Consistency-Aware Cross-Modal Noisy Correspondence Rectification via Perturbation-Invariant Modeling

1st Zhongmei Wang
School of Transportation and Electrical Engineering, Hunan University of Technology
Zhuzhou, China
wangzhongmei@hut.edu.cn

2nd Shenao Peng
School of Transportation and Electrical Engineering, Hunan University of Technology
Zhuzhou, China
m23081100009@stu.hut.edu.cn

3rd Jianhua Liu
School of Transportation and Electrical Engineering, Hunan University of Technology
Zhuzhou, China
jhliu@hut.edu.cn

4th Wenxiu AO
School of Transportation and Electrical Engineering, Hunan University of Technology
Zhuzhou, China
m23081100020@stu.hut.edu.cn

***Abstract*—Large-scale image-text data has advanced cross-modal retrieval but inevitably contains noisy correspondences. Most robust training methods rely on the small-loss principle for sample selection, which often discards hard positives semantically matched pairs with large losses caused by semantic complexity. In addition, existing label rectification approaches mainly depend on the model's own predictions, making them vulnerable to error accumulation due to confirmation bias. To mitigate these issues, we propose a Consistency-Aware Noisy Correspondence Rectification (CA-NCR) framework. First, we introduce consistency under perturbations as a second criterion and construct a 2D-Gaussian Mixture Model (2D-GMM) in the loss-consistency space, which effectively separates hard positives from true noise. Second, we design an anchor-guided rectification scheme that utilizes a high-precision clean set as an external semantic anchor to reconstruct robust soft labels through bi-directional geometric consistency, thereby breaking the self-confirmation loop. Extensive experiments on Flickr30K, MS-COCO, and CC152K demonstrate that CA-NCR significantly outperforms state-of-the-art methods under both varying noise ratios and real world noise scenarios.**

***Index Terms*—Cross-Modal Retrieval, Mismatched Pairs, Noisy Labels, Multimodal Learning**

I. Introduction

Cross-modal matching (image-text matching) is a core task that connects vision and language, supporting a wide range of applications such as cross-modal retrieval [1]–[4], visual question answering [5], and image captioning [6]. As deep learning increasingly relies on large-scale data, web-crawled image-text pairs have become a major source for training. However, such low-cost data inevitably contains noisy correspondences, i.e., semantically mismatched pairs. Prior studies show that deep neural networks exhibit a memorization effect [7] and tend to overfit to noise in later training stages, which degrades generalization. Therefore, learning robustly under noisy supervision remains a key challenge.

Recent methods for noisy correspondence learning, including NCR [8] and CTPR [9], exploit the observation that DNNs learn simple patterns first [10]. They select samples based on the "small-loss" principle or fit Gaussian Mixture Models (GMM) over losses [11], [12] to partition the training set into clean and noisy subsets. Despite their effectiveness, two issues persist. First, using loss as the sole selection criterion is often too coarse: hard positives—genuinely aligned pairs with complex visual content or semantically challenging descriptions—can incur high losses similar to true noise and are thus mistakenly discarded, which hinders the learning of fine-grained semantics. Second, many label-correction approaches [8], [13] rely on the model's own predicted probabilities as soft labels. This self-relabeling is susceptible to confirmation bias: once early predictions are biased by noise, subsequent training may reinforce these errors and lead to error accumulation.

To address these limitations, we propose a Consistency-Aware Noisy Correspondence Rectification (CA-NCR) framework. For sample selection, we introduce perturbation-based semantic consistency as a second dimension. Motivated by the fact that true semantic alignment tends to be stable under mild perturbations [14], [15], we construct a 2D-GMM in the loss-consistency space, which effectively separates high-loss hard positives (high consistency) from true noise (low consistency), enabling more reliable sample partitioning. For soft-label correction, we move beyond conventional self-prediction. Inspired by [16], we design an anchor-guided rectification mechanism

This work was supported in part by the National Natural Science Foundation of China under Grants 62573435 and 52272347, in part by the Young Scientists Fund of the National Natural Science Foundation of China under Grant 62303178, in part by the Key Project of the Hunan Provincial Department of Education under Grant 25A0423, and in part by the Natural Science Foundation of Hunan Province under Grant 2026JJ80046.

that utilizes a high-precision clean set as external semantic anchors. By comparing each noisy sample to the anchor set in the embedding space (in both image-to-text and text-to-image directions), this design breaks the self-confirmation loop and better leverages informative signals from noisy pairs.

In summary, the main contributions of this work are as follows:

- We propose a novel consistency-aware data partitioning module that constructs a 2D-Gaussian Mixture Model (2D-GMM) by jointly modeling loss values and perturbation invariance. This approach effectively distinguishes hard positives from true noise in the geometric space, addressing the limitations of coarse, single-dimensional loss-based selection.
- We design an anchor-guided bi-directional rectification mechanism to correct noisy labels. By leveraging high-purity clean samples as external semantic anchors, this mechanism generates robust soft labels based on geometric consistency, fundamentally avoiding the error accumulation caused by confirmation bias in self-prediction methods.
- Extensive experiments on three benchmark datasets (Flickr30K, MS-COCO and CC152K) demonstrate that our proposed CA-NCR significantly outperforms state-of-the-art methods, showing superior robustness and accuracy particularly in scenarios with high noise ratios and real-world noise.

II. Related Work

A. Image-Text Matching

The goal of image–text matching is to learn a shared embedding space where semantic similarity across modalities can be measured, enabling downstream tasks such as cross-modal retrieval. Early studies [17], [18] typically adopted convolutional neural networks (CNNs) for visual feature extraction and recurrent encoders (e.g., RNN/GRU) for text representation. With the advancement of deep learning, fine-grained interaction between image regions and textual tokens has become a dominant paradigm. For instance, VSRN [19] leverages a graph convolutional network (GCN) to model semantic relations among image regions, while Transformer-based language encoders such as BERT [20] provide contextualized representations that strengthen cross-modal alignment.

In terms of learning objectives, metric learning with ranking-based losses is widely used. The triplet ranking loss remains a standard choice, and VSE++ [21] further improves discriminability with online hard-negative mining, emphasizing challenging negatives during training. To alleviate sensitivity to a fixed margin, several works explore adaptive-margin formulations and geometric constraints [22]. However, most image–text matching methods implicitly assume that training pairs are correctly matched; their performance can degrade significantly when noisy correspondences exist in web-crawled data.

B. Learning with Noisy Correspondence

Learning with noisy supervision has been extensively studied in image classification, with representative strategies including loss re-weighting, label correction, and Co-teaching [13]. Different from noisy class labels, noisy correspondence in cross-modal matching arises from incorrectly paired image–text samples, which makes the noise structure more implicit and harder to identify.

Existing solutions for noisy correspondence learning generally follow two directions: denoising/partitioning and label rectification. NCR [8] models the distribution of training losses with a Gaussian Mixture Model (GMM) to split data into clean and noisy subsets, and then performs rectification using model predictions. Building upon this line, DECL [23] introduces evidential learning to estimate uncertainty for more reliable noise handling, while CTPR [9] proposes a tri-partition scheme to better treat borderline samples close to the decision boundary. BiCro [16] further exploits bidirectional cross-modal consistency for label rectification, improving robustness under noisy supervision. Nevertheless, two limitations remain. First, sample selection based on one-dimensional loss statistics is often insufficient to distinguish high-loss hard positives from true noisy pairs, which may discard informative aligned samples. Second, many rectification methods heavily rely on self-predicted probabilities, making them vulnerable to confirmation bias and error accumulation. These observations motivate our consistency-aware selection and anchor-guided bi-directional rectification framework.

III. Methodology

We describe CA-NCR for robust image-text matching. The framework consists of four stages: (1) Warm-up: we first train the model for a few epochs to obtain a reasonable cross-modal alignment (Sec. III-A). (2) Consistency-Aware Partitioning: we fit a 2D GMM over training loss and perturbation-based semantic consistency to split pairs into clean, hard, and noisy subsets (Sec. III-B). (3) Anchor-Guided Rectification: we use a high-confidence clean set as semantic anchors to generate soft labels for the remaining pairs via bidirectional consistency in the embedding space (Sec. III-C). (4) Joint Optimization: we leverage the rectified soft labels to construct an adaptive-margin objective for robust training (Sec. III-D).

A. Warm-up

The primary goal of learning from noisy correspondence is to partition the training set into different subsets based on semantic relevance. To achieve this, it is essential to first train the Deep Neural Networks (DNNs) to acquire basic feature discrimination capabilities. This subsection demonstrates how the warm-up module achieves initial network convergence.

Given a dataset $\mathcal{D} = \{(I_i, T_i)\}_{i=1}^{N}$ containing N image-text pairs, we employ a standard dual-stream architecture for feature extraction. Specifically, the image encoder (based on the region feature extractor of Faster-RCNN [24]–[26]) and the text encoder (BERT [20]) map the raw image I_i and text T_i into d-dimensional feature vectors v_i and t_i, respectively.

Prior studies [13] indicate that relying on a single model for data partitioning is prone to error accumulation. To avoid this issue, we introduce a co-teaching paradigm, maintaining two networks, A and B, with identical structures but different initializations. The two networks parallelly compute the cosine similarity $S(I_i, T_i)$ between image and text features to measure the semantic relevance across modalities.

During the warm-up stage, networks A and B are trained on the training set by minimizing the Triplet Ranking Loss:

$$\mathcal{L}_{\text{warm}}(I_i, T_i) = \sum_{j \neq i} \Big(\big[\alpha - S(I_i, T_i) + S(I_i, T_j)\big]_+ + \big[\alpha - S(I_i, T_i) + S(I_j, T_i)\big]_+ \Big) \tag{1}$$

where α is the margin and $[x]_+ = \max(x, 0)$. After several warm-up epochs on $\mathcal{D}$, the model provides a stable starting point for the subsequent consistency-aware partitioning.

B. Consistency-Aware Data Partitioning

Although the warm-up stage endows the model with initial discriminative power, relying solely on the one-dimensional distribution of warm-up loss for sample partitioning has inherent limitations. Specifically, hard positives—matched pairs with complex visual content or abstract textual descriptions—often exhibit high loss values, making them prone to be misclassified as noise by loss-based mixture models. To overcome this, inspired by consistency regularization [14] and view invariance [15], we introduce perturbation invariance as a second discriminative dimension, distinguishing hard samples from true noise based on the assumption that genuine semantic associations remain invariant under weak perturbations.

1) Cross-View Consistency Metric: We first construct cross-view augmentations to measure sample stability. To prevent semantic drift caused by aggressive transformations, we employ a weak augmentation strategy. Given the i-th image-text pair (I_i, T_i), we generate augmented views $\tilde{I}_i$ and $\tilde{T}_i$ using mild random cropping and synonym replacement, respectively. We then compute the set of cross-similarities between original and augmented views:

$$\mathcal{S}_i = \{S(I_i, T_i),\, S(\tilde{I}_i, T_i),\, S(I_i, \tilde{T}_i),\, S(\tilde{I}_i, \tilde{T}_i)\}. \tag{2}$$

The consistency score c_i is defined as the stability of the similarity distribution within this set:

$$c_i = \exp\left(-\frac{\sigma_i^2}{\tau}\right), \tag{3}$$

where σ_i^2 is the variance of the similarity distribution in $\mathcal{S}_i$, and τ is a temperature parameter controlling sensitivity (set to $\tau = 1$ for simplicity). For hard positives, despite their potentially low raw similarity (resulting in high loss), their semantic structure is robust, leading to small fluctuations in cross-view similarity and thus a high consistency score c_i. Conversely, the matching of noisy pairs is typically random and sensitive to perturbations, resulting in low consistency.

2) 2D-Gaussian Mixture Modeling (2D-GMM): To jointly leverage loss and consistency information for precise partitioning, we model the i-th sample as a two-dimensional observation vector $\mathbf{u}_i = [\ell_i,\, 1 - c_i]$, where ℓ_i is the normalized warm-up loss. Compared to traditional 1D modeling, the 2D space utilizes geometric structure to effectively separate hard samples (high loss, high consistency) from noisy samples (high loss, low consistency). We employ a two-component 2D-GMM [27] to fit the distribution of all samples. The probability density function for the i-th sample is defined as:

$$p(\mathbf{u}_i) = \sum_{k=1}^{K} \lambda_k \mathcal{N}(\mathbf{u}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\,, \tag{4}$$

where $K = 2$ denotes the number of Gaussian components (corresponding to clean and noisy sets, respectively). λ_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$ represent the mixture coefficient, mean vector, and covariance matrix of the k-th component, respectively. These parameters are estimated via the Expectation-Maximization (EM) algorithm [28]. Based on the memorization effect of DNNs [29], clean samples exhibit faster loss decay during warm-up [7], [30]. Using Bayes' theorem, we compute the posterior probability w_i that the i-th sample belongs to the "clean" component (the component with smaller loss mean and higher consistency mean):

$$w_i = p(k^* \mid \mathbf{u}_i) = \frac{\lambda_{k^*} \mathcal{N}(\mathbf{u}_i \mid \boldsymbol{\mu}_{k^*}, \boldsymbol{\Sigma}_{k^*})}{p(\mathbf{u}_i)}. \tag{5}$$

3) Data Partitioning Strategy: To distill high-confidence pairs from the noisy dataset, we adopt the co-teaching paradigm [9]. Specifically, for network A, we first divide the dataset into an initial clean set $\mathcal{S}_c^A$ and noisy set $\mathcal{S}_n^A$ based on its posterior probability w_i^A and a threshold β:

$$\begin{cases} \mathcal{S}_c^A = \{(I_i, T_i) \mid w_i^A \geq \beta\}, \\ \mathcal{S}_n^A = \{(I_i, T_i) \mid w_i^A < \beta\}. \end{cases} \tag{6}$$

Similarly, for network B, we obtain $\mathcal{S}_c^B$ and $\mathcal{S}_n^B$. Subsequently, we leverage the consensus between the two networks to define the final shared partitions. The data partitioning is formulated as:

$$\begin{cases} \mathcal{S}_c^{AB} = \mathcal{S}_c^A \cap \mathcal{S}_c^B, \\ \mathcal{S}_n^{AB} = \mathcal{D} \setminus \mathcal{S}_c^{AB}. \end{cases} \tag{7}$$

Where $\mathcal{S}_c^{AB}$ and $\mathcal{S}_n^{AB}$ represent the high-purity clean set (shared anchors) and the rectification candidate set, respectively. The obtained clean subset $\mathcal{S}_c^{AB}$ serves as highly reliable semantic anchors, laying a solid foundation for the subsequent anchor-guided bidirectional rectification based on geometric consistency.

C. Anchor-Guided Bi-Directional Rectification

To mitigate the confirmation bias introduced by self-prediction–based rectification, we propose an anchor-guided mechanism. Specifically, we use the high-confidence clean set $\mathcal{S}_c^{AB}$ obtained in the previous stage as an external anchor pool (Semantic Anchors) and reconstruct soft labels for samples

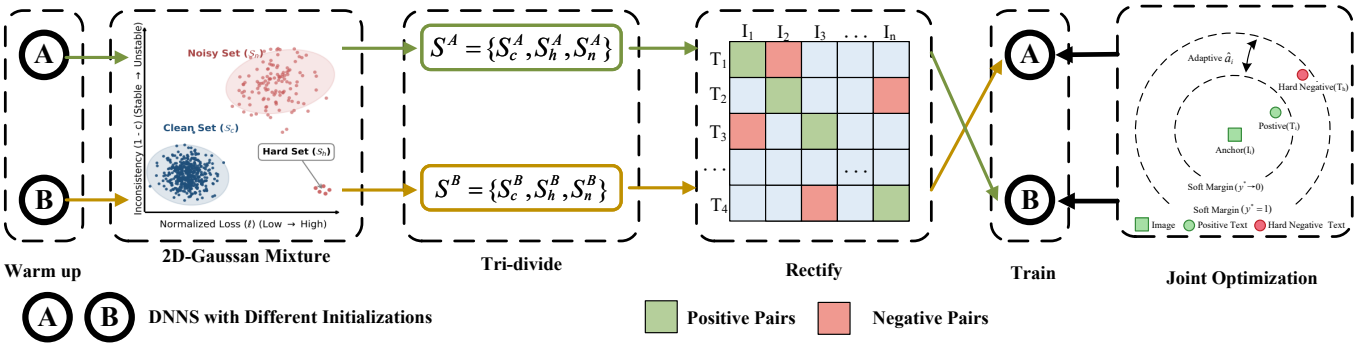

Fig. 1. Structure of the CA-NCR Network.

in the rectification candidate set $\mathcal{S}_n^{AB}$ based on bidirectional geometric consistency.

Inspired by the observations in Yang et al. [16] and Wang et al. [17], we leverage the principle that semantic topology is consistent across modalities: if two images are semantically similar in the feature space, their corresponding textual descriptions should exhibit a similar semantic distance.Based on this premise, for a candidate pair (I_i, T_i), we first retrieve an anchor pair $(I_{\text{anc}}, T_{\text{anc}}) \in \mathcal{S}_c^{AB}$ whose image embedding is nearest to I_i. We then quantify image-to-text consistency by the ratio of distances:

$$\tilde{y}_{i\to t} = \frac{D(I_i, I_{\text{anc}})}{D(T_i, T_{\text{anc}})}. \tag{8}$$

Similarly, taking T_i as the query, we retrieve the nearest anchor text T'_{anc} from the anchor pool and compute the reverse text-to-image consistency:

$$\tilde{y}_{t\to i} = \frac{D(T_i, T'_{\text{anc}})}{D(I_i, I'_{\text{anc}})}. \tag{9}$$

Where $D(\cdot)$ denotes a distance function in the embedding space. The rectified soft label is given by averaging the two directional estimates:

$$y_i^* = \frac{1}{2}\left(\tilde{y}_{i\to t} + \tilde{y}_{t\to i}\right). \tag{10}$$

Because the rectification is grounded on an external anchor pool rather than the model's instantaneous prediction probabilities, it provides a more stable supervisory signal for noisy pairs and facilitates subsequent joint optimization.

D. Joint Optimization

To effectively incorporate the rectified soft labels into feature learning, we introduce an Adaptive Soft Margin mechanism building upon the standard triplet ranking loss. For an arbitrary training pair (I_i, T_i), the loss function is defined as:

$$\begin{aligned}\mathcal{L}_{\text{soft}}(I_i, T_i) = &\left[\hat{\alpha}_i - S(I_i, T_i) + S(I_i, \hat{T}_h)\right]_+ \\ &+ \left[\hat{\alpha}_i - S(I_i, T_i) + S(\hat{I}_h, T_i)\right]_+.\end{aligned} \tag{11}$$

where $[x]_+ = \max(x, 0)$, and $S(\cdot)$ denotes the cosine similarity function in the joint embedding space. The terms $(I_i, \hat{T}_h)$ and $(\hat{I}_h, T_i)$ represent the hardest negative pairs mined within the mini-batch, defined by $\hat{T}_h = \arg\max_{T' \neq T_i} S(I_i, T')$ and $\hat{I}_h = \arg\max_{I' \neq I_i} S(I', T_i)$.

The adaptive margin $\hat{\alpha}_i$ is dynamically determined by the refined label $\hat{y}_i$:

$$\hat{\alpha}_i = \frac{m^{\hat{y}_i} - 1}{m - 1}\alpha, \tag{12}$$

where α is the base margin parameter and m is the curvature hyperparameter [8]. Note that $\hat{y}_i$ unifies the partitioning results (set to 1 for the clean set and the rectified score y_i^* for the candidate set), allowing the margin to smoothly decay for noisy samples to mitigate gradient error accumulation.

IV. Experiments

A. Datasets and Evaluation Metrics

To verify the effectiveness of the proposed method, we evaluate our model on three publicly available datasets: MS-COCO [30], Flickr30K [31], and CC152K [32].

- MS-COCO contains 123,287 images, each accompanied by five corresponding textual descriptions. Following the split in [21], we use 113,287 images for training, 5,000 for validation, and 5,000 for testing.
- Flickr30K consists of 31,783 images collected from Flickr, with each image associated with five captions. Following the data partition in [21], we use 1,000 images for validation, 1,000 for testing, and the remaining 29,000 for training.
- CC152K is a subset curated by NCR from the Conceptual Captions dataset [32], containing approximately 150,000 training pairs. Unlike the two manually annotated datasets mentioned above, CC152K is harvested from the web and inherently contains about 3% to 20% real-world noisy correspondences, making it more representative of practical application scenarios.

Noise Simulation. To comprehensively evaluate the model's robustness under varying noise intensities, we simulate synthetic noise on the training sets of Flickr30K and MS-COCO. Specifically, we randomly shuffle the correspondence between images and texts for a certain proportion of the data to construct training sets containing incorrect labels.

Evaluation Metrics. Image-text retrieval performance is typically evaluated using the Recall@K ($K = 1, 5, 10$) metric,

TABLE I
COMPARISON OF EXPERIMENTAL RESULTS ON THE FLICKR30K AND MS-COCO DATASET

Noise	Methods	Flickr30K							MS-COCO						
		Image→Text			Text→Image				Image→Text			Text→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%	SCAN	58.5	81.0	90.8	35.5	65.0	75.2	406.0	62.2	90.0	96.1	46.2	80.8	89.2	464.5
	VSRN	33.4	59.5	71.3	25.0	47.6	58.6	295.4	61.8	87.3	92.9	50.0	80.3	88.3	460.6
	IMRAM	22.7	54.0	67.8	16.6	41.8	54.1	257.0	69.9	93.6	97.4	55.9	84.4	89.6	490.8
	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
	RCL	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
	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
	GSC-SGR	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
	CA-NCR	**78.5**	**95.6**	**98.0**	**62.0**	**85.0**	**91.0**	**510.1**	80.2	**96.6**	98.2	**65.0**	**91.2**	**96.0**	**527.3**
40%	SCAN	26.0	57.4	71.8	17.8	40.5	51.4	264.9	42.9	74.6	85.1	24.2	52.6	63.8	343.2
	VSRN	2.6	10.3	14.8	3.0	9.3	15.0	55.0	29.8	62.1	76.6	17.1	46.1	60.3	292.0
	IMRAM	5.3	25.4	37.6	5.0	13.5	19.6	106.4	51.8	82.4	90.9	38.4	70.3	78.9	412.7
	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
	RCL	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
	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
	GSC-SGR	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
	CA-NCR	**77.5**	**94.2**	97.0	**62.0**	**83.5**	**89.5**	**503.7**	**79.4**	**96.2**	98.1	**64.3**	**90.2**	**95.7**	**523.9**
60%	SCAN	13.6	36.5	50.3	4.8	13.6	19.8	138.6	29.9	60.9	74.8	0.9	2.4	4.1	173.0
	VSRN	0.8	2.5	5.3	1.2	4.2	6.9	20.9	11.6	34.0	47.5	4.6	16.4	25.9	140.0
	IMRAM	1.5	8.9	17.4	1.9	5.0	7.8	42.5	18.2	51.6	68.0	17.9	43.6	54.6	253.9
	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
	RCL	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
	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
	GSC-SGR	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
	CA-NCR	**72.0**	91.0	95.0	**55.0**	79.5	86.6	**479.1**	77.0	**95.5**	97.5	**61.0**	87.0	94.0	**512.0**

denoted as R@1, R@5, and R@10. Recall@K represents the percentage of queries for which the correct ground-truth match is found within the top-K retrieved items; higher values indicate better retrieval accuracy. We calculate the sum of the three recall metrics for both image-to-text and text-to-image retrieval tasks and combine them into an overall evaluation metric, referred to as rSum:

$$rSum = (R@1 + R@5 + R@10)_{\text{image}\to\text{text}} + (R@1 + R@5 + R@10)_{\text{text}\to\text{image}}. \tag{13}$$

B. Implementation Details

For fair comparison, we adopt the standard dual-stream architecture consistent with previous robust learning methods [8]. All experiments are implemented using the PyTorch framework and conducted on NVIDIA Tesla T4 GPUs. During training, the network is optimized via the Adam optimizer with a mini-batch size of 128. We initialize the learning rate at 1×10^{-4}, decaying it by a factor of 0.1 every 10 epochs. Regarding hyperparameters, the partitioning threshold β is set to 0.5. Following the protocol in [8], the warm-up stage lasts for 10 epochs, while the base margin α and the curvature parameter m are fixed at 0.2 and 6, respectively.

C. Comparison Experiments

In this section, we comprehensively compare CA-NCR with two main categories of mainstream methods to verify its effectiveness. The baselines include general cross-modal matching methods (e.g., SCAN [1], VSRN [19], IMRAM [2]) and robust learning methods specifically designed for noisy correspondence (e.g., BiCro [16], RCL [33], L2RM [34], GSC-SGR [35]).

Results on Simulated Datasets. Table I reports the retrieval results on Flickr30K and MS-COCO 1K datasets under different noise ratios (20%, 40%, 60%). By analyzing the experimental data, we can draw the following conclusions:

The performance of general methods drops precipitously under noise interference. For instance, as the noise ratio increases from 20% to 60%, the rSum of SCAN on Flickr30K plummets from 406.0 to 138.6. This confirms that the forceful fitting of noisy samples by deep models severely disrupts the semantic structure of the feature space.

CA-NCR demonstrates superior robustness on the large-scale MS-COCO dataset, achieving significant improvements over existing models. Specifically, under the high noise setting of 60%, CA-NCR achieves an rSum score of 512.0, improving the R@1 metric for text-to-image retrieval by 4.3% compared to the recent best method, GSC-SGR. This significant improvement indicates that the 2D-GMM partitioning strategy effectively prevents the model from overfitting to massive noise.

On the Flickr30K dataset, CA-NCR maintains a lead in the vast majority of metrics. Notably, under the extreme condition of 60% noise, although the rSum of CA-NCR is comparable to that of GSC-SGR, CA-NCR outperforms GSC-SGR by approximately 1.2% on the most critical Image-to-Text R@1 metric. Since R@1 reflects the model's ability to identify the

TABLE II
COMPARISON OF EXPERIMENTAL RESULTS ON CC152K.

Methods	Image→Text			Text→Image			
	R@1	R@5	R@10	R@1	R@5	R@10	rSum
SCAN	30.5	55.3	65.3	26.9	53.0	64.7	295.7
VSRN	32.6	61.3	70.5	32.5	59.4	70.4	326.7
IMRAM	33.1	57.6	68.1	29.0	56.8	67.4	312.0
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
L2RM	**43.0**	67.5	75.7	42.8	68.0	77.2	374.2
GSC-SGR	42.1	68.4	77.7	42.2	67.6	77.1	375.1
CA-NCR	42.7	**68.8**	**78.0**	**43.1**	**67.9**	**77.6**	**378.1**

TABLE III
ABLATION STUDIES ON FLICKR30K WITH 40% NOISE.

Method	Image→Text			Text→Image			rSum
	R@1	R@5	R@10	R@1	R@5	R@10	
w/o Img-Aug	75.8	93.7	96.8	60.9	82.0	86.3	495.5
w/o Txt-Aug	75.5	93.4	96.6	60.6	81.9	86.1	494.1
w/o Consistency	75.3	92.9	96.3	59.6	80.9	85.7	490.7
w/o Anchor-Guided	77.3	94.0	**97.0**	61.6	83.2	88.1	501.2
CA-NCR	**77.5**	**94.2**	**97.0**	**62.0**	**83.5**	**89.5**	**503.7**

"most relevant" match, this result validates that the anchor-guided rectification mechanism can accurately preserve the most critical semantic cues even when the vast majority of data is mismatched.

Results on Real-World Datasets. To further verify the capability of CA-NCR in handling complex real-world noise, we conducted comparative experiments on the CC152K dataset, as shown in Table II. It can be seen that CA-NCR achieves impressive performance, with an overall rSum reaching 378.1. Compared to the strong baseline BiCro and the latest GSC-SGR, our method improves by 7.9 and 3.0 percentage points, respectively.

This superior performance is mainly attributed to the anchor-guided mechanism, which utilizes high-purity clean samples as a geometric reference frame. Unlike previous methods that rely on self-prediction, our strategy effectively "distills" correct semantic alignment relationships from mismatched data, thereby avoiding error accumulation caused by confirmation bias.

D. Ablation Study

To investigate the contribution of each core component in the CA-NCR framework, we conducted detailed ablation studies on the Flickr30K dataset with 40% noise. The results are reported in Table III, and we analyze the impact of consistency-aware partitioning and anchor-guided rectification separately.

Effectiveness of Consistency-Aware Partitioning. To verify the contribution of perturbation invariance to noise identification, we removed image augmentation (w/o Img-Aug), text augmentation (w/o Txt-Aug), and both augmentations (w/o Consistency), respectively. As shown in Table III, removing single-modality augmentation leads to a decrease in rSum by 8.2 and 11.6 points, indicating that semantic stability in both visual and linguistic modalities is crucial for judging sample quality. More critically, when both augmentations are removed (degenerating to 1D-GMM), the performance drops significantly, with rSum falling to 490.7. This result strongly demonstrates that relying solely on loss values cannot effectively distinguish hard positives from noise, whereas introducing cross-view semantic consistency captures the intrinsic robustness of samples under perturbation, thereby significantly improving the precision of data partitioning.

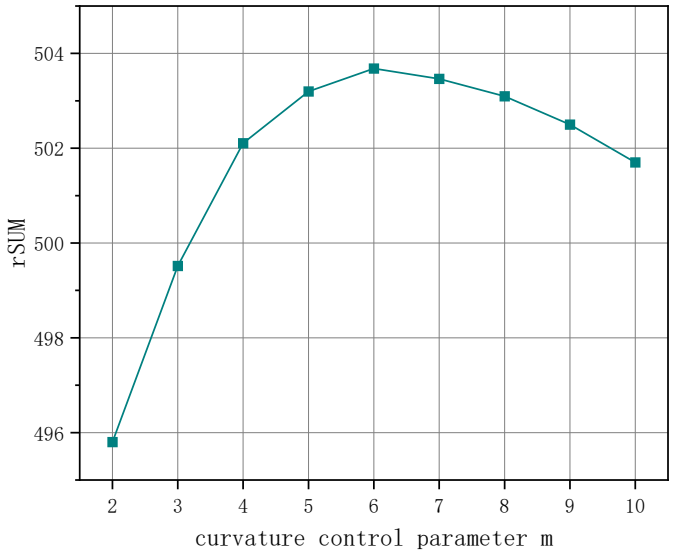

Fig. 2. Rsum Performance under Different Curvature Control Parameter m.

Importance of Anchor-Guided Rectification. To validate the role of the anchor mechanism in label correction, we replaced "Anchor-Guided Rectification" with the traditional "Adaptive Prediction" strategy (i.e., self-prediction). The results show that the rSum of this variant drops by 2.5 points. Although the decline is smaller than that caused by removing the consistency module, there is still a noticeable loss in high-precision retrieval metrics (e.g., R@1). This performance gap confirms that the adaptive prediction strategy is prone to confirmation bias, leading the model to reinforce early errors. In contrast, using high-purity clean samples as external geometric anchors provides more objective and robust supervisory signals, effectively breaking the vicious cycle of error accumulation.

In summary, the full model (CA-NCR) achieves the best performance, proving that 2D-GMM partitioning and anchor-guided rectification are complementary and indispensable, jointly ensuring robust learning under noisy conditions.

Effect of the curvature control parameter m. To investigate the specific impact of the curvature parameter m in the adaptive margin, we conducted sensitivity experiments on the Flickr30K dataset (40% noise ratio). As shown in Fig. 2, the model performance increases first and then decreases as m increases, reaching a peak at $m = 6$. This phenomenon indicates that moderately increasing m imposes stricter margin penalties on low-confidence noisy samples, effectively enhancing noise

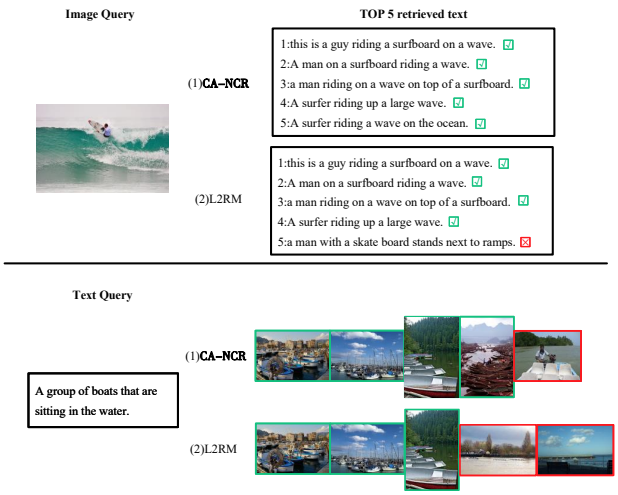

Fig. 3. Visualization of top-5 retrieval results on Flickr30K. Green and red borders denote matched and mismatched results, respectively.

robustness. However, when m is too large, the non-linear relationship approaches saturation, and the computation function tends towards a step function, causing the soft margin to lose its smooth transition capability. Therefore, $m = 6$ maintains the optimal balance between "suppressing true noise" and "preserving hard samples," avoiding the failure of distance constraints caused by excessive non-linearity.

E. Visualization

To provide intuitive insights into the robustness of CA-NCR beyond numerical metrics, we visualize the top-5 retrieval rankings on the Flickr30K dataset in Figure 3. The comparison with the robust baseline L2RM reveals two key observations regarding semantic alignment.

First, in the Image-to-Text task (top row), when querying with an image of a surfer, CA-NCR successfully retrieves five semantically correct captions. In contrast, while L2RM retrieves four correct items, it fails at rank 5 by returning a description involving a "skateboard." This error suggests that the baseline is easily confused by visual similarities (surfboard vs. skateboard) when semantic discriminability is compromised by noise. CA-NCR, however, maintains precise alignment due to our consistency-aware partitioning.

Second, in the Text-to-Image task (bottom row), given the query "A group of boats...", CA-NCR retrieves four highly relevant images depicting clusters of boats. Although the fifth image represents a mismatch (showing a single boat), it remains thematically consistent. Conversely, L2RM retrieves irrelevant scenery images at ranks 4 and 5, indicating a significant semantic drift. These results collectively demonstrate that our anchor-guided rectification mechanism effectively filters out noisy patterns and enforces stricter semantic consistency than existing methods.

V. Conclusion

In this paper, we propose a Consistency-Aware Noisy Correspondence Rectification (CA-NCR) framework to address the prevalent challenge of noisy correspondence in cross-modal matching. To overcome the limitations of relying solely on loss distribution for sample selection, we introduce perturbation invariance as a second discriminative dimension and construct a 2D-Gaussian Mixture Model (2D-GMM). This approach accurately distinguishes high-loss hard positives from true noise within the geometric feature space. Furthermore, to address the confirmation bias inherent in traditional adaptive prediction strategies, we design an anchor-guided rectification mechanism. By discarding the self-referential prediction manner and instead leveraging filtered high-purity clean samples as an external reference, this mechanism reconstructs objective and robust soft labels for noisy data based on bidirectional geometric consistency. Extensive experiments on Flickr30K, MS-COCO, and CC152K datasets demonstrate that CA-NCR not only significantly outperforms existing state-of-the-art methods but also exhibits superior robustness in scenarios with high noise ratios and complex real-world noise.

References

[1] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[2] H. Chen, G. Ding, X. Liu, Z. Lin, J. Liu, and J. 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 (CVPR)*, June 2020.

[3] K. Wen and X. Gu, "Dual semantic relationship attention network for image-text matching," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–7.

[4] K. Zhou, Y. Xing, Y. Geng, and J. Zhao, "Cross-modal retrieval based on attention embedded variational auto-encoder," in *2024 International Joint Conference on Neural Networks (IJCNN)*, 2024, pp. 1–7.

[5] J. Yu, W. Zhang, Y. Lu, Z. Qin, Y. Hu, J. Tan, and Q. Wu, "Reasoning on the relation: Enhancing visual representation for visual question answering and cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 22, no. 12, pp. 3196–3209, 2020.

[6] H. Ben, Y. Pan, Y. Li, T. Yao, R. Hong, M. Wang, and T. Mei, "Unpaired image captioning with semantic-constrained self-learning," *IEEE Transactions on Multimedia*, vol. 24, pp. 904–916, 2022.

[7] D. Arpit, S. Jastrzębski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, and S. Lacoste-Julien, "A closer look at memorization in deep networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 233–242.

[8] Z. Huang, G. Niu, X. Liu, W. Ding, X. Xiao, H. Wu, and X. Peng, "Learning with noisy correspondence for cross-modal matching," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 29 406–29 419.

[9] Z. Feng, Z. Zeng, C. Guo, Z. Li, and L. Hu, "Learning from noisy correspondence with tri-partition for cross-modal matching," *IEEE Transactions on Multimedia*, vol. 26, pp. 3884–3896, 2024.

[10] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Commun. ACM*, vol. 64, no. 3, p. 107–115, Feb. 2021.

[11] J. Li, R. Socher, and S. C. H. Hoi, "Dividemix: Learning with noisy labels as semi-supervised learning," 2020.

[12] E. Yang, D. Yao, T. Liu, and C. Deng, "Mutual quantization for cross-modal search with noisy labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 7551–7560.

[13] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman,

N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018.

[14] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 596–608.

[15] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 1597–1607.

[16] S. Yang, Z. Xu, K. Wang, Y. You, H. Yao, T. Liu, and M. Xu, "Bicro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 19 883–19 892.

[17] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[18] L. Wang, Y. Li, J. Huang, and S. Lazebnik, "Learning two-branch neural networks for image-text matching tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 394–407, 2019.

[19] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Visual semantic reasoning for image-text matching," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[21] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improving visual-semantic embeddings with hard negatives," 2018.

[22] A. F. Biten, A. Mafla, L. Gómez, and D. Karatzas, "Is an image worth five sentences? a new look into semantics for image-text matching," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2022, pp. 1391–1400.

[23] Y. Qin, D. Peng, X. Peng, X. Wang, and P. Hu, "Deep evidential learning with noisy correspondence for cross-modal retrieval," in *Proceedings of the 30th ACM International Conference on Multimedia*, ser. MM '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 4948–4956.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015.

[26] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[27] E. Arazo, D. Ortego, P. Albert, N. O'Connor, and K. Mcguinness, "Unsupervised label noise modeling and loss correction," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 312–321.

[28] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[29] X. Xia, T. Liu, B. Han, C. Gong, N. Wang, Z. Ge, and Y. Chang, "Robust early-learning: Hindering the memorization of noisy labels," in *International Conference on Learning Representations*, 2021.

[30] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755.

[31] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[32] P. Sharma, N. Ding, S. Goodman, and R. 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 (Volume 1: Long Papers)*, I. Gurevych and Y. Miyao, Eds. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2556–2565.

[33] P. Hu, Z. Huang, D. Peng, X. Wang, and X. Peng, "Cross-modal retrieval with partially mismatched pairs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 9595–9610, 2023.

[34] H. Han, Q. Zheng, G. Dai, M. Luo, and J. Wang, "Learning to rematch mismatched pairs for robust cross-modal retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 26 679–26 688.

[35] Z. Zhao, M. Chen, T. Dai, J. Yao, B. Han, Y. Zhang, and Y. Wang, "Mitigating noisy correspondence by geometrical structure consistency learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 27 381–27 390.