HaNa: Hardness and Noise-Aware Robust Cross-modal Retrieval

Fangming Zhong, Haiquan Yu, Cun Zhu, Suhua Zhang*

School of Software, Dalian University of Technology
fmzhong@dlut.edu.cn, {yhq, zhucun, suhua_zhang}@mail.dlut.edu.cn

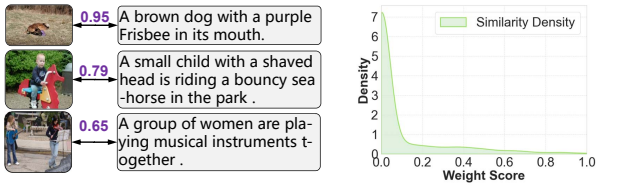
Abstract

Noisy correspondence in cross-modal retrieval introduces significant challenges due to its inherent difficulty in identification and correction. Although existing methods attempt to minimize the influence of noisy samples by the weighting mechanism, these methods still struggle with performance degradation under increasing noise levels. Specifically, the clean samples are assigned the same weight of 1, which ignores the sample hardness. In addition, the weights for noisy samples are approaching 0, leading to the overlook of sample diversity. To address these issues, we propose a Hardness and Noise-aware (HaNa) robust cross-modal retrieval method. HaNa introduces a momentum-based reweighting mechanism to adaptively balance learning difficulty across clean samples, avoiding overfitting risk and accumulative partitioning bias. Moreover, HaNa addresses the limitation that weights for noisy data are approaching 0 from a new perspective to fully employ the diversity of samples to further improve its generalization. It employs an Asymmetric Noise-aware Regularization Loss (ANRL) to treat identified noisy data as negative samples for optimization. Extensive experiments demonstrate that HaNa achieves superior matching accuracy and stability, especially in high-noise scenarios, outperforming state-of-the-art methods.

Introduction

Recently, image-text matching (Anderson et al. 2018; Li et al. 2019; Sun et al. 2024) aiming to establish semantic connections between images and texts for cross-modal retrieval, has gained much attention due to the widespread existence of multimodal data. Different from traditional image-text matching that assumes all training data are correctly matched, which is expensive in practice, recent researches have drawn considerable interests in the cross-modal retrieval with noisy correspondence, i.e., training data contains wrongly matched image-text pairs. It poses a critical yet challenging problem in robust cross-modal retrieval.

Numerous efforts have been devoted to tackling such issue (Huang et al. 2021; Yang et al. 2023; Qin et al. 2022; Zha et al. 2025a; Liu et al. 2024; Pu et al. 2025). While most of them share a common goal, i.e., distinguishing noisy data from clean data and then devising robust loss functions to effectively learn from noisy data, thereby enhancing model performance and generalization in complex scenarios. For example, NCR (Huang et al. 2021) is a pioneer work in exploring cross-modal retrieval with noisy correspondence by adopting a collaborative teaching scheme. NCR divides samples into clean and noisy subsets based on the memorization effect of neural networks and proposes a novel triplet loss by recasting the rectified labels as the soft margin. Similarly, BiCro (Yang et al. 2023) leverages the assumption that similar images should have similar textual descriptions. Most of the existing methods adopt the similar triplet loss with soft margin as in NCR to prevent deep neural networks from overfitting to noisy data. However, the soft margin acts symmetrically on both positive and negative pairs. Such indiscriminate penalization of both noisy positive and negative samples undermines the learning process, ultimately degrading model performance.



(a) Similarities of positive samples (b) Distribution of weights for noisy samples

Figure 1: Observations from existing methods.

Subsequent research has focused on loss reweighting for noisy pairs to mitigate their influence during optimization. For instance, NPC (Zhang, Li, and Ye 2024) proposes a negative pre-aware learning paradigm, which adaptively assesses the potential negative impact of each sample before model learning and assigns lower confidence weights to samples with high negative impact. In (Liu et al. 2024), a novel self-drop and dual-weight (SDD) approach is introduced, which adopts a dual-weight strategy to ensure that the model focuses more on significant samples while appropriately leveraging vague samples. Despite the success of these methods, they still struggle with performance degradation under increasing noise levels. Specifically, the uniform weighting of clean positive samples (fixed weight = 1) disregards intrinsic variations in learning difficulty, i.e., the hardness varies across different positive samples as shown in

*Corresponding author.

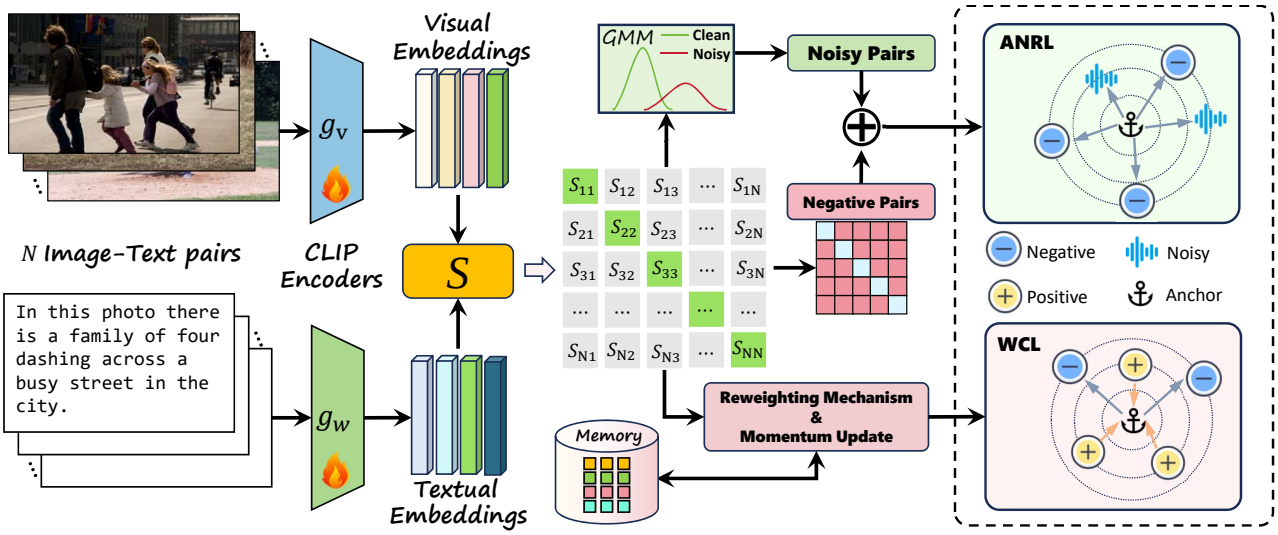


Figure 2: Illustration of the training pipeline of HaNa. S is the similarity matrix computed via visual embeddings and textual embeddings. Based on S, a reweighting mechanism including momentum update is proposed for weighted contrastive learning (WCL). Noisy samples collaborated negative pairs also contribute positively to model optimization by an Asymmetric Noise-aware Regularization Loss (ANRL).

Figure 1(a), which easily leads to the overfitting risk. Moreover, we observe that the weights assigned for the noisy samples are almost approaching zero (Figure 1(b)). As a result, these noisy samples do not contribute effectively to the optimization process, leading to the overlook of sample diversity, as well as the low-generalization problem. Additionally, the selected "clean" positive samples in previous methods cannot be guaranteed to be truly noise-free. Hence, any errors in the partitioning process will cumulatively introduce additional learning bias, which further degrades the model performance.

To address the aforementioned issues, we propose a novel Hardness and Noise-Aware (HaNa) method for robust cross-modal retrieval. As illustrated in Figure 2, the framework consists of two key components: a simple yet effective reweighting mechanism and an asymmetric noise-aware regularization. In contrast to previous methods, HaNa fully explores the hardness of clean samples and introduces an insightful theoretical analysis on the gradients of positive and negative samples in contrastive loss. To this end, HaNa exploits the strong zero-shot discriminability of CLIP to compute initial weights uniformly without partitioning clean and noisy data. However, directly using the raw weights computed from each batch may lead to significant fluctuations during training. Therefore, we further introduce a momentum-based updating mechanism. Our reweighting mechanism provides a superior approximation of the true data distribution, avoiding overfitting risk and accumulative partitioning bias. In addition, HaNa addresses the limitation that weights for noisy data are approaching 0 from a new perspective to fully employ the diversity of samples to further improve its generalization. It proposes the asymmetric noise-aware regularization loss (ANRL) to treat identified noisy data as negative samples for optimization. Specifically, we employ Gaussian Mixture Model (GMM) (Huang et al. 2021) to identify potential noisy samples. ANRL aims to push noisy samples away from positive ones in the representation space, thereby improving the effectiveness of contrastive learning. Though partitioning is used in ANRL, the ratio of noisy data is small among the negative samples. Consequently, the impact of mispartitioned data on negative optimization of our model is significantly less substantial than the effect of partitioning errors on positive optimization in traditional methods. Extensive experiments on MS-COCO, Flickr30K, and CC120K datasets demonstrate that our method achieves superior performance over state-of-the-art approaches and exhibits strong robustness.

Our main contributions can be summarized as follows:

- A novel hardness and noise-aware robust cross-modal retrieval method is proposed that for the first time integrates reweighting clean samples and regularizing noisy and negative samples.
- A sample reweighting mechanism based on contrastive loss analysis and momentum-based smoothing is introduced, which explores and analyzes the hardness of positive samples, alleviating overfitting risk and cumulative partitioning bias.
- We further introduce an asymmetric noise-aware regularization loss, explicitly modeling noisy samples from the negative perspective, which takes full advantage of the diversity of samples to improve generalization, rather than dropping out noisy samples.

Related Works

Image-text Matching

Most of the previous image-text matching approaches can be broadly categorized into two types: global-level matching

(Zhang et al. 2024; Li et al. 2025, 2022) methods and local-level matching (Diao et al. 2021; Pan, Wu, and Zhang 2023; Qu et al. 2021) methods. Global-level methods focus on capturing and aligning the overall semantic representations of images and texts. For instance, GPO (Chen et al. 2021) enhances extraction via adaptive pooling, and CORA (Pham et al. 2024) improves semantic structure capture by building text-based scene graphs. On the local level, methods prioritize fine-grained alignments between image regions and textual words to achieve precise semantic correspondence. Among these, CHAN (Pan, Wu, and Zhang 2023) utilizes a simple but effective hard alignment mechanism, directly associating the most relevant image regions with specific words to enable detailed semantic matching.

In recent years, CLIP (Radford et al. 2021) has gained widespread attention in various cross-modal matching tasks due to its strong zero-shot learning capability. However, when fine-tuned for downstream tasks, its performance remains sensitive to noisy data, making it difficult to achieve satisfactory results in the presence of noise.

Noisy Correspondence Learning

Different from traditional image-text matching that assumes all training data are correctly matched, recent researches have drawn considerable interests in the cross-modal retrieval with noisy correspondence, i.e., training data contains wrongly matched image-text pairs.

Among the pioneering explorations, NCR (Huang et al. 2021) utilizes a co-teaching strategy, capitalizing on the DNN memorization effect to identify high per-sample loss as noises. These samples are subsequently trained with estimated soft labels. Similarly, BiCro (Yang et al. 2023) addresses the noisy correspondence problem by assuming that semantically similar images should align with semantically similar textual descriptions. Further approaches, such as CREAM (Ma et al. 2024) and CTPR (Feng et al. 2024), employ dataset partitioning into three subsets (clean, noisy, and hard) to better leverage image-text pairs exhibiting partial semantic relevance. Recent work includes L2RM (Han et al. 2024), which leverages optimal transport to explore underlying semantic similarities among unpaired samples, representing the inaugural application of optimal transport theory to noisy correspondence tasks. Additionally, NPC (Zhang, Li, and Ye 2024) introduces a negative sample pre-aware and reweighting mechanism designed to alleviate performance instability stemming from noisy pair learning. NPC first employs the pretrained CLIP as a backbone, making a breakthrough for this task. Furthermore, GSC (Zhao et al. 2024) distinguishes noisy samples by analyzing intra-modality geometric structure consistency discrepancies.

Despite these advances, existing methods still suffer from several limitations, such as uniform treatment of clean samples and excessive discarding of noisy samples. To overcome these challenges, we propose a novel hardness and noise-aware robust cross-modal retrieval (HaNa). Different from previous methods, our HaNa not only mitigates the adverse effects of noisy data but also takes into account the hardness of positive samples and maximizes the utility of noisy samples.

Methodology

Problem Definition

Here, we formally define the image-text matching task with noisy correspondence and present fundamental formulations with symbol interpretations. Given a dataset $\mathcal{D} = (I_i, T_i, y_i)_{i=1}^{N}$ where N denotes the dataset size, (I_i, T_i) represents the i-th image-text pair, and $y_i \in 0, 1$ is the ground-truth correspondence label. This label explicitly indicates whether the pair exhibits positive semantic correspondence ($y_i = 1$) or noisy correspondence ($y_i = 0$). The objective of image-text matching is to project multimodal embeddings into a shared latent space and measure their alignment, for which cosine similarity serves as the prevalent metric as formulated in Eq. (1):

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

where $f(\cdot)$ and $g(\cdot)$ denote feature extractors for image and text modalities, respectively.

Reweighting Mechanism for Clean Samples

Mathematical Proof Firstly, following the same theoretical approach as RDE (Qin et al. 2024a), we outline the methodology for assigning adaptive weights to clean samples. We follow the settings in (Zhang, Li, and Ye 2024) that leverages CLIP as the image-text encoders and employs a contrastive loss function as the objective to train our model. The initial objective can be formulated as:

$$CE(I_i, T_i) = -\frac{1}{B}\sum_{i=1}^{B} \log\left(\frac{\exp(S(I_i, T_i)/\tau)}{\sum_{j=1}^{B}\exp(S(I_i, T_j)/\tau)}\right), \tag{2}$$

$$\mathcal{L}_{InfoNCE}(I_i, T_i) = CE(I_i, T_i) + CE(T_i, I_i), \tag{3}$$

where B refers to batch size and τ is the temperature coefficient. As can be seen from Eqs. (2) and (3), the optimization is to increase the similarity between positive samples while decreasing the similarity between negative samples.

However, with the noisy correspondence, the model will be influenced to learn incorrect pairing information, degrading performance. To address this issue, several existing methods mitigate the impact of noisy samples by assigning a weighting factor to them, thereby reducing their gradient contributions during backpropagation, which can be formulated as follows:

$$RCE(I_i, T_i) = -\frac{1}{B}\sum_{i=1}^{B} w_i \log\left(\frac{\exp(S(I_i, T_i)/\tau)}{\sum_{j=1}^{B}\exp(S(I_i, T_j)/\tau)}\right), \tag{4}$$

where w_i is the optimization weight. For clean samples, w_i is set as 1, while for noisy data, w_i is computed by the performance change of the model after the sample (I_i, T_i) is trained.

Nevertheless, such a weighting mechanism still exhibits a limitation. In particular, even within clean samples, intrinsic variations in learning difficulty persist, allowing them to be stratified into easy and hard samples. More critically, enforcing uniform learning intensity across both hard

and easy samples may amplify overfitting risks, particularly because hard samples frequently contain ambiguous, boundary-related patterns. This argument can be supported by further theoretical analysis on the gradients of positive and negative samples in RCE loss, which are also stated as follows,

$$\nabla_{T_i} RCE = \frac{1}{B\tau}\sum_{i=1}^{B} w_i \left(\frac{\exp(S(I_i,T_i)/\tau)}{\sum_{k=1}^{B}\exp(S(I_i,T_k)/\tau)} - 1 \right) \cdot \frac{\partial S(I_i,T_i)}{\partial T_i}, \quad (5)$$

$$\nabla_{T_j} RCE = \frac{1}{B\tau}\sum_{i=1}^{B} w_i \frac{\exp(S(I_i,T_j)/\tau)}{\sum_{k=1}^{B}\exp(S(I_i,T_k)/\tau)} \cdot \frac{\partial S(I_i,T_j)}{\partial T_j}. \quad (6)$$

Eqs. (5) and (6) delineate the gradient formulations of the RCE loss for positive and negative text samples, respectively. The gradients for image samples exhibit an analogous structure and are omitted herein for conciseness. It is worth noting that if the weight of clean samples is fixed at 1, their gradient contributions, including that of their corresponding negative pairs, are entirely governed by the softmax-derived probabilities calculated in the current epoch. For hard samples, the similarity scores between positive and negative pairs are often nearly identical, resulting in comparable gradient contributions from both, which undermines the model's discriminative learning ability.

Therefore, it is necessary to take into account the hardness of positive samples. Though some existing studies focus on explicitly detecting hard samples, they suffer from the increasing model complexity and computational cost. To this end, we introduce a momentum-based weight prediction approach, which is detailed as follows.

Momentum-based Weights Update Leveraging the strong zero-shot capability of CLIP, we propose a simple yet effective way to compute the weights in Eq. (4), which can be stated as follows:

$$\tilde{w}_i = \left(\frac{\exp(S(I_i,T_i)/\tau)}{\sum_{j=1}^{B}\exp(S(I_i,T_j)/\tau)} + \frac{\exp(S(I_i,T_i)/\tau)}{\sum_{j=1}^{B}\exp(S(I_j,T_i)/\tau)}\right)/2. \quad (7)$$

As can be seen, the weights are derived from the mean of bidirectional probabilities. This symmetric mechanism integrates confidence information from both directions, effectively reducing the bias inherent in unidirectional predictions. Furthermore, compared with methods that partition the sample set into clean and noisy subsets and process them independently, our reweighting mechanism via Eq. (7) provides a superior approximation of the true data distribution. This advantage arises because accurate subset partitioning is often infeasible in practice, and any errors in the partitioning process will cumulatively introduce additional learning bias.

However, directly using the raw weights computed from each batch may lead to significant fluctuations during training. This is because of the inherent instability in data distribution and model predictions across batches. To address this issue, we further introduce a momentum update. Specifically, the weights obtained from the current batch are stored and updated in a smoothed manner through momentum, ensuring more stable gradient updates and reducing training instability. This is particularly beneficial for achieving robust convergence in noisy environments. The momentum-based weights update can be formulated as,

$$w_i^{(t)} = \alpha \cdot w_i^{(t-1)} + (1-\alpha)\cdot \tilde{w}_i^{(t)}, \quad (8)$$

where α is the parameter balancing current and historical weight values, t is the training epoch. Here, we employ a memory bank as shown in Figure 2 to store the weights. Furthermore, during both the weight computation and momentum update phases, we freeze the model parameters, enabling score generation without optimization. In doing so, it further enhances the stability and reliability of the weight estimation process.

Finally, with the reweighting mechanism, a new Weighted Contrastive Learning (WCL) loss $\mathcal{L}_{WCL}$ similar to Eq. (4) is introduced.

Asymmetric Noise-Aware Regularization Loss

We argue that noisy samples can also contribute positively to model optimization when utilized appropriately. From the previous methods, such as SDD (Liu et al. 2024), we observe that the weights assigned to noisy samples are typically very small, leading to negligible gradient contributions and nearly vanishing backpropagation signals. It indicates that noisy samples are almost discarded at training. This ignores the essential diversity of samples. In addition, if noisy samples are mistakenly categorized as clean samples, they may interfere with the learning process of genuine negative instances. Moreover, contrastive learning frameworks often apply a uniform treatment to all negative samples regardless of their semantic relevance or difficulty, which is a limitation commonly referred to as the "one-size-fits-all" problem.

Inspired by Asymmetric Loss (Wang et al. 2025) and Complementary Learning (Qin et al. 2024b), we propose a novel asymmetric loss function that treats detected noisy samples as negatives and explicitly pushes them away from target samples. A cosine similarity-guided weighting mechanism is introduced during the optimization process to adaptively control the influence of these noisy samples, thereby enhancing the overall optimization effectiveness and improving model robustness. First, we partition the training set into a clean subset and a noisy subset. Following NCR, we model the loss distribution of each sample in the training set using a two-component Gaussian Mixture Model (GMM). The loss is computed using the original contrastive loss function.

$$p(l|\theta) = \sum_{k=1}^{K} \alpha_k \phi(l|\theta_k), \quad (9)$$

where α_k denotes the mixing coefficient of the k^{th} component, and $\phi(l|\theta_k)$ represents its corresponding probability density function. Based on Eq. (9), the k-th resulting posterior probability is taken as the clean probability p_i for the i-th training sample as follows:

$$p_i = p(\theta_k|l_i) = \frac{p(\theta_k)p(l_i|\theta_k)}{p(l_i)}, \quad (10)$$

where $k = 0/1$ denotes whether the data pair (I_i, T_i, y_i) is clean/noisy. Using the estimated clean probability for each sample pair, the dataset is divided into a clean subset $\mathcal{D}_{\text{c}}$ and a noisy subset $\mathcal{D}_{\text{n}}$ as:

$$\mathcal{D}_{\text{c}} = \{(I_i, T_i) \mid p(k = 0 \mid \ell_i) > \delta, \forall (I_i, T_i) \in \mathcal{D}\}, \quad (11)$$

$$\mathcal{D}_{\text{n}} = \{(I_i, T_i) \mid p(k = 0 \mid \ell_i) \leq \delta, \forall (I_i, T_i) \in \mathcal{D}\}, \quad (12)$$

where δ denotes the threshold for separating clean samples from noisy ones. In practice, it is set to 0.5 in our experiments.

After filtering via the GMM model, we design a distinct loss function for the negative pairs associated with noisy and clean samples, as formulated below:

$$\mathcal{L}_{\text{ANRL}} = \frac{1}{B} \sum_{i=1}^{B} \log \left(1 + \sum_{j=1}^{B} e^{\lambda(S_{ij} - \gamma)} \cdot S_{ij} \cdot M_{ij} \right), \quad (13)$$

where λ denotes the scaling coefficient and γ indicates a penalty margin. The masking matrix $M \in \{0, 1\}^{B \times B}$ is used to select negative pairs, which is defined as:

$$M_{ij} = \begin{cases} 1, & \text{if } i = j \text{ and } (I_i, T_i) \in \mathcal{D}_n \\ 1, & \text{if } i \neq j \\ 0, & \text{otherwise} \end{cases}. \quad (14)$$

Objective Function

Thus, the total objective function in the training process can be denoted as:

$$\mathcal{L}_{total} = \mathcal{L}_{WCL} + \mu \cdot \mathcal{L}_{ANRL}, \quad (15)$$

where μ is the trade-off parameter.

Experiments

Experimental Settings

Datasets and Evaluation Metrics. Our method is evaluated on three benchmark datasets, MS-COCO (Lin et al. 2014), Flickr30K (Young et al. 2014), and CC120K (Sharma et al. 2018):

- MS-COCO contains 123,287 images, each annotated with five textual descriptions. Following standard practices, we split the dataset into 113,287 training images, 5,000 validation images, and 5,000 test images.
- Flickr30K contains 31,783 images, each associated with five descriptive sentences. According to established protocols, we divide the data set into 29,783 training images, 1,000 validation images, and 1,000 test images.
- CC120K, derived from the Conceptual Captions dataset, includes 120,851 noisy image-text pairs, each with one caption. Due to real-world data collection, 3%-20% of pairs have mismatched annotations. We use a split of 118,851 training, 1,000 validation, and 1,000 test samples to assess robustness under noisy conditions.

For evaluation, we use Recall@K (R@K) and the sum (i.e., RSUM) to measure the percentage of ground truth matches in the top K retrieved results ($K \in \{1, 5, 10\}$), with higher values indicating a stronger alignment capability. Additionally, the variance of R@1 scores ($\text{Var}(R@1)$) across different noise levels that measures the robustness to data distribution changes is also used, with lower variance indicating better stability in real-world settings.

Implementation Details HaNa is a robust vision-language matching method that enhances noise resistance in cross-modal models, using CLIP with ViT-B/32 as the baseline visual encoder. Experiments are conducted on a single RTX 4090 GPU, optimized with AdamW (default parameters) (Loshchilov and Hutter 2019), with an initial learning rate of 1×10^{-5} and a weight decay of 0.2. The batch size is fixed at 256. Training epochs are set to 5 for MS-COCO and Flickr30K, and 10 for CC120K. The hyperparameters are set as: contrastive loss temperature $\tau = 0.07$, momentum update coefficient $\alpha = 0.8$, penalty margin $\gamma = 0.2$, objective function scaling factor $\lambda = 64$, and the trade-off parameter $\mu = 0.01$. The code is available at: https://github.com/haiquan-yu/HaNa.

Comparison with State of the Arts

In this section, we conduct comprehensive comparisons with several state-of-the-art approaches on three benchmark datasets, including NCR (Huang et al. 2021), DECL (Qin et al. 2022), BiCro (Yang et al. 2023), ESC (Yang et al. 2024), GSC (Zhao et al. 2024), L2RM (Han et al. 2024), ReCon (Zha et al. 2025b), NPC (Zhang, Li, and Ye 2024), and the fine-tuned CLIP model (Radford et al. 2021). Among them, NPC and our HaNa are based on CLIP encoders. To evaluate our method's effectiveness and robustness, the experiments are evaluated under three noise settings, increasing the noise ratio from 0.2 to 0.6 in 0.2 increments. We simulate the noise by randomly shuffling matched image-text pairs from MS-COCO and Flickr30K at varying proportions, creating controlled environments to analyze performance under mismatched data.

Experiments on Synthetic Noise As shown in Table 1, our proposed method demonstrates superior performance

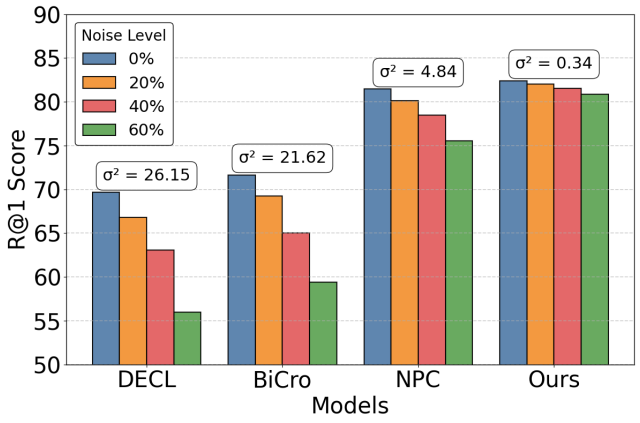


Figure 3: Variance analysis of several methods on the Flickr30K dataset.

Noise	Methods	MS-COCO 1K						Flickr30K					
		Image-to-Text			Text-to-Image			Image-to-Text			Text-to-Image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
20%	NCR	77.7	95.5	98.2	62.5	89.3	95.3	73.5	93.2	96.6	56.9	82.4	88.5
	DECL	77.5	95.9	98.4	61.7	89.3	95.4	77.5	93.8	97.0	56.1	81.8	88.5
	BiCro	78.8	96.1	98.6	63.7	90.3	95.7	78.1	94.4	97.5	60.4	84.4	89.9
	ESC	79.2	97.0	**99.1**	64.8	90.7	96.0	79.0	94.8	97.5	59.1	83.8	89.1
	GSC	79.5	96.4	98.9	64.4	90.6	95.9	78.3	94.6	97.8	60.1	84.5	90.5
	L2RM	80.2	96.3	98.5	64.2	90.1	95.4	77.9	95.2	97.8	59.8	83.6	89.5
	ReCon	80.9	**96.6**	98.8	65.2	91.0	96.0	80.3	95.3	97.8	61.6	85.5	91.3
	CLIP	75.0	93.1	97.2	58.7	86.1	97.2	82.3	95.5	98.3	66.0	88.5	93.5
	NPC	79.9	95.9	98.4	66.3	90.8	**98.4**	87.3	97.5	98.8	72.9	92.1	95.8
	Ours	**81.7**	**96.6**	98.8	**69.0**	**92.2**	96.6	**89.0**	**97.9**	**99.3**	**75.1**	**93.7**	**96.8**
40%	NCR	74.7	94.6	98.0	59.6	88.1	94.7	68.1	89.6	94.8	51.4	78.4	84.8
	DECL	75.6	95.5	98.3	59.5	88.3	94.8	72.7	92.3	95.4	53.4	79.4	86.4
	BiCro	77.0	95.9	98.3	61.8	89.2	94.9	74.6	92.7	96.2	55.5	81.1	87.4
	ESC	78.6	96.6	**99.0**	63.2	90.6	95.9	76.1	93.1	96.4	56.0	80.8	87.2
	GSC	78.2	95.9	98.2	62.5	89.7	95.4	76.5	94.1	97.6	57.5	82.7	88.9
	L2RM	77.5	95.8	98.4	62.0	89.1	94.9	75.8	93.2	96.9	56.3	81.0	87.3
	ReCon	79.9	96.2	98.6	63.5	90.5	95.9	79.4	94.3	97.6	59.9	83.9	90.1
	CLIP	70.7	91.7	96.2	54.7	83.4	96.2	76.2	93.3	96.5	59.4	85.0	90.9
	NPC	79.4	95.1	98.3	65.0	90.1	**98.3**	85.6	97.5	98.4	71.3	91.3	95.3
	Ours	**81.9**	**96.7**	98.9	**68.2**	**91.6**	96.5	**88.5**	**98.3**	**99.3**	**75.1**	**93.6**	**96.8**
60%	NCR	0.1	0.3	0.4	0.1	0.5	1.0	13.9	37.7	50.5	11.0	30.1	41.4
	DECL	73.0	94.2	97.9	57.0	86.6	93.8	65.2	88.4	94.0	46.8	74.0	82.2
	BiCro	73.9	94.4	97.8	58.3	87.2	93.9	67.6	90.8	94.4	51.2	77.6	84.7
	ESC	77.2	95.1	98.1	61.1	89.3	95.2	72.6	90.9	94.6	53.0	78.6	85.3
	GSC	75.6	95.1	98.0	60.0	88.3	94.6	70.8	91.1	95.9	53.6	79.8	86.8
	L2RM	75.4	94.7	97.9	59.2	87.4	93.8	70.0	90.8	95.4	51.3	76.4	83.7
	ReCon	77.2	95.9	98.4	61.8	89.3	95.2	74.3	93.6	96.6	55.7	81.6	88.1
	CLIP	67.0	88.8	95.0	49.7	79.6	95.0	66.3	87.3	93.0	52.1	78.8	87.4
	NPC	78.2	94.4	97.7	63.1	89.0	**97.7**	83.0	95.9	98.6	68.1	89.6	94.2
	Ours	**80.8**	**96.1**	**98.6**	**67.0**	**91.1**	96.1	**87.1**	**98.0**	**99.4**	**74.6**	**93.2**	**96.4**

Table 1: Image-Text Matching on MS-COCO 1K and Flickr30K.

across various noise levels compared to existing approaches. Specifically, compared to the state-of-the-art NPC, our framework exhibits significant advantages in both retrieval accuracy and noise robustness. In low-noise scenarios (20% noise ratio), our method achieves substantial improvements in the RSUM, increasing from 487.2 to 551.8 on Flickr30K and from 517.5 to 534.9 on MS-COCO datasets relative to the pioneering NCR approach. Under moderate noise conditions (40% noise ratio), we observe performance gains of 12.2 points on Flickr30K and 7.3 points on MS-COCO over the second-best NPC. Most notably, in high-noise environments (60% noise ratio), our framework consistently outperforms the CLIP baseline across all critical metrics (R@1, R@5, R@10), with particular emphasis on maintaining stable R@1 performance, the most crucial indicator for text-image retrieval tasks. As for the stability, our method achieves significantly lower variance compared to the baseline approaches shown in Figure 3, demonstrating improved consistency and reliability under varying noise conditions.

Experiments on Real-World Noise Table 2 shows the experimental results on the realistic noisy dataset CC120K. As can be seen, our method achieves more competitive retrieval performance compared to baseline methods under real-world noise conditions. Specifically, compared to the baseline CLIP model, our method achieves performance gains of 5.6 (R@1), 5.2 (R@5), and 2.9 (R@10) in image-to-text retrieval, and gains of 4.7 (R@1), 5.6 (R@5), and 3.7 (R@10) in text-to-image retrieval. Compared to the current state-of-the-art NPC model, the RSUM increases from 517.6 to 521.5. These results further demonstrate the effectiveness of our reweighting mechanism and asymmetric noise-aware regularization, which completely differ from NPC.

Methods	Image-to-Text			Text-to-Image			RSUM
	@1	@5	@10	@1	@5	@10	
CLIP	68.8	87.0	92.9	67.8	86.4	90.9	493.8
NPC	71.1	92.0	**96.2**	**73.0**	90.5	**94.8**	517.6
Ours	**74.4**	**92.2**	95.8	72.5	**92.0**	94.6	**521.5**

Table 2: Comparison with baselines on CC120K.

Experiments on ViT-B/32 Backbone Methods Since the CLIP-based ViT-B/32 backbone (Devlin et al. 2019; Dosovitskiy et al. 2021) inherently offers superior performance, to ensure a fair comparison, we also present experimental results from other methods that adopt the same ViT-B/32 architecture on MS-COCO for 1K and 5K evaluation protocols, respectively. The methods includes VSE∞ (Chen et al. 2021), PCME (Chun et al. 2021), PCME++ (Chun 2023),

Noise	Methods	1K R@1	5K R@1	1K RSUM
20%	VSE∞	72.0	51.4	520.2
	PCME	69.9	48.1	519.3
	PCME++	70.8	49.5	522.4
	PAU	71.4	51.7	521.5
	CLIP	66.8	47.2	507.2
	NPC	73.1	53.8	529.8
	Ours	**75.4**	**57.4**	**536.7**
50%	VSE∞	38.5	18.4	390.5
	PCME	65.8	43.0	505.7
	PCME++	65.7	44.0	503.9
	PAU	69.3	49.6	513.4
	CLIP	60.9	41.4	486.0
	NPC	71.3	51.9	523.4
	Ours	**74.8**	**56.4**	**533.0**

Table 3: Comparison of methods with ViT-B/32 backbone on MS-COCO 5K.

Configuration	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
HaNa	**88.5**	**98.3**	**99.3**	**75.1**	**93.6**	**96.8**
w/o WCL	76.0	92.7	96.3	57.5	81.2	87.1
w/o Mo	87.6	97.6	99.2	74.1	93.0	96.6
w/o WCS	86.5	97.6	99.2	73.5	92.8	96.4
w/o ANRL	87.3	97.9	99.0	74.3	93.3	96.4

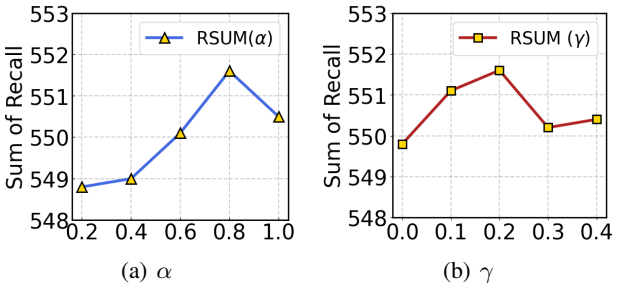
Table 4: Ablation studies on Flickr30K with 40% noise ratio.

PAU (Li et al. 2023), and our baseline model NPC. The comparison results for these models are all taken from the original NPC paper. As shown in Table 3, our method achieves performance improvements to varying degrees under both 20% and 50% noise ratios. Compared to existing methods based on CLIP ViT-B/32, our approach significantly reduces the sensitivity to noisy data during training, demonstrating enhanced robustness and practicality, which further highlights the value of our work.

Ablation Study

To comprehensively validate the contribution of different components in our framework, we conduct detailed ablation studies on the Flickr30K dataset with 40% noise rate, as shown in Table 4. The experiment includes four variants of HaNa. HaNa w/o weighted contrastive loss (w/o WCL), which replaces the weighted contrastive loss with standard contrastive loss. HaNa w/o momentum update (w/o Mo), which directly uses weights obtained from Eq. (7) without the momentum update mechanism. HaNa w/o weighted clean simples (w/o WCS), which means a weight of 1 is assigned to the clean samples identified by GMM. HaNa w/o ANRL, which excludes the asymmetric noise-aware regularization loss.

Compared with HaNa w/o WCL, the proposed weighted contrastive loss demonstrates significant performance improvement in cross-modal retrieval tasks, empirically validating its effectiveness in handling noisy annotations. The momentum update mechanism enables smoother weight transitions across batches, as evidenced by the performance degradation in HaNa w/o Mo. This smoothing strategy proves particularly beneficial for noise correlation tasks, where abrupt weight changes might amplify error propagation. The results of HaNa w/o WCS validate the importance of assigning different weights to clean samples, as demonstrated by improved performance metrics and alignment with the dynamic weighting mechanism described in our framework. While HaNa w/o ANRL shows comparable results, the consistent improvements across all six evaluation metrics when incorporating ANRL confirm its validity.



(a) α (b) γ

Figure 4: Analysis of hyper-parameters α and γ on Flickr30K with 40% noise.

Hyperparameter Analysis

The selection of parameters α and γ plays a crucial role in the performance of HaNa: α represents the weight contribution from the memory bank during momentum update, while γ determines the threshold for applying strong penalties in the ANRL loss. Therefore, we conduct a systematic analysis of these two key hyperparameters. As shown in Figure 4, HaNa can achieve the optimal performance (RSUM) when $\alpha = 0.8$ and $\gamma = 0.2$. It can be clearly seen that our model accepts different values of α and γ, and can always find the optimals, suggesting that it can be conveniently trained and the results are highly reproducible.

Discussion

Although the reweighting mechanism is effective in handling the hardness of clean samples, it is still limited without partitioning and correcting noisy data. This is precisely one of the reasons that constrain the performance improvement of our HaNa. It motivates our future work to focus on classifying and correcting noisy data through novel methods such as the prompts from LLMs and the Mixup scheme.

Conclusion

This paper proposes HaNa, a novel robust cross-modal retrieval method designed to mitigate the negative impact of noisy correspondences. Different from previous methods, HaNa introduces a reweighting mechanism for clean samples and addresses the problem of noisy weights approaching 0 from a negative perspective. The experimental results have validated the effectiveness of HaNa in alleviating overfitting risk and improving generalization.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 62476036, 62006035), and Science and Technology Major Project on Artificial Intelligence of Liaoning Province (2023JH26/10100008).

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 6077–6086. Computer Vision Foundation / IEEE Computer Society.

Chen, J.; Hu, H.; Wu, H.; Jiang, Y.; and Wang, C. 2021. Learning the Best Pooling Strategy for Visual Semantic Embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chun, S. 2023. Improved Probabilistic Image-Text Representations. arXiv:2305.18171.

Chun, S.; Oh, S. J.; de Rezende, R. S.; Kalantidis, Y.; and Larlus, D. 2021. Probabilistic Embeddings for Cross-Modal Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR-21)*, 8415–8424. virtual: IEEE.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. 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 (NAACL HLT-19)*, 4171–4186. Minneapolis, MN, USA: Association for Computational Linguistics.

Diao, H.; Zhang, Y.; Ma, L.; and Lu, H. 2021. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI conference on artificial intelligence (AAAI-21)*, 1218–1226. Palo Alto, California: AAAI Press.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

Feng, Z.; Zeng, Z.; Guo, C.; Li, Z.; and Hu, L. 2024. Learning From Noisy Correspondence With Tri-Partition for Cross-Modal Matching. *IEEE Transactions on Multimedia*, 26: 3884–3896.

Han, H.; Zheng, Q.; Dai, G.; Luo, M.; and Wang, J. 2024. Learning to Rematch Mismatched Pairs for Robust Cross-Modal Retrieval. arXiv:2403.05105.

Huang, Z.; Niu, G.; Liu, X.; and et al. 2021. Learning with noisy correspondence for cross-modal matching. In *NeurIPS*.

Li, H.; Song, J.; Gao, L.; Zhu, X.; and Shen, H. T. 2023. Prototype-based Aleatoric Uncertainty Quantification for Cross-modal Retrieval. In *NeurIPS*.

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

Li, S.; Tao, Z.; Li, K.; and Fu, Y. 2019. Visual to Text: Survey of Image and Video Captioning. *IEEE Trans. Emerg. Top. Comput. Intell.*, 3(4): 297–312.

Li, Z.; Guo, C.; Wang, X.; Zhang, H.; and Hu, L. 2025. Multi-view visual semantic embedding for cross-modal image–text retrieval. *Pattern Recognition*, 159: 111088.

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 *Proceedings of the European conference on computer vision (ECCV-14)*, 740–755. Zurich, Switzerland: Springer.

Liu, F.; Dong, C.; Zhang, C.; Zhou, H.; and Zhou, J. 2024. Robust Noisy Correspondence Learning via Self-Drop and Dual-Weight. arXiv:2412.06172.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations (ICLR-19)*. New Orleans, LA, USA: OpenReview.net.

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 transactions on image processing*.

Pan, Z.; Wu, F.; and Zhang, B. 2023. Fine-grained Image-text Matching by Cross-modal Hard Aligning Network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pham, K.; Huynh, C.; Lim, S.-N.; and Shrivastava, A. 2024. Composing object relations and attributes for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14354–14363.

Pu, R.; Sun, Y.; Qin, Y.; Ren, Z.; Song, X.; Zheng, H.; and Peng, D. 2025. Robust Self-Paced Hashing for Cross-Modal Retrieval with Noisy Labels. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, 19969–19977. AAAI Press.

Qin, Y.; Chen, Y.; Peng, D.; Peng, X.; Zhou, J. T.; and Hu, P. 2024a. Noisy-Correspondence Learning for Text-to-Image Person Re-identification. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Qin, Y.; Peng, D.; Peng, X.; Wang, X.; and Hu, P. 2022. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM-22)*, 4948–4956. New York, NY, United States: Association for Computing Machinery.

Qin, Y.; Sun, Y.; Peng, D.; Zhou, J. T.; Peng, X.; and Hu, P. 2024b. Cross-modal Active Complementary Learning with Self-refining Correspondence. *Advances in Neural Information Processing Systems*, 36.

Qu, L.; Liu, M.; Wu, J.; Gao, Z.; and Nie, L. 2021. Dynamic modality interaction modeling for image-text retrieval. In

Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 1104–1113.

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 *Proceedings of the 38th International Conference on Machine Learning ICML-21*, 8748–8763. Virtual Event: PMLR.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. 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 (ACL-18)*, 2556–2565. Melbourne, Australia: Association for Computational Linguistics.

Sun, Y.; Liu, K.; Li, Y.; Ren, Z.; Dai, J.; and Peng, D. 2024. Distribution Consistency Guided Hashing for Cross-Modal Retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, 5623–5632. ACM.

Wang, Y.; Wu, Y.; Dai, Z.; Tian, C.; Long, J.; and Chen, J. 2025. Noisy Correspondence Rectification via Asymmetric Similarity Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(20): 21384–21392.

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-23*.

Yang, Y.; Wang, L.; Yang, E.; and Deng, C. 2024. Robust noisy correspondence learning with equivariant similarity consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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. *Transactions of the Association for Computational Linguistics*, 2: 67–78.

Zha, Q.; Liu, X.; Cheung, Y.-M.; Peng, S.-J.; Xu, X.; and Wang, N. 2025a. UCPM: Uncertainty-Guided Cross-Modal Retrieval With Partially Mismatched Pairs. *IEEE Transactions on Image Processing*, 34: 3622–3634.

Zha, Q.; Liu, X.; Peng, S.-J.; ming Cheung, Y.; Xu, X.; and Wang, N. 2025b. ReCon: Enhancing True Correspondence Discrimination through Relation Consistency for Robust Noisy Correspondence Learning. arXiv:2502.19962.

Zhang, X.; Li, H.; and Ye, M. 2024. Negative Pre-aware for Noisy Cross-Modal Matching. In *AAAI*.

Zhang, Y.; Ji, Z.; Wang, D.; Pang, Y.; and Li, X. 2024. USER: Unified semantic enhancement with momentum contrast for image-text retrieval. *IEEE Transactions on Image Processing*.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27381–27390.