Seeking Proxy Point via Stable Feature Space for Noisy Correspondence Learning

Yucheng Xie[1] , **Songyue Cai**[1] , **Tao Tong**[1] , **Ping Hu**[1*] , **Xiaofeng Zhu**[1]

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China

xyemrsnon@gmail.com, sonnycai@std.uestc.edu.cn, {tongqtao, chinahuping, seanzhuxf}@gmail.com

Abstract

To meet the growing demand for cross-modal training data, directly collecting multimodal data from the Internet has become prevalent. However, such data inevitably suffer from Noisy Correspondence. Previous works focused on recasting soft labels to mitigate noise's negative impact. We explore a novel perspective to solve this problem: pursuing proxy representation for noisy data to enable reliable feature learning. To this end, we propose a novel framework: Seeking Proxy Point via Stable Feature Space (SPS). This framework employs a fine-grained partitioning strategy to obtain a high-confidence reliable set. By imposing intermodal cross-transformation consistency constraints and intramodal metric consistency constraints, a stable feature space is constructed. Building on this foundation, SPS seeks proxy points for noisy data, enabling even noisy data to be accurately embedded into appropriate positions within the feature space. Combined with partial alignment for partially matched data pairs, SPS ultimately achieves robust learning under Noisy Correspondence. Experiments on three widely used cross-modal datasets demonstrate that SPS significantly outperforms previous methods. Our code is available at https://github.com/C-TeaRanger/SPS.

1 Introduction

Cross-modal retrieval, a key task in mutilmodal learning, is gaining significant attention as a crucial cornerstone to achieve Artificial General Intelligence[Morris *et al.*, 2024], which focuses on aligning and comparing data across modalities (e.g., images, text, audio) to enable efficient and accurate retrieval of diverse information. Most existing methods[Yang *et al.*, 2024][Chen *et al.*, 2020b][Hu *et al.*, 2021] rely on contrastive learning, aiming to reduce the feature space distance between matching pairs (positive examples) to bridge semantic gaps. However, these methods often assume perfect alignment in training data pairs, which is unrealistic. Unlike unimodal tasks, annotating cross-modal data is costly, especially for large-scale, high-quality pairs, making it labor intensive and sometimes unfeasible. To address this, a more economical and efficient method has been proposed[Sharma *et al.*, 2018][Jia *et al.*, 2021]: directly crawling native cross-modal data pairs from the Internet. Yet, this method inevitably introduces N Noisy Correspondence[Huang *et al.*, 2021b][Qin *et al.*, 2022], that is, some data pairs that are inherently mismatched, significantly compromising the retrieval performance of the model.

In recent years, the issue of Noisy Correspondence has garnered significant attention and has been extensively researched. Initially, NCR[Huang *et al.*, 2021b] reformulates the binary hard labels of the original data into soft labels within the [0, 1] interval by estimating the relevance degree of each data pair and employs a soft-margin triplet loss constructed based on these soft labels. Subsequently, MSCN[Han *et al.*, 2023] adopted meta-learning[Li *et al.*, 2019a] to perform consistency correction. UGNCL[Zha *et al.*, 2024] partitioned the data under the guidance of uncertainty. L2RM[Han *et al.*, 2024], based on the Optimal Transport[Cuturi, 2013] theory, selects reliable samples from negative examples to compensate for the reduced training information caused by noise. However, noise data reweighted by soft labels only mitigate noise's negative impact without providing reliable supervision signals. Additionally, whether using the triplet loss[Faghri *et al.*, 2018] or cross-modal InfoNCE[Hu *et al.*, 2023] contrastive loss, the representation space lacks geometric stability, making it hard to ensure prediction consistency between image and text domains under Noisy Correspondence. This instability significantly degrades the retrieval performance.

To address the aforementioned issues, we propose a novel Noise Correspondence robust learning framework, Seeking Proxy point via Stable feature space (SPS). This framework leverages the Memory Effect[Arpit *et al.*, 2017] of deep neural networks (DNN) to extract a high-confidence reliable set from the original dataset and further divides the remaining data into a quasi-clean set and a noisy set using posterior probabilities, enabling subsequent fine-grained learning strategies. Specifically, we begin by adequately mining the supervisory signals latent within the reliable set. By imposing intermodal cross-transformation consistency constraints and intramodal metric consistency constraints, we construct a stable joint feature space, effectively alleviating the inconsis-

*Corresponding author

tency in predictions across modalities caused by Noise Correspondence. Subsequently, based on this stable representation space, we seek proxy points for the noisy pairs, enabling them to be accurately embedded into appropriate positions within the feature space, even if they are noisy. This provides additional reliable supervisory signals and eliminates the impact of Noise Correspondence. Furthermore, the quasi-clean pairs predominantly represent the partial matching problem. We address this issue by refining the internal consistency of quasi-clean pairs to achieve partial alignment. Finally, we integrate these constraints and employ a cross-modal bidirectional contrastive loss and co-teaching paradigm to achieve robust learning under Noisy Correspondence conditions. The main contributions and innovations of this paper can be summarized as follows:

- We propose a novel noisy robust learning framework, SPS, which effectively resolves the Noisy Correspondence problem through fine-grained data partitioning and stable feature space constraint learning.
- Our framework innovatively introduces a noise proxy representation learning method based on a stable feature space, eliminating the impact of Noisy Correspondence while providing reliable supervision signals for model training.
- Extensive experiments across multiple cross-modal retrieval tasks validate the effectiveness of SPS, demonstrating that it significantly outperforms existing methods, particularly under high noise rates.

2 Related Work

2.1 Cross-modal Retrieval

Cross-modal retrieval[Cheng *et al.*, 2022], essential in multimodal learning[Huang *et al.*, 2021a][Zolfaghari *et al.*, 2021] and information retrieval[Hambarde and Proenca, 2023], enables mutual retrieval across modalities like images and text by bridging the modality gap and aligning features semantically. A common approach uses contrastive learning to create a shared embedding space, bringing similar cross-modal samples closer. Existing methods to improve retrieval include SCAN[Lee *et al.*, 2018] (stacked cross-attention for image-text similarity), VSRN[Li *et al.*, 2019b] (GCNs for semantic reasoning), IMRAM[Chen *et al.*, 2020a] (iterative matching with attention), and SGRAF[Diao *et al.*, 2021] (similarity graph reasoning). However, these methods assume perfectly aligned training data and overlook Noisy Correspondence—mismatched image-text pairs common in real-world data due to collection and annotation challenges. When noise is present, model performance drops as mismatched samples are incorrectly aligned, corrupting the feature space.

2.2 Noisy Correspondence Learning

Noisy Correspondence Learning (NCL) is a novel paradigm addressing semantically mismatched or partially matched cross-modal data pairs, unlike traditional noisy label[Liu and Tao, 2016][Xia *et al.*, 2020] learning focused on single-modality data (e.g., image classification). Introduced by[Huang *et al.*, 2021b], NCL leverages DNNs' memorization effect to partition datasets and correct labels adaptively. Subsequent improvements include: DECL[Qin *et al.*, 2022] (combining Cross-modal Evidence Learning and Robust Dynamic Hinge Loss), BiCro[Yang *et al.*, 2023] (using bidirectional similarity consistency for soft labels), CREAM[Ma *et al.*, 2024] (adapting InfoNCE loss to uncover consistency within mismatched pairs), L2RM[Han *et al.*, 2024] (using Optimal Transport to filter reliable samples), and PC2[Duan *et al.*, 2024] (pseudo-caption methods with oscillation for relevance correction). However, limitations persist: insufficient feature space stability due to noise-induced drift, and limited utilization of noisy data, as methods only re-weight rather than extract reliable supervision. Our proposed SPS framework effectively tackles these challenges.

3 Methodology

3.1 Problem Formulation

In cross-modal retrieval, we consider a dataset $\mathcal{D} = \{(I_i, T_i, y_i)\}_{i=1}^{N}$ of N samples, where each triplet (I_i, T_i, y_i) consists of an image-text pair (I_i, T_i) and a binary label y_i. The label y_i indicates whether I_i and T_i are positively correlated ($y_i = 1$) or uncorrelated ($y_i = 0$). Uncorrelated examples can introduce erroneous supervisory signals, misleading the model during training and degrading its overall performance. Thus, our objective is to learn a robust feature embedding that faithfully captures the true relevance between images and text, even in the presence of mislabeled or partially incorrect pairs.

3.2 Vanilla Cross-modal Loss

The core of cross-modal retrieval lies in accurately measuring the consistency between different modalities. We employ modality-specific encoders $f(\cdot)$ and $g(\cdot)$ to map an image I and its corresponding text T into a shared feature space, with $f(I)$ and $g(T)$ representing their respective feature embeddings. The semantic similarity between image I and text T is measured as $\mathcal{S}(f(I), g(T))$. To learn the encoders $f(\cdot)$ and $g(\cdot)$, we adopt a variant of the InfoNCE [Radford *et al.*, 2021] loss tailored for cross-modal scenario. This contrastive loss, derived from mutual information, aims to bring positive pairs closer while simultaneously pushing negative pairs as far apart as possible:

$$\mathcal{L}(I_i, T_i) = \mathcal{H}(y_i, \mathcal{P}_i^{i2t}) + \mathcal{H}(y_i, \mathcal{P}_i^{t2i}), \tag{1}$$

where $\mathcal{H}$ denotes the unidirectional cross-entropy function. Since cross-modal retrieval involves bidirectional queries, $\mathcal{L}$ consists of two symmetric terms $\mathcal{H}$. $\mathcal{P}_i^{i2t}$ represents the matching probability of query I_i with respect to T_i in the image-text pair (I_i, T_i), $\mathcal{P}_i^{t2i}$ represents retrieval in the opposite direction, expressed as:

$$\begin{aligned} \mathcal{P}_i^{i2t} &= \frac{\exp(\mathcal{S}(I_i, T_i)/\tau)}{\sum_{j=1}^{N} \exp(\mathcal{S}(I_i, T_j)/\tau)} \\ \mathcal{P}_i^{t2i} &= \frac{\exp(\mathcal{S}(I_i, T_i)/\tau)}{\sum_{j=1}^{N} \exp(\mathcal{S}(I_j, T_i)/\tau)}, \end{aligned} \tag{2}$$

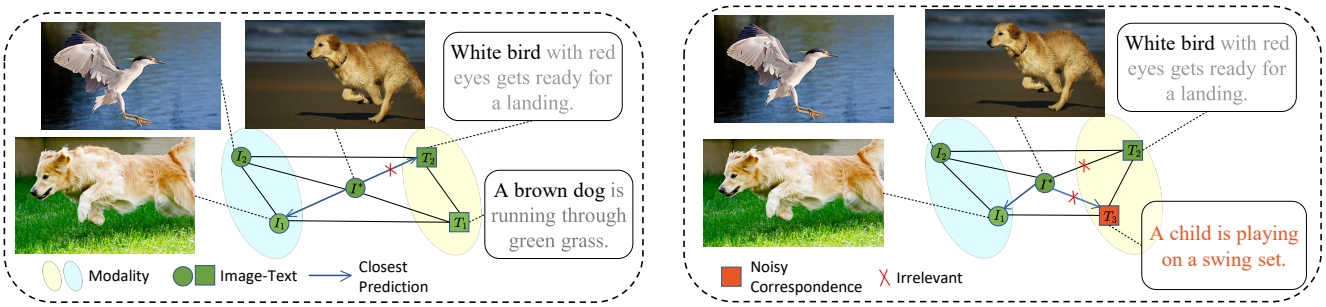

Figure 1: Illustrations of prediction inconsistency and its exacerbation by Noisy Correspondence. (a) Traditional contrastive loss functions create feature spaces with prediction inconsistency due to structural instability. For example, I^* aligns with I_1 (a brown dog) in the image domain but with T_2 ("white bird...") in the text domain, degrading retrieval performance. (b) Noisy Correspondence erroneously link irrelevant image-text pairs, further destabilizing the structure. This exacerbates the prediction errors for I^*, making both T_2 and T_3 incorrect.

where τ is the temperature parameter, which is fixed at 0.07 in this experiment. When Noisy Correspondence occurs, there exist triplets $(I_a, T_b, y^* = 1) \in \mathcal{D}$, where I_a and T_b are not actually matched, but the corresponding label y^* is incorrectly annotated as 1, which leads the model to erroneously align originally irrelevant cross-modal features.

3.3 Fine-grained Data Partitioning

Directly training the model on the raw dataset mentioned in 3.1 without preprocess will lead to overfitting on noisy samples, significantly degrading cross-modal retrieval performance. The memorization effect of DNNs, indicates that DNNs tend to prioritize memorizing clean training data before memorizing noisy training data. Leveraging this characteristic, we can achieve dataset partitioning by analyzing the loss distribution differences among sample pairs.

Given the raw dataset $\mathcal{D}$, we compute the loss value $\mathcal{L}_i$ for each sample using Equation (1):

$$\mathcal{L}_{\mathcal{D}} = \{\mathcal{L}_i\}_{i=1}^{N} = \{\mathcal{L}(I_i, T_i)\}_{i=1}^{N}. \tag{3}$$

Then we employ a two-component Gaussian Mixture Model (GMM)[Li *et al.*, 2020][Permuter *et al.*, 2006] to fit the probability distribution of the loss values L_i:

$$p(\mathcal{L}_i) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathcal{L}_i | \mu_k, \sigma_k), \tag{4}$$

where $K = 2$, and π_k and $\mathcal{N}(\mathcal{L}_i|\mu_k, \sigma_k)$ represent the mixing coefficient and probability density function of the k-th component with parameters $\mu_k > 0, \sigma_k > 0$. We employ the Expectation-Maximization (EM) algorithm to ensure the convergence of the GMM. Subsequently, we calculate the probability p_i that $\mathcal{L}_i$ belongs to the component k' with a smaller mean loss (i.e., the probability that (I_i, T_i) is a positive pair):

$$p_i = p(k'|\mathcal{L}_i) = \frac{p(k')p(\mathcal{L}_i|k')}{p(\mathcal{L}_i)}. \tag{5}$$

Given that the construction of a stable feature space highly depends on reliable data, we design a fine-grained data partitioning strategy. Specifically, we use the posterior probability p_i to approximate the matching degree of iamge-caption pairs. By setting thresholds ϵ_1 and ϵ_2, the original training set $\mathcal{D}$ is meticulously divided into three subsets: the reliable set $\mathcal{D}_{re}$, the quasi-clean set $\mathcal{D}_{qc}$ and the noisy set $\mathcal{D}_n$:

$$\mathcal{D}_{re} = \{(I_i, T_i, y_i = 1) | p(k'|\mathcal{L}_i) > \epsilon_1, \forall (I_i, T_i) \in \mathcal{D}\}. \tag{6}$$

For the quasi-clean set D_{qc}, we treat it as partially matched samples, remove the original labels, and use the adjusted consistency coefficients $\hat{y_{qc}}$ as the new labels:

$$\mathcal{D}_{qc} = \{(I_i, T_i, \hat{y_{qc}}) | \epsilon_2 < p(k'|\mathcal{L}_i) \leq \epsilon_1, \forall (I_i, T_i) \in \mathcal{D}\}. \tag{7}$$

The noisy set D_n is identified as mismatched sample pairs, and their original labels are discarded entirely.

$$\mathcal{D}_n = \{(I_i, T_i) | p(k'|\mathcal{L}_i) \leq \epsilon_2, \forall (I_i, T_i) \in \mathcal{D}\}. \tag{8}$$

To address the issue of bias accumulation in single-model training, which is common in noisy label learning, we improve the co-teaching [Yu *et al.*, 2019][Han *et al.*, 2018] framework by training two networks $\mathcal{M}_a = \{f_a, g_a, \mathcal{S}_a\}$ and $\mathcal{M}_b = \{f_b, g_b, \mathcal{S}_b\}$ with identical architectures simultaneously and adjusting the data interaction mechanism: at the batch level, only the data partitioning strategy is shared, rather than the raw data. This approach enhances training efficiency while adaptively correcting training errors and avoiding bias accumulation in a single model.

3.4 Seeking Proxy Point via Stable Feature Space

Traditional contrastive loss functions, such as InfoNCE and Triple Loss, primarily construct the feature representation space from two perspectives: alignment and uniformity[Wang and Isola, 2022][Pu *et al.*, 2022]. These loss functions aim to maximize the alignment of positive sample pairs while simultaneously pursuing a uniform distribution of features on the hypersphere, with the goal of preserving the maximum amount of information. However, in the cross-modal retrieval tasks we study, due to the presence of noise and the inherent modality gap[Liang *et al.*, 2022], simply pursuing alignment and uniformity struggles to effectively address this challenge[Jiang *et al.*, 2023]. In such cases, the constructed feature space often suffers from prediction inconsistency[Goel *et al.*, 2022] issues owing to the lack

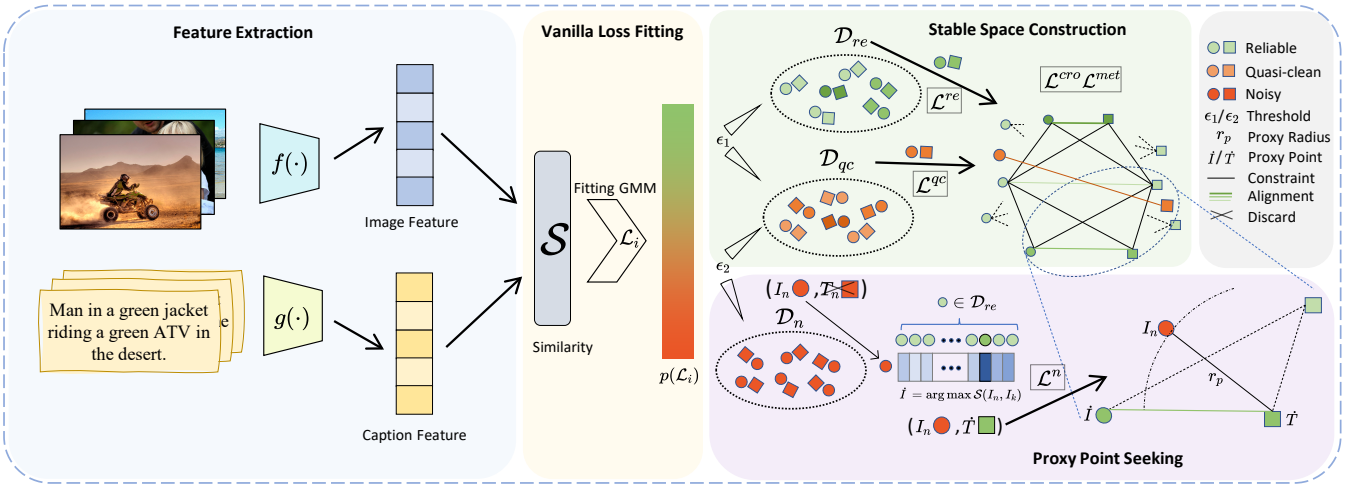

Figure 2: Overview of the proposed SPS. The given dataset is processed through modality-specific encoders to extract features, that are then used to fit GMM based on the memorization effect of DNNs. By applying multiple thresholds and employing a fine-grained partitioning strategy, the dataset is divided into reliable set, quasi-clean set, and noisy set. Next, by imposing intermodal cross-transformation consistency constraints and intramodal metric consistency constraints on the reliable set, a stable joint feature space is constructed. Building on this foundation, proxy points are sought for the noisy set, leveraging proxy features to accurately embed noisy data into appropriate positions. Finally, by combining partial alignment for the quasi-clean set, robust cross-modal retrieval is achieved.

of structural stability. When Noisy Correspondence are introduced, the model erroneously align irrelevant features, further destabilizing the model structure and exacerbating this problem ,as illustrated in Figure 1. Inspired by this, we propose a novel framework that addresses the NC problem by constructing a stable cross-modal feature space, enabling proxy point learning for the noisy set and partial alignment for the quasi-clean set.

Stable Feature Space Constraint

For data pairs identified as clean samples with high confidence, we should fully exploit the supervisory information they contain. Additionally, in the cross-modal joint embedding space, the feature representations of different modalities should maintain metric consistency, without significant discrepancies. Based on these two intuitive insights, we introduce an intermodal cross-transformation consistency constraint $\mathcal{L}^{cro}$ and an intramodal metric consistency constraint $\mathcal{L}^{met}$ on the reliable set D_{re}, aiming to construct a stable joint feature space.

The constraint $\mathcal{L}^{cro}$ ensures that the cross-transformation between clean sample pairs remains consistent, which contributes to achieving a stable geometric structure. It is formulated as:

$$\mathcal{L}^{cro} = \frac{1}{|\mathcal{B}_{re}|^2} \sum_i \sum_j [\|\mathcal{S}(I_i, T_j) - \mathcal{S}(I_j, T_i)\|_2^2 - \alpha]_+. \quad (9)$$

The constraint $\mathcal{L}^{met}$ requires that the metric differences within each modality should be consistent. It is formulated as:

$$\mathcal{L}^{met} = \frac{1}{|\mathcal{B}_{re}|^2} \sum_i \sum_j [\|\mathcal{S}(I_i, I_j) - \mathcal{S}(T_j, T_i)\|_2^2 - \alpha]_+, \quad (10)$$

where $S(I_i, I_j)$ denotes $S(f(I_i), f(I_j))$, and $S(T_i, T_j)$ denotes $S(g(T_i), g(T_j))$, B_{re} represents a batch sampled from $\mathcal{D}_{re}$, $[x]_+ = \max(x, 0)$ and $\|\cdot\|_2$ denotes the L_2 norm. Following[Wang *et al.*, 2023], we utilize a margin parameter α to control the strength of regularization.

Proxy Representation Learning

Unlike previous approaches that widely adopted re-weighting strategies for noisy data, as discussed in the 1, such methods merely mitigate the negative impact of noisy data on the model, essentially providing low-quality and unreliable supervisory signals. Instead, we propose to search for proxy points for noisy data based on the stable feature space constructed in 3.4, thereby providing reliable supervisory signals.

Without loss of generality, we use the image as a query to find a proxy caption, while a similar method is applied when using the caption as a query to find a proxy image. Specifically, given a noisy data pair $(I_n, T_n) \in \mathcal{D}_n$, we first discard the caption T_n. Then, using the image I_n as a query, we search for the image $\dot{I}$ with the highest similarity to I_n within the current batch B_{re}, along with the corresponding caption $\dot{T}$ of $\dot{I}$. This process can be formulated as:

$$\dot{I} = \arg \max_{I_k \in B_{re}} \mathcal{S}(I_n, I_k), \quad (11)$$

yielding a clean triplet $(\dot{I}, \dot{T}, y = 1)$. Subsequently, we derive the proxy label y_p based on the similarity $\mathcal{S}(I_n, \dot{I})$:

$$y_p = \frac{1}{\gamma + \exp(-\beta \cdot \mathcal{S}(I_n, \dot{I}))}, \quad (12)$$

where γ and β are hyperparameters. For demonstration purposes, we can define the proxy radius r_p as $r_p = 1 - y_p$, as illustrated in the figure 2. Next, we treat $\dot{T}$ as the proxy caption for the image I, constructing a new triplet $(I, \dot{T}, \hat{y} = y_p)$.

These triplets, formed by noisy samples and proxy points, constitute a new dataset $\hat{D}_n$, which is used for more robust training. The loss function for the noisy set is defined as:

$$\mathcal{L}^n = \frac{1}{|\mathcal{B}_n|} \sum_i \hat{y_n} \mathcal{L}(I_i, \dot{T}_i), \tag{13}$$

where $\mathcal{B}_n$ is a batch sampled from $\hat{D}_n$. Consequently, even noisy data initially provided as mismatched pairs can be accurately embedded into appropriate positions within the stable feature space, determined jointly by the proxy point and the proxy label y_p.

Partial Alignment

The quasi-clean sample pairs primarily reflect partial matching issues. Based on the previously constructed stable feature spave, we only need to adjust the consistency coefficients according to the internal similarity of the quasi-clean sample pairs to achieve partial alignment. Given a quasi-clean sample pair $(I_i, T_i, \hat{y}_{qc}) \in \mathcal{D}_{qc}$, like[Ma *et al.*, 2024], we adjust $\hat{y_{qc}}$ using the posterior probability p_i and the binary hard label y_i collectively:

$$\hat{y_{qc}} = p_i y_i + (1 - p_i)\hat{y_i}, \tag{14}$$

where $\hat{y}_i$ represents the predicted matching degree of the sample pair (I_i, T_i) by the model, obtained by averaging the bidirectional cross-modal matching probabilities:

$$\hat{y_i} = \frac{1}{2}[\mathcal{P}_i^{i2t} + \mathcal{P}_i^{t2i}]. \tag{15}$$

Finally, we compute the loss for the quasi-clean set using the following formula:

$$\mathcal{L}^{qc} = \frac{1}{|\mathcal{B}_{qc}|} \sum_i \hat{y_{qc}} \mathcal{L}(I_i, T_i), \tag{16}$$

where $\mathcal{B}_{qc}$ denotes a batch sampled from $\mathcal{D}_{qc}$.

3.5 Global Training Objective

Before implementing co-teaching, we need to perform Warmup training for models $\mathcal{M}_a$ and $\mathcal{M}_b$ to achieve preliminary parameter convergence. Details are provided in the Appendix. For the fine-grained partitioned reliable set $\mathcal{D}_{re}$, we minimize the following loss function using batch data $\mathcal{B}_{re} \subset \mathcal{D}_{re}$:

$$\mathcal{L}^{re} = \frac{1}{|\mathcal{B}_{re}|} \sum_i y_i \mathcal{L}(I_i, T_i). \tag{17}$$

In summary, the overall loss function of the SPS method can be expressed as:

$$\mathcal{L}^{Overall} = \mathcal{L}^{re} + \mathcal{L}^{qc} + \mathcal{L}^{n} + \lambda_1 \mathcal{L}^{cro} + \lambda_2 \mathcal{L}^{met}, \tag{18}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are hyperparameters used to balance the weights of the two spatial constraints, thereby achieving optimal performance. The entire framework of SPS is illustrated in the figure 2.

4 Experiments

4.1 Datasets and Evaluation Metrics

We employ three widely used cross-modal datasets to validate our proposed framework.

- **Flickr30K** [Young *et al.*, 2014] consists of 31,000 images sourced from the Flickr platform, with each image annotated by five distinct captions, yielding a total of 155,000 image-text pairs for training purposes. Following [Lee *et al.*, 2018], we allocate 1,000 images for validation, 1,000 images for testing, and utilize the remaining images for training.
- **MS-COCO** [Lin *et al.*, 2014] contains 123,287 images, each paired with five captions, resulting in a total of 616,434 image-text pairs for training. We divide the dataset into 566,435 pairs for training, 25,000 for validation, and 25,000 for testing.
- **Conceptual Captions** [Sharma *et al.*, 2018] is a large-scale dataset comprising 3.3 million image-text pairs, which inherently exhibit real-world Noisy Correspondence challenges. Each image is paired with a single caption. Following [Huang *et al.*, 2021c], we utilize a subset of the dataset, CC152K, for our experiments. Within CC152K, 150,000 images are designated for training, while 1,000 images each are reserved for validation and testing.

Given that Flickr30K and MS-COCO are meticulously annotated, we simulate Noisy Correspondence by randomly shuffling the captions of training images at predefined noise ratios. In contrast, Conceptual Captions, being automatically curated from the Internet, inherently contain approximately 3% to 20% of native mismatched noise, thus requiring no additional artificial shuffling. Following [Huang *et al.*, 2021b], the retrieval performance is assessed using the recall at K (R@K) metric, which quantifies the percentage of relevant items correctly identified within the top K retrieved results. Our experimental evaluation includes R@1, R@5, R@10, as well as the cumulative recall scores RSum for bidirectional matching tasks.

4.2 Implementation Details

The proposed SPS is a universal Noisy Correspondence robust framework that can be directly applied to most existing cross-modal retrieval models. We adopt the widely used cross-modal retrieval model SGR[Diao *et al.*, 2021] as the backbone network, integrating SPS to enhance its robustness to noise. In all our experiments, we use the Adam [Kingma, 2014] optimizer with default parameters for updating the model parameters. To better control the learning progress, we initially train the model exclusively on the reliable set and then gradually incorporate the quasi-clean set and the noisy set into the training process. We select the best checkpoint on the validation set to evaluate performance on the test set. All experiments were conducted on Linux using NVIDIA A100 GPUs.

4.3 Comparisons with State-of-The-Art

In this section, we validate the effectiveness of our method on both artificially noisy and natively noisy datasets. For

Noise	Methods	Flickr30K							MS-COCO 1K						
		Image → Text			Text → Image				Image → Text			Text → Image			
		R@1	R@5	R@10	R@1	R@5	R@10	RSum	R@1	R@5	R@10	R@1	R@5	R@10	RSum
20%	SGRAF	72.8	90.8	95.4	56.4	82.1	88.6	486.1	75.4	95.2	97.9	60.1	88.5	94.8	511.9
	NCR	73.5	93.2	96.6	56.9	82.4	88.5	491.1	76.6	95.6	98.2	60.8	88.8	95.0	515.0
	DECL	77.5	93.8	97.0	56.1	81.8	88.5	494.7	77.5	95.9	98.4	61.7	89.3	95.4	518.2
	BiCro	78.1	94.4	97.5	60.4	**84.4**	**89.9**	504.7	78.8	96.1	**98.6**	63.7	90.3	95.7	523.2
	L2RM	77.9	**95.2**	97.8	59.8	83.6	89.5	503.8	**80.2**	96.3	98.5	64.2	90.1	95.4	524.7
	PC2	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
	CREAM	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
	SPS	**79.5**	95.0	**98.0**	**60.5**	84.3	89.8	**507.1**	79.8	**96.4**	**98.6**	**64.3**	**90.5**	**95.8**	**525.5**
40%	SGRAF	8.3	18.1	31.4	5.3	16.7	21.3	101.1	15.8	23.4	54.6	17.8	43.6	54.1	209.3
	NCR	68.1	89.6	94.8	51.4	78.4	84.8	467.1	74.7	94.6	98.0	59.6	88.1	94.7	509.7
	DECL	72.7	92.3	95.4	53.4	79.4	86.4	479.6	75.6	95.5	98.3	59.5	88.3	94.8	512.0
	BiCro	74.6	92.7	96.2	55.5	81.1	87.4	487.5	77.0	**95.9**	98.3	61.8	89.2	94.9	517.1
	L2RM	75.8	93.2	96.9	56.3	81.0	87.3	490.5	77.5	95.8	98.4	62.0	89.1	94.9	517.7
	PC2	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
	CREAM	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
	SPS	**77.8**	**93.6**	**97.1**	57.3	**83.5**	**89.6**	**498.9**	**79.2**	**95.9**	**98.5**	**63.3**	**89.8**	**95.4**	**522.1**
60%	SGRAF	2.3	5.8	10.9	1.9	6.1	8.2	35.2	0.2	3.6	7.9	1.5	5.9	12.6	31.7
	NCR	13.9	37.7	50.5	11.0	30.1	41.4	184.6	0.1	0.3	0.4	0.1	0.5	1.0	2.4
	DECL	65.2	88.4	94.0	46.8	74.0	82.2	450.6	73.0	94.2	97.9	57.0	86.6	93.8	502.5
	BiCro	67.6	90.8	94.4	51.2	77.6	84.7	466.3	73.9	94.4	97.8	58.3	87.2	93.9	505.5
	L2RM	70.0	90.8	95.4	51.3	76.4	83.7	467.6	75.4	94.7	97.9	59.2	87.4	93.8	508.4
	PC2	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
	CREAM	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
	SPS	**73.4**	**92.7**	**96.3**	**53.7**	**80.2**	**87.7**	**484.1**	**77.6**	**95.7**	**98.3**	**61.6**	**89.0**	**95.1**	**517.2**

Table 1: The retrieval performance on Flickr30K and MS-COCO 1K under 20%, 40% and 60% noise rates separately. The best results and the second best results are respectively marked by **bold** and underline.

Method	Image → Text			Text → Image			
	R@1	R@5	R@10	R@1	R@5	R@10	RSum
SCAN	30.5	55.3	65.3	26.9	53.0	64.7	295.7
IMRAM	33.1	57.6	68.1	29.0	56.8	67.4	312.0
SAF	31.7	59.3	68.2	31.9	59.0	67.9	318.0
SGR	35.0	63.4	73.3	34.9	63.0	72.8	342.4
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
BiCro	40.8	67.2	76.1	42.1	67.6	76.4	370.2
PC2	39.3	66.4	75.4	39.8	66.4	73.2	355.6
CREAM	40.3	68.5	77.1	40.2	68.2	78.3	372.6
SPS	**40.8**	**67.9**	**77.7**	**42.4**	**69.5**	**78.0**	**376.3**

Table 2: The retrieval performance on CC152K, The best results are marked by **bold**.

the former, we report experimental results for noise rates of 20%, 40%, and 60%. The baselines include standard cross-modal retrieval model that lack robustness to noise, SCAN[Lee *et al.*, 2018], VSRN[Li *et al.*, 2019b], IMRAM[Chen *et al.*, 2020a], SAF and SGRAF[Diao *et al.*, 2021], noise-correcting methods NCR[Huang *et al.*, 2021b] and DECL[Qin *et al.*, 2022], and recently proposed noise-resistant approaches BiCro[Yang *et al.*, 2023], L2RM[Han *et al.*, 2024], PC2[Duan *et al.*, 2024], and CREAM[Ma *et al.*, 2024].

Results on Simulated Noise

Following [Huang *et al.*, 2021b], we adopt two evaluation protocols to validate the performance on MS-COCO: 5-fold cross-validation on 1,000 test images (referred to as MS-COCO 1K), and evaluation on the full 5,000 test images (referred to as MS-COCO 5K). Due to space limitations, Table 1 only reports the bidirectional retrieval results on Flickr30K and MS-COCO 1K compared to recent models. For the complete experimental results, please refer to the appendix. The results for MS-COCO 5K are reported in Table 3. The experimental results demonstrate that SPS significantly outperforms existing methods in terms of robustness to Noisy Correspondence, achieving notably higher RSum scores than all baselines. Specifically, SPS outperforms the previous best baseline, CREAM, on Flickr30K and MS-COCO at noise ratios of 20%, 40%, and 60% by margins of 4.8, 3.8, and 6.7 on Flickr30K and 2.5, 5.3, and 7.3 on MSCOCO, respectively. Notably, as the noise rate increases to 60%, SPS maintains high stability, particularly on MS-COCO 5K, where it surpasses the second-best method by 13.5, demonstrating a significant improvement.

Results on Inherent Noise

To further validate the noise robustness of SPS in real-world application scenarios, we report results on the CC152K dataset, as shown in Table 2. According to the results, SPS achieves the best performance with an overall RSum score of 376.3. Compared to the backbone network SGR and the robustness method NCR, SPS achieves improvements of 9.9% and 4.8%, respectively. The experiments demonstrate that SPS effectively handles both simulated and real-world noisy environments.

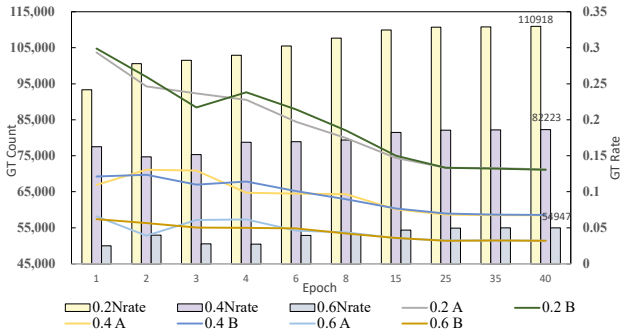

Figure 3: Illustration of the distribution changes of GT in Flickr30K under different noise rates throughout the training process.

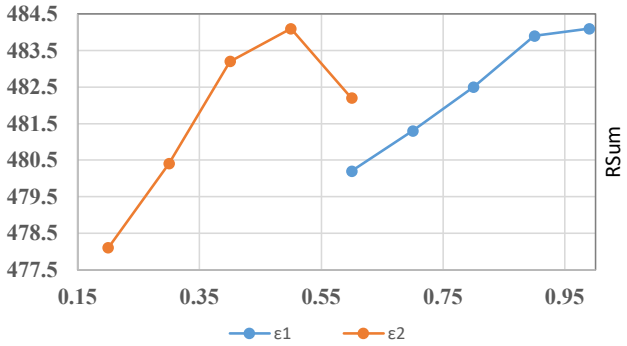

Figure 4: Illustration of the relationship between ϵ_1 and RSum when ϵ_2 is fixed at 0.5, and the relationship between ϵ_2 and RSum when ϵ_1 is fixed at 0.99.

4.4 Ablation Study

In this section, we conduct ablation experiments on the proposed SPS to validate the contribution of each major module within the framework, including WarmUp , Space Constraint (SCT), L^{qc}, and L^{n}. All experiments are performed on Flickr30K with a 40% noise rate, and to ensure a fair comparison, the same parameter settings are used across all experiments. The results are reported in Table 4. From the results, we observe that the best noise robustness is achieved only when all components are included, which fully demonstrates the effectiveness of each module.

4.5 Analytical Experiments

We analyze the impacts of hyper-parameters ϵ_1 and ϵ_2 under a 60% noise rate on Flick30K, as illustrated in the Figure 4 . It can be observed that the best retrieval performance is achieved when the threshold ϵ_1 is set to a relatively strict value, specifically $\epsilon_1 = 0.99$. We attribute this to the fact that constructing a stable feature space highly depends on reliable positive pairs. Therefore, when the dataset permits, it is reasonable to impose stricter requirements on the data partitioned into the reliable set.Additionally, for the parameter ϵ_2, experimental results indicate that maintaining a moderate value, specifically $\epsilon_2 = 0.5$ yields the best performance. This is intuitive, as it effectively distinguishes the quasi-clean set from the noisy set. Next, we investigated the performance of SPS in terms of dataset partitioning accuracy. Specifically, we observed the distribution changes of Ground Truth (GT) (i.e., truly clean pairs) in the Flickr30K dataset with different noise rates (20%, 40%, and 60%) throughout the training process, as shown in Figure 3. This includes the number of GT samples classified into the reliable set (denoted as GT Count) and the proportion of GT samples in the noisy set (denoted as GT Rate). The results demonstrate that SPS can accurately partition the data even as the noise rate increases, which aligns with the superior retrieval performance observed in the previous experiments.

Noise	Method	Image → Text			Text → Image			RSum
		R@1	R@5	R@10	R@1	R@5	R@10	
20%	SCAN	11.6	32.2	44.8	7.3	23.5	35.9	155.4
	IMRAM	17.0	44.4	59.4	15.6	38.0	50.8	225.1
	NCR	55.0	82.2	90.7	39.6	68.8	79.8	416.1
	DECL	57.3	83.3	90.7	40.0	69.1	79.8	420.1
	CREAM	57.6	84.1	91.6	41.4	71.1	81.2	427.0
	L2RM	**59.6**	**85.1**	**92.0**	42.5	71.5	81.3	432.0
	SPS	**59.6**	84.9	91.8	**42.6**	**71.7**	**81.9**	**432.5**
40%	SCAN	12.5	33.1	46.0	6.7	21.1	32.5	151.9
	IMRAM	13.5	34.9	49.5	13.6	34.6	47.4	193.5
	NCR	55.5	82.2	89.8	39.5	68.3	79.1	414.4
	DECL	53.4	81.4	89.4	38.6	67.2	78.3	408.3
	CREAM	55.3	82.3	90.6	39.8	69.3	80.1	417.3
	L2RM	57.1	83.4	91.0	40.8	69.4	79.7	421.4
	SPS	**58.1**	**84.0**	**91.8**	**41.6**	**70.6**	**81.0**	**427.1**
60%	SCAN	10.8	30.0	42.4	5.6	18.7	29.5	136.9
	IMRAM	10.7	30.8	44.2	11.6	30.4	42.6	170.3
	NCR	49.9	78.5	87.9	36.1	65.4	76.5	394.3
	DECL	39.1	69.1	80.5	28.4	56.4	68.6	342.0
	CREAM	52.1	80.4	89.0	37.8	66.9	78.0	404.3
	L2RM	53.5	81.0	88.9	37.3	65.7	76.7	403.1
	SPS	**55.6**	**82.8**	**90.5**	**39.4**	**68.8**	**79.5**	**416.6**

Table 3: The retrieval performance on MS-COCO 5K. The best results are marked by **bold**.

SCT	L^{qc}	L^{n}	WarmUp	Image → Text		Text → Image		RSum
				R@1	R@10	R@1	R@10	
✓	✓	✓	✓	**77.8**	**97.1**	**57.3**	**89.6**	**498.9**
	✓	✓	✓	76.1	97.0	55.8	88.9	493.4
✓		✓	✓	75.9	96.8	56.4	88.7	494.5
✓	✓		✓	75.3	97.0	56.7	88.3	494.1
			✓	70.9	95.3	52.3	86.7	475.6
✓	✓	✓		0.5	5.0	0.4	6.4	17.6

Table 4: Ablation study on the major components of SPS using the Flicker30K with 40% noise. The best results are marked by **bold**.

5 Conclusion

In this paper, we investigate a relatively underexplored issue in the field of cross-modal retrieval: the Noisy Correspondence problem. To address this challenge, we propose a reliable set-driven approach to construct a stable feature space. Building on this foundation, we seek proxy points for noisy data to enable reliable feature learning. Extensive experiments on several cross-modal datasets demonstrate the effectiveness of the proposed method. In the future, we will explore the implementation and improvement of this framework in other cross-modal domains.

Acknowledgments

This work was supported in part by the National Key Research & Development Program of China under Grant 2022YFA1004100, the Medico-Engineering Cooperation Funds from University of Electronic Science and Technology of China (ZYGX2022YGRH009 and ZYGX2022YGRH014), and supported by National Natural Science Foundation of China under Grant 62476048.

References

[Arpit *et al.*, 2017] Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks, 2017.

[Chen *et al.*, 2020a] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. IMRAM: iterative matching with recurrent attention memory for cross-modal image-text retrieval. *CoRR*, abs/2003.03772, 2020.

[Chen *et al.*, 2020b] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. *CoRR*, abs/2011.04305, 2020.

[Cheng *et al.*, 2022] Mengjun Cheng, Yipeng Sun, Longchao Wang, Xiongwei Zhu, Kun Yao, Jie Chen, Guoli Song, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Vista: Vision and scene text aggregation for cross-modal retrieval, 2022.

[Cuturi, 2013] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances, 2013.

[Diao *et al.*, 2021] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. *CoRR*, abs/2101.01368, 2021.

[Duan *et al.*, 2024] Yue Duan, Zhangxuan Gu, Zhenzhe Ying, Lei Qi, Changhua Meng, and Yinghuan Shi. Pc2: Pseudo-classification based pseudo-captioning for noisy correspondence learning in cross-modal retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 9397–9406. ACM, October 2024.

[Faghri *et al.*, 2018] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives, 2018.

[Goel *et al.*, 2022] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A. Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining, 2022.

[Hambarde and Proenca, 2023] Kailash A Hambarde and Hugo Proenca. Information retrieval: recent advances and beyond. *IEEE Access*, 2023.

[Han *et al.*, 2018] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-sampling: Training robust networks for extremely noisy supervision. *CoRR*, abs/1804.06872, 2018.

[Han *et al.*, 2023] Haochen Han, Kaiyao Miao, Qinghua Zheng, and Minnan Luo. Noisy correspondence learning with meta similarity correction, 2023.

[Han *et al.*, 2024] Haochen Han, Qinghua Zheng, Guang Dai, Minnan Luo, and Jingdong Wang. Learning to rematch mismatched pairs for robust cross-modal retrieval, 2024.

[Hu *et al.*, 2021] Peng Hu, Xi Peng, Hongyuan Zhu, Liangli Zhen, and Jie Lin. Learning cross-modal retrieval with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5403–5413, 2021.

[Hu *et al.*, 2023] Peng Hu, Zhenyu Huang, Dezhong Peng, Xu Wang, and Xi Peng. Cross-modal retrieval with partially mismatched pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9595–9610, 2023.

[Huang *et al.*, 2021a] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 10944–10956. Curran Associates, Inc., 2021.

[Huang *et al.*, 2021b] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. *Advances in Neural Information Processing Systems*, 34:29406–29419, 2021.

[Huang *et al.*, 2021c] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29406–29419. Curran Associates, Inc., 2021.

[Jia *et al.*, 2021] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *CoRR*, abs/2102.05918, 2021.

[Jiang *et al.*, 2023] Qian Jiang, Changyou Chen, Han Zhao, Liqun Chen, Qing Ping, Son Dinh Tran, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Understanding and constructing latent modality structures in multi-modal representation learning, 2023.

[Kingma, 2014] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Lee *et al.*, 2018] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. *CoRR*, abs/1803.08024, 2018.

[Li *et al.*, 2019a] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan Kankanhalli. Learning to learn from noisy labeled data, 2019.

[Li *et al.*, 2019b] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. *CoRR*, abs/1909.02701, 2019.

[Li *et al.*, 2020] Junnan Li, Richard Socher, and Steven C. H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *CoRR*, abs/2002.07394, 2020.

[Liang *et al.*, 2022] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning, 2022.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[Liu and Tao, 2016] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):447–461, March 2016.

[Ma *et al.*, 2024] Xinran Ma, Mouxing Yang, Yunfan Li, Peng Hu, Jiancheng Lv, and Xi Peng. Cross-modal retrieval with noisy correspondence via consistency refining and mining. *IEEE Transactions on Image Processing*, 33:2587–2598, 2024.

[Morris *et al.*, 2024] Meredith Ringel Morris, Jascha Sohl-dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. Levels of agi for operationalizing progress on the path to agi, 2024.

[Permuter *et al.*, 2006] Haim Permuter, Joseph Francos, and Ian Jermyn. A study of gaussian mixture models of color and texture features for image classification and segmentation. *Pattern recognition*, 39(4):695–706, 2006.

[Pu *et al.*, 2022] Shi Pu, Kaili Zhao, and Mao Zheng. Alignment-uniformity aware representation learning for zero-shot video classification, 2022.

[Qin *et al.*, 2022] Yang Qin, Dezhong Peng, Xi Peng, Xu Wang, and Peng Hu. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 4948–4956, New York, NY, USA, 2022. Association for Computing Machinery.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021.

[Sharma *et al.*, 2018] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[Wang and Isola, 2022] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere, 2022.

[Wang *et al.*, 2023] Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Equivariant similarity for vision-language foundation models, 2023.

[Xia *et al.*, 2020] Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *International conference on learning representations*, 2020.

[Yang *et al.*, 2023] Shuo Yang, Zhaopan Xu, Kai Wang, Yang You, Hongxun Yao, Tongliang Liu, and Min Xu. Bicro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency, 2023.

[Yang *et al.*, 2024] Mouxing Yang, Yunfan Li, Changqing Zhang, Peng Hu, and Xi Peng. Test-time adaptation against multi-modal reliability bias. In *The Twelfth International Conference on Learning Representations*, 2024.

[Young *et al.*, 2014] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 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, 2014.

[Yu *et al.*, 2019] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor W. Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? *CoRR*, abs/1901.04215, 2019.

[Zha *et al.*, 2024] Quanxing Zha, Xin Liu, Yiu-ming Cheung, Xing Xu, Nannan Wang, and Jianjia Cao. 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*, SIGIR '24, page 852–861, New York, NY, USA, 2024. Association for Computing Machinery.

[Zolfaghari *et al.*, 2021] Mohammadreza Zolfaghari, Yi Zhu, Peter Gehler, and Thomas Brox. Crossclr: Cross-modal contrastive learning for multi-modal video representations, 2021.