Decoupled Prototype Contrastive Alignment Hashing for Cross-Modal Retrieval

Yunfei Chen[1], **Renwei Xia**[2] , **Shangchong Gao**[3] , **Zhan Yang**[2*]

[1]Information Support Force Engineering University, Wuhan, China
[2]Central South University, Changsha, China
[3]Xinjiang University, Urumqi, China
yunfeichen_csu@126.com, xiarenwei@csu.edu.cn,
gaoshangchong@stu.xju.edu.cn, zyang22@csu.edu.cn

Abstract

Cross-modal hashing is essential for efficient large-scale retrieval in multimedia analysis. However, existing unsupervised methods often struggle to fully capture interactions between modalities, enforce multi-level alignment, and maintain stable optimization. To overcome these challenges, we propose Decoupled Prototype Contrastive Alignment Hashing (DPCAH), which employs a cross-modal semantic contrastive learning module that concatenates image and text features and encodes their interactions via a transformer to produce unified representations. Feature-level and hash-level contrastive objectives jointly align semantic information and guide discriminative hash code learning across modalities. A decoupled prototype consistency module further models cross-modal correlations independently, enhancing semantic alignment while ensuring stable and robust optimization. Experiments on three benchmark datasets demonstrate that DPCAH outperforms state-of-the-art methods in cross-modal retrieval.

1 Introduction

With the rapid growth of multimedia data, such as images, texts, and videos, cross-modal retrieval [11, 34] has become an essential task in multimedia analysis. This task aims to search relevant items across different modalities, such as finding images based on text queries or retrieving textual descriptions for images. Hashing-based methods are particularly attractive for cross-modal retrieval [1] because binary codes enable compact storage and fast similarity computation, making them well suited for large-scale datasets. By learning compact representations for both modalities, cross-modal hashing [9, 26] can reduce memory usage and speed up retrieval while maintaining good accuracy.

Existing cross-modal hashing [13] methods can be broadly divided into supervised and unsupervised approaches. Supervised methods [19, 21] benefit from explicit semantic guidance provided by labels or paired annotations, which helps to learn discriminative and semantically consistent hash codes. However, their reliance on large amounts of labeled data limits scalability[4], and obtaining high-quality annotations is often costly and time-consuming. In contrast, unsupervised methods [14, 35] do not require manual labels, making them more flexible and scalable while still achieving promising retrieval performance. Despite these advantages, unsupervised cross-modal hashing methods still face several key challenges[17]. Many approaches fail to effectively capture interactions between visual and textual modalities at both the feature and hash levels[16], leading to suboptimal semantic alignment. Alignment constraints are often applied at a single representation stage[28], which limits the consistency of semantic correspondence across different levels. In addition, existing methods do not explicitly model cross-modal correlations in a way that preserves the main feature space,

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

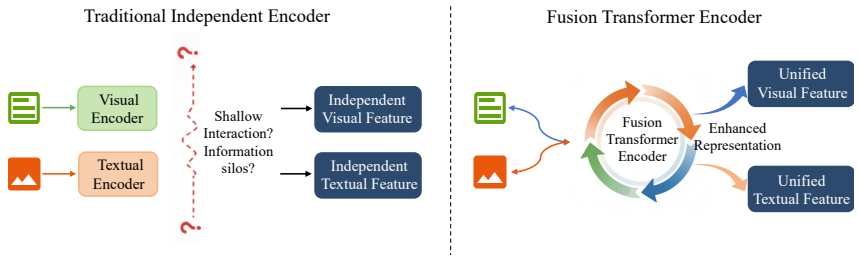


Figure 1: Comparison between the traditional independent encoder and the proposed fusion transformer encoder for cross-modal representation learning.

resulting in unstable training and less robust hash codes[24]. These limitations indicate the need for unsupervised cross-modal hashing methods that can achieve richer feature fusion[23, 30], robust multi-level alignment, and stable optimization, producing discriminative and scalable hash codes suitable for large-scale retrieval.

To address these challenges, this work proposes a new unsupervised cross-modal hashing method Decoupled Prototype Contrastive Alignment Hashing (DPCAH). As illustrated in Fig. 1, instead of adopting the traditional independent encoding paradigm, our method employs a Fusion Transformer Encoder (FTE) to enable early and effective cross-modal interaction, thereby facilitating richer semantic representation learning. Building upon this fusion-based encoding scheme, we design two key modules to further enhance cross-modal representation learning. The first is the Cross-Modal Semantic Contrastive Learning (CMSCL) module, which concatenates image and text features and applies both feature-level and hash-level contrastive objectives to produce discriminative and semantically aligned hash codes. The second is the Decoupled Prototype Consistency (DPC) module, designed as an auxiliary branch to model cross-modal correlations independently of the main feature space, improving alignment without interfering with primary representation learning and ensuring stable optimization. The main contributions of this paper are as follows:

- We propose a cross-modal semantic contrastive learning module that concatenates image and text features, models their interactions with a transformer, and applies feature-level and hash-level contrastive objectives to learn discriminative and semantically aligned hash codes.
- We propose a decoupled prototype consistency module that captures semantic structures via ensemble prototypes and enforces alignment through a consistency objective, improving robustness and stability in cross-modal representation learning.
- We conduct comprehensive ablation studies that verify the effectiveness of each proposed component and demonstrate that their joint use further improves the discriminability and stability of the learned hash codes.

2 Related Work

Cross-modal hashing methods can be broadly categorized into supervised and unsupervised approaches.

2.1 Supervised Cross-Modal Hashing

Supervised cross-modal hashing methods exploit label information to learn unified hash codes and reduce modality gaps, yet still face challenges in discrete optimization, semantic alignment, and multi-instance modeling. Collaboratively Semantic alignment and Metric learning for cross-modal Hashing (CSMH) [10] projects nonlinear features into a shared latent subspace to align cross-modal distributions while embedding label similarity for semantic preservation. However, its global subspace alignment limits the modeling of fine-grained instance-level correspondences. Diverse Instances Matching for Cross-modal Hashing (DIMCH) [22] alleviates this issue by learning diverse instance features and adopting a weighted triplet loss to capture multi-label correlations. Robust Hashing with Bilinear Drift (RHBD) [36] improves robustness and discrimination by integrating bilinear drift

with semantic autoencoding and multiple supervision signals[5]. These methods demonstrate the effectiveness of supervised semantic guidance, but their strong dependence on labels and tightly coupled optimization restricts scalability and flexibility.

2.2 Unsupervised Cross-Modal Hashing

Unsupervised cross-modal hashing methods remove reliance on annotations, but must address semantic alignment and reliable similarity estimation without supervision. Unsupervised Dual Deep Hashing (UDDH) [32] introduces dual deep hash codes combining semantic indexing with modality-specific content representations to enable collaborative optimization. Dual Aggregation-based Joint-modal Similarity Hashing (DAJSH) [29] fuses image and text features via a transformer encoder with contrastive learning and constructs a dual-aggregation affinity matrix to guide hash learning. Vision-guided Text Mining for Unsupervised Cross-modal Hashing (VTM-UCH) [3] leverages CLIP-based vision–language representations to mine fine-grained correspondences and refine similarity structures through community-aware adjustment. Although effective[33], existing unsupervised methods often entangle representation learning and alignment objectives, which may limit training stability and semantic consistency[25].

3 Proposed Method

The overall framework of Decoupled Prototype Contrastive Alignment Hashing (DPCAH) is shown in Fig. 2. It comprises a cross-modal semantic contrastive learning module for joint representation learning and contrastive alignment, and a decoupled prototype consistency module that models cross-modal correlations to enhance semantic alignment and training stability.

3.1 Notations

Suppose we are given a training set $\mathcal{D} = \{(\mathbf{x}_i^v, \mathbf{x}_i^t)\}_{i=1}^n$, where $\mathbf{x}_i^v \in \mathbb{R}^{1\times d_v}$ and $\mathbf{x}_i^t \in \mathbb{R}^{1\times d_t}$ denote the visual and textual features of the i-th sample, respectively. Here, n denotes the number of samples, d_v and d_t represent the dimensions of the visual and textual features. The goal of cross-modal hashing is to learn a compact binary representation $\mathbf{b}_i \in \{-1, 1\}^{1\times k}$ for each sample, where k is the hash code length. The hash codes of all samples are denoted as $\mathbf{B} = \{\mathbf{b}_i\}_{i=1}^n$.

3.2 Cross-Modal Semantic Contrastive Learning

We present our cross-modal semantic contrastive learning module, which uses a fusion transformer encoder to capture interactions between visual and textual modalities, and applies a hierarchical contrastive learning strategy for alignment at both feature and hash levels.

3.2.1 Fusion Transformer Encoder

To extract robust cross-modal representations, we begin by leveraging the powerful CLIP model to encode both visual and textual inputs. Specifically, given the training set $\mathcal{D}$, we extract CLIP features for each modality as follows:

$$\mathbf{f}_i^* = \text{CLIP}(\mathbf{x}_i^*), \quad \text{s.t.} * \in \{v, t\}, \tag{1}$$

where $\mathbf{f}_i^v$ and $\mathbf{f}_i^t$ are the visual and textual feature vectors encoded by CLIP, respectively. To capture deep cross-modal interactions, we concatenate the visual and textual features, $\mathbf{f}_i^v$ and $\mathbf{f}_i^t$, into a single representation $\mathbf{f}_i = \text{concat}[\mathbf{f}_i^v, \mathbf{f}_i^t]$ and feed it into a transformer encoder. The self-attention mechanism ($SAtt$) models global dependencies across modalities and produces the intermediate fused features $\bar{\mathbf{f}}_i$:

$$\bar{\mathbf{f}}_i = SAtt(q_i, K, V) = \sum_{j=1}^{m} \text{softmax}\left(\frac{\mathbf{q}_i \mathbf{k}_j^T}{\sqrt{d_k}}\right)\mathbf{v}_j, \tag{2}$$

where $\mathbf{q}_i, \mathbf{k}_j, \mathbf{v}_j$ are queries, keys, and values derived from $\mathbf{f}_i$ and $\mathbf{f}_j$, and m is the batch size. The output of the transformer $\mathbf{z}_i \in \mathbb{R}^{1\times 2d}$ is obtained by processing $\bar{\mathbf{f}}_i$ through a feed-forward network with residual connections and layer normalization, and is formed by concatenating the visual and textual features, $\mathbf{z}_i = \text{concat}[\mathbf{z}_i^v, \mathbf{z}_i^t]$, where $\mathbf{z}_i^v$ and $\mathbf{z}_i^t$ are the visual and textual representations, respectively.

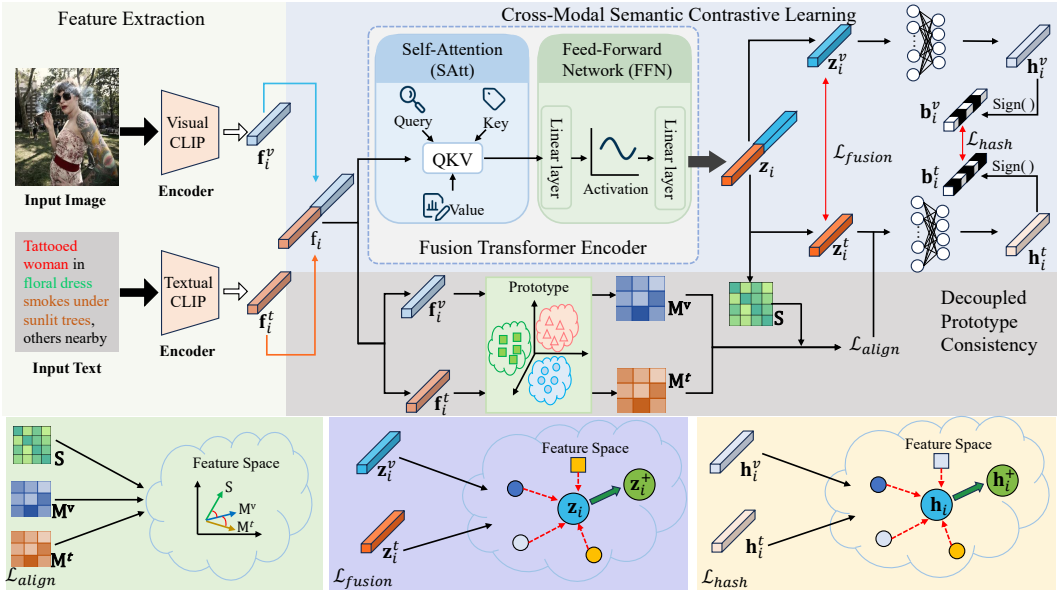


Figure 2: The architecture of the DPCAH method primarily consists of two core components: the cross-modal semantic contrastive learning module, the decoupled prototype consistency module.

3.2.2 Multi-Level Contrastive Loss

To enforce semantic alignment between modalities in an unsupervised manner, we employ contrastive learning at two hierarchical levels: the fusion feature level and the hash code level.

Fusion-Level Contrastive Loss: At the fusion feature level, each image-text pair $(\mathbf{z}_i^v, \mathbf{z}_i^t)$ from the same sample is treated as a positive pair, with all other samples in the mini-batch as negatives. Visual and textual features are normalized and concatenated to form $2m$ embeddings, and the fusion contrastive loss is computed as:

$$\mathcal{L}_{fusion} = -\frac{1}{2m}\sum_{i=1}^{2m}\log\frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_i^+)/\tau)}{\sum_{k=1}^{2m}\mathbb{I}_{[k\neq i]}\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}, \tag{3}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, τ is the temperature coefficient, and m is the batch size.

Hash-Level Contrastive Loss: While the fusion-level loss ensures semantic alignment in the continuous feature space, we also need to maintain cross-modal consistency when mapping features to compact binary hash codes. To achieve this, we introduce modality-specific hash functions implemented by Multi-Layer Perceptrons (MLP) that transform the fused features into continuous hash representations:

$$\mathbf{h}_i^v = \mathcal{H}_v(\mathbf{z}_i^v; \theta_v), \quad \mathbf{h}_i^t = \mathcal{H}_t(\mathbf{z}_i^t; \theta_t), \tag{4}$$

where $\mathbf{h}_i^v, \mathbf{h}_i^t$ are the continuous hash representations, and θ_v, θ_t denote the parameters of the visual and textual hash networks, respectively. We further apply a hash-level contrastive loss:

$$\begin{aligned}\mathcal{L}_{hash}^v &= -\frac{1}{m}\sum_{i=1}^{m}\log\frac{\exp(\langle\mathbf{h}_i^v, \mathbf{h}_i^{t+}\rangle/\tau)}{\sum_{j=1}^{m}\exp(\langle\mathbf{h}_i^v, \mathbf{h}_j^t\rangle/\tau)},\\ \mathcal{L}_{hash}^t &= -\frac{1}{m}\sum_{i=1}^{m}\log\frac{\exp(\langle\mathbf{h}_i^t, \mathbf{h}_i^{v+}\rangle/\tau)}{\sum_{j=1}^{m}\exp(\langle\mathbf{h}_i^t, \mathbf{h}_j^v\rangle/\tau)}.\end{aligned} \tag{5}$$

The total hash contrastive loss combines both directions:

$$\mathcal{L}_{hash} = \mathcal{L}_{hash}^v + \mathcal{L}_{hash}^t. \tag{6}$$

Algorithm 1 Training Algorithm for Proposed Method

Input: Training dataset $\mathcal{D} = \{(\mathbf{x}_i^v, \mathbf{x}_i^t)\}_{i=1}^n$, batch size m, epochs T, learning rates l_{rm} for the transformer encoder, l_{rv} and l_{rt} for the image and text hash functions, l_{rd} for the DPC module, and hyperparameters λ_1, λ_2, and λ_3
Output: Trained parameters θ_e, θ_v, θ_t

```
Initialize θ_e, θ_v, and θ_t with pre-trained CLIP
for e = 1 to T do
    Obtain fused features z_i^v and z_i^t by fusion transformer encoder
    Calculate fusion contrastive loss L_fusion using Eq. (3)
    Obtain hash representations: h_i^v ← H_v(z_i^v; θ_v), h_i^t ← H_t(z_i^t; θ_t) using Eq. (4)
    Calculate hash contrastive loss L_hash using Eq. (6)
    Compute the co-association matrix M* for similarity modeling using Eq. (7)
    Calculate alignment loss L_align using Eq. (9)
    Calculate total loss using Eq. (10)
    Update parameters θ_e, θ_v, and θ_t by backpropagation
end for
return θ_e, θ_v, θ_t
```

3.3 Decoupled Prototype Consistency

While the contrastive learning framework ensures semantic alignment, directly modifying the main feature space may introduce training instability. To address this, we develop a Decoupled Prototype Consistency (DPC) module to jointly capture latent semantic structures and enforce reliable alignment for cross-modal representation learning in an unsupervised setting. The proposed module integrates prototype-level structural modeling and consistency-driven optimization in a decoupled yet complementary manner, thereby enhancing the robustness and discriminability of learned features while mitigating the impact of noise during training[7].

To begin with, the module exploits ensemble clustering to discover prototype representations from heterogeneous modalities. By aggregating clustering results obtained from multiple runs, the model is able to extract more stable and representative semantic structures embedded in deep feature spaces. Specifically, K-means is employed as the base clustering algorithm due to its simplicity and effectiveness. Considering that K-means is sensitive to random initialization, different runs can produce diverse clustering outcomes, which naturally satisfies the diversity requirement of ensemble clustering. In practice, multiple base partitions are constructed with independently initialized cluster centers. The number of clusters is typically bounded by $K = \sqrt{n}$[37] and is randomly selected within the interval $[\lceil \sqrt{n}/2 \rceil, \lfloor \sqrt{n} \rfloor]$, achieving a balance between diversity and computational complexity. Let the set of base partitions be denoted as $\{\pi_1, \pi_2, \ldots, \pi_m\}$, where each partition $\pi_i := \{C_{i1}, C_{i2}, \ldots, C_{iK_i}\}$ consists of K_i clusters.

Based on these base partitions, an intra-modal co-association matrix is constructed to characterize the similarity relationships among samples. This matrix reflects how frequently two instances are assigned to the same cluster across different partitions, thereby providing a robust estimation of semantic similarity. Formally, it is defined as:

$$M^* = \{M_{ij}\}_{i=1,j=1}^n = \frac{1}{R} \sum_{c}^{R} \sum_{k}^{K_R} \psi\left(C_{ck}(f^*(i)), C_{ck}(f^*(j))\right) \quad \text{s.t.} * \in \{v, t\}, \tag{7}$$

where M_{ij} denotes an entry in either the visual similarity matrix M_v or the textual similarity matrix M_t, and C_{ck} represents the k-th cluster in the c-th base partition. The indicator function $\psi(\cdot)$ evaluates whether two feature representations belong to the same cluster:

$$\psi\left(C_{ck}(f_i), C_{ck}(f_j)\right) = \begin{cases} 1, & \text{if } C_{ck}(f_i) = C_{ck}(f_j) \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

The indicator function reflects whether the deep features f_i and f_j from image or text modalities are assigned to the same cluster under C_{ck}. Specifically, $\psi(C_{ck}(f_i), C_{ck}(f_j)) = 1$ if both features

belong to an identical prototype category, and $\psi(C_{ck}(f_i), C_{ck}(f_j)) = 0$ otherwise. As a similarity measure in clustering, the co-association matrix captures how frequently a pair of samples is grouped into the same cluster across multiple clustering results.

On top of the structural modeling, the module further introduces a consistency-driven optimization objective to align similarity patterns obtained from different sources. In unsupervised cross-modal retrieval, a critical challenge lies in noise propagation during the early stages of training, which may lead the model to converge to suboptimal solutions. To address this issue, the proposed formulation enforces a soft alignment between the feature-level similarity matrix and the prototype-based co-association matrices, encouraging consistency across different representations while avoiding overly strict constraints. The corresponding objective function is defined as:

$$\underset{\theta_v, \theta_t, M^v, M^t}{\arg\min} \mathcal{L}_{align} = \left\| S - \mu \frac{(M^v + M^t)}{2} \right\|_F^2, \tag{9}$$

where $0 < \mu < 1$ is a hyperparameter that controls the strength of the alignment. The matrix S is a similarity matrix computed from the fused sample feature vectors Z_i using cosine similarity, which is utilized to guide the optimization of hash codes. By adopting a soft consistency constraint, the model reduces the risk of introducing additional noise that may arise from hard alignment, thereby facilitating stable convergence during optimization.

Overall, by decoupling prototype-based structural learning and consistency-based optimization while allowing them to interact through a moderated alignment mechanism, the proposed module effectively enhances the quality of cross-modal representations. It not only captures comprehensive semantic relationships through ensemble prototypes but also improves the reliability of similarity learning, leading to more accurate and robust cross-modal retrieval performance.

The total training objective combines all three loss components with appropriate weighting factors:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{fusion} + \lambda_2 \mathcal{L}_{hash} + \lambda_3 \mathcal{L}_{align}, \tag{10}$$

where λ_1, λ_2 and λ_3 are weighting hyperparameters that balance the contribution of each loss term.

During inference, the DPC module is completely discarded, and binary hash codes are generated directly from the hash networks via the sign function:

$$\mathbf{b}_i^* = \text{sign}(\mathbf{h}_i^*), \quad \text{s.t.} * \in \{v, t\}, \tag{11}$$

where $\text{sign}(\cdot)$ maps continuous values to $\{-1, 1\}$.

3.4 Theoretical Analysis

We provide a brief analysis of the optimization stability of the proposed Transformer + DPC framework.

Bounded Alignment Updates. The alignment loss is:

$$\mathcal{L}_{align} = \left\| S - \mu \frac{(M^v + M^t)}{2} \right\|_F^2, \tag{12}$$

where S is the cosine similarity matrix and M^v, M^t are co-association matrices. Since $S_{ij} \in [-1, 1]$ and $M_{ij} \in [0, 1]$, we have:

$$\|S - \mu M\|_F \leq n(1 + \mu), \tag{13}$$

implying that the gradient magnitude of $\mathcal{L}_{align}$ is bounded. Therefore, μ controls the update strength and helps stabilize optimization.

Error Reduction via Ensemble Prototypes. Let S^* denote the ideal similarity structure. Then:

$$\|S - S^*\| \leq \|S - \mu M\| + \|\mu M - S^*\|. \tag{14}$$

The first term is minimized during training, while the second term depends on the quality of M. Since M is constructed by ensemble clustering, the prototype-based similarity estimation becomes more robust and reduces approximation error.

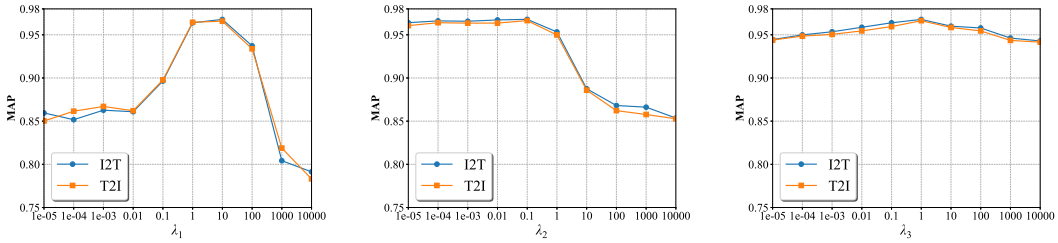
Effect of Transformer Fusion. The Transformer encoder captures global cross-modal interactions and produces smoother similarity structures. Combined with the alignment constraint:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{fusion} + \lambda_2 \mathcal{L}_{hash} + \lambda_3 \mathcal{L}_{align}, \tag{15}$$

the model achieves balanced semantic alignment and regularized similarity learning. Overall, the stability of DPCAH benefits from bounded alignment updates, robust prototype estimation, and Transformer-based feature fusion.

Table 1: The mAP scores for all methods on three datasets. **Bold** denotes the best result, and underlined denotes the second-best result.

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
I → T	DJSRH [20]	0.810	0.843	0.862	0.876	0.724	0.773	0.798	0.817	0.678	0.724	0.743	0.768
	SACH [31]	0.875	0.892	0.916	0.924	0.769	0.808	0.839	0.843	0.778	0.806	0.812	0.824
	SCCGDH [18]	0.882	0.892	0.893	0.911	0.812	0.837	0.840	0.871	0.732	0.777	0.808	0.822
	CIRH [38]	0.901	0.913	0.929	0.937	0.815	0.836	0.854	0.862	0.797	0.819	0.830	0.849
	RICH [8]	0.869	0.875	0.908	0.925	0.790	0.806	0.842	0.852	—	—	—	—
	ESAFH [15]	0.886	0.891	0.931	0.935	0.783	0.809	0.821	0.839	0.761	0.805	0.829	0.859
	UDDH [32]	0.836	0.844	0.899	0.912	0.773	0.791	0.801	0.822	0.789	0.808	0.862	0.894
	RSRH [27]	0.907	0.918	0.933	0.943	0.817	0.836	0.860	0.864	0.871	0.882	0.892	0.925
	DPCAH	**0.926**	**0.955**	**0.957**	**0.968**	**0.853**	**0.884**	**0.892**	**0.895**	**0.889**	**0.917**	**0.935**	**0.937**
T → I	DJSRH [20]	0.786	0.822	0.835	0.847	0.712	0.744	0.771	0.789	0.650	0.753	0.805	0.823
	SACH [31]	0.842	0.878	0.896	0.896	0.739	0.801	0.808	0.814	0.798	0.832	0.861	0.890
	SCCGDH [18]	0.820	0.856	0.872	0.880	0.767	0.773	0.772	0.801	0.774	0.781	0.832	0.878
	CIRH [38]	0.867	0.885	0.900	0.901	0.774	0.803	0.810	0.817	0.811	0.847	0.872	0.895
	RICH [8]	0.830	0.843	0.885	0.902	0.771	0.777	0.802	0.822	—	—	—	—
	ESAFH [15]	0.891	0.895	0.911	0.919	0.791	0.813	0.823	0.836	0.851	0.870	0.879	0.897
	UDDH [32]	0.824	0.835	0.858	0.869	0.769	0.771	0.785	0.802	0.799	0.815	0.847	0.881
	RSRH [27]	0.874	0.887	0.897	0.905	0.784	0.811	0.810	0.820	0.862	0.878	0.881	0.902
	DPCAH	**0.924**	**0.952**	**0.954**	**0.966**	**0.864**	**0.884**	**0.899**	**0.901**	**0.885**	**0.920**	**0.938**	**0.941**



(a) λ_1 (b) λ_2 (c) λ_3

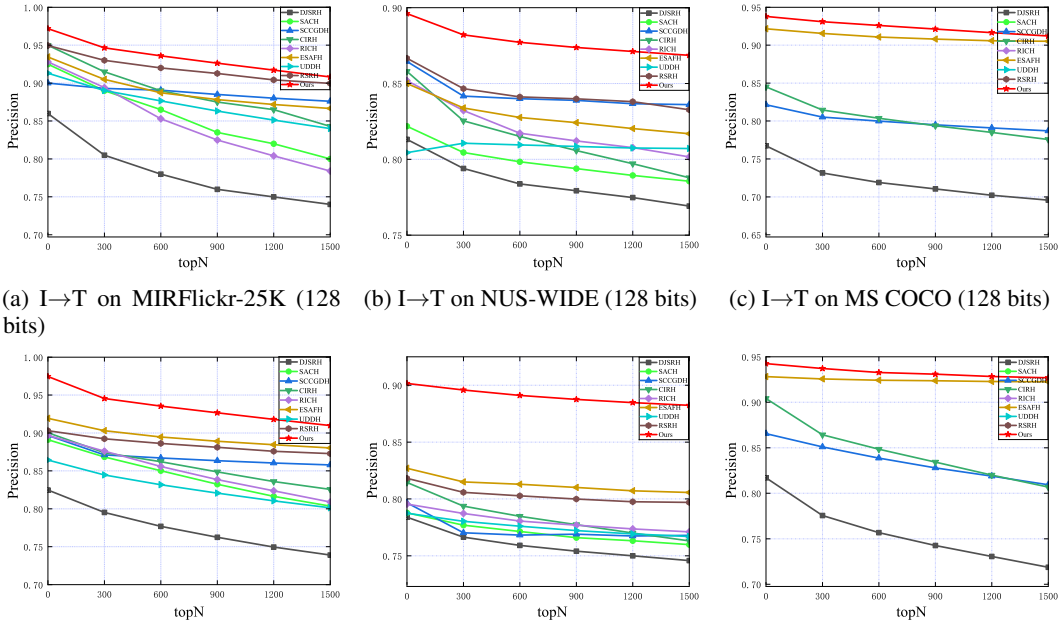
Figure 3: Impact of parameters on I2T and T2I performance with 128-bit hash codes on the MIRFlickr-25K dataset.

4 Experiments

To evaluate the retrieval performance, we adopt mAP@50 as the evaluation metric for both image-to-text (I→T) and text-to-image (T→I) retrieval tasks. Comprehensive comparisons are conducted across different hash code lengths. As shown in Table 1, in the I→T retrieval task, DPCAH consistently outperforms the best-performing baseline by 2.1% to 4.0%, 2.8% to 5.6%, and 1.3% to 4.8% on the MIRFlickr-25K, NUS-WIDE, and MSCOCO datasets, respectively. For the T→I retrieval task, DPCAH achieves performance gains of 3.7% to 6.4%, 7.8% to 9.2%, and 2.7% to 6.5% over the strongest baseline on the same three datasets. Overall, the results demonstrate that DPCAH yields more pronounced improvements in the I→T retrieval setting, indicating its stronger capability in learning discriminative cross-modal representations from visual to textual modalities.

Table 2: Statistics of the experimental datasets.

Dataset	Categories	Training	Query	Retrieval
MIRFlickr-25K	24	5,000	2,000	18,015
NUS-WIDE	10	5,000	2,000	184,577
MS COCO	80	5,000	2,000	121,287



(a) I→T on MIRFlickr-25K (128 bits) (b) I→T on NUS-WIDE (128 bits) (c) I→T on MS COCO (128 bits)

(d) T→I on MIRFlickr-25K (128 bits) (e) T→I on NUS-WIDE (128 bits) (f) T→I on MS COCO (128 bits)

Figure 4: Precision@TopN curves on MIRFlickr-25K, NUS-WIDE, and MS COCO with 128-bit hash codes.

4.1 Datasets

To comprehensively evaluate the effectiveness and robustness of the proposed method DPCAH, we conduct experiments on three widely used cross-modal benchmark datasets, MIRFlickr-25K [6], NUS-WIDE [2], and MS COCO [12], under varying dataset scales, category sizes, and semantic diversities. The detailed dataset statistics are provided in Table 2.

4.2 Implementation Details

All experiments were conducted on a single NVIDIA GeForce RTX 4090 GPU with 24 GB of memory. We pre-extract 512-dimensional visual and textual features using frozen CLIP encoders. These features remain fixed during training, and no gradients are backpropagated to CLIP. The subsequent trainable DPCAH components, including the FTE and modality-specific hash networks, are optimized under the overall objective. Training uses a batch size of 128 for 120 epochs with a random seed of 2025. All trainable components use Adam with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The initial learning rates are 10^{-4} for the FTE and DPC branches and 10^{-3} for both hash networks. The hash-network learning rates are multiplied by 1.2 after 30 and 80 mini-batch updates on MIRFlickr-25K and NUS-WIDE, and by 0.6 after 200 updates on MS COCO; the FTE and DPC learning rates remain fixed. The same feature-extraction protocol is used across datasets, hash-code lengths, and ablation configurations.

The FTE comprises two Transformer encoder layers, each with four attention heads, a 1024-dimensional embedding, and a 2048-dimensional feed-forward hidden layer. Both contrastive losses use $\tau = 0.5$, balancing semantic discrimination and optimization stability. For each modality,

Table 3: The mAP results of the ablation experiments on the three datasets. **Red** denotes the best results.

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
I → T	FTE only	0.566	0.590	0.598	0.607	0.508	0.496	0.506	0.532	0.551	0.542	0.561	0.579
	Con only	0.880	0.886	0.894	0.897	0.786	0.784	0.783	0.803	0.839	0.836	0.862	0.873
	DPC only	0.558	0.577	0.583	0.582	0.497	0.482	0.489	0.504	0.539	0.531	0.530	0.545
	w/o FTE	0.899	0.914	0.918	0.920	0.814	0.827	0.829	0.835	0.858	0.871	0.883	0.888
	w/o Con	0.585	0.618	0.622	0.630	0.536	0.539	0.552	0.564	0.570	0.577	0.582	0.594
	w/o DPC	0.907	0.927	0.933	0.945	0.825	0.841	0.846	0.863	0.870	0.882	0.914	0.922
	DPCAH	**0.926**	**0.955**	**0.957**	**0.968**	**0.853**	**0.884**	**0.892**	**0.895**	**0.889**	**0.917**	**0.935**	**0.937**
T → I	FTE only	0.626	0.640	0.654	0.661	0.465	0.472	0.475	0.517	0.522	0.524	0.550	0.567
	Con only	0.865	0.869	0.881	0.886	0.769	0.768	0.765	0.793	0.833	0.838	0.869	0.878
	DPC only	0.607	0.621	0.625	0.631	0.436	0.436	0.437	0.467	0.508	0.512	0.515	0.532
	w/o FTE	0.885	0.901	0.903	0.911	0.802	0.808	0.813	0.822	0.852	0.873	0.886	0.892
	w/o Con	0.646	0.672	0.676	0.686	0.498	0.512	0.523	0.546	0.541	0.559	0.567	0.581
	w/o DPC	0.904	0.920	0.932	0.941	0.831	0.844	0.851	0.872	0.866	0.885	0.921	0.927
	DPCAH	**0.924**	**0.952**	**0.954**	**0.966**	**0.864**	**0.884**	**0.899**	**0.901**	**0.885**	**0.920**	**0.938**	**0.941**

Table 4: The time cost (seconds) of the DPCAH method and various baseline methods with 128 bits on the MIRFlickr-25K, NUS-WIDE, and MS COCO datasets. The best-performing results are highlighted in **Red**.

Method	Training time (s)			Query time (s)		
	MIRFlickr-25K	NUS-WIDE	MSCOCO	MIRFlickr-25K	NUS-WIDE	MSCOCO
DJSRH [20]	743.68	783.42	935.41	12.67	91.34	85.64
RICH [8]	817.74	861.13	947.37	20.15	112.84	94.3
CIRH [38]	309.86	304.43	377.33	11.35	93.74	72.12
ESAFH [15]	82.69	117.26	101.16	14.22	117.84	241.47
DPCAH	**62.43**	**90.32**	**73.95**	**10.36**	**81.54**	**63.11**

DPC constructs $R = 20$ independently initialized K-means partitions, balancing consensus stability and clustering cost. The cluster number is sampled uniformly from the integers in $[\lceil\sqrt{n}/2\rceil, \lfloor\sqrt{n}\rfloor]$, where n is the number of training samples used for prototype construction. The averaged consensus $(M^v + M^t)/2$ provides the soft target for prototype-consistency regularization. The FTE architecture, temperature, and clustering settings are shared across all datasets and hash-code lengths.

4.3 Parameter Sensitivity Analysis

In the parameter sensitivity experiments, we evaluate the influence of key loss weights by varying λ_1, λ_2, and λ_3 on a logarithmic scale from 10^{-5} to 10^4, with the results shown in Fig. 3. The three loss components exhibit clearly different sensitivity behaviors. The fusion-level contrastive loss weight λ_1 strongly affects retrieval performance and shows high sensitivity, indicating its critical role in cross-modal representation learning. The hash-level contrastive loss weight λ_2 remains relatively stable within a moderate range but leads to performance degradation when overly emphasized. In contrast, the alignment loss weight λ_3 demonstrates strong robustness across a wide range, consistently providing complementary supervision without disrupting contrastive learning. Overall, these results indicate that contrastive objectives dominate performance, while the alignment objective acts as a stable auxiliary component that improves training robustness.

Parameter Sensitivity and Generalization. λ_1 should be relatively large to ensure strong cross-modal alignment, λ_2 should be moderate to avoid over-constraining the hash space, and λ_3 works best with small but non-zero values as a soft regularizer. These trends remain consistent across datasets, indicating low sensitivity to precise tuning. Based on empirical observations, we use

$\lambda_1 \in [10^{-1}, 10^{1}]$, $\lambda_2 \in [10^{-4}, 10^{-1}]$, and $\lambda_3 \in [10^{-2}, 10^{0}]$, which cover the optimal regions in Fig. 3.

4.4 Ablation Experiments

We evaluate all seven non-empty combinations of FTE, multi-level contrastive learning (Con), and DPC. Con denotes the joint use of $\mathcal{L}_{fusion}$ and $\mathcal{L}_{hash}$. All configurations use the same input features, data splits, hash-code lengths, training protocol, and evaluation criteria, with only the specified components enabled. Table 3 reports the results for both retrieval directions on all three datasets. Con achieves the highest performance among the single-component configurations, indicating that explicit multi-level alignment is central to retrieval accuracy. w/o DPC is the strongest pairwise configuration across the reported settings. Adding DPC to this configuration further improves performance, while w/o FTE also consistently outperforms Con alone. The relatively lower performance of w/o Con indicates that feature interaction and prototype consistency do not fully replace the contrastive objectives. Full DPCAH achieves the highest mAP in every reported setting, supporting the complementary roles of the three components.

5 Conclusion

In this paper, we propose DPCAH, a novel framework for cross-modal retrieval that integrates cross-modal feature fusion with a transformer encoder to capture global cross-modal dependencies, contrastive learning to enhance semantic alignment, and a decoupled prototype consistency module to further refine representation consistency across modalities. Extensive experiments on three benchmark datasets demonstrate that DPCAH consistently outperforms existing methods under different hash code lengths. Ablation studies further confirm that each component contributes positively to the overall performance, highlighting their complementary roles in achieving robust and effective cross-modal retrieval. These results indicate that DPCAH offers a reliable solution for bridging visual and textual modalities in large-scale retrieval tasks.

6 Acknowledgment

This work was supported by the Natural Science Foundation of Hunan Province (2025JJ40057).

References

[1] Shuai Cheng, Lin Wang, Xiaoshuai Hao, Wanqian Zhang, Xiaohua Chen, and Wei Wang. Multi-granularity based collaborative learning for semi-supervised hashing. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2025.

[2] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. 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*, pages 1–9, 2009.

[3] Haozhi Fan and Yuan Cao. Vision-guided text mining for unsupervised cross-modal hashing with community similarity quantization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 2843–2851, 2025.

[4] Tuan Hoang, Thanh-Toan Do, Tam V. Nguyen, and Ngai-Man Cheung. Multimodal mutual information maximization: A novel approach for unsupervised deep cross-modal hashing. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):6289–6302, 2023.

[5] Cheng Huang, Wenzhe Liu, Jinghua Wang, Jinrong Cui, and Jie Wen. Dual-driven cross-modal contrastive hashing retrieval network via structural feature and semantic information. *Information Fusion*, 123:103252, 2025.

[6] Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 39–43, 2008.

[7] L.I. Kuncheva and D.P. Vetrov. Evaluation of stability of k-means cluster ensembles with respect to random initialization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1798–1808, 2006.

[8] Bo Li, Dan Yao, and Zhixin Li. Rich: A rapid method for image-text cross-modal hash retrieval. *Displays*, 79:102489, 2023.

[9] Fengling Li, Yang Sun, Tianshi Wang, Lei Zhu, and Xiaojun Chang. Fast partial-modal online cross-modal hashing. *IEEE Transactions on Image Processing*, 34:4440–4455, 2025.

[10] Jiaxing Li, Wai Keung Wong, Lin Jiang, Kaihang Jiang, Xiaozhao Fang, Shengli Xie, and Jie Wen. Collaboratively semantic alignment and metric learning for cross-modal hashing. *IEEE Transactions on Knowledge and Data Engineering*, 37(5):2311–2328, 2025.

[11] Huilong Lin, Yangtao Wang, Meie Fang, Yanzhao Xie, Da Chen, Xiaocui Li, Weilong Peng, Siyuan Chen, Maobin Tang, and Ping Li. Enhancing cross-modal semantic consistency via key token alignment for image-text retrieval. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2025.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C.L. Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, volume 8693, pages 740–755, 2014.

[13] Xiaoqing Liu, Zhiwen Yu, Kaixiang Yang, Jun Yu, Huanqiang Zeng, and CL Philip Chen. Enhanced cross-modal hashing via hybrid distillation and structural refinement. *IEEE Transactions on Image Processing*, 2025.

[14] Xingbo Liu, Jiamin Li, Xiushan Nie, Xuening Zhang, and Yilong Yin. Fast unsupervised cross-modal hashing with robust factorization and dual projection. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(12):1–21, 2024.

[15] Li Mingyong and Ge Mingyuan. Enhanced-similarity attention fusion for unsupervised cross-modal hashing retrieval. *Data Science and Engineering*, 10(2):258–276, 2025.

[16] Shu-Juan Peng, Jinhan Yi, Xin Liu, Yiu ming Cheung, Zhen Cui, and Taihao Li. Olch: Online label consistent hashing for streaming cross-modal retrieval. *Pattern Recognition*, 150:110335, 2024.

[17] Qibing Qin, Lei Wu, Wenfeng Zhang, Lei Huang, and Jie Nie. Deep semantic-consistent penalizing hashing for cross-modal retrieval. *IEEE Transactions on Multimedia*, 27:4613–4626, 2025.

[18] Zhenqiu Shu, Yibing Bai, Donglin Zhang, Jun Yu, Zhengtao Yu, and Xiao-Jun Wu. Specific class center guided deep hashing for cross-modal retrieval. *Information Sciences*, 609:304–318, 2022.

[19] Zhenqiu Shu, Julong Zhang, and Zhengtao Yu. Adaptive centroid guided hashing for cross-modal retrieval. *Pattern Recognition*, 176:113186, 2026.

[20] S. Su, Z. Zhong, and C. Zhang. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3027–3035, 2019.

[21] Yuan Sun, Jian Dai, Zhenwen Ren, Yingke Chen, Dezhong Peng, and Peng Hu. Dual self-paced cross-modal hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15184–15192, 2024.

[22] Junfeng Tu, Xueliang Liu, Zhen Huang, Yanbin Hao, Richang Hong, and Meng Wang. Cross-modal hashing via diverse instances matching. *IEEE Transactions on Image Processing*, 34:2737–2749, 2025.

[23] Rong-Cheng Tu, Xian-Ling Mao, Qinghong Lin, Wenjin Ji, Weize Qin, Wei Wei, and Heyan Huang. Unsupervised cross-modal hashing via semantic text mining. *IEEE Transactions on Multimedia*, 25:8946–8957, 2023.

[24] Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. Cross-modal retrieval: A systematic review of methods and future directions. *Proceedings of the IEEE*, 112(11):1716–1754, 2024.

[25] Fei Wu, Shuaishuai Li, Guangwei Gao, Yimu Ji, Xiao-Yuan Jing, and Zhiguo Wan. Semi-supervised cross-modal hashing via modality-specific and cross-modal graph convolutional networks. *Pattern Recognition*, 136:109211, 2023.

[26] Qingpeng Wu, Zheng Zhang, Yishu Liu, Jingyi Zhang, and Liqiang Nie. Contrastive multi-bit collaborative learning for deep cross-modal hashing. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):5835–5848, 2024.

[27] You Wu, Bo Li, and Zhixin Li. Revising similarity relationship hashing for unsupervised cross-modal retrieval. *Neurocomputing*, 614:128844, 2025.

[28] Xinyu Xia, Guohua Dong, Fengling Li, Lei Zhu, and Xiaomin Ying. When clip meets cross-modal hashing retrieval: A new strong baseline. *Information Fusion*, 100:101968, 2023.

[29] Le Xu and Jun Yin. Dual aggregation based joint-modal similarity hashing for cross-modal retrieval. *Neural Networks*, page 108069, 2025.

[30] Dan Yao, Zhixin Li, Bo Li, Canlong Zhang, and Huifang Ma. Similarity graph-correlation reconstruction network for unsupervised cross-modal hashing. *Expert Systems with Applications*, 237:121516, 2024.

[31] Heng Yu, Shuyan Ding, Lunbo Li, and Jiexin Wu. Self-attentive clip hashing for unsupervised cross-modal retrieval. In *Proceedings of the 4th ACM International Conference on Multimedia in Asia*, pages 1–7, 2022.

[32] Bin Zhang, Yue Zhang, Junyu Li, Jiazhou Chen, Tatsuya Akutsu, Yiu-Ming Cheung, and Hongmin Cai. Unsupervised dual deep hashing with semantic-index and content-code for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1):387–399, 2025.

[33] Donglin Zhang, Zhikai Hu, Xiao-Jun Wu, and Josef Kittler. Rsph: Robust self-paced hashing for cross-modal retrieval. *Pattern Recognition*, 171:112072, 2026.

[34] Huaiwen Zhang, Yang Yang, Fan Qi, Shengsheng Qian, and Changsheng Xu. Active supervised cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[35] Qian Zhang, Jiarui Zhao, Xiao-Qian Liu, Yu-Wei Zhan, Zhen-Duo Chen, Xin Luo, and Xin-Shun Xu. Hypergraph-based clip hashing for unsupervised cross-modal retrieval. *Knowledge-Based Systems*, page 114508, 2025.

[36] Huan Zhao, Zeyi Li, Song Wang, Zixing Zhang, and Keqin Li. Robust hashing with bilinear drift for image-text retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2025.

[37] Caiming Zhong, Xiaodong Yue, Zehua Zhang, and Jingsheng Lei. A clustering ensemble: Two-level-refined co-association matrix with path-based transformation. *Pattern Recognition*, 48(8):2699–2709, 2015.

[38] Lei Zhu, Xize Wu, Jingjing Li, Zheng Zhang, Weili Guan, and Heng Tao Shen. Work together: Correlation-identity reconstruction hashing for unsupervised cross-modal retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):8838–8851, 2022.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification:The abstract and introduction (Sec. 1) clearly state the paper's contributions: proposing a cross-modal semantic contrastive learning module and a decoupled prototype consistency module for unsupervised cross-modal hashing. All claims are directly supported by the experimental results and ablation studies in Sec. 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification:The limitations of this work are explicitly discussed. First, the method relies on pre-trained CLIP features, which may limit adaptability to domains with significant distribution shifts. Second, experiments are conducted on three benchmark datasets (MIRFlickr-25K, NUS-WIDE, MS COCO), and generalization to other modalities or large-scale real-world systems remains to be further validated. Third, the theoretical analysis mainly provides insights into optimization stability rather than formal convergence guarantees. Future work will explore domain adaptation, broader datasets, and more rigorous theoretical analysis.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper provides theoretical analysis in Sec. 3.5, including bounded gradient analysis of the alignment loss and error decomposition of the prototype-based similarity estimation. The assumptions (e.g., bounded cosine similarity and co-association values) are explicitly stated, and the analysis provides theoretical insights into the optimization stability of the proposed method.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All datasets used (MIRFlickr-25K, NUS-WIDE, MS COCO) are clearly specified with statistics in Table 1. The full architecture is detailed in Sec. 3, the training algorithm is provided as pseudocode in Algorithm 1, and all loss functions are formulated (Eq. 3-10). In addition, parameter sensitivity and generalization analysis (Fig. 4)provides practical guidance for hyperparameter selection, ensuring sufficient information for independent implementation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The datasets used in this work are publicly available benchmark datasets, including MIRFlickr-25K, NUS-WIDE, and MS COCO. The source code, training scripts, and implementation details for the proposed method are provided in the supplementary material and will be publicly released upon publication. The submission includes sufficient instructions for reproducing the main experimental results, including data preprocessing, model configuration, training procedures, and evaluation protocols. In addition, the paper provides detailed architectural descriptions (Sec. 3), pseudocode (Algorithm 1), and explicit loss formulations to facilitate reproducibility.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides detailed experimental settings, including dataset splits (Table 1), model architecture (Sec. 3), training procedure (Algorithm 1), and loss formulations (Eq. 3–10). In addition, parameter sensitivity analysis (Sec. 4.1) provides guidance for hyperparameter selection. Full training details such as optimizer type, learning rates,

and training epochs will be included in the supplementary material to ensure complete reproducibility.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Following standard protocol in deep cross-modal hashing literature, the main results (Table 2, Table 1) report standard mAP scores on fixed public splits, consistent with prior work. Due to the computational cost of training deep models, statistical significance tests or error bars over multiple random runs are not reported, but the comparison is made fairly under identical dataset conditions.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports both training and query time costs (in seconds) for the proposed method and key baselines in Table 4, offering an empirical estimate of the computational resources and efficiency required.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper focuses on multimedia retrieval. It uses standard, publicly available benchmark datasets, does not involve human subjects, and does not release data/models that pose immediate high-stakes risks. The authors have reviewed and adhere to the Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This work contributes to efficient cross-modal retrieval, which can benefit applications such as multimedia search, recommendation systems, and information organization. However, potential risks include misuse in surveillance or large-scale content filtering. These risks are not unique to this work but are inherent to retrieval technologies. Future work may explore fairness-aware and privacy-preserving retrieval mechanisms.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper offers an algorithmic framework for generating binary hash codes for retrieval, not a generative model or a scraped dataset. The released artifacts (algorithm/code) do not pose a direct high risk for misuse nor require specific safeguarding measures.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All three datasets are correctly cited in the references: MIRFlickr-25K [6], NUS-WIDE [2], and MS COCO [12]. These are standard public research benchmarks, and their usage adheres to their respective research licenses.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not introduce new datasets or pre-trained model checkpoints at this stage. The novel architecture is fully described in Sec. 3, Algorithm 1, and associated figures. If code is released in the future, it will be accompanied by usage instructions and a license file.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper exclusively uses standard benchmark datasets and does not involve any crowdsourcing tasks, human subjects experiments, or participant data collection.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification:This computational study uses only existing public data and does not involve human subject research, thus IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Large Language Models are not utilized as a component of the proposed cross-modal hashing architecture or any original methodological step in this research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.