

# NeighborRetr: Balancing Hub Centrality in Cross-Modal Retrieval

Zengrong Lin<sup>1\*</sup> Zheng Wang<sup>1\*†</sup> Tianwen Qian<sup>2</sup> Pan Mu<sup>1</sup> Sixian Chan<sup>1</sup> Cong Bai<sup>1</sup>  
 {linzengrong, zhengwang, panmu, sxchan congbai}@zjut.edu.cn, twqian@cs.ecnu.edu.cn  
<sup>1</sup>College of Computer Science, Zhejiang University of Technology, Zhejiang, China  
<sup>2</sup>College of Computer Science and Technology, East China Normal University, Shanghai, China

## Abstract

Cross-modal retrieval aims to bridge the semantic gap between different modalities, such as visual and textual data, enabling accurate retrieval across them. Despite significant advancements with models like CLIP that align cross-modal representations, a persistent challenge remains: the hubness problem, where a small subset of samples (hubs) dominate as nearest neighbors, leading to biased representations and degraded retrieval accuracy. Existing methods often mitigate hubness through post-hoc normalization techniques, relying on prior data distributions that may not be practical in real-world scenarios. In this paper, we directly mitigate hubness during training and introduce NeighborRetr, a novel method that effectively balances the learning of hubs and adaptively adjusts the relations of various kinds of neighbors. Our approach not only mitigates the hubness problem but also enhances retrieval performance, achieving state-of-the-art results on multiple cross-modal retrieval benchmarks. Furthermore, NeighborRetr demonstrates robust generalization to new domains with substantial distribution shifts, highlighting its effectiveness in real-world applications. We make our code publicly available at: <https://github.com/NeighborRetr>.

## 1. Introduction

Cross-modal retrieval has long been the touchstone of cross-modal representation learning, with the primary goal of bridging the semantic gap between different modalities for accurate retrieval. In recent years, visual-language models such as CLIP [39] have gained significant attention due to their remarkable performance in aligning cross-modal representations. By leveraging large-scale paired datasets, these models enable contrastive learning to align visual and textual data distributions, which has notably advanced cross-modal retrieval tasks.

\* Equal Contribution.

† Corresponding Author: Zheng Wang <zhengwang@zjut.edu.cn>.

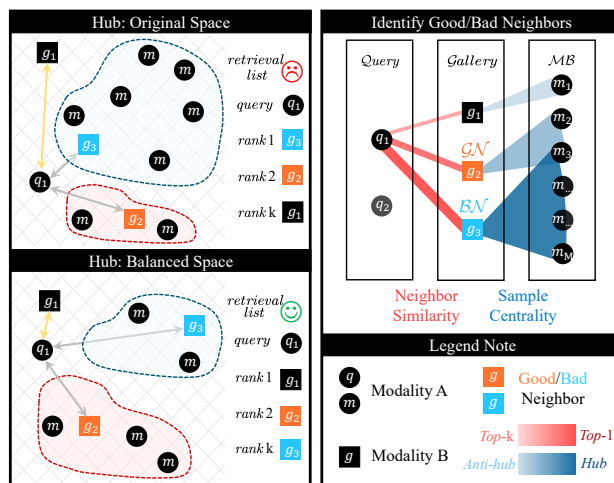


Figure 1. **Left: Hubness Balancing.** In the original embedding space (top left), bad hubs such as  $g_3$  dominate the neighborhood, leading to unsatisfying retrieval performance. NeighborRetr rebalances the neighborhood relations (bottom left) by adaptively bringing good hubs closer, such as  $g_2$ , effectively mitigating the hubness problem. **Right: Good/Bad Neighbors Identification.** NeighborRetr identifies cross-modality hubs based on sample centrality with Memory bank (MB), and distinguishes between good neighbors (GN) and bad neighbors (BN) by considering both the neighbor similarity and the sample centrality.

However, despite these advancements, a significant challenge remains: hubness. Hubness is a common phenomenon in high-dimensional embedding spaces [47], where a small subset of samples, referred to as *hubs*, frequently emerge as the nearest neighbors to other samples, while the majority of other samples rarely appear as neighbors. As illustrated in the top left of Fig. 1, hubs such as  $g_3$  dominate the nearest-neighbor rankings, leading to biased representations and degraded retrieval accuracy.

Even large-scale pretrained models like CLIP are not immune to the hubness problem [3]. Recently, some methods have attempted to mitigate hubness by applying post-hoc similarity normalization techniques during inference. These methods either leverage the distribution of the test set [5] or

construct additional modality-specific banks from training samples [3, 50]. Lately, NNN [6] normalizes the retrieval score using the k-closest queries from a reference dataset. However, these approaches rely heavily on knowing the prior data distribution of either the training or test sets, which can be problematic in real-world scenarios where the test set distribution may shift or remain unknown. This limitation highlights the importance of addressing hubness during training to ensure robust performance across varying distributions. To tackle hubness during training, earlier methods, such as HAL [35], scale up the loss of hubs of both positive and negative samples. However, this strategy does not differentiate between semantically relevant and irrelevant hubs, potentially punishing beneficial hubs. A more nuanced approach that distinguishes between relevant (good) hubs and irrelevant (bad) hubs is essential for effectively mitigating the hubness problem while maximizing retrieval accuracy.

To achieve nuanced hubness reduction during training, we empirically show in Sec. 3.2 and Fig. 2(a) that although semantical relevant hubs and irrelevant hubs co-exist, while relevant hubs exhibit lower retrieval frequency than irrelevant ones, making it possible to distinguish them. Inspired by this, we propose NeighborRetr, a novel method that aims to balance the centrality [19] of hubs in the embedding space. Rather than categorically rejecting all hubs, we introduce sample centrality for identifying hubs and further distinguish between good neighbors (semantically relevant samples) and bad neighbors (irrelevant or noisy samples) incorporating neighbor similarity with sample centrality. By selectively adjusting the affinity strength with all neighbors, NeighborRetr rebalances the neighbor relationships to bring good hubs closer than bad ones. Additionally, we leverage the uniform retrieving objective to further enforce the retrieval of anti-hubs.

Our main contributions are as follows: (i) We conduct an in-depth investigation into the cross-modal hubness problem, revealing the presence of both beneficial good hubs and detrimental bad hubs, as well as overlooked anti-hubs. (ii) We introduce NeighborRetr, a novel method that effectively balances the learning of hubs and adjusts neighborhood relationships to mitigate hubness. (iii) Our approach not only mitigates the hubness problem but also enhances retrieval performance across multiple benchmarks, achieving state-of-the-art results on four text-video benchmarks and three text-image benchmarks. In addition, our method demonstrates robust generalization to new domains with substantial distribution shifts.

## 2. Related Works

**Cross-Modal Matching.** Triplet ranking loss [30] takes the sum of all negative samples for visual semantic learning. VSE++ [13] emphasizes hard negatives and contrasts the

hardest negative sample. Wray et al. [52] argue that visual similarity does not imply semantic similarity, and propose several proxies measuring text similarity to allow relative ranking. PVSE [42] represents each sample as a set of embedding vectors. PCME [8] introduces probabilistic embedding representing each sample as a set of vectors sampled from a normal distribution. Li et al. [32] utilize language pre-training models to identify semantic similarity and adjust the margin of false negatives. Kim et al. [27] regard semantic alignment as a set prediction problem and propose the smooth-Chamfer similarity to address the set alignment problem. MSRM [51] introduces hypergraph into the one-to-many correspondence of image-text retrieval.

**Hubness Problem.** The hubness is a fundamental property observed in high-dimensional representation. Jian et al. [22] highlight representation degeneration in multimodal data, hindering retrieval performance. It has been empirically observed that hubness also exists for cross-modal matching. Liu et al. [34] looked into the hubness problem in cross-modal matching, and proposed the kNN-margin loss that considers all k-hardest samples as negatives. Furthermore, HAL [35] alleviates the hubness problem by using global measurements for hubness level and scaling up hub losses during learning. In addition to training adjustments, recently, more methods handle hubness during inference. Dual Softmax [5] performs prior normalization in both directions as a postprocessing step. QB-NORM [3] reduces the prominence of the hub with the query bank built from queries in the training set. DBNORM [50] leverages two banks of query and gallery samples from the training set to reduce hubness during inference.

**Visual-Text Retrieval.** ViLT [28] introduces a minimal vision-and-language transformer, simplifying visual input processing. SGRAF [10] further enhances image-text retrieval by effectively capturing and filtering both local and global alignments. To enhance video representations for text-video retrieval, approaches like CLIP4Clip [37] have sought to overcome this obstacle by leveraging knowledge transferred from CLIP [39], while still regards videos as a whole representation. X-Pool [18] allows a text to focus on its most semantically similar video frames, subsequently creating an aggregated video representation conditioning on these specific frames. TS2-Net [36] introduces the token shift and selection transformer, enhancing the temporal and spatial video representations. To enable fine-grained cross-modal matching for text-video retrieval, UATVR [14] allows for both global and local probabilistic alignment between text and video feature representations. HBI [24] incorporates Hierarchical Banzhaf Interaction to value possible correspondence between video frames and text words. DiCoSA [25] aligns text and video via semantic concepts, using adaptive pooling for set-to-set and partial matching. To regularize semantic samples, Tomas et al. [43] propose a

novel approach to prioritize loosely aligned samples.

### 3. Problem Setup

#### 3.1. Preliminary

**Task definition.** Generally, given a corpus of visual-text pairs  $(v, t)$ , cross-modal representation learning aims to learn a visual encoder  $\Phi_v$  and a text encoder  $\Phi_t$ . The optimization target is formulated as a cross-modality similarity measurement  $S(\Phi_v(v), \Phi_t(t))$  by contrasting cross-modal samples, where maximize the similarity for matched pairs and minimize it for mismatched pairs. For cross-modal retrieval, a query  $x$  from one modality is compared with gallery  $y$  from another modality in the joint embedding space. The goal is to generate a ranked list of galleries that best match the query  $x$ .

**Hubness Problem.** In a dataset  $\mathcal{X} \subset \mathbb{R}^d$ , the affinity between a sample  $x \in \mathcal{X}$  with other samples  $x' \in \mathcal{X}$  can be measured by inner product similarity, with the nearest neighbors being those with the maximum similarities. The  $k$ -occurrence of  $x$ , denoted as  $N_k(x)$ , is the number of times  $x$  appears among the  $k$ -nearest neighbors of other samples. A sample is considered a *hub* when  $N_k(x)$  is significantly higher than others. It further leads to the phenomenon known as *hubness*, where a small portion of samples become hubs. As described by Radovanovic et al. [40], the distribution of  $N_k(\mathcal{X})$  often exhibits high skewness, indicating the presence of hubness. Oppositing to hubs, *anti-hubs* are those samples that appear in very few neighbors of other samples.

**Good/Bad Hubs.** In the presence of class labels, Tomavsev et al. [47] introduce the concepts of *good hub* and *bad hub*. Specifically,  $x$  is a *good (or bad) neighbor* of another sample  $x'$  if: (i)  $x$  is among the  $k$ -nearest neighbors of  $x'$ , and (ii)  $x$  and  $x'$  have the same (or different) class label. The good (or bad)  $k$ -occurrence of  $x$ , denoted as  $GN_k(x)$  (or  $BN_k(x)$ ), counts how many other samples have  $x$  as one of their good (or bad)  $k$ -nearest neighbors. A sample  $x$  is a *good hub (or bad hub)* if  $GN_k(x)$  (or  $BN_k(x)$ ) is exceptionally large.

#### 3.2. Observation and Analysis

**Hubness Observation.** In typical cross-modal retrieval tasks, each sample has a single labeled positive counterpart, with all other counterparts treated as negatives. This setup contrasts with real-world scenarios where multiple positive counterparts may exist. To categorize unlabeled cross-modal pairs as pseudo-positive or negative, we propose using intra-text similarity as a probe to identify potential cross-modal multi-correlations. Concretely, we use CLIP’s text encoder to obtain the representations of the text originally paired with the image and consider these with high textual similarity as additional positives (further ra-

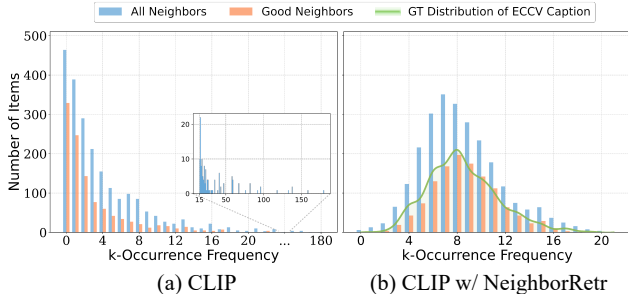


Figure 2. Distribution of  $k$ -occurrence frequency among all textual similarities for (a) vanilla CLIP and (b) CLIP finetuned with NeighborRetr. The x-axis represents how often an image appears in queries’ top-15 nearest neighbors. The y-axis counts the number of images at each frequency. Blue bars indicate all neighbors, orange bars highlight positive pairs among neighbors, and the green line represents the ground truth distribution.

tionale for this is provided in the supplementary material). Then, to investigate the hubness issue of cross-modal embedding space, we compare cross-modal similarities across all unlabeled samples and count how often each image appears among the top 15 nearest neighbors ( $k$ -occurrence) of text queries in the test set of ECCV Caption [9]. The results are presented in Fig. 2 and analyzed below.

**Hubness Analysis.** In Fig.2 (a), we show the  $k$ -occurrence frequency of text-to-image retrieval. From this plot, we draw three observations: (i) The distribution of all neighbors exhibits an extreme long-tail pattern, with a high presence of bad hubs (undesirable large  $N_k(x)$ ), as highlighted in the inset. (ii) Good neighbors are distributed across a range of lower  $N_k(x)$  frequencies. (iii) Many anti-hubs (with zero or very small  $N_k(x)$ ) seldom appear among the top-15 nearest neighbors of any query.

To address these issues, rather than indiscriminately eliminating all hubs and ignoring anti-hubs, our proposed NeighborRetr method mitigates the hubness problem in multiple aspects. In Fig.2 (b), we analyze the CLIP model fine-tuned with NeighborRetr, showing that (i) bad hubs are reduced, (ii) good hubs are enhanced, and (iii) anti-hubs are minimized. These changes lead to a distribution of good neighbors that better aligns with the ground truth. This indicates that NeighborRetr creates a more balanced embedding space, capturing true semantic correlations of ECCV Caption. In the following sections, we introduce the method of NeighborRetr and further explain how it achieves such changes.

## 4. Methodology

### 4.1. Sample Centrality

Centrality, introduced in [40], refers to the tendency of samples closer to the global centroid to exhibit greater affinity to all other samples. Hubness can be understood as a

kind of local centrality [45], reflecting a localized centering phenomenon. Since hubs frequently appear among the  $k$ -nearest neighbors of other samples, we define a *centrality score* to measure a sample’s affinity with others. Directly calculating the centrality score across all training set samples can be computationally intensive, we manage this by maintaining a queue [20] as a memory bank  $\mathcal{M}$  to efficiently compute the uni-modality centrality score:

$$C(x_i) = \frac{1}{|\mathcal{M}|} \sum_{j=1}^{|\mathcal{M}|} \frac{x_i^\top x_j}{\|x_i\| \cdot \|x_j\|}, \quad (1)$$

where  $|\mathcal{M}|$  is the size of the memory bank.  $C(x_i)$  measures the average similarity of  $x_i$  to all memory samples  $x_j$ . A high centrality score suggests a strong global affinity.

Similarly, in the joint embedding space, the centrality score for a sample  $y_j$  in one modality can be defined as its average similarity with all samples  $x_i$  in another modality:

$$\hat{C}(y_j) = \frac{1}{|\mathcal{M}|} \sum_{i=1}^{|\mathcal{M}|} \frac{y_j^\top x_i}{\|y_j\| \cdot \|x_i\|}. \quad (2)$$

## 4.2. Centrality Weighting

Previous methods for mitigating cross-modal hubness [35] apply inter-modality centrality scores for softly weighting various hubs. Here, we propose to weight hubs leveraging centrality within the intra-modality similarity. We define the centrality weight as:

$$w(x_i) = \exp(C(x_i)/\kappa), \quad (3)$$

where  $C(x_i)$  is the uni-modality centrality score as in Eq. 1, and  $\kappa$  is a scaling hyperparameter. To emphasize the learning of hubs, we incorporate the centrality weight into the contrastive loss, defining the Centrality Weighting Loss as:

$$\mathcal{L}_{Wti} = -\frac{1}{B} \sum_{i=1}^B w(x_i) \log \frac{\exp(S(x_i, y_i))}{\sum_j^B \exp(S(x_i, y_j))}. \quad (4)$$

where  $B$  is the batch size, and  $S$  is the cross-modal similarity. Here, we weight the contrastive loss with intra-modality centrality scores instead of cross-modal centrality scores because intra-modality scores provide a more stable representation of hub within each modality, enabling precise hub identification, whereas cross-modal scores may conflate interdependencies within different modalities and obscure modality-specific hub centrality. By incorporating intra-modal weights into the loss, we can accurately identify hubs with high centrality scores and emphasize the centrality of hubs within each modality.

## 4.3. Neighbor Adjusting

During contrastive learning, it is crucial to handle hubs effectively by bringing good hubs closer to the anchor sample and distancing it from bad ones. Ideally, if multi-correlations are available, a supervised contrastive learning function can be employed. However, in most cases, each sample has only a single labeled positive counterpart. In addition, the cross-modal similarity measure is often biased in positive-negative discrimination due to the hubness issue within the embedding space. Moreover, centrality scores alone can only identify whether a sample is a hub, without distinguishing between good and bad hubs. To address this, we incorporate the centrality score into the cross-modal similarity measure, rewriting it as:

$$\tilde{S}(x_i, y_j) = S(x_i, y_j) - \hat{C}(y_j). \quad (5)$$

With this de-centrality cross-modal similarity, we focus on balancing contrastive neighbors within data batches. Specifically, for a query  $x_i$ , we define a set of gallery samples  $\mathcal{N}(x_i)$  as its closest neighbors. For convenience, we also define the neighbor set including the ground truth  $y_i$ :  $\mathcal{N}^+(x_i) = \{y_i\} \cup \mathcal{N}(x_i)$ . To differentiate various kinds of affinity in the neighborhood, we introduce the Neighbor Adjusting Loss  $\mathcal{L}_{Nbi}$  as:

$$\mathcal{L}_{Nbi}(x_i) = - \sum_{y_j \in \mathcal{N}^+(x_i)} \mathcal{H}(y_j) \log P(y_j | \mathcal{N}^+(x_i)), \quad (6)$$

where

$$P(y_j | \mathcal{N}^+(x_i)) = \frac{\exp(S(x_i, y_j))}{\sum_{k=1}^{|\mathcal{N}^+|} \exp(S(x_i, y_k))}, \quad (7)$$

and

$$\mathcal{H}(y_j) = \frac{\exp(\tilde{S}(x_i, y_j))}{\sum_{y_k \in \mathcal{N}(x_i)} \exp(\tilde{S}(x_i, y_k))}, \quad (8)$$

and  $\mathcal{H}(y_j) = 1.0$  if  $y_j = y_i$ . Here,  $P(y_j | \mathcal{N}^+(x_i))$  quantifies the likelihood of  $x_i$  matching  $y_j$  among neighbors, normalized over  $\mathcal{N}^+(x_i)$  using their similarity scores.  $\mathcal{H}(y_j)$  adaptively adjusts the contribution of each neighbor based on  $\tilde{S}(x_i, y_j)$ . Intuitively,  $\mathcal{H}(y_j)$  penalizes bad hubs with high centrality while promoting good hubs. And the gradient of  $\mathcal{L}_{Nbi}$  with respect to the similarity scores  $S_{i,j}$  is computed as:

$$\frac{\partial \mathcal{L}_{Nbi}(x_i)}{\partial S(x_i, y_j)} = P(y_j | \mathcal{N}^+(x_i)) - \mathcal{H}(y_j). \quad (9)$$

The gradient adaptively adjusts similarity scores to promote good hubs and penalize bad ones. Specifically, when  $P < \mathcal{H}(y_j)$  (good hub), the negative gradient increases  $S(x_i, y_j)$ , promoting the hub. Conversely, when  $P > \mathcal{H}(y_j)$  (bad hub), the positive gradient decreases  $S(x_i, y_j)$ , penalizing the hub. This adaptively balances the representation of good and bad hubs.

#### 4.4. Uniform Regularization

While  $\mathcal{L}_{Wti}$  and  $\mathcal{L}_{Nbi}$  losses mitigate hubness by targeting hubs, they overlook anti-hubs that are rarely or never retrieved. To address the anti-hub issue [46], inspired by [11], we propose a uniform marginal constraint that treats all samples with equal probability of retrieval, ensuring anti-hubs have retrieval probabilities comparable to normal samples. This uniform retrieving objective enforces balanced retrieval probabilities across all samples:

$$\begin{aligned} \max_{\mathbf{Q} \in \mathbb{R}_+^{n \times m}} \quad & \langle \mathbf{Q}, \mathbf{S} \rangle = \sum_{i=1}^n \sum_{j=1}^m \mathbf{Q}_{i,j} \mathbf{S}_{i,j}, \\ \text{s.t.} \quad & \mathbf{Q} \mathbf{1}_m = \frac{1}{n} \mathbf{1}_n, \quad \mathbf{Q}^\top \mathbf{1}_n = \frac{1}{m} \mathbf{1}_m, \end{aligned} \quad (10)$$

where  $\langle \mathbf{Q}, \mathbf{S} \rangle$  denotes the Frobenius inner product, and  $n$ ,  $m$  are sizes of the query and gallery sets, respectively. The constraints  $\frac{1}{n} \mathbf{1}_n$  and  $\frac{1}{m} \mathbf{1}_m$  enforce uniform marginal distributions, ensuring bidirectional retrieval by equally matching each query with all gallery samples and vice versa. This guarantees that every sample, including overlooked anti-hubs, contributes equally to the Uniform Retrieving Matrix  $\mathbf{Q}$ , making anti-hubs as likely to be retrieved as other normal samples. Based on this equality, the proposed Uniformity Regularization loss is formulated as:

$$\mathcal{L}_{Opt} = -\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m \mathbf{Q}_{i,j} \log P(j|i), \quad (11)$$

where

$$P(j|i) = \frac{\exp(\mathbf{S}(x_i, y_j))}{\sum_{k=1}^m \exp(\mathbf{S}(x_i, y_k))}. \quad (12)$$

And the gradient of  $\mathcal{L}_{Opt}$  with respect to the similarity scores  $\mathbf{S}_{i,j}$  is computed as:

$$\frac{\partial \mathcal{L}_{Opt}}{\partial \mathbf{S}_{i,j}} = -\frac{1}{n} (\mathbf{Q}_{i,j} - P(j|i)). \quad (13)$$

Here,  $\mathbf{Q}$  is obtained by solving Eq. (10). By updating  $\mathbf{S}_{i,j}$  to minimize  $\mathcal{L}_{Opt}$ ,  $P(j|i)$  aligns with  $\mathbf{Q}_{i,j}$ , leading to uniform probabilities  $P(j|\cdot) \rightarrow \frac{1}{m}$ . More details and proof of the losses are provided in the supplementary material.

#### 4.5. Visual-Text Learning

The framework of NeighborRetr is similar to fine-grained matching methods [14, 24], as shown in Fig. 3. To achieve fine-grained semantic alignment, the visual input  $v$  is embedded into patch sequence  $\mathbf{V}_p = \{v_p^i\}_{i=1}^{N_v}$ , where  $N_v$  is the length of patch  $v$ . The textual input  $t$  is embedded into word sequence  $\mathbf{T}_w = \{t_w^j\}_{j=1}^{N_t}$ , where  $N_t$  is the length of text  $t$ . The alignment matrix is defined as:  $A = [a_{ij}]^{N_v \times N_t}$ , where  $a_{ij} = \frac{v_p^i \cdot t_w^j}{\|v_p^i\| \|t_w^j\|}$  represents the alignment score between the  $i$ -th visual patch and the  $j$ -th text word. For

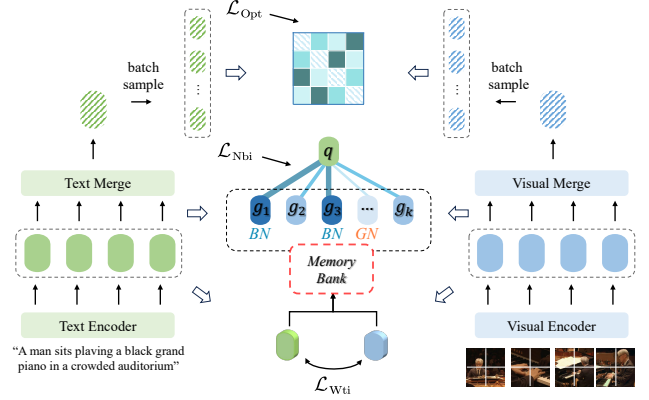


Figure 3. Overview of NeighborRetr on two-level hierarchy framework. At the low level,  $\mathcal{L}_{Wti}$  emphasizes the centrality of hubs and  $\mathcal{L}_{Nbi}$  balances neighborhood relations via comparing a memory bank. After token merging,  $\mathcal{L}_{Opt}$  enforces equal retrieval probabilities for overall matching at the high level.

the  $i$ -th visual patch, we calculate its maximum alignment score as  $\max_j a_{ij}$ . Then, we use the weighted average maximum alignment score overall visual patch as the visual-to-text similarity using the Weighted Token-wise Interaction (WTI) module [48]. Similarly, we can obtain the text-to-visual similarity. The total similarity score can be defined as:

$$\mathbf{S}_{v,t} = \frac{1}{2} \left( \underbrace{\sum_{i=1}^{N_v} \omega_v^i \max_j a_{ij}}_{\text{visual-to-text}} + \underbrace{\sum_{j=1}^{N_t} \omega_t^j \max_i a_{ij}}_{\text{text-to-visual}} \right), \quad (14)$$

where  $[\omega_v^1, \dots, \omega_v^{N_v}] = \text{Softmax}(\text{MLP}_v(\mathbf{V}_p))$  are the weights of the visual patches. We can similarly obtain the text weight. To obtain the high-level representation as a single token, we utilize DPC-KNN [12], a k-nearest neighborhood density peaks clustering algorithm, to cluster the visual (textual) tokens for token merging.

To stabilize the optimization, following [24], we introduce the KL divergence loss between the distribution of low level  $\mathbf{S}_{v,t}^l$  and high level  $\mathbf{S}_{v,t}^g$ , which can be formulated as:

$$\mathcal{L}_{KL} = \text{KL}(\mathbf{S}_{v,t}^g \| \mathbf{S}_{v,t}^l). \quad (15)$$

Therefore, the total training objective can be defined as:

$$\mathcal{L} = \frac{1}{2} \sum_{m \in \{2v, v2t\}} \{ \mathcal{L}_{Wti}^m + \mathcal{L}_{Nbi}^m + \mathcal{L}_{Opt}^m + \mathcal{L}_{KL}^m \}. \quad (16)$$

## 5. Experiments

### 5.1. Experimental Settings

**Datasets.** MSR-VTT [53] contains 10K videos, each with 20 text descriptions. We follow the training proto-

Methods	Text-to-Video						Video-to-Text					
	R@1↑	R@5↑	R@10↑	Rsum↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	Rsum↑	MdR↓	MnR↓
T2VLAD [49] <small>CVPR21</small>	29.5	59.0	70.1	158.6	4.0	-	31.8	60.0	71.1	162.9	3.0	-
CLIP4Clip [37] <small>Neurocomputing22</small>	44.5	71.4	81.6	197.5	<b>2.0</b>	15.3	42.7	70.9	80.6	194.2	<b>2.0</b>	11.6
X-Pool [18] <small>CVPR22</small>	46.9	72.8	82.2	201.9	<b>2.0</b>	14.3	44.4	73.3	84.0	201.7	<b>2.0</b>	9.0
TS2-Net [36] <small>ECCV22</small>	47.0	74.5	83.8	205.3	<b>2.0</b>	13.0	45.3	74.1	83.7	203.1	<b>2.0</b>	9.2
EMCL-Net [23] <small>NeurIPS22</small>	46.8	73.1	83.1	203.0	<b>2.0</b>	12.8	46.5	73.5	83.5	203.5	<b>2.0</b>	8.8
DiCoSA [25] <small>IJCAI23</small>	47.5	74.7	83.8	206.0	2.0	13.2	46.7	75.2	84.3	206.2	2.0	8.9
HBI [24] <small>CVPR23</small>	48.6	74.6	83.4	206.6	<b>2.0</b>	<b>12.0</b>	46.8	<b>74.3</b>	84.3	205.4	<b>2.0</b>	8.9
Diffusion [26] <small>ICCV23</small>	49.0	<b>75.2</b>	82.7	206.9	<b>2.0</b>	12.1	47.7	73.8	84.5	206.0	<b>2.0</b>	8.8
EERCF [44] <small>AAAI24</small>	47.8	74.1	<b>84.1</b>	206.0	-	-	44.7	74.2	83.9	202.8	-	-
MPT [55] <small>ACM MM24</small>	48.3	72.0	81.7	202.0	-	14.9	46.5	74.1	82.6	203.2	-	11.8
<b>NeighborRetr (Ours)</b>	<b>49.5</b>	74.1	<b>84.1</b>	<b>207.7</b>	<b>2.0</b>	12.8	<b>48.7</b>	74.2	<b>84.7</b>	<b>207.5</b>	<b>2.0</b>	<b>8.4</b>

Table 1. Comparisons to state-of-the-art methods on the MSR-VTT dataset. “↑” denotes higher is better. “↓” denotes lower is better. DiCoSA [25] utilizes QB-Norm [3] for inference and is grayed out for a fair comparison.

MSVD						ActivityNet Captions						DiDeMo					
Methods	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	Methods	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	Methods	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
FROZEN	33.7	64.7	76.3	3.0	-	CLIP4Clip	40.5	72.4	83.6	<b>2.0</b>	7.5	FROZEN	34.6	65.0	74.7	3.0	-
EMCL-Net	42.1	71.3	81.1	<b>2.0</b>	17.6	TS2-Net	41.0	73.6	84.5	<b>2.0</b>	8.4	TS2-Net	41.8	71.6	82.0	<b>2.0</b>	14.8
CLIP4Clip	45.2	75.5	84.3	<b>2.0</b>	10.3	DiCoSA	42.1	73.6	84.6	2.0	6.8	CLIP4Clip	42.8	68.5	79.2	<b>2.0</b>	18.9
UATVR	46.0	76.3	85.1	<b>2.0</b>	10.4	MPT	41.4	70.9	82.9	-	7.8	Diffusion	46.7	74.7	82.7	<b>2.0</b>	14.3
Diffusion	46.6	75.9	84.1	<b>2.0</b>	15.7	HBI	42.2	73.0	84.6	<b>2.0</b>	6.6	HBI	46.9	74.9	82.7	<b>2.0</b>	12.1
<b>Ours</b>	<b>47.9</b>	<b>77.3</b>	<b>86.0</b>	<b>2.0</b>	<b>9.2</b>	<b>Ours</b>	<b>46.0</b>	<b>75.9</b>	<b>86.5</b>	<b>2.0</b>	<b>6.1</b>	<b>Ours</b>	<b>48.2</b>	<b>76.7</b>	<b>84.9</b>	<b>2.0</b>	<b>11.9</b>

Table 2. Comparisons to other methods on the Text-to-Video task on the MSVD, ActivityNet Captions, and DiDeMo datasets.

col in [17] and evaluate on the 1K-A testing split. **ActivityNet Captions** [31] consists of densely annotated temporal segments of 20K videos. We use the 10K training split and test on the 5K “val1” split. **DiDeMo** [1] contains 10K videos with 40K text descriptions. We follow the training and evaluation protocol in [37]. **MSVD** [4] contains 1,970 videos, we split the train, validation, and test set with 1,200, 100, and 670 videos. **ECCV Caption** [9] is a machine-and-human-verified test, containing  $\times 8.5$  positive images and  $\times 3.6$  positive captions compared to MS-COCO. Experiments on **MS-COCO** [33] and **Flickr30K** [54] are listed in the supplementary material.

**Retrieval Metrics.** We choose Recall at rank K (R@K), Median Rank (MdR), Sum of Recall at rank {1, 5, 10} (Rsum), and mean rank (MnR) to evaluate the retrieval performance. We utilize mAP@R [38] and R-Precision (R-P) [38] metrics on the ECCV Caption to evaluate the ability of the model to recall incorrect negatives.

**Hubness Metrics.** To assess the hubness and the effectiveness of our mitigation strategies, we utilize two kinds of hubness metrics: **i) Distribution-based Metrics:** K-Skewness (skew) [40] and K-Skewness Truncated Normal (trunc) [45]. Atkinson Index (atkinson) [16] and Robin Hood Index (robin) [15]. **ii) Occurrence-based Metrics:** Antihub Occurrence (anti) [41] and Hub Occurrence (hub) [40]. More details on the definition of hubness met-

rics are in the supplementary material.

**Implementation Details.** Following previous works [23, 37], we initial the text encoder and visual encoder with CLIP (ViT-B/32) [39]. The dimensions of the visual and textual representation feature are 512. We use the Adam optimizer [29] and set the batch size to 128. The initial learning rate is  $1e-4$  for MSR-VTT, MSVD, DiDeMo, and MS-COCO, and  $3e-4$  for ActivityNet Captions. For MSR-VTT and MSVD, the word length is 24 and the frame length is 12, and the network is optimized for 5 epochs. For ActivityNet Captions and DiDeMo, the word length is 64 and the frame length is 64, and the network is optimized for 10 epochs. For image datasets, i.e., MS-COCO, the network is optimized for 10 epochs.

## 5.2. Comparison with other Methods

**Compared Methods.** We compare the proposed NeighborRetr with two kinds of text-video retrieval methods: **i) non-CLIP methods:** T2VLAD [49], FROZEN [2]; and **ii) CLIP-based methods:** CLIP4Clip [37], X-Pool [18], TS2-Net [36], EMCL-Net [23], UATVR [14], DiCoSA [25], HBI [24], MPT [55]. We also compare NeighborRetr with text-image retrieval methods: ViLT [28], PVSE [42], PCME [8], SGRAF [10], CUSA [21], PCME++ [7].

**Retrieval Performance.** In Table. 1, we compare NeighborRetr with existing methods on the MSR-VTT dataset.

Methods	Text-to-Image			Image-to-Text		
	R@1↑	R-P↑	mAP@R↑	R@1↑	R-P↑	mAP@R↑
ViLT	82.7	50.0	41.7	73.4	38.6	27.5
PVSE	83.9	53.4	44.6	62.6	35.6	23.4
PCME	84.1	56.3	47.9	65.5	38.7	26.2
SGRAF	85.7	53.1	44.6	71.1	39.3	27.4
CUSA	88.2	55.8	47.6	80.9	44.1	33.6
PCME++	90.8	57.0	49.5	81.3	45.0	<b>34.4</b>
<b>Ours</b>	<b>92.1</b>	<b>57.7</b>	<b>49.8</b>	<b>82.1</b>	<b>45.6</b>	<b>34.4</b>

Table 3. Comparisons to others on the ECCV Caption datasets.

Methods	skew↓	trunc↓	atkinson↓	robin↓	anti↓	hub↓
X-Pool	5.14	1.24	0.71	0.70	0.51	0.80
EMCL-Net	4.67	1.23	0.71	0.70	0.53	0.78
DiCoSA	7.37	1.23	0.59	0.64	0.34	0.71
HBI	5.19	1.26	0.75	0.75	0.55	0.84
DiffusionRet	5.20	1.26	0.75	0.75	0.55	0.84
UCOFIA	7.10	1.40	0.92	0.90	0.78	0.94
Baseline†	5.97	1.30	0.81	0.79	0.63	0.80
<b>Ours</b>	<b>3.20</b>	<b>1.04</b>	<b>0.44</b>	<b>0.45</b>	<b>0.23</b>	<b>0.58</b>

Table 4. Comparing with other methods on the Text-to-Video task on MSR-VTT for hubness metrics. †the same two-level hierarchy model with only a vanilla contrastive loss and a KL loss.

NeighborRetr achieves the best performance across most metrics in both text-to-video (T2V) and video-to-text (V2T) retrieval tasks. Compared to the generative diffusion-based method DiffusionRet [26], NeighborRetr shows a clear improvement of 1.0% on V2T R@1 metric. When compared to the fine-grained interaction method HBI [24], which employs a three-level framework, NeighborRetr with only a two-level framework achieves 0.9% improvement on T2V R@1 and 1.9% improvement on V2T R@1. Table 2 compares NeighborRetr with other methods on the text-to-video retrieval task on MSVD, ActivityNet Captions, and DiDeMo datasets. On the short video dataset MSVD, NeighborRetr surpasses other CLIP-based methods by 1.3%. On long video datasets ActivityNet and DiDeMo, NeighborRetr consistently outperforms others, advancing the R@1 by 3.8% and 1.3%. Additionally, as shown in Table 3, NeighborRetr surpasses the previous method by 1.3% in R@1 for text-to-image retrieval and by 0.8% in image-to-text retrieval on the ECCV Caption dataset. More retrieval results are listed in the supplementary material.

**Hubness Mitigation.** Alongside its strong retrieval performance, NeighborRetr significantly mitigates the hubness problem as evidenced in Table 4. NeighborRetr reduces skew by 46% compared to the baseline, effectively shortening the long tail of the distribution of retrieval frequency. The 43% decrease in robin indicates a more balanced and equitable distribution. Additionally, the 28% reduction of

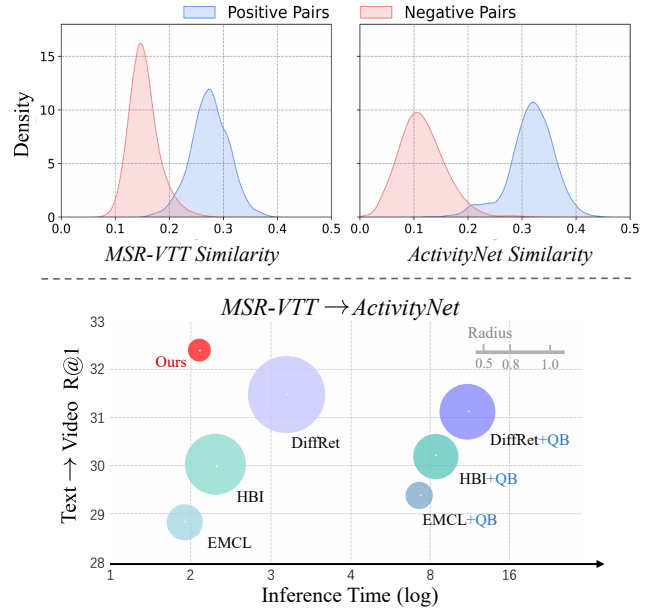


Figure 4. Top: Distributions of cross-modal similarity of MSR-VTT and ActivityNet. Bottom: Cross-domain adaptation on MSR-VTT to ActivityNet. The radius denotes hub occurrence [40].

	$\mathcal{L}_{Wti}$	$\mathcal{L}_{Opt}$	$\mathcal{L}_{Nbi}$	$\mathcal{L}_{KL}$	R@1↑	R@5↑	R@10↑	Rsum↑	MnR↓	anti↓	hub↓
✓	✗	✗	✗	✗	47.2	74.0	83.5	204.7	12.9	0.63	0.83
✓	✓	✗	✗	✗	48.0	73.8	83.7	205.5	12.5	0.32	0.76
✓	✗	✓	✗	✗	48.3	74.5	83.9	206.7	12.5	0.45	0.62
✓	✓	✗	✓	✓	48.4	73.9	83.7	206.0	<b>12.3</b>	0.28	0.74
✓	✓	✓	✓	✓	<b>49.5</b>	<b>74.1</b>	<b>84.1</b>	<b>207.7</b>	12.8	<b>0.23</b>	<b>0.52</b>

Table 5. Retrieval performance and hubness metric on the Text-to-Video task on MSR-VTT for all losses ablation.

hub demonstrates NeighborRetr’s ability to prevent over-centralized samples, enhancing retrieval fairness.

**Cross-domain Adaptation.** At the top of Fig. 4, we show the substantial domain gap between MSR-VTT (source) and ActivityNet (target), urging methods for robust cross-domain generalization. To evaluate this, we train NeighborRetr on the MSR-VTT and test it on the ActivityNet test set. As shown at the bottom of Fig. 4, our method achieves the lowest hub occurrence and the best retrieval performance on MSR-VTT to ActivityNet, indicating that addressing hubness during training benefits cross-domain adaptation. Moreover, we compare with QB-Norm [3], a test-time adjusting method that recalibrates retrieval scores by referencing a query bank from the training set. Due to the large distribution shift, QB-Norm can not effectively use training set priors to alleviate the hubness problem in the test domain, limiting its retrieval performance.

### 5.3. Ablation Studies

**Contribution of each loss of NeighborRetr.** In Table 5, we compare different combinations of loss functions:  $\mathcal{L}_{Wti}$ ,

$C_{\mathcal{L}_{Wti}}$	$C_{\mathcal{L}_{Nbi}}$	R@1↑	R@5↑	R@10↑	Rsum↑	MdR↓	MnR↓
intra	intra	48.3	73.7	83.4	205.4	2.0	<b>12.2</b>
inter	inter	47.9	73.8	83.7	205.4	2.0	13.0
inter	intra	47.7	73.2	83.2	204.1	2.0	12.5
<b>intra</b>	<b>inter</b>	<b>49.5</b>	<b>74.1</b>	<b>84.1</b>	<b>207.7</b>	<b>2.0</b>	12.8

Table 6. Retrieval performance on the Text-to-Video task on MSR-VTT for diverse centrality.

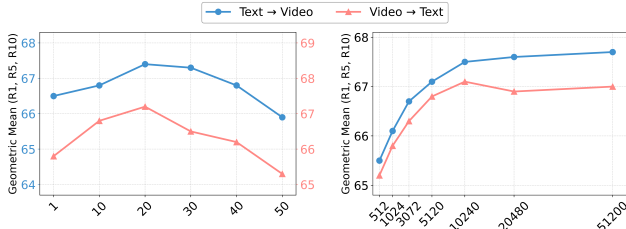


Figure 5. Retrieval performance on MSR-VTT: The y-axis shows the geometric mean of  $R@{1,5,10}$ . *Left*: Ablation study on the size of neighbors; the x-axis represents the size of neighbors. *Right*: Ablation study on the size of the Memory Bank during training; the x-axis represents the Memory Bank size.

$\mathcal{L}_{Opt}$ ,  $\mathcal{L}_{Nbi}$ , and  $\mathcal{L}_{KL}$ . The ablation reveals that incorporating  $\mathcal{L}_{Wti}$  (Row 1) improves Rsum to 204.7, boosting discriminative power by centralizing samples. Adding  $\mathcal{L}_{Opt}$  (Row 2) increases Rsum by 0.8%, with *anti* decreasing by 49% and *hub* decreasing by 8% compared to  $\mathcal{L}_{Wti}$  (Row 1), which in particular affects the uniform probability of anti-hub. The  $\mathcal{L}_{Nbi}$  loss increases Rsum by 2.0%, with *anti* decreasing by 29% and *hub* decreasing by 25% when paired with  $\mathcal{L}_{Wti}$  (Row 3), balancing good and bad hubs in neighborhood relations. Utilizing all these losses (Row 5) yields the best results, showcasing their complementary strengths:  $\mathcal{L}_{Wti}$  enhances central samples,  $\mathcal{L}_{Opt}$  realigns retrieval distribution, and  $\mathcal{L}_{Nbi}$  balances neighborhood relations.

**Intra/Inter-Modal Centrality for  $\mathcal{L}_{Wti}$  and  $\mathcal{L}_{Nbi}$ .** To validate the effectiveness of our different measurements of centrality for  $\mathcal{L}_{Wti}$  and  $\mathcal{L}_{Nbi}$ , we compare our methods trained with different choices of centralities  $C_{\mathcal{L}_{Wti}}$  and  $C_{\mathcal{L}_{Nbi}}$ . In Table. 6, we opt for measuring intra/inter-modal centrality ( $C$  or  $\hat{C}$ ) in  $\mathcal{L}_{Wti}$  and  $\mathcal{L}_{Nbi}$ . Results demonstrate that employing intra-modal centrality for  $\mathcal{L}_{Wti}$  and inter-modal centrality for  $\mathcal{L}_{Nbi}$  achieves the best performance compared to other measurement methods, with the final Rsum addition improvement by 3.6%. Integrating intra-modal centrality into  $\mathcal{L}_{Wti}$  preserves each modality’s structure, by promoting central samples as hubs, thereby capturing intrinsic representations. Meanwhile, applying inter-modal centrality to  $\mathcal{L}_{Nbi}$  leverages cross-modal embedding density, improving the discrimination of neighborhoods based on inter-modal centrality.

**Different Formulations of  $\mathcal{L}_{Nbi}$ .** To evaluate our neighborhood loss design (Simi-Cent), we ablate various formu-

Methods	R@1↑	R@5↑	R@10↑	Rsum↑	MdR↓	MnR↓
w/o $\mathcal{L}_{Nbi}$	48.4	73.9	83.7	206.0	<b>2.0</b>	12.3
Simi	48.6	74.1	83.8	206.5	<b>2.0</b>	<b>12.2</b>
Cent	48.2	74.0	83.8	206.0	<b>2.0</b>	12.4
Simi+Cent	48.0	73.4	83.6	205.0	<b>2.0</b>	12.8
<b>Simi-Cent</b>	<b>49.5</b>	<b>74.1</b>	<b>84.1</b>	<b>207.7</b>	<b>2.0</b>	12.8

Table 7. Retrieval performance on the Text-to-Video task on MSR-VTT for the neighborhood loss  $\mathcal{L}_{Nbi}$  ablation.

lations in Table 7. We compared the individual effects of cross-modal similarity (Simi) and centrality (Cent) scores, as well as their additive combination (Simi+Cent), against a baseline model without  $\mathcal{L}_{Nbi}$ . The results in Rows 2&3 show that using Simi or Cent alone provides minimal improvement over the baseline, indicating that adjusting rankings by either similarity or centrality alone is insufficient. However, the additive combination Simi+Cent yields only a slight boost in retrieval performance (Row 4), underscoring the limited effectiveness of this approach compared to subtraction. Our proposed Simi-Cent formulation, which subtracts centrality from similarity, achieves the best results (Row 5). This formulation effectively brings beneficial neighbors closer while distancing detrimental hubs, enhancing the embedding space and improving retrieval accuracy. **Size of Neighborhood and Memory Bank.** As illustrated on the left of Fig. 5, increasing the number of neighbors initially enhances retrieval performance, reaching an optimal value at 20 neighbors, beyond which performance declines. This suggests that incorporating too many neighbors may weaken the regularization due to sparse relevant examples. On the right of Fig. 5, we observe that performance steadily improves with larger memory bank sizes, gradually saturating beyond 10,240. Consequently, we set the memory bank size to 10,240 in our experiments.

## 6. Conclusions

In this paper, we introduced NeighborRetr, a method to systematically address the hubness problem in cross-modal retrieval by distinguishing between good hubs and bad hubs. By identifying hubs based on sample centrality, we perform centrality weighting and adaptively balance neighborhood relationships using a neighbor adjusting loss to promote good hubs and penalize bad ones. Additionally, we incorporate a uniform regularization loss enforcing uniform marginal constraints to treat anti-hubs equally. Experiments on various benchmarks demonstrate that NeighborRetr effectively mitigates hubness, resisting distribution shifting, and achieves state-of-the-art performance.

**Acknowledgements** This work was supported by Natural Science Foundation of China (No. 62302453) and Zhejiang Provincial Natural Science Foundation of China (No. LMS25F020003). Thank reviewer and Peng Jin for advice.

## References

- [1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017.
- [2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021.
- [3] Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross modal retrieval with querybank normalisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5194–5205, 2022.
- [4] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011.
- [5] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*, 2021.
- [6] Neil Chowdhury, Franklin Wang, Sumedh Shenoy, Douwe Kiela, Sarah Schwettmann, and Tristan Thrush. Nearest neighbor normalization improves multimodal retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22571–22582, 2024.
- [7] Sanghyuk Chun. Improved probabilistic image-text representations. *arXiv preprint arXiv:2305.18171*, 2023.
- [8] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8415–8424, 2021.
- [9] Sanghyuk Chun, Wonjae Kim, Song Park, Minsuk Chang, and Seong Joon Oh. Eccv caption: Correcting false negatives by collecting machine-and-human-verified image-caption associations for ms-coco. In *Proceedings of the European Conference on Computer Vision*, pages 1–19. Springer, 2022.
- [10] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1218–1226, 2021.
- [11] Cuturi M Sinkhorn Distances. Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26:2292–2300, 2013.
- [12] Mingjing Du, Shifei Ding, and Hongjie Jia. Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowledge-Based Systems*, 99:135–145, 2016.
- [13] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: improving visual-semantic embeddings with hard negatives. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 12. BMVA Press, 2018.
- [14] Bo Fang, Wenhao Wu, Chang Liu, Yu Zhou, Yuxin Song, Weiping Wang, Xiangbo Shu, Xiangyang Ji, and Jingdong Wang. Uatvr: Uncertainty-adaptive text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13723–13733, 2023.
- [15] Roman Feldbauer, Maximilian Leodolter, Claudia Plant, and Arthur Flexer. Fast approximate hubness reduction for large high-dimensional data. In *2018 IEEE International Conference on Big Knowledge (ICBK)*, pages 358–367. IEEE, 2018.
- [16] Thomas Fischer and Frederik Lundtofte. Unequal returns: Using the atkinson index to measure financial risk. *Journal of Banking & Finance*, 116:105819, 2020.
- [17] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *Proceedings of the European Conference on Computer Vision*, pages 214–229, 2020.
- [18] Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-pool: Cross-modal language-video attention for text-video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5006–5015, 2022.
- [19] Kazuo Hara, Ikumi Suzuki, Kei Kobayashi, Kenji Fukumizu, and Milos Radovanovic. Flattening the density gradient for eliminating spatial centrality to reduce hubness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016.
- [20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.
- [21] Hailang Huang, Zhijie Nie, Ziqiao Wang, and Ziyu Shang. Cross-modal and uni-modal soft-label alignment for image-text retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18298–18306, 2024.
- [22] Xiangru Jian and Yimu Wang. Invgc: Robust cross-modal retrieval by inverse graph convolution. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 836–865, 2023.

- [23] Peng Jin, Jinfa Huang, Fenglin Liu, Xian Wu, Shen Ge, Guoli Song, David A. Clifton, and Jie Chen. Expectation-maximization contrastive learning for compact video-and-language representations. In *NeurIPS*, pages 30291–30306, 2022.
- [24] Peng Jin, Jinfa Huang, Pengfei Xiong, Shangxuan Tian, Chang Liu, Xiangyang Ji, Li Yuan, and Jie Chen. Video-text as game players: Hierarchical banzhaf interaction for cross-modal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2472–2482, 2023.
- [25] Peng Jin, Hao Li, Zesen Cheng, Jinfa Huang, Zhennan Wang, Li Yuan, Chang Liu, and Jie Chen. Text-video retrieval with disentangled conceptualization and set-to-set alignment. *arXiv preprint arXiv:2305.12218*, 2023.
- [26] Peng Jin, Hao Li, Zesen Cheng, Kehan Li, Xiangyang Ji, Chang Liu, Li Yuan, and Jie Chen. Diffusionret: Generative text-video retrieval with diffusion model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2470–2481, 2023.
- [27] Dongwon Kim, Namyup Kim, and Suha Kwak. Improving cross-modal retrieval with set of diverse embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23422–23431, 2023.
- [28] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021.
- [29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [30] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.
- [31] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017.
- [32] Zheng Li, Caili Guo, Zerun Feng, Jenq-Neng Hwang, and Zhongtian Du. Integrating language guidance into image-text matching for correcting false negatives. *IEEE Transactions on Multimedia*, 2023.
- [33] 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 *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014.
- [34] Fangyu Liu and Rongtian Ye. A strong and robust baseline for text-image matching. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 169–176, 2019.
- [35] Fangyu Liu, Rongtian Ye, Xun Wang, and Shuaipeng Li. Hal: Improved text-image matching by mitigating visual semantic hubs. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11563–11571, 2020.
- [36] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. Ts2-net: Token shift and selection transformer for text-video retrieval. In *Proceedings of the European Conference on Computer Vision*, pages 319–335, 2022.
- [37] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022.
- [38] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *Proceedings of the European Conference on Computer Vision*, pages 681–699. Springer, 2020.
- [39] 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. In *International Conference on Machine Learning*, pages 8748–8763, 2021.
- [40] Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(sept):2487–2531, 2010.
- [41] Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Reverse nearest neighbors in unsupervised distance-based outlier detection. *IEEE transactions on knowledge and data engineering*, 27(5): 1369–1382, 2014.
- [42] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1979–1988, 2019.
- [43] Christopher Thomas and Adriana Kovashka. Emphasizing complementary samples for non-literal cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4632–4641, 2022.
- [44] Kaibin Tian, Yanhua Cheng, Yi Liu, Xinglin Hou, Quan Chen, and Han Li. Towards efficient and effective text-to-video retrieval with coarse-to-fine visual representation learning. In *Proceedings of the AAAI*

- Conference on Artificial Intelligence*, pages 5207–5214, 2024.
- [45] Nenad Tomasev. The role of hubness in high-dimensional data analysis. *Informatica (Slovenia)*, 38(4), 2014.
- [46] Nenad Tomasev, Miloa Radovanović, Dunja Mladenić, and Mirjana Ivanović. A probabilistic approach to nearest-neighbor classification: Naive hubness bayesian knn. In *Proceedings of the 20th ACM International Conference on Information and knowledge management*, pages 2173–2176, 2011.
- [47] Nenad Tomašev, Krisztian Buza, Kristóf Marussy, and Piroska B Kis. Hubness-aware classification, instance selection and feature construction: Survey and extensions to time-series. *Feature selection for data and pattern recognition*, pages 231–262, 2015.
- [48] Qiang Wang, Yanhao Zhang, Yun Zheng, Pan Pan, and Xian-Sheng Hua. Disentangled representation learning for text-video retrieval, 2022.
- [49] Xiaohan Wang, Linchao Zhu, and Yi Yang. T2vlad: Global-local sequence alignment for text-video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5079–5088, 2021.
- [50] Yimu Wang, Xiangru Jian, and Bo Xue. Balance act: Mitigating hubness in cross-modal retrieval with query and gallery banks. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.
- [51] Zheng Wang, Zhenwei Gao, Kangshuai Guo, Yang Yang, Xiaoming Wang, and Heng Tao Shen. Multi-lateral semantic relations modeling for image text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2830–2839, 2023.
- [52] Michael Wray, Hazel Doughty, and Dima Damen. On semantic similarity in video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3650–3660, 2021.
- [53] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.
- [54] 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.
- [55] Haonan Zhang, Pengpeng Zeng, Lianli Gao, Jingkuan Song, and Heng Tao Shen. Mpt: Multi-grained prompt tuning for text-video retrieval. In *ACM Multimedia 2024*, 2024.