Polysemic Semantic Instance Network for Cross-Modal Hashing

Shuo Han[1], **Qibing Qin**[2*], **Kezhen Xie**[3], **Wenfeng Zhang**[4], **Lei Huang**[5]
[1]Qufu Normal University
[2]Weifang University
[3]Qingdao University of Technology
[4]Chongqing Normal University
[5]Ocean University of China
hs_hanshuo@163.com, qinbing@wfu.edu.cn, xiekezhen@qut.edu.cn, itzhangwf@cqnu.edu.cn, huangl@ouc.edu.cn

Abstract

Hashing techniques are widely adopted in large-scale cross-modal retrieval due to their efficiency and low storage cost. However, semantic ambiguities, including polysemy, multi-object images, and missing semantic descriptions, significantly degrade the accuracy of alignment and retrieval performance. Most existing methods rely on one-to-one mappings that preserve only global average semantics, which fail to capture the intrinsic polysemous structures embedded within individual samples. To address this issue, we propose a novel Deep Polysemic Semantic Instance Hashing (DPSIH) method and design a Diverse Semantic Instance Embedding (DSIE) module. This module integrates local and global features through multi-head self-attention and residual learning, generating multiple diverse embeddings per sample to effectively capture fine-grained and polysemous semantic structures. Furthermore, we design a multi-embedding semantic correlation constraint that relaxes strict alignment restrictions to improve robustness under partial alignment, and introduce Maximum Mean Discrepancy (MMD) regularization to alleviate cross-modal distribution shifts. Additionally, an embedding diversity mechanism is proposed to prevent all embeddings from collapsing into a central or averaged representation, thereby enhancing semantic diversity. Extensive experiments on four benchmark datasets demonstrate that DPSIH significantly outperforms state-of-the-art methods and effectively improves the modeling of semantic ambiguity in cross-modal retrieval tasks.

Code — https://github.com/QinLab-WFU/DPSIH

1 Introduction

With the increasing dimensionality and scale of multimedia data, achieving efficient retrieval in complex data scenarios has become a pressing and significant challenge. Hashing techniques, owing to their advantages in storage and retrieval efficiency, have been widely adopted in large-scale retrieval tasks (Zhu et al. 2023; Dubey 2021; Chen et al. 2022). Their primary objective is to compress high-dimensional features into low-dimensional discrete representations while preserving the similarity relationships among the original data as much as possible (Wang et al. 2017; Qin et al. 2023b). Compared to single-modal hashing methods(Qin et al. 2020, 2023a) that are limited to retrieval within a single data source, cross-modal hashing (Shen et al. 2020; Huo et al. 2024b; Cheng et al. 2025) demonstrates greater flexibility in handling heterogeneous data. The key challenge lies in mapping heterogeneous data into a unified discrete space while achieving semantic alignment. Most existing cross-modal hashing methods typically project each heterogeneous sample into a single point in the discrete space and compute semantic similarity between samples using Euclidean distance or cosine similarity to perform cross-modal retrieval (Wang et al. 2016; Zhu et al. 2023).

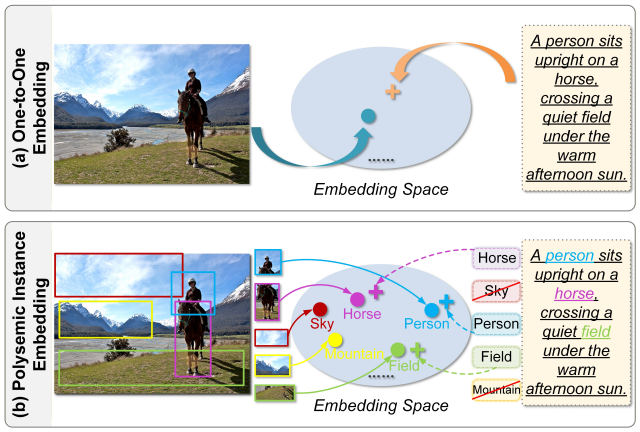


Figure 1: Motivation of the proposed DPSIH method. (a) One-to-one embedding compresses rich semantics into a single representation, making polysemous information hard to preserve and causing information loss under semantic misalignment. (b) In contrast, our polysemic instance embedding learns multiple embeddings to capture intrinsic semantic diversity and effectively mitigate semantic ambiguity.

However, multi-modal data often suffers from semantic ambiguity and incomplete descriptions. On the one hand, multi-object scenes in images or polysemous expressions in text result in samples with complex and diverse semantic structures (Xu et al. 2015; Shi and Jain 2019; Thomas and Kovashka 2022). On the other hand, in real-world applications, cross-modal pairs frequently exhibit partial semantic alignment, where text may describe only local regions

*Corresponding Author.

or specific attributes of an image, leading to unclear correspondences between modalities (Song and Soleymani 2018; Chun et al. 2021; Tu et al. 2025). Although previous one-to-one embedding methods that map each sample to a single vector representation have made notable progress, they inherently average out multiple semantic meanings during the projection process, resulting in semantic compression and the loss of potentially important information (Kim, Kim, and Kwak 2023; Song and Soleymani 2019). As illustrated in Fig. 1-(a), an image sample may contain multiple semantic components such as person, horse, sky, mountain, and field, while the corresponding text may only describe person, horse, and field. The one-to-one embedding approach represents the sample using a single coarse-grained vector, which fails to capture the full semantic content. Particularly under incomplete alignment, unaligned but semantically relevant information can be weakened or even lost during mapping, with no means of recovery. In contrast, our proposed DPSIH, as illustrated in Fig. 1-(b), generates multiple embeddings for each sample, with each embedding corresponding to a different semantic component (e.g., image region or text token), thereby enabling fine-grained cross-modal semantic alignment. Therefore, one-to-one embedding methods are inherently inadequate for handling multi-modal samples with polysemous semantics.

To address these limitations, we propose a novel deep supervised cross-modal hashing method, termed Deep Polysemic Semantic Instance Hashing (DPSIH). By integrating multi-head self-attention and residual learning, we design a Diverse Semantic Instance Embedding (DSIE) module that fuses token-level local features with global semantic representations, generating multiple diverse embeddings to learn a more discriminative embedding space. As illustrated in Fig. 2, the DPSIH framework mainly consists of three key components: (1) The DSIE, which leverages multi-head self-attention and residual learning to generate multiple diverse embeddings for each heterogeneous sample. (2) A multi-embedding semantic correlation constraint, which relaxes strict alignment restrictions to enhance robustness under partial semantic alignment, while Maximum Mean Discrepancy (MMD) regularization is introduced to reduce distributional differences between modalities. (3) A binary embedding diversity mechanism, which penalizes redundancy among local features to improve the discriminability and diversity of the learned representations.

In summary, our main contributions are as follows:

1. Firstly, we propose the DPSIH framework featuring the Diverse Semantic Instance Embedding (DSIE) module that integrates multi-head self-attention with residual learning to fuse token-level local features and global representations, thereby generating multiple diverse embeddings per heterogeneous sample and enhancing the modeling of semantic ambiguities.
2. Secondly, to improve robustness under partially aligned semantics, the multi-embedding semantic correlation constraint is proposed by relaxing strict alignment restrictions, while a Maximum Mean Discrepancy (MMD) regularization is employed to mitigate distribution discrepancies between modalities in the shared embedding space.
3. Thirdly, to prevent the collapse of multiple embeddings into redundant representations, the embedding diversity mechanism is designed to penalize redundancy among local features, thereby improving the modeling of semantic diversity.
4. Finally, extensive experiments on four benchmark datasets demonstrate that the proposed DPSIH method outperforms existing techniques in cross-modal hashing tasks.

2 Related Works

2.1 Cross-modal Hashing

In recent years, deep learning-based cross-modal hashing has made notable progress, aiming to map heterogeneous data into a unified discrete space while preserving semantic consistency. In the absence of label information, unsupervised methods exploit intrinsic cross-modal correlations to learn hash functions. For instance, DGCPN preserves local semantic similarity via graph neighborhood structures without explicit supervision (Yu et al. 2021), while BiCMR employs reinforcement learning to optimize bidirectional retrieval and guide hash function learning (Li et al. 2022). In contrast, supervised methods focus more on leveraging prior label information to learn compact discrete representations. To alleviate the modality discrepancy problem, DAGNN combines graph neural networks and adversarial learning to propose a dual-adversarial graph framework, which enhances both multi-label modeling and cross-modal alignment capabilities (Qian et al. 2021). MIAN builds a modality-invariant asymmetric structure to learn shared semantic spaces, achieving fine-grained alignment across modalities (Zhang et al. 2022). In addition, some recent methods have incorporated Transformer architectures to more effectively model deep semantic associations between heterogeneous data. For instance, DHaPH constructs hierarchical semantic proxies in hyperbolic space and utilizes self-paced learning to enhance neighborhood modeling (Huo et al. 2024b). To further improve inter-class separability, DDBH introduces discriminative boundary supervision to strengthen the representation of class margins in the cross-modal discrete space (Qin et al. 2025).

2.2 Embedding Beyond Vector Representation

The above methods focus on one-to-one projection in the embedding space. This coarse-grained embedding often ignores the inherent semantic ambiguity within samples. Even when a sample contains multiple semantics that could be projected to different positions, a single vector only retains an averaged semantic expression, which is insufficient for modeling semantic diversity. To address this limitation, one-to-many embedding approaches based on multi-instance learning have emerged in recent years, aiming to generate a set of embeddings for each sample to capture its multiple semantic components. For example, the BSE method maps each sample to multiple binary vectors and uses set-level similarity measures to enhance the modeling of semantic

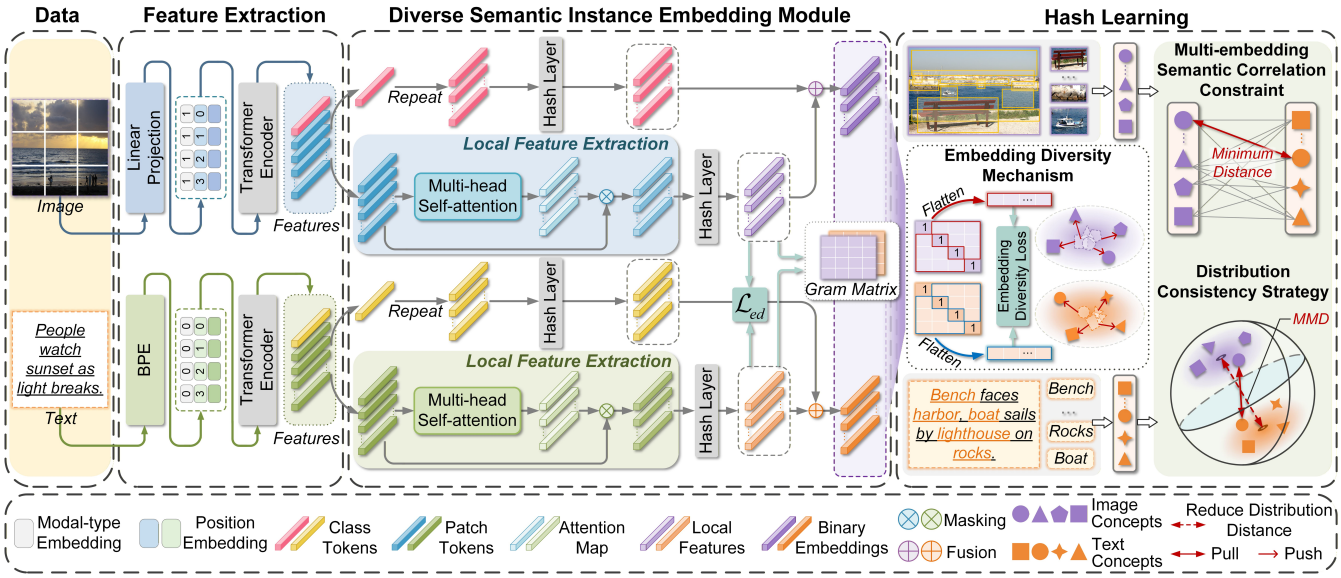


Figure 2: Overview of the DPSIH framework. The proposed method comprises three key components: (1) Feature Extraction: CLIP-based encoders extract modality-specific semantics. (2) Diverse Semantic Instance Embedding Module: multi-head self-attention and residual learning fuse local and global semantics to generate multiple binary embeddings per sample. (3) Hash Learning: By relaxing strict alignment constraints, a multi-embedding semantic correlation constraint is proposed, while Maximum Mean Discrepancy (MMD) regularization is employed to mitigate cross-modal distribution differences, and an embedding diversity mechanism is designed to enhance the discriminability and diversity of the learned hash codes.

ambiguity (Yu, Liu, and Shao 2016). PCME embeds each sample as a Gaussian distribution to capture uncertainty and semantic fuzziness (Chun et al. 2021). DMIH introduces a multi-instance pooling module into the network, enabling the generation of multiple discrete representations for each image and effectively capturing fine-grained semantics in multi-object scenarios (Zhao et al. 2021). In (Kim, Kim, and Kwak 2023), a set prediction module is proposed to generate a diverse embedding set for each sample through slot attention, enabling richer semantic representation. MiViSE further adopts a multi-instance learning mechanism to uncover implicit semantic units within samples and enhance semantic alignment (Song and Soleymani 2018). MIVSE partitions an image into multiple sub-regions, learns an embedding for each, and aggregates them via a multi-instance strategy to form a sample-level representation, thereby improving the modeling of semantic ambiguity (Ren et al. 2017).

3 Method

3.1 Notation

In this paper, we focus on cross-modal retrieval between image and text modalities. Let the dataset containing image-text pairs be denoted as $O = \{o_i\}_{i=1}^{N}$, where each instance $o_i = \{\mathcal{V}_i, \mathcal{T}_i, L_i\}$. Here, $\mathcal{V}_i$ and $\mathcal{T}_i$ represent the i-th image and text sample respectively, and $L = \{l_i\}_{i=1}^{N}$ denotes the label information of instance o_i, where $l_i \in \{0,1\}^C$, and C is the total number of classes. By mapping high-dimensional heterogeneous samples into a common discrete space, we obtain binary codes for the image and text modalities: $H^{\mathcal{V}} = \left[h_1^{\mathcal{V}}, h_2^{\mathcal{V}}, \cdots, h_N^{\mathcal{V}}\right] \in \{-1,1\}^{N\times K}$ and $H^{\mathcal{T}} = \left[h_1^{\mathcal{T}}, h_2^{\mathcal{T}}, \cdots, h_N^{\mathcal{T}}\right] \in \{-1,1\}^{N\times K}$, where K denotes the length of the binary codes. Other notations will be introduced as needed.

3.2 Feature Extraction

Transformers capture global semantic information efficiently through cross-modal self-attention mechanisms, enabling more effective and accurate feature extraction (Tu et al. 2022). Inspired by prior work (Radford et al. 2021; Huo et al. 2024b), we employ two Transformer encoders to extract semantic features from image data $x^{\mathcal{V}}$ and text data $x^{\mathcal{T}}$, producing image modality features $F^{\mathcal{V}} \in \mathbb{R}^{B\times 50\times D}$ and text modality features $F^{\mathcal{T}} \in \mathbb{R}^{B\times 33\times D}$, where B denotes the batch size and D represents the feature embedding dimension. Specifically, each encoder consists of 12 layers, with each layer composed of Layer Normalization (LN), Multi-head Self-Attention (MSA), and a Multi-Layer Perceptron (MLP). For the text encoder, we adopt Byte-Pair Encoding (BPE) to tokenize the input text $x^{\mathcal{T}}$.

3.3 Diverse Semantic Instance Embedding Module

Existing cross-modal hashing methods tend to rely on global semantic representations, which overlook fine-grained and complementary information inherent in multimodal data (Shi and Jain 2019; Chun et al. 2021). In contrast, we design the Diverse Semantic Instance Embedding (DSIE) module by combining the global semantics of heterogeneous modalities and multiple local features for residual learning to ob-

tain the final η binary embedding representations $p^{\mathcal{V}}$ and $p^{\mathcal{T}}$. As illustrated in Fig. 2, the DSIE receives image and text features $F^{\mathcal{V}} \in \mathbb{R}^{B\times 50\times D}$ and $F^{\mathcal{T}} \in \mathbb{R}^{B\times 33\times D}$, respectively. Global features $F^{\mathcal{V}}_{global} \in \mathbb{R}^{B\times 1\times D}$ and $F^{\mathcal{T}}_{global} \in \mathbb{R}^{B\times 1\times D}$ are extracted via the class token and max pooling, while the remaining patch tokens and end-of-text vectors form the local features $F^{\mathcal{V}}_{local} \in \mathbb{R}^{B\times 49\times D}$ and $F^{\mathcal{T}}_{local} \in \mathbb{R}^{B\times 32\times D}$. Since different combinations of local features can obtain diverse and fine-grained representations, a local feature extraction module is designed to recalculate multiple local features $F^{\mathcal{V}}_{local}$ and $F^{\mathcal{T}}_{local}$ to obtain η $\bar{F}^{\mathcal{V}}_{local}$ and $\bar{F}^{\mathcal{T}}_{local}$.

Specifically, a two-layer perceptron implements multi-head self-attention over the local features $F^{\mathcal{V}}_{local}$ and $F^{\mathcal{T}}_{local}$, generating η attention maps $\gamma^{\mathcal{V}} \in \mathbb{R}^{\eta\times B}$ and $\gamma^{\mathcal{T}} \in \mathbb{R}^{\eta\times B}$, defined as:

$$\begin{aligned}\gamma^{\mathcal{V}} &= \text{softmax}\left(\theta_2 \tanh\left(\theta_1 (F^{\mathcal{V}}_{local})^{\top}\right)\right)\\ \gamma^{\mathcal{T}} &= \text{softmax}\left(\theta_2 \tanh\left(\theta_1 (F^{\mathcal{T}}_{local})^{\top}\right)\right)\end{aligned} \tag{1}$$

where $\theta_1 \in \mathbb{R}^{\omega\times D}$, $\theta_2 \in \mathbb{R}^{\eta\times\omega}$, and $\omega = D/2$. The softmax operation is applied row-wise to normalize each attention map.

Subsequently, we multiply $\gamma^{\mathcal{V}}$ with $F^{\mathcal{V}}_{local}$, then apply a sigmoid activation to obtain η new local features $\bar{F}^{\mathcal{V}}_{local} \in \mathbb{R}^{\eta\times D}$. The text modality is constructed in the same manner, and the new local feature representations are computed as in Eq. (2).

$$\begin{aligned}\bar{F}^{\mathcal{V}}_{local} &= \sigma\left(\beta + \left(\gamma^{\mathcal{V}} F^{\mathcal{V}}_{local}\right)\theta_3\right)\\ \bar{F}^{\mathcal{T}}_{local} &= \sigma\left(\beta + \left(\gamma^{\mathcal{T}} F^{\mathcal{T}}_{local}\right)\theta_3\right)\end{aligned} \tag{2}$$

where $\theta_3 \in \mathbb{R}^{D\times D}$ and $\beta \in \mathbb{R}^{D}$.

Next, the global and local semantic information are fused. To avoid redundancy due to shared sample origins across modalities, we employ residual learning to encourage the learning of meaningful local features. Residual learning ensures that local features contribute non-redundant and discriminative semantics, rather than replicating information already captured by the global embedding. For the image modality, $F^{\mathcal{V}}_{global}$ is the main input and $\bar{F}^{\mathcal{V}}_{local}$ is treated as residual. The parameters $(\theta_1, \theta_2, \theta_3, \beta)$ are optimized independently. Therefore, we repeat $F^{\mathcal{V}}_{global}$ η times to form $\bar{F}^{\mathcal{V}}_{global}$, and the final η binary embeddings of the image modality $p^{\mathcal{V}}$ are calculated as shown in Eq. (3).

$$p^{\mathcal{V}} = \text{LN}\left(\text{HL}\left(\bar{F}^{\mathcal{V}}_{global}\right) + \text{HL}\left(\bar{F}^{\mathcal{V}}_{local}\right)\right) \tag{3}$$

where HL is composed of a fully connected layer and a continuous relaxation function (tanh) for learning binary-like codes of length K (Huo et al. 2023), and LN denotes standard layer normalization (Ba, Kiros, and Hinton 2016). Then, during the hash code generation stage, we employ the $sign$ function to convert the binary-like code into a discrete code. Similarly, the text modality is computed in Eq. (4).

$$p^{\mathcal{T}} = \text{LN}\left(\text{HL}\left(\bar{F}^{\mathcal{T}}_{global}\right) + \text{HL}\left(\bar{F}^{\mathcal{T}}_{local}\right)\right) \tag{4}$$

3.4 Hash Learning

Multi-embedding Semantic Correlation Constraint Unlike traditional one-to-one mapping methods, we allow each sample to generate η binary embeddings. Consequently, the semantic correlation between cross-modal sample pairs no longer relies on a single embedding but is determined collectively by all $\eta \times \eta$ embedding pairs. Enforcing strict alignment for all embedding pairs may result in unnecessary penalties, especially when the semantic correlation between samples is not perfect. Inspired by (Amores 2013), we design a multi-embedding semantic correlation constraint. Specifically, for positive sample pairs, we assume that at least one embedding pair is matched, while for negative sample pairs, all embedding pairs should be mismatched. This constraint design allows the model to automatically select the binary embedding pairs with the highest semantic consistency for learning, disregarding mismatched binary embedding pairs. This approach maintains robustness in scenarios with imperfect semantic alignment. This is expressed as:

$$\min_{a,b} d\left(p^{\mathcal{V}}_{i,a}, p^{\mathcal{T}}_{i,b}\right) < d\left(p^{\mathcal{V}}_{i,a}, p^{\mathcal{T}}_{j,b}\right), \forall i \neq j, \forall a, b \tag{5}$$

where $a, b = 1, \ldots, \eta$ and $d(\cdot,\cdot)$ denotes the cosine similarity distance metric. This is further transformed into the multi-embedding semantic correlation loss. Introducing a margin parameter m (default value set to 0.25 in this work), the loss is formulated as follows:

$$\begin{aligned}\mathcal{L}_{msc} = \frac{1}{N^2}\sum_{i,j}^{N} \max\big(&\min_{a,b} d(p^{\mathcal{V}}_{i,a}, p^{\mathcal{T}}_{i,b})\\ &- \min_{a,b} d(p^{\mathcal{V}}_{i,a}, p^{\mathcal{T}}_{j,b}) + m,\ 0\big)\end{aligned} \tag{6}$$

Distribution Consistency Strategy As only the closest pair of cross-modal embeddings is optimized, the remaining $(\eta \times \eta - 1)$ embedding pairs are unconstrained, making the $\min_{a,b} d\left(p^{\mathcal{V}}_{a}, p^{\mathcal{T}}_{b}\right)$ optimization susceptible to distributional bias during training. Given multiple binary embeddings $p^{\mathcal{V}}$ and $p^{\mathcal{T}}$ from DSIE, their overall distributions $p^{\mathcal{V}} \sim \mathcal{V}$ and $p^{\mathcal{T}} \sim \mathcal{T}$ may deviate. Therefore, we introduce Maximum Mean Discrepancy (MMD) (Gretton et al. 2006) as a regularizer to minimize cross-modal distribution divergence, computed as:

$$\text{MMD}(\mathcal{V}, \mathcal{T}) = \sup_{f\in\mathcal{F}}\left(\mathbb{E}_{p^{\mathcal{V}}\sim\mathcal{V}}\left[f\left(p^{\mathcal{V}}\right)\right] - \mathbb{E}_{p^{\mathcal{T}}\sim\mathcal{T}}\left[f\left(p^{\mathcal{T}}\right)\right]\right) \tag{7}$$

where $\mathcal{F}$ is a reproducing kernel Hilbert space (RKHS) with kernel function $\zeta(\cdot)$. As shown in prior work (Gretton et al. 2006), the maximum is achieved when: $f(x) = \mathbb{E}_{X'\sim\mathcal{V}}\left[\zeta(x, X')\right] - \mathbb{E}_{X'\sim\mathcal{T}}\left[\zeta(x, X')\right]$. Then, bringing in Eq. (7) and squaring it, the radial basis function (RBF) kernel is used as the kernel function, and the final distributional consistency loss is shown below.

$$\begin{aligned}\mathcal{L}_{dc} = \frac{1}{N^2\eta^2}\sum_{i,j=1}^{N}\sum_{a,b=1}^{\eta}\big(&\zeta(p^{\mathcal{V}}_{i,a}, p^{\mathcal{V}}_{j,b}) + \zeta(p^{\mathcal{T}}_{i,a}, p^{\mathcal{T}}_{j,b})\\ &- 2\zeta(p^{\mathcal{V}}_{i,a}, p^{\mathcal{T}}_{j,b})\big)\end{aligned} \tag{8}$$

Embedding Diversity Mechanism To encourage the model to learn discriminative and diverse binary embeddings $p^{\mathcal{V}}$ and $p^{\mathcal{T}}$, and to prevent the embeddings from collapsing to a single center, we introduce an embedding diversity mechanism in DSIE. We penalize redundancy in local features $\bar{F}^{\mathcal{V}}_{local}$ and $\bar{F}^{\mathcal{T}}_{local}$, thereby improving semantic diversity. Specifically, we ℓ_2-normalize the local features to lie on a unit sphere, and compute their Gram matrices: $G^{\mathcal{V}} = \bar{F}^{\mathcal{V}}_{local} \cdot (\bar{F}^{\mathcal{V}}_{local})^{\top}$ and $G^{\mathcal{T}} = \bar{F}^{\mathcal{T}}_{local} \cdot (\bar{F}^{\mathcal{T}}_{local})^{\top}$. Since we normalize the features, the diagonal elements of the Gram matrix are one, and the sum of off-diagonal elements reflects local feature redundancy. Therefore, the embedding diversity loss is calculated by Eq. (9).

$$\mathcal{L}_{ed} = \frac{1}{\eta^2} \left(\left\| G^{\mathcal{V}} - I \right\|_2 + \left\| G^{\mathcal{T}} - I \right\|_2 \right) \tag{9}$$

where I denotes the identity matrix. It is worth emphasizing that since the final binary embeddings $p^{\mathcal{V}}$ and $p^{\mathcal{T}}$ generated by the DSIE encapsulate global semantic information, directly imposing orthogonality constraints on them would cause the optimization objective to fail, thereby hindering global semantic modeling. To address this, our designed embedding diversity loss is computed exclusively on the local features $\bar{F}^{\mathcal{V}}_{local}$ and $\bar{F}^{\mathcal{T}}_{local}$. This ensures that the learned $p^{\mathcal{V}}$ and $p^{\mathcal{T}}$ encode diverse semantic information.

Objective Function Combining the above multi-embedding semantic correlation loss, distribution consistency loss, and embedding diversity loss, our overall objective is defined as:

$$\mathcal{L} = \mathcal{L}_{msc} + \alpha_1 \mathcal{L}_{dc} + \alpha_2 \mathcal{L}_{ed} \tag{10}$$

where α_1 and α_2 are hyperparameters to balance the loss terms. By optimizing this objective, the proposed DPSIH effectively captures the complex semantic correlations between heterogeneous modalities and learns compact and discriminative binary codes for efficient cross-modal retrieval.

4 Experiment

We evaluate DPSIH on four public datasets: MIRFLICKR-25K (Huiskes and Lew 2008), NUS-WIDE (Chua et al. 2009), MS COCO (Lin et al. 2014), and IAPR TC-12 (Grubinger et al. 2006). MIRFLICKR-25K contains 24,581 image-text pairs across 24 categories. NUS-WIDE has 195,834 pairs from 21 classes. Following (Qin et al. 2025), MS COCO combines training and validation sets, covering 80 concepts. IAPR TC-12 includes 20,000 samples from 291 classes, each with multiple captions. All datasets follow standard splits as in (Huo et al. 2023): 10,000 for training, 5,000 for query, and the rest for retrieval.

4.1 Experimental Settings

Baselines We compare our method with several representative deep cross-modal hashing approaches, including DCMH (Jiang and Li 2017), SSAH (Li et al. 2018), DCHMT (Tu et al. 2022), MIAN (Zhang et al. 2022), DNPH (Huo et al. 2024a), DHaPH (Huo et al. 2024b), BiLGSEH (Zhu et al. 2025), and DDBH (Qin et al. 2025). For fairness, we adopt the official code released by the authors and strictly follow their recommended parameter settings.

Implementation Details Our DPSIH method is implemented based on the PyTorch framework and trained on an NVIDIA RTX 4090 GPU. During training, we use the Adam optimizer with a learning rate of 0.001 and a weight decay of 0.2. The batch size is set to 128. As for hyperparameters, the loss balance factors α_1 and α_2 are both set to 0.01, and the number of binary embeddings η is set to 4.

Evaluation Metrics To evaluate the performance of DPSIH, we adopt two commonly used metrics: Mean Average Precision (mAP) and Precision@Hamming radius 2 (P@H$\leq$2). Moreover, two types of retrieval tasks are conducted: image-to-text (I2T) and text-to-image (T2I). These evaluation protocols comprehensively validate the effectiveness of DPSIH in cross-modal retrieval scenarios.

4.2 Comparison with the Baselines

Tab. 1 summarizes the mAP results of DPSIH and other baseline methods on MIRFLICKR-25K, NUS-WIDE, MS COCO, and IAPR TC-12 datasets. It can be observed that DPSIH consistently outperforms other methods across all datasets. This demonstrates the effectiveness of our approach in cross-modal hashing by generating multiple binary embeddings per multimodal sample, thereby enhancing its robustness and discriminative ability in scenarios involving semantic ambiguity and partial alignment. Furthermore, to visually present the advantages of our method, we plot the Precision@Hamming radius 2 (P@H$\leq$2) on MIRFLICKR-25K, NUS-WIDE, MS COCO, and IAPR TC-12 datasets, as shown in Fig. 3, further validating the superiority of DPSIH.

4.3 Parameter Analysis

We examine the impact of key hyperparameters on the MIRFLICKR-25K dataset using 32-bit hash codes, as shown in Fig. 4. In Fig. 4-(a), we vary the number of binary embeddings η from 1 to 10 for both I2T and T2I tasks. The performance peaks at $\eta = 4$, confirming its critical role. Fig. 4-(b) investigates the sensitivity of the hyperparameters α_1 and α_2, where the reported results are based on the average mAP of the two retrieval tasks. The model exhibits robustness across a broad range, achieving optimal performance when $\alpha_1 = \alpha_2 = 0.01$.

4.4 Ablation Study

To evaluate the contribution of each component in DPSIH on the MIRFLICKR-25K and IAPR TC-12 datasets using 32-, 64-, and 128-bit codes. As shown in Tab. 2, we compare the full model with four variants: (1) without DSIE module; (2) replacing $\mathcal{L}_{msc}$ with triplet loss; (3) removing $\mathcal{L}_{dc}$; and (4) removing $\mathcal{L}_{ed}$. DPSIH consistently outperforms all variants, verifying the effectiveness of multi-embedding learning and the contribution of each loss term to alignment and representation.

4.5 Robustness Study Under Semantic Ambiguity

To assess robustness under real-world semantic ambiguity, we simulate label noise by randomly perturbing 20%, 50%, and 80% of training labels on MIRFLICKR-25K. We report the average mAP under 32-, 64-, and 128-bit codes. As

Task	Method	MIRFLICKR-25K			NUS-WIDE			MS COCO			IAPR TC-12		
		32bits	64bits	128bits	32bits	64bits	128bits	32bits	64bits	128bits	32bits	64bits	128bits
I2T	DCMH	0.7736	0.7797	0.7881	0.5513	0.5617	0.5703	0.5444	0.5627	0.5659	0.4875	0.5063	0.5249
	SSAH	0.8129	0.8220	0.8059	0.6058	0.6095	0.6069	0.4855	0.5395	0.5482	0.5240	0.5415	0.5203
	DCHMT	0.8253	0.8222	0.8259	0.6706	0.6863	0.6983	0.6216	0.6553	0.6782	0.6196	0.6254	0.6269
	MIAN	0.8444	0.8501	0.8566	0.6548	0.6625	0.6738	0.5987	0.6121	0.6273	0.5657	0.5774	0.5957
	DNPH	0.8269	0.8289	0.8370	0.6811	0.6939	0.7093	0.6910	0.7294	0.7251	0.5147	0.5723	0.6336
	DHaPH	0.8437	0.8531	0.8549	0.7035	0.7136	0.7155	0.7415	0.7475	0.7543	0.6285	0.6326	0.6399
	BiLGSEH	0.8116	0.8194	0.8207	0.6934	0.7090	0.7210	0.7333	0.7596	0.7485	0.6164	0.6215	0.6212
	DDBH	0.8534	0.8610	0.8650	0.7145	0.7229	0.7251	0.7454	0.7681	0.7824	0.6920	0.7031	0.7288
	DPSIH	**0.8703**	**0.8755**	**0.8780**	**0.7235**	**0.7364**	**0.7422**	**0.7801**	**0.7956**	**0.8014**	**0.7215**	**0.7366**	**0.7466**
T2I	DCMH	0.7998	0.8029	0.8083	0.5810	0.5853	0.5904	0.5424	0.5450	0.5526	0.5376	0.5472	0.5708
	SSAH	0.8127	0.8017	0.7473	0.6058	0.6167	0.6620	0.4798	0.5053	0.5147	0.5383	0.5311	0.4075
	DCHMT	0.8048	0.8031	0.8070	0.6837	0.6943	0.7087	0.6212	0.6486	0.6778	0.6286	0.6387	0.6313
	MIAN	0.8151	0.8182	0.8245	0.6922	0.6936	0.6974	0.5493	0.5365	0.5296	0.5516	0.5817	0.5870
	DNPH	0.8176	0.8166	0.8232	0.6994	0.7182	0.7191	0.7012	0.7388	0.7298	0.4950	0.5665	0.6303
	DHaPH	0.8165	0.8229	0.8279	0.7054	0.6998	0.7042	0.7069	0.7154	0.7187	0.6266	0.6454	0.6414
	BiLGSEH	0.8241	0.8343	0.8347	0.7089	0.7202	0.7324	0.7316	0.7543	0.7472	0.6048	0.6097	0.6065
	DDBH	0.8318	0.8390	0.8433	0.7211	0.7325	0.7428	0.7394	0.7595	0.7707	0.6906	0.7005	0.7287
	DPSIH	**0.8404**	**0.8410**	**0.8464**	**0.7319**	**0.7420**	**0.7455**	**0.7779**	**0.7950**	**0.8024**	**0.7402**	**0.7665**	**0.7747**

Table 1: mAP results of DPSIH and baseline methods w.r.t. 32bits, 64bits, and 128bits on four datasets.

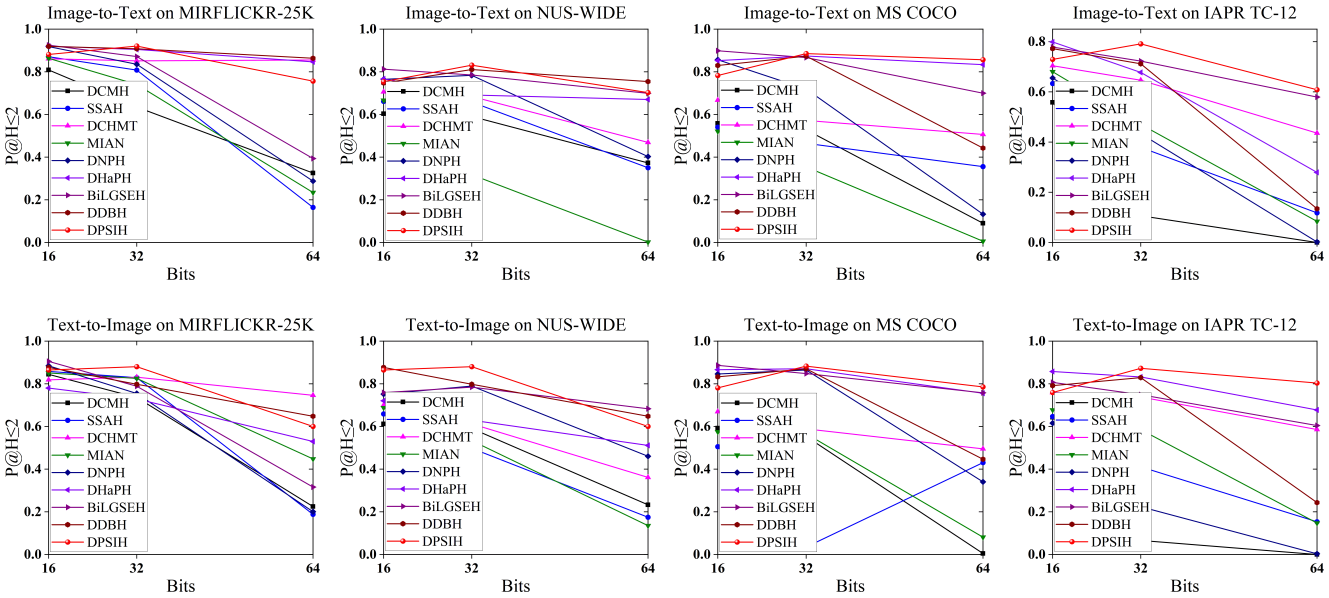


Figure 3: P@H≤2 results with different code lengths on MIRFLICKR-25K, NUS-WIDE, MS COCO and IAPR TC-12.

shown in Tab. 3, DPSIH consistently surpasses DCHMT (Tu et al. 2022), DHRL (Shu et al. 2024), NRCH (Wang et al. 2024), DNPH (Huo et al. 2024a), and DHaPH (Huo et al. 2024b), confirming its robustness under semantic ambiguity and partial modality alignment, enabled by diverse multi-embedding modeling.

4.6 Algorithm Efficiency

To evaluate the efficiency of the model, we compare DPSIH with representative baselines in terms of training and encoding time on the MS COCO dataset with 32-bit hash codes. As shown in Fig. 5, DPSIH achieves shorter training time, attributed to its lightweight DSIE design and efficient objective. Encoding is completed within milliseconds for all methods, with DPSIH maintaining competitive speed, further validating the efficiency of hashing-based retrieval.

4.7 Visualization

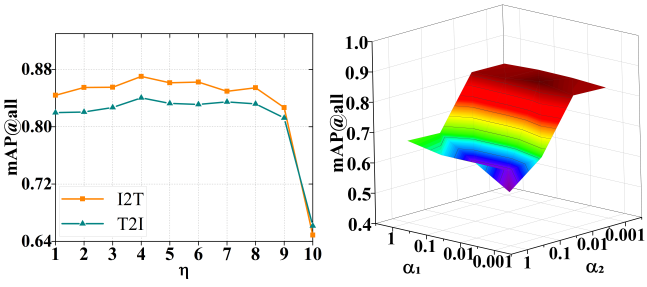
To analyze lexical ambiguity in multimodal data, we visualize frequent nouns and verbs in MS COCO captions, as illustrated in Fig. 6. Common nouns like person and man, and verbs such as sitting and holding, exhibit overlapping or

Method	Image → Text (I2T)			Text → Image (T2I)		
	32	64	128	32	64	128
MIRFLICKR-25K						
w/o $DSIE$	0.8265	0.8437	0.8571	0.8068	0.8105	0.8167
w/o $\mathcal{L}_{msc}$	0.8554	0.8593	0.8611	0.8241	0.8314	0.8344
w/o $\mathcal{L}_{dc}$	0.8634	0.8671	0.8705	0.8213	0.8259	0.8306
w/o $\mathcal{L}_{ed}$	0.8608	0.8646	0.8697	0.8296	0.8325	0.8391
DPSIH	**0.8703**	**0.8755**	**0.8780**	**0.8404**	**0.8410**	**0.8464**
IAPR TC-12						
w/o $DSIE$	0.6318	0.6526	0.6825	0.6122	0.6453	0.6749
w/o $\mathcal{L}_{msc}$	0.7063	0.7192	0.7287	0.7251	0.7367	0.7488
w/o $\mathcal{L}_{dc}$	0.7175	0.7204	0.7391	0.7334	0.7510	0.7631
w/o $\mathcal{L}_{ed}$	0.7146	0.7219	0.7274	0.7386	0.7452	0.7502
DPSIH	**0.7215**	**0.7366**	**0.7466**	**0.7402**	**0.7665**	**0.7747**

Table 2: Ablation study of DPSIH and its variants on MIRFLICKR-25K and IAPR TC-12 at 32, 64, and 128 bits.

Method	32 bits			64 bits			128 bits		
	20%	50%	80%	20%	50%	80%	20%	50%	80%
DCHMT	.8094	.7990	.7848	.8122	.8055	.8039	.8108	.8087	.8056
DHRL	.7023	.6909	.6849	.7085	.6965	.6711	.7052	.6784	.6623
NRCH	.7614	.7544	.7368	.7646	.7619	.7533	.7677	.7645	.7538
DNPH	.7915	.7972	.7805	.8112	.8131	.8013	.8240	.8204	.8027
DHaPH	.8173	.7704	.7452	.8039	.7778	.7554	.8098	.7838	.7583
DPSIH	**.8525**	**.8424**	**.8152**	**.8595**	**.8539**	**.8286**	**.8621**	**.8578**	**.8455**

Table 3: Average mAP of different methods under various noise rates and hash code lengths on MIRFLICKR-25K.

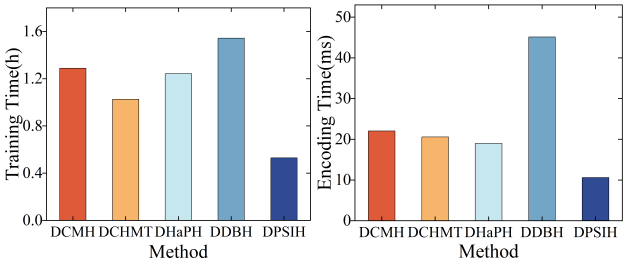


(a) (b)

Figure 4: Parameter sensitivity analysis for η, α_1 and α_2 on MIRFLICKR-25K with 32-bits.

context-dependent meanings. This reveals the inherent semantic ambiguity in object and action descriptions, posing challenges for accurate cross-modal alignment.

We further visualize attention maps to assess the effect of DSIE on semantic representation. As illustrated in Fig. 7, the DSIE-based model captures more diverse and semantically relevant regions, indicating a better ability to capture multiple visual cues. These results show that generating multiple embeddings improves fine-grained semantic modeling under complex scenes.



(a) Training time (b) Encoding time

Figure 5: Training time and encoding time of DPSIH and baselines on MS COCO with 32-bits.

(a) Nouns (b) Verbs

Figure 6: Distributional analysis of nouns and verbs in MS COCO reveals significant lexical ambiguity and conceptual abstraction.

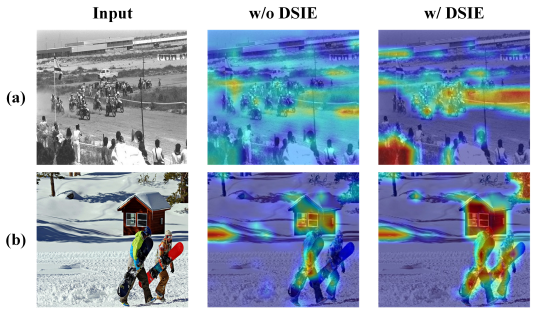


Figure 7: Attention maps of models w/ and w/o DSIE.

5 Conclusion

This paper presents a novel Deep Polysemic Semantic Instance Hashing (DPSIH) framework that integrates local and global semantics through multi-head self-attention and residual learning. A Diverse Semantic Instance Embedding (DSIE) module generates multiple diverse embeddings, thereby learning a discriminative embedding space. By relaxing strict alignment restrictions, a multi-embedding semantic correlation constraint is designed, with Maximum Mean Discrepancy (MMD) regularization mitigating cross-modal distribution discrepancies. Furthermore, an embedding diversity mechanism penalizes redundancy among local features, enhancing the model's capacity to capture semantic diversity. Extensive experiments show the significant advantages of DPSIH in handling semantic ambiguity and partially aligned scenarios. Future work will extend it to zero-shot and few-shot retrieval under limited supervision.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.62302338, No.62472390), Shandong Provincial Natural Science Foundation (No.ZR2022QF046, No.ZR2025MS1067), Natural Science Foundation of Chongqing (No.CSTB2023NSCQ-MSX0407) and Science and Technology Research Program of Chongqing Municipal Education Commission (No.KJQN202200551).

References

Amores, J. 2013. Multiple instance classification: Review, taxonomy and comparative study. *Artificial intelligence*, 201: 81–105.

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Chen, W.; Liu, Y.; Wang, W.; Bakker, E. M.; Georgiou, T.; Fieguth, P.; Liu, L.; and Lew, M. S. 2022. Deep learning for instance retrieval: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6): 7270–7292.

Cheng, K.; Qin, Q.; Zhang, W.; Huang, L.; and Nie, J. 2025. Deep Probabilistic Binary Embedding via Learning Reliable Uncertainty for Cross-Modal Retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, 6393–6402.

Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, 1–9.

Chun, S.; Oh, S. J.; De Rezende, R. S.; Kalantidis, Y.; and Larlus, D. 2021. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8415–8424.

Dubey, S. R. 2021. A decade survey of content based image retrieval using deep learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5): 2687–2704.

Gretton, A.; Borgwardt, K.; Rasch, M.; Schölkopf, B.; and Smola, A. 2006. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19.

Grubinger, M.; Clough, P.; Müller, H.; and Deselaers, T. 2006. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International workshop ontoImage*, volume 2.

Huiskes, M. J.; and Lew, M. S. 2008. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, 39–43.

Huo, Y.; Qibing, Q.; Dai, J.; Zhang, W.; Huang, L.; and Wang, C. 2024a. Deep Neighborhood-aware Proxy Hashing with Uniform Distribution Constraint for Cross-modal Retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(6): 1–23.

Huo, Y.; Qin, Q.; Dai, J.; Wang, L.; Zhang, W.; Huang, L.; and Wang, C. 2023. Deep semantic-aware proxy hashing for multi-label cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(1): 576–589.

Huo, Y.; Qin, Q.; Zhang, W.; Huang, L.; and Nie, J. 2024b. Deep hierarchy-aware proxy hashing with self-paced learning for cross-modal retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 36(11): 5926–5939.

Jiang, Q.-Y.; and Li, W.-J. 2017. Deep cross-modal hashing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3232–3240.

Kim, D.; Kim, N.; and Kwak, S. 2023. Improving cross-modal retrieval with set of diverse embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 23422–23431.

Li, C.; Deng, C.; Li, N.; Liu, W.; Gao, X.; and Tao, D. 2018. Self-supervised adversarial hashing networks for cross-modal retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4242–4251.

Li, T.; Yang, X.; Wang, B.; Xi, C.; Zheng, H.; and Zhou, X. 2022. Bi-CMR: Bidirectional reinforcement guided hashing for effective cross-modal retrieval. In *Proceedings of the AAAI conference on artificial intelligence*, 10275–10282.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Qian, S.; Xue, D.; Zhang, H.; Fang, Q.; and Xu, C. 2021. Dual adversarial graph neural networks for multi-label cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2440–2448.

Qin, Q.; Huang, L.; Wei, Z.; Xie, K.; and Zhang, W. 2020. Unsupervised deep multi-similarity hashing with semantic structure for image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(7): 2852–2865.

Qin, Q.; Huang, L.; Xie, K.; Wei, Z.; Wang, C.; and Zhang, W. 2023a. Deep adaptive quadruplet hashing with probability sampling for large-scale image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12): 7914–7927.

Qin, Q.; Huo, Y.; Zhang, W.; Huang, L.; and Nie, J. 2025. Deep Discriminative Boundary Hashing for Cross-modal Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*.

Qin, Q.; Xie, K.; Zhang, W.; Wang, C.; and Huang, L. 2023b. Deep neighborhood structure-preserving hashing for large-scale image retrieval. *IEEE Transactions on Multimedia*, 26: 1881–1893.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763.

Ren, Z.; Jin, H.; Lin, Z.; Fang, C.; and Yuille, A. L. 2017. Multiple Instance Visual-Semantic Embedding. In *BMVC*.

Shen, H. T.; Liu, L.; Yang, Y.; Xu, X.; Huang, Z.; Shen, F.; and Hong, R. 2020. Exploiting subspace relation in semantic labels for cross-modal hashing. *IEEE Transactions on Knowledge and Data Engineering*, 33(10): 3351–3365.

Shi, Y.; and Jain, A. K. 2019. Probabilistic face embeddings. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6902–6911.

Shu, Z.; Bai, Y.; Yong, K.; and Yu, Z. 2024. Deep cross-modal hashing with ranking learning for noisy labels. *IEEE Transactions on Big Data*, 11: 553–565.

Song, Y.; and Soleymani, M. 2018. Cross-modal retrieval with implicit concept association. *arXiv preprint arXiv:1804.04318*.

Song, Y.; and Soleymani, M. 2019. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1979–1988.

Thomas, C.; and Kovashka, A. 2022. Emphasizing complementary samples for non-literal cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4632–4641.

Tu, J.; Liu, X.; Huang, Z.; Hao, Y.; Hong, R.; and Wang, M. 2025. Cross-Modal Hashing via Diverse Instances Matching. *IEEE Transactions on Image Processing*, 34: 2737–2749.

Tu, J.; Liu, X.; Lin, Z.; Hong, R.; and Wang, M. 2022. Differentiable cross-modal hashing via multimodal transformers. In *Proceedings of the 30th ACM International Conference on Multimedia*, 453–461.

Wang, J.; Zhang, T.; Sebe, N.; Shen, H. T.; et al. 2017. A survey on learning to hash. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 769–790.

Wang, K.; Yin, Q.; Wang, W.; Wu, S.; and Wang, L. 2016. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*.

Wang, L.; Qin, Y.; Sun, Y.; Peng, D.; Peng, X.; and Hu, P. 2024. Robust Contrastive Cross-modal Hashing with Noisy Labels. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5752–5760.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2048–2057. JMLR.org.

Yu, J.; Zhou, H.; Zhan, Y.; and Tao, D. 2021. Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing. In *Proceedings of the AAAI conference on artificial intelligence*, 4626–4634.

Yu, M.; Liu, L.; and Shao, L. 2016. Binary set embedding for cross-modal retrieval. *IEEE transactions on neural networks and learning systems*, 28(12): 2899–2910.

Zhang, Z.; Luo, H.; Zhu, L.; Lu, G.; and Shen, H. T. 2022. Modality-invariant asymmetric networks for cross-modal hashing. *IEEE Transactions on Knowledge and Data Engineering*, 35(5): 5091–5104.

Zhao, W.; Guan, Z.; Luo, H.; Peng, J.; and Fan, J. 2021. Deep multiple instance hashing for fast multi-object image search. *IEEE Transactions on Image Processing*, 30: 7995–8007.

Zhu, L.; Wu, R.; Zhu, X.; Zhang, C.; Wu, L.; Zhang, S.; and Li, X. 2025. Bi-Direction Label-Guided Semantic Enhancement for Cross-Modal Hashing. *IEEE Transactions on Circuits and Systems for Video Technology*, 35: 3983–3999.

Zhu, L.; Zheng, C.; Guan, W.; Li, J.; Yang, Y.; and Shen, H. T. 2023. Multi-modal hashing for efficient multimedia retrieval: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(1): 239–260.