DEMO: A Statistical Perspective for Efficient Image-Text Matching

Fan Zhang[1], **Xian-Sheng Hua**[2], **Chong Chen**[2], **Xiao Luo**[3†]
[1]Georgia Tech Shenzhen Institute, Tianjin University (GTSI)
[2]Terminus Group [3]University of California, Los Angeles
fanzhang@gatech.edu, huaxiansheng@gmail.com, chenchong.cz@gmail.com, xiaoluo@cs.ucla.edu

Abstract

Image-text matching has been a long-standing problem, which seeks to connect vision and language through semantic understanding. Due to the capability to manage large-scale raw data, unsupervised hashing-based approaches have gained prominence recently. They typically construct a semantic similarity structure using the natural distance, which subsequently provides guidance to the model optimization process. However, the similarity structure could be biased at the boundaries of semantic distributions, causing error accumulation during sequential optimization. To tackle this, we introduce a novel hashing approach termed Distribution-based Structure Mining with Consistency Learning (DEMO) for efficient image-text matching. From a statistical view, DEMO characterizes each image using multiple augmented views, which are considered as samples drawn from its intrinsic semantic distribution. Then, we employ a non-parametric distribution divergence to ensure a robust and precise similarity structure. In addition, we introduce collaborative consistency learning which not only preserves the similarity structure in the Hamming space but also encourages consistency between retrieval distribution from different directions in a self-supervised manner. Through extensive experiments on three benchmark image-text matching datasets, we demonstrate that DEMO achieves superior performance compared with many state-of-the-art methods.

1 Introduction

Image-text matching (Sun et al., 2023; Zhang et al., 2022b; Huang et al., 2022; Liu and Ye, 2019; Hu et al., 2023b) is a pivotal task in both computer vision and natural language processing, which bridges data across heterogeneous modalities. The objective is to return images correlated with a given textual description and detect texts corresponding to a given image. Considering explosively growing web data (Krotov and Johnson, 2023), there is a significant demand for an efficient approach that can select a small candidate set from a comprehensive dataset. Towards this end, hashing has become prevalent in information retrieval (Luo et al., 2021a), particularly image-text matching (Hu et al., 2023a; Sun et al., 2022a; Tu et al., 2023a; Zeng et al., 2023; Cao et al., 2022), which involves mapping both texts and images into a shared binary space (Hamming space), and then determining cross-modal similarity scores by comparing their binary codes.

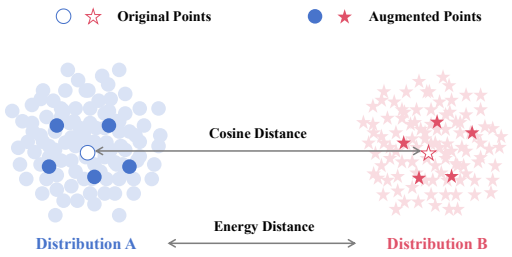

Figure 1: Comparison between cosine distance and energy distance. We leverage the randomness of data augmentation to estimate the latent semantics distributions, and then use energy distance between distributions as a substitute for cosine distance between data points.

In literature, numerous approaches have been developed for cross-modal hashing (Jiang and Li, 2017; Kaur et al., 2021), which can broadly be categorized into supervised and unsupervised methods. Supervised methods (Chen et al., 2019; Jia et al., 2021; Gu et al., 2019) typically incorporate ground truth similarities into a pairwise (Fan et al., 2023) or rankwise (Liu et al., 2023) loss objective. However, due to the high costs associated with label annotation, unsupervised approaches (Tu et al., 2023a; Zeng et al., 2023; Cao et al., 2022) tend to be more appreciated in real-world applications. Unsuper-

† Corresponding author.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 355–369

vised cross-modal hashing approaches typically begin by reconstructing the similarity structure between different modalities, which subsequently provides guidance during the learning process of the hashing model.

Despite the notable advancements, prevailing unsupervised cross-modal hashing approaches (Tu et al., 2023a; Zeng et al., 2023; Cao et al., 2022) still suffer from two major limitations: (1) *Biased Similarity Structure.* These approaches typically employ natural distances (e.g., cosine distance) to generate the semantic similarity structure. Since deep features with the same semantics should be from a high-dimensional distribution, utilizing cosine distance would be imprecise at the distribution boundaries, which generates noisy supervision, and serious error accumulation during subsequent optimization procedures. (2) *Distribution Discrepancy Across Modalities.* Given the inherent heterogeneity, different networks are utilized to generate binary codes, which could obey distinct distributions in the Hamming space. This distribution discrepancy inherently undermines the effectiveness of cross-modal retrieval and brings suboptimal results.

To handle these limitations, in this work, we propose a new hashing approach named Distribution-based Structure Mining with Consistency Learning (DEMO) for efficient image-text matching. The core of our DEMO revolves around exploring the latent semantic distribution of each sample using multiple random augmentations. In particular, given that data augmentation generally maintains the semantics (Dai et al., 2023), we consider each augmented view of an image as samples drawn from its intrinsic semantic distribution. Then a non-parametric metric (i.e., energy distance (Rizzo and Székely, 2016)) is incorporated to precisely measure the distribution divergence (see Figure 1), thereby reconstructing a robust and accurate semantic structure. The subsequent optimization of the hashing network is achieved by preserving this semantic structure in the Hamming space. Furthermore, to diminish the distribution shift across modalities, we generate cross-modal retrieval distributions given both queries of images and texts and their consistency are promoted in a self-supervised manner. In addition, we employ a sharpening operation to refine retrieval results by emphasizing points with high degrees of similarity. We conduct comprehensive experiments on three benchmark image-text matching datasets, and the results show that our DEMO outperforms a wide range of competing methods. In brief, the main contribution of this paper can be summarized as follows:

- *Innovative Perspective.* We explore the latent semantics distribution and adopt the distribution divergence to construct a robust and accurate semantics structure to guide unsupervised cross-modal hashing through a statistical perspective.
- *Coherent Framework.* DEMO optimizes the modality-specific hashing networks by preserving the semantics structure in the Hamming space. Additionally, DEMO promotes consistency between cross-modal retrieval distributions, resulting in modality-invariant binary descriptors.
- *Outstanding Performance.* Comprehensive experiments reveal that DEMO outperforms various state-of-the-art hashing-based methods on image-text matching benchmark datasets.

2 Related Work

2.1 Image-text Matching

Image-text matching is a fundamental problem which can bridge computer vision and natural language processing (Sun et al., 2023; Zhang et al., 2022b; Huang et al., 2022; Liu and Ye, 2019; Hu et al., 2023b). Recent approaches can be divided into local-level and global-level approaches. Local-level matching approaches (Liu et al., 2019a; Chen et al., 2020; Zhang et al., 2022a; Dong et al., 2022; Fu et al., 2023; Bhattacharyya et al., 2022) take the input of image-text pairs to learn fine-grained relationships, such as region-word alignments. In contrast, global-level matching approaches (Tu et al., 2021; Lu et al., 2022; Radford et al., 2019; Jia et al., 2021) map both images and texts into a shared space and then calculate their latent embedding similarities. To enhance the efficiency of image-text matching, this paper proposes a novel hashing method termed DEMO for binary descriptors, which enables the calculation of similarity using the efficient “XOR" operation (Gu et al., 2022).

2.2 Unsupervised Cross-modal Hashing

Cross-modal hashing (Hu et al., 2023a; Sun et al., 2022a; Tu et al., 2023a; Zeng et al., 2023; Cao et al., 2022) attempts to project samples from various modalities into a shared binary space in which samples with similar semantics should be close. Early efforts typically investigate hand-crafted features for hash codes (Song et al., 2013; Zhou et al.,

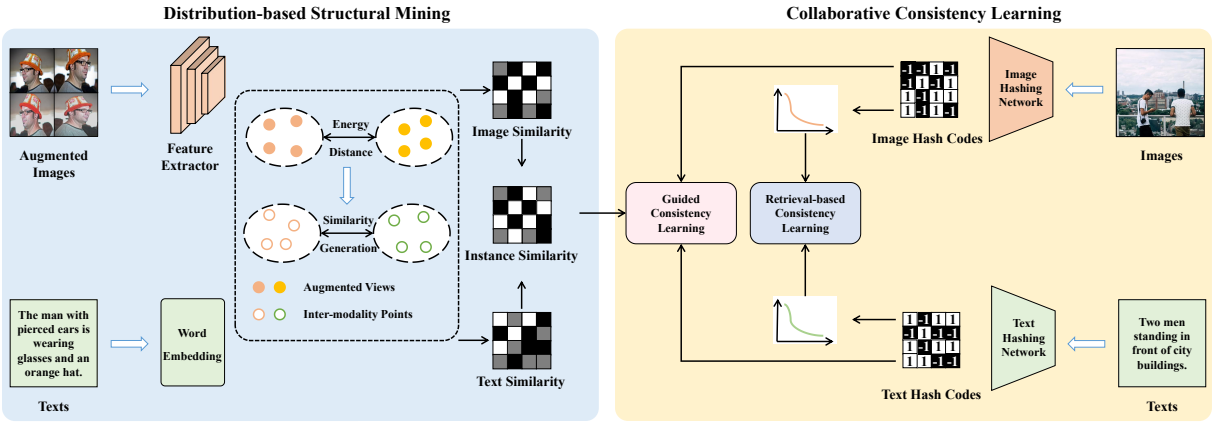

Figure 2: An overview of our proposed DEMO. DEMO first calculates the energy distance between latent semantics distributions to generate an instance similarity matrix. Then DEMO simultaneously optimizes the modality-specific hashing networks by preserving the similarity with guided consistency learning. In addition, retrieval distributions using both image and text queries are encouraged to be consistent to obtain modality-invariant binary codes.

2014), which are typically not discriminative to preserve similarity structure. Recently, various deep unsupervised cross-modal hashing approaches have been developed (Gao et al., 2023; Mikriukov et al., 2022), which typically reconstruct the similarity structure based on cosine distances to optimize the process of learning to hash. However, these methods are incapable of producing precise supervision signals, resulting in inferior binary hash codes. Towards this end, we investigate the latent distribution for each sample and adopt the distribution divergence for enhanced semantic structures.

3 The Proposed Approach

3.1 Problem Definition and Overview

Problem Definition. We begin with notations and the formal definition. $X = \{\boldsymbol{x}_i\}_{i=1}^{N}$ represents a dataset consisting of N images and $Y = \{\boldsymbol{y}_i\}_{i=1}^{N}$ represent a dataset with N texts. Each $\boldsymbol{y}_i$ is associated with text embeddings $\boldsymbol{t}_i$. The objective is to map these samples into a shared Hamming space. We expect the matched samples between two modalities to be encoded into similar binary codes with small Hamming distances. This mapping can guarantee effective and efficient image-text matching.

Framework Overview. This work proposes a new cross-modal hashing approach named DEMO for efficient image-text matching. As depicted in Figure 2, DEMO first employs a pre-trained feature extractor $F^v(\cdot)$ for images, which removes the last layer of a well-known classification neural network (Tu et al., 2023a). We also extract text embeddings using an embedding layer $F^t(\cdot)$. Then, two feed-forward networks (FFNs), $\phi^v(\cdot)$ and $\phi^t(\cdot)$ are adopted to map features of images and texts into binary codes, respectively. Formally, we have:

$$\boldsymbol{b}_i^v = sgn(\phi^v(F^v(\boldsymbol{x}_i))), \quad (1)$$

$$\boldsymbol{b}_i^t = sgn(\phi^t(F^t(\boldsymbol{y}_i))), \quad (2)$$

where $sgn(\cdot)$ is the sign function. Our DEMO mainly consists of two modules, (1) *Distribution-based Structural Mining.* We delve into the inherent semantics distribution behind each image using random data augmentation and utilize the distribution divergence to reconstruct an accurate semantic structure, which would effectively guide the optimization of hashing networks. (2) *Collaborative Consistency Learning.* On the one hand, we maximize the consistency of similarity scores between the semantic structure and hash codes. On the other hand, we produce cross-modal retrieval distributions given texts and images and encourage their consistency from opposing directions.

3.2 Distribution-based Structural Mining

A pivotal challenge in unsupervised cross-modal hashing lies in the lack of supervised information. Previous approaches (Yu et al., 2021; Tu et al., 2023b; Hu et al., 2023a) typically reconstruct the similarity structure as supervision by measuring the natural distance (e.g., cosine distance) of deep features. However, the reconstructed structure may introduce noise, leading to significant error accumulation throughout subsequent optimization stages. In particular, we observe that deep features with the same semantics should originate from a high-dimensional distribution (Sun et al., 2022b; Yang

et al., 2018; Tu et al., 2020), and the natural distance could be inaccurate at the boundaries of latent distributions. Consequently, we aim to measure the distribution divergence for effective structural mining, ensuring high-quality hash codes for efficient image-text matching.

Firstly, we take the image dataset as an example of similarity structure mining. In particular, the random vector of each example $\boldsymbol{x}_i$ in the embedding space is represented as $\boldsymbol{\xi}_i$ with the cumulative distribution function G_i. Then, the distribution divergence between the underlying semantic distributions of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is formulated as:

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \psi(G_i, G_j), \tag{3}$$

in which ψ is a given metric. However, due to immense complexity, parameterizing the high-dimensional distributions remains a considerable challenge. Therefore, classic methods such as KL divergence and JS divergence are not inappropriate here. Towards this end, we turn to a non-parametrized metric, i.e., energy distance (Székely and Rizzo, 2013) . This metric enables modeling of the distribution divergence without the derivation of specific distribution functions, providing an effective alternative for handling the challenges in the high-dimensional space.

Definition 1 *(Energy Distance). Given two independent random vectors $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ with the cumulative distribution functions $G_{\boldsymbol{\xi}}$ and $G_{\boldsymbol{\zeta}}$, respectively. We construct two independent copies $\boldsymbol{\xi}'$ and $\boldsymbol{\zeta}'$ from these cumulative distribution functions. Then, the energy distance is defined as:*

$$D^2(G_{\boldsymbol{\xi}}, G_{\boldsymbol{\zeta}}) = 2\mathrm{E}\rho(\xi, \zeta) - \mathrm{E}\rho\left(\xi, \xi'\right) - \mathrm{E}\rho\left(\zeta, \zeta'\right), \tag{4}$$

where $\rho(\cdot, \cdot)$ is a pointwise distance metric such as cosine distance.

When random variables are real-valued, we can rewrite Eqn. 4 into:

$$D^2(G_{\boldsymbol{\xi}}, G_{\boldsymbol{\zeta}}) = \int_{-\infty}^{\infty} \rho^2(G_{\boldsymbol{\xi}}(x), G_{\boldsymbol{\zeta}}(x)) dx. \tag{5}$$

From Eqn. 5, we can infer that $D^2(G_{\boldsymbol{\xi}}, G_{\boldsymbol{\zeta}}) \geq 0$ and the equality holds when two distributions are identical. In non-parametric test, we generate statistical samples $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_M\}$ and $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_M\}$ from $G_{\boldsymbol{\xi}}$ and $G_{\boldsymbol{\zeta}}$, respectively. Then, we explore the statistics for the null hypothesis, i.e., $G_{\boldsymbol{\xi}} = G_{\boldsymbol{\zeta}}$ by calculating the following averages:

$$\begin{aligned} A &= \frac{1}{M^2} \sum_{m=1}^{M} \sum_{m'=1}^{M} \rho\left(\boldsymbol{u}_m, \boldsymbol{v}_{m'}\right) \\ B &= \frac{1}{M^2} \sum_{m=1}^{M} \sum_{m'=1}^{M} \rho\left(\boldsymbol{u}_m, \boldsymbol{u}_{m'}\right) \\ C &= \frac{1}{M^2} \sum_{m=1}^{M} \sum_{m'=1}^{M} \rho\left(\boldsymbol{v}_m, \boldsymbol{v}_{m'}\right) \end{aligned}. \tag{6}$$

The statistics (Székely and Rizzo, 2013) can be formulated as:

$$\mathcal{E}\left(\{\boldsymbol{u}_m\}_{m=1}^{M}, \{\boldsymbol{v}_m\}_{m=1}^{M}\right) = 2A - B - C, \tag{7}$$

where $\mathcal{E}(\cdot, \cdot)$ denotes energy distance. A large energy distance would reject the null hypothesis, indicating different distribution functions. Since the labels of unlabeled samples cannot acquired, we turn to data augmentation (Sun et al., 2022b; Luo et al., 2021b; He et al., 2020). In particular, we view the augmented view of each image $\boldsymbol{x}_i$ as the samples from its underlying semantic distribution G_i since data augmentation would typically retain the semantics. Therefore, the distribution divergence between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ can be estimated as:

$$d\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \mathcal{E}\left(\{\boldsymbol{z'}_{im}\}_{m=1}^{M}, \{\boldsymbol{z'}_{jm}\}_{m=1}^{M}\right), \tag{8}$$

where $\boldsymbol{z'}_{im} = F^v(\boldsymbol{x}'_{im})$ is the deep feature of the augmented view $\boldsymbol{x}'_{im}$. In our implementation, we use cosine distance for $\rho(\cdot, \cdot)$. Finally, we set a threshold τ to reject the null hypothesis and thus the pair with the distance below the threshold is considered as positive. Moreover, we notice there are still fine-grained differences among dissimilar pairs. Towards this end, we introduce image and text similarities in the semantics structure:

$$S_{ij}^v = \rho(\sum_{m=1}^{M} \boldsymbol{z'}_{im}, \sum_{m=1}^{M} \boldsymbol{z'}_{jm}), \tag{9}$$

$$S_{ij}^t = \rho(\boldsymbol{t}_i, \boldsymbol{t}_j), \tag{10}$$

where $\boldsymbol{t}_i$ is the text embedding of $\boldsymbol{y}_i$. We combine S_{ij}^v and S_{ij}^t to depict the similarities when $d\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) \geq \tau$. In formulation, we construct the instance similarity structure as follows:

$$S_{ij} = \begin{cases} 1, & d\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) < \tau \\ \alpha S_{ij}^v + (1-\alpha) S_{ij}^t, & otherwise, \end{cases} \tag{11}$$

where α is a coefficient to balance two similarities (Tu et al., 2023b, 2020; Ma et al., 2022). It can be noticed that when $M = 1$, our distribution divergence would be degraded to the fundamental cosine distance. The incorporation of multiple augmented views makes it more robust against random attacks. Moreover, it alleviates biases for examples at the boundary of latent semantic distributions, ensuring the accuracy of structural mining.

3.3 Optimization with Collaborative Consistency Learning

In this part, we jointly optimize the image and text hashing networks using collaborative consistency learning which mainly includes guided consistency learning and retrieval-based consistency learning.

Guided Consistency Learning. After constructing the similarity structure (Luo et al., 2021b; Yang et al., 2018; Tu et al., 2020), we aim to preserve this in produced hash codes. In particular, we generate hash codes for both images and texts and then produce their similarities, which would be consistent with the reconstructed structure. Formally, we have:

$$\mathcal{L}_{gui} = \sum_{i,j=1}^{N} \sum_{e_1,e_2 \in \{v,t\}} ||\rho(\boldsymbol{b}_i^{e_1}, \boldsymbol{b}_j^{e_2}) - S_{ij}||^2, \tag{12}$$

where e_1 and e_2 indicate the selected modalities. Therefore, image-image, text-text, and image-text consistency are jointly considered and mapped to the similarity structure under the guidance.

Retrieval-based Consistency Learning. To further reduce the potential distribution discrepancy between the two modalities (Lu et al., 2022; Wei et al., 2021), we simulate the cross-modal retrieval procedure in different directions and enforce the consistency between the retrieval results. In formulation, given a batch, the probability distribution corresponding to text-to-image retrieval is written as:

$$\boldsymbol{p}_i^{T2I} = [\rho(\boldsymbol{b}_i^t, \boldsymbol{b}_1^v), \cdots, \rho(\boldsymbol{b}_i^t, \boldsymbol{b}_B^v)], \tag{13}$$

where B denotes the batch size. Similarly, the probability distribution corresponding to image-to-text retrieval is:

$$\boldsymbol{p}_i^{I2T} = [\rho(\boldsymbol{b}_i^v, \boldsymbol{b}_1^t), \cdots, \rho(\boldsymbol{b}_i^v, \boldsymbol{b}_B^t)]. \tag{14}$$

Then, we utilize the sharpening operator (Xie et al., 2016; Assran et al., 2021; Wang et al., 2023) to refine the soft distributions with:

$$\delta(\boldsymbol{p})_b = \frac{[\boldsymbol{p}]_b^{1/T}}{\sum_{b'=1}^{B}[\boldsymbol{p}]_{b'}^{1/T}}, b = 1, \cdots, B. \tag{15}$$

Our sharpening operation is capable of enhancing the purification of the retrieval results and emphasizing the samples with high similarities. Finally, we conduct consistency learning across two directions in a self-supervised fashion using:

$$\begin{aligned}\mathcal{L}_{ret} = \sum_{i=1}^{B} (&KL(\delta(\boldsymbol{p}_i^{I2T})||\boldsymbol{p}^{T2I}) \\ &+ KL(\delta(\boldsymbol{p}_i^{T2I})||\boldsymbol{p}^{I2T})),\end{aligned} \tag{16}$$

where $KL(\cdot||\cdot)$ returns the KL divergence of two distributions and T is a temperature coefficient that controls the sharp degree set to 0.25 empirically.

Besides, we leverage the co-occurrence knowledge embedded in the dataset, which enforces binary codes of images and texts with identical objects to be close. In particular, we have:

$$\mathcal{L}_{co} = \sum_{i=1}^{N} ||\rho(\boldsymbol{b}_i^v, \boldsymbol{b}_j^t) - \gamma||^2, \tag{17}$$

where γ is set to 1.5 empirically (Tu et al., 2023a) to emphasize this accurate embedding knowledge.

In a nutshell, we summarize our framework by combining all these objectives:

$$\mathcal{L} = \mathcal{L}_{gui} + \mathcal{L}_{ret} + \mathcal{L}_{co}. \tag{18}$$

However, directly minimizing Eqn. 18 is infeasible since $sgn(\cdot)$ is not differentiable at zero and its derivative is zero at the other point. To tackle this problem, we replace $sgn(\cdot)$ with $tanh(\cdot)$ during optimization, which results in approximate hash codes $\hat{\boldsymbol{b}}_i^v = tanh(\phi^v(F^v(\boldsymbol{x}_i)))$ and $\hat{\boldsymbol{b}}_i^v = tanh(\phi^t(F^t(\boldsymbol{y}_i)))$. We summarize the whole training algorithm of DEMO in Algorithm 1.

3.4 Model Inference

After the optimization procedure, we would feed each sample into the hashing network for a binary descriptor. Then, given each query text y_q (image x_q) with a binary code $\boldsymbol{b}_q^t$ ($\boldsymbol{b}_q^v$), we rank the Hamming distances between $\boldsymbol{b}_q^t$ ($\boldsymbol{b}_q^v$) and $\{\boldsymbol{b}_i^v\}_{i=1}^N$ ($\{\boldsymbol{b}_i^t\}_{i=1}^N$), which can produce the nearest examples efficiently. In practice, we consider the returned samples as candidates and conduct fine-grained matching for the final results (Tu et al., 2021).

4 Experiment

4.1 Experimental Settings

Datasets and Evaluation Metrics. To assess the performance of our DEMO, we employ three public and widely-used benchmark datasets to conduct experiments, including MIRFlickr-25K, NUS-WIDE, and MS-COCO. *MIRFlickr-25K* comprises

Task	Method	MIRFlickr-25K				NUS-WIDE				MS-COCO			
		16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
I2T	CVH	0.620	0.608	0.594	0.583	0.487	0.495	0.456	0.419	0.503	0.504	0.471	0.425
	LSSH	0.597	0.609	0.606	0.605	0.442	0.457	0.450	0.451	0.484	0.525	0.542	0.551
	CMFH	0.557	0.557	0.556	0.557	0.339	0.338	0.343	0.339	0.366	0.369	0.370	0.365
	FSH	0.581	0.612	0.635	0.662	0.557	0.565	0.598	0.635	0.539	0.549	0.576	0.587
	MTFH	0.507	0.512	0.558	0.554	0.297	0.297	0.272	0.328	0.399	0.293	0.295	0.395
	FOMH	0.575	0.640	0.691	0.659	0.305	0.305	0.306	0.314	0.378	0.514	0.571	0.601
	DCH	0.596	0.602	0.626	0.636	0.392	0.422	0.430	0.436	0.422	0.420	0.446	0.468
	DGCPN	0.651	0.683	0.718	0.724	0.601	0.618	0.631	0.640	0.556	0.569	0.578	0.580
	UCHSTM	0.701	0.715	0.724	0.723	0.625	0.635	0.646	0.644	0.558	0.572	0.576	0.573
	UCCH	0.716	0.726	0.728	0.732	0.621	0.623	0.640	0.645	0.560	0.562	0.566	0.574
	DEMO	**0.718**	**0.733**	**0.734**	**0.743**	**0.646**	**0.648**	**0.662**	**0.664**	**0.575**	**0.578**	**0.586**	**0.605**
T2I	CVH	0.629	0.615	0.599	0.587	0.470	0.475	0.444	0.412	0.506	0.508	0.486	0.429
	LSSH	0.602	0.598	0.598	0.597	0.473	0.482	0.471	0.457	0.490	0.522	0.547	0.560
	CMFH	0.553	0.553	0.553	0.553	0.306	0.306	0.306	0.306	0.346	0.346	0.346	0.346
	FSH	0.576	0.607	0.635	0.660	0.569	0.604	0.651	0.666	0.537	0.524	0.564	0.573
	MTFH	0.514	0.524	0.518	0.581	0.353	0.314	0.399	0.410	0.335	0.374	0.300	0.334
	FOMH	0.585	0.648	0.719	0.688	0.302	0.304	0.300	0.306	0.368	0.484	0.559	0.595
	DCH	0.612	0.623	0.653	0.665	0.379	0.432	0.444	0.459	0.421	0.428	0.454	0.471
	DGCPN	0.653	0.682	0.712	0.715	0.605	0.626	0.637	0.644	0.550	0.566	0.578	0.577
	UCHSTM	0.695	0.711	0.713	0.723	0.632	0.643	0.651	0.652	0.555	0.567	0.578	0.573
	UCCH	0.703	0.712	0.720	0.721	0.625	0.637	0.650	0.652	0.564	0.573	0.572	0.581
	DEMO	**0.708**	**0.719**	**0.722**	**0.728**	**0.654**	**0.655**	**0.669**	**0.671**	**0.572**	**0.579**	**0.583**	**0.597**

Table 1: MAP scores comparison with code length varying from 16 to 128 bits. I2T refers to the image-to-text matching task, and T2I signifies the text-to-image task. The highest scores are shown in **boldface**.

25,000 pairs of image-text data, and each sample is manually annotated with multiple labels from a set of 24 distinct classes. We remove samples lacking class information, resulting in 20,015 samples for our experiments. We divide these samples into two sets: a query database containing 2,000 paired samples and a retrieval database containing the remaining samples. We employ bag-of-words (BoW) vectors with a dimension of 1,386 to represent the text samples. *NUS-WIDE* consists of 269,498 paired image-text samples and each sample is assigned to a multilabel category from 81 categories. We select 186,557 samples from the top 10 frequent classes for our experiments. These samples are split into a query database with 2,100 image-text pairs and a retrieval database with the remaining samples. Similarly, we employ 1,000-dimensional BoW vectors to represent the text samples. *MS-COCO* is a benchmark dataset which consists of 123,287 images. Each image is associated with 5 annotations from 80 categories. After deleting the samples without label annotations, 122,218 pairs remain during the experiment. We choose 5,000 paired image-text samples randomly as the query database and the remaining pairs are left as the retrieval database. Correspondingly, the text samples are represented by 2026-dimensional BoW vectors.

We evaluate the matching performance based on two protocols: the Hamming ranking protocol and the hash lookup protocol. The former is evaluated by the widely used metric Mean Average Precision (MAP) score, and the latter is evaluated by three types of curves: Precision-Recision curve, Precision-top N curve, and Recall-top N curve. For a fair comparison, we report MAP@All scores as default.

Baselines and Implementation Details. We employ 10 state-of-the-art hashing-based image-text matching approaches as baseline methods, including three supervised cross-modal hashing methods (MTFH (Liu et al., 2019b), FOMH (Lu et al., 2019), DCH (Xu et al., 2017)), four shallow unsupervised cross-modal hashing methods (CVH (Kumar and Udupa, 2011), LSSH (Zhou et al., 2014), CMFH (Ding et al., 2016), FSH (Liu et al., 2017)), and three deep unsupervised cross-modal hashing methods (DGCPN (Yu et al., 2021), UCHSTM (Tu et al., 2023b), UCCH (Hu et al., 2023a)). We randomly select 5,000/10,000/all samples from the retrieval database as the training samples for supervised/deep unsupervised/shallow unsupervised cross-modal hashing methods. For a fair comparison, we follow previous works (Tu et al., 2023b; Hu et al., 2023a) and reimplement the deep unsupervised methods, utilizing VGG-19 pre-trained on the ImageNet dataset and a two-layer MLP as the backbone of the image hashing network and text hashing network, respectively. We adopt the SGD

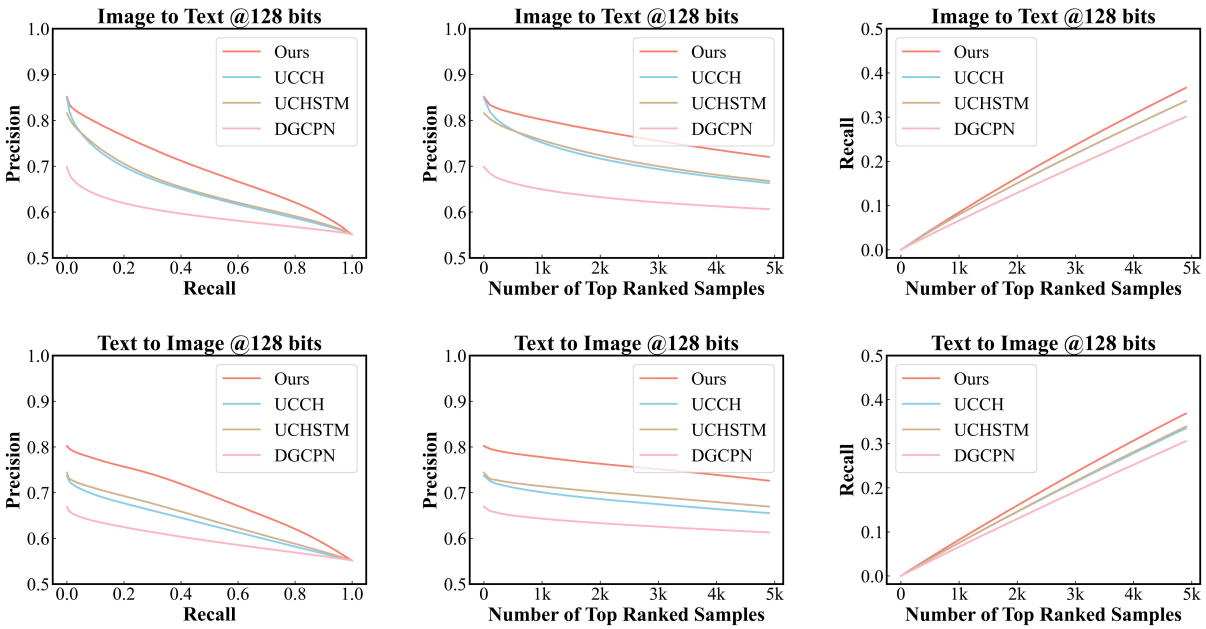

Figure 3: The Precision-Recall curve, Precision-top N curve, and Recall-top N curve with 128 bits on MIRFlickr-25K. The first row plots image-to-text results, and the second row plots text-to-image results.

algorithm with a learning rate of 1e-3 to optimize the networks. The batch size is set to 128. More hyper-parameters are set according to Section 4.4.

4.2 Main Results

Hamming Ranking Protocol. We showcase the MAP scores of all compared baseline methods and our DEMO in Table 1. From these results, the following observations can be attained: *First*, deep unsupervised cross-modal hashing methods outperform shallow unsupervised cross-modal hashing approaches even with insufficient amounts of training data, indicating the superiority of deep neural networks in generating high-quality and modality-invariant hash codes. *Next*, supervised methods excel due to their reliance on expensive labeled data. However, when labeled data is scarce, these methods fall short compared to deep unsupervised approaches. Consequently, deep unsupervised cross-modal hashing emerges as the fundamental technique for image-text matching in the presence of vast amounts of unlabeled multimodal data. *Furthermore*, DEMO outperforms all the compared state-of-the-art hashing-based image-text matching methods, revealing the effectiveness of our proposed distribution-based structural mining and collaborative consistency learning. Additionally, our approach exhibits consistent and significant improvements across three datasets, highlighting the success of addressing previously overlooked distribution divergence combined with collaborative consistency. The proposed components can enhance the performance of unsupervised hashing-based image-text matching in a robust manner.

Hash Lookup Protocol. We also incorporate the hash lookup protocol to generate Precision-Recall, Precision-top N, and Recall-top N curves for our DEMO and three reproduced deep unsupervised baselines using 128 bits on MirFlickr-25K, as illustrated in Figure 3. Due to space limitations, curves for other code lengths can be seen in Section D. The Precision-Recall curve represents the relationship between the varying precision and recall scores. The Precision-top N and Recall-top N curves depict precision and recall values as the retrieval numbers vary from 1 to $5,000$ with a step size of 100. In brief, for these three types of curves, the higher-performing method's curve is usually above the curves of other methods. These curves clearly illustrate that our DEMO consistently outperforms the other baselines, underscoring its superiority. The hash lookup results are consistent with the Hamming ranking results, further validating the exceptional performance and robustness of our DEMO in image-text matching.

4.3 Ablation Study

In Table 2, we investigate the contributions of each proposed component with 16 bits on three datasets. *Firstly*, we remove the distribution-based struc-

Task	Method	MIRF-25K	NUS-WIDE	MS-COCO
I2T	DEMO w/o D	0.698	0.627	0.560
	DEMO w/o R	0.705	0.632	0.565
	DEMO w/o S	0.712	0.636	0.571
	Full Model	**0.718**	**0.646**	**0.575**
T2I	DEMO w/o D	0.696	0.630	0.559
	DEMO w/o R	0.695	0.634	0.564
	DEMO w/o S	0.699	0.642	0.569
	Full Model	**0.708**	**0.654**	**0.572**

Table 2: Ablation on each proposed component. The highest MAP scores are shown in **boldface**.

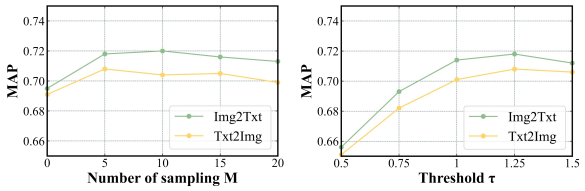

Figure 4: Sensitivity analysis of sampling times M and threshold τ with 16 bits on MIRFlickr-25K.

tural mining process and replace it with sample-based structural mining. The comparison between DEMO w/o D in the first row and the full model in the last row highlights the significant improvement achieved by our distribution-based objective. *Next*, we assess the significance of the retrieval-based consistency learning by removing it. Without this module, the retrieved distributions given images and texts as queries are not encouraged to be consistent. The performance degradation observed in DEMO w/o R in the second row emphasizes the effectiveness of this component. *Moreover*, we conduct an experiment where we remove only the proposed sharpening operation from the retrieval-based consistency learning module. The results of DEMO w/o S in the third row reveal slight differences compared to the full model, underscoring the importance of the sharpening operation. Furthermore, the performance of DEMO w/o S falls between DEMO w/o R and the full model, which is reasonable since DEMO w/o S removes only parts of the retrieval-based consistency learning module while still retaining the ability to promote consistency between the retrieved distributions. *Finally*, we evaluate the full model which incorporates all the components. Results in the last row exhibit the best performance across all scenarios. These experiments successfully verify the significance of each proposed component in DEMO.

4.4 Sensitivity Analysis

To assess the impact of the hyper-parameter M and τ, Figure 4 plots the MAP scores with respect to

Method	LSSH	UGACH	UCCH	DEMO
Inference Time	7.78s	26.59s	0.41s	0.41s

Table 3: Comparison with other methods on the inference speed.

M ranging from 0 to 20, and τ ranging from 0.5 to 1.5. From the results, we can observe that increasing M from 0 to 5 yields a significant performance improvement, but further increasing M from 5 to 20 does not lead to any noticeable enhancement. This phenomenon demonstrates that our DEMO is not sensitive to M within the range of $[5, 20]$. Therefore, M is fixed at 5 and we proceed to investigate the threshold τ, varying from 0.5 to 1.5. The threshold τ plays a crucial role as it controls the percentage of the image-text pairs categorized as positive samples, thereby influencing the quality of the generated similarity matrix. A large value of τ will mistakenly consider numerous incorrect image-text pairs as the matching ones, while a small value of τ will classify many matching image-text pairs as non-matching pairs. From the results, it can be found that 1.25 is the most suitable for the threshold τ. Consequently, we obtain the optimal value of $M = 5$ and $\tau = 1.25$, respectively.

4.5 Efficiency Analysis

We make experimental verification on the inference speed. In particular, we compare our DEMO with state-of-the-art hashing-based image-text matching approaches with 128 bits on MIRFlickr-25K. As shown in Table 3, our DEMO can achieve much higher efficiency compared with LSSH (Zhou et al., 2014) and UGACH (Zhang et al., 2018). Even though the inference time of UCCH (Hu et al., 2023a) and our DEMO is the same, our retrieval performance is much better. In summary, our DEMO is superior to these baselines taking into both efficiency and effectiveness.

4.6 Visualization

We present the t-SNE (van der Maaten and Hinton, 2008) visualization of hash codes from two different modalities generated by four methods with 128 bits on MirFlickr-25K in Figure 5. The results of the comparison with the other three approaches reveal that our DEMO demonstrates a significantly higher degree of similarity and overlap between the image and text modalities. This observation serves as a strong indication that combined with collaborative consistency learning, distribution-based struc-

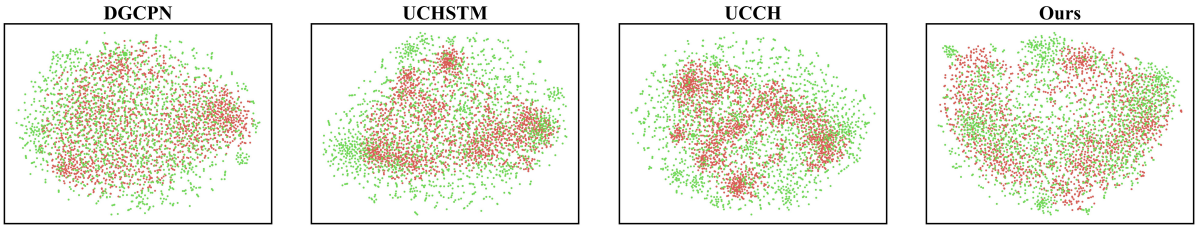

Figure 5: The t-SNE visualization with 128 bits on the MIRFlickr-25K. The image modality is colored red, and the text modality is colored green. The overlap degree represents the degree of modality-invariant hash codes.

tural mining is superior to sample-based structural mining. The visualization results also provide compelling evidence of the exceptional quality and modality-invariant hash codes learned by our approach.

5 Conclusion

In this paper, we investigate the problem of image-text matching and propose a novel deep unsupervised hashing-based approach termed DEMO. The crux of our DEMO is to explore the latent semantic distributions of each sample for effective semantics structure mining. Specifically, we characterize each image with multiple augmented views, which are regarded as samples from its intrinsic semantic distribution. Then, a non-parametric distribution divergence is employed to ensure a robust and accurate similarity structure in the process of similarity generation, which serves as guidance for the optimization. Extensive experimental results across multiple datasets substantiate the efficacy of DEMO.

6 Limitation

Although our DEMO achieves promising results, it still has some limitations. First, there could be different complicated scenarios in real-world applications such as data contamination and domain shift. We would extend our DEMO to more generalization scenarios in our future works. Second, our unsupervised hashing approach DEMO targets at coarse-level retrieval. How to improve unsupervised cross-modal hashing for fine-grained cross-modal retrieval remains an open problem.

References

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Armand Joulin, Nicolas Ballas, and Michael Rabbat. 2021. Semi-supervised learning of visual features by non-parametrically predicting view assignments with support samples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8443–8452.

Abhidip Bhattacharyya, Cecilia Mauceri, Martha Palmer, and Christoffer Heckman. 2022. Aligning images and text with semantic role labels for fine-grained cross-modal understanding. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 4944–4954. European Language Resources Association.

Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. 2022. Image-text retrieval: A survey on recent research and development. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 5410–5417. ijcai.org.

Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. 2020. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12655–12663.

Zhen-Duo Chen, Chuan-Xiang Li, Xin Luo, Liqiang Nie, Wei Zhang, and Xin-Shun Xu. 2019. Scratch: A scalable discrete matrix factorization hashing framework for cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(7):2262–2275.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Zihao Wu, Lin Zhao, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. Chataug: Leveraging chatgpt for text data augmentation. *CoRR*, abs/2302.13007.

Guiguang Ding, Yuchen Guo, Jile Zhou, and Yue Gao. 2016. Large-scale cross-modality search via collective matrix factorization hashing. *IEEE Transactions on Image Processing*, 25(11):5427–5440.

Xinfeng Dong, Huaxiang Zhang, Lei Zhu, Liqiang Nie, and Li Liu. 2022. Hierarchical feature aggregation based on transformer for image-text matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):6437–6447.

Wentao Fan, Chao Zhang, Huaxiong Li, Xiuyi Jia, and Guoyin Wang. 2023. Three-stage semisupervised cross-modal hashing with pairwise relations exploitation. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14.

Zheren Fu, Zhendong Mao, Yan Song, and Yongdong Zhang. 2023. Learning semantic relationship among instances for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15159–15168.

Zijun Gao, Jun Wang, Guoxian Yu, Zhongmin Yan, Carlotta Domeniconi, and Jinglin Zhang. 2023. Long-tail cross modal hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7642–7650.

Wen Gu, Xiaoyan Gu, Jingzi Gu, Bo Li, Zhi Xiong, and Weiping Wang. 2019. Adversary guided asymmetric hashing for cross-modal retrieval. In *Proceedings of the 2019 on international conference on multimedia retrieval*, pages 159–167.

Wenchao Gu, Yanlin Wang, Lun Du, Hongyu Zhang, Shi Han, Dongmei Zhang, and Michael Lyu. 2022. Accelerating code search with deep hashing and code classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2534–2544.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Peng Hu, Hongyuan Zhu, Jie Lin, Dezhong Peng, Yin-Ping Zhao, and Xi Peng. 2023a. Unsupervised contrastive cross-modal hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3877–3889.

Xuming Hu, Zhijiang Guo, Zhiyang Teng, Irwin King, and Philip S. Yu. 2023b. Multimodal relation extraction with cross-modal retrieval and synthesis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 303–311. Association for Computational Linguistics.

Yan Huang, Yuming Wang, Yunan Zeng, and Liang Wang. 2022. MACK: multimodal aligned conceptual knowledge for unpaired image-text matching. In *NeurIPS*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Qing-Yuan Jiang and Wu-Jun Li. 2017. Deep cross-modal hashing. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3270–3278. IEEE Computer Society.

Parminder Kaur, Husanbir Singh Pannu, and Avleen Kaur Malhi. 2021. Comparative analysis on cross-modal information retrieval: A review. *Computer Science Review*, 39:100336.

Vlad Krotov and Leigh Johnson. 2023. Big web data: Challenges related to data, technology, legality, and ethics. *Business Horizons*, 66(4):481–491.

Shaishav Kumar and Raghavendra Udupa. 2011. Learning hash functions for cross-view similarity search. In *Twenty-second international joint conference on artificial intelligence*.

Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. 2019a. Focus your attention: A bidirectional focal attention network for image-text matching. In *Proceedings of the 27th ACM international conference on multimedia*, pages 3–11.

Fangyu Liu and Rongtian Ye. 2019. A strong and robust baseline for text-image matching. In *Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 169–176. Association for Computational Linguistics.

Hong Liu, Rongrong Ji, Yongjian Wu, Feiyue Huang, and Baochang Zhang. 2017. Cross-modality binary code learning via fusion similarity hashing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7380–7388.

Xiaoqing Liu, Huanqiang Zeng, Yifan Shi, Jianqing Zhu, Chih-Hsien Hsia, and Kai-Kuang Ma. 2023. Deep cross-modal hashing based on semantic consistent ranking. *IEEE Transactions on Multimedia*, pages 1–12.

Xin Liu, Zhikai Hu, Haibin Ling, and Yiu-ming Cheung. 2019b. Mtfh: A matrix tri-factorization hashing framework for efficient cross-modal retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):964–981.

Haoyu Lu, Nanyi Fei, Yuqi Huo, Yizhao Gao, Zhiwu Lu, and Ji-Rong Wen. 2022. Cots: Collaborative two-stream vision-language pre-training model for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15692–15701.

Xu Lu, Lei Zhu, Zhiyong Cheng, Jingjing Li, Xiushan Nie, and Huaxiang Zhang. 2019. Flexible online multi-modal hashing for large-scale multimedia retrieval. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1129–1137.

Xiao Luo, Daqing Wu, Zeyu Ma, Chong Chen, Minghua Deng, Jianqiang Huang, and Xian-Sheng Hua. 2021a. A statistical approach to mining semantic similarity for deep unsupervised hashing. In *Proceedings of the 29th ACM International Conference on Multimedia*, page 4306–4314.

Xiao Luo, Daqing Wu, Zeyu Ma, Chong Chen, Minghua Deng, Jinwen Ma, Zhongming Jin, Jianqiang Huang, and Xian-Sheng Hua. 2021b. CIMON: towards high-quality hash codes. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 902–908. ijcai.org.

Zeyu Ma, Wei Ju, Xiao Luo, Chong Chen, Xian-Sheng Hua, and Guangming Lu. 2022. Improved deep unsupervised hashing via prototypical learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 659–667.

Georgii Mikriukov, Mahdyar Ravanbakhsh, and Begüm Demir. 2022. Unsupervised contrastive hashing for cross-modal retrieval in remote sensing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4463–4467. IEEE.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Maria L Rizzo and Gábor J Székely. 2016. Energy distance. *Wiley Interdisciplinary Reviews: Computational Statistics*, 8(1):27–38.

Jingkuan Song, Yang Yang, Yi Yang, Zi Huang, and Heng Tao Shen. 2013. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *Proceedings of the 2013 ACM SIGMOD international conference on management of data*, pages 785–796.

Changchang Sun, Hugo Latapie, Gaowen Liu, and Yan Yan. 2022a. Deep normalized cross-modal hashing with bi-direction relation reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4941–4949.

Jinan Sun, Haixin Wang, Xiao Luo, Shikun Zhang, Wei Xiang, Chong Chen, and Xian-Sheng Hua. 2022b. Heart: Towards effective hash codes under label noise. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 366–375.

Rui Sun, Zhecan Wang, Haoxuan You, Noel Codella, Kai-Wei Chang, and Shih-Fu Chang. 2023. Unifine: A unified and fine-grained approach for zero-shot vision-language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 778–793.

Gábor J Székely and Maria L Rizzo. 2013. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272.

Rong-Cheng Tu, Lei Ji, Huaishao Luo, Botian Shi, He-Yan Huang, Nan Duan, and Xian-Ling Mao. 2021. Hashing based efficient inference for image-text matching. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 743–752.

Rong-Cheng Tu, Jie Jiang, Qinghong Lin, Chengfei Cai, Shangxuan Tian, Hongfa Wang, and Wei Liu. 2023a. Unsupervised cross-modal hashing with modality-interaction. *IEEE Transactions on Circuits and Systems for Video Technology*.

Rong-Cheng Tu, Xian-Ling Mao, Qinghong Lin, Wenjin Ji, Weize Qin, Wei Wei, and Heyan Huang. 2023b. Unsupervised cross-modal hashing via semantic text mining. *IEEE Transactions on Multimedia*, pages 1–12.

Rong-Cheng Tu, Xianling Mao, and Wei Wei. 2020. Mls3rduh: Deep unsupervised hashing via manifold based local semantic similarity structure reconstructing. In *IJCAI*, pages 3466–3472.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Haixin Wang, Huiyu Jiang, Jinan Sun, Shikun Zhang, Chong Chen, Xian-Sheng Hua, and Xiao Luo. 2023. Dior: Learning to hash with label noise via dual partition and contrastive learning. *IEEE Transactions on Knowledge and Data Engineering*.

Chen Wei, Huiyu Wang, Wei Shen, and Alan L. Yuille. 2021. CO2: consistent contrast for unsupervised visual representation learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Junyuan Xie, Ross B. Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 478–487. JMLR.org.

Xing Xu, Fumin Shen, Yang Yang, Heng Tao Shen, and Xuelong Li. 2017. Learning discriminative binary codes for large-scale cross-modal retrieval. *IEEE Transactions on Image Processing*, 26(5):2494–2507.

Erkun Yang, Cheng Deng, Tongliang Liu, Wei Liu, and Dacheng Tao. 2018. Semantic structure-based unsupervised deep hashing. In *Proceedings of the 27th international joint conference on artificial intelligence*, pages 1064–1070.

Jun Yu, Hao Zhou, Yibing Zhan, and Dacheng Tao. 2021. Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 4626–4634.

Donghuo Zeng, Jianming Wu, Gen Hattori, Rong Xu, and Yi Yu. 2023. Learning explicit and implicit dual common subspaces for audio-visual cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2s):1–23.

Huatian Zhang, Zhendong Mao, Kun Zhang, and Yongdong Zhang. 2022a. Show your faith: Cross-modal confidence-aware network for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3262–3270.

Jian Zhang, Yuxin Peng, and Mingkuan Yuan. 2018. Unsupervised generative adversarial cross-modal hashing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Kun Zhang, Zhendong Mao, Quan Wang, and Yongdong Zhang. 2022b. Negative-aware attention framework for image-text matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15640–15649. IEEE.

Jile Zhou, Guiguang Ding, and Yuchen Guo. 2014. Latent semantic sparse hashing for cross-modal similarity search. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 415–424.

A Algorithm

Algorithm 1 Training Algorithm of DEMO

Require: Image dataset X; text dataset Y; number of augmented views M, threshold τ.
Ensure: Parameters of the hashing network.

Generate M augmented views for every $\boldsymbol{x}_i$;
Calculate the distribution divergence using Eqn. 8;
Generate the instance similarity structure using Eqn. 11;
repeat
  Sample a mini-batch randomly;
  Output approximate binary codes for both images and texts;
  Generate cross-modal retrieval results using Eqn. 13 and Eqn. 14;
  Calculate the whole loss using Eqn. 18;
  Update the hashing network by backpropagation;
until convergence

B Data Augmentation Strategy

We leverage the randomness of data augmentations to convert sample-based structural mining to distribution-based structural mining. The detailed data augmentation strategy is illustrated below. First, we resize the image to 256×256 and randomly crop a size of 224×224. Then we employ strategies such as Random Horizontal Flip, Random Color Jitter with $p = 0.7$, Random Grayscale with $p = 0.2$, and Gaussian Blur with $kernelsize = 3$. Finally, we normalize the data with pre-computed mean and standard values. With this augmentation strategy, the intrinsic semantic distribution of a data sample is established for future semantic structure mining.

C Compared Methods

Many state-of-the-art cross-modal hashing-based methods are employed for comparison, including three supervised methods, four shallow unsupervised methods, and three deep unsupervised methods. The detailed introduction of these methods is as follows:

- **CVH** (Kumar and Udupa, 2011) introduces a novel relaxation technique that transforms the learning-to-hash process into a tractable eigenvalue problem. To address this challenge, they utilize techniques such as Locality Sensitive Indexing and Canonical Correlation Analysis.

- **LSSH** (Zhou et al., 2014) extracts the latent semantics from textual samples by matrix factorization. It also leverages sparse coding techniques to capture essential image structures. Introducing an effective iterative method, it analyzes the correlation between multimodal representations, thereby narrowing the semantic gap within the latent semantic space.

- **CMFH** (Ding et al., 2016) builds robust connections via cross-modal factorization, integrating locally linear embedding to uphold the Euclidean structure. Additionally, it employs a classifier-like loss function to leverage semantic label information effectively.

- **FSH** (Liu et al., 2017) defines the similarity between different modalities by introducing a graph-based framework, and then utilizing it to learn modality-invariant hash codes.

- **MTFH** (Liu et al., 2019b) proposes to learn semantic correlations between modalities and aligns heterogeneous data to obtain modality-specific hash codes.

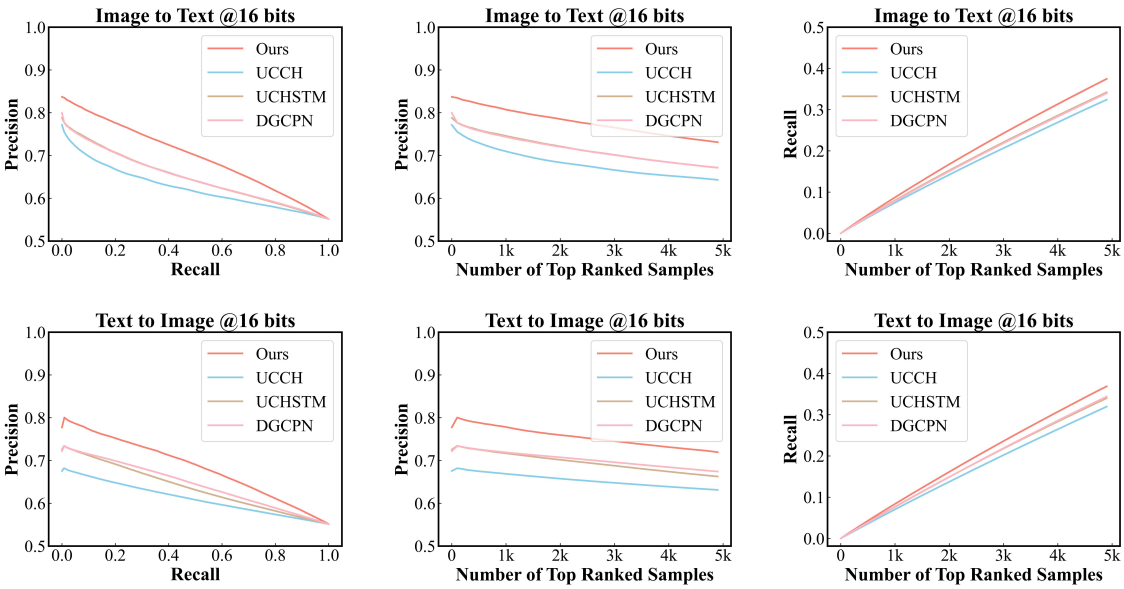

Figure 6: The Precision-Recall curve, Precision-top N curve, and Recall-top N curve with 16 bits on the MIRFlickr-25K dataset. Image-to-text results are plotted in the first row, and text-to-image results are plotted in the second row.

- **FOMH** (Lu et al., 2019) introduces a multi-modal fusion framework to fuse representations when modalities are missing, and then constructs discriminative hash codes.
- **DCH** (Xu et al., 2017) optimizes the network to get modality-specific and modality-invariant hash codes simultaneously. Moreover, it refines the hash codes by iterative training to enhance efficiency.
- **DGCPN** (Yu et al., 2021) investigates the correlations between data samples and their neighbors to improve the quality of similarity generation. It employs a hybrid optimization strategy, combining real and binary components, to minimize discrepancies between the Hamming space and the continuous latent space, thus enhancing similarity and value consistency.
- **UCHSTM** (Tu et al., 2023b) explores correlations among words in textual data points, facilitating the creation of a text modality-specific similarity matrix derived from these correlations. Furthermore, it introduces a self-redefined similarity loss to rectify inaccuracies in the instance similarity matrix, thereby improving the accuracy of similarity measurements.
- **UCCH** (Hu et al., 2023a) introduces contrastive learning, aiming to align various modalities with unified binary representations. It emphasizes leveraging discrimination from all pairs rather than solely focusing on the hardest negative pairs.

D Detailed Hash Lookup Protocol

We showcase the hash lookup results on MIRFlickr-25K with varying code lengths in Figure 6, Figure 7, and Figure 8. From the results of the hash lookup protocol, several conclusions can be observed:

1. Firstly, from the Precision-Recall curve, we can notice the correlation between precision and recall scores. These two metrics are contradictory to each other, as an increase in one often leads to a decrease in the other. From the results in the first column, it can be found that as the recall score increases, the precision score of our DEMO consistently surpasses the other three compared baseline hashing-based image-text matching methods.
2. Secondly, the Precision-top N curve represents the correlation between the precision score and the top N number of results returned in a single retrieval process. As the number of retrieved samples increases, the precision score tends to decrease. From the results in the second column, it can be observed that as

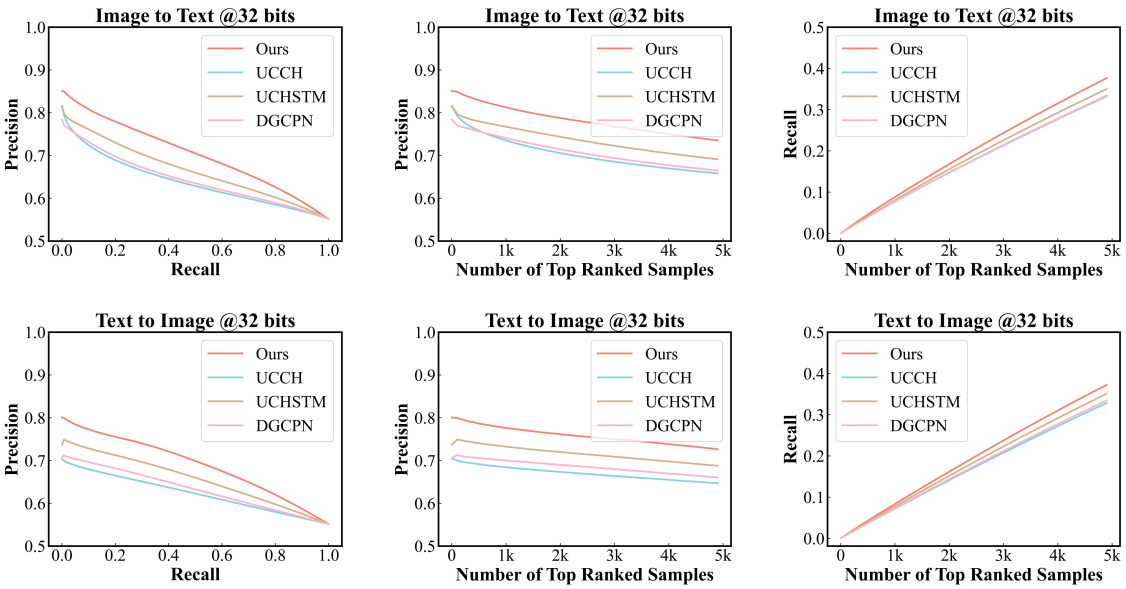

Figure 7: The Precision-Recall curve, Precision-top N curve, and Recall-top N curve with 32 bits on the MIRFlickr-25K dataset. Image-to-text results are plotted in the first row, and text-to-image results are plotted in the second row.

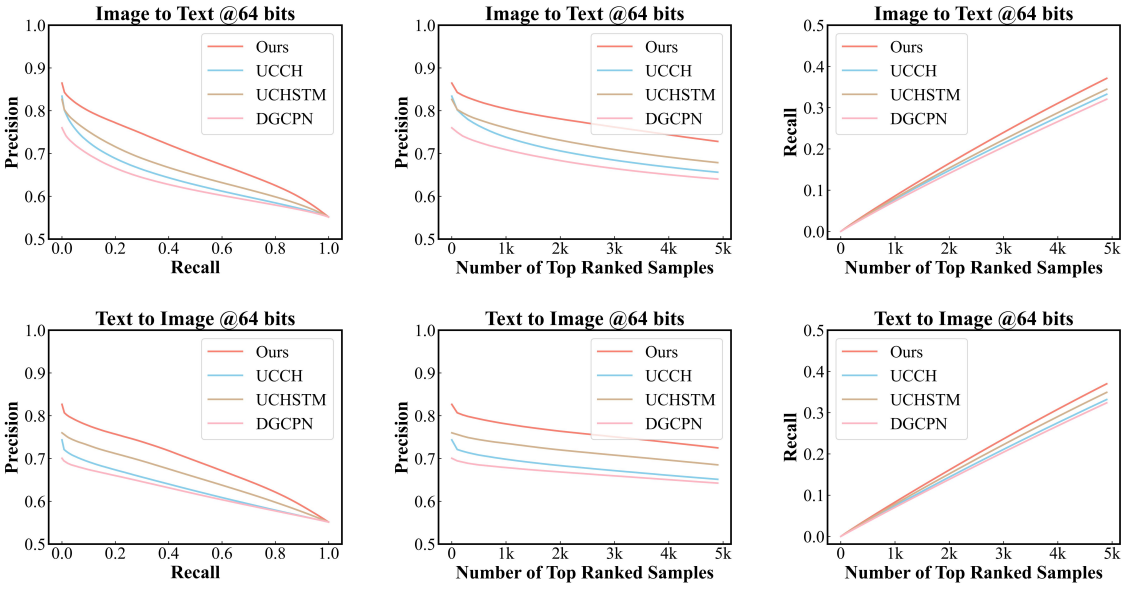

Figure 8: The Precision-Recall curve, Precision-top N curve, and Recall-top N curve with 64 bits on the MIRFlickr-25K dataset. Image-to-text results are plotted in the first row, and text-to-image results are plotted in the second row.

N increases from 1 to 5000, our DEMO consistently outperforms the other three methods.

3. Furthermore, similar to the Precision-top N curve, the Recall-top N curve represents the correlation between the recall score and the top N results returned in a single retrieval. Different from the precision score, the recall score tends to increase as N increases. From the results in the third column, it can be seen that as N increases from 1 to 5000, our curve consistently remains above the other three curves.

4. Lastly, a large number of results demonstrate the robustness of our method from different perspectives. Whether it pertains to various code lengths or different modalities, all the results indicate that we have successfully explored a more suitable similarity structure for

unsupervised cross-modal hashing from a distribution perspective. By combining collaborative consistency learning, DEMO effectively improves the image-text matching quality.