UDCH: Unsupervised Dynamic Weighted Cluster-cooperative Hashing for Cross-modal Retrieval

Yuanzhi Zhao, Fan Yang*, Yudong Zhao, Xiaoyu Li

School of Computer and Artificial Intelligence, Nanjing University of Finance and Economics, Nanjing, 210023, China
1120241393@stu.nufe.edu.cn, nufe_yf@163.com, 1120241370@stu.nufe.edu.cn, lixxyu1017@163.com

Abstract

In cross-modal retrieval tasks, unsupervised hash code learning still faces key challenges, including the difficulty of modeling shared semantic structures across modalities and the inability to adaptively balance multiple supervision objectives during optimization. To address these issues, we propose a novel Unsupervised Dynamic Weighted Cluster-Cooperative Hashing (UDCH) framework, which jointly models feature-level alignment and cluster-level semantic structure to guide consistency learning across modalities under label-free conditions. Specifically, we design an instance-level contrastive loss in the feature branch to align the embedding spaces of images and texts, while employing K-Means clustering to generate pseudo-labels and construct a cluster-center contrast mechanism that captures semantic grouping information. Furthermore, we integrate cross-modal feature similarity to construct a high-order structure matrix, enabling fine-grained structural supervision. To enhance the synergy of multi-objective optimization, we introduce a dynamic weighting strategy that adaptively adjusts the contributions of the feature and cluster branches based on the degree of modal alignment and semantic compactness. Extensive experiments on multiple cross-modal retrieval benchmarks demonstrate that UDCH achieves superior semantic alignment and retrieval performance under unsupervised settings, validating the effectiveness of multi-level semantic modeling and adaptive collaboration mechanisms in unsupervised hashing tasks.

Introduction

Cross-modal hashing (CMH) has emerged as a vital technique for large-scale multimedia retrieval due to its outstanding efficiency in storage and retrieval speed (Qin et al. 2025a). Although recent deep CMH methods have demonstrated remarkable progress (Yang et al. 2024; Qin et al. 2025b), they typically rely heavily on abundant annotated data, limiting their real-world applicability (Liu et al. 2024). Hence, effective unsupervised CMH techniques without explicit supervision remain critically important.

Despite recent advancements, existing unsupervised CMH approaches still face several major challenges that limit performance and generalization. First, many methods overly emphasize instance-level contrastive learning while ignoring global semantic patterns and higher-order relationships among samples (Li, Long, and Yang 2025; Hu et al. 2023), which leads to modality-specific inconsistencies and semantic drift, undermining the quality of the learned hash codes. Second, although two-stage discrete optimization and correlation–identity reconstruction can improve discrete code quality and correlation preservation (Zhang et al. 2023; Zhu et al. 2023), explicit exploitation of clustering structures to supervise semantic learning remains underused (Zhang et al. 2025); without cluster-level semantic regularization, models struggle to maintain intra-class compactness and inter-class separability (Chen et al. 2025; Yang et al. 2025), especially in heterogeneous multimodal spaces (Huang et al. 2024; Tu et al. 2024). Third, most existing frameworks rely on static loss-weighting strategies and lack adaptive optimization mechanisms (Yang et al. 2023; Xie et al. 2024); in practice, the relative importance of instance-level alignment versus structure-level modeling varies throughout training, and failing to rebalance these objectives often causes unstable convergence or suboptimal modality fusion (Pu et al. 2025; Li, Long, and Yang 2025).

To address these limitations, we propose UDCH, a unified framework that integrates hash-guided instance-level contrastive learning with cluster-aware semantic modeling via K-means, builds a high-order semantic matrix through second-order propagation, and employs a gated fusion mechanism with a dynamic weighting strategy to align global and local cues, enhancing optimization robustness and adaptability.

The main contributions of our work are as follows:

- We propose a hash-guided instance-level contrastive learning module to capture fine-grained alignment across modalities, laying a discriminative semantic foundation without any label supervision.
- We design a dual-level cluster-aware semantic modeling scheme, in which both inter-cluster separability and intra-cluster consistency are enforced through center-based and pair-based contrastive objectives, leading to more coherent structure-guided hash codes.
- We further introduce a dynamic weighting strategy based on pseudo-label consistency and cross-modal divergence, which adaptively balances feature-level and cluster-level learning during training.

*Corresponding author: Fan Yang

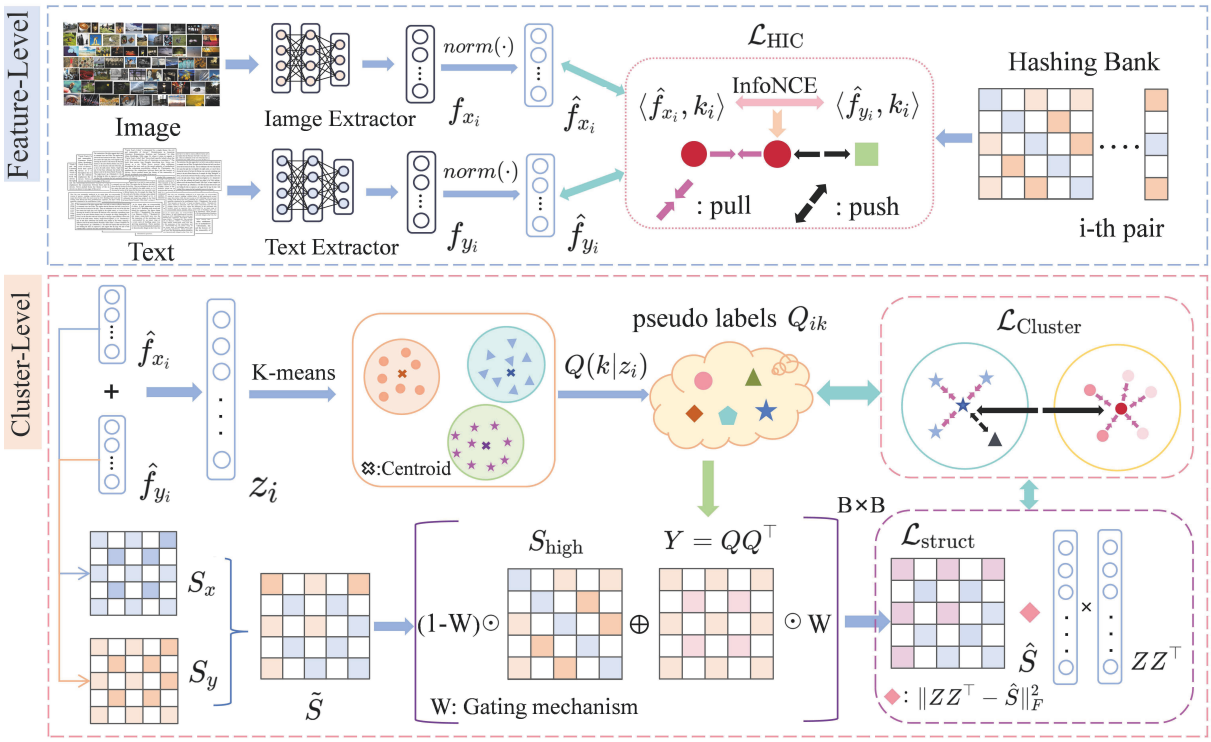


Figure 1: The overall structure of the proposed unsupervised cross-modal hashing framework. First, image and text features are semantically aligned through instance-level contrastive learning, providing a stable foundation for subsequent clustering. Next, coarse-grained and fine-grained structural supervision signals are constructed through the pairing of cluster centers and cross-modal pairs. Finally, the fused high-order semantic matrix guides the hash encoding to retain a more consistent semantic topological structure.

- Experiments on three public benchmarks confirm that UDCH achieves state-of-the-art performance in cross-modal retrieval with consistent gains under varying hash lengths.

Methodology

Problem Definition

Given a cross-modal training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^{d_x}$ and $y_i \in \mathbb{R}^{d_y}$ denote the image and text feature representations of the i-th sample pair respectively, our goal is to learn unified binary hash codes $b_{x_i}, b_{y_i} \in \{-1, +1\}^r$, where r is the length of hash code.

The learned hash codes are expected to preserve both inter-modal semantic similarity and intra-modal structural consistency. To model the rich semantic structure across modalities, we first extract modality-specific feature embeddings f_{x_i}, f_{y_i}, and then apply K-means clustering on the concatenated features to obtain soft pseudo-labels $Q_i \in [0, 1]^K$ and cross-modal cluster centers $c \in \mathbb{R}^{K \times r}$. Here, K is the predefined number of clusters, and each cluster center c_k serves as a semantic anchor in the Hamming space.

Feature-level Contrastive Multi-modal Hash Learning

Recently, contrastive learning has demonstrated strong potential in unsupervised representation learning, particularly in capturing fine-grained instance-level semantics. Motivated by this, we introduce a hash-aware instance-level contrastive learning strategy to align image and text features within a shared embedding space. This alignment not only provides semantic stability for subsequent cluster-level modeling but also enables fine-grained feature-level supervision for cross-modal hashing. First, two modality-specific encoders are employed to map the image modality x_i and the text modality y_i into a unified continuous feature space with dimension r. The normalized feature representations are computed as:

$$\hat{f}_{x_i} = \text{norm}(f_{x_i}), \quad \hat{f}_{y_i} = \text{norm}(f_{y_i}) \ . \tag{1}$$

To capture instance-level semantic similarity without relying on explicit class labels, we adopt the InfoNCE loss to optimize the discriminative ability of hash codes. For a given query $q \in \{f_{x_i}, f_{y_i}\}$, the probability of retrieving its corresponding positive hash key k_i is defined as:

$$p_i(q) = \frac{\exp(\langle q, k_i \rangle / \tau)}{\exp(\langle q, k_i \rangle / \tau) + \sum_{j \in \mathcal{N}_i} \exp(\langle q, k_j \rangle / \tau)}, \tag{2}$$

Here, the numerator $\exp(\langle q, k_i \rangle / \tau)$ measures the semantic similarity between the query q and its positive key k_i (with $\langle \cdot, \cdot \rangle$ denoting cosine similarity); the denominator aggregates both the similarity to the positive sample k_i and all negatives k_j, and normalizes the distribution via a Softmax function. $\tau > 0$ is the temperature scaling factor, and $\mathcal{N}_i$ denotes the set of negative keys. The contrastive loss $\mathcal{L}_{\text{HIC}}$ is designed to minimize the expectation of the log-probability for positive pairs while suppressing that of negative pairs. It is defined as:

$$\mathcal{L}_{\text{HIC}} = \frac{1}{B} \sum_{i=1}^{B} \left[-\log p_i(\hat{f}_{x_i}) - \log p_i(\hat{f}_{y_i}) \right], \quad (3)$$

where B denotes the batch size. The terms $-\log p_i(\hat{f}_{x_i})$ and $-\log p_i(\hat{f}_{y_i})$ respectively optimize the image and text modalities by maximizing the similarity between image and text embeddings and their corresponding hash codes, while simultaneously minimizing their similarity to negative hash keys.

Cluster-aware Semantic Alignment

While the feature-level contrastive loss promotes instance-wise alignment, it is limited in capturing global semantic structures and is vulnerable to noise. To overcome this, we propose a Cluster-driven Semantic Learning module to enhance intra-cluster compactness and inter-cluster separability, and a cross-modal clustering-aware contrastive loss to align semantic structures across modalities.

We first define the shared semantic representation and modality-specific representations for each pair as:

$$z_i = \frac{1}{2} \left(\hat{f}_{x_i} + \hat{f}_{y_i} \right). \quad (4)$$

and we perform K-means clustering on the representations $\{z_i\}_{i=1}^{B}$ to obtain hard pseudo-labels $Q(k|z_i) \in [0, 1]$, where 1 indicates that sample i belongs to cluster k. Based on the assignments, we define the shared cluster center c_k and the modality-specific cluster center $c_k^{(v)}$, both normalized by ℓ_2 norm:

$$c_k = \frac{\sum_{i=1}^{B} Q(k|z_i)\, z_i}{\left\| \sum_{i=1}^{B} Q(k|z_i)\, z_i \right\|_2}, \quad (5)$$

$$c_k^{(v)} = \frac{\sum_{i=1}^{B} Q(k|z_i)\, z_i^{(v)}}{\left\| \sum_{i=1}^{B} Q(k|z_i)\, z_i^{(v)} \right\|_2}. \quad (6)$$

To effectively capture the potential structural information in cross-modal data, we first introduce a structure-guided contrastive loss based on cluster centers. This method uses each cluster center as an anchor point and encourages contrastive learning between multiple perspective encodings within a cluster, promoting higher similarity among representations within the same cluster while increasing the distance between representations of different cluster centers:

$$\begin{aligned} \mathcal{L}_{\text{CSL}} &= \frac{1}{K} \sum_{v=1}^{V} \sum_{k=1}^{K} -\log \frac{a_k^{(v)}}{a_k^{(v)} + \sum_{j \neq k}^{K} a_j}, \\ a_k^{(v)} &= \exp\left(\frac{s(c_k, c_k^{(v)})}{\tau_c} \right), \\ a_j &= \exp\left(\frac{s(c_k, c_j)}{\tau_c} \right), \end{aligned} \quad (7)$$

where the first term pulls together the shared and modality-specific centers within the same cluster, and the second term pushes apart centers from different clusters via softmax-based contrast.

We further introduce a cluster-aware contrastive loss that focuses on modeling intra-cluster semantic alignment across modalities. Specifically, we construct a cross-modal contrastive objective based on each image-text pair, aiming to maximize their similarity if they belong to the same cluster, while contrasting against other non-matching texts within the same cluster. The proposed loss is formulated as:

$$\mathcal{L}_{\text{CSL}}^{\text{cross}} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\sum_{j=1}^{B} \mathbb{I}[Q_i = Q_j] \cdot \exp\left(\frac{s(\hat{f}_{x_i}, \hat{f}_{y_j})}{\beta} \right)}{\sum_{j=1}^{B} \exp\left(\frac{s(\hat{f}_{x_i}, \hat{f}_{y_j})}{\beta} \right)}, \quad (8)$$

Here, β denotes the temperature hyperparameter used to control the smoothness of the softmax distribution. The indicator function $\mathbb{I}[Q_i = Q_j]$ equals 1 if the image-text pair (i, j) shares the same cluster label, and 0 otherwise. The optimization objective of this loss function lies in two key aspects: On one hand, when $(\hat{f}_{x_i}, \hat{f}_{y_j})$ are assigned to the same semantic cluster, the model encourages them to be close in the shared embedding space, i.e., maximizing the similarity of matched pairs within the same cluster. On the other hand, for samples from different clusters, the objective pushes them apart in the embedding space to form well-separated cluster structures.

To comprehensively optimize both intra-cluster consistency and inter-modal semantic alignment, we introduce a unified total objective that combines the cluster-level semantic loss and the cross-modal clustering-aware contrastive loss:

$$\mathcal{L}_{\text{Cluster}} = \lambda_1 \cdot \mathcal{L}_{\text{CSL}} + \lambda_2 \cdot \mathcal{L}_{\text{CSL}}^{\text{cross}}, \quad (9)$$

Here, λ_1 and λ_2 are two balancing weights that control the importance of intra-cluster discrimination and cross-modal alignment respectively. By jointly optimizing the above total loss, our method encourages both fine-grained semantic structure learning within clusters and robust semantic matching across modalities.

Constructing Advanced Multi-level Semantic Matrix

Although instance-level contrastive learning can align the local semantics of image-text samples in the embedding space, this mechanism only models semantic constraints at

the level of positive and negative sample pairs, making it difficult to explicitly capture higher-order structural relationships between cross-modal samples.

To further enhance fine-grained semantic correlation modeling in unsupervised hashing, we first construct intra-modal similarity matrices for the image and text modalities separately to quantify local semantic similarity among samples. Specifically, the image intra-modal similarity matrix $S_x \in \mathbb{R}^{B\times B}$ is computed as the cosine similarity of normalized embeddings between every pair of samples (i, j) in a batch, formulated as:

$$S_x = \langle \hat{f}_{x_i}, \hat{f}_{x_j} \rangle, \quad S_y = \langle \hat{f}_{y_i}, \hat{f}_{y_j} \rangle, \tag{10}$$

Here, $\langle \cdot, \cdot \rangle$ denotes the cosine similarity, and all values are bounded in the range $[-1, 1]$. To further fuse structural information from both modalities, we introduce a tunable weighting factor $\gamma \in [0, 1]$ and *fuse* the two matrices via a convex combination to form a unified similarity structure:

$$\widetilde{S} = \gamma S_x + (1 - \gamma)\, S_y, \tag{11}$$

This operation allows flexible balancing of intra-modal and inter-modal structure contributions, thereby facilitating unified modeling of semantic relations across modalities.

Based on the fused intra-modal and inter-modal similarity matrix $\widetilde{S} \in \mathbb{R}^{B\times B}$, we further introduce a second-order propagation strategy to construct the high-order semantic structure matrix. This aims to enhance the structural connectivity between cross-modal samples. Let $\eta \in [0, 1]$ be the propagation weight coefficient and B the batch size. The final high-order mixed similarity matrix is defined as:

$$S_{\text{high}} = (1 - \eta)\, \widetilde{S} + \frac{\eta}{B}\, \widetilde{S}\, \widetilde{S}^{\top}, \tag{12}$$

In this formulation, the first term $(1 - \eta)\widetilde{S}$ preserves the original first-order structural relations, while the second term $\frac{\eta}{B}\widetilde{S}\widetilde{S}^{\top}$ captures the reinforced high-order semantics via matrix self-propagation. This facilitates second-hop semantic connection modeling between samples across modalities. The propagation mechanism not only unifies intra-modal and inter-modal structure modeling but also promotes global coherence and fine-grained semantic relation construction.

To enhance the structural expressiveness of pseudo-labels during supervision, we first utilize soft cluster assignments to construct a global coarse-grained semantic relation matrix $Y \in [0, 1]^{B\times B}$, which captures global structural similarities between samples. Specifically, the pseudo label matrix Y is computed as:

$$Y = QQ^{\top}, \quad Q_{ik} = p(k \mid z_i), \tag{13}$$

where $Q_{ik} \in [0, 1]^{B\times K}$ represents the probability distribution of sample i over K clusters, providing a coarse-grained view of global semantics.

However, relying solely on such coarse semantic information makes it difficult to capture fine-grained semantic pathways between samples. Therefore, we further construct a high-level semantic structure matrix $\hat{S} \in (0, 1)^{B\times B}$, which models the fine-grained semantic similarity structure in the embedding space across modalities.

To merge the coarse-grained cluster relation Y with the fine-grained second-order similarity S_{high}, we adopt an element-wise gated fusion: we compute an adaptive weight matrix $W = \sigma(Y - S_{\text{high}}) \in (0, 1)^{B\times B}$. Interpreting W as a measure of local agreement, $W \to 1$ favors the cluster structure Y, $W \to 0$ favors the propagated fine-grained structure S_{high}, and $W \to 0.5$ performs a smooth compromise between the two. The final structure-aware supervision matrix is:

$$\hat{S} = W \odot Y + (1 - W) \odot S_{\text{high}}, \tag{14}$$

Here, $\odot$ denotes the element-wise (Hadamard) product. This fusion strategy achieves a semantic balance between coarse and fine structures: it preserves confident structure from Y and refines uncertain regions through S_{high}, thereby improving both the expressiveness and adaptability of the learned supervision signal.

To preserve semantic consistency between the learned representations and the fused supervision matrix $\hat{S}$, we propose a structure-aware regression objective $\mathcal{L}_{\text{struct}}$ to guide semantic layout modeling in the shared space.

Specifically, based on the previously defined consensus embeddings $z_i = \frac{1}{2}(\hat{f}_{x_i} + \hat{f}_{y_i})$, we collect all paired embeddings from image and text into a matrix:

$$Z = \left[z_1^{\top}, z_2^{\top}, \ldots, z_B^{\top}\right]^{\top} \in \mathbb{R}^{B\times r}. \tag{15}$$

We minimize the Frobenius norm as follows:

$$\mathcal{L}_{\text{struct}} = \left\| ZZ^{\top} - \hat{S} \right\|_F^2, \tag{16}$$

This regression objective explicitly enforces the semantic alignment between the modality-consensus embeddings and the final fused structure supervision matrix $\hat{S}$. By minimizing their discrepancy in the Frobenius norm, this loss ensures that the learned embedding similarity structure conforms to the refined semantic topology encoded in $\hat{S}$, thereby enhancing semantic discriminability and structure-awareness in cross-modal representation learning.

Dynamic Weighting Strategy for Loss Balancing

To address the varying contributions of feature-level and cluster-level objectives during training, we design a dual dynamic weighting strategy. Specifically, we assign weights based on two criteria: the consistency between current cross-modal pseudo labels and the divergence of modality-specific feature distributions. We also apply exponential moving average (EMA) to enhance stability.

We define the semantic consistency using the Normalized Mutual Information (NMI) between current image and text pseudo labels:

$$\text{NMI}(Y^x, Y^y) = \frac{2\, I(Y^x, Y^y)}{H(Y^x) + H(Y^y)} \in [0, 1], \tag{17}$$

where $Y^x, Y^y \in \{0, 1\}^{B\times K}$ represent pseudo cluster labels for the image and text modalities, respectively. $I(\cdot, \cdot)$ denotes mutual information and $H(\cdot)$ is the entropy.

Task	Method	MIRFLICKR-25K				NUS-WIDE				MS COCO			
		16-bit	32-bit	64-bit	128-bit	16-bit	32-bit	64-bit	128-bit	16-bit	32-bit	64-bit	128-bit
I→T	DJSRH	0.7878	0.8180	0.8320	0.8569	0.7110	0.7535	0.7811	0.8001	0.6714	0.7371	0.7565	0.7821
	JDSH	0.8253	0.8577	0.8731	0.8740	0.7400	0.7952	0.7983	0.8186	0.6792	0.7263	0.7600	0.7842
	AGCH	0.8554	0.8880	0.8892	0.9079	0.7816	0.8251	0.8298	0.8520	0.7433	0.7692	0.7699	0.8072
	CIRH	0.8666	0.8850	0.9011	0.9115	0.7838	0.8110	0.8272	0.8389	0.7452	0.7933	0.8151	0.8235
	SCH	0.8430	0.8808	0.9052	**0.9176**	0.7931	0.8251	0.8472	0.8540	0.6451	0.7568	0.8000	0.8180
	VLKD	0.8848	0.8847	0.9022	0.9073	0.7948	0.8088	0.8489	0.8435	0.7824	0.8232	0.8259	0.8249
	UDCH	**0.8930**	**0.9025**	**0.9054**	0.9134	**0.8569**	**0.8773**	**0.8838**	**0.8866**	**0.8561**	**0.8860**	**0.9011**	**0.9135**
T→I	DJSRH	0.7575	0.7950	0.8132	0.8183	0.7281	0.7439	0.7568	0.7721	0.7011	0.7674	0.7933	0.8160
	JDSH	0.8000	0.8165	0.8249	0.8574	0.7121	0.7320	0.7750	0.7776	0.7009	0.7577	0.7908	0.8218
	AGCH	0.8289	0.8344	0.8447	0.8699	0.7384	0.7674	0.7851	0.7901	0.7138	0.7496	0.7897	0.8173
	CIRH	0.8420	0.8642	0.8769	0.8810	0.7584	0.7736	0.7912	0.8048	0.7439	0.8228	0.8495	0.8512
	SCH	0.7624	0.8181	0.8487	0.8578	0.7568	0.7759	0.7901	0.8090	0.6026	0.7173	0.7512	0.7712
	VLKD	0.8143	0.8454	0.8465	0.8603	0.7445	0.7448	0.7719	0.7762	0.7496	0.8053	0.8172	0.8238
	UDCH	**0.8843**	**0.8964**	**0.9025**	**0.9085**	**0.8306**	**0.8460**	**0.8551**	**0.8582**	**0.8555**	**0.8852**	**0.9005**	**0.9115**

Table 1: mAP results of all methods on MIRFLICKR-25K, NUS-WIDE, and MS COCO across different bit lengths.

The feature divergence is measured via the Maximum Mean Discrepancy (MMD) between cross-modal features:

$$\begin{aligned}\mathrm{MMD}^2(f^x, f^y) = \mathbb{E}[g(f^x, f^x)] + \mathbb{E}[g(f^y, f^y)] \\ - \ 2\,\mathbb{E}[g(f^x, f^y)]\,,\end{aligned} \tag{18}$$

where $f^x, f^y \in \mathbb{R}^{B\times d}$ denote the batch-level features for image and text modalities. $\mathbb{E}$ takes expectation over all feature pairs within and across modalities, and $g(\cdot,\cdot)$ is a Gaussian kernel that quantifies pairwise similarity.

We compute the raw confidence scores based on the above metrics as:

$$\begin{aligned}s_{\mathrm{fea}} &= \exp(\beta \cdot \mathrm{NMI}(Y^x, Y^y))\,, \\ s_{\mathrm{clu}} &= \exp\big(-\beta \cdot \mathrm{MMD}^2(f^x, f^y)\big)\,,\end{aligned} \tag{19}$$

where $\beta > 0$ controls the sharpness; NMI is left unsquared to retain $[0, 1]$ scale sensitivity, while using MMD^2 provides proper scaling and smoother gradients for stable weight updates.

These scores are normalized to obtain the current batch weights:

$$w_{\mathrm{fea}} = \frac{s_{\mathrm{fea}}}{s_{\mathrm{fea}} + s_{\mathrm{clu}}}, \quad w_{\mathrm{clu}} = \frac{s_{\mathrm{clu}}}{s_{\mathrm{fea}} + s_{\mathrm{clu}}}. \tag{20}$$

To suppress early-stage instability, we apply EMA updates:

$$\begin{aligned}\widehat{w}_{\mathrm{fea}} &\leftarrow \rho \cdot w_{\mathrm{fea}} + (1-\rho)\cdot \widehat{w}_{\mathrm{fea}}, \\ \widehat{w}_{\mathrm{clu}} &\leftarrow \rho \cdot w_{\mathrm{clu}} + (1-\rho)\cdot \widehat{w}_{\mathrm{clu}},\end{aligned} \tag{21}$$

where $\rho \in (0, 1)$ is the smoothing coefficient. We also use a warm-up strategy by fixing the feature-branch weight to 0.5 during the first few epochs to avoid unstable updates caused by noisy or uninformative pseudo labels at early stages.

Objective Function

To achieve an effective balance between semantic alignment and structural consistency, we formulate the overall optimization objective by integrating the proposed loss components, as shown in Equation (22):

$$\min\ \mathcal{L}_{\mathrm{total}} = \underbrace{\widehat{w}_{\mathrm{fea}}\,\mathcal{L}_{\mathrm{HIC}}}_{\text{Feature-guided}} + \underbrace{\widehat{w}_{\mathrm{clu}}\big(\mathcal{L}_{\mathrm{Cluster}} + \mathcal{L}_{\mathrm{struct}}\big)}_{\text{Cluster-guided}}. \tag{22}$$

By minimizing the feature-level alignment loss $\mathcal{L}_{\mathrm{HIC}}$, the cluster-level contrastive loss $\mathcal{L}_{\mathrm{Cluster}}$, and the structure-preserving loss $\mathcal{L}_{\mathrm{struct}}$, our framework ensures that both the fine-grained instance semantics and the coarse-grained structural semantics are fully captured and aligned across modalities.

Experiments

Datasets

To evaluate the effectiveness of our proposed UDCH framework, we conduct extensive experiments on three widely used cross-modal retrieval benchmarks: MIRFLICKR-25K (Li et al. 2025a), NUS-WIDE (Wong et al. 2025), and MS COCO (Li et al. 2025b).

MIRFLICKR-25K: a widely used benchmark containing 25,000 image–text pairs annotated with 24 semantic categories. After removing unlabeled samples, 20,015 pairs are retained. We randomly select 2,000 pairs as the query set, and the rest are used as the gallery set.

NUS-WIDE: consists of 269,498 images annotated with labels from 81 concept categories. We select 186,577 image–text pairs from the top 10 most frequent classes for evaluation. Among them, 2,100 pairs are randomly chosen as the query set and the remaining as the gallery set.

MS COCO: contains 123,287 image–text pairs categorized into 80 semantic classes, each described by five textual annotations. After filtering out unlabeled data, 122,218 valid samples remain. We randomly select 5,000 pairs as the query set, and the rest form the gallery set.

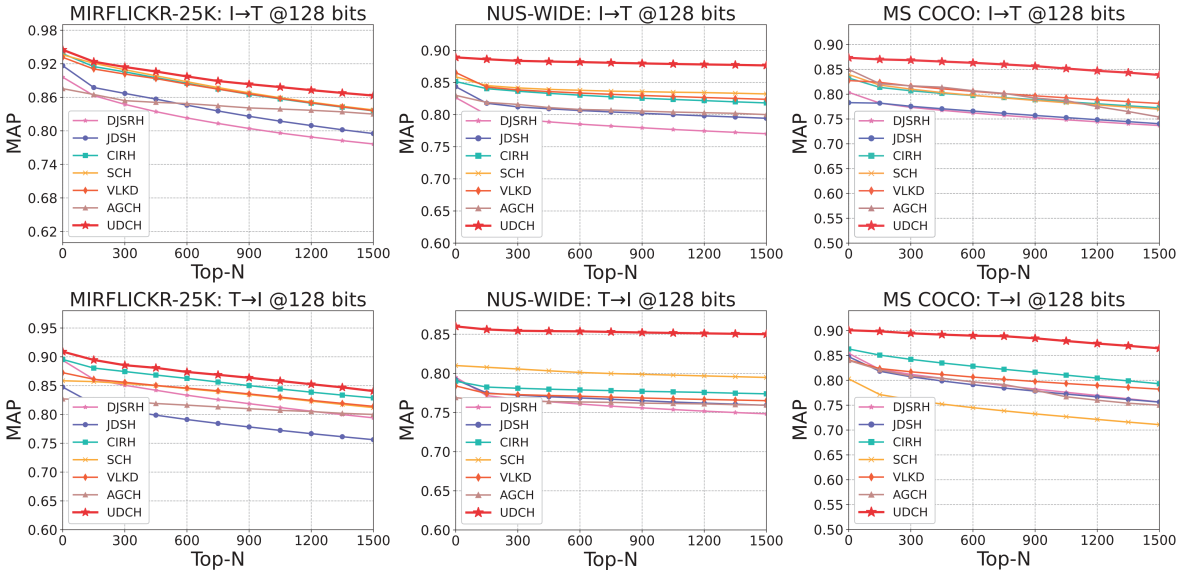


Figure 2: Top-N precision curves (128 bits) on MIRFLICKR-25K, NUS-WIDE, and MS COCO.

Baselines

We compare UDCH with six representative unsupervised cross-modal hashing methods, including DJSRH (Su, Zhong, and Zhang 2019), JDSH (Liu et al. 2020), AGCH (Zhang et al. 2021), CIRH (Zhu et al. 2023), VLKD (Sun, Li, and Dong 2023), and SCH (Hu et al. 2024). All methods are evaluated under the same protocol using mAP and top-N precision metrics.

Evaluation Metrics

To comprehensively evaluate UDCH for cross-modal retrieval, we consider two tasks: image-to-text (I→T) and text-to-image (T→I). We plot iteration curves to assess optimization stability, and we report mean average precision (mAP) and top-N precision curves: mAP summarizes overall retrieval accuracy, while top-N shows how precision varies with the list length. In addition, we provide t-SNE visualizations of the learned embeddings colored by K-means clusters to assess cluster compactness and separability.

Implementation Details

All implementations are based on the PyTorch framework and executed on a single NVIDIA GeForce RTX 3080 Ti GPU.

Training Details: We adopt Adam optimizer with a learning rate of 5×10^{-4} and weight decay of 1×10^{-6}. The batch size is set to 256, and the InfoNCE temperature τ is set to 0.9. For dynamic weighting, we set $\beta = 0.3$ and EMA smoothing factor $\rho = 0.9$. The number of clusters K is typically set in proportion to the number of categories in each dataset: K=8 for MIRFLICKR-25K, K=4 for NUS-WIDE, and K=26 for MS COCO.

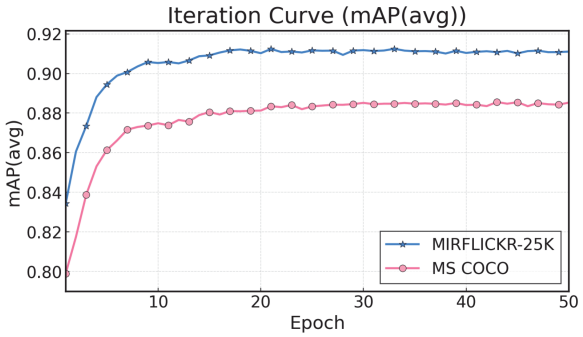
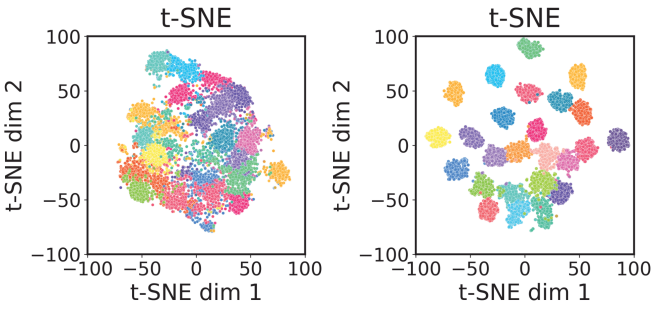


Figure 3: Iteration curves of mAP across epochs on MIRFLICKR-25K (blue stars) and MS COCO (pink circles).

Comparison Results and Discussions

Results on top-N Precision Curves: As illustrated in Figure 2, UDCH consistently outperforms all baselines across three datasets. On MIRFLICKR-25K and NUS-WIDE, UDCH maintains the top precision from N=1 to 1500, showing strong generalization. Notably, on MS COCO, UDCH preserves a clear lead in both I→T and T→I, reflecting its superior scalability and robustness. Additionally, the UDCH curves start higher in the small-N region and decrease more slowly as N increases, indicating earlier retrieval of high-confidence items and greater stability under tail noise.

Results on Convergence Curve: Figure 3 illustrates the convergence behavior of our proposed UDCH method. We select MIRFLICKR-25K and MS COCO as representative datasets to evaluate the training dynamics. As training progresses, the mAP values steadily improve and stabilize. These results confirm the effectiveness and stable optimization behavior of UDCH.



(a) Epoch 0 (b) Final

Figure 4: Visualization of clustering alignment via t-SNE on MIRFLICKR-25K.

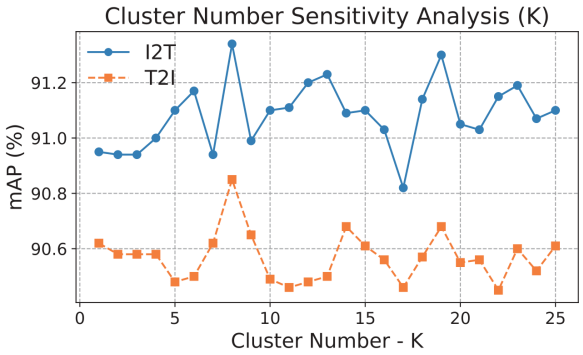


Figure 5: Cluster number sensitivity analysis on MIRFLICKR-25K (128 bits).

mAP Results Analysis: In Table 1, UDCH leads across three datasets, four code lengths, and both retrieval directions. At 16-bit, on MS COCO the T $\rightarrow$ I score improves from 0.7496 to 0.8555 (+10.59%); on NUS-WIDE, I $\rightarrow$ T and T $\rightarrow$ I rise by +6.21% and +7.22%, respectively. At 32- and 64-bit, the gains remain steady: for NUS-WIDE (32-bit), I $\rightarrow$ T and T $\rightarrow$ I improve by +5.22% and +7.01%; for MS COCO (64-bit), I $\rightarrow$ T and T $\rightarrow$ I improve by +7.52% and +5.10%. At 128-bit, MS COCO reaches 0.9135 (I $\rightarrow$ T) and 0.9115 (T $\rightarrow$ I), surpassing the strongest baselines by +8.86% and +6.03%; on NUS-WIDE the margins are +3.26% and +4.92%. Overall, UDCH is robust across code lengths, with advantages amplified on larger and more semantically complex datasets.

Results on t-SNE Visualization: As shown in Figure 4, subfigure (a) depicts the initial distribution before training, where clusters are loose and heavily entangled, indicating poor semantic separability. After training, subfigure (b) shows that clusters become compact and well-separated, validating the effectiveness of our cluster-aware semantic modeling.

Cluster Number Sensitivity Analysis

To investigate the impact of the cluster number K on retrieval performance, we conduct a sensitivity analysis by varying K from 1 to 25. As shown in Fig. 5, the performance curves of both Image-to-Text and Text-to-Image retrieval tasks fluctuate with different K values. Overall, the results remain relatively stable when K is around 8. When K is too small, the semantic representation tends to be overly coarse, potentially failing to distinguish fine-grained content. These findings align with the common empirical strategy of setting the number of clusters to approximately one-third of the number of semantic categories, which provides a reasonable trade-off between semantic abstraction and structural stability.

Ablation Experiments

To evaluate the contribution of each loss in UDCH, we conduct ablations on MS COCO with 128-bit codes (Table 2). Using only the instance-level loss $\mathcal{L}_{\text{HIC}}$ yields 89.76% (I$\rightarrow$T) and 89.93% (T$\rightarrow$I). Adding the cluster-level term ($\mathcal{L}_{\text{Cluster}} + \mathcal{L}_{\text{HIC}}$) gives 90.11% and 89.63% — a small fluctuation on T$\rightarrow$I but clear gain on I$\rightarrow$T. Introducing the structure-aware regression on top of both ($\mathcal{L}_{\text{struct}} + \mathcal{L}_{\text{Cluster}} + \mathcal{L}_{\text{HIC}}$) further improves to 90.54% and 90.05%. The full model (UDCH (Full)) reaches 91.35% and 91.15%, showing that combining instance, cluster, and structure-level cues produces the most discriminative and semantically aligned hash codes.

Task	Method (Loss Composition)	128-bits
I$\rightarrow$T	UDCH-$\mathcal{L}_{\text{HIC}}$	89.76%
	UDCH-$\mathcal{L}_{\text{Cluster}} + \mathcal{L}_{\text{HIC}}$	90.11%
	UDCH-$\mathcal{L}_{\text{struct}} + \mathcal{L}_{\text{Cluster}} + \mathcal{L}_{\text{HIC}}$	90.54%
	UDCH (Full)	**91.35%**
T$\rightarrow$I	UDCH-$\mathcal{L}_{\text{HIC}}$	89.93%
	UDCH-$\mathcal{L}_{\text{Cluster}} + \mathcal{L}_{\text{HIC}}$	89.63%
	UDCH-$\mathcal{L}_{\text{struct}} + \mathcal{L}_{\text{Cluster}} + \mathcal{L}_{\text{HIC}}$	90.05%
	UDCH (Full)	**91.15%**

Table 2: Ablation study of UDCH and its loss combinations on the MS COCO dataset at 128 bits.

Conclusion

In this paper, we propose UDCH, an Unsupervised Dynamic Weighted Cluster-Cooperative Hashing framework for cross-modal retrieval. Instance-level alignment, cluster-aware semantic modeling, and structure-preserving regression are jointly optimized under a single objective, forming a global–local–structural loop. The cluster-guided term shapes the global semantic geometry and sharpens inter-class boundaries, thereby stabilizing instance alignment; the structural term captures fine-grained higher-order local relations while suppressing geometric drift in Hamming space to improve cross-modal consistency; and the dynamic weighting mechanism adaptively reallocates emphasis across training stages.

In future work, we will extend UDCH to continual learning in streaming and open-vocabulary settings, introduce lightweight multimodal priors as consistency constraints within the unsupervised loop, and systematically evaluate its scalability and robustness at web scale.

References

Chen, H.; Zou, Z.; Liu, Y.; and Zhu, X. 2025. Deep Class-Guided Hashing for Multi-Label Cross-Modal Retrieval. *Applied Sciences*, 15(6): 3068.

Hu, P.; Zhu, H.; Lin, J.; Peng, D.; Zhao, Y.-P.; and Peng, X. 2023. Unsupervised Contrastive Cross-Modal Hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3877–3889.

Hu, Z.; Cheung, Y.-m.; Li, M.; and Lan, W. 2024. Cross-Modal Hashing Method with Properties of Hamming Space: A New Perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 7636–7650.

Huang, J.; Kang, P.; Fang, X.; Han, N.; Xie, S.; and Gao, H. 2024. Efficient Discriminative Hashing for Cross-Modal Retrieval. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 54(6): 3865–3878.

Li, J.; Jiang, L.; Ma, Z.; Jiang, K.; Fang, X.; and Wen, J. 2025a. Lightweight Contrastive Distilled Hashing for Online Cross-Modal Retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(5): 4779–4787.

Li, Y.; Long, J.; and Yang, Z. 2025. Asymmetric Cross-Modal Hashing Based on Formal Concept Analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1392–1401.

Li, Y.; Zhen, L.; Sun, Y.; Peng, D.; Peng, X.; and Hu, P. 2025b. Deep Evidential Hashing for Trustworthy Cross-Modal Retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(17): 18566–18574.

Liu, S.; Qian, S.; Guan, Y.; Zhan, J.; and Ying, L. 2020. Joint-Modal Distribution-Based Similarity Hashing for Large-Scale Unsupervised Deep Cross-Modal Retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 1379–1388. ACM.

Liu, X.; Li, J.; Nie, X.; Zhang, X.; and Yin, Y. 2024. Fast Unsupervised Cross-Modal Hashing with Robust Factorization and Dual Projection. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 20(12): 370.

Pu, R.; Qin, Y.; Peng, D.; Song, X.; and Zheng, H. 2025. Deep Reversible Consistency Learning for Cross-Modal Retrieval. *IEEE Transactions on Multimedia*, 27: 4095–4106.

Qin, Q.; Huo, Y.; Zhang, W.; Huang, L.; and Nie, J. 2025a. Deep Discriminative Boundary Hashing for Cross-Modal Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, early access: 1–1.

Qin, Q.; Wu, L.; Zhang, W.; et al. 2025b. Deep Semantic-Consistent Penalizing Hashing for Cross-Modal Retrieval. *IEEE Transactions on Multimedia*, 27: 4613–4626.

Su, S.; Zhong, Z.; and Zhang, C. 2019. Deep Joint-Semantics Reconstructing Hashing for Large-Scale Unsupervised Cross-Modal Retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3027–3035.

Sun, L.; Li, Y.; and Dong, Y. 2023. Learning From Expert: Vision-Language Knowledge Distillation for Unsupervised Cross-Modal Hashing Retrieval. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval (ICMR)*, 499–507. ACM.

Tu, J.; Liu, X.; Hao, Y.; Hong, R.; and Wang, M. 2024. Two-Step Discrete Hashing for Cross-Modal Retrieval. *IEEE Transactions on Multimedia*, 26: 8730–8741.

Wong, W. K.; Fei, L.; Qin, J.; Zhao, S.; Wen, J.; and He, Z. 2025. Heterogeneous Pairwise-Semantic Enhancement Hashing for Large-Scale Cross-Modal Retrieval. *IEEE Transactions on Multimedia*, 27: 3238–3250.

Xie, H.; Jiang, Y.; Zhang, L.; Li, P.; Zhang, D.; and Zhang, Y. 2024. Semantic-Enhanced Proxy-Guided Hashing for Long-Tailed Image Retrieval. *IEEE Transactions on Multimedia*, 26: 9499–9514.

Yang, F.; Han, M.; Ma, F.; et al. 2024. Disperse Asymmetric Subspace Relation Hashing for Cross-Modal Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(1): 603–617.

Yang, F.; Liu, X.; Ma, F.; et al. 2025. Online Asymmetric Supervised Discrete Cross-Modal Hashing for Streaming Multimedia Data. *Pattern Recognition*, 165: 111604.

Yang, Z.; Deng, X.; Guo, L.; et al. 2023. Asymmetric Supervised Fusion-Oriented Hashing for Cross-Modal Retrieval. *IEEE Transactions on Cybernetics*, 54(2): 851–864.

Zhang, D.; Wu, X.-J.; Xu, T.; and Kittler, J. 2023. WATCH: Two-Stage Discrete Cross-Media Hashing. *IEEE Transactions on Knowledge and Data Engineering*, 35(6): 6461–6474.

Zhang, P.-F.; Li, Y.; Huang, Z.; and Xu, X.-S. 2021. Aggregation-Based Graph Convolutional Hashing for Unsupervised Cross-Modal Retrieval. *IEEE Transactions on Multimedia*, 24: 466–479.

Zhang, T.; Xue, Z.; Mahmood, A.; Du, J.; Dong, Y.; Ou, S.; Feng, L.; Yang, M.-H.; and Qi, Y. 2025. Generating Synthetic Data for Unsupervised Federated Learning of Cross-Modal Retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(21): 22569–22577.

Zhu, L.; Wu, X.; Li, J.; Zhang, Z.; Guan, W.; and Shen, H. T. 2023. Work Together: Correlation-Identity Reconstruction Hashing for Unsupervised Cross-Modal Retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 35(9): 8838–8851.