Robust Semi-paired Multimodal Learning for Cross-modal Retrieval

Yang Qin[1], Yuan Sun[2], Xi Peng[1,3], Dezhong Peng[1,3], Joey Tianyi Zhou[4,5], Xiaomin Song[6], Peng Hu[1*]

[1]College of Computer Science, Sichuan University, Chengdu, China
[2]National Key Laboratory of Fundamental Algorithms and Models for Engineering Numerical Simulation, Sichuan University, Chengdu, China
[3]Tianfu Jincheng Laboratory, Chengdu, China
[4]CFAR, Agency for Science, Technology and Research, Singapore
[5]IHPC, Agency for Science, Technology and Research, Singapore
[6]Sichuan Newstrong UHD Video Technology Co., Ltd, Chengdu, China
{qinyang.gm, pengx.gm, joey.tianyi.zhou, penghu.ml}@gmail.com, sunyuanwork@163.com, pengdz@scu.edu.cn, songxiaomin@uptcsc.com

Abstract

Cross-modal retrieval is a fundamental application of multimodal learning that has achieved remarkable success with large-scale well-paired data. However, in practice, it is costly to collect large-scale well-paired data. To alleviate the dependence on the amount of paired data, in this paper, we study a practical learning paradigm: semi-paired cross-modal learning (SPL), which utilizes both a small amount of paired data and a large amount of unpaired data to enhance cross-modal learning directly and is more accessible in practice. To achieve this, we take image-text retrieval as an example and propose a novel Robust Cross-modal Semi-paired Learning method (RCSL) by addressing two challenges. To be specific, i) to overcome the under-optimization issue caused by too little paired data, we present Semi-paired Discriminative Learning (SDL) to fully learn visual-semantic associations from a small amount of image-text pairs by preserving the alignment and uniformity of modality representations. ii) To mine visual-semantic correspondences from unpaired data, RCSL first constructs pseudo-paired correlations across different modalities by nearest neighbor association. However, this may introduce noisy correspondences (NCs) due to inaccurate pseudo signals, which could degrade the model's performance. To tackle NCs, we devise Robust Cross-correlation Mining (RCM) based on the risk minimization criterion to robustly and explicitly learn visual-semantic associations from pseudo-paired data, thus boosting cross-modal learning. Finally, we conduct extensive experiments on four datasets, *i.e.*, three widely used benchmark datasets of Flickr30K, MS-COCO, CC152K, and a newly constructed real-world dataset Drone-SP, to demonstrate the effectiveness of RCSL under semi-paired and noisy settings.

Code — https://github.com/QinYang79/RCSL

Introduction

As one of the most fundamental tasks in multimodal community (Li and Pun 2023; Hu et al. 2023; Qin, Feng, and Zhang 2025; Qin et al. 2025; Hu et al. 2025; Feng et al. 2025b,a; Yang et al. 2025; Zou et al. 2025), cross-modal retrieval (Fartash et al. 2018; Pu et al. 2025; Lee et al. 2018; Chen et al. 2021; Qin et al. 2022) has made significant progress by bridging heterogeneous modalities (*e.g.*, vision and language) for a comprehensive understanding in the real world. For example, the primary challenge of image-text retrieval is to overcome the *heterogeneity gap* across image and text modalities, and accurately learn visual-semantic similarity to measure the matching degree.

To tackle the challenge, numerous methods (Li et al. 2019; Diao et al. 2021; Zhang et al. 2022) have been proposed and achieved remarkable performance. These methods could be roughly classified into two groups according to the alignment types, *i.e.*, global-level methods and local-level methods. The first group typically learns holistic visual and textual representations by Deep Neural Net-

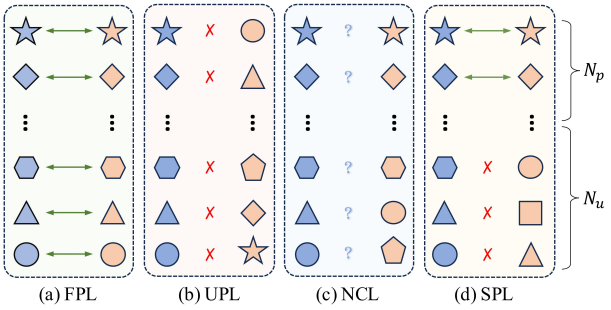


Figure 1: Examples of different cross-modal learning paradigms. Each shape represents an instance, each color denotes a modality (*e.g.*, image or text), the green solid line indicates paired correspondence, the red cross indicates unpaired correspondence, and the blue question mark means that the correspondence is unknown/noisy. (a) fully-paired cross-modal learning (FPL), where all pairs are perfectly matched; (b) unpaired cross-modal learning (UPL), where all data are unpaired; (c) noisy correspondence learning (NCL), where alignment information of all data is noisy and unknown; (d) our studied paradigm, *i.e.*, semi-paired cross-modal learning (SPL), where the number of matched pairs is much smaller than that of unmatched ones, *i.e.*, $N_p \ll N_u$.

*Corresponding author

works (DNNs) in a latent common space for global-level matching. The second one attempts to achieve fine-grained semantic alignments between image regions and word tokens for local-level matching. Although these prior methods for image-text retrieval have demonstrated promising results, almost all of them require the image-text training data to be fully paired for fully-paired cross-modal learning (FPL, see Figure 1). In other words, FPL requires all training pairs to be carefully semantically aligned between vision and language. However, collecting such a well-aligned dataset is usually cost-prohibitive or even infeasible in some real-world scenarios due to the labor-intensive manual filtering and post-processing (Sharma et al. 2018; Jia et al. 2021).

In contrast to well-paired data, unpaired data or noisy pairs are more readily available and less expensive to obtain without careful manual processing. Unlike FPL, unpaired cross-modal learning (UPL) (Li et al. 2021; Zhou et al. 2022; Huang et al. 2022) aims to build associations between vision and language without aligned image-text data. However, most of them rely on pre-trained models to bridge the visual-semantic gap. A more realistic setting is concerned with the presence of mismatched pairs in the image-text pairs, especially those collected from real-world scenarios (*e.g.*, the Internet), *i.e.*, Noisy Correspondences (NCs) (Huang et al. 2021). To address this problem, some recent studies (Huang et al. 2021; Qin et al. 2022, 2023; Zha et al. 2024; Yang et al. 2024; Zha et al. 2025a,b) leverage the memorization effect of DNNs (Arpit et al. 2017) to identify or recast noisy correspondences of image-text pairs, thus mitigating the adverse effects of NCs. However, most methods lack explicit and robust cross-correlation mining for these unpaired/noisy data, thereby limiting the improvement of retrieval performance. Motivated by this, we explore the semi-paired cross-modal learning (SPL) for image-text retrieval. Different from existing learning paradigms as illustrated in Figure 1, the proposed paradigm not only focuses on learning visual-semantic associations from limited paired data but also explores ways to gain performance improvement from a large amount of unpaired data. Although the most relevant to the paradigm is semi-supervised cross-modal retrieval, most existing works (Wang, Gong, and Yan 2024; Shen et al. 2024) mainly focus on category-based cross-modal retrieval or cross-modal hashing rather than instance-level image-text retrieval, and cannot be directly used to align unpaired instance data.

To achieve SPL, we propose a novel Robust Cross-modal Semi-paired Learning method (RCSL) for image-text retrieval as shown in Figure 2. RCSL aims to learn visual-semantic associations from limited paired data and mine cross-correlations from a large amount of unpaired data to enrich semantic information. Specifically, RCSL consists of Semi-paired Discriminative Learning (SDL) and Robust Cross-correlation Mining (RCM) to address two urgent issues for robust SPL, respectively. *First*, SDL endows intra-modal contrastive learning with the properties of alignment and uniformity to make representations stable and diverse, thus overcoming the under-optimization issue caused by insufficient paired data. *Second*, we build pseudo-paired correlations across different modalities by nearest neighbor association and treat the pseudo-paired data as paired data to learn from unpaired data. However, these pseudo-paired data inevitably introduce image-text pairs with noisy correspondences (NCs) (Huang et al. 2021) due to the lack of careful post-processing, which would mislead the model. To address such a noise problem, we proposes Robust Cross-correlation Mining (RCM) based on the risk minimization criterion to prevent the model from overfitting noisy pseudo-paired pairs, thus enhancing the robustness against NCs. This can help RCSL robustly and explicitly learn cross-modal associations from pseudo-paired data, thus mining reliable visual-semantic information from unpaired data for performance improvement. Our contributions are summarized as:

- We explore semi-paired cross-modal learning for cross-modal retrieval. Unlike traditional paradigms, this paradigm leverages both paired and unpaired cross-modal data to reduce the dependence on large-scale paired data.
- A novel Robust Cross-modal Semi-paired Learning method (RCSL) is proposed to achieve robust image-text retrieval by tackling two tricky problems, *i.e.*, under-optimization and noisy correspondence. Our RCSL reveals that image-text retrieval can benefit not only from paired inter-modal contrastive learning but also from cross-correlation mining.
- Extensive experiments are conducted on three widely-used benchmarks, namely Flickr30K, MS-COCO, and CC152K, as well as a newly constructed semi-paired dataset termed Drone-SP, to verify the robustness of our RCSL under the semi-paired and noisy setting.

Methodology

Problem Formulation

Let $\mathcal{D} = \mathcal{D}_p \cup \mathcal{D}_u = \{(I_i, T_i, y_i)\}_{i=1}^{N}$ be a semi-paired image-text training dataset, where (I_i, T_i) indicates an image-text pair, $\mathcal{D}_p = \{(I_i, T_i, 1)\}_{i=1}^{N_p}$ is the paired image-text set, and $\mathcal{D}_u = \{(I_i, T_i, 0)\}_{i=1}^{N_u}$ is the unpaired set, $y_i \in \{0, 1\}$ is the pairing label to indicate whether the pair is matched and $N = N_p + N_u$ ($N_p \ll N_u$). As illustrated in Figure 1, the training dataset used for semi-paired cross-modal learning (SPL) is different from that of existing paradigms. For example, fully-paired cross-modal learning (FPL) (Lee et al. 2018; Chen et al. 2021) is suitable for the case where all labels are known, unpaired cross-modal learning (UPL) (Huang et al. 2022) is suitable for the case where all labels are unknown, while noisy correspondence learning (NCL) is suitable for the case where labels are noisy. Among them, current FPL techniques (Lee et al. 2018; Chen et al. 2021) heavily rely on the assumption that the training pairs are correctly matched. However, a large number of unpaired image-text pairs, which are often easier and cheaper to obtain than well-paired data, are ignored. Besides, UCL methods typically rely on pre-trained models for alignment modeling, which limits the application scenarios, especially lightweight or sparsity applications.

For NCL, existing methods almost recommend discarding noisy pairs or reducing the negative impact of noisy ones,

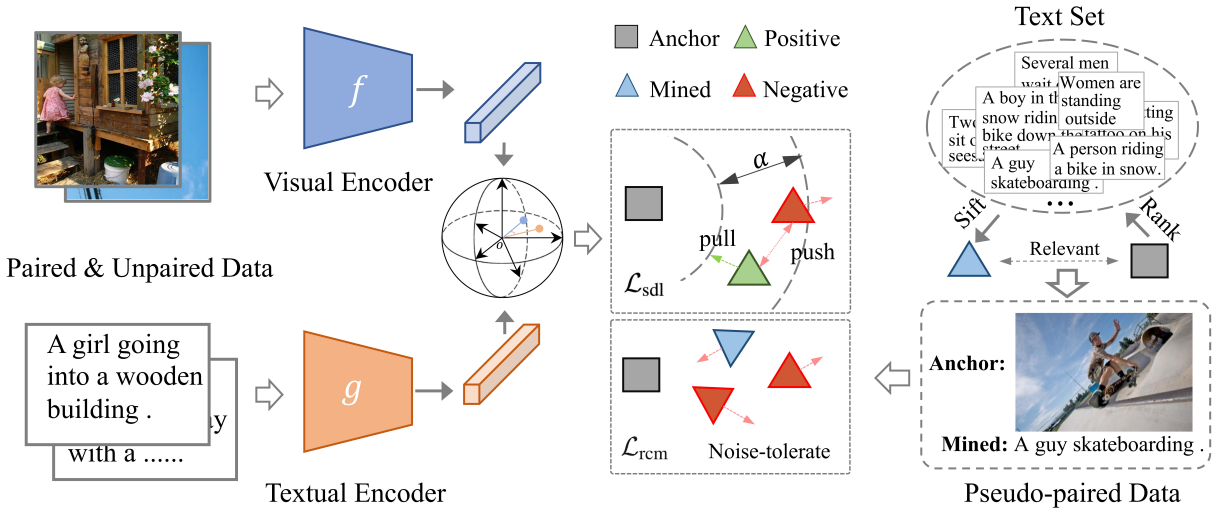


Figure 2: The overview of our RCSL includes Semi-paired Discriminative Learning (SDL, $\mathcal{L}_{\text{sdl}}$) and Robust Cross-correlation Mining (RCM, $\mathcal{L}_{\text{rcm}}$). SDL aims to project different modalities into a latent common space by keeping good properties of contrastive learning, thus capturing rich cross-modal associations. RCM mines the visual-semantic correlations from unpaired data to robustly learn the potential paired information, thereby facilitating cross-modal learning.

thus not fully and explicitly utilizing the semantic information of valuable unpaired data (*e.g.*, confident ones). In this paper, we explore SPL, where the training set $\mathcal{D}$ includes both paired and unpaired data.

To achieve SPL, we propose a Robust Cross-modal Semi-paired Learning method (RCSL) to learn the visual-semantic associations from paired and unpaired data, which could be trained in a batch-by-batch manner. For a mini-batch $\mathcal{B} = \{\mathcal{I}_B, \mathcal{T}_B, \mathcal{Y}_B\} = \{(I_i, T_i, y_i)\}_{i=1}^{K}$ randomly sampled from $\mathcal{D}$, the joint overall loss is

$$\mathcal{L}_{\text{overall}}(\mathcal{B}) = \mathcal{L}_{\text{sdl}}(\mathcal{B}) + \mathcal{L}_{\text{rcm}}(\mathcal{B}) \tag{1}$$

where $\mathcal{L}_{\text{sdl}}$ is the loss for semi-paired discriminative learning to fully learn accurate visual-semantic associations from the semi-paired data and $\mathcal{L}_{\text{rcm}}$ is the robust cross-correlation mining loss to learn the potential cross-modal alignments in unpaired data. The framework of our RCSL is illustrated in Figure 2, and the full training process is provided in the supplementary material

Semi-paired Discriminative Learning

Given a cross-modal model $\mathcal{M}(\Theta_v, \Theta_t) = \{f, g\}$, where f and g are the visual and textual encoders and Θ_v, Θ_t are the corresponding network parameters, respectively. For any image-text pair $(I_i, T_j) \in \mathcal{D}$, the visual-semantic similarity could be measured by the cosine similarity, *i.e.*, $S(I_i, T_j) = f(I_i)^\top g(T_j) / \|f(I_i)\| \|g(T_j)\|$, where $f(I_i) \in \mathbb{R}^{d\times 1}$ and $g(T_j) \in \mathbb{R}^{d\times 1}$ are the holistic representations computed by modality-specific encoders and d is the dimensionality of the latent common space. Thus, the learning objective of $\mathcal{M}$ is to maximize the similarities of positive pairs while minimizing those of negative ones, which is commonly realized by the inter-modal contrastive learning loss, *e.g.*, the Triplet Ranking Loss (Fartash et al. 2018) (TRL). TRL exploits online hard negative mining to enforce positive pairs to be distant from any negative ones by at least a given positive margin α, which is defined as:

$$\begin{aligned}\mathcal{L}_{\text{trl}}(I_i, T_i) = &\left[\alpha - S(I_i, T_i) + S(I_i, \hat{T}_h)\right]_+ \\ &+ \left[\alpha - S(I_i, T_i) + S(\hat{I}_h, T_i)\right]_+ ,\end{aligned} \tag{2}$$

where $\hat{T}_h$ and $\hat{I}_h$ are the corresponding hardest samples of I_i and T_i, respectively, and $[x]_+ \equiv \max(x, 0)$. However, some studies (Zhang et al. 2023b) have shown that TRL is prone to poor/insufficient learning in early training, thus leading to suboptimal performance. This issue is trickier in SPL (*e.g.*, the results of CHAN in Table 2) due to the sparsity of limited paired data. We think the reason is that TRL cannot effectively learn a feature distribution that preserves as much information as possible from semi-paired data. Inspired by the instance-wise contrastive learning (Wang and Isola 2020), we propose to reconsider the properties of *Alignment* ($\mathcal{A}$) and *Uniformity* ($\mathcal{U}$) to alleviate the above problem. More specifically, under the semi-paired setting, the two properties could be re-expressed as follows:

- *Alignment*: The positive samples should be close together on the hypersphere. For the semi-paired mini-batch $\mathcal{B}$, the alignment is defined as $\mathcal{A}(\mathcal{B}) = \mathbb{E}_{(I_i,T_i)^{y_i=1}_{\sim}\mathcal{B}}\left[\text{Dist}\left(\boldsymbol{x}_i, \boldsymbol{y}_i\right)\right]$, where $\mathbb{E}[\cdot]$ is the expected operator, $\boldsymbol{x}_i = \ell_2(f(I_i))$, $\boldsymbol{y}_i = \ell_2(g(T_i))$, $\ell_2(\cdot)$ is the L2-normalized function, and $\text{Dist}(\boldsymbol{x}_i, \boldsymbol{y}_i) = \|\boldsymbol{x}_i - \boldsymbol{y}_i\|_2^2$ is the function to measure the distance of samples in the hypersphere. The smaller $\mathcal{A}(\mathcal{B})$ is, the better for cross-modal learning.
- *Uniformity*: It indicates that the cross-modal samples should be uniformly distributed on the hypersphere to seek a feature distribution that preserves maximal information. For the semi-paired mini-batch $\mathcal{B} =$

$\{\mathcal{I}_B, \mathcal{T}_B, \mathcal{Y}_B\}$, The uniformity is defined as:

$$\mathcal{U}(\mathcal{B}) = \frac{1}{2}\Big(\log\big(\mathbb{E}_{(I_i,I_j)^{i\neq j}_{\sim}\mathcal{I}_B}\big[\exp(-2\,\mathrm{Dist}(\boldsymbol{x}_i,\boldsymbol{x}_j))\big]\big) + \log\big(\mathbb{E}_{(T_i,T_j)^{i\neq j}_{\sim}\mathcal{T}_B}\big[\exp(-2\,\mathrm{Dist}(\boldsymbol{y}_i,\boldsymbol{y}_j))\big]\big)\Big). \tag{3}$$

Likewise, the smaller $\mathcal{U}(\mathcal{B})$ is, the better for cross-modal learning.

Then, we utilize the *Alignment* and *Uniformity* as regularizations to empower semi-paired learning for better properties of representations. For the training mini-batch $\mathcal{B}$ with the size of K, the learning objective is defined as:

$$\mathcal{L}_{\text{sdl}}(\mathcal{B}) = \mathcal{L}^{\dagger}_{\text{trl}}(\mathcal{B}) + \mathcal{L}_{\text{reg}}(\mathcal{B}), \tag{4}$$

where $\mathcal{L}^{\dagger}_{\text{trl}}$ is used to capture inter-modal relevance from paired data and $\mathcal{L}_{\text{reg}}$ is the regularization loss to keep good alignment and uniformity on the hypersphere. $\mathcal{L}^{\dagger}_{\text{trl}}$ is written as:

$$\mathcal{L}^{\dagger}_{\text{trl}}(\mathcal{B}) = \sum_{i=1}^{K} y_i \Big(\big[\alpha - S(I_i,T_i) + S(I_i,\bar{T}_i)\big]_+ + \big[\alpha - S(I_i,T_i) + S(\bar{I}_i,T_i)\big]_+ \Big), \tag{5}$$

where $\bar{T}_i = \underset{\bar{T}\neq T_i}{\operatorname{argmax}} S(I_i,\bar{T}), \bar{T} \in \mathcal{B}\cap\mathcal{D}_p$, and $\bar{I}_i = \underset{\bar{I}\neq I_i}{\operatorname{argmax}} S(\bar{I},T_i), \bar{I} \in \mathcal{B}\cap\mathcal{D}_p$. The regularization loss is:

$$\mathcal{L}_{\text{reg}}(\mathcal{B}) = \mathcal{A}(\mathcal{B}) + \mathcal{U}(\mathcal{B}), \tag{6}$$

Thanks to $\mathcal{L}_{\text{sdl}}$, our RCSL could fully learn the accurate visual semantic associations from limited paired data and keep good properties of contrastive learning, which will be the basis and guarantee for subsequent cross-correlation mining. While the $\mathcal{L}_{\text{sdl}}$ is good at exploiting known paired data to fully learn the multimodal information, it does not explicitly model and learn possible potential associations from a large amount of unpaired data, thus limiting the performance improvement. In the next section, we provide a solution for directly learning visual-semantic associations from large amounts of unpaired data to robustly exploit inter-modal knowledge, substantially enhancing the ability to benefit from semi-paired image-text data.

Robust Cross-correlation Mining

To leverage unpaired data and facilitate semi-paired cross-modal learning, we establish pseudo pairs from the unpaired data and mine the latent cross-modal associations among them. To be specific, we pair all the visual and textual samples from $\mathcal{D}_u$ according to the ranking results of image-text similarities. For example, given an image I_i in $\mathcal{D}_u$, the textual sample pseudo-paired with it is $T'_i = \text{argmax}_{T'}\left(\{S(I_i,T') \mid T' \in \mathcal{T}_u\}\right)$, and $\mathcal{T}_u$ is the text set in $\mathcal{D}_u$. Similarly, we could easily obtain the pseudo-paired image I'_i of T_i, *i.e.*, $I'_i = \text{argmax}_{I'}\left(\{S(I',T_i) \mid I' \in \mathcal{I}_u\}\right)$, and $\mathcal{I}_u$ is the image set in $\mathcal{D}_u$.

Although we can construct pseudo-paired associations for unpaired data through similarity ranking directly, the constructed pseudo-paired data may have a large number of image-text pairs with noisy correspondences (Huang et al. 2021; Qin et al. 2022), which would mislead the model to conduct harmful cross-modal learning. To refine cross-modal learning robustly, we propose a robust cross-correlation mining (RCM) loss $\mathcal{L}_{\text{rcm}}$. Given a mini-batch $\mathcal{B} = \{\mathcal{I}_B, \mathcal{T}_B, \mathcal{Y}_B\} = \{(I_i,T_i,y_i)\}_{i=1}^{K}$, we can re-construct two minded mini-batches with pseudo cross-correlations, which are represented as:

$$\begin{cases} \mathcal{B}_t = (\mathcal{B}\cap\mathcal{D}_p)\cup\{(I_i,T'_i)\}_{i=1}^{K_u}, & \forall I_i \in \mathcal{I}_B, \\ \mathcal{B}_v = (\mathcal{B}\cap\mathcal{D}_p)\cup\{(I'_i,T_i)\}_{i=1}^{K_u}, & \forall T_i \in \mathcal{T}_B, \end{cases} \tag{7}$$

where T'_i/I'_i are the pseudo-paired samples for I_i/T_i and K_u is the number of unmatched pairs. Take the image-to-text direction as an example, the robust mining loss for the pseudo-paired texts in the mini-batch $\mathcal{B}_t$ is

$$\mathcal{L}^{t}_{\text{rcm}}(\mathcal{B}) = \mathcal{L}_{\text{m}}(\mathcal{B}_t) = \frac{1}{2}\sum_{i=1}^{K}\Big[(1-p^{i2t}_{ii}) + (1-p^{t2i}_{ii})\Big], \tag{8}$$

where p^{t2i}_{ii} are the bidirectional matching probabilities for pair $(I_i,T_i) \in \mathcal{B}_t$, which are defined as:

$$p^{i2t}_{ii} = \frac{\exp(S'_{ii}/\tau)}{\sum_{T_j\in\mathcal{B}_t}\exp(S'_{ij}/\tau)},\ p^{t2i}_{ii} = \frac{\exp(S'_{ii}/\tau)}{\sum_{I_j\in\mathcal{B}_t}\exp(S'_{ji}/\tau)}. \tag{9}$$

Note that S'_{ij} is defined as:

$$S'_{ij} = \begin{cases} S(I_i,T_j), & \text{if } (I_i,T_j) \in \mathcal{D}_p, \\ S(I_i,T'_j), & \text{if } I_i \in \mathcal{I}_p \text{ and } T_j \in \mathcal{T}_u, \\ S(I'_i,T_j), & \text{if } I_i \in \mathcal{I}_u \text{ and } T_j \in \mathcal{T}_p, \end{cases} \tag{10}$$

where $\mathcal{I}_p/\mathcal{T}_p$ is the image/text set in $\mathcal{D}_p$ and $\mathcal{I}_u/\mathcal{T}_u$ is the image/text set in $\mathcal{D}_u$, respectively. Likewise, the robust mining loss $\mathcal{L}^{v}_{\text{rcm}}(\mathcal{B})$ for the pseudo-paired images in the mini-batch $\mathcal{B}_v$ could be computed like Equation (8). Without loss of generality, we formulate the final robust cross-correlation mining loss as follows:

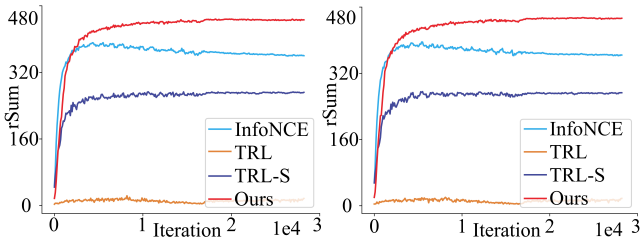
$$\mathcal{L}_{\text{rcm}}(\mathcal{B}) = \mathcal{L}^{t}_{\text{rcm}}(\mathcal{B}) + \mathcal{L}^{v}_{\text{rcm}}(\mathcal{B}). \tag{11}$$

The intuition behind using $\mathcal{L}_{\text{rcm}}(\mathcal{L}_m)$ lies in improving the noise tolerance of the loss function for those mined pseudo-paired data that may be accompanied by noisy correspondence. To be convincing, we conduct the following theoretical analysis, *i.e.*, $\mathcal{L}_{\text{m}}$ is robust against noisy correspondences. Thus, our RCSL could robustly learn the accurate visual-semantic associates from the mined informative cross-correlations, thus improving the performance of image-text retrieval.

Theoretical Analysis. First, to facilitate analysis, we provide a definition for noisy correspondence as shown in Definition 1. Based on the risk minimization theory (Manwani and Sastry 2013), the loss function is noise-tolerant *i.f.f.* the risk of learning with noisy annotations has a shared global minimizer of the risk under no-noise case.

Definition 1 *Flowing (Qin et al. 2023), we assume that noisy correspondence is uniform and define it for any pair* (I_i,T_j) *as:*

$$\tilde{c}_{ij} = \begin{cases} c_{ij} & \textit{with probability } (1-\eta_{ij}), \\ 1-c_{ik} & \textit{with probability } \bar{\eta}_{ik}, \forall k \neq j. \end{cases} \tag{12}$$



(a) Validation (b) Testing

Figure 3: The performance versus iterations on the Flickr30K dataset with 60% noise. 'InfoNCE' means the InfoNCE loss, 'TRL' is the triplet ranking loss with the hardest sample mining (Fartash et al. 2018), and 'TRL-S' is the summation version of TRL that considers all negative samples.

For all pairs, conditioned on that if $i = j$ then $c_{ij} = 1$ else $c_{ij} = 0$, we have $\sum_{j \neq k} \bar{\eta}_{ik} = \eta_{ij}$, $\eta_{ij} = \eta$, and $\bar{\eta}_{ik} = \frac{\eta}{N-1}, \forall k \neq j$, where η represents the noise rate.

For the loss $\mathcal{L}_m(I_i, T_i) = \frac{1}{2}\Big[(1 - p_{ii}^{i2t}) + (1 - p_{ii}^{t2i})\Big]$, the relationship between the noisy risk and the clean risk is

$$R_{\mathcal{L}_m}^{\eta}(f) = (1 - \frac{N\eta}{N-1}) R_{\mathcal{L}_m}(f) + \eta, \tag{13}$$

where f is the decision function. Given a global minimizer f_η^* of $R_{\mathcal{L}_m}^{\eta}(f)$, we have $R_{\mathcal{L}_m}^{\eta}(f_\eta^*) - R_{\mathcal{L}_m}^{\eta}(f) = (1 - \frac{N\eta}{N-1})(R_{\mathcal{L}_m}(f_\eta^*) - R_{\mathcal{L}_m}(f)) \leq 0$. Thus, f_η^* is also the global minimizer of $R_{\mathcal{L}_m}(f)$ *i.f.f.* $n \leq \frac{N-1}{N}$, which gives Lemma 1. Different from $\mathcal{L}_m$, none of the widely-used contrastive losses, *i.e.*, InfoNCE, TRL, and TRL-S, can share a minimizer on clean risk and noise risk, thus they are not robust. To demonstrate this, we visualize their performances versus iterations on the noisy dataset in Figure 3. From the results, except for our $\mathcal{L}_m$, all others more or less produced noise overfitting, and TRL even brings a failed cross-modal learning, which is consistent with the above theoretical analysis.

Lemma 1 *In an instance-level image-text retrieval problem, $\mathcal{L}_m$ is noise-tolerant against uniform noisy correspondences i.f.f. noise rate $\eta < \frac{N-1}{N}$.*

RCSL for Noisy Correspondence Learning

Obviously, our RCSL can be used to handle NCs by partitioning the data. We perform a preliminary partitioning of the data through a warm-up process and treat the partitioned clean set as a paired set and the noisy set as an unpaired set to implement semi-paired learning. We call the variant RCSL-NC for a clearer expression. Due to space limitations, we put more details in the supplementary material.

Experiments

In this section, we conduct comprehensive experiments on three widely-used benchmark datasets, *i.e.*, Flickr30K (Young et al. 2014), MS-COCO (Lin et al. 2014), and CC152K (Huang et al. 2021), to demonstrate the superiority and effectiveness of our RCSL. Besides, we construct a new benchmark to evaluate semi-paired learning, *i.e.*, Drone-SP, which is a special dataset for drone image-text retrieval. Note that due to space limitations, we put more comparative experiments and exploratory experimental results in the supplementary material.

Datasets	Paired	#Training	#Validation	#Testing
Flickr30K	25K	5,000/24,000	1,000	1,000
	2.5K	500/28,500	1,000	1,000
MS-COCO	25K	5,000/108,287	5,000	5,000
	2.5K	5,00/112,787	5,000	5,000
Drone-SP	962	962/5,310	500	500

Table 1: Brief statistics of the semi-paired datasets used in our experiments. "Paired" means the paired image-text pairs in the semi-paired dataset. The former of "/" is the number of training images in the paired set, and the latter is the number of training images in the unpaired set. Each image in the Flickr30K and MS-COCO datasets has five paired captions.

Datasets and Evaluation Protocols

Datasets For the semi-paired setting, we utilize the Flickr30K (Young et al. 2014), MS-COCO (Lin et al. 2014), and Drone-SP datasets to evaluate our methods. For Flickr30K and MS-COCO, we construct unpaired data pairs by randomly shuffling the captions. Table 1 shows the brief statistics of the partitions under the semi-paired setting. For noise setting, following (Huang et al. 2021), we inject unmatched correspondences of different ratios by proportionally shuffling the captions on the Flickr30K and MS-COCO training datasets, *i.e.*, 20%, 40%, 60%, and 80% noise rates. Besides, we evaluate our method on the real noisy dataset of CC152K to further verify the robustness. All dataset details are provided in the supplementary material.

Evaluation Protocols Following (Qin et al. 2022), we adapt Recall at K (R@K=1, 5, and 10) and their sum (rSum) to evaluate the performance of bidirectional retrievals. Specifically, R@K is defined as the proportion of the correct items in the top-K retrieved results. Our experiments were conducted on Nvidia GeForce RTX 3090 and A800 GPUs.

Comparisons with State-of-the-Arts

In this section, we evaluate our RCSL by comparing it with ten baselines on the two benchmarks with the same semi-paired setting, including the global-level methods: VSE∞ (CVPR'21) (Chen et al. 2021), 2AD (ACL'23) (Zhang et al. 2023b), HREM (CVPR'23) (Fu et al. 2023), ESA (TCSVT'23) (Zhu et al. 2023), and FEM (ICASSP'24) (Wang, Yin, and Ramakrishnan 2024); The local-level methods: NAAF (CVPR'22) (Zhang et al. 2022), RCAR (TIP'23) (Diao et al. 2023), CHAN (CVPR'23) (Pan, Wu, and Zhang 2023), LAPS (Fu et al. 2024), and X-Dim (Zhang et al. 2023a). Since these baselines cannot exploit unpaired data, we train them with the paired training set. To verify the effectiveness of the variant RCSL-NC, we compare it with four robust baselines against noisy corre-

		Flickr30K 1K							MS-COCO 5-fold 1K							MS-COCO 5K						
		Image⟶Text			Text⟶Image				Image⟶Text			Text⟶Image				Image⟶Text			Text⟶Image			
Paired	Methods	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	R@1	R@5	R@10	R@1	R@5	R@10	rSum
25K	VSE∞('21)	51.4	77.2	85.8	35.2	63.3	73.9	386.8	49.8	80.9	89.6	36.7	70.2	81.8	409.0	27.5	54.8	67.4	18.2	42.1	54.7	264.7
	NAAF('22)	56.1	83.8	91.1	44.6	72.1	80.8	428.5	54.1	85.5	93.0	44.3	77.0	87.4	441.3	29.4	59.8	73.3	23.9	50.6	63.0	300.0
	RCAR('23)	35.9	65.7	78.4	25.1	54.5	67.6	327.2	41.4	75.1	86.9	31.7	66.9	80.4	382.4	19.6	46.0	59.2	14.5	36.6	49.9	225.8
	2AD('23)	31.9	60.5	71.7	45.0	73.1	82.0	364.2	34.4	68.4	80.8	46.1	**78.6**	**89.3**	397.6	16.3	39.8	52.6	23.5	50.5	64.1	246.8
	CHAN('23)	56.5	81.0	89.8	40.8	70.7	80.4	419.2	53.6	84.3	92.5	41.8	76.2	87.5	435.9	29.1	59.2	72.0	22.1	47.5	61.2	291.1
	HREM('23)	57.9	82.7	89.5	41.3	69.6	79.1	420.1	54.1	84.0	92.2	39.8	73.8	85.4	429.3	31.0	59.6	71.7	20.2	45.6	58.7	286.8
	ESA('23)	60.4	84.1	89.7	41.1	69.9	78.9	424.1	33.2	61.6	73.1	24.7	52.5	66.0	311.1	17.8	36.6	46.5	11.6	28.1	38.2	178.8
	LAPS('24)	45.3	75.5	84.0	34.2	64.0	75.0	378.0	46.5	79.3	89.8	35.5	71.6	84.3	407.3	22.7	50.7	64.0	16.6	40.8	54.7	249.5
	X-Dim('24)	56.9	81.4	88.4	41.0	67.9	78.0	413.6	55.3	85.1	92.9	42.6	75.8	86.8	438.5	31.0	61.3	73.5	22.9	48,2	61.3	298.2
	FEM('24)	56.8	83.1	90.3	38.8	67.4	77.4	413.8	55.8	84.8	92.8	40.0	74.6	85.9	433.9	31.7	60.2	72.3	20.3	45.7	58.7	288.9
	RCSL	**63.6**	**88.1**	**92.9**	**46.2**	**73.9**	**82.7**	**447.4**	**60.6**	**88.2**	**94.7**	**46.4**	**78.6**	87.7	**456.2**	**36.7**	**66.2**	**78.5**	**25.7**	**53.2**	**65.6**	**325.9**
2.5K	VSE∞('21)	6.5	16.6	24.9	6.6	18.9	27.6	101.1	11.8	34.1	47.8	9.1	28.4	40.9	172.1	3.7	12.7	20.5	3.0	10.2	16.7	66.8
	NAAF('22)	23.2	49.1	62.0	18.5	42.2	53.9	248.9	24.4	56.7	72.4	21.6	52.4	68.3	295.8	9.4	26.8	39.2	8.7	24.5	35.3	143.9
	RCAR('23)	2.5	8.5	14.9	2.8	10.5	16.3	55.5	10.6	32.0	45.1	6.9	22.7	34.2	151.5	3.0	11.3	18.2	2.2	7.8	12.8	55.3
	2AD('23)	6.3	18.9	27.6	8.7	24.5	35.2	121.2	9.4	29.1	42.4	12.3	35.9	49.8	178.9	2.9	10.4	16.9	3.8	13.1	21.0	68.1
	CHAN('23)	1.2	4.5	9.0	1.1	4.4	8.4	28.6	16.6	48.0	64.7	16.9	45.8	62.6	254.6	4.1	17.2	28.4	6.2	18.9	29.2	104.0
	HREM('23)	20.7	44.6	56.7	15.1	38.1	51.3	226.5	19.7	48.2	63.6	15.7	42.9	59.3	249.4	7.3	21.0	31.7	5.5	17.1	26.7	109.3
	ESA('23)	16.6	40.2	52.2	11.9	32.4	45.6	198.9	11.7	32.1	44.8	9.5	28.0	40.6	166.7	3.7	12.4	19.5	3.2	10.3	16.5	65.6
	LAPS('24)	15.2	38.8	52.3	12.8	32.7	44.7	196.5	19.6	48.7	63.7	15.6	44.5	60.8	252.9	6.8	20.1	30.3	5.2	17.6	27.5	107.5
	X-Dim('24)	13.0	35.9	47.6	11.2	30.6	42.4	180.7	14.0	40.7	57.3	12.8	37.4	53.5	215.7	4.4	15.1	24.1	3.9	13.9	22.2	83.6
	FEM('24)	16.9	42.0	54.5	8.6	25.3	36.0	183.3	21.1	51.1	66.5	13.1	37.9	53.0	242.7	7.8	22.2	33.8	7.8	22.2	33.8	127.6
	RCSL	**25.5**	**51.8**	**63.5**	**18.2**	**42.6**	**54.9**	**256.5**	**34.6**	**66.4**	**79.3**	**26.5**	**59.3**	**73.5**	**339.6**	**15.7**	**37.4**	**49.8**	**11.3**	**30.4**	**42.6**	**187.2**

Table 2: Performance on the Flickr30K and MS-COCO datasets under the semi-paired setting. The best results are in **bold**.

		Flickr30K							MS-COCO						
		Image → Text			Text → Image				Image → Text			Text → Image			
Noise	Methods	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%	NCR('21)	76.7	93.9	96.9	57.5	82.8	89.2	497.0	77.0	95.6	98.1	61.5	89.3	95.1	516.6
	DECL('22)	75.6	93.8	97.4	58.5	82.9	89.4	497.6	77.1	95.9	98.4	61.6	89.1	95.2	517.3
	MSCN('23)	77.4	94.9	97.6	59.6	83.2	89.2	501.9	78.1	**97.2**	**98.8**	**64.3**	**90.4**	**95.8**	**524.6**
	CREAM('24)	77.4	95.0	97.3	58.7	84.1	89.8	502.3	**78.9**	96.3	98.6	63.3	90.1	**95.8**	523.0
	RCSL-NC	**80.6**	**95.3**	**97.9**	**60.2**	**85.4**	**91.2**	**510.6**	78.8	96.0	98.4	63.0	90.3	95.6	522.1
40%	NCR('21)	75.3	92.1	95.2	56.2	80.6	87.4	486.8	76.5	95.0	98.2	60.7	88.5	95.0	513.9
	DECL('22)	72.5	93.1	97.0	55.8	81.2	88.1	487.7	77.1	95.7	98.3	61.5	89.2	**95.3**	517.1
	MSCN('23)	74.4	**94.4**	96.9	57.2	81.7	87.6	492.2	74.8	94.9	98.0	60.3	88.5	94.4	510.9
	CREAM('24)	76.3	93.4	97.1	57.0	82.6	88.7	495.1	76.5	95.6	98.3	**61.7**	89.4	**95.3**	516.8
	RCSL-NC	**79.1**	94.0	**97.4**	**58.5**	**84.1**	**90.1**	**503.2**	**77.9**	**95.9**	**98.5**	**61.7**	**89.6**	95.2	**518.8**
60%	NCR('21)	68.7	89.9	95.5	52.0	77.6	84.9	468.6	72.7	94.0	97.6	57.9	87.0	94.1	503.3
	DECL('22)	69.4	89.4	95.2	52.6	78.8	85.9	471.3	73.8	94.7	97.7	59.6	87.9	94.5	508.2
	MSCN('23)	70.4	91.0	94.9	53.4	77.8	84.1	471.6	74.4	**95.1**	97.9	59.2	87.1	92.8	506.5
	CREAM('24)	70.6	91.2	96.1	53.3	79.2	87.0	477.4	74.7	94.8	98.0	59.7	88.0	**94.6**	509.9
	RCSL-NC	**75.6**	**93.6**	**97.2**	**56.4**	**82.1**	**88.9**	**493.8**	**75.5**	95.0	**98.3**	**59.9**	**88.4**	94.4	**511.5**
80%	NCR('21)	1.4	7.1	11.7	1.5	5.4	9.3	36.4	21.6	52.6	67.6	15.1	38.1	49.8	244.8
	DECL('22)	60.7	84.6	91.2	42.1	69.6	78.6	426.8	65.6	91.6	96.6	52.0	83.0	91.3	480.1
	MSCN('23)	1.0	4.4	9.1	0.4	1.4	2.5	18.8	66.8	91.6	96.2	52.7	83.0	90.9	481.2
	CREAM('24)	56.1	81.2	88.4	39.2	66.7	76.2	407.8	68.6	92.0	96.4	54.3	84.8	92.5	488.7
	RCSL-NC	**70.3**	**89.8**	**94.3**	**50.6**	**77.8**	**85.7**	**468.5**	**72.3**	**94.0**	**97.4**	**56.5**	**86.1**	**93.1**	**499.4**

Table 3: Performance on the Flickr30K and MS-COCO datasets under the noisy setting. The best results are in **bold**.

spondence: NCR (NeurIPS'21) (Huang et al. 2021), DECL (ACM MM'22) (Qin et al. 2022), MSCN (CVPR'23) (Han et al. 2023), and CREAM (TIP'24) (Ma et al. 2024). To be fair, the backbones of all baselines are unified.

Results on the semi-paired data: From the results shown in Table 2, (1) the performance of all evaluated baselines remarkably deteriorates as the number of paired data decreases, indicating a heavy dependency on a large amount of paired data. As demonstrated by the experimental results, when the paired number is decreased from 25K to 2.5K, the overall performance (rSum) of the latest baseline FEM decreases by 230.5 and 191.2 on the Flickr30K 1K test and the MS-COCO 5-fold 1K test, respectively. Although our RCSL also suffers performance degradation due to the reduction of paired data, compared to FEM, the amplitude is smaller (*i.e.*, 230.5 *vs.* 190.9 and 191.2 *vs.* 116.6) and shows promising performance. **(2)** When there is very little pairing data (2.5K), most baselines struggle to converge

due to the insufficient number of paired data, resulting in inadequate performance, *e.g.*, 2AD, CHAN, and *etc*. In contrast, our RCSL method addresses the limitation of learning knowledge solely from paired data by mining potential visual-semantic associations from a large amount of unpaired data. As seen in Table 2, our method achieves the best performance in almost all metrics with a notable advantage, demonstrating the effectiveness and superiority of semi-paired cross-modal learning.

To further verify the necessity of our method, we also conduct experiments with CLIP on a constructed real-world dataset (*i.e.*, Drone-SP) that includes a limited number of carefully annotated instances as well as a large amount of unpaired data from the Internet. From the results shown in Table 4, one can see that the zero-shot performance of CLIP is poor, with only 12.3% in terms of R@1. Although fine-tuning CLIP on the paired dataset significantly improves the performance, our method, even our RCSL without RCM can remarkably outperform fine-tuned CLIP by a large margin. Furthermore, since RCM can robustly extract visual semantic associations from pseudopaired data, the performance of RCSL equipped with RCM is further improved by 13.2% absolutely in terms of rSum, further demonstrating the effectiveness and practicality of our RCSL in real-world scenarios.

Methods	Image⟶Text			Text⟶Image			rSum
	R@1	R@5	R@10	R@1	R@5	R@10	
CLIP⋆	11.2	28.2	39.8	12.6	28.6	38.0	158.4
CLIP	37.0	73.2	85.4	32.6	68.2	79.8	376.2
RCSL w/o RCM	**40.8**	74.4	87.0	34.6	67.0	82.8	386.6
RCSL	**40.8**	**75.8**	**88.4**	**38.2**	**71.4**	**85.2**	**399.8**

Table 4: Performance (R@K %) on the Drone-SP dataset. '⋆' means the zero-shot results.

Methods	Image → Text			Text → Image			rSum
	R@1	R@5	R@10	R@1	R@5	R@10	
NCR	39.5	64.5	73.5	40.3	64.6	73.2	355.6
DECL	39.0	66.1	75.5	40.7	66.3	76.7	364.3
MSCN	40.1	65.7	76.6	40.6	67.4	76.3	366.7
CREAM	40.3	68.5	77.1	40.2	68.2	78.3	372.6
RCSL-NC	**45.7**	**70.4**	**77.9**	**43.5**	**70.8**	**79.9**	**388.2**

Table 5: Performance (R@K %) on the CC152K dataset with real noisy correspondence. The best results are in **bold**.

Results on the noisy data: We report the quantitative results on the synthetic noisy data in Table 3 and the results on the real noisy data in Table 5. Note that we do not report the ensemble performance of multiple models as in (Huang et al. 2021; Qin et al. 2022; Yang et al. 2023). From the results in Table 3, one can see that our RCSL-NC shows excellent performance on both datasets and almost all exceed the compared baselines. Besides, our RCSL-NC is very stable under high noise, *i.e.*, 60% and 80%. Specifically, our method exceeds the best baseline in terms of rSum by 41.7% and 9.7% on the two datasets with 80% noise. Besides, our method also performs all baselines excellently under the real noisy dataset CC152K. The quantitative results are shown in Table 5. From the results, our RCSL-NC prevails over CREAM by 15.6% in terms of rSum absolutely.

No.	Methods	Image → Text			Text → Image			rSum
		R@1	R@5	R@10	R@1	R@5	R@10	
#1	**RCSL (FULL)**	63.6	88.1	92.9	46.2	73.9	82.7	447.4
#2	- w/o $\mathcal{L}_{\text{reg}}$	62.8	85.6	92.2	45.9	73.4	82.2	442.2
#3	- w/o $\mathcal{L}_{\text{sdl}}^{\dagger}$	48.8	72.8	82.5	34.1	62.5	73.7	374.4
#4	- w/o $\mathcal{L}_{\text{rcm}}^{t}$	62.5	85.7	92.3	44.5	72.4	81.9	439.3
#5	- w/o $\mathcal{L}_{\text{rcm}}^{v}$	62.5	85.1	91.9	44.5	72.9	81.4	438.3
#6	- w/o $\mathcal{L}_{\text{rcm}}$	60.0	84.6	90.9	42.3	71.6	80.8	430.2
#7	- w/o $\mathcal{L}_{\text{reg}}$ & $\mathcal{L}_{\text{rcm}}$	58.4	83.7	90.5	41.8	71.3	80.5	426.2
#8	- w/o $\mathcal{L}_{\text{sdl}}^{\dagger}$ & $\mathcal{L}_{\text{rcm}}$	37.0	64.6	74.3	20.7	47.6	60.3	304.5

Table 6: Ablation studies on the Flickr30K dataset with 5K paired data. Default settings are marked in gray.

Ablation Study

In this section, we perform ablation experiments to investigate the individual contributions of each component to our RCSL. All experiments are conducted under the semi-paired settings on the Flickr30K dataset with 25K paired data. The retrieval results of various RCSL variants with different configurations are presented in Table 6. From the results, our full version of RCSL is the best, which shows that all components are important to obtaining promising performance. #2 shows that maintaining the good properties of representation learning can bring performance gains. #4-6 are the experiments to explore the effect of the proposed RCM, wherein #4 means only learning with the mined texts during RCM, and #5 means only learning with the mined images. #4-6 verify the effectiveness of learning with the mined cross-correlations through RCM. Compared with #1 and #6, the full version of RCSL (*i.e.*, #1) is greatly improved (over 10% on rSum) through robust cross-association mining (RCM), which shows that it is feasible to mine visual-semantic information from unpaired data in an end-to-end manner. #7-8 shows that each proposed item is beneficial for the performance of cross-modal retrieval, and our standard version #1 is the balanced and optimal result of each item. In conclusion, the experimental results demonstrate the effectiveness of each component and show that it is helpful for semi-paired cross-modal learning.

Conclusion

In this paper, we reveal and study a new learning paradigm for image-text retrieval, called semi-paired cross-modal learning. Unlike the existing paradigms, this paradigm attempts to mitigate the requirement of fully-paired data by endowing the model with exploiting unpaired data. To this end, we propose a Robust Cross-modal Semi-paired learning method (RCSL) to extract visual-semantic associations from paired as well as unpaired data. Besides, we proved that SPL is also a feasible solution to solve the noisy correspondence problem, which will provide a new perspective for future image-text retrieval. We perform extensive experiments on four datasets to verify the effectiveness and superiority of the proposed method in the semi-paired and noisy setting.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants U25A201523, U25B6003, 62372315, 62472295, 62176171 and U21B2040; in part by Sichuan Science and Technology Planning Project under Grants 2024ZDZX0004 and 2024NSFTD0038; in part by Central Government's Guide to Local Science and Technology Development Fund under Grant 2025ZYDF101; in part by the Fundamental Research Funds for the Central Universities under Grants CJ202303 and CJ202403; in part by System of Systems and Artificial Intelligence Laboratory pioneer fund grant under Grant HLJGGG20240327517-15; This research is also supported by the Japan Science and Technology Agency (JST) and the Agency for Science, Technology and Research (A*STAR) under the Japan-Singapore Joint Call (Project No. R24I6IR133).

References

Arpit, D.; Jastrzębski, S.; Ballas, N.; Krueger, D.; Bengio, E.; Kanwal, M. S.; Maharaj, T.; Fischer, A.; Courville, A.; Bengio, Y.; and Lacoste-Julien, S. 2017. A Closer Look at Memorization in Deep Networks. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 233–242. PMLR.

Chen, J.; Hu, H.; Wu, H.; Jiang, Y.; and Wang, C. 2021. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15789–15798.

Diao, H.; Zhang, Y.; Liu, W.; Ruan, X.; and Lu, H. 2023. Plug-and-Play Regulators for Image-Text Matching. *IEEE Transactions on Image Processing*.

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.

Fartash, F.; Fleet, D.; Kiros, J.; and Fidler, S. 2018. VSE++: Improved visual semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference (BMCV)*.

Feng, Y.; Li, Y.; Sun, Y.; Qin, Y.; Peng, D.; and Hu, P. 2025a. Interactive Cross-modal Learning for Text-3D Scene Retrieval. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Feng, Y.; Qin, Y.; Peng, D.; Zhu, H.; Peng, X.; and Hu, P. 2025b. PointCloud-Text Matching: Benchmark Dataset and Baseline. *IEEE Transactions on Multimedia*, 27: 6986–6995.

Fu, Z.; Mao, Z.; Song, Y.; and Zhang, Y. 2023. Learning Semantic Relationship Among Instances for Image-Text Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15159–15168.

Fu, Z.; Zhang, L.; Xia, H.; and Mao, Z. 2024. Linguistic-Aware Patch Slimming Framework for Fine-grained Cross-Modal Alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 26307–26316.

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.

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.

Hu, P.; Qin, Y.; Gou, Y.; Li, Y.; Yang, M.; and Peng, X. 2025. Probabilistic multimodal learning with von Mises-Fisher distributions. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 5390–5398.

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.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 4904–4916. PMLR.

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 (ECCV)*, 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.

Li, L. H.; You, H.; Wang, Z.; Zareian, A.; Chang, S.-F.; and Chang, K.-W. 2021. Unsupervised Vision-and-Language Pre-training Without Parallel Images and Captions. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5339–5350.

Li, W.; and Pun, C.-M. 2023. Asymmetric Scalable Cross-Modal Hashing. In *2023 IEEE International Conference on Image Processing (ICIP)*, 316–320. IEEE.

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 *European conference on computer vision*, 740–755. Springer.

Ma, X.; Yang, M.; Li, Y.; Hu, P.; Lv, J.; and Peng, X. 2024. Cross-modal Retrieval with Noisy Correspondence via Consistency Refining and Mining. *IEEE transactions on image processing*.

Manwani, N.; and Sastry, P. 2013. Noise tolerance under risk minimization. *IEEE transactions on cybernetics*, 43(3): 1146–1151.

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.

Pu, R.; Qin, Y.; Song, X.; Peng, D.; Ren, Z.; and Sun, Y. 2025. SHE: Streaming-media Hashing Retrieval. In *Forty-second International Conference on Machine Learning*.

Qin, Y.; Feng, G.; and Zhang, X. 2025. Scalable One-Pass Incomplete Multi-View Clustering by Aligning Anchors. In *Advancement of Artificial Intelligence*, 20042–20050.

Qin, Y.; Peng, D.; Peng, X.; Wang, X.; and Hu, P. 2022. Deep Evidential Learning with Noisy Correspondence for Cross-modal Retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, 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. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Qin, Y.; Zhang, X.; Yu, S.; and Feng, G. 2025. A survey on representation learning for multi-view data. *Neural Networks*, 181: 106842.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2556–2565.

Shen, X.; Yu, G.; Chen, Y.; Yang, X.; and Zheng, Y. 2024. Graph Convolutional Semi-Supervised Cross-Modal Hashing. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5930–5938.

Wang, J.; Gong, T.; and Yan, Y. 2024. Semi-supervised Prototype Semantic Association Learning for Robust Cross-modal Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 872–881.

Wang, T.; and Isola, P. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, 9929–9939. PMLR.

Wang, Z.; Yin, Y.; and Ramakrishnan, I. 2024. Enhancing Image-Text Matching with Adaptive Feature Aggregation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8245–8249. IEEE.

Yang, A.; Feng, Y.; Sun, Y.; Peng, D.; Duan, G.; and Qin, Y. 2025. Noise-Robust Cross-modal Learning for Reliable 2D-3D Retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 8154–8163.

Yang, S.; Xu, Z.; Wang, K.; You, Y.; Yao, H.; Liu, T.; and Xu, M. 2023. BiCro: Noisy Correspondence Rectification for Multi-modality Data via Bi-directional Cross-modal Similarity Consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19883–19892.

Yang, Y.; Wang, L.; Yang, E.; and Deng, C. 2024. Robust Noisy Correspondence Learning with Equivariant Similarity Consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17700–17709.

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

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

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 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 852–861.

Zha, Q.; Liu, X.; Peng, S.-J.; Cheung, Y.-m.; Xu, X.; and Wang, N. 2025b. ReCon: Enhancing True Correspondence Discrimination through Relation Consistency for Robust Noisy Correspondence Learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 29680–29689.

Zhang, K.; Mao, Z.; Wang, Q.; and Zhang, Y. 2022. Negative-Aware Attention Framework for Image-Text Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15661–15670.

Zhang, K.; Zhang, L.; Hu, B.; Zhu, M.; and Mao, Z. 2023a. Unlocking the Power of Cross-Dimensional Semantic Dependency for Image-Text Matching. In *Proceedings of the 31st ACM International Conference on Multimedia*, 4828–4837.

Zhang, Z.; Shu, C.; Xiao, Y.; Shen, Y.; Zhu, D.; Chen, Y.; Xiao, J.; Lau, J. H.; Zhang, Q.; and Lu, Z. 2023b. Improving Visual-Semantic Embedding with Adaptive Pooling and Optimization Objective. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 1217–1229.

Zhou, M.; Yu, L.; Singh, A.; Wang, M.; Yu, Z.; and Zhang, N. 2022. Unsupervised vision-and-language pre-training via retrieval-based multi-granular alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16485–16494.

Zhu, H.; Zhang, C.; Wei, Y.; Huang, S.; and Zhao, Y. 2023. ESA: External Space Attention Aggregation for Image-Text Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zou, H.; Zhang, R.; Zhou, X.; and Zou, J. 2025. GEA: Generation-Enhanced Alignment for Text-to-Image Person Retrieval. *arXiv preprint arXiv:2511.10154*.