to its alt text), a problem known as Noisy Correspondence [Huang *et al.*, 2021]. Training standard models (*e.g.,* CLIP [Radford *et al.*, 2021]) on such noisy data forces them to fit spurious correlations, severely degrading retrieval performance [Zhang *et al.*, 2024]. Therefore, developing robust learning algorithms that can resist noisy correspondence is of critical practical importance.

Well-aligned: Jockeys racing horses down a grass track.

Partial-aligned: A group of horses and their riders are racing each other.

Mismatched: Several people working outside to construct a wood frame.

Figure 1: Examples of data pairs with different matching degrees.

This paper studies image-text matching under noisy correspondence. A widely adopted strategy is to estimate a soft matching degree for each pair and adjust its training weight accordingly [Huang *et al.*, 2021; Zhang *et al.*, 2024; Li *et al.*, 2025]. As shown in Figure 1, pairs range from well-aligned (accurate description) to partially aligned (missing details such as the background grass track) to fully mismatched, and should receive high, medium, and zero weight, respectively. Accurately estimating this matching degree is challenging. Existing methods are trained with contrastive learning, which only enforces relative ordering rather than calibrated absolute scores. Consequently, model-predicted similarities are unreliable measures of true semantic alignment. To address this, prior works map raw similarities to $[0, 1]$ weights [Huang *et al.*, 2021; Ma *et al.*, 2024] or construct local graphs for label propagation [Li *et al.*, 2025]. Despite their progress, these approaches fundamentally rely on the model's own similarity predictions, assuming higher scores indicate better matches. This assumption fails under noisy training: the model can be misled to assign deceptively high similarity to partially aligned pairs (see Section 4.6), causing inaccurate re-weighting.

To overcome this limitation, we evaluate matching quality through *ranking consistency* rather than absolute similarity. Our key insight is that a well-aligned pair (I, T) should occupy semantically consistent positions in the embedding space; hence, using I and T as queries to retrieve other texts (or images) should yield consistent rankings. Figure 2 illustrates this: although T_1 and T_2 are equidistant from I_1, T_2

*Corresponding Author.

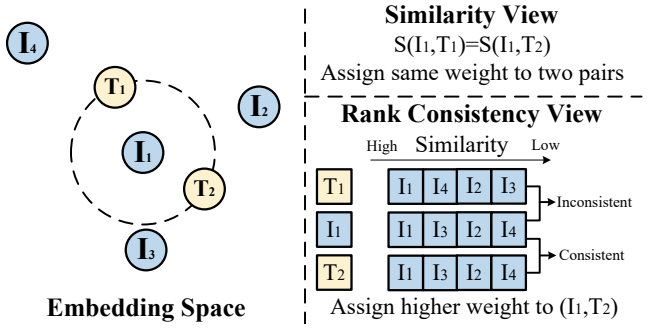

Figure 2: An illustration of estimating the matching degree between text and image by comparing the consistency between cross-modal retrieval and intra-modal retrieval.

produces rankings more consistent with I_1's neighborhood structure and therefore deserves a higher weight. Case studies in Section 4.6 provide further intuitive examples.

In practice, we treat intra-modal retrieval as the ideal ranking and cross-modal retrieval as the actual ranking, then measure their agreement via Normalized Discounted Cumulative Gain (NDCG) [Järvelin and Kekäläinen, 2000]. Because this weight derives from the global manifold structure rather than point-wise predictions, it remains stable even when individual similarity scores are unreliable. To sum up, our main contributions are as follows:

- We propose a new re-weighting method based on ranking consistency between cross-modal and intra-modal retrieval, offering a novel solution to handle noisy correspondence in an end-to-end manner.
- Our method can finely distinguish between well-aligned and partially aligned pairs even when they receive similar similarity scores, enabling more precise soft-label estimation in noisy correspondence.
- Extensive experiments on three benchmarks (Flickr30K, MSCOCO, and CC120k) under both synthetic and real-world noise demonstrate our method achieves state-of-the-art performance.

2 Related Work

2.1 Cross-modal Retrieval

Cross-modal retrieval aims to align data from different modalities within the same embedding space. Current methods mainly focus on how to measure the similarity between images and text, which can be broadly categorized into two types: (1) coarse-grained measurement [Kiros *et al.*, 2014; Faghri *et al.*, 2018], which measures whether the global features of images and text have the same semantics. (2) fine-grained measurement [Lee *et al.*, 2018; Li *et al.*, 2019; Chen *et al.*, 2020; Diao *et al.*, 2021; Xie *et al.*, 2025b; Xie *et al.*, 2025a], which measures the semantic consistency of fine-grained fragments and considers the relationships between fragments. Existing models have achieved significant success in cross-modal retrieval. Pre-trained models like CLIP [Radford *et al.*, 2021] demonstrating powerful zero-shot cross-modal retrieval performance. However, prior work reveal they remain sensitive to noisy correspondence issues. When the fine-tuning data in downstream applications contains mismatched data, model performance is severely compromised.

2.2 Noisy Correspondence Learning

Noisy Correspondence Learning is first introduced by Huang *et al.* [2021], aims to develop robust cross-modal alignment algorithms to reduce the impact of mismatched and partially matched data on model performance. The core of existing methods is to accurately estimate a soft label reflecting the semantic matching degree of each pair. To achieve this, NCR [Huang *et al.*, 2021] first divides samples into clean and noisy subsets using the sample selection method in DivideMix [Han *et al.*, 2018; Arazo *et al.*, 2019; Li *et al.*, 2020], then converts the model's predicted image-text similarity into soft labels using an adaptive prediction function, and finally trains the model using triple loss with soft margin. Subsequent work has investigated more fine-grained sample partitioning schemes [Ma *et al.*, 2024; Zha *et al.*, 2024; Lyu *et al.*, 2025] and considered the relationships between samples, estimating soft labels through optimal transmission processes [Han *et al.*, 2024] or label propagation [Li *et al.*, 2025]. These methods typically assume that image-text pairs with higher predicted similarity (or lower loss) are more matched, making them sensitive to inaccurate image-text similarity predictions by the model.

Other methods consider image-text pairs with smaller losses as clean anchors to estimate soft labels. BiCro [Yang *et al.*, 2023] assumes that similar images should correspond to similar text and estimates soft labels by comparing the distances of other image-text pairs to the clean anchors. Recent work NPC [Zhang *et al.*, 2024] estimates soft labels by analyzing changes in the model's predictions on clean anchors before and after gradient updates. The performance of these methods is highly dependent on the selection of clean anchors. If the clean anchors contains partially matched pairs, the estimated soft labels are unreliable. Furthermore, because the selection of clean anchors depends on model predictions, these methods are also sensitive to inaccurate image-text similarity predictions by the model.

In addition to soft label estimation, existing work has also explored other directions, including developing loss functions robust to noisy correspondence [Qin *et al.*, 2022; Hu *et al.*, 2023; Qin *et al.*, 2023], introducing regularization terms to constrain the feature space [Yang *et al.*, 2024; Zhao *et al.*, 2024], and generating pseudo captions for mismatched pairs [Duan *et al.*, 2024; Han and Cao, 2025; Xie *et al.*, 2025c]. These methods are orthogonal to soft label estimation, improving performance from different aspects.

Our method offers a novel approach for soft label estimation: comparing the relative relationships between image and text with other data points. This method focuses on relative ranking structure, which is more robust to inaccurate similarity predictions by the model, and eliminates the need for clean anchor selection, offering a simple and effective method for noisy correspondence learning.

3 Method

3.1 Preliminary

Cross-modal retrieval [Kiros *et al.*, 2014] aligns different modalities into a shared embedding space. Formally, given a training set $\{(I_i, T_i), y_i\}_{i=1}^{N}$, (I_i, T_i) denotes the i-th image-text pair with label $y_i \in \{0, 1\}$ ($y_i = 1$ for matched, 0 for mismatched). The noisy correspondence problem [Huang *et al.*, 2021] occurs when partially or fully irrelevant pairs are incorrectly labeled as matched ($y_i = 1$). The goal is to learn robust correspondence despite such label noise. Following prior work [Zhang *et al.*, 2024; Li *et al.*, 2025], we adopt CLIP [Radford *et al.*, 2021] as the backbone, comprising an image encoder $f(\cdot)$ and a text encoder $g(\cdot)$. The similarity between a pair is measured by cosine similarity:

$$S(I_i, T_i) = \frac{f(I_i) \cdot g(T_i)}{\|f(I_i)\| \cdot \|g(T_i)\|}. \tag{1}$$

CLIP is trained with in-batch negative sampling strategy and optimized by the InfoNCE loss [van den Oord *et al.*, 2018]:

$$P_i^{i2t} = \frac{\exp(S(I_i, T_i)/\tau)}{\sum_{j=1}^{b} \exp(S(I_i, T_j)/\tau)}, \tag{2}$$

$$P_i^{t2i} = \frac{\exp(S(I_i, T_i)/\tau)}{\sum_{j=1}^{b} \exp(S(I_j, T_i)/\tau)}, \tag{3}$$

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{N} \sum_{i=1}^{N} \left(\log P_i^{i2t} + \log P_i^{t2i} \right), \tag{4}$$

where b is the batch size, τ is the temperature parameter which controls the sharpness of the probability distribution.

3.2 Data Partitioning

To prevent the model from memorizing the noisy correspondences, we propose a two-phase strategy: first filtering out obviously mismatched image-text pairs based on loss distribution; then re-weighting the remaining pairs according to their semantic matching degrees. This subsection details the first phase: filtering mismatched pairs by fitting Gaussian Mixture Models (GMM) [Arazo *et al.*, 2019] on per-sample losses.

Sample Loss Computation. Memorization effect in Deep Neural Networks (DNNs) indicates DNNs typically learn from correct labels before fitting noisy ones [Arpit *et al.*, 2017; Liu *et al.*, 2020]. Therefore, we treat pairs with higher losses as mismatched data. For an image-text pair (I_i, T_i), we compute its contrastive loss in both directions:

$$\ell_i^{\text{i2t}} = -\log P_i^{\text{i2t}}, \quad \ell_i^{\text{t2i}} = -\log P_i^{\text{t2i}}. \tag{5}$$

Loss Distribution Modeling. To distinguish high-loss samples, we fit the loss distributions using one-dimension two-component Gaussian Mixture Models: the component with the higher mean corresponds to mismatched sample pairs, and the other component corresponds to matched or partially matched sample pairs. The distribution is formally defined as:

$$p(\ell) = \sum_{k=1}^{2} \pi_k \cdot \mathcal{N}(\ell \mid \mu_k, \sigma_k^2), \tag{6}$$

where π_k, μ_k, and σ_k^2 denote the weight, mean, and variance of the k-th Gaussian component, respectively. We fit this GMM independently to the sets ℓ_i^{i2t} and ℓ_i^{t2i} using the Expectation-Maximization (EM) algorithm.

Sample Filtering. After fitting the GMM, for each sample i, we calculate the posterior probability that ℓ_i belongs to the lower-mean Gaussian component k', which is computed via Bayes' theorem:

$$\alpha_i = \frac{\pi_{k'} \cdot \mathcal{N}(\ell_i \mid \mu_{k'}, \sigma_{k'})}{p(\ell_i)}. \tag{7}$$

We require a sample to exhibit low loss in both directions to be considered a candidate for subsequent re-weighting. The filtered candidate set C is defined as:

$$C = \{(I_i, T_i) \mid \alpha_i^{\text{i2t}} > \delta \text{ and } \alpha_i^{\text{t2i}} > \delta\}, \tag{8}$$

where δ is a probability threshold. Samples outside C are considered clear mismatched pairs and are assigned zero weight during the subsequent re-weighting stage.

3.3 Sample Reweighting by Ranking Consistency

This section address the core challenge: reliably re-weighting pairs in C, which contain both well matched and partially matched pairs.

Figure 3 illustrates our basic idea: if image I_i and text T_i describe the same semantics, they should occupy similar positions in the feature space. Consequently, using I_i and T_i as queries to retrieve other texts (or images) in the dataset should yield two similar rankings lists. We measure the consistency of ranking lists by Normalized Discount Cumulative Gain (NDCG), which is a standard information retrieval metric that evaluates the quality of a ranking list against an ideal order. Take estimating the matching degree of (I_i, T_i) via text retrieval as an example. We compare following two ranking:

- Ideal Ranking: The ranking of all texts based on their similarity with T_i. We denote this ranked list as $\mathcal{R}^{t2t}(T_i)$, and $\mathcal{R}^{t2t}(T_i)_j$ represents the text in the training set with the j-th largest similarity to T_i.
- Actual Ranking: The ranking of all texts based on their similarity with I_i. We denote this ranked list as $\mathcal{R}^{i2t}(I_i)$ and $\mathcal{R}^{i2t}(I_i)_j$ refers to the text in the training set with the j-th largest similarity to I_i.

We quantify this top-rank consistency between $\mathcal{R}^{i2t}(I_i)$ and $\mathcal{R}^{t2t}(T_i)$ using NDCG. For a given truncation level K, the Discounted Cumulative Gain (DCG) of the actual ranking is:

$$\text{DCG@}K_{text}(I_i, T_i) = \sum_{j=1}^{K} \frac{2^{S(T_i, \mathcal{R}^{i2t}(I_i)_j)} - 1}{\log_2(j+1)}. \tag{9}$$

We treat $S(T_i, \mathcal{R}^{i2t}(I_i)_j)$ as the "relevance" of the text at position j in the actual ranking. This design is crucial: it uses the text-to-text similarity as the ideal relevance score to assess whether the image I_i has retrieved the texts that are semantically close to T_i. Generally, the higher the similarity between the top-ranked texts in $R^{i2t}(I_i)$ and T_i, the higher the DCG score. Since the DCG score is unbounded thus unsuitable as sample weights, we utilize the DCG score under ideal ranking

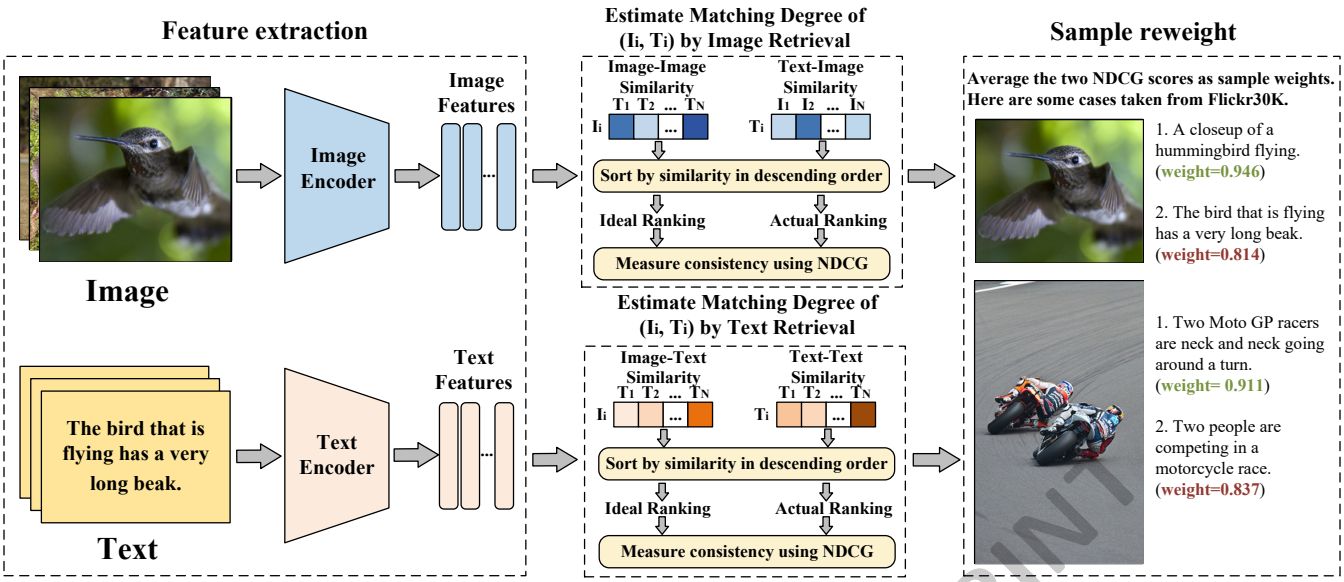

Figure 3: An illustration of our method: For a image-text pair (I_i, T_i), we use I_i and T_i as queries to perform text retrieval and image retrieval, respectively, and regard the consistency of the retrieval results as the semantic matching degree of (I_i, T_i).

(*i.e.,* IDCG) to normalize it to $[0, 1]$, which generates NDCG score in text retrieval:

$$\text{IDCG@}K_{text}(I_i, T_i) = \sum_{j=1}^{K} \frac{2^{S(T_i, \mathcal{R}^{t2t}(T_i)_j)} - 1}{\log_2(j+1)}, \quad (10)$$

$$\text{NDCG@}K_{text}(I_i, T_i) = \frac{\text{DCG@}K_{text}(I_i, T_i)}{\text{IDCG@}K_{text}(I_i, T_i)}. \quad (11)$$

Symmetrically, when estimate matching degree of (I_i, T_i) by image retrieval, we use $S(I_i, I_j)$ as the ideal relevance to evaluate the ranking produced by using text T_i to retrieve images. Denote the NDCG@K score for (I_i, T_i) in image retrieval as $\text{NDCG@}K_{img}(I_i, T_i)$, the final ranking consistency weight for the pair (I_i, T_i) is the average of the assessments in both image and text retrieve:

$$w_i = \frac{1}{2}(\text{NDCG@}K_{text}(I_i, T_i) + \text{NDCG@}K_{img}(I_i, T_i)). \quad (12)$$

A high w_i indicates that I_i and T_i have highly consistent neighborhood structures across modalities, implying strong semantic matching. During training, we assign zero weights to samples outside set C, avoiding the learning from mismatched pairs. For samples from set C, we reweight them using w_i to prioritize learning information from reliable pairs:

$$\hat{w}_i = \mathbb{I}((I_i, T_i) \in C) w_i, \quad (13)$$

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \hat{w}_i \left(\log P_i^{i2t} + \log P_i^{t2i} \right). \quad (14)$$

We term our method as Ranking Consistency Re-weighting (RCR).

4 Experiment

4.1 Datasets and Evaluation Metrics

Datasets. We evaluate our method on three widely-used image-text matching datasets: Flickr30K [Young *et al.*, 2014], MS-COCO [Lin *et al.*, 2014], and Conceptual Captions [Sharma *et al.*, 2018]. For well-annotated Flickr30K and MS-COCO, we randomly shuffling the captions of training images to simulate noisy correspondence. Conceptual Captions is crawled from the Internet, which contains $3\% \sim 20\%$ mismatched pairs. Following NPC [Zhang *et al.*, 2024], we use the CC120K subset in evaluation.

Evaluation Protocol. We use the recall rate at K (R@K) metric to evaluate retrieval performance. R@K measures the proportion of relevant items retrieved within the top K items. We report R@1, R@5, R@10 for both image-to-text and text-to-image retrieval, and their sum (rSum).

4.2 Implementation Details

Following prior work NPC [Zhang *et al.*, 2024] and GLP [Li *et al.*, 2025], we take CLIP with ViT-B/32 as backbone and train it using AdamW optimizer [Kingma and Ba, 2015; Loshchilov and Hutter, 2019] with a weight decay of 0.2. The initial learning rate is set to 5e-7, and cosine annealing is used for training. On all datasets, we train the model for 5 epochs with a mini-batch size of 256. The hyperparameter δ is chosen from $\{0.5, 0.9\}$, and K is chosen from $\{5, 10\}$ based on the validation performance.

4.3 Comparison with Advanced Methods

Following previous work GLP [Li *et al.*, 2025], we compare our Ranking Consistency Re-weighting (RCR) method with a comprehensive set of state-of-the-art baselines. These include robust learning methods built upon the SGRAF [Diao *et al.*, 2021]: NCR [Huang *et al.*, 2021], DECL [Qin *et al.*,

noise	method	Flickr30K							MSCOCO 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
0%	NCR	77.3	94.0	97.5	59.6	84.4	89.9	502.7	78.3	95.8	98.5	63.3	90.4	95.8	522.1
	DECL	79.8	94.9	97.4	59.5	83.9	89.5	505.0	79.1	96.3	98.7	63.3	90.1	95.6	523.1
	BiCro	81.7	95.3	98.4	61.6	85.6	90.8	513.4	79.1	96.4	98.6	63.8	90.4	96.0	524.3
	PC^2	78.7	94.8	97.0	60.0	84.4	89.8	504.7	79.1	96.5	**98.8**	64.0	90.3	95.6	524.3
	CLIP	86.2	97.6	99.2	72.9	92.3	96.0	544.2	79.6	95.3	98.2	64.5	89.9	95.8	523.3
	NPC	87.9	98.1	99.4	75.0	93.7	**97.2**	551.3	81.6	96.2	98.5	67.4	91.6	**96.7**	532.0
	GLP	**88.9**	98.1	99.4	75.1	93.4	96.7	551.6	81.4	95.7	98.4	67.9	91.7	96.5	531.6
	RCR	**88.9**	**98.2**	**99.5**	**75.7**	**94.2**	96.9	**553.4**	**82.4**	**96.6**	98.6	**68.7**	**92.1**	**96.7**	**535.2**
20%	NCR	73.5	93.2	96.6	56.9	82.4	88.5	491.1	77.7	95.5	98.2	62.5	89.3	95.3	518.5
	DECL	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
	BiCro	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
	L2RM	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
	PC^2	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
	CREMA	77.4	95.0	97.3	58.7	84.1	89.8	502.3	78.9	96.3	**98.6**	63.3	90.1	95.8	523.0
	SPS	79.5	95.0	98.0	60.5	84.3	89.8	507.1	79.8	**96.4**	**98.6**	64.3	90.5	95.8	525.4
	CLIP	82.3	95.5	98.3	66.0	88.5	93.5	524.1	74.4	93.1	97.2	58.3	85.7	93.2	501.9
	NPC	87.3	97.5	98.8	72.9	92.1	95.8	544.4	79.3	95.5	98.5	65.3	90.2	95.7	524.4
	GLP	88.1	**98.2**	99.4	74.1	93.3	96.3	549.4	81.8	96.3	98.4	67.5	91.5	**96.7**	532.1
	RCR	**88.4**	97.8	**99.6**	**75.8**	**93.8**	**97.0**	**552.5**	**81.9**	96.1	**98.6**	**68.1**	**91.9**	**96.7**	**533.4**
40%	NCR	68.1	89.6	94.8	51.4	78.4	84.8	467.1	74.7	94.6	98.0	59.6	88.1	94.7	509.7
	DECL	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
	BiCro	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
	L2RM	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
	PC^2	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
	CREMA	76.3	93.4	97.1	57.0	82.6	88.7	495.1	76.5	95.6	98.3	61.7	89.4	95.3	516.8
	SPS	77.8	93.6	97.1	57.3	83.5	89.6	498.9	79.2	95.9	**98.5**	63.3	89.8	95.4	522.1
	CLIP	76.2	93.3	96.5	59.4	85.0	90.9	501.3	70.6	91.4	96.2	54.7	83.3	91.5	487.7
	NPC	85.6	97.5	98.4	71.3	91.3	95.3	539.4	78.5	95.0	98.2	64.7	90.0	95.6	522.1
	GLP	87.9	98.1	**99.4**	73.8	92.9	96.5	548.6	80.8	**96.0**	**98.5**	66.8	91.0	96.3	529.4
	RCR	**88.5**	**98.3**	**99.4**	**74.4**	**93.9**	**96.8**	**551.3**	**81.5**	**96.0**	**98.5**	**67.6**	**91.5**	**96.6**	**531.7**
60%	NCR	13.9	37.7	50.5	11.0	30.1	41.4	184.6	0.1	0.3	0.4	0.1	0.5	1.0	2.4
	DECL	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
	BiCro	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
	L2RM	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
	PC^2	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
	CREMA	70.6	91.2	96.1	53.3	79.2	87.0	477.4	74.7	94.8	98.0	59.7	88.0	94.6	509.8
	SPS	73.4	92.7	96.3	53.7	80.2	87.7	484.0	77.6	95.7	98.3	61.6	89.0	95.1	517.3
	CLIP	66.3	87.3	93.0	52.1	78.8	87.4	464.9	67.0	88.8	95.0	49.7	79.6	75.0	455.1
	NPC	83.0	95.9	98.6	68.1	89.6	94.2	529.4	77.8	94.2	97.7	63.6	88.9	94.9	517.0
	GLP	88.2	**98.0**	99.2	72.2	92.2	96.2	546.0	79.7	95.7	**98.5**	66.3	91.0	96.1	527.1
	RCR	**88.6**	97.8	**99.5**	**74.4**	**93.0**	**96.5**	**549.8**	**80.3**	**96.0**	98.4	**66.6**	**91.2**	**96.3**	**528.8**

Table 1: Retrieval performance comparison on Flickr30K and MS-COCO 1k datasets under various noise levels. The best results and the second best results are respectively marked by **bold** and underline.

2022], BiCro [Yang *et al.*, 2023], L2RM [Han *et al.*, 2024], PC^2 [Duan *et al.*, 2024], CREMA [Ma *et al.*, 2024], and SPS [Xie *et al.*, 2025c]. To ensure a direct and fair comparison on the same foundation, we evaluate against SOTA methods based on the CLIP backbone: fine-tuned CLIP [Radford *et al.*, 2021], NPC [Zhang *et al.*, 2024], and GLP [Li *et al.*, 2025]. Notably, NPC and GLP are highly relevant baselines as they also address the problem through sample re-weighting schemes, making the performance comparison particularly indicative of the effectiveness of our proposed ranking consistency paradigm versus existing weighting strategies.

Results on Simulate Noise. Table 1 compares performance on Flickr30K and MS-COCO 1K under varying noise ratios. Our method achieves highest rSum scores in all configuration, reaching state-of-the-art performance across noise levels. A particularly insightful finding is that RCR also demonstrate significant improvements on the original, clean splits (0% noise). This indicates that even well-annotated datasets contain partially-matched or ambiguous pairs. Existing methods that rely on similarity predicted by the model often overestimate the reliability of these pairs. In contrast, our global ranking consistency measure naturally down-weights such pairs because their cross-modal and intra-modal retrieval results exhibit inherent inconsistencies, thereby refining the learning signal even in "clean" datasets. Our case study in Section 4.6 intuitively shows this advantage.

Method	image-to-text			text-to-image			
	R@1	R@5	R@10	R@1	R@5	R@10	rSum
CLIP	67.9	88.9	93.1	67.6	88.0	92.8	498.3
NPC	71.2	90.9	94.7	71.8	90.8	94.2	513.6
GLP	72.0	90.9	94.9	72.2	90.7	**94.3**	515.0
RCR	**73.3**	**91.0**	**95.5**	**73.0**	**90.8**	93.7	**517.3**

Table 2: Retrieval performance comparison on CC120K datasets.

Results on Real-world Noise. To validate the effectiveness of our method in real-world scenarios, we further evaluate it on the CC120K dataset, which contains a significant proportion of naturally occurring noisy correspondences. As shown in Table 2, our RCR method consistently achieves the best performance across nearly all metrics. Specifically, RCR attains the highest R@1 scores for both image-to-text (73.3) and text-to-image (73.0) retrieval, which are key indicators of precise matching. The overall retrieval sum (rSum) of 517.3 also represents a new state-of-the-art, surpassing the previous best method (GLP) by +2.3 points.

4.4 Ablation Study

Re-weight strategy	image-to-text			text-to-image			
	R@1	R@5	R@10	R@1	R@5	R@10	rSum
No Re-weight	70.4	90.4	94.5	71.6	89.3	93.5	509.7
Image-Text Similarity	65.7	87.4	92.8	65.0	88.0	92.7	491.6
Loss Distribution	70.7	90.7	94.9	71.0	89.6	93.6	510.5
Model Prediction	71.6	90.7	94.5	72.0	90.3	93.1	512.2
RCR	**73.3**	**91.0**	**95.5**	**73.0**	**90.8**	**93.7**	**517.3**

Table 3: Comparison of different re-weight strategy on CC120K.

To further analysis the effectiveness of our method, we evaluate the performance when replacing the w_i calculate in Eq. (12) with following strategies:

- No Re-weight: $w_i = 1$ for all pairs in C.
- Image-Text Similarity: Using the cosine similarity between image and text features as weight, *i.e.,* $w_i = S(I_i, T_i)$.
- Loss Distribution: Using the posterior probability from the GMM (Eq. 7) as the reliability weight, *i.e.,* $w_i = \frac{1}{2}(\alpha_i^{t2i} + \alpha_i^{i2t})$. This typically assigns higher weights to samples with lower losses.
- Model Prediction: Using the model's own in-batch prediction probabilities as weights. P_i^{i2t} in Eq. (2) denotes the probability that given the image I_i, its paired text T_i is correctly matched among all texts in the batch. Conversely, P_i^{t2i} in Eq. (3) represents the probability that given the text T_i, its paired image I_i is correctly matched among all images in the batch. We average these two probability and use it as the sample weights, *i.e.,* $w_i = \frac{1}{2}(P_i^{i2t} + P_i^{t2i})$.

Table 3 compares the performance of different re-weight strategy on CC120K dataset. The results lead to two critical findings that reinforce our motivation.

Firstly, using the raw image-text similarity as weight results in the worst performance (rSum: 491.6), even below the no re-weight baseline (rSum: 509.7). This stark outcome empirically confirms the challenge we highlighted in the introduction: models trained with contrastive objectives (like CLIP) are optimized to produce relative orderings which only ensures matched pairs have higher scores than mismatched ones, rather than output absolute, calibrated measures of semantic matching degree. Consequently, the similarity scores (*e.g.,* typically within a narrow range like 0.2-0.4 in our experiment) are unreliable as direct confidence weights.

Secondly, strategies that derive weights from processed model outputs like loss distribution and model prediction only yield marginal improvements over the no re-weight baseline. Also, their performance fall short of the previous state-of-the-art, GLP (rSum: 515.0 in Tab. 2). This indicates that while they mitigate some noise, they remain fundamentally constrained by the inherent biases and unreliability of the model's direct predictions in noisy settings. In contrast, our method derives weights from the global agreement of relative ranking orders rather than any absolute score, achieves a substantial gain (rSum: 517.3). It is the only re-weighting paradigm that surpasses GLP. This key distinction demonstrates that our method's effectiveness stems from its novel weighting criterion, rather than from a generic re-weighting effect.

4.5 Training Overhead Comparison

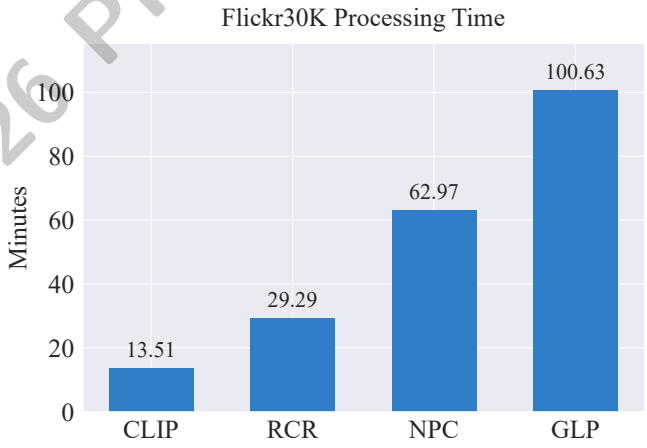

Figure 4: Efficiency Comparison of different methods.

Figure 4 compare the training time on Flickr30K for fine-tuning CLIP, NPC, GLP and our method (all methods train CLIP ViT-B/32 for 5 epochs with a batch size of 256.). While our RCR introduces inevitable overhead compared to vanilla CLIP for noise handling, it is significantly more efficient than other state-of-the-art re-weighting methods. Specifically, RCR is only about 2.2x slower than CLIP, whereas NPC and GLP are 4.7x and 7.4x slower, respectively. This efficiency stems from our method's non-iterative nature. GLP conduct costly iterative label propagation and NPC need additional gradient steps for weight-estimation. In contrast, our method the additional computational overhead of our method comes from NDCG calculations, which mainly involve similarity computation and ranking, both are highly parallelizable matrix operations.

Figure 5: Examples taken from Flickr30K dataset with 20% noise. Upper part shows weight given by different re-weight strategy. The lower part shows the Top-5 results when performing image retrieval using I2, T3, and T4 as queries.

4.6 Case Study

Figure 5 shows cases from Flickr30K with 20% injected noise. The upper panel compares different re-weighting strategies for two image-text pairs. For image I_1, text T_1 is more descriptively accurate than T_2, since T_2 omits the grass-track context. Similarly, for image I_2, text T_3 is more precise than T_4, as T_4 fails to mention the cornering action.

Despite these semantic differences, the fine-tuned model assigns nearly identical cosine similarities to all four pairs: $S(I_1, T_1) = 0.316$ vs. $S(I_1, T_2) = 0.324$, and $S(I_2, T_3) = 0.322$ vs. $S(I_2, T_4) = 0.325$. In fact, the partially aligned pairs (I_1, T_2) and (I_2, T_4) receive marginally *higher* similarity scores. This miscalibration also misleads GLP [Li *et al.*, 2025], which assigns the maximal weight 1.0 to all four pairs because its graph-based propagation is fundamentally anchored to these unreliable point-wise similarities.

Our RCR method successfully distinguishes the partially aligned pairs by inspecting the global neighborhood structure. Specifically, RCR assigns (I_1, T_1) a weight of 0.916 while down-weighting (I_1, T_2) to 0.827; likewise, (I_2, T_3) receives 0.911 whereas (I_2, T_4) receives 0.837. The lower panel of Figure 5 reveals why: when I_2, T_3, and T_4 are used to retrieve images, the Top-5 results for I_2 and T_3 are both dominated by motorcycle-racing images featuring sharp cornering scenes. In contrast, the Top-5 results for T_4 are less focused, mixing generic motorcycle images with other racing contexts. Because T_3's retrieval ranking aligns more closely with I_2's intra-modal neighborhood than T_4's does, NDCG assigns T_3 a higher consistency score. This example illustrates that RCR's reliance on *relative* ranking agreement, rather than *absolute* similarity magnitude, enables finer-grained discrimination between well-aligned and partially aligned pairs.

5 Conclusion

In this paper, we address the noisy correspondence problem in cross-modal retrieval by proposing a novel sample re-weighting method based on ranking consistency. Unlike existing methods that heavily rely on the unreliable image-text similarity predicted by the model, our key insight is to assess the quality of an image-text pair by measuring the consistency between cross-modal and intra-modal retrieval results. This design makes the weighting signal inherently more robust as it leverages the global manifold structure rather than point-wise, noisy model outputs. Extensive experiments on Flickr30K, MS-COCO, and CC120K under various synthetic and real-world noise levels demonstrate that our method consistently achieves state-of-the-art performance. The ablation and case study further validate that our ranking consistency strategy is more effective compared to weighting strategies based on model prediction or loss distribution. In conclusion, by shifting the focus from absolute similarity scores to relative ranking consistency, we provide a robust solution for learning with noisy correspondence, offering a fresh perspective for building reliable cross-modal retrieval systems.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62276110 and in part by the fund of Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL). The authors would also like to thank the anonymous reviewers for their comments on improving the quality of this paper.

References

[Arazo *et al.*, 2019] Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 312–321. PMLR, 2019.

[Arpit *et al.*, 2017] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron C. Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 233–242, 2017.

[Chen *et al.*, 2020] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12655–12663, 2020.

[Diao *et al.*, 2021] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, pages 1218–1226. AAAI Press, 2021.

[Duan *et al.*, 2024] Yue Duan, Zhangxuan Gu, Zhenzhe Ying, Lei Qi, Changhua Meng, and Yinghuan Shi. Pc^2: Pseudo-classification based pseudo-captioning for noisy correspondence learning in cross-modal retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9397–9406. ACM, 2024.

[Faghri *et al.*, 2018] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: improving visual-semantic embeddings with hard negatives. In *British Machine Vision Conference 2018, BMVC 2018*, page 12, 2018.

[Girdhar *et al.*, 2023] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind one embedding space to bind them all. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023.

[Han and Cao, 2025] Ao Han and Yang Cao. Cross-modal matching with noisy correspondence via neighbor replacing. In *Advanced Intelligent Computing Technology and Applications - 21st International Conference, ICIC 2025*, volume 15846 of *Lecture Notes in Computer Science*, pages 309–320, 2025.

[Han *et al.*, 2018] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018*, pages 8536–8546, 2018.

[Han *et al.*, 2024] Haochen Han, Qinghua Zheng, Guang Dai, Minnan Luo, and Jingdong Wang. Learning to rematch mismatched pairs for robust cross-modal retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, pages 26669–26678, 2024.

[Hu *et al.*, 2023] Peng Hu, Zhenyu Huang, Dezhong Peng, Xu Wang, and Xi Peng. Cross-modal retrieval with partially mismatched pairs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(8):9595–9610, 2023.

[Huang *et al.*, 2021] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, pages 29406–29419, 2021.

[Järvelin and Kekäläinen, 2000] Kalervo Järvelin and Jaana Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In *SIGIR 2000: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 41–48, 2000.

[Jia *et al.*, 2021] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916, 2021.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, 2015.

[Kiros *et al.*, 2014] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014.

[Lee *et al.*, 2018] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Computer Vision - ECCV 2018 - 15th European Conference*, volume 11208, pages 212–228, 2018.

[Li *et al.*, 2019] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019*, pages 4653–4661, 2019.

[Li *et al.*, 2020] Junnan Li, Richard Socher, and Steven C. H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *8th International Conference on Learning Representations, ICLR 2020*, 2020.

[Li *et al.*, 2025] Ruoxuan Li, Xiangyu Wu, and Yang Yang. Noise self-correction via relation propagation for robust cross-modal retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 4748–4757, 2025.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755, 2014.

[Liu *et al.*, 2020] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. In *Advances in Neural Information Processing Systems*, 2020.

[Loshchilov and Hutter, 2019] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations*, 2019.

[Lyu *et al.*, 2025] Shuai Lyu, Zijing Tian, Zhonghong Ou, Yifan Zhu, Xiao Zhang, Qiankun Ha, Haoran Luo, and Meina Song. TSVC: tripartite learning with semantic variation consistency for robust image-text retrieval. In *AAAI*, pages 19269–19277, 2025.

[Ma *et al.*, 2024] Xinran Ma, Mouxing Yang, Yunfan Li, Peng Hu, Jiancheng Lv, and Xi Peng. Cross-modal retrieval with noisy correspondence via consistency refining and mining. *IEEE Trans. Image Process.*, 33:2587–2598, 2024.

[Qin *et al.*, 2022] Yang Qin, Dezhong Peng, Xi Peng, Xu Wang, and Peng Hu. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *MM '22: The 30th ACM International Conference on Multimedia*, pages 4948–4956. ACM, 2022.

[Qin *et al.*, 2023] Yang Qin, Yuan Sun, Dezhong Peng, Joey Tianyi Zhou, Xi Peng, and Peng Hu. Cross-modal active complementary learning with self-refining correspondence. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023*, 2023.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763, 2021.

[Sharma *et al.*, 2018] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2556–2565, 2018.

[van den Oord *et al.*, 2018] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

[Xie *et al.*, 2025a] Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Ji Ao, Dawei Leng, and Yuhui Yin. Fg-clip 2: A bilingual fine-grained vision-language alignment model. *arXiv preprint arXiv:2510.10921*, 2025.

[Xie *et al.*, 2025b] Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Gengshen Zhang, Dawei Leng, and Yuhui Yin. Fg-clip: Fine-grained visual and textual alignment. *arXiv preprint arXiv:2505.05071*, 2025.

[Xie *et al.*, 2025c] Yucheng Xie, Songyue Cai, Tao Tong, Ping Hu, and Xiaofeng Zhu. Seeking proxy point via stable feature space for noisy correspondence learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 2072–2080. ijcai.org, 2025.

[Yang *et al.*, 2023] Shuo Yang, Zhaopan Xu, Kai Wang, Yang You, Hongxun Yao, Tongliang Liu, and Min Xu. Bicro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023*, pages 19883–19892, 2023.

[Yang *et al.*, 2024] Yuchen Yang, Likai Wang, Erkun Yang, and Cheng Deng. Robust noisy correspondence learning with equivariant similarity consistency. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17700–17709. IEEE, 2024.

[Young *et al.*, 2014] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2:67–78, 2014.

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

[Zhang *et al.*, 2024] Xu Zhang, Hao Li, and Mang Ye. Negative pre-aware for noisy cross-modal matching. In *Thirty-Eighth AAAI Conference on Artificial Intelligence*, pages 7341–7349, 2024.

[Zhao *et al.*, 2024] Zihua Zhao, Mengxi Chen, Tianjie Dai, Jiangchao Yao, Bo Han, Ya Zhang, and Yanfeng Wang. Mitigating noisy correspondence by geometrical structure consistency learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27371–27380, 2024.