NeighborRetr: Balancing Hub Centrality in Cross-Modal Retrieval Supplementary Material

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

Contents

A. Method Details

A.1. Hubness Observation

In Sec. 3.2, 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. Here, we proved further rationale for using the text encoder in the CLIP for probing. In [19], to calculate the relevance degree between unlabeled samples, they compare across several ways for false negative identification and find that intra-text similarities are the best way to discourse false negatives from true negatives. Especially, as shown in Fig.4 of [19], i) cross-modal comparison (text encoder in CLIP v.s. ViT in CLIP) can judge true negatives and false negatives but the two data distributions do not separate by a large margin. ii) Intra-visual comparison (ViT in CLIP v.s. ViT in CLIP) cannot accurately distinguish true negatives from false negatives. iii) Intra-text comparison (text encoder in CLIP v.s. text encoder in CLIP) has the ability to discriminate false negatives, as text is an abstraction of semantic content and the text similarity is closer to the semantic content similarity. In our work, we find that the distribution of intra-text comparison is well-aligned with the ground truth positive-negative annotations. In a broader context, especially without multi-correlations annotations, intra-text similarity can be used as a probe of semantic alignment.

A.2. Proof of Gradient

Gradient of $\mathcal{L}_{\text{Nbi}}$. The gradient of the Neighbor Adjusting Loss $\mathcal{L}_{\text{Nbi}}(x_i)$ with respect to the similarity score $\mathrm{S}(x_i, y_j)$ is derived as follows:

$$
\begin{aligned}
&\frac{\partial \mathcal{L}_{\text{Nbi}}(x_i)}{\partial \mathrm{S}(x_i, y_j)} \\
&= \frac{\partial}{\partial \mathrm{S}(x_i, y_j)} \left(- \sum_{y_k \in \mathcal{N}^+(x_i)} \mathcal{H}(y_k) \log P(y_k \mid \mathcal{N}^+(x_i)) \right) \\
&= - \sum_{y_k \in \mathcal{N}^+(x_i)} \mathcal{H}(y_k) \frac{\partial}{\partial \mathrm{S}(x_i, y_j)} \log P(y_k \mid \mathcal{N}^+(x_i)) \\
&= - \sum_{y_k \in \mathcal{N}^+(x_i)} \mathcal{H}(y_k) \left(\delta_{kj} - P(y_j \mid \mathcal{N}^+(x_i)) \right) \\
&= -\mathcal{H}(y_j)(1 - P(y_j \mid \mathcal{N}^+(x_i))) \\
&\quad + P(y_j \mid \mathcal{N}^+(x_i)) \sum_{y_k \in \mathcal{N}^+(x_i)} \mathcal{H}(y_k) \\
&= -\mathcal{H}(y_j) + P(y_j \mid \mathcal{N}^+(x_i)) \sum_{y_k \in \mathcal{N}^+(x_i)} \mathcal{H}(y_k) \\
&= P(y_j \mid \mathcal{N}^+(x_i)) - \mathcal{H}(y_j).
\end{aligned}
$$

The third step involves recognizing that the derivative of $\log P(y_k \mid \mathcal{N}^+(x_i))$ with respect to $\mathrm{S}(x_i, y_j)$ is $\delta_{kj} -$

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

$P(y_j \mid \mathcal{N}^+(x_i))$, where δ_{kj} is the Kronecker delta function. The derivation reveals that the gradient equals $P(y_j \mid \mathcal{N}^+(x_i)) - \mathcal{H}(y_j)$, indicating that it is the difference between the predicted probability and the adjusted target. This guides the update of similarity scores $\mathrm{S}(x_i, y_j)$: a positive gradient, occurring when the predicted probability $P(y_j \mid \mathcal{N}^+(x_i))$ exceeds the adjusted target $\mathcal{H}(y_j)$, decreases the similarity score to penalize bad hubs, while a negative gradient increases it to promote good hubs, thereby balancing the representation space.

Gradient of $\mathcal{L}_{\text{Opt}}$. Derive the gradient of the Uniformity Regularization loss $\mathcal{L}_{\text{Opt}}$ with respect to the similarity score $\mathbf{S}_{i,j}$ for query i and gallery j.

$$
\begin{aligned}
&\frac{\partial \mathcal{L}_{\text{Opt}}}{\partial \mathbf{S}_{i,j}} \\
&= \frac{\partial}{\partial \mathbf{S}_{i,j}} \left(-\frac{1}{n} \sum_{i'=1}^{n} \sum_{l=1}^{m} \mathbf{Q}_{i',l} \log P(l|i') \right) \\
&= -\frac{1}{n} \frac{\partial}{\partial \mathbf{S}_{i,j}} \left(\sum_{i'=1}^{n} \sum_{l=1}^{m} \mathbf{Q}_{i',l} \log P(l|i') \right) \\
&= -\frac{1}{n} \left(\sum_{l=1}^{m} \mathbf{Q}_{i,l} \frac{\partial}{\partial \mathbf{S}_{i,j}} \log P(l|i) \right) \\
&= -\frac{1}{n} \left(\sum_{l=1}^{m} \mathbf{Q}_{i,l} \left(\delta_{j,l} - P(j|i) \right) \right) \\
&= -\frac{1}{n} \left(\mathbf{Q}_{i,j} - P(j|i) \sum_{l=1}^{m} \mathbf{Q}_{i,l} \right) \\
&= -\frac{1}{n} \left(\mathbf{Q}_{i,j} - P(j|i) \right).
\end{aligned}
$$

The fourth step involves recognizing that the derivative of $\log P(l|i)$ with respect to $\mathbf{S}_{i,j}$ is $\delta_{j,l} - P(j|i)$, where $\delta_{j,l}$ is the Kronecker delta function. Since $\mathbf{Q}i, j$ is designed to uniform probability retrieval, the gradient encourages the model to adjust $P(j|i)$ towards a uniform distribution over all possible retrievals. By minimizing the loss, the optimization process reduces this discrepancy, effectively guiding the predicted probabilities to match the uniform target. This mechanism ensures that the model will not favor hub or ignore anti-hub.

A.3. Uniform Regularization

In Sec. 4.4, we utilize a uniform marginal constraint inspired by Sinkhorn Distances [5] to enforces equal retrieval probabilities for overall matching at the high level.

We start by deriving high-level representations for each modality using the Deep Projection Clustering with K-Nearest Neighbors (DPC-KNN) [7] module. To learn high semantic alignment, high-level representations are denoted as $\mathbf{V_g} = \{v_g^i\}_{i=1}^N$ for the visual modality and $\mathbf{T_g} = \{t_g^j\}_{j=1}^N$ for the textual modality, where N represent the lengths of high-level representation. The high similarity between a visual high-level representation v_g^i and a textual high-level representation t_g^j is then computed using a cosine similarity metric. The high alignment matrix G is defined as $G = [g_{ij}]^{N \times N}$, where:

$$
g_{ij} = \frac{{v_g^i}^\top t_g^j}{||v_g^i||\ ||t_g^j||}, \tag{1}
$$

which is the high similarity score between the i-th visual high representation and the j-th textual high representation.

Uniform Retrieving Matrix $\mathbf{Q}^*$ has a simple normalized exponential matrix solution by Sinkhorn fixed point iterations [5]. By extensively reordering relationships, optimal transport aligns semantically related pairs more closely, balancing the centrality of hubs and anti-hubs. To further enhance robustness, realigned targets $\mathbf{Q}$ are incorporated:

$$
\mathbf{Q} = (1 - \beta)\mathbf{I}_B + \beta \mathbf{Q}^*. \tag{2}
$$

Here, $\mathbf{I}_B$ is the identity matrix of size B, where B represents the batch size, and $\beta \in [0, 1]$ controls the balance between the identity matrix and the uniform retrieving matrix $\mathbf{Q}^*$. By redefining $\mathbf{Q}$ in this manner, each query or gallery is realigned based on uniform regularization, thereby enforcing balanced retrieval probabilities across all samples.

B. More Experiments

B.1. Experimental Settings

Datasets. We also evaluate our method on two text-image retrieval benchmarks: Flickr30K and MS-COCO. **Flickr30K** [31] contains 31K images, each image is annotated with 5 sentences. Following the data split of [18], we use 1K images for validation, 1K images for testing, and the remaining 29K for training. **MS-COCO** [20] contains 123K images, and each image comes with 5 sentences. We mirror the data split setting of [8]. We use 113K images for training, 5K images for validation, and 5K images for testing.

Implementation Details. For Flickr30K and MS-COCO, we utilize the CLIP (ViT-B/16) [26] as the pre-trained model. The dimensions of the visual and textual representation feature are 512. We use the Adam optimizer [17] and set the batch size to 128. The initial learning rate is 1e-4 for Flickr30K, and MS-COCO. For Flickr30K, the network is optimized for 5 epochs. For MS-COCO, the network is optimized for 10 epochs.

B.2. Comparison with other Methods

Compared Methods. We compare NeighborRetr with other text-image retrieval methods: UNITER [2],

Model	Flickr30K (1K Test Set)						
	Text-to-Image			Image-to-Text			Rsum↑
	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑	
CSIC	88.5	98.1	99.4	75.3	93.6	96.7	551.6
IAIS	88.3	98.4	99.4	76.9	93.3	95.7	552.0
ViSTA	89.5	98.4	99.6	75.8	94.2	96.9	554.4
ViLEM	92.4	99.2	99.7	78.1	94.6	97.0	561.0
I-CLIP	91.2	99.2	99.7	74.3	92.3	95.5	552.2
CUSA	90.8	99.1	99.7	77.4	95.5	97.7	560.2
Ours	**92.7**	**99.3**	**99.8**	**79.5**	**95.7**	**97.8**	**564.8**

Model	MSCOCO (5K Test Set)						
	Text-to-Image			Image-to-Text			Rsum↑
	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑	
UNITER	63.3	87.0	93.1	48.4	76.7	85.9	454.4
ViSTA	63.9	87.8	93.6	47.8	75.8	84.5	453.4
SOHO	66.4	88.2	93.8	50.6	78.0	86.7	463.7
PCME	65.3	89.2	94.5	51.2	79.1	87.5	474.2
ViLEM	69.0	**90.7**	95.1	52.6	79.4	87.2	474.0
CUSA	67.9	90.3	94.7	52.4	**79.8**	88.1	473.2
Ours	**69.5**	90.5	**95.3**	**53.2**	79.7	**88.2**	**476.4**

Table A. Experimental results of image-text retrieval on MSCOCO and Flickr30K datasets. "↑" denotes higher is better.

MSVD					
Methods	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
EMCL-Net	54.3	81.3	88.1	**1.0**	5.6
CLIP4Clip	62.0	87.3	92.6	**1.0**	4.5
X-Clip	60.9	87.8	-	-	4.7
Diffusion	61.9	88.3	92.9	**1.0**	4.5
Ours	**63.3**	**89.6**	**95.4**	**1.0**	**3.3**

ActivityNet Captions					
Methods	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
CLIP4Clip	41.4	73.7	85.3	**2.0**	6.7
EMCL-Net	42.7	74.0	-	**2.0**	-
ECLIPSE	42.3	73.2	83.8	-	8.2
HBI	42.2	73.0	86.0	**2.0**	6.5
Ours	**43.5**	**74.6**	**86.7**	**2.0**	**6.1**

DiDeMo					
Methods	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
X-Clip	43.1	77.2	-	-	10.9
EMCL-Net	45.7	74.3	82.7	**2.0**	10.9
Diffusion	46.2	74.3	82.2	**2.0**	10.7
HBI	46.2	73.0	82.7	**2.0**	**8.7**
Ours	**48.4**	**74.6**	**83.7**	**2.0**	9.0

Table B. Comparisons to other methods on the Video-to-Text task on the MSVD, ActivityNet Captions, and DiDeMo datasets. "↑" denotes higher is better, "↓" denotes lower is better.

SOHO [12], PCME [4], IAIS [29], ViLT [16], CSIC [22], ViSTA [3], ViLEM [1], I-CLIP [6], CUSA [11]. We also compare NeighborRetr with other methods on the video-to-text retrieval task: CLIP4Clip [23], X-Clip [25],ECLIPSE [21] EMCL-Net [13], HBI [14], Diffusion [15].

Image-Text Retrieval. Table A presents a comprehensive comparison of our method against state-of-the-art approaches on the Flickr30K and MSCOCO datasets for both text-to-image and image-to-text retrieval tasks. On the Flickr30K, compared to the recent reinforcement learning-enhanced method I-CLIP [6], our approach shows a clear 1.5% improvement on the text-to-image R@1 metric and a 5.2% improvement on image-to-text R@1, even with I-CLIP's additional steps to boost cross-modal alignment. On the MSCOCO dataset, compared to the soft-label supervision strategy of CUSA [11], which focuses on enhancing similarity recognition through uni-modal samples, our approach achieves a notable 2.1% improvement on image-to-text R@1.

Video-to-Text Retrieval. We compare the proposed NeighborRetr method with other state-of-the-art methods on the Video-to-Text Retrieval task using the MSVD, DiDeMo, and ActivityNet Captions datasets, as shown in Table B. The results demonstrate that NeighborRetr outperforms the competing methods across all datasets. Specifically, in the MSVD dataset, NeighborRetr improves the R@1 score by 1.3%. For the DiDeMo dataset, NeighborRetr achieves an R@1 improvement of 2.2% over HBI, indicating a significant advancement in the retrieval accuracy of long videos. These results consistently highlight the effectiveness of NeighborRetr in handling video-to-text retrieval tasks ranging from short clips to long videos.

MSR-VTT					
Methods	R@1↑	R@5↑	R@10↑	Rsum↑	MnR↓
CLIP4Clip	44.5	71.4	81.6	197.5	15.3
w/ NeighborRetr	**46.4**	**73.0**	**82.8**	**202.2**	**13.1**

Table C. Comparison of retrieval performance on the Text-to-Video task on MSR-VTT, demonstrating improved efficacy under the CLIP4Clip framework with our proposed NeighborRetr method.

B.3. Ablation Studies

CLIP4Clip w/ NeighborRetr. Integrating NeighborRetr into the CLIP4Clip [24] framework yields significant performance improvements, as evidenced in Table C. Specifically, the R@1 score increases by 1.9%, the Rsum score improves by 4.7%, and MnR decreases by 2.2%. These metrics underscore the adaptability and robustness of NeighborRetr, demonstrating its potential for enhancing retrieval performance across different backbone architectures.

Hubness ablation of each loss. Table D demonstrates the effect of each loss function on hubness mitigation. The baseline model, without the proposed losses, exhibits

$\mathcal{L}_{Wti}$	$\mathcal{L}_{Opt}$	$\mathcal{L}_{Nbi}$	$\mathcal{L}_{KL}$	skew↓	trunc↓	atkinson↓	robin↓	anti↓	hub↓
✓	✗	✗	✗	5.37	1.3	0.73	0.76	0.63	0.83
✓	✓	✗	✗	4.87	1.27	0.65	0.68	0.32	0.76
✓	✗	✓	✗	4.05	1.16	0.55	0.51	0.45	0.62
✓	✓	✓	✓	**3.20**	**1.04**	**0.44**	**0.45**	**0.23**	**0.58**

Table D. Hubness metrics in the Text-to-Video retrieval task on the MSR-VTT dataset for all losses ablation. "↓" denotes lower is better.

Methods	skew↓	trunc↓	atkinson↓	robin↓	anti↓	hub↓
w/o $\mathcal{L}_{Nbi}$	5.14	1.24	0.71	0.70	0.51	0.80
Simi	4.47	1.17	0.50	0.66	0.37	0.72
Cent	4.76	1.21	0.62	0.64	0.40	0.76
Simi+Cent	4.12	1.15	0.54	0.60	0.30	0.70
Simi-Cent	**3.20**	**1.04**	**0.44**	**0.45**	**0.23**	**0.58**

Table E. Hubness on the Text-to-Video task on MSR-VTT for $\mathcal{L}_{Nbi}$ loss ablation. "↓" denotes lower is better.

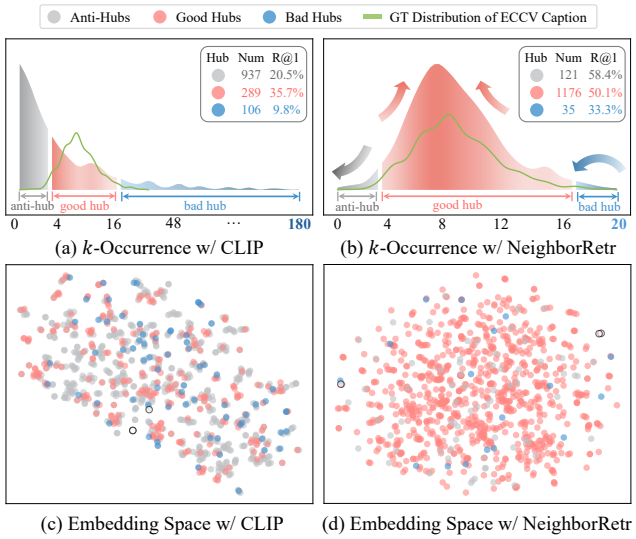
the highest hubness values, indicating severe neighborhood relationship issues. Introducing $\mathcal{L}_{Wti}$ reduces *skew* by 10%, boosting discriminative power by centralizing samples. Adding $\mathcal{L}_{Opt}$ significantly reduces hubness indicators, with *anti* decreasing by 49%, in particular affects the uniform probability of anti-hub. Finally, including $\mathcal{L}_{Nbi}$ alongside the other losses yields the most substantial reductions: *skew* drops by 17%, and *hub* by 25%. These results indicate that $\mathcal{L}_{Nbi}$ is crucial in refining neighbor relationships and minimizing hub formation. Overall, the combined losses provide the most effective solution, balancing and diversifying neighbor interactions.

Hubness ablation for formulations of $\mathcal{L}_{Nbi}$. In Table E, we analyze different formulations of cross-modal similarity (Simi) and centrality (Cent) for the hubness problem. The results show that incorporating Simi alone results in a modest skewness reduction of 13%, while Cent alone slightly decreases skewness by 7%, indicating the limitation of these formulations individually. The combined Simi+Cent formulation yields a certain decrease in skewness but remains suboptimal. However, the Simi-Cent formulation outperforms all other formulations, achieving a 38% reduction in skewness, which highlights its effectiveness in reducing hubness across all metrics by better aligning the embedding space and fine-tuning more equitable neighborhood relations. This substantial improvement underscores the importance of the subtractive formulation in mitigating high-dimensional data biases.

Test-Time Hubness Mitigation Analysis. Table F demonstrates that NeighborRetr achieves superior retrieval performance over the baseline while maintaining computational efficiency. For training, NeighborRetr employs an online queue to maintain updated representations, mitigating staleness and removing outdated hubs. In inference, NeighborRetr is memorybank-free. Thus in Table F, NeighborRetr achieves 300× speedup than the QB-Norm method, which adopts a memory-bank as a test-time similarity normalization strategy. Notably, NeighborRetr can improve R@1 by 2.7% and R@Sum by 4.6% with minimal inference time impact with the simplest similarity measure (Simi). The Simi-Cent variant enhances R@Sum by 0.2% using only 1k validation samples compared to Simi. NeighborRetr effectively decouples representation quality from hubness bias during training, making test-time similarity adjustments redundant. This may explain why the test-time Simi-Cent metric does not translate to practical gains in NeighborRetr. In contrast, QB-Norm methods require the entire training set (10k samples) and incur prohibitive inference costs. While both methods address the hubness problem, NeighborRetr operates during training for distribution-robust representations, whereas QB-Norm applies post-hoc similarity adjustments during

Method (Metric)	MB (Source/Size)	R@1↑	R@5↑	R@10↑	R@Sum↑	Inf. Time↓
Baseline (Simi)	-	46.8	73.6	82.7	203.1	**19.30**
Ours (Simi)	-	**49.5**	74.1	84.1	207.7	20.88
Ours (Simi-Cent)	Val (1k)	49.1	**74.5**	**84.3**	**207.9**	23.37
Baseline (+QB-Norm)	Train (10k)	47.7	74.0	83.7	205.4	5592.15
Ours (+QB-Norm)	Train (10k)	48.6	73.8	82.8	205.2	5950.97

Table F. Comparison of different similarity measures for Text-to-Video retrieval on MSR-VTT during inference, showing retrieval performance and inference time. "↑" denotes higher is better, "↓" denotes lower is better.



(a) k-Occurrence w/ CLIP (b) k-Occurrence w/ NeighborRetr

(c) Embedding Space w/ CLIP (d) Embedding Space w/ NeighborRetr

Figure A. Visualization of hub frequency distribution and embedding space comparing CLIP vs. NeighborRetr, illustrating three hub types: **anti-hub** ($k < 4$)**, good hub** ($4 \leq k \leq 16$)**, bad hub** ($k > 16$). NeighborRetr shows improved hub distribution and embedding characteristics.

inference, making their combination counterproductive (-2.5% R@Sum) due to conflicting hubness mitigation strategies.

B.4. Hub Distribution & Embedding Space Analysis

Fig. A illustrates hub distribution characteristics of CLIP versus NeighborRetr. We categorize hubs into three disjoint subsets: anti-hubs ($k < 4$), good hubs ($4 \leq k \leq 16$), and bad hubs ($k > 16$). NeighborRetr achieves a more balanced k-occurrence distribution, aligning more closely with the actual data distribution. reducing isolated anti-hubs from 937 to 121, yielding a 37.9% R@1 gain, and decreasing dominant bad hubs by 71, leading to a 23.5% improvement. The embedding space shown in Fig. A(d) demonstrates a more uniform distribution, with good hubs gaining a 14.4% improvement. This result confirms NeighborRetr effectively rebalances neighbor relationships by drawing relevant points closer while preventing irrelevant hubs from dominating neighborhoods.

B.5. Visualization of Text-to-Video Retrieval

Figure. B demonstrates the effectiveness of our method in ranking videos that are highly relevant to the query while encompassing a range of semantically related scenarios. Higher-ranked videos show larger gaps between similarity (*Simi*) and centrality (*Cent*) scores, indicating the model's ability to prioritize less central samples, thereby reducing potential bias towards over-represented data. The lower-ranked videos showcase diversity, reflecting the model's adaptability to various semantic contexts and its capability to match different topics within the same activity. As centrality scores increase with higher rankings, it becomes evident that our method not only captures good neighborhood relationships but also improves them dynamically, enhancing the overall retrieval quality.

C. Hubness Metric Details

Definition of k-Occurrence. In high-dimensional embedding spaces prevalent in cross-modal retrieval tasks, the metric of hubness manifests through certain samples' recurrent appearance as nearest neighbors. We quantify this phenomenon using the k-occurrence metric [30], $N_k(\mathbf{x})$, which indicates the frequency a sample $\mathbf{x}$ appears among the k-nearest neighbors of other samples:

$$N_k(\mathbf{x}) = \sum_{i=1}^{n} p_{i,k}(\mathbf{x}), \tag{3}$$

where $p_{i,k}(\mathbf{x})$ is a binary indicator function, defined as:

$$p_{i,k}(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \text{ is } k \text{ nearest neighbors of } q_i, \\ 0 & \text{otherwise.} \end{cases}$$

This metric provides insights into the biases and centrality of samples within the embedding space, facilitating the analysis of hub formation and its impact on retrieval effectiveness.

C.1. Distribution-based Metrics

Skewness. We measure the asymmetry of the k-occurrence distribution using skewness [27], denoted by S_{N_k}. It is calculated as:

$$S_{N_k} = \frac{E\left(N_k - \mu_{N_k}\right)^3}{\sigma_{N_k}^3}, \tag{4}$$

where μ_{N_k} and σ_{N_k} are the mean and standard deviation of the k-occurrences, respectively. A high skewness value indicates a distribution with pronounced tails, suggesting the presence of hub samples that disproportionately appear as nearest neighbors, exacerbating the hubness issue.

Skewness Truncated Normal. The Skewness Truncated Normal [30], $S_{N_k}^{trunc}$, addresses skewness by truncating the k-occurrence distribution to enhance accuracy in central tendency analysis:

$$S_{N_k}^{trunc} = \frac{E\left[(N_k^{trunc} - \mu_{N_k^{trunc}})^3\right]}{(\sigma_{N_k^{trunc}})^3}, \tag{5}$$

where N_k^{trunc} represents the adjusted k-occurrences, with $\mu_{N_k^{trunc}}$ and $\sigma_{N_k^{trunc}}$ detailing the mean and standard deviation of this truncated distribution, sharpening the focus on typical data behavior by excluding outliers.

Atkinson Index. The Atkinson Index [10], A_{N_k}, specifically measures the extent of inequality in k-occurrence distributions with a unique focus on the tails. This focus is modulated by the ϵ parameter, which provides a tunable sensitivity:

$$A_{N_k} = 1 - \frac{E[(N_k(x_i))^{1-\epsilon}]^{\frac{1}{1-\epsilon}}}{\mu_{N_k}}, \tag{6}$$

where ϵ enhances the index's responsiveness to the presence of hubs and antihubs, crucial for nuanced analysis in cross-modal retrieval tasks. This adaptability allows for a tailored examination of how extremes centrality overall system fairness and data representation equity.

Robin Hood Index. The Robin Hood Index [9], also referred to as the Hoover index, measures the inequality of the k-occurrence distribution. It reflects the extent of redistribution required to achieve uniformity in neighbor assignments across the dataset D:

$$\mathcal{H}^k = \frac{1}{2} \cdot \frac{E\left|N_k - \mu_{N_k}\right|}{\sum_{x \in D} N_k(x)}, \tag{7}$$

a higher index value signals greater inequality, which is indicative of significant disparities in the distribution of nearest neighbors, thereby highlighting potential issues in the fairness of retrieval system distributions.

Figure B. Examples of text-to-video retrieval results with associated similarity (Simi) and centrality (Cent) scores. The ground truth video is highlighted in red. Each candidate video is assessed to demonstrate accurate text-to-video matching. For neighborhoods in Rank 2 to 5, NeighborRetr effectively identifies good neighbors, highlighting the effectiveness of our approach.

C.2. Occurrence-based Metrics

Antihub Occurrence. Antihub [28], A_{anti}, identifies the proportion of samples that are not recognized as nearest neighbors by any other objects in the dataset:

$$A_{anti} = E[\mathbf{1}_{\{N_k(x_i)=0\}}], \tag{8}$$

where $N_k(x_i)$ is the k-occurrence of sample x_i, and $\mathbf{1}_{\{\cdot\}}$ indicates whether x_i is completely ignored, emphasizing exclusion in the embedding space and spotlighting representational biases.

Hub Occurrence. Hub [27], A_{hub}, measures the prevalence of certain samples overly dominating nearest neighbor selections, affecting data representation equity:

$$A_{hub} = E[N_k(x_i) \cdot \mathbf{1}_{\{N_k(x_i) > k \cdot hub\_size\}}]. \tag{9}$$

where hub_size sets the threshold defining a hub if it appears frequently as a nearest neighbor. This metric, by quantifying the proportion of nearest neighbor slots occupied by hubs, highlights the potential skew in data representation.

References

[1] Yuxin Chen, Zongyang Ma, Ziqi Zhang, Zhongang Qi, Chunfeng Yuan, Ying Shan, Bing Li, Weiming Hu, Xiaohu Qie, and JianPing Wu. Vilem: Visual-language error modeling for image-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11018–11027, 2023.

[2] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[3] Mengjun Cheng, Yipeng Sun, Longchao Wang, Xiongwei Zhu, Kun Yao, Jie Chen, Guoli Song, Junyu Han, Jingtuo Liu, Errui Ding, et al. Vista: Vision and scene text aggregation for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5184–5193, 2022.

[4] 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.

[5] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[6] Xinfeng Dong, Longfei Han, Dingwen Zhang, Li Liu, Junwei Han, and Huaxiang Zhang. Giving text more imagination space for image-text matching. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6359–6368, 2023.

[7] 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.

[8] 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.

[9] 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.

[10] Thomas Fischer and Frederik Lundtofte. Unequal returns:

Using the atkinson index to measure financial risk. *Journal of Banking & Finance*, 116:105819, 2020.

[11] 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.

[12] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12976–12985, 2021.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision*, pages 201–216, 2018.

[19] 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.

[20] 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.

[21] Yan-Bo Lin, Jie Lei, Mohit Bansal, and Gedas Bertasius. Eclipse: Efficient long-range video retrieval using sight and sound. In *European Conference on Computer Vision*, pages 413–430. Springer, 2022.

[22] Zejun Liu, Fanglin Chen, Jun Xu, Wenjie Pei, and Guangming Lu. Image-text retrieval with cross-modal semantic importance consistency. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(5):2465–2476, 2022.

[23] 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.

[24] 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.

[25] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 638–647, 2022.

[26] 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.

[27] 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.

[28] 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.

[29] Shuhuai Ren, Junyang Lin, Guangxiang Zhao, Rui Men, An Yang, Jingren Zhou, Xu Sun, and Hongxia Yang. Learning relation alignment for calibrated cross-modal retrieval. *arXiv preprint arXiv:2105.13868*, 2021.

[30] Nenad Tomasev. The role of hubness in high-dimensional data analysis. *Informatica (Slovenia)*, 38(4), 2014.

[31] 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.