# Noisy Correspondence Rectification via Asymmetric Similarity Learning

**Yunbo Wang[1*], YuJie Wu[2*], Zhien Dai[3], Can Tian[4†], Jun Long[1], Jianhai Chen[5]**

[1]Big Data Institute, Central South University, China
[2]School of Computer Science and Engineering, Central South University, China
[3]School of Automation, Central South University, China
[4]School of Computer Science and Cyber Engineering, Guangzhou University, China
[5]College of Computer Science and Technology, Zhejiang University, China
{wangyunbo, 8102211218, zhiendai, jlong}@csu.edu.cn, tiancan@gzhu.edu.cn, chenjh919@zju.edu.cn

## Abstract

Cross-modal matching shows enormous potential to recognize objects across different sensory modalities, which is fundamental to numerous visual-language tasks like image-text retrieval and visual captioning. Existing works generally rely on massive and well-aligned data pairs for model training. Unfortunately, multimodal datasets are extremely difficult to annotate and collect. As an alternative, the co-occurred data pairs collected from the internet have been widely exploited to train a cross-modal matching model. However, the cheaply-collected dataset unavoidably contains mismatched pairs (i.e., noisy correspondence), which are detrimental to the matching model. In this paper, we propose an alternative method termed noisy correspondence rectification via Asymmetric Similarity Learning (ASL), and it allows for dealing with insufficient learning of positive and negative pairs caused by the popular triplet-based symmetric learning fashion. Specifically, the learning of positive or negative pairs within a triplet is conducted in an asymmetric fashion, and the self-paced weighting boundary is imposed on positive pairs to mitigate the effect of noise. Meanwhile, the optimization of negative samples will not be affected in the process of punishing potentially-noisy positive samples. To verify the effectiveness of our proposed approach, a series of experiments are conducted on three widely-used benchmarks (i.e., Flick30k, MS-COCO and CC152k), and the results show superior performance compared to the state-of-the-art methods.

## Introduction

Cross-modal matching (Pan, Wu, and Zhang 2023; Wei et al. 2023a; Zhang et al. 2022) aims to establish the relationship between different modalities (e.g, image and text), facilitating the downstream tasks such as image/video captioning (Liu et al. 2023; Nguyen et al. 2024), cross-modal retrieval (Wang and Peng 2022; Kim, Kim, and Kwak 2023; Yang et al. 2023b; Wu et al. 2023), and visual question answering (Shao et al. 2023). The common practice in cross-modal matching is to project the representations of different modalities into a common subspace, so that the semantic correlation across modalities can be obtained by similarity measures such as cosine similarity and Euclidean distance.

With the rapid advance of computing resources, recent works have achieved impressive performance by training models on larger-scale datasets. Unfortunately, it is costly to collect and construct a well-annotated dataset in a real-world scene, especially multimodal datasets. Thus, current multimodal datasets unavoidably contain mismatched pairs, resulting in noisy correspondence (Huang et al. 2025). For example, the popular cross-modal datasets MS-COCO (Lin et al. 2014) and CC152K (Sharma et al. 2018) include a great number of inaccurate descriptions in image-text pairs. Different from the traditional noisy label (Song et al. 2022; Huang, Zhang, and Shan 2023; Wei et al. 2023b), the noisy correspondence refers to mismatched cross-modal data pairs, while these mismatched pairs are regarded as matched pairs when training model. Therefore, the performance of existing cross-modal tasks has nearly reached a bottleneck. Recently, some methods (Yang et al. 2023a; Han et al. 2023; Zha et al. 2024) have been presented to deal with noisy correspondence in cross-modal tasks, mitigating the adverse impact of noise. These works seek to model the difference of distribution about per-sample loss at the early period of training based on the memory effect of deep neural networks (DNNs), where DNNs tend to fit simple samples at this early stage. Next, the dataset would be classified into clean subset and noisy subset, in which the small-loss samples are more likely to be clean samples. Subsequently, soft label estimation strategies are employed for the two subsets, and the soft label is expressed as a margin in the triplet-based matching loss to penalize the noisy pairs, preventing the deep neural networks from over-fitting noisy data. However, the effect from such margin in triplet loss acted on the positive and negative pairs is symmetrical, so the penalty for the potentially-noisy positive sample will act on the negative sample with the same force, resulting in insufficient learning and suboptimal performance. Therefore, how to effectively exploit the supervision from positive and negative pairs remains a critical challenge in noisy correspondence.

To address the above issues, we propose a method dubbed noisy correspondence rectification via Asymmetric Similarity Learning (ASL) for robust image-text matching. It constructs an alternative paradigm that enables positive and negative pairs to be optimized in an asymmetric way, achieving independently penalizing potentially-noisy positive pairs without impacting the optimization of negative pairs. The

*These authors contributed equally.
†Corresponding author (tiancan@gzhu.edu.cn).

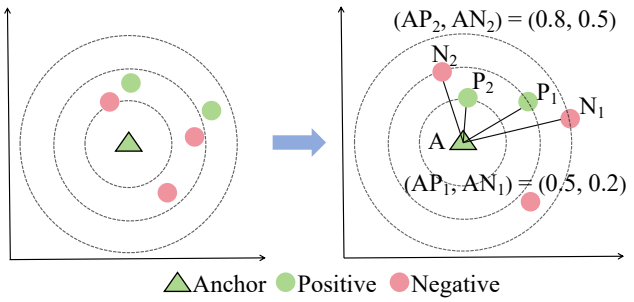


Figure 1: illustrates the typical triplet loss, which mainly focuses on maintaining the margin between positive pair and negative pair. Thus it produces a suboptimal performance when the score of positive pairs and negative pairs is ambiguous or close.

traditional triplet loss is designed to narrow the distance of positive pairs and enlarge the distance of negative pairs in a symmetric fashion, where the single margin in triplet determines the decision boundary of model convergence (Wang et al. 2019b; Yang et al. 2023b). However, such a learning strategy inevitably leads to insufficient optimization of negative and positive pairs. For example, assuming the value of decision boundary is 0.3 in the triplet unit $(AP_i, AN_i)$ shown in Fig. 1, where $AP_i$ and $AN_i$ denote the similarity of positive pair and negative pair, respectively. The set (0.5, 0.2) can be considered as an ideal convergence, while another set (0.8, 0.5) is also an ideal state for the same decision boundary. This situation would inevitably result in a suboptimal performance. In our proposed ASL, the positive and negative pairs are decoupled, and they can be optimized in an asymmetric way. It allows for penalizing potentially-noisy positive samples without impacting the optimization of negative ones.

The main contributions of this work are summarized as follows:

- We propose a method termed noisy correspondence rectification with Asymmetric Similarity Learning, which aims to correct the noisy correspondence via a general asymmetric learning paradigm, thereby mitigating the effect of noise on image-text matching.
- We propose to model the distribution of per-sample loss using a Variational Bayesian Gaussian Mixture Model, which incorporates prior distribution and Bayesian inference to fit the distribution of loss in noisy scenarios, obtaining better probability density parameters.
- A novel asymmetric dynamic matching loss is presented to independently regulate the optimization of positive and negative pairs. It enables accurately punishing potentially-noisy positive samples without impacting the optimization of negative ones.
- Extensive experiments are conducted on three widely-used benchmarks, and the results demonstrate the effectiveness of our proposed method in both synthetic and real noise scenarios.

# Related Works

## Cross-modal Matching

Cross-modal matching is a fundamental research in the field of multimodal learning (Lee et al. 2018; Jia et al. 2022; Huang et al. 2022; Schlarmann and Hein 2023), and it usually projects the representations of different modalities into a common subspace, thereby retrieving readily similar items of other modality given the query modality (Pan, Wu, and Zhang 2023). According to different matching fashions, existing works can be classified into two categories in image-text matching: 1) Coarse-grained matching. The images and texts are matched from a global feature computed by deep neural networks. VSE++ (Faghri et al. 2017) attempts to employ the hard negatives in triplet loss to enhance the global feature matching. TIMAM (Sarafianos, Xu, and Kakadiaris 2019) proposes to learn the modality-invariant global representation by Generative Adversarial Networks, improving the image-text matching. The similar works (Wang et al. 2019a; Zhang et al. 2020) employ a two-stream global feature learning network and compute the pairwise similarity according to global feature. 2) Fine-grained matching. The regions within an image and words in a sentence correspond with each other so that fine-grained information is captured by embedding semantically-similar regions and works. For example, DVSA (Karpathy and Fei-Fei 2015) proposes to take the maximized matching score between regions of image and words of sentence as the image-text matching score. VSRN (Li et al. 2019) tries to reason the visual semantic relationship between the regions of image via graph convolutional network for enhancing cross-modal matching. SGRAF (Diao et al. 2021) further constructs a graph structure for multimodal data to infer relation-aware similarities via graph reasoning. However, all these alignment methods assume that the multimodal data are perfectly aligned in training, while it is impossible to satisfy due to a high-cost collection and annotation.

## Noisy Correspondence Learning

Different from the traditional noisy label, noisy correspondence refers to the mismatch issue in paired data. NCR (Huang et al. 2021) first investigates this issue and proposes a noise-robust solution according to the memorization effect of DNNs. It considers mismatched cross-modal pairs instead of incorrect annotations. After that, a series of methods are presented to rectify noisy correspondence in various visual-language tasks, like graph matching (Lin et al. 2023), image captioning (Kang et al. 2023), multi-view learning (Yang et al. 2021), person re-identification (Yang et al. 2022) and cross-modal retrieval (Feng et al. 2024; Han et al. 2024; Zhang, Li, and Ye 2024). DECL (Qin et al. 2022) proposes a cross-modal evidential learning solution and dynamically filters out noisy correspondences within each batch. MSCN (Han et al. 2023) utilizes meta-learning to distinguish the clean and noisy pairs via jointing the triplet-based matching objective. BiCro (Yang et al. 2023a) constructs a soft label estimation strategy and expresses it as a margin in triplet loss to correct the noisy correspondence in image-text matching. Recently, L2RM (Han et al. 2024) leverages op-

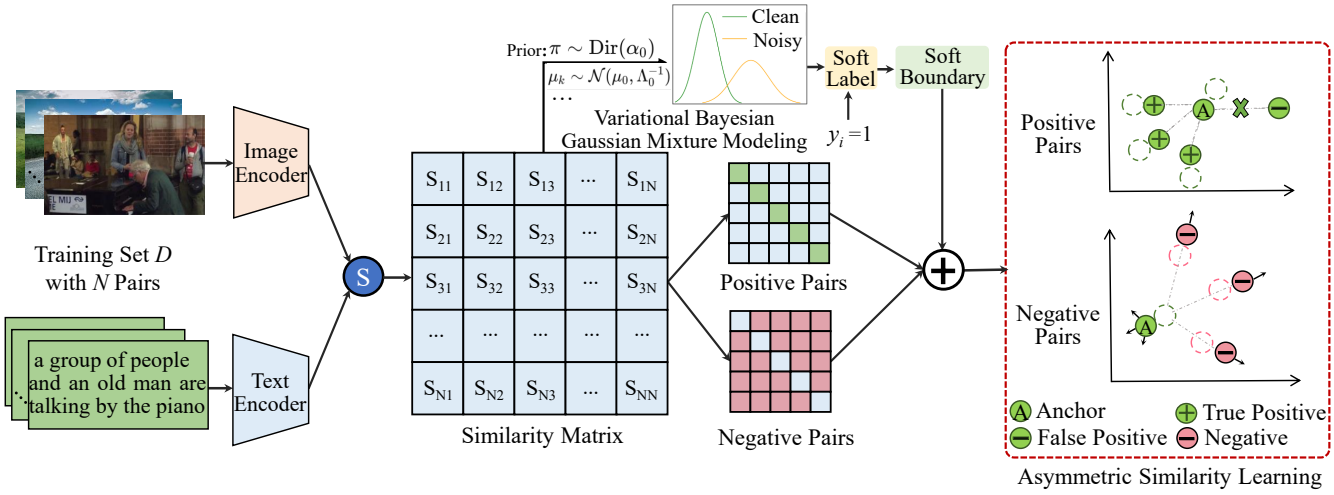


Figure 2: the framework of the proposed ASL for robust image-text matching, which mainly consists of soft boundary computing and asymmetric similarity learning. In soft boundary computing, the VBGMM is exploited to model the distribution of per-sample loss for estimating its label, and then the estimated label is leveraged to compute the optimized boundary of per-sample. In asymmetric similarity learning, a novel asymmetric dynamic matching loss is presented to independently regulate the optimization of positive and negative pairs.

timal transport to discover potential matching relationships among mismatched pairs, thus utilizing useful knowledge from mismatched pairs for improving image-text matching. However, most of existing works leverage the triplet-based symmetric learning fashion to correct noise correspondence. The penalty enforced on the potentially-noisy positive samples would react on the negative ones with the same intensity, leading to insufficient learning of image-text pairs. In this paper, we present an asymmetric learning fashion to conduct noisy correspondence rectification, which effectively exploits the supervision from positive and negative pairs, achieving the potentially-noisy positive samples penalty without affecting the optimization of negative ones.

## Methodology

### Problem Definition

Following the common practice, we utilize image-text matching as a proxy for revealing noise correspondence. Fig. 2 gives the framework of our proposed ASL, which mainly consists of soft boundary computing and asymmetric similarity learning.

Given a dataset $D = \{(I_i, T_i, y_i)\}_{i=0}^{N}$, where $N$ denotes the total number of training samples, $(I_i, T_i)$ represents an image-text pair and $y_i \in \{0, 1\}$ is the binary label. The label $y_i$ is a hard-defined correspondence that indicates the image-text pair $(I_i, T_i)$ is positively correlated $(y_i = 1)$ or not $(y_i = 0)$. The objective of image-text matching is to embed the features of image and text in the original space into a unified common representation space via feature encoders, where the similarity of positive pairs should be maximized as far as possible. The similarity of an image-text pair can be generally written as $S(f(I_i), g(T_i))$, in which $S$ is a similarity metric like cosine similarity, $f$ and $g$ are the feature encoder of image and text modalities, respectively. For ease of description, we denote $S(f(I_i), g(T_i))$ as $S(I_i, T_i)$ in the following context. The typical triplet loss is widely used to train the feature encoders:

$$\begin{aligned} l_{tri}(I_i, T_i) =& [S(I_i, \hat{T}_j) - S(I_i, T_i) + \alpha]_+ \\ &+ [S(\hat{I}_h, T_i) - S(I_i, T_i) + \alpha]_+, \end{aligned} \tag{1}$$

where $\alpha > 0$ denotes a given margin, $[x]_+ = max(x, 0)$. In the triplet loss, the first term treats $I_i$ as a query taking over all negative texts $\hat{T}_j$, and the second term treats $T_i$ as a query taking over all negative images $\hat{I}_h$.

The application of triplet loss highly relies on the assumption that image-text pairs are correctly aligned. However, in a real-world scenario, the cheaply-collected multimodal data usually contain a part of mismatched pairs, which are erroneously labeled as matched pairs, i.e., the label $y_i = 1$. However, the above objective tends to over-fit these noisy data, and it inevitably leads to a performance decline in image-text matching.

### Cross-modal Matching with Noisy Correspondence

To address the noisy correspondence, existing works propose to correct the original hard-label correspondence $y_i = 1$ to a soft label $\hat{y}_i$, where $\hat{y}_i$ is a continuous value within the range [0, 1] describing the correspondence in image-text pair $(I_i, T_i)$. To be specific, if the correlation of data pair is strong, $\hat{y}_i$ would be close to 1, and vice versa. Subsequently, the soft label is transformed into a soft margin in the triplet loss, thereby achieving noise-robust cross-modal matching. The triplet loss with a soft margin can be defined as follows:

$$\begin{aligned} l'_{tri}(I_i, T_i) =& [S(I_i, \hat{T}_j) - S(I_i, T_i) + \alpha']_+ \\ &+ [S(\hat{I}_h, T_i) - S(I_i, T_i) + \alpha']_+, \end{aligned} \tag{2}$$

where $\alpha' = \frac{z^{\hat{y}_i} - 1}{z - 1}\alpha$, $\hat{y}_i$ represents the soft-label correspondence of the $i$-th data pair established on the correlation be-

tween $I_i$ and $T_i$, $z$ is a relaxation factor, and $\alpha$ is a constant margin.

The above formula is conducive to facilitating model convergence when positive pairs are with noise, and it decreases the likelihood of over-fitting on noisy pairs by reducing the margin. However, the above optimization formula has a significant short in noisy scenes. From the perspective of gradient, the gradients of positive and negative samples in triplet loss are listed as follows:

$$\frac{\partial l'_{tri}}{\partial T_i} = \begin{cases} -\partial S(I_i,T_i)/\partial T_i & S(I_i,T_i) < S(I_i,\hat{T}_j) + \alpha' \\ 0 & S(I_i,T_i) \geq S(I_i,\hat{T}_j) + \alpha' \end{cases} \quad (3)$$

$$\frac{\partial l'_{tri}}{\partial \hat{T}_j} = \begin{cases} \partial S(I_i,\hat{T}_j)/\partial \hat{T}_j & S(I_i,T_i) < S(I_i,\hat{T}_j) + \alpha' \\ 0 & S(I_i,T_i) \geq S(I_i,\hat{T}_j) + \alpha' \end{cases} \quad (4)$$

For ease of observation, we only give the gradients of the first term in Eq. (2) about the positive sample $T_i$ and negative sample $\hat{T}_j$. From Eq. (3) and Eq. (4), it can be observed that the gradients about positive and negative samples are symmetrical, and the gradient is always 0 when the condition holds $S(I_i,T_i) \geq S(I_i,\hat{T}_j) + \alpha'$. As we penalize noisy positive samples by decreasing the margin, the symmetrical nature of the gradient can cause premature termination of optimization for negative ones, resulting in insufficient optimization.

## Asymmetric Similarity Learning

Since the symmetrical structure of triplet loss leads to insufficient optimization on the positive and negative samples, our method proposes to minimize the impact on negative samples optimization while penalizing potentially-noisy positive ones in an asymmetric learning paradigm. Thus, it allows penalizing noisy positive samples while maintaining proper optimization for negative samples. Inspired by (Sun et al. 2020), a novel asymmetric dynamic matching loss is presented, achieving reliable penalization of noisy positive samples without affecting the optimization of negative ones.

$$l_{asy}(I_i,T_i,\hat{T}_j) = \log\Big(1 + \sum_{i=1}^{N} exp(-\lambda \mu_p^i (S(I_i,T_i) - m_p)) \sum_{j=1}^{N-1} exp(\lambda \mu_n^j (S(I_i,\hat{T}_j) - m_n))\Big) \quad (5)$$

where $\lambda$ represents a scaling factor, $m_p$ and $m_n$ denote the margin for positive and negative pairs, respectively. With the above formula, the optimization expectation of positive pair is $S(I_i,T_i) > m_p$, and negative pair is $S(I_i,\hat{T}_j) < m_n$. Empirically, $m_p$ is set to $1-m_0$ and $m_n$ is set to $m_0$, where $m_0$ is a constant margin with the default 0.2. $\mu_p$ and $\mu_n$ are non-negative weight factors defined as follows:

$$\begin{cases} \mu_p^i = [\sigma U_p - S(I_i,T_i)]_+ \\ \mu_n^j = [(S(I_i,\hat{T}_j) - U_n]_+ \end{cases} \quad (6)$$

where $U_p$ represents the ideal optimal boundary for positive pairs, and $U_n$ represents the ideal optimal boundary for negative pairs. Empirically, $U_p$ is set to $1+m_0$ and $U_n$ is set to $-m_0$. $[\cdot]_+$ indicates a truncation operation to zero, ensuring the weight factors $\mu_p^i$ and $\mu_n^j$ are non-negative. Since it is unclear whether the positive pair is clean or noisy, a soft boundary denoted as $\sigma U_p$ is computed according to the soft label $\hat{y}_i$. $\sigma U_p$ denotes the dynamic optimization upper boundary for positive pair, $\sigma = \frac{z^{\hat{y}^i}-1}{z-1}$, and z is a relaxation factor. When the soft label $\hat{y}^i = 1$, the dynamic optimization upper boundary equals to the ideal optimal boundary. Obviously, the soft label $\hat{y}^i$ is crucial to the learning of positive pairs, we will elaborate detailedly the estimation of soft label $\hat{y}^i$ in the next subsection.

With Eq. (5), we can accurately penalize the potentially-noisy positive samples. The gradients for positive and negative samples in $l_{asy}(I_i,T_i,\hat{T}_j)$ are listed as follows:

$$\frac{\partial l_{asy}(I_i,T_i,\hat{T}_j)}{\partial (T_i)} = \lambda A(2S(I_i,T_i) - \sigma U_p - (1-m)) \frac{\partial S(I_i,T_i)}{\partial (T_i)} \quad (7)$$

$$\frac{\partial l_{asy}(I_i,T_i,\hat{T}_j)}{\partial \hat{T}_j} = 2A \frac{exp(\lambda((s(I_i,\hat{T}_j)^2 - m^2))}{\sum_{l=1}^{N-1} exp(\lambda(s(I_i,\hat{T}_j)^2 - m^2))} \lambda(s(I_i,\hat{T}_j)) \frac{\partial S(I_i,\hat{T}_j)}{\partial (\hat{T}_j)} \quad (8)$$

where $A = 1 - exp^{-l_{asy}(I_i,T_i,\hat{T}_j)}$. From the above two formulas, we can observe that due to the asymmetric optimization fashion, penalizing noisy positive samples with the adaptive upper boundary would have a minimal impact on the optimization of negative ones. As the model stops optimizing noisy positive samples, it does not cease optimization for the corresponding negative samples, enhancing the utilization of supervision from negative ones.

## Soft Label Estimation with VBGMM

Soft label estimation aims to distinguish the clean pairs from noisy ones. Due to the memorization effect of DNNs, DNNs tend to memorize simple pairs first, and the loss of clean pair is lower than that of noisy pair at the initial stage of training. Therefore, like (Yang et al. 2023a), a warmup training is conducted for training our image-text matching model, and the loss of image-text pairs is denoted as $\ell_i$:

$$\ell_{(f,g,S)} = \{\ell_i\}_{i=1}^N = l_{tri}(I_i,T_i)_{i=1}^N \quad (9)$$

According to the difference about loss distribution of clean pair and noisy pair, the original dataset is divided into clean subset and noisy subset. Existing works widely use the Gaussian Mixture Model (GMM) (Huang et al. 2021) to model the difference about loss distribution, while the vanilla GMM can't exploit the priori for distribution modeling, bringing a certain uncertainty to the parameters of GMM. Recent research demonstrates that Variational Bayesian Gaussian Mixture Model (VBGMM) shows better performance in dividing different categories of data (Chakladar, Roy, and Chang 2024), which introduces prior distribution of parameters and Bayesian inference to approximate a posteriori distribution $p(\pi,Z,\mu,\Sigma/X)$, and it allows for dealing with the uncertainty of model parameters, generating more reliable parameters of probability density.

$$p(\theta,Z|X,m) = \prod_k p(\theta_k|X_k,m) p(Z_k|X_k,m) \quad (10)$$

where $Z$ is the hidden variable of model m, $p(Z|X,m)$ is the approximate posterior distribution of model parameters, and

$p(\theta|X,m)$ refers to the solution of VBGMM, where Expectation Maximization algorithm is used to perform optimization. Through the above formula, we can get precious parameters of probability density, and the loss of per-sample is fitted as follows:

$$p(\ell_i|\theta) = \sum_{k=1}^{K} \pi_k \phi(\ell_i|\theta_k) \tag{11}$$

where $K = 2$, $\pi_k$ and $\phi_k$ represent the mixture coefficient and the probability density function of the $k$-th component in VBGMM. Then, it computes the posterior probability $p(k|\ell_i)$ as the clean probability of $i$-th sample:

$$p(k|\ell_i) = p(k)p(\ell_i|k)/p(\ell_i) \tag{12}$$

where $k \in \{0,1\}$ denotes the data pair $(I_i, T_i, y_i)$ is clean or noisy. According to the above computed probability of each pair, we can partition the dataset into clean subset $D_{clean}$ and noisy subset $D_{noise}$ as follows:

$$\begin{aligned} D_{clean} &= \{(I_i,T_i)|p(k=0|\ell_i) > \delta, \forall (I_i,T_i) \in D\} \\ D_{noise} &= \{(I_i,T_i)|p(k=0|\ell_i) \leq \delta, \forall (I_i,T_i) \in D\} \end{aligned} \tag{13}$$

where $\delta$ is a threshold for partitioning the dataset into clean and noisy subsets, which is generally empirically set to 0.5. Since the model's predictions can lead to self-reinforcing errors and error accumulation in training, the co-training strategy is employed to mitigate this error. Specifically, we train two models simultaneously, and each model is equipped with a different initialization and batch sequence. The detailed calculation of soft label can be referred to (Huang et al. 2021).

## Robust Training

After obtaining the estimated soft label, we adopt the Eq. (5) to perform cross-modal matching. To ensure consistent performance across image and text modalities, we employ bidirectional matching to encompass both image-to-text and text-to-image tasks as follows:

$$L(I_i,T_i) = l_{asy}(I_i,T_i,\hat{T}_j) + l_{asy}(T_i,I_i,\hat{I}_h) \tag{14}$$

With the above formula, we can accurately penalize the potentially-noisy positive samples while having a minimal impact on the optimization of negative ones, achieving noise-robust image-text matching.

# Experiments

## Experimental Setting

**Datasets** To comprehensively evaluate the effectiveness of our proposed method, a series of experiments are conducted on three mainstream image-text matching benchmarks. Specifically, the performances of our method are shown under synthetic noise conditions on Filckr30K (Young et al. 2014) and MS-COCO (Lin et al. 2014), and under real-world noise conditions on Conceptual Captions (Sharma et al. 2018). More details of the three datasets are listed as follows:

- Flickr30K contains 31,000 images collected from the Flickr website, and each image is accompanied with 5 corresponding descriptive captions. Following (Qin et al. 2022), we use 1,000 pairs for validation, 1,000 pairs for testing, and 29,000 pairs for training.
- MS-COCO is another popular dataset for cross-modal learning, which collects 123,287 images associated with 5 captions. Following the split in (Lee et al. 2018), 5000 pairs are used for validation, 5,000 pairs for testing, and 113,287 pairs for training.
- Conceptual Captions contains 3,334,173 images with one caption. It is a large scale real-world dataset with about $3\% \sim 20\%$ pairs being mismatched, because it is automatically collected from the Internet. Following (Qin et al. 2022), we use a subset of Conceptual Captions, i.e., CC152K. In our experiment setting, 1,000 pairs are used for validation, 1,000 pairs for testing, and 150,000 pairs for training.

**Evaluation Metrics** Recall at K(R@K) is a widely-used metric to evaluate the performance of text-image retrieval. R@K primarily reports the proportion of successfully retrieved relevant items in the top-K results given a query. In our experiments, we mainly report R@1, R@5, R@10 and the sum of the three Recalls (rSum) for both text-to-image and image-to-text retrieval.

## Implementation Details

The experiments are conducted on NVIDIA RTX 3090 in Pytorch-2.0.1. For fairness in our experiments, our approach employs the same backbone as NCR, and the SGR model (Diao et al. 2021) is adopted at capturing the relationship between local alignments. Specifically, we first leverage the Faster-RCNN (Ren et al. 2015) to extract the top 36 regions for every image as a preprocess, and incorporate a fully-connected layer to serve as image feature encoder f. The Bi-GRU (Schuster and Paliwal 1997) serves as text feature encoder g. The similarity function S is computed by combining local and global features using graph reasoning techniques (Diao et al. 2021). To mitigate errors from self-reinforcement, we employed a co-training strategy. For the Flickr30K, we conduct 5 warmup training epochs, while 10 warmup epochs for MS-COCO and CC152K. The total number of training epochs after warmup is 40, 30, and 40 for Flickr30K, MS-COCO and CC152K, respectively. The Adam optimizer is used for training with a batch size of 180, and the initial learning rate is 0.0002. In hyperparameter settings, the margin value $m_0$ is set to 0.2, the scaling factor $\lambda$ is set to 64, and the relaxation factor $z$ is 3. For fitting VBGMM, we set the maximum number of iterations to 10 with a regularization penalty coefficient of 0.0005.

## Comparison with State-of-the-art Methods

To demonstrate the effectiveness of our proposed ASL, we compared our approach with several state-of-the-art methods, including baseline models like SCAN (Lee et al. 2018), SGR, SAF (Diao et al. 2021), and robust noisy learning methods including NCR (Huang et al. 2021), DECL (Qin et al. 2022), RCL (Hu et al. 2023), MSCN (Han et al. 2023),

| Noise ratio | 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 |
| | SGR | 55.9 | 81.5 | 88.9 | 40.2 | 66.8 | 75.3 | 408.6 | 25.7 | 58.8 | 75.1 | 23.5 | 58.9 | 75.1 | 317.1 |
| | SAF | 62.8 | 88.7 | 93.9 | 49.7 | 73.6 | 78.0 | 446.7 | 71.5 | 94.0 | 97.5 | 57.8 | 86.4 | 91.9 | 499.1 |
| | NCR | 75.0 | 93.9 | 97.5 | 58.3 | 83.0 | 89.0 | 496.7 | 76.6 | 95.6 | 98.2 | 62.5 | 89.3 | 95.3 | 517.5 |
| | DECL | 74.5 | 92.9 | 97.1 | 53.6 | 79.5 | 86.8 | 484.4 | 75.6 | 95.1 | 98.3 | 59.9 | 88.3 | 94.7 | 511.9 |
| | RCL | 74.2 | 91.8 | 96.9 | 55.6 | 81.2 | 87.5 | 487.2 | 77.0 | 95.5 | 98.1 | 61.3 | 88.8 | 94.8 | 515.5 |
| | MSCN | 76.4 | 94.5 | 97.6 | 58.8 | 83.5 | 89.2 | 500.0 | 78.1 | **97.2** | 98.8 | 64.3 | 90.4 | 95.8 | 524.6 |
| | BiCro | 76.5 | 93.1 | 97.4 | 58.1 | 82.3 | 88.5 | 495.9 | 76.6 | 95.4 | 98.2 | 61.3 | 88.8 | 94.8 | 515.1 |
| | CRCL | 78.9 | **94.8** | **97.9** | 58.7 | 83.0 | 89.2 | 502.5 | 77.8 | 96.1 | 98.5 | 63.4 | 90.3 | 95.9 | 522.0 |
| | L2RM | 76.5 | 93.7 | 97.3 | 55.5 | 81.5 | 88.0 | 492.5 | 78.4 | 95.7 | 98.3 | 62.1 | 89.1 | 94.9 | 518.5 |
| | **ASL(Ours)** | **79.2** | 93.3 | 97.0 | **59.6** | **84.3** | **90.2** | **503.8** | **79.7** | 96.9 | **98.9** | **65.3** | **90.7** | **95.9** | **527.4** |
| 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 |
| | SGR | 4.1 | 16.6 | 24.1 | 4.1 | 13.2 | 19.7 | 81.8 | 1.3 | 3.7 | 6.3 | 0.5 | 2.5 | 4.1 | 18.4 |
| | SAF | 7.4 | 19.6 | 26.7 | 4.4 | 12.2 | 17.0 | 87.3 | 13.5 | 43.8 | 48.2 | 16.0 | 39.0 | 50.8 | 211.3 |
| | NCR | 68.1 | 89.6 | 94.8 | 51.4 | 78.4 | 84.8 | 467.1 | 76.6 | 95.6 | 98.2 | 61 | 88.9 | 94.9 | 515.2 |
| | 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 |
| | RCL | 71.3 | 91.1 | 95.3 | 51.4 | 78.0 | 85.2 | 472.3 | 73.9 | 94.9 | 97.9 | 59 | 87.4 | 93.9 | 507.0 |
| | MSCN | 74.4 | 93.2 | 96.0 | 55.3 | 80.4 | 86.8 | 486.1 | 74.5 | 96 | 98.1 | 60.8 | 89.0 | 95.0 | 513.4 |
| | BiCro | 74.6 | 92.7 | 96.2 | 55.5 | 81.1 | 87.4 | 487.5 | 75.1 | 95.9 | 98.3 | 59.8 | 89.1 | 94.9 | 513.1 |
| | CRCL | 74.1 | 92.6 | 96.9 | 55.5 | 80.9 | 87.6 | 487.6 | 76.6 | 95.6 | 98.5 | 62.3 | 89.7 | 95.4 | 518.1 |
| | L2RM | 75.8 | 93.2 | 96.9 | 56.3 | 81.0 | 87.3 | 490.5 | 75.2 | 94.8 | 98.1 | 59.4 | 87.8 | 94.1 | 509.4 |
| | **ASL(Ours)** | **77.5** | **93.3** | **97.4** | **58.3** | **83.0** | **89.4** | **498.9** | **79.0** | **96.6** | **98.6** | **63.5** | **89.7** | **95.6** | **523.0** |
| 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 |
| | SGR | 1.5 | 6.6 | 9.6 | 0.3 | 2.3 | 4.2 | 24.5 | 0.1 | 0.6 | 1.0 | 0.1 | 0.5 | 1.1 | 3.4 |
| | SAF | 0.1 | 1.5 | 2.8 | 0.4 | 1.2 | 2.3 | 8.3 | 0.1 | 0.5 | 0.7 | 0.8 | 3.5 | 6.3 | 11.9 |
| | NCR | 13.9 | 37.7 | 50.5 | 11.0 | 30.1 | 41.4 | 184.6 | 0.1 | 0.3 | 0.4 | 0.5 | 1.0 | 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 |
| | RCL | 62.3 | 86.3 | 92.9 | 45.1 | 71.3 | 80.2 | 438.1 | 62.3 | 86.3 | 92.9 | 45.1 | 71.3 | 80.2 | 438.1 |
| | MSCN | 67.5 | 88.4 | 93.1 | 48.7 | 76.1 | 82.3 | 456.1 | 73.7 | 95.1 | 98.5 | 57.0 | 86.9 | 94.0 | 505.2 |
| | BiCro | 67.6 | 90.8 | 94.4 | 51.2 | 77.6 | 84.7 | 466.3 | 73.9 | 94.7 | 97.9 | 58.7 | 87.0 | 93.8 | 506.0 |
| | CRCL | 70.4 | 90.4 | 94.9 | 52.6 | 78.1 | 85.1 | 471.5 | 75.2 | 94.9 | 98.0 | 60.1 | 88.5 | 94.8 | 511.5 |
| | 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 |
| | **ASL(Ours)** | **73.9** | **92.4** | **96.2** | **54.8** | **80.7** | **87.1** | **485.1** | **77.1** | **94.8** | **98.2** | **60.7** | **88.6** | **94.8** | **514.2** |

Table 1: Comparison of different methods with noise 20%, 40%, and 60% on the Flickr30K and MS-COCO 1K. Best results are highlighted in each column.

BiCro (Yang et al. 2023a), CRCL (Qin et al. 2023) and L2RM (Han et al. 2024). For fairness in comparison, the SGR model is employed to capture the relationship between local alignments in the compared methods. Since the data in Flickr30K and MS-COCO are considered as correctly-matched pairs, we synthesize noisy pairs by randomly shuffling the descriptions corresponding to the images, where three kinds of noise rates are set by 20%, 40% and 60%. The CC152K dataset collected from the web reflects the noise conditions of real-world dataset.

**Experiment on Synthetic Noise** Table 1 shows the experimental results of different methods on the Flickr30K and MS-COCO datasets with synthetic noise rate of 20%, 40% and 60%. Note that we report the average results over 5 folds of 1K test images for MS-COCO. The results show that our proposed ASL is significantly superior to the compared baselines like NCR, DECL, CRCL and L2RM. Specifically, compared to the pioneer method NCR, our ASL improves the sum of Recall from 487.2 to 503.6 on the Flick30k with noise rate 20%, and 517.5 to 527.4 on the MS-COCO with noise rate 20%. Under the medium noise rate 40%, our ASL achieves the improvements about sum of Recall by 11.3 and 8.4 on the Flick30k compared to the baselines CRCL and L2RM, as well as the improvements by 4.9 and 13.6 on the MS-COCO. Under the high noise rate 60%, the ASL consistently outperforms the baselines at all of the metrics R@1, R@5 and R@10. The R@1 metric also demonstrates the highest accuracy, and our ASL exceeds the baseline methods on the two datasets. From the above results, we can observe ASL shows a significant gain with the noise rate increasing, and this means that our method is robust to noisy data. The explanation is that (1) we are the first to construct an asymmetric learning fashion to penalty the potentially-noisy positive samples while having a minimal effect on the optimization of negative ones; (2) The VBGMM is presented to model the distribution of per-sample loss for effectively distinguishing clean pairs from noisy ones.

**Experiment on Real-World Noise** Table 2 gives the experimental results of our proposed ASL and the baselines on the CC152K dataset in both Image→Text and Text→Image retrieval tasks. From this table, as can be seen that our method demonstrates more competitive retrieval perfor-

| 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 |
| SGR | 11.3 | 29.7 | 39.6 | 13.1 | 30.1 | 41.6 | 165.4 |
| SAF | 31.7 | 59.3 | 68.2 | 31.9 | 59.0 | 67.9 | 318.0 |
| NCR | 39.5 | 64.5 | 73.5 | 40.3 | 64.6 | 73.2 | 355.6 |
| RCL | 38.3 | 63.0 | 70.4 | 39.2 | 63.2 | 72.3 | 346.4 |
| DECL | 36.2 | 63.6 | 73.2 | 37.1 | 63.6 | 73.7 | 347.4 |
| MSCN | **40.1** | 65.7 | 76.6 | 40.6 | 67.4 | 76.3 | 366.7 |
| BiCro | 39.7 | 64.6 | 72.6 | 39.2 | 65.0 | 74.1 | 355.2 |
| L2RM | 39.5 | **66.2** | 76.0 | 41.8 | 65.9 | 74.9 | 364.3 |
| **ASL(Ours)** | 39.1 | 65.2 | **76.9** | **42.7** | **67.4** | **76.6** | **367.9** |

Table 2: Comparison of different methods on the CC152K.

mance compared to the baselines under the real-world noise condition. Specifically, our ASL achieves improvement by 1.2, 12.7 and 3.6 about the sum of Recall compared to the methods MSCN, BiCro and L2RM, respectively. In the Text→Image task, the most stringent retrieval metric R@1 achieves improvement by 3.5 and 0.9 compared to the methods BiCro and L2RM, respectively. Similarly, our ASL also shows notable retrieval performance improvements about the metrics R@5 and R@10. Additionally, compared to the pioneering NCR, our approach surpasses NCR with a 12.3 increase about the sum of Recall. These results further demonstrate that our method effectively enhances retrieval accuracy in the real-world noisy scenario.

## Experimental Analysis

**Ablation Study** To verify the effectiveness of each component in our proposed ASL, we give the Recall value under different combinations of VGBMM, the proposed loss $l_{asy}$ as well as the Warmup in Table 3. The symbol ✓ in this table means that the corresponding component is used. Noting that, if the VBGMM or $l_{asy}$ is not selected, we employ the plain GMM or triplet loss $l'_{tri}$ as its corresponding alternative. For example, the fifth row denotes that it uses the plain GMM as an alternative of VBGMM. From this table, we can observe that the retrieval Recall suffers from a dramatic degradation when the Warmup operation isn't performed, and it demonstrates the Warmup is crucial to the model's training. Meanwhile, the performance of the combination Warmup + VBGMM or Warmup + $l_{asy}$ is inferior to the proposed ASL, and these results demonstrate that all components are important to achieve competitive results. Additionally, the performance of combination Warmup + VBGMM or Warmup + $l_{asy}$ still outperforms most of existing robust methods like NCR, DECL and MSCN, which further proves the effectiveness of the proposed ASL.

**Hyper-parameter Analysis** The ASL method incorporates two main hyper-parameters including $\lambda$ and $z$ for robust image-text matching, and Fig. 3 shows its effect on the Flick30k with noise 40%. $z$ is a relaxation factor to compute the self-paced weighting optimized boundary of positive pairs. According to the results, ASL shows stability when $z$ is in the range $[2,5]$, while $z = 3$ is chosen for optimal performance. When $z$ takes a bigger value, it would

| Methods | | | Image→Text | | |
|---|---|---|---|---|---|
| Warmup | VBGMM | $l_{asy}$ | R@1 | R@5 | R@10 |
| ✓ | | | 13.9 | 37.7 | 50.5 |
| ✓ | ✓ | | 70.6 | 89.5 | 93.7 |
| ✓ | | ✓ | 71.3 | 90.4 | 94.1 |
| | ✓ | ✓ | 16.4 | 28.1 | 35.9 |
| ✓ | ✓ | ✓ | 73.9 | 92.4 | 96.2 |
| Methods | | | Text→Image | | |
| Warmup | VBGMM | $l_{asy}$ | R@1 | R@5 | R@10 |
| ✓ | | | 11.0 | 30.1 | 41.4 |
| ✓ | ✓ | | 51.2 | 77.4 | 84.6 |
| ✓ | | ✓ | 52.5 | 78.1 | 85.8 |
| | ✓ | ✓ | 13.4 | 25.9 | 30.5 |
| ✓ | ✓ | ✓ | 54.8 | 80.7 | 87.1 |

Table 3: Ablation study about different components on the Flickr30K with noise 60%.

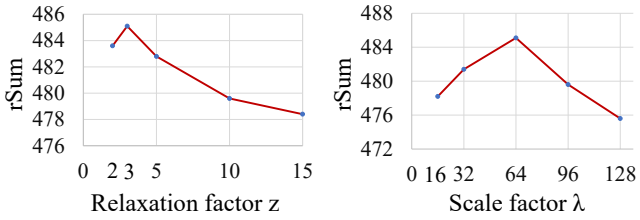


Figure 3: Analysis of different hyper-parameters on Flicker30K with 60% noise. Left: z is a relaxation factor to compute the soft boundary of each sample. Right: $\lambda$ is a scale factor to control the robustness of our proposed loss.

impose a heavy penalty on optimizing positive sample, resulting in a suboptimal performance. $\lambda$ is a scale factor to control the robustness of our proposed asymmetric dynamic matching loss. We provide the sum of Recall with the varied value from 16 to 128. The results indicate a better performance can be obtained when $\lambda = 64$. It becomes unstable with a larger value in our implementation, because a higher value would bring a stronger gradient in the proposed loss, causing the model unstable and underfit.

## Conclusion

This paper reveals a common issue that insufficient optimization of positive and negative samples is caused by the symmetric optimization in current noisy correspondence rectification methods. It proposes a novel noise-robust learning framework named ASL for image-text matching. Specifically, through independently optimizing positive and negative samples, we achieve separate optimization for potentially-noisy positive samples, thereby avoiding inappropriate or insufficient penalty on negative samples. Additionally, we present a variational Bayesian strategy in the EM algorithm to enhance the modeling capability of GMM for the distribution of per-sample loss, incorporating the prior distribution of parameters for modeling. Extensive experiments are conducted on three benchmarks, and The effectiveness of the proposed method has been validated in both synthetic and real noise scenarios.

## Acknowledgments

This work is supported in part by the National Nature Science Foundation of China (No. 62402532), in part by the Hunan Provincial Natural Science Foundation of China (No. 2024JJ6526), in part by the Science and Technology Plan of Hunan Province (No. 2023GK2013), and in part by the High Performance Computing Center of Central South University.

## References

Chakladar, D. D.; Roy, P. P.; and Chang, V. 2024. Integrated spatio-temporal deep clustering for cognitive workload assessment. *Biomedical Signal Processing and Control*, 89: 105703.

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*, volume 35, 1218–1226.

Faghri, F.; Fleet, D. J.; Kiros, J. R.; and Fidler, S. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.

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.; Miao, K.; Zheng, Q.; and Luo, M. 2023. Noisy correspondence learning with meta similarity correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7517–7526.

Han, H.; Zheng, Q.; Dai, G.; Luo, M.; and Wang, J. 2024. Learning to Rematch Mismatched Pairs for Robust Cross-Modal Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26679–26688.

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

Huang, Y.; Wang, Y.; Zeng, Y.; and Wang, L. 2022. MACK: multimodal aligned conceptual knowledge for unpaired image-text matching. *Advances in Neural Information Processing Systems*, 35: 7892–7904.

Huang, Z.; Niu, G.; Liu, X.; Ding, W.; Xiao, X.; Wu, H.; and Peng, X. 2021. Learning with noisy correspondence for cross-modal matching. *Advances in Neural Information Processing Systems*, 34: 29406–29419.

Huang, Z.; Yang, M.; Xiao, X.; Hu, P.; and Peng, X. 2025. Noise-robust vision-language pre-training with positive-negative learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1): 338–350.

Huang, Z.; Zhang, J.; and Shan, H. 2023. Twin contrastive learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11661–11670.

Jia, X.; Yang, S.; Wang, Y.; Zhang, J.; Peng, Y.; and Chen, S. 2022. Dual-view 3D reconstruction via learning correspondence and dependency of point cloud regions. *IEEE Transactions on Image Processing*, 31: 6831–6846.

Kang, W.; Mun, J.; Lee, S.; and Roh, B. 2023. Noise-aware learning from web-crawled image-text data for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2942–2952.

Karpathy, A.; and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3128–3137.

Kim, D.; Kim, N.; and Kwak, S. 2023. Improving cross-modal retrieval with set of diverse embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23422–23431.

Lee, K.-H.; Chen, X.; Hua, G.; Hu, H.; and He, X. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision*, 201–216.

Li, K.; Zhang, Y.; Li, K.; Li, Y.; and Fu, Y. 2019. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4654–4662.

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 European Conference on Computer Vision*, 740–755.

Lin, Y.; Yang, M.; Yu, J.; Hu, P.; Zhang, C.; and Peng, X. 2023. Graph matching with bi-level noisy correspondence. In *Proceedings of the IEEE/CVF international conference on computer vision*, 23362–23371.

Liu, X. B.; Kirilyuk, V.; Yuan, X.; Olwal, A.; Chi, P.; Chen, X. A.; and Du, R. 2023. Visual captions: augmenting verbal communication with on-the-fly visuals. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–20.

Nguyen, T.; Gadre, S. Y.; Ilharco, G.; Oh, S.; and Schmidt, L. 2024. Improving multimodal datasets with image captioning. *Advances in Neural Information Processing Systems*, 36.

Pan, Z.; Wu, F.; and Zhang, B. 2023. Fine-grained image-text matching by cross-modal hard aligning network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19275–19284.

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 ACM International Conference on Multimedia*, 4948–4956.

Qin, Y.; Sun, Y.; Peng, D.; Zhou, J. T.; Peng, X.; and Hu, P. 2023. Cross-modal active complementary learning with self-refining correspondence. *Advances in Neural Information Processing Systems*, 36: 24829–24840.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28.

Sarafianos, N.; Xu, X.; and Kakadiaris, I. A. 2019. Adversarial representation learning for text-to-image matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5814–5824.

Schlarmann, C.; and Hein, M. 2023. On the adversarial robustness of multi-modal foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3677–3685.

Schuster, M.; and Paliwal, K. K. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11): 2673–2681.

Shao, Z.; Yu, Z.; Wang, M.; and Yu, J. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14974–14983.

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 Annual Meeting of the Association for Computational Linguistics*, 2556–2565.

Song, H.; Kim, M.; Park, D.; Shin, Y.; and Lee, J.-G. 2022. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11): 8135–8153.

Sun, Y.; Cheng, C.; Zhang, Y.; Zhang, C.; Zheng, L.; Wang, Z.; and Wei, Y. 2020. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6398–6407.

Wang, L.; Li, Y.; Huang, J.; and Lazebnik, S. 2019a. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 394–407.

Wang, Y.; Liang, J.; Cao, D.; and Sun, Z. 2019b. Local semantic-aware deep hashing with hamming-isometric quantization. *IEEE Transactions on Image Processing*, 28(6): 2665–2679.

Wang, Y.; and Peng, Y. 2022. MARS: Learning modality-agnostic representation for scalable cross-media retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7): 4765–4777.

Wei, J.; Yang, Y.; Xu, X.; Song, J.; Wang, G.; and Shen, H. T. 2023a. Less is better: Exponential loss for cross-modal matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9): 5271–5280.

Wei, Q.; Feng, L.; Sun, H.; Wang, R.; Guo, C.; and Yin, Y. 2023b. Fine-grained classification with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11651–11660.

Wu, W.; Luo, H.; Fang, B.; Wang, J.; and Ouyang, W. 2023. Cap4video: What can auxiliary captions do for text-video retrieval? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10704–10713.

Yang, M.; Huang, Z.; Hu, P.; Li, T.; Lv, J.; and Peng, X. 2022. Learning with twin noisy labels for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14308–14317.

Yang, M.; Li, Y.; Huang, Z.; Liu, Z.; Hu, P.; and Peng, X. 2021. Partially view-aligned representation learning with noise-robust contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1134–1143.

Yang, S.; Xu, Z.; Wang, K.; You, Y.; Yao, H.; Liu, T.; and Xu, M. 2023a. 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*, 19883–19892.

Yang, W.; Chen, Y.; Li, Y.; Cheng, Y.; Liu, X.; Chen, Q.; and Li, H. 2023b. Cross-view semantic alignment for livestreaming product recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13404–13413.

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.; Xu, X.; Wang, N.; and Cao, J. 2024. UGNCL: Uncertainty-Guided Noisy Correspondence Learning for Efficient Cross-Modal Matching. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 852–861.

Zhang, H.; Mao, Z.; Zhang, K.; and Zhang, Y. 2022. Show your faith: Cross-modal confidence-aware network for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3262–3270.

Zhang, Q.; Lei, Z.; Zhang, Z.; and Li, S. Z. 2020. Context-aware attention network for image-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3536–3545.

Zhang, X.; Li, H.; and Ye, M. 2024. Negative pre-aware for noisy cross-modal matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7341–7349.