# Noise-Robust Cross-Modal retrieval via Noisy Label Re-Matching and Pseudo-Classification

Yukai Wang*, Mingyong Li*†‡
*College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China
†Chongqing Key Laboratory of Brain-Inspired Cognitive Computing
and Educational Rehabilitation for Children with Special Needs,
Chongqing Normal University, Chongqing 401331, China
‡Corresponding author: Mingyong Li, limingyong@cqnu.edu.cn
Email: 2024210516126@stu.cqnu.edu.cn

***Abstract*—Cross-modal matching relies on large-scale data, which researchers often collect through crowd-sourcing or web crawling to reduce annotation costs. However, this process inevitably introduces mismatched pairs, leading to the noisy correspondence problem. Since noisy correspondences are difficult to identify and correct, noise-robust cross-modal learning becomes a challenging task. To achieve robust cross-modal retrieval, we propose a noisy label re-matching and pseudo-classification framework. The noisy label re-matching module identifies and corrects mismatched pairs by estimating correspondence confidence in the learned representation space and refining alignment with a pre-trained vision–language model. The pseudo-classification module treats titles as category labels and generates pseudo-labels from clean samples to promote confident and balanced predictions. In addition, a confidence-based weighting strategy is introduced to adaptively suppress harmful samples during training. Extensive experiments on Flickr30K, MS-COCO, and CC120K demonstrate that our method outperforms state-of-the-art approaches.**

***Index Terms*—Cross-modal retrieval, Noisy correspondence, Noise-robust learning**

## I. Introduction

Cross-modal matching is used to establish semantic correspondence between different types of modalities, such as images and text, which lays the foundation for many applications, such as image captioning, image retrieval from text, and visual question answering [1]–[3]. In recent years, this concept has been extended to more advanced systems, such as large-scale multimodal models [4], [5] and diffusion-based generative models [6], which leverage powerful cross-modal alignment capabilities to enhance the performance of tasks like image generation, zero-sampling learning, and multimodal pre-training. [4]–[6]. Although large-scale labeled datasets like Flickr30K [7] and MSCOCO [8] provided support for early development, they were constructed manually, which was costly and inefficient. To scale up, recent datasets such as Conceptual Captions [9], Open Images [10], Visual Genome [11], and LAION [12] have been built through web crawling or crowdsourcing, but this inevitably introduces noise correspondence. That is to say, those mismatched pairs were wrongly aligned. Unlike traditional label noise that only occurs in a single modality, cross-modal noise is caused by semantic mismatch between modalities [13]–[15], which makes it more challenging to identify and correct. If this problem is not solved, this kind of noise will accumulate more and more during the training process and significantly reduce the model's anti-interference ability in high-noise scenarios [16], [17].

Many existing studies address noisy correspondences by filtering suspicious image–text pairs or correcting labels based on model predictions [18]–[20]. While these approaches achieve partial success, they suffer from inherent limitations under high-noise conditions. Specifically, their reliance on uncertain predictions often causes deep neural networks to memorize indistinguishable noisy samples, leading to performance degradation [21], [22]. Moreover, as the noise ratio increases, most methods fail to maintain stable discriminative learning, resulting in significant performance fluctuations [23], [24]. Although CLIP [4] provides strong pretrained representations, CLIP-based fine-tuning becomes increasingly unstable as noise grows: deep neural networks easily overfit noisy correspondences, resulting in degraded performance under high noise. As a result, existing methods primarily focus on noise filtering and discrimination, while largely overlooking training stability, as illustrated in Fig. 1. We argue that noise-robust fine-tuning of large-scale vision–language models requires two core capabilities: (1) effectively distinguishing noisy samples from clean ones, and (2) maintaining stable discriminative learning as noise increases.

To address these challenges, we propose a new approach termed Noisy Label Re-matching and Pseudo-classification (NRP). Under noisy correspondence settings, a small number of clean samples are more valuable than the majority of noisy ones. Motivated by this observation, NRP employs a Gaussian Mixture Model (GMM) to fit the loss distribution and distinguish noisy pairs from clean ones. To further improve reliability, high-confidence clean samples are stored in a Knowledge Base ($KB$) to help the model estimate potential negative impacts. A pretrained vision–language model is then used for semantic alignment to refine noisy samples. To fully exploit clean pairs, we introduce an auxiliary task that treats titles as class labels and generates pseudo-labels from clean samples to guide confident and balanced predictions.

Our main contributions can be summarized as follows:

- We propose Noisy Label Re-matching and Pseudo-

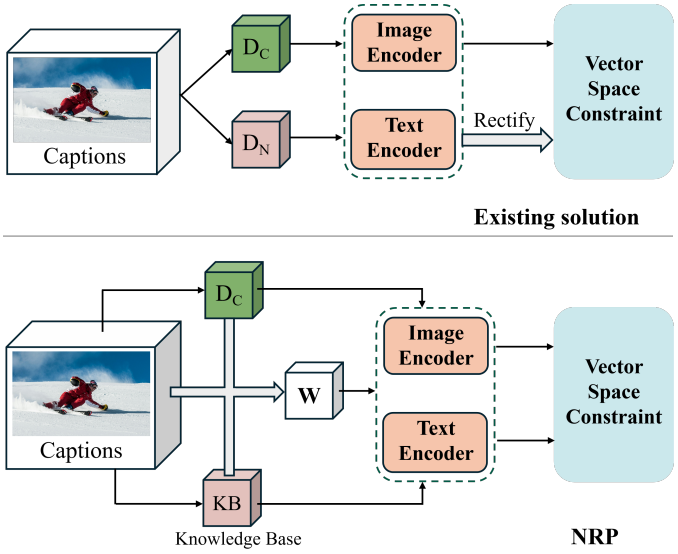


Fig. 1. Differences between existing methods and NRP. $D_c$ denotes the clean set, $D_n$ denotes the noisy set, and $w$ represents the sample weight.

classification (NRP), an adaptive strategy that leverages confidence estimation from pre-trained vision–language models and semantic alignment to identify and correct noisy correspondences.

- We introduce a pseudo-classification strategy that leverages clean-data pseudo-labels to stabilize discriminative learning under noisy supervision.
- We conduct extensive experiments on multiple datasets with both synthetic and real-world noise, demonstrating that NRP consistently outperforms state-of-the-art methods and remains stable as noise levels increase.

## II. Related Work

### A. Image-text Matching

Conventional image text matching approaches bring together data from diverse modalities to evaluate their resemblance. Early works such as [25]–[28] mainly focused on global alignment by mapping images and texts into a shared embedding space. Subsequent attention-based approaches [29]–[32] enabled fine-grained region–word matching, while recent methods [33]–[35] further modeled many-to-many relationships.

However, most of these methods implicitly assume clean and accurate image–text correspondences during training, and their performance can degrade significantly when mismatched or noisy pairs are present. Transformer-based vision–language pre-training models, such as CLIP [4], have demonstrated strong zero-shot performance. However, when adapted to downstream image-text matching tasks, their contrastive learning objective becomes sensitive to noisy supervision [16], [36]. In this paper, we employ CLIP as our backbone and introduce an anti-noise learning strategy.

### B. Cross-modal Noise-robust Learning

Some studies have addressed noisy correspondences in cross-modal tasks. Early methods [13], [15] leveraged the DNN memory effect to construct error correction networks for identifying mismatched pairs. Later approaches [14], [18] focused on improving loss functions through dynamic weighting or bidirectional similarity, while soft correspondence labels were introduced to handle uncertain matching scenarios [14], [37]. More recent works [20], [38] adopt dynamic filtering or uncertainty-aware strategies to reduce the impact of noisy supervision during training. However, many of these methods still rely on strong assumptions about alignment quality or impose uniform penalties on positive and negative pairs, which may be suboptimal in highly noisy environments.

## III. Methodology

### A. Problem Definition

We study noisy correspondences in cross-modal retrieval using image–text retrieval as a representative task. Given a dataset $D = \{(I_i, T_i)\}_{i=1}^{N}$, where $(I_i, T_i)$ denotes the $i$-th image–text pair and $N$ is the dataset size, existing approaches assume perfect alignment between image–text pairs, while real-world datasets often contain annotation noise. Image–text matching seeks to align both modalities in a shared embedding space, with similarity computed as in (1).

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

Here, $f(\cdot)$ and $g(\cdot)$ denote the feature extractors for the visual and textual modalities, respectively. Typically, matched (positive) pairs produce higher similarity scores, whereas mismatched (negative) pairs result in lower ones.

### B. CLIP in Noisy Environments

Given the strong performance of the vision–language pre-trained model CLIP [4] in cross-modal tasks, we adopt it as the backbone of the proposed NRP framework. CLIP jointly learns visual and textual representations by minimizing a symmetric cross-entropy loss $\mathcal{L}_{CE}(I_i, T_i)$, which promotes alignment between matched image–text pairs while separating mismatched ones. The loss is formulated as follows:

$$\begin{aligned} &\mathcal{L}_{CE}(I_i, T_i) = CE(I_i, T_i) + CE(T_i, I_i), \\ &CE(x_i, y_i) = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{\exp(S(x_i, y_i))}{\sum_{j=1}^{N}\exp(S(x_i, y_j))}\right). \end{aligned} \tag{2}$$

The effectiveness of this loss function relies on the assumption that $(I_i, T_i)$ forms a correct pair. However, when noisy correspondences are present in the training data, relying solely on (2) can significantly degrade model performance [21].

### C. Knowledge Base

We aim to estimate the potential negative impact of individual training samples. A straightforward approach is to compare model performance before and after training. However, evaluating the test set is computationally costly. To address

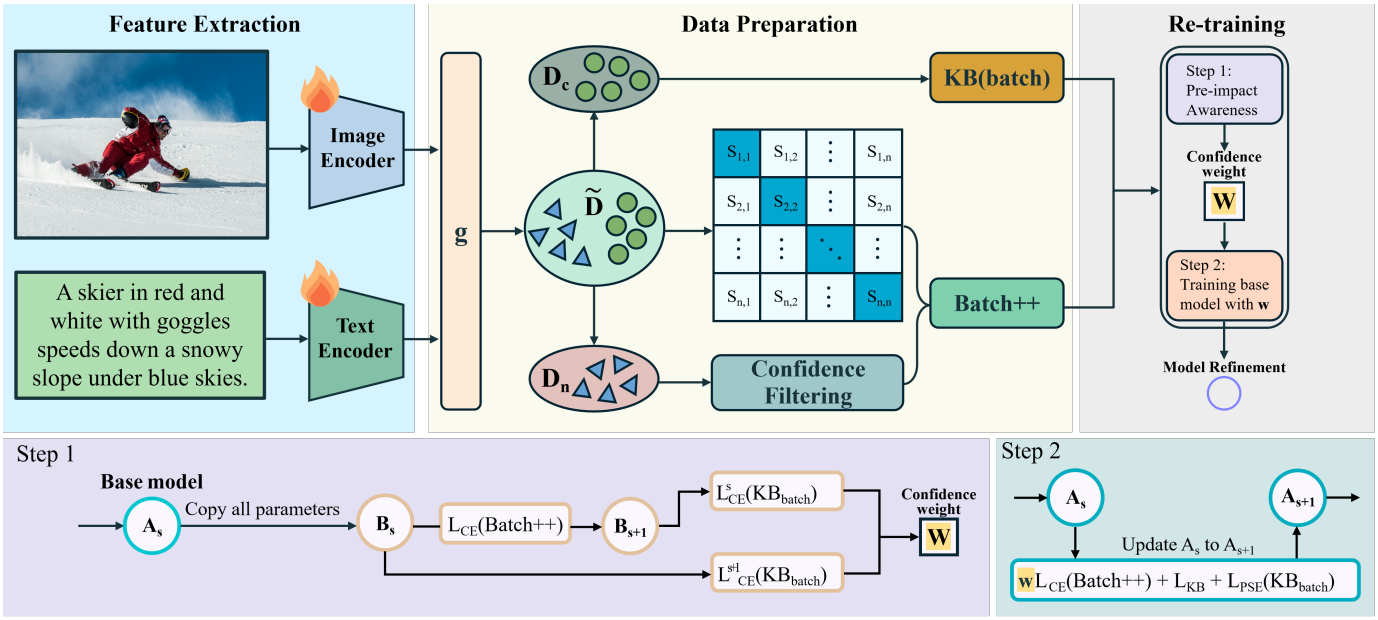


Fig. 2. Overview of the **NRP**. The framework encompasses three key components: 1) **Feature Extraction:** Training samples are processed by image and text encoders to obtain modality-specific representations. 2) **Data Preparation:** We divide the data into a strict clean set ($D_c$) and a noisy set ($D_n$) using GMM ($g$). For the $D_c$, the corresponding knowledge entries ($KB$) are selected ($KB(batch)$) as inputs. For the $D_n$, severely noisy samples are refined through noisy label re-matching, and the resulting optimized correspondences ($Batch++$) are used as inputs. 3) **Re-training:** The base model is optimized in two steps. The first step estimates a confidence weight $w$ for each sample, where a smaller $w$ indicates a higher negative impact. The second step trains the base model using the samples weighted by $w$.

this issue, we construct a Knowledge Base (KB) that provides reference evaluation entries for each sample, selected from a reliable, clean subset to enable stable performance estimation [39]. This design leverages the observation that clean samples tend to incur lower training losses than noisy ones [21]. Motivated by this observation, we analyze the loss distribution

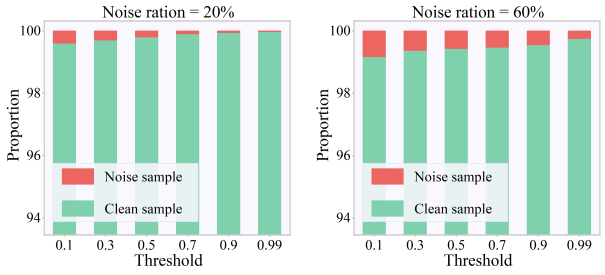


Fig. 3. Proportion of clean and noisy samples selected by GMM under varying thresholds $\beta$. A strict threshold $\beta = 0.99$ yields an almost noise-free clean set.

of all samples to identify clean pairs. Following NCR [13], we model the per-sample loss distribution in the training set using a two-component Gaussian Mixture Model (GMM).

$$p(l|\theta) = \sum_{k=1}^{K} \alpha_k \phi(l|\theta_k) \tag{3}$$

Here, $\alpha_k$ denotes the mixture coefficient, and $\phi(z|\theta_k)$ stands for the probability density function of the $k^{th}$ Gaussian element. Subsequently, the posterior probability obtained from (4) is regarded as the clean probability $p_i$ associated with the $i^{th}$ sample.

$$p_i = p(\theta_k|l_i) = \frac{p(\theta_k)p(l_i|\theta_k)}{p(l_i)} \tag{4}$$

Here, $\theta_k$ represents the Gaussian component that has a lower mean value. Samples with $p_i \geq \beta$ are regarded as clean, as shown in (5).

$$D_c = \{(I_j, T_j)|p_j \geq \beta\} \tag{5}$$

Fig. 3 shows the proportion of clean and noisy samples in dataset $D_c$ under different threshold values of $\beta$. We adopt a strict threshold of $\beta = 0.99$ to construct a nearly noise-free clean set $D_c$.

For each training sample $(I_i, T_i)$, we then retrieve corresponding evaluation entries from the clean set $D_c$ to build the KB. Specifically, for an image $I_i$, we select an image-text pair $(I_i^I, T_i^I)$ from $D_c$, where $I_i^I$ has the highest cosine similarity (1) with $I_i$. Likewise, for a text $T_i$, we choose a pair $(I_i^T, T_i^T)$ following the same principle. The resulting Knowledge Base can be expressed as $KB = (I_i^I, T_i^I), (I_i^T, T_i^T)_{i=1}^N$.

### *D. Noisy Label Re-Matching*

In this work, we propose a GMM-based framework to detect noisy pairs via confidence estimation in the learned representation space. The GMM models the distribution of loss values and assigns each sample to a mixture component, thereby enabling the identification of samples that deviate markedly from the expected distribution. Leveraging its clustering capability, the GMM designates as noisy those samples that diverge

most from the dominant distribution. The clean probability $p_i$ of the $i^{th}$ sample is determined by the posterior probability calculated using (4). A lower value of this probability implies a greater probability that the sample is noisy. We then select the lowest-confidence $\alpha N$ samples to form the noise set, defined as

$$\mathcal{S}_{\text{noise}} = \text{Top}-\lfloor \alpha N \rfloor \left( \{p_i\}_{i=1}^{N} \right). \tag{6}$$

where $\text{Top} - \lfloor \alpha N \rfloor$ denotes the indices of the $\lfloor \alpha N \rfloor$ samples with the lowest confidence scores. We set $\alpha = 0.1$ to achieve a balance between precision and coverage, yielding a representative noise set that improves the reliability of the subsequent noisy label re-matching process.

We next process the identified noisy set. Pre-trained multimodal models have already learned rich cross-modal semantic correspondences [4], which can be used to assess the semantic consistency between images and texts. When noisy labels are semantically mismatched with their corresponding images, they exhibit lower similarity in the feature space. Based on this principle, we leverage the semantic knowledge of pre-trained models by calculating the feature similarity between noisy images and all candidate texts, thereby matching the noisy image with a semantically more consistent textual description for label correction. For an image $I_h$ in the noisy set $S_{noise}$, its similarity with all candidate texts is calculated using (1).

However, multimodal datasets often contain noisy annotations, and blindly replacing them may introduce additional noise and negative effects. To avoid this, we introduce an adaptive thresholding mechanism. For each detected noisy sample, we calculate an improvement score as follows:

$$\Delta_h = S(I_h, \hat{T}_h) - S(I_i, T_i). \tag{7}$$

Here, $\hat{T}_h$ denotes the best matching text retrieved for image $I_h$, and $T_i$ represents the original noisy text. The samples with $\Delta_h \geq \beta$ are regarded as a valid match. Conversely, samples with $\Delta_h < \beta$ are not re-matched.

### *E. Pseudo-Classification*

In noisy training scenarios, mismatched image–text pairs severely hinder the model's ability to learn reliable representations from clean data, an issue often overlooked by existing methods [13]. To address this, we introduce an auxiliary objective that reinforces learning from reliable samples. The core idea is to reinterpret each image-text pair as a classification instance, where the title of the image serves as the classification label, categorical label $y \in 1, \cdots, K$, with $K$ being a predefined hyperparameter. Under such a setting, the image-text alignment task can be regarded as a $K$-way classification problem. For example, if $K = 2$, captions may be grouped into two semantic categories—such as natural scenes and biological activities. The model is then trained to correctly classify images depicting landscapes or living beings into their respective categories. To achieve this, we introduce a pseudo-classifier $C(\cdot)$ that makes use of captions from untainted data to produce pseudo-labels for supervision.

Specifically, given a clean dataset of size $M$, denoted as $(I_i^C, T_i^C)i = 1^B$, we compute pseudo-predictions $p_i^c = C(f(I_i^c))$ and $q_i^c = C(g(T_i^c))$, where $p_i^c, q_i^c \in \mathbb{R}^K+$ represent probability distributions (soft labels). Then, we compute the cross-entropy loss between the hard pseudo-labels $\hat{q}_i^c = \arg\max(q_i^c)$ and the image pseudo-predictions $p_i^c$:

$$\mathcal{L}^{\text{pse}} = \frac{1}{B} \sum_{i=1}^{B} H(\hat{q}_i^{\text{c}}, p_i^{\text{c}}). \tag{8}$$

Here, $H(P, Q)$ denotes the standard cross-entropy loss between distributions $Q$ and $P$. We adopt hard pseudo-labeling [40] to encourage confident predictions, and introduce an entropy regularization term to prevent classifier collapse by enforcing a balanced pseudo-label distribution across categories.

$$\mathcal{L}^{\text{ent}} = -\frac{1}{B} \sum_{i=1}^{B} p_i^{\text{c}} \log(\frac{1}{B} \sum_{i=1}^{B} p_i^{\text{c}}). \tag{9}$$

Our pseudo-classification loss promotes the model to learn from clean samples by modeling inter-sample similarity, thereby enhancing robustness against noisy interference from $D_n$.

### *F. Impact-aware Learning*

An important observation is that when the model is trained using a noisy dataset, its performance on corresponding clean datasets typically experiences a substantial decline [21]. Therefore, after training on a sample, we can quantify its negative impact on the model's performance by evaluating how it affects related clean samples. To measure this impact, we construct corresponding clean evaluation items for every sample, which collectively constitute a Knowledge Base ($KB$).

As illustrated in Fig. 2, during training, a batch of size $m$ and its corresponding knowledge entry set $KB_{batch} = b_1, b_2, \ldots, b_m$ are fed into the model simultaneously. At the beginning of each batch, the base model $A$ shares all parameters with model $B$. It is worth mentioning that $A$ and $B$ are updated independently throughout training. The objective of model $B$ is to measure the adverse influence of every sample within the batch by evaluating the change in model performance on $KBbatch$ after training, following previous approaches for assessing noisy correspondences [13], [16]. The loss function is used as the performance indicator, where a lower loss implies better performance on $KBbatch$.For each image–text pair $(I_k, T_k)$, the corresponding evaluation entry $b_k$ computes the losses for both image-to-text ($i2t$) and text-to-image ($t2i$) retrieval as follows:

$$\begin{aligned} q_{i2t}^k[s] &= CE(I_k^I, T_k^I) + CE(I_k^T, T_k^T), \\ q_{t2i}^k[s] &= CE(T_k^I, I_k^I) + CE(T_k^T, I_k^T). \end{aligned} \tag{10}$$

We denote states of the model prior to and subsequent to training as $B_s$ and $B_{s+1}$, correspondingly. The change in model performance resulting from training on the sample $(I_k, T_k)$ can then be calculated as follows:

$$c_k = \frac{1}{2} \left( \frac{q_{i2t}^k[s]}{q_{i2t}^k[s+1]} + \frac{q_{t2i}^k[s]}{q_{t2i}^k[s+1]} \right). \tag{11}$$

When $c_k < 1$, the post-training losses increase, indicating degraded performance on the clean pairs associated with $(I_k, T_k)$ and thus a negative influence of the sample on the model. To quantify this effect, we assign a confidence weight $w_k$ according to (12), where samples with smaller $c_k$ are down-weighted using the $\tanh(\cdot)$ function, following [40]. For samples with $c_k \geq 1$, we set $w_k = 1$. This weighting scheme enables effective estimation of each sample's adverse impact within the batch as follows:

$$w_k = \begin{cases} \tanh(c_k) & , c_k < 1 \\ 1 & , otherwise. \end{cases} \tag{12}$$

### G. Model Re-training for Robustness

After estimating the adverse effects of each sample, we update the base model $A$ to obtain a more robust version, denoted as $A_{s+1}$. To reduce the detrimental influence of negatively impactful samples during training, we re-weight the symmetric cross-entropy loss as follows:

$$\mathcal{L}_{PCE} = \frac{1}{m} \sum_{k=1}^{m} w_k \mathcal{L}_{CE}(I_k, T_k). \tag{13}$$

For these harmful samples, the associated labels frequently lack reliability. To further mitigate their negative impact on the model, we utilize the corresponding knowledge entries to guide the model toward learning accurate semantic correspondences, as shown in (14).

$$\mathcal{L}_{KB} = \frac{1}{m} \sum_{k=1}^{m} \left[ \mathcal{L}_{CE}(I_k^I, T_k^I) + \mathcal{L}_{CE}(I_k^T, T_k^T) \right]. \tag{14}$$

Therefore, the complete optimization objective for the re-training process is given by:

$$\mathcal{L}_{total} = \mathcal{L}_{PCE} + \mathcal{L}_{KB} + \mathcal{L}_{PSE} + \mathcal{L}_{ENT}. \tag{15}$$

## IV. Experiments

### A. Datasets

We evaluate our method using three widely used multi-modal datasets:

- MS-COCO: A widely recognized dataset consisting of 123,287 images, each annotated with five captions. Following previous works [13], we use the standard split with 113,287 images for training, 5,000 for validation, and 5,000 for testing.
- Flickr30K: Another well-known dataset for cross-modal learning, containing 31,783 images, each with five textual descriptions. Following the split in [13], 29,783 images are used for training, 1,000 for validation, and 1,000 for testing.
- CC120K: This dataset is a randomly sampled subset of the large-scale Conceptual Captions dataset [9], which is collected from the web and contains approximately 3%–20% noisy image–text pairs. CC120K comprises 120,851 images, each accompanied by a single caption. In our experiments, we use 118,851 images for training, while 1,000 images are allocated for validation and another 1,000 for testing.

### B. Evaluation Metrics

We use Recall@K (R@K) as the leading indicator to evaluate the retrieval performance, where R@K represents the proportion of relevant items retrieved within the top K results. To conduct a thorough evaluation, we present R@1, R@5, and R@10 for both image-to-text and text-to-image retrieval. Subsequently, we combine these metrics into a single performance indicator, rSum.

### C. Implementation Details

We adopt CLIP [4] with a ViT-B/32 backbone as the baseline and integrate our NRP module to improve robustness under noisy conditions. Both the baseline CLIP and our NRP variant are trained on a single NVIDIA RTX 3090 GPU using the AdamW optimizer [45]. We set the learning rates to $5e-7$ for CLIP and $2e-7$ for NRP, with a weight decay of 0.2. All models are trained for 5 epochs with a batch size of 256, the momentum coefficient $\beta$ is fixed at 0.99, and the hyperparameter $\alpha$ is fixed at 0.1.

### D. Comparison with State of the Arts

*1) Experiments on Simulated Noise:* To verify the efficacy of NRP, we carried out comparative trials against eight state-of-the-art methods, namely NCR [13], BiCro [18], PC2 [38], NPC [41], ReCon [42], TSVC [43], ASL [44], and fine-tuned CLIP [4]. It is crucial to emphasize that CLIP serves as the baseline for our methodology. Since the Flickr30K and MS-COCO datasets contain correctly aligned image-text pairs, we introduced noisy correspondences by randomly shuffling the captions of 20%, 40%, and 60% of the training images. The results under different noise ratios on both datasets are presented in Table I.

*2) Experiments on Real-World Noise:* We assess the proposed method using the CC120K dataset under real-world noisy conditions. This dataset contains partially mismatched pairs, making it a demanding benchmark. The comprehensive outcomes are presented in Table II. As shown, NRP still delivers competitive performance even in the presence of real noise. In particular, NRP surpasses the second-best method, NPC, by 2.4%, which confirms the effectiveness of NRP in improving retrieval accuracy.

### E. Ablation Study

*1) Impact of Components:* We conducted ablation studies on Flickr30K with 60% noise to evaluate the contribution of three components in our framework: the confidence weight $w$, the pseudo-classification loss $\mathcal{L}_{PSE}$, and the knowledge base loss $\mathcal{L}_{KB}$.

Table IV indicates that the complete model attains the best retrieval performance. This outcome showcases the supplementary impacts of all three elements. Removing either $\mathcal{L}_{PSE}$ or $\mathcal{L}_{KB}$ results in a clear drop in R@1 and overall rSum, indicating that both components provide useful additional

TABLE I

PERFORMANCE OF IMAGE–TEXT MATCHING ON THE FLICKR30K AND MSCOCO 1K ACROSS VARYING NOISE RATIOS. THE BEST SCORES ARE SHOWN IN BOLD, WHILE THE SECOND-BEST ARE UNDERLINED.

| Noise | Method | Flickr30K | | | | | | | MSCOCO 1K | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Image-to-Text | | | Text-to-Image | | | rSum | Image-to-Text | | | Text-to-Image | | | rSum |
| | | 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 [13] | 75.0 | 93.9 | 97.5 | 58.3 | 83.0 | 89.0 | 496.7 | 77.7 | 95.6 | 98.2 | 62.6 | 89.3 | 95.3 | 518.7 |
| | BiCro [18] | 78.1 | 94.4 | 97.5 | 60.4 | 84.4 | 89.9 | 504.7 | 78.8 | 96.1 | 98.6 | 63.7 | 90.3 | 95.7 | 523.2 |
| | PC2 [38] | 78.7 | 94.9 | 96.9 | 59.8 | 83.9 | 89.6 | 503.8 | 77.8 | 95.7 | 98.4 | 62.8 | 89.7 | 95.3 | 519.7 |
| | NPC [41] | 87.3 | 97.5 | 98.8 | 72.9 | 92.1 | **95.8** | 544.4 | 79.9 | 95.9 | 98.4 | 66.3 | 90.8 | **98.4** | 529.7 |
| | ReCon [42] | 80.3 | 95.3 | 97.8 | 61.6 | 85.5 | 91.3 | 511.8 | 80.9 | 96.6 | 98.8 | 65.2 | 91.0 | 96.0 | 528.6 |
| | TSVC [43] | 79.6 | 95.7 | 98.6 | 60.9 | 84.1 | 90.9 | 509.8 | 79.9 | **97.5** | 98.6 | 65.3 | 91.8 | 97.3 | 530.4 |
| | ASL [44] | 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 |
| | CLIP [4] | 82.3 | 95.5 | 98.3 | 66.0 | 88.5 | 93.5 | 524.1 | 75.0 | 93.1 | 97.2 | 58.7 | 86.1 | 97.2 | 507.3 |
| | **NRP** | **89.2** | **98.6** | **99.3** | **73.7** | **92.6** | 95.6 | **549.0** | **81.9** | 96.7 | **99.0** | **67.9** | **92.5** | 96.8 | **534.8** |
| 40% | NCR [13] | 73.5 | 92.6 | 95.8 | 55.7 | 80.3 | 86.9 | 484.8 | 76.6 | 95.6 | 98.2 | 61.0 | 88.9 | 94.9 | 515.2 |
| | BiCro [18] | 74.6 | 92.7 | 96.2 | 55.5 | 81.1 | 87.4 | 487.5 | 77.0 | 95.9 | 98.3 | 61.8 | 89.2 | 94.9 | 517.1 |
| | PC2 [38] | 75.8 | 93.5 | 96.9 | 57.5 | 81.9 | 88.2 | 493.8 | 77.4 | 95.8 | 98.4 | 62.1 | 89.4 | 95.1 | 518.2 |
| | NPC [41] | 85.6 | 97.5 | 98.4 | 71.3 | **91.3** | **95.3** | 539.4 | 79.4 | 95.1 | 98.3 | 65.0 | 90.1 | **98.3** | 526.2 |
| | ReCon [42] | 79.4 | 94.3 | 97.6 | 59.9 | 83.9 | 90.1 | 505.2 | 79.9 | 96.2 | 98.6 | 63.5 | 90.5 | 95.9 | 524.5 |
| | TSVC [43] | 77.7 | 95.3 | 98.3 | 58.8 | 83.1 | 87.1 | 500.3 | 79.0 | **97.3** | **98.6** | 65.5 | 91.3 | 96.2 | 527.9 |
| | ASL [44] | 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 |
| | CLIP [4] | 76.2 | 93.3 | 96.5 | 59.4 | 85.0 | 90.9 | 501.3 | 70.7 | 91.7 | 96.2 | 54.7 | 83.4 | 96.2 | 492.9 |
| | **NRP** | **87.0** | **97.7** | **99.1** | **71.9** | 91.1 | 95.0 | **541.8** | **80.8** | 96.2 | 98.2 | **66.0** | **91.4** | 96.5 | **529.1** |
| 60% | NCR [13] | 70.0 | 91.0 | 94.4 | 52.3 | 76.9 | 84.0 | 468.6 | 72.6 | 93.8 | 97.4 | 57.0 | 86.4 | 93.6 | 500.8 |
| | BiCro [18] | 67.6 | 90.8 | 94.4 | 51.2 | 77.6 | 84.7 | 466.3 | 73.9 | 94.4 | 97.8 | 58.3 | 87.2 | 93.9 | 505.5 |
| | PC2 [38] | 70.8 | 90.3 | 94.4 | 53.1 | 79.0 | 85.9 | 473.5 | 74.2 | 94.4 | 97.8 | 58.9 | 87.5 | 93.8 | 506.6 |
| | NPC [41] | 83.0 | 95.9 | 98.6 | 68.1 | 89.6 | 94.2 | 529.4 | **78.2** | 94.4 | 97.7 | 63.1 | 89.0 | **97.7** | 520.1 |
| | ReCon [42] | 74.3 | 93.6 | 96.6 | 55.7 | 81.6 | 88.1 | 489.9 | 77.2 | 95.9 | 98.4 | 61.8 | 89.3 | 95.2 | 517.8 |
| | TSVC [43] | 73.2 | 92.0 | 95.1 | 54.8 | 78.5 | 86.2 | 479.8 | 77.4 | **96.8** | **99.5** | 61.7 | 89.0 | 94.9 | 519.3 |
| | ASL [44] | 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 |
| | CLIP [4] | 66.3 | 87.3 | 93.0 | 52.1 | 78.8 | 87.4 | 474.9 | 67.0 | 88.8 | 95.0 | 49.7 | 79.6 | 95.0 | 475.1 |
| | **NRP** | **85.8** | **97.8** | **99.3** | **69.7** | **90.6** | **94.3** | **537.5** | 78.0 | 96.4 | 98.7 | **65.0** | **90.5** | 95.5 | **524.1** |

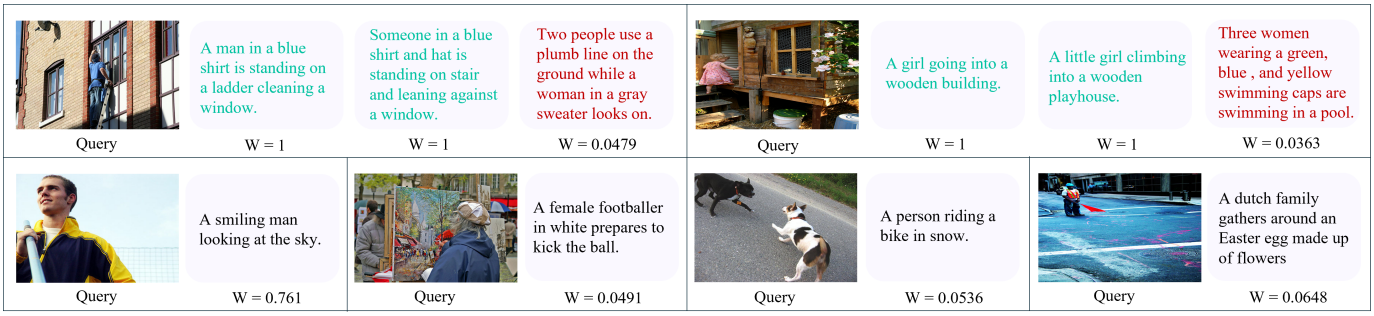


Fig. 4. Noisy correspondences in the Flickr30K training set with 40% noise. Correct and incorrect pairs are highlighted in green and red, respectively. The example illustrates the average confidence weight ($w$) across all epochs, showing that correct matches exhibit significantly higher $w$ values than noisy pairs.

TABLE II

COMPARISONS WITH REAL-WORLD NOISE ON CC120K. THE BEST INDICATORS ARE REPRESENTED IN BOLD.

| Methods | Image to Text | | | Text to Image | | | rSum |
|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | |
| NPC | 71.1 | **92.0** | **96.2** | 73.0 | 90.5 | 94.8 | 517.6 |
| CLIP | 68.8 | 87.0 | 92.9 | 67.8 | 86.4 | 90.0 | 492.9 |
| NRP | **72.9** | 91.1 | 95.6 | **73.2** | **91.3** | **95.9** | **520.0** |

supervision under noisy correspondence settings. In addition, eliminating the confidence weight $w$ significantly degrades performance, indicating that down-weighting samples with strong negative impact is essential for robust learning. These results confirm that each component independently contributes to robustness, while their integration yields the most stable and reliable performance in noisy cross-modal scenarios.

*2) Impact of hyper-parameters:* Our framework introduces the hyperparameter $\alpha$, which plays a crucial role in controlling the strictness of noise sample detection. These selected samples will serve as samples for the noise set $\mathcal{S}_{\text{noise}}$ in (6). If the value of $\alpha$is relatively small, it indicates that the selection

TABLE III
ABLATION STUDY OF THRESHOLD $\alpha$ ON FLICKR30K.

| Noise | Threshold $\alpha$ | Image to Text | | | Text to Image | | | rSum |
|---|---|---|---|---|---|---|---|---|
| | | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | |
| 20% | 0.01 | 87.8 | 98.2 | 99.2 | 73.6 | 92.2 | 95.6 | 546.6 |
| | 0.05 | 88.9 | 98.3 | 99.2 | 73.6 | 92.3 | 95.4 | 547.7 |
| | 0.1 | **89.2** | **98.6** | **99.3** | **73.7** | **92.6** | **95.6** | **549.0** |
| 60% | 0.01 | 85.2 | 96.6 | 98.7 | 69.4 | 90.0 | 93.9 | 533.8 |
| | 0.05 | **86.1** | 96.6 | 98.9 | 69.7 | 90.5 | 94.1 | 535.9 |
| | 0.1 | 85.8 | **97.8** | **99.3** | **69.7** | **90.6** | **94.3** | **537.5** |

TABLE IV
ABLATION STUDY OF DIFFERENT COMPONENTS ON THE FLICKR30K DATASET WITH 60% NOISE.

| Components | | | Image to Text | | | rSum |
|---|---|---|---|---|---|---|
| $w$ | $\mathcal{L}_{PSE}$ | $\mathcal{L}_{KB}$ | R@1 | R@5 | R@10 | |
| ✓ | ✓ | ✓ | **85.8** | **97.8** | **99.3** | **537.5** |
| ✓ | ✓ | | 84.2 | 95.9 | 98.9 | 532.1 |
| ✓ | | ✓ | 84.1 | 96.3 | 98.7 | 531.9 |
| | ✓ | ✓ | 78.5 | 95.5 | 97.7 | 506.9 |
| ✓ | | | 78.7 | 95.3 | 97.6 | 506.3 |
| | | | Text to Image | | | |
| ✓ | ✓ | ✓ | **69.7** | **90.6** | **94.3** | **537.5** |
| ✓ | ✓ | | 68.7 | 89.6 | 93.8 | 532.1 |
| ✓ | | ✓ | 68.8 | 89.7 | 93.5 | 531.9 |
| | ✓ | ✓ | 59.7 | 84.9 | 90.6 | 506.9 |
| ✓ | | | 59.3 | 84.6 | 90.8 | 506.3 |

criteria are stricter, and only the samples with the most severe noise will be retained. This ensures that the subsequent noisy label re-matching has relatively good robustness. In the experiments we conducted, we adopted a relatively conservative setting of $\alpha = 0.1$. and the corresponding experimental results are illustrated in Table III. This way, while maintaining stability in the training, it can also effectively filter out those extreme outliers.

### *F. Visualization*

To illustrate the effectiveness of NRP, Fig. 4 shows examples from Flickr30K. Noisy pairs consistently receive much lower confidence weights ($w$), forming a clear contrast with correctly annotated pairs of the same image.

## V. CONCLUSION

This paper proposes a simple yet effective Robust noisy label re-matching and pseudo-classification strategy to address the key challenges in noisy correspondence learning. Our findings indicate that, for vision-language pre-trained models, a limited quantity of clean samples holds greater value than the vast majority of noisy samples. Drawing from this observation, we apply a Noisy Label Re-Matching method that uses confidence estimation with a Gaussian Mixture Model (GMM) and an adaptive thresholding mechanism to effectively detect and correct noisy pairs, thereby enhancing the reliability of the supervision signal. Additionally, we introduce a pseudo-classification method to assist in the classification task, generating pseudo-labels from clean data to help the model capture semantic structures and suppress noise interference. Extensive experimental results demonstrate that the proposed method exhibits superior robustness in both simulated and real-world noise scenarios and significantly outperforms state-of-the-art methods across multiple benchmarks.

## REFERENCES


[1] S. Li, Z. Tao, K. Li, and Y. Fu, "Visual to text: Survey of image and video captioning," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 3, pp. 297–312, 2019.

[2] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, "From show to tell: A survey on deep learning-based image captioning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 539–559, 2022.

[3] F. Gao, Q. Ping, G. Thattai, A. Reganti, Y. N. Wu, and P. Natarajan, "Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering," in *CVPR*, 2022, pp. 5067–5077.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[5] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," https://arxiv.org/abs/2308.12966, 2023.

[6] Z. Tang, Z. Yang, C. Zhu, M. Zeng, and M. Bansal, "Any-to-any generation via composable diffusion," *Advances in Neural Information Processing Systems*, vol. 36, pp. 16 083–16 099, 2023.

[7] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[8] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[9] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *ACL*, 2018, pp. 2556–2565.

[10] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, A. Veit *et al.*, "Openimages: a public dataset for large-scale multi-label and multi-class image classification. dataset (2017)," 2017.

[11] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, no. 1, pp. 32–73, 2017.

[12] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," *arXiv preprint arXiv:2111.02114*, 2021.

[13] Z. Huang, G. Niu, X. Liu, W. Ding, X. Xiao, H. Wu, and X. Peng, "Learning with noisy correspondence for cross-modal matching," in *NeurIPS*, vol. 34, 2021, pp. 29 406–29 419.

[14] Y. Qin, D. Peng, X. Peng, X. Wang, and P. Hu, "Deep evidential learning with noisy correspondence for cross-modal retrieval," in *ACM MM*, 2022, pp. 4948–4956.

[15] H. Han, K. Miao, Q. Zheng, and M. Luo, "Noisy correspondence learning with meta similarity correction," in *CVPR*, 2023, pp. 7517–7526.

[16] X. Ma, M. Yang, Y. Li, P. Hu, J. Lv, and X. Peng, "Cross-modal retrieval with noisy correspondence via consistency refining and mining," *IEEE transactions on image processing*, vol. 33, pp. 2587–2598, 2024.

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

[18] S. Yang, Z. Xu, K. Wang, Y. You, H. Yao, T. Liu, and M. Xu, "Bicro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency," in *CVPR*, 2023, pp. 19 883–19 892.

[19] Q. Zha, X. Liu, Y.-m. Cheung, X. Xu, N. Wang, and J. Cao, "Ugncl: Uncertainty-guided noisy correspondence learning for efficient cross-modal matching," in *SIGIR*, 2024, pp. 852–861.

[20] H. Han, Q. Zheng, G. Dai, M. Luo, and J. Wang, "Learning to rematch mismatched pairs for robust cross-modal retrieval," in *CVPR*, 2024, pp. 26 679–26 688.

[21] D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. C. Courville, Y. Bengio, and S. Lacoste-Julien, "A closer look at memorization in deep networks," in *ICML*, 2017, pp. 233–242.

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

[23] Z. Dang, M. Luo, C. Jia, G. Dai, X. Chang, and J. Wang, "Noisy correspondence learning with self-reinforcing errors mitigation," in *AAAI*, vol. 38, no. 2, 2024, pp. 1463–1471.

[24] Y. Yang, L. Wang, E. Yang, and C. Deng, "Robust noisy correspondence learning with equivariant similarity consistency," in *CVPR*, 2024, pp. 17 700–17 709.

[25] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improving visual-semantic embeddings with hard negatives," in *BMVC*. Newcastle, UK: BMVA Press, 2018, p. 12.

[26] Y. Song and M. Soleymani, "Polysemous visual-semantic embedding for cross-modal retrieval," in *CVPR*. Long Beach, CA, USA: IEEE, 2019, pp. 1979–1988.

[27] L. Wang, Y. Li, J. Huang, and S. Lazebnik, "Learning two-branch neural networks for image-text matching tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 394–407, 2018.

[28] S. Qian, D. Xue, H. Zhang, Q. Fang, and C. Xu, "Dual adversarial graph neural networks for multi-label cross-modal retrieval," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2440–2448.

[29] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *ECCV*. Munich, Germany: Springer, 2018, pp. 201–216.

[30] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Visual semantic reasoning for image-text matching," in *ICCV*. Seoul, Korea (South): IEEE, 2019, pp. 4654–4662.

[31] H. Diao, Y. Zhang, L. Ma, and H. Lu, "Similarity reasoning and filtration for image-text matching," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 2, 2021, pp. 1218–1226.

[32] X. Zhang, X. Niu, P. Fournier-Viger, and X. Dai, "Image-text retrieval via preserving main semantics of vision," in *ICME*, 2023, pp. 1967–1972.

[33] S. Chun, S. J. Oh, R. S. de Rezende, Y. Kalantidis, and D. Larlus, "Probabilistic embeddings for cross-modal retrieval," in *CVPR*. Virtual: IEEE, 2021, pp. 8415–8424.

[34] S. Chun, "Improved probabilistic image-text representations," 2023.

[35] H. Li, J. Song, L. Gao, P. Zeng, H. Zhang, and G. Li, "A differentiable semantic metric approximation in probabilistic embedding for cross-modal retrieval," in *NeurIPS*, vol. 35, 2022, pp. 11 934–11 946.

[36] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *ICML*, 2021, pp. 4904–4916.

[37] Y. Qin, Y. Sun, D. Peng, J. T. Zhou, X. Peng, and P. Hu, "Cross-modal active complementary learning with self-refining correspondence," *NeurIPS*, pp. 24 829–24 840, 2024.

[38] Y. Duan, Z. Gu, Z. Ying, L. Qi, C. Meng, and Y. Shi, "Pc2: Pseudo-classification based pseudo-captioning for noisy correspondence learning in cross-modal retrieval," in *ACM MM*, 2024, pp. 9397–9406.

[39] H. Permuter, J. Francos, and I. Jermyn, "A study of gaussian mixture models of color and texture features for image classification and segmentation," *Pattern recognition*, vol. 39, no. 4, pp. 695–706, 2006.

[40] J. Li, R. Socher, and S. C. Hoi, "Dividemix: Learning with noisy labels as semi-supervised learning," *arXiv preprint arXiv:2002.07394*, 2020.

[41] X. Zhang, H. Li, and M. Ye, "Negative pre-aware for noisy cross-modal matching," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 7341–7349.

[42] Q. Zha, X. Liu, S.-J. Peng, Y.-m. Cheung, X. Xu, and N. Wang, "Recon: Enhancing true correspondence discrimination through relation consistency for robust noisy correspondence learning," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 29 680–29 689.

[43] S. Lyu, Z. Tian, Z. Ou, Y. Zhu, X. Zhang, Q. Ha, H. Luo, and M. Song, "Tsvc: Tripartite learning with semantic variation consistency for robust image-text retrieval," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 18, 2025, pp. 19 269–19 277.

[44] Y. Wang, Y. Wu, Z. Dai, C. Tian, J. Long, and J. Chen, "Noisy correspondence rectification via asymmetric similarity learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 20, 2025, pp. 21 384–21 392.

[45] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *7th International Conference on Learning Representations (ICLR-19)*. New Orleans, LA, USA: OpenReview.net, 2019.