DSSL-Hash: Dynamic Semantic Structure Learning for Unsupervised Cross-Modal Hashing

Fan Yang, **Tongxuan Pei**, **Yuanzhi Zhao** and **Yudong Zhao***

School of Computer and Artificial Intelligence, Nanjing University of Finance and Economics, Nanjing 210023, Jiangsu, China

nufe_yf@163.com, creazyptx@163.com, 2185215287@qq.com, ydzhao1127@163.com

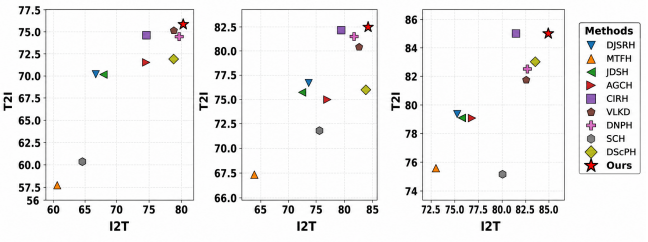
Abstract

A central bottleneck in unsupervised cross-modal hashing is that both similarity guidance and objective weighting are typically static, while the semantic structure induced during training is inherently dynamic. We present DSSL-Hash, which turns semantic structure learning into an evolving process rather than a one-shot preprocessing step. Specifically, DSSL-Hash distills semantic cues from a vision–language foundation model and constructs a dynamic perception graph via an adaptive similarity matrix, enabling graph propagation to capture higher-order relations beyond pairwise matching. We further reshape Hamming-space learning with a semantic-channel constraint that explicitly regulates distances for partially similar pairs, reducing semantic distortion after binarization. Finally, we develop a Retrieval-Aware Adaptive Scheduler (RAAS) that leverages retrieval feedback to co-adjust modality interactions and objective weights, achieving robust optimization without extensive manual hyper-parameter search. Extensive experiments on MS COCO, NUS-WIDE, and MIRFLICKR-25K across multiple code lengths demonstrate that DSSL-Hash consistently outperforms recent state-of-the-art unsupervised cross-modal hashing methods.

1 Introduction

Driven by the rapid proliferation and widespread adoption of the Internet, multimodal data has experienced an unprecedented explosion. However, efficiently processing and storing such vast volumes of heterogeneous information remains a formidable challenge. To address this, cross-modal retrieval has long been regarded as the touchstone for representation learning. By learning a unified feature space, it effectively unlocks the immense potential of multimodal data while bridging the semantic gap and heterogeneity gap inherent in disparate modalities, such as text and video [Zhang *et al.*, 2025].

Cross-modal hashing (CMH) facilitates efficient multimodal retrieval by mapping heterogeneous data into a unified Hamming space [Ding *et al.*, 2025; Tu *et al.*, 2025; Xu *et al.*, 2025; Yang *et al.*, 2025a; Zhu *et al.*, 2025].



(a) 16 bits (b) 32 bits (c) 64 bits

Figure 1: mAP Comparison on MS COCO Dataset.

Among its variants, unsupervised CMH (UCMH) is particularly valuable for real-world applications due to its independence from semantic labels [Zhao *et al.*, 2026]. However, UCMH still faces three critical challenges: 1) Traditional methods often fail to capture fine-grained semantics and higher-order correlations when aligning diverse modalities [Huo *et al.*, 2024; Qin *et al.*, 2025a]. 2) Most approaches rely on predefined static similarity matrices [Zou *et al.*, 2025], which cannot adapt to evolving semantic relationships during training, leading to suboptimal hash code representation. 3) Existing frameworks typically depend on manual, static hyper-parameter settings [Lin *et al.*, 2025]. Such inflexibility ignores training dynamics and limits the model's potential, as hash code learning is highly sensitive to hyper-parameter configurations [Yang *et al.*, 2024].

To address the above challenges, this paper proposes a dynamic optimization cross-modal hashing method that integrates knowledge distillation and reinforcement learning—Dynamic Semantic Structure Learning for Unsupervised Cross-Modal Hashing (DSSL-Hash). Within this framework, we first leverage the CLIP vision-language pretrained model as the teacher network and innovatively introduce a dynamic perception matrix. Then, through a knowledge distillation mechanism, the student network is guided to capture deep semantic correlations across heterogeneous modalities. Furthermore, we introduce a reinforcement learning-based optimization strategy by employing the Deep Deterministic Policy Gradient (DDPG) algorithm to dynamically adjust both the modality feature weights and the loss function ratios based on real-time model performance. This enables the network to maintain strong adaptability under dy-

*Corresponding author.

namically changing multimodal data distributions. The main contributions are as follows:

- To mitigate cross-modal semantic loss, we propose Dynamic Graph-Based Cognitive Optimization (DGO). DGO leverages CLIP-guided distillation for fine-grained learning and GCNs to capture multi-scale structural relations in the perception graph. Furthermore, a Hamming distance optimization strategy is employed to minimize semantic distortion during hash mapping.
- To surpass static configuration constraints, we propose RAAS. RAAS employs the DDPG algorithm to adaptively adjust modal weights and loss ratios based on real-time feedback. By bypassing manual empirical tuning, RAAS autonomously explores hyper-parameter dependencies and optimal strategies, significantly enhancing parameter interaction and optimization efficiency.
- Sufficient experiments on three benchmark datasets show that the proposed method DSSL-Hash outperforms other state-of-the-art unsupervised cross-modal hashing methods.

2 Related Work

Deep cross-modal hashing (DCMH) maps multimodal data into a unified Hamming space and is categorized into supervised [Zou *et al.*, 2025; Cao *et al.*, 2024; Huo *et al.*, 2024; Liang *et al.*, 2024] and unsupervised [Dufumier *et al.*, 2025; Wen *et al.*, 2025; Wang *et al.*, 2024; Cui *et al.*, 2024; Cai *et al.*, 2024; Liu *et al.*, 2024] methods.

Supervised Deep Cross-Modal Hashing. utilizes semantic labels to maximize discriminative power. Several studies enhance uncertainty perception: FUME [Duan *et al.*, 2025] employs fuzzy set theory for category credibility, TCL [Qin *et al.*, 2025b] introduces Evidential Deep Learning (EDL) for self-assessment, and PDF [Cao *et al.*, 2024] proposes a Predictive Dynamic Fusion framework. Other methods prioritize semantic discrimination: PromptHash [Zou *et al.*, 2025] introduces prompt learning, DNPH [Huo *et al.*, 2024] utilizes learnable category proxies, and DScPH [Qin *et al.*, 2025a] explores imbalanced correlations for semantic consistency. Additionally, SCH [Hu *et al.*, 2024] and MRDH [Liang *et al.*, 2024] refine similarity constraints via differential constraints and global similarity metrics.

Unsupervised Deep Cross-Modal Hashing. extracts similarity relationships without labels. To bridge semantic gaps, dual-coding mechanisms like UDC [Cai *et al.*, 2024] and UDDH [Zhang *et al.*, 2024] jointly optimize semantic and content hashes. For similarity matrix robustness, VTM-UCH [Fan and Cao, 2025] uses vision-guided text mining, SMSH [Kuai *et al.*, 2025] introduces Jaccard similarity, SACH [Cui *et al.*, 2024] uses structure-aware contrastive learning, and FUCH [Liu *et al.*, 2024] employs collective matrix factorization with Cauchy Loss. To capture fine-grained information, HuggingHash [Wang *et al.*, 2024] and USFTH [Yang *et al.*, 2025b] leverage Transformer architectures, while CoMM [Dufumier *et al.*, 2025] and InfMasking [Wen *et al.*, 2025] utilize information-theoretic frameworks to maximize collaborative interaction.

To address hyper-parameter complexity and semantic modeling gaps, we propose DSSL-Hash. It leverages DDPG for auto-tuning, while combining Knowledge Distillation (KD) and Graph Convolutional Networks (GCNs) to capture global and local semantic structures.

3 Methodology

3.1 Problem Definition

The proposed DSSL-Hash method targets unsupervised cross-modal retrieval. Let $\mathcal{O} = \{v_i, t_i\}_{i=1}^{n}$ be a dataset of n image-text pairs, where $v_i \in R^{1\times d_v}$ and $t_i \in R^{1\times d_t}$ denote image and text features. Guided by a teacher model, we map these to c-bit binary codes $\mathbf{H}_v, \mathbf{H}_t \in \{-1, +1\}^{n\times c}$ by optimizing Hamming distance.

We adopt a mini-batch training strategy with m pairs per batch. Throughout this paper, uppercase and lowercase letters denote matrices and vectors, respectively. For an image query q_v, the cross-modal retrieval task aims to return the top M most relevant instances from $\mathcal{O}$, defined as:

$$\mathcal{R}_t(q_v) = \text{Top-}M_{t_j\in\mathcal{O}}\left(-D_H(h_q, \mathbf{H}_t^{j*})\right) \tag{1}$$

where $\mathcal{R}_t$ denotes the result set of the retrieval task, h_q denotes the hash code of the query sample, and $\mathbf{H}_t^{j*}$ denotes the binary code of the j-th candidate text sample in $\mathbf{H}_t$. The operator Top-$M(\cdot)$ returns M samples with the largest values, which is equivalent to selecting samples with the minimum Hamming distances D_H. Figure 2 illustrates the overall framework of DSSL-Hash.

3.2 Dynamic Graph-Based Cognitive Optimization

In order to solve the problem of static similarity matrix which is difficult to adapt to the dynamic change of semantic structure during the training process and the semantic loss caused by the mapping of real-valued space to Hamming space, we developed the DGO component.

Dynamic Perception Matrix. The dynamic perception matrix in DGO is composed of two hierarchical units: intra-modal feature perception and cross-modal interaction. By leveraging the RAAS component to regulate similarities across these dimensions, we construct a dynamic perception matrix $\mathbf{S}_{DS}$. This matrix serves as the structural foundation for effective feature propagation within GCNs. The formal expressions are defined as follows:

$$\begin{aligned}\mathbf{S}_{DS} &= \alpha\phi(\mathcal{M},\mathcal{M}) + \beta\phi(\mathcal{N},\mathcal{N}) + \gamma\phi(\mathcal{M},\mathcal{N}) \\ &\text{s.t. } \alpha+\beta+\gamma = 1, \quad \alpha,\beta,\gamma \geq 0\end{aligned} \tag{2}$$

where $\phi(X,Y) = \frac{X^\top Y}{\|X\|_2\|Y\|_2}$ represents the cosine similarity, and $\mathcal{M} \in \mathbb{R}^{m\times d_v}, \mathcal{N} \in \mathbb{R}^{m\times d_t}$ represent CLIP-guided teacher features, with $\|\cdot\|_2$ denoting the L_2 norm used to normalize vectors to unit length. Crucially, $\mathbf{S}_{DS}$ is reconstructed at every mini-batch from the current teacher features $(\mathcal{M},\mathcal{N})$ and the modality weights (α,β,γ) updated by RAAS, so the perception graph continuously evolves with the local semantic state during training.

DGO refines CLIP features $\mathcal{Z}_*^{(0)} \in \{\mathcal{M},\mathcal{N}\}$ through layer-wise GCN updates: $\mathcal{Z}_*^{(l+1)} =$

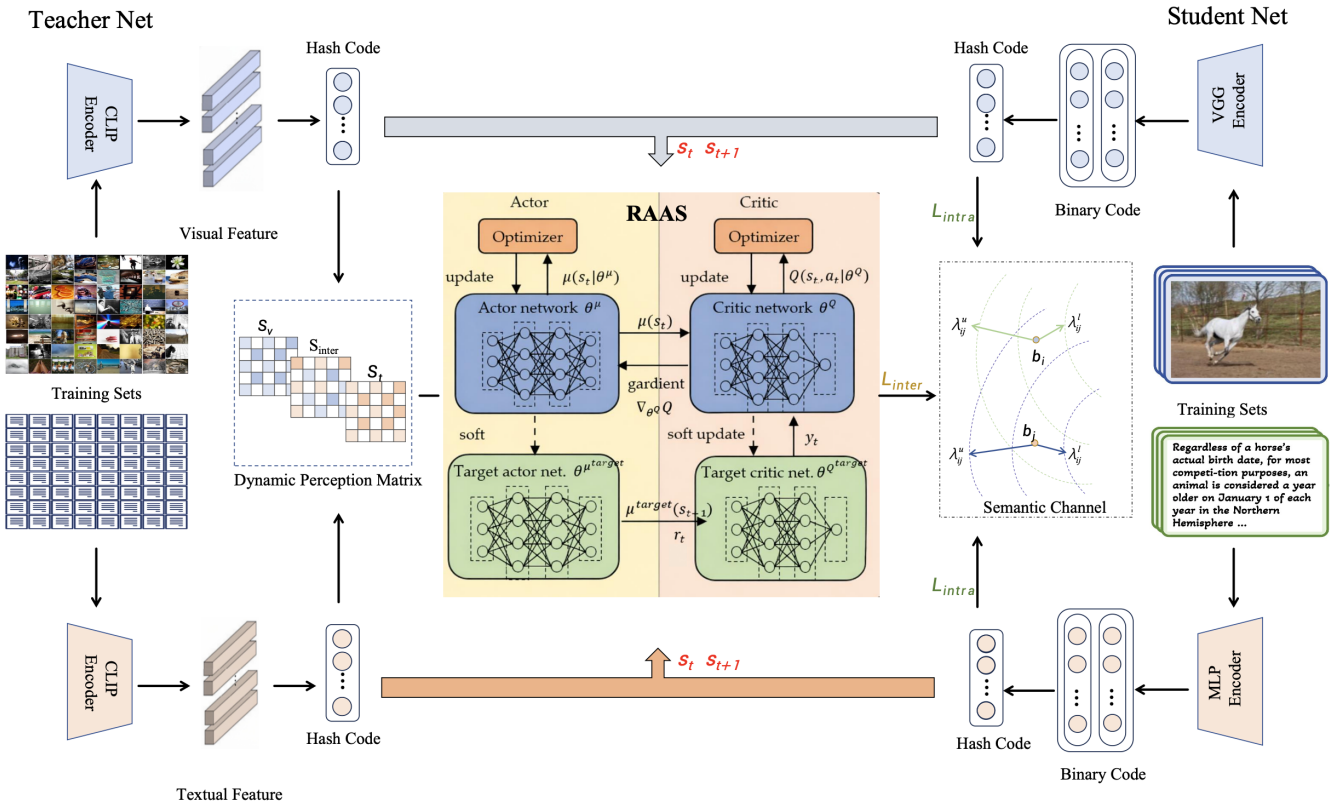


Figure 2: DSSL-Hash comprises two synergistic modules: DGO and RAAS. The DGO module employs CLIP-based distillation and GCNs to learn fine-grained cross-modal semantics and dynamic structural relationships. Concurrently, the RAAS module utilizes reinforcement learning based on DDPG to adaptively optimize feature weights and loss ratios.

$GCN(\mathcal{Z}_*^{(l)}, \mathbf{S}_{DS}, \mathbf{W}^{(l+1)})$, where $\mathbf{W}^{(l+1)}$ denotes the learnable layer weight matrix. An iterative non-linear quantization strategy is further introduced to reduce quantization loss and enhance code discriminability:

$$\mathbf{H}_*^A = \lim_{\omega \to \infty} \tanh(\omega \cdot GCN(\mathcal{Z}_*^{(2)})), \quad \text{s.t.} * \in \{v, t\} \quad (3)$$

where $\mathbf{H}_*^A \in \{-1, +1\}^{m \times c}$ denotes the discrete teacher hash codes, and ω is a RAAS-optimized hyper-parameter controlling the quantization steepness. Larger ω pushes continuous features closer to compact bipolar hash codes.

Semantic-Enhanced Knowledge Distillation. The student model adopts a VGG-16 visual encoder for the image branch and an MLP textual encoder for the text branch, whose outputs are projected into the Hamming space as hash codes $\mathbf{H}_v$ and $\mathbf{H}_t$. In order to improve the accuracy and discriminative power of the student model in generating hash codes, we introduce a distillation loss function $\mathcal{L}_{\text{hash}}$, whose loss is defined as:

$$\mathcal{L}_{\text{hash}} = \sum_{x \in \{v,t\}} \sum_{i=1}^{m} \|h_{x,i}^A - h_{x,i}\|_2^2, \text{ s.t. } h_{x,i}^A, h_{x,i} \in [-1, +1]^c \quad (4)$$

where $h_{x,i}^A, h_{x,i}$ are teacher and student hash codes for modality $x \in \{v, t\}$.

To further enhance the structural coherence within each modality, we introduce an intra-modal consistency loss, denoted as $\mathcal{L}_{\text{intra}}$. The loss is defined as:

$$\mathcal{L}_{\text{intra}} = \sum_{x \in \{v,t\}} \sum_{i,j} \left(\rho h_{x,i}^\top h_{x,j} - h_{x,i}^{A\top} h_{x,j}^A\right)^2 \quad (5)$$

where $h_{x,i}^A$ and $h_{x,i}$ denote the teacher and student hash codes for modality x. A scaling factor ρ is introduced to adjust the similarity magnitude computed by the student model for better matching. To further improve the quality of hash encoding, we construct two cross-modal loss terms: the semantic alignment term $\mathcal{L}_1$ and the structure-preserving term $\mathcal{L}_2$, to jointly optimize the cross-modal hashing representations for image and text modalities. The semantic alignment term $\mathcal{L}_1$ aims to align image and text representations in the hash space under the guidance of $\mathbf{S}_{DS}$. This loss can be rewritten in a unified sample-wise form:

$$\begin{aligned} \mathcal{L}_1 = -\sum_{i,j} \big[& s_{ij} \log \sigma(h_{v,i}^\top h_{t,j}) \\ & + (1 - s_{ij}) \log(1 - \sigma(h_{v,i}^\top h_{t,j})) \big] \end{aligned} \quad (6)$$

where s_{ij} is the (i, j)-th element of the dynamic perceptual similarity matrix S_{DS}, and $\sigma(h_{v,i}^\top h_{t,j}) = (1 + e^{-h_{v,i}^\top h_{t,j}})^{-1}$. The structure-preserving term $\mathcal{L}_2$ is designed to retain the intra-modal structural relationships within each modality dur-

ing cross-modal alignment. It is defined as:

$$\mathcal{L}_2 = \frac{1}{m}(\sum_{i=1}^{m} \|h_{v,i}\|^2 \cdot \|h_{t,i}\|^2 + 2 \sum_{1 \le i < j \le m} (h_{v,i}^\top h_{v,j} \cdot h_{t,i}^\top h_{t,j})) \tag{7}$$

where $\frac{1}{m}$ is a normalization factor.

To encourage structural alignment between visual and textual modalities, we minimize the loss term $\mathcal{L}_2$, which serves as a regularizer that promotes consistency in patterns of intramodal similarity. The final cross-modal consistency loss is defined as:

$$\mathcal{L}_{\text{inter}} = \mathcal{L}_1 + \mathcal{L}_2, \tag{8}$$

To reduce semantic distortion and quantization errors in Hamming mapping, the DGO module introduces a semantically regulated Hamming distance optimization. By redefining Hamming distance to align with sample relationships, we define the target distance based on the similarity score $S_{ij} \in [0, 1]$ as:

$$\lambda_{ij} = \frac{\varepsilon}{2}(1 - S_{ij}), \tag{9}$$

where ε is a tunable hyper-parameter that controls the scale. Additionally, the semantic channel is bounded by an upper limit of $\lambda_{ij}^u = \lambda_{ij}$ and a lower limit of $\lambda_{ij}^l = \max(0, \lambda_{ij} - \tau)$, where $\tau \in \mathbb{R}_+$ denotes the width of the semantic channel.

To model semantic relationships more precisely, we categorize sample pairs into three types and apply different constraint strategies. For partially semantic-positive pairs (i.e., $0 < S_{ij} < 1$), their Hamming distance should lie within a bounded interval:

$$\lambda_{ij}^l \le d_H(h_i, h_j) \le \lambda_{ij}^u, \tag{10}$$

where $d_H(h_i, h_j)$ denotes the Hamming distance between a pair of samples in the hamming space. The overall Hamming distance optimization loss is defined as:

$$\mathcal{L}_{\text{Hamm}} = \sum_{i,j} \left(\eta_{ij} \cdot \Delta_{ij}^l + \xi_{ij} \cdot \Delta_{ij}^u \right), \tag{11}$$

where $\Delta_{ij}^l = (\max(0, \lambda_{ij}^l - d_H(h_i, h_j)))^2$ and $\Delta_{ij}^u = (\max(0, d_H(h_i, h_j) - \lambda_{ij}^u))^2$. The coefficients (η_{ij}, ξ_{ij}) are set based on semantic relations: $(1, 1)$ for partial, $(0, 1)$ for $S_{ij} = 1$, and $(1, 0)$ for $S_{ij} = 0$. While the overall Hamming loss combines constraints for different semantic pair types, we further refine this by explicitly modeling deviations from target distance intervals. The resulting multi-level semantic constraint loss is given by:

$$\mathcal{L}_{sc} = \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{* \in \{l,u\}} W_{ij}^* (d_H(h_i, h_j) - \Lambda_{ij}^*)^2, \tag{12}$$

where $\Lambda^* = [\lambda_{ij}^*]^{m \times m}$ for $* \in \{l, u\}$ are matrices of target Hamming distances. The weights W_{ij}^* are: $W_{ij}^l, W_{ij}^u = 1$ if $S_{ij} \in (0, 1)$; $W_{ij}^l = \theta, W_{ij}^u = 0$ if $S_{ij} = 1$; and $W_{ij}^l = 0, W_{ij}^u = \eta$ if $S_{ij} = 0$. The total loss function of DGO is defined as follows:

$$\mathcal{L}_{\text{DGO}} = \chi_1 \mathcal{L}_{\text{hash}} + \chi_2 \mathcal{L}_{\text{intra}} + \chi_3 \mathcal{L}_{\text{inter}} + \chi_4 \mathcal{L}_{\text{sc}}, \tag{13}$$

Algorithm 1 Dynamic Semantic Structure Learning for Unsupervised Cross-Modal Hashing

Input: multimodal dataset O, hash length c.
Initialize modality weights (α, β, γ) and loss coefficients $\{\chi_k\}_{k=1}^4$.
while the training process has not converged **do**
  DGO Optimization:
    perform graph-based representation learning under current configuration.
    compute $\mathcal{L}_{\text{DGO}}$ according to Eq. (13) and update student networks.
  RAAS Scheduling:
    observe training state and evaluate retrieval performance.
    update actor–critic networks according to Eq. (18).
    reconfigure (α, β, γ) and $\{\chi_k\}_{k=1}^4$ for subsequent optimization.
end while
Output: learned cross-modal hash codes H_v, H_t.

where $\{\chi_k\}_{k=1}^4$ are the dynamic weight coefficients adjusted by the RAAS agent to balance each optimization constraint.

3.3 Retrieval-Aware Adaptive Scheduler

RAAS uses reinforcement learning to adaptively tune modality weights and loss coefficients. At each epoch, it performs an N-order search, filters unstable candidates through gradient-protected rollback, and retains the best configuration via elite selection.

Deep Deterministic Policy Gradient. RAAS formulates hyper-parameter tuning as a reinforcement learning task, employing a dynamic referencing mechanism to navigate the continuous search space. In this framework, the agent's actions are guided by a differential reward that quantifies the performance gap between the current exploration trial and a dynamic baseline, which is updated upon discovering superior configurations. Specifically, let mAP_t denote the validation mAP after the agent applies action a_t at step t, and mAP_{best} the highest mAP observed before step t; the reward is

$$r_t = \text{mAP}_t - \text{mAP}_{\text{best}}, \tag{14}$$

followed by $\text{mAP}_{\text{best}} \leftarrow \max(\text{mAP}_{\text{best}}, \text{mAP}_t)$. To optimize this process, RAAS leverages the DDPG algorithm within an actor-critic architecture. Both the actor and critic networks utilize a two-hidden-layer ResNet backbone. To stabilize training, the critic adopts a target network to calculate the cumulative Q-value as follows:

$$y = r + \zeta Q(s', a'; \theta'), \quad \zeta \in [0, 1], \tag{15}$$

where r is the immediate reward, ζ is the discount factor, and $Q(s', a'; \theta')$ estimates the future value based on the next state s' and the target actor's action a'.

The critic network is trained by minimizing the squared error between the target y and the predicted Q-value:

$$\mathcal{L}_{\text{critic}} = \mathbb{E}_{(s,a,r,s') \sim \mathcal{D}}[(y - Q(s, a; \theta^Q))^2], \tag{16}$$

where y is the target value. Subsequently, the actor network

	Methods	Reference	MIRFLICKR-25K				NUS-WIDE				MS COCO			
			16bits	32bits	64bits	mean	16bits	32bits	64bits	mean	16bits	32bits	64bits	mean
I→T	DCMH	CVPR'2017	0.7240	0.7292	0.7325	0.7286	0.5540	0.5662	0.5887	0.5696	0.5055	0.5215	0.5578	0.5283
	DJSRH	ICCV'2019	0.7878	0.8180	0.8320	0.8126	0.7110	0.7535	0.7811	0.7485	0.6714	0.7371	0.7565	0.7217
	MTFH	TPAMI'2019	0.8379	0.8701	0.8997	0.8692	0.7094	0.7455	0.7649	0.7399	0.6070	0.6370	0.7374	0.6605
	JDSH	SIGIR'2020	0.8253	0.8577	0.8731	0.8520	0.7400	0.7952	0.7983	0.7778	0.6792	0.7263	0.7600	0.7218
	AGCH	TMM'2021	0.8554	0.8880	0.8892	0.8775	0.7816	0.8251	0.8298	0.8122	0.7433	0.7692	0.7699	0.7608
	CIRH	TKDE'2022	0.8666	0.8850	0.9011	0.8842	0.7838	0.8110	0.8272	0.8073	0.7452	0.7933	0.8151	0.7845
	VLKD	ICMR'2023	0.8848	0.8847	0.9022	0.8906	0.7948	0.8088	0.8489	0.8175	0.7824	0.8232	0.8259	0.8105
	SCH	TPAMI'2024	0.8430	0.8808	0.9052	0.8763	0.7931	0.8251	0.8472	0.8218	0.6451	0.7568	0.8000	0.7340
	DNPH	TOMM'2024	0.8679	0.8849	0.9026	0.8851	0.7667	0.7885	0.8009	0.7854	0.7904	0.8118	0.8267	0.8096
	DScPH	TMM'2025	0.8839	0.8891	0.8999	0.8910	0.7880	0.8150	0.8244	0.8091	0.7848	0.8346	0.8354	0.8183
	DSSL-Hash	Ours	**0.8852**	**0.9030**	**0.9174**	**0.9019**	**0.8016**	**0.8273**	**0.8540**	**0.8276**	**0.7956**	**0.8370**	**0.8489**	**0.8272**
		Increased↑	**0.05%**	**1.56%**	**1.35%**	**1.22%**	**0.86%**	**0.27%**	**0.60%**	**0.71%**	**0.66%**	**0.29%**	**1.62%**	**1.09%**
T→I	DCMH	CVPR'2017	0.7633	0.7634	0.7724	0.7664	0.5540	0.5933	0.6144	0.5872	0.5491	0.5716	0.6042	0.5750
	DJSRH	ICCV'2019	0.7575	0.7950	0.8132	0.7886	0.7281	0.7439	0.7568	0.7429	0.7011	0.7674	0.7933	0.7539
	MTFH	TPAMI'2019	0.8410	0.8688	0.8793	0.8630	0.6940	0.7440	0.7469	0.7283	0.5769	0.6743	0.7564	0.6692
	JDSH	SIGIR'2020	0.8000	0.8165	0.8249	0.8138	0.7121	0.7320	0.7750	0.7397	0.7009	0.7577	0.7908	0.7498
	AGCH	TMM'2021	0.8289	0.8344	0.8447	0.8360	0.7384	0.7674	0.7851	0.7636	0.7138	0.7496	0.7897	0.7510
	CIRH	TKDE'2022	0.8420	0.8642	0.8769	0.8610	0.7584	0.7736	0.7912	0.7744	0.7439	0.7933	0.8495	0.7956
	VLKD	ICMR'2023	0.8143	0.8454	0.8465	0.8354	0.7445	0.7448	0.7719	0.7537	0.7496	0.8053	0.8172	0.7907
	SCH	TPAMI'2024	0.7624	0.8181	0.8487	0.8097	0.7568	0.7759	0.7901	0.7743	0.6026	0.7173	0.7512	0.6904
	DNPH	TOMM'2024	0.8378	0.8538	0.8784	0.8567	0.7490	0.7629	0.7728	0.7616	0.7430	0.8162	0.8254	0.7949
	DScPH	TMM'2025	0.8366	0.8573	0.8641	0.8527	0.7424	0.7756	0.7890	0.7690	0.7177	0.7599	0.8297	0.7691
	DSSL-Hash	Ours	**0.8443**	**0.8701**	**0.8811**	**0.8652**	**0.7626**	**0.7789**	**0.7924**	**0.7780**	**0.7561**	**0.8255**	**0.8496**	**0.8104**
		Increased↑	**0.27%**	**0.15%**	**0.20%**	**0.25%**	**0.55%**	**0.39%**	**0.15%**	**0.46%**	**0.87%**	**1.14%**	**0.01%**	**1.86%**

Table 1: Comparison of mAP scores on MIRFLICKR-25K, NUS-WIDE, and MS COCO datasets. The best results are shown in bold, and the second-best results are underlined. The "Increased" row reports the relative gain of DSSL-Hash over the underlined result in each column.

is updated by minimizing the following loss to maximize the expected Q-value:

$$\mathcal{L}_{\text{actor}} = -\mathbb{E}_{s\sim\mathcal{D}}[Q(s, \mu(s;\theta_u);\theta^Q)]. \tag{17}$$

The actor's objective is to optimize the policy μ by generating actions that yield higher rewards as estimated by the critic. The objective functions of RAAS are defined as follows:

$$\mathcal{L}_{\text{RAAS}} = \mathcal{L}_{\text{critic}} + \mathcal{L}_{\text{actor}}. \tag{18}$$

Combining all constraints, the total objective function is:

$$\min_{H_v,H_t} \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{DGO}} + \mathcal{L}_{\text{RAAS}}. \tag{19}$$

This joint objective optimizes binary codes by preserving structural consistency and enhancing semantic discriminability, while the proposed framework remains fully unsupervised. No semantic labels are used to construct $\mathbf{S}_{DS}$, supervise the teacher network, or optimize the student hashing networks. RAAS only accesses the validation mAP at the *scheduling level* to update (α, β, γ) and $\{\chi_k\}_{k=1}^4$, without introducing label supervision into the hashing loss or network optimization. Algorithm 1 summarizes the procedure.

4 Experiments

4.1 Datasets

We evaluate DSSL-Hash on three widely used cross-modal retrieval benchmarks. MIRFLICKR-25K contains 25,000 Flickr image-text pairs annotated with 24 categories. NUS-WIDE includes 269,648 web image-tag pairs with 81 semantic concepts; following common practice, we use the 10 selected categories and 186,577 corresponding pairs. MS COCO contains 123,287 image-text pairs covering 80 categories, where each image is paired with textual descriptions of real-world scenes [Zhu *et al.*, 2023].

4.2 Baselines and Evaluation Metrics

To evaluate the cross-modal retrieval performance of DSSL-Hash, we conduct image-to-text (I→T) and text-to-image (T→I) tasks using mAP and Top-N precision. We compare DSSL-Hash against 11 SOTA baselines: DJSRH [Su *et al.*, 2019], JDSH [Liu *et al.*, 2020], DCMH [Jiang and Li, 2017], MTFH [Liu *et al.*, 2019], AGCH [Zhang *et al.*, 2021], CIRH [Zhu *et al.*, 2022], SCH [Hu *et al.*, 2024], VLKD [Sun *et al.*, 2023], DScPH [Qin *et al.*, 2025a], and DNPH [Huo *et al.*, 2024] on three benchmark datasets. The results are summarized in the following tables and figures.

4.3 Implementation Details

Experiments were conducted on a Win11 server with an RTX 3090Ti GPU using 16–128 bit hash codes. We report mAP@500 and top-1500 precision. Following standard protocol, each dataset is split into database, query, and training sets. We further reserve 5% of the training set as a validation subset, which is used *only* by RAAS to compute r_t and never for representation learning. The N-order RAAS search is set to $N = 10$. The initial weights are $\alpha = 0.5$, $\beta = 0.3$, $\gamma = 0.2$, $\chi_1 = \chi_2 = 0.1$, and $\chi_3 = \chi_4 = 1$, and are then adapted by RAAS. Other fixed hyper-parameters are set as

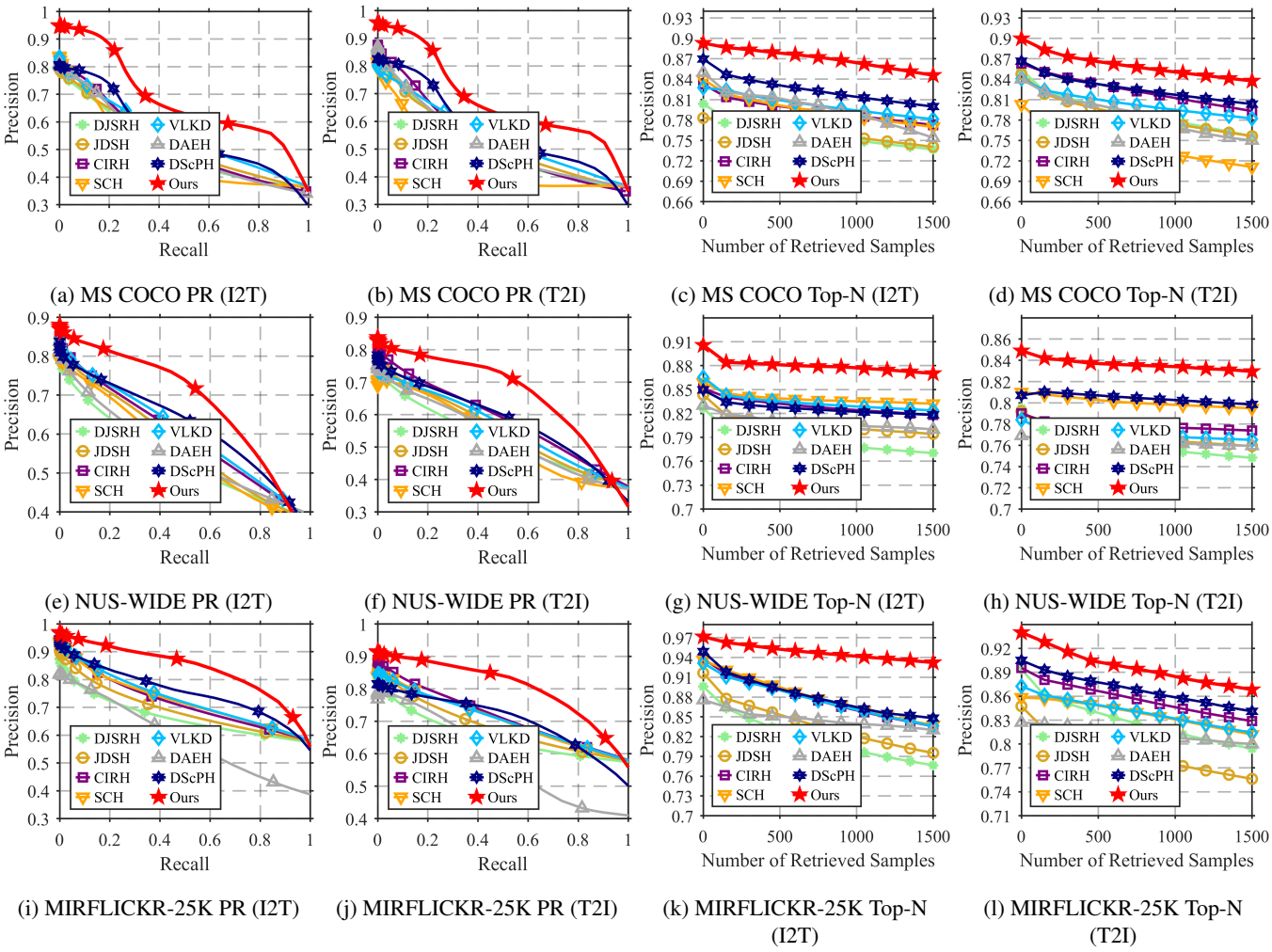


Figure 3: Precision-Recall curves and Top-N precision curves of DSSL-Hash and baselines on three datasets at 128 bits.

$\tau = 0.2$, $\rho = 0.5$, $\varepsilon = 1.0$, $\theta = 0.3$, $\eta = 1.0$, and $\zeta = 0.99$. The student and GCN networks are optimized by SGD with LR=10^{-3}, while the actor and critic use Adam with LRs 10^{-4} and 10^{-3}, respectively. All baselines are reproduced under identical settings on three benchmark datasets.

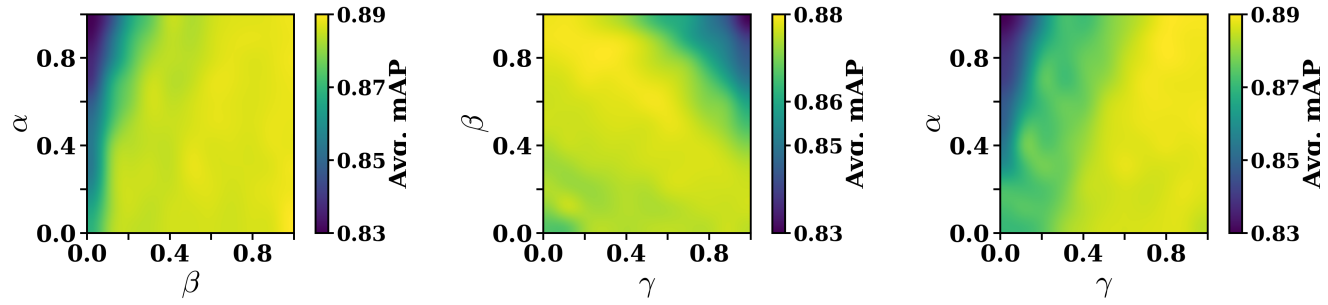
4.4 Comparison Results and Discussions

Results on mAP: As shown in Tables 1, DSSL-Hash consistently achieves superior performance on both the I→T and T→I retrieval tasks across different datasets and hash code lengths. Compared with recent state-of-the-art methods from 2024 and 2025, including DNPH (TOMM2024), SCH (TPAMI2024), and DScPH (TMM2025), DSSL-Hash delivers significant improvements. In addition, the PR curves in the first two columns of Figure 3 further support the mAP results. DSSL-Hash maintains higher precision across most recall ranges on the three datasets, showing stable retrieval performance under different recall levels.

Results on Top-N Precision Curves. The Top-N precision curves in the last two columns of Figure 3 evaluate retrieval stability under different numbers of returned samples. DSSL-Hash maintains competitive precision across the three datasets, especially in the I→T task, indicating its ability to preserve reliable nearest-neighbor rankings under larger retrieval ranges. This benefits from the unsupervised modeling of cross-modal semantic relationships, which helps learn more discriminative modality-specific representations.

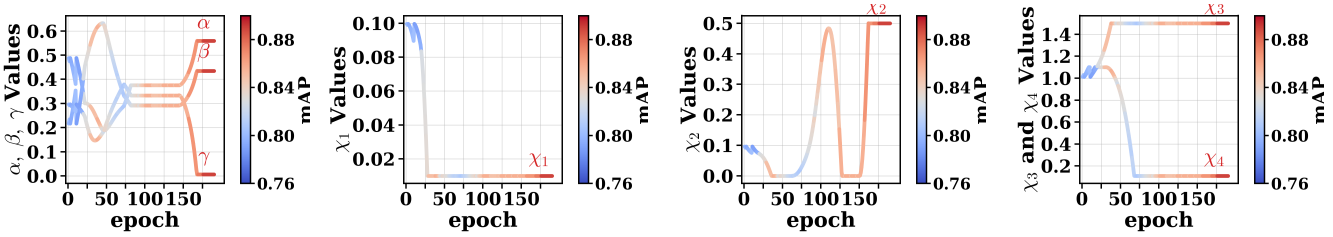
4.5 Ablation Experiments

In order to gain more insight into our proposed DSSL-Hash method, we performed ablation experiments as shown in Table 2. We ablated the three modules of DSSL-Hash, and in the table DSSL-Hash-D which shows that the method deletes the distillation-related loss function in the DGO component, the performance decreases by up to 13.85% due to the loss of high-quality guidance. DSSL-Hash-DGO shows that the method removes the loss function related to semantic conditioning in the DGO component, and the Hamming space is no longer reasonable, leading to a performance drop of about 7%. DSSL-Hash-RAAS shows that we delete the RAAS component part of the dynamic adaptive parameter tuning of the reinforcement learning, which leads to the model not



(a) α vs. β (b) β vs. γ (c) α vs. γ

Figure 4: Heatmaps showing the average mAP under fixed parameters.



(a) α, β and γ (b) χ_1 (c) χ_2 (d) χ_3 and χ_4

Figure 5: Parameter tuning trajectories guided by the RAAS component.

being able to dynamically and adaptively learn the required modal feature information, resulting in a performance drop.

Task	Method	16-bits	32-bits	64-bits	128-bits
I→T	DSSL-Hash-D	72.74	78.32	80.11	80.24
	DSSL-Hash-DGO	86.60	88.10	90.11	90.90
	DSSL-Hash-RAAS	82.24	83.89	85.64	86.49
	DSSL-Hash	**88.52**	**90.30**	**91.74**	**92.87**
T→I	DSSL-Hash-D	75.79	77.98	80.13	80.20
	DSSL-Hash-DGO	82.49	83.72	85.01	86.13
	DSSL-Hash-RAAS	77.98	80.96	83.21	85.11
	DSSL-Hash	**84.43**	**87.01**	**88.11**	**88.63**

Table 2: The mAP results of DSSL-Hash and its variants

4.6 Adaptive Parameter Analysis

Figure 4 shows performance heat maps under fixed weight combinations of α, β, and γ, with $\alpha+\beta+\gamma = 1$. Subfigure (a) shows that the best performance occurs around $\alpha \approx 0.5$ and $\beta \approx 0.3$, indicating the importance of balanced intra-modal similarities. Subfigures (b) and (c) further show that moderate cross-modal interaction ($\gamma \approx 0.2$) is beneficial, while excessive reliance on it without intra-modal guidance degrades performance. Additionally, Figure 5 visualizes the dynamic parameter trajectories guided by the RAAS module. The trajectories clearly show how RAAS adaptively updates the parameters throughout training. As the policy is iteratively refined, the model's mAP improves steadily, confirming the effectiveness of the reinforcement learning mechanism in discovering superior hyper-parameter configurations.

The performance gains verify the effectiveness of RAAS, which adopts a DDPG-based reinforcement learning strategy to search for suitable hyper-parameter configurations. Compared with static or grid-based tuning, RAAS explores the continuous high-dimensional search space more efficiently and avoids the burden of discrete enumeration. During training, the agent adjusts modality weights and loss coefficients using iterative feedback such as mAP, and refines its policy through a closed-loop actor–critic mechanism. The heat map evolution further shows that RAAS shifts from broad exploration to high-performance exploitation, improving accuracy and robustness by balancing modality-specific and cross-modal signals. In terms of overhead, RAAS uses an actor–critic network with two hidden layers and is invoked only N times per epoch rather than per mini-batch, introducing moderate additional training cost. At inference, RAAS is disabled, and retrieval remains standard Hamming lookup over c-bit codes.

5 Conclusion

We propose DSSL-Hash, a robust cross-modal retrieval framework that synergizes CLIP-guided distillation (DGO) for complex dependency modeling with reinforcement learning based on DDPG for autonomous hyper-parameter tuning. This collaborative design enables the model to maintain consistent performance by employing the DGO module to refine semantic alignment and the DDPG-driven RAAS mechanism to search for optimal training configurations. Extensive experiments show that DSSL-Hash significantly outperforms state-of-the-art baselines in accuracy across three benchmarks, demonstrating its robust generalization and efficiency in various retrieval scenarios. Future work will extend the RAAS mechanism to broader foundational models and more heterogeneous data modalities.

References

[Cai *et al.*, 2024] Hongmin Cai, Bin Zhang, Junyu Li, Bin Hu, and Jiazhou Chen. Unsupervised dual hashing coding (udc) on semantic tagging and sample content for cross-modal retrieval. *IEEE Transactions on Multimedia*, 26:9109–9120, 2024.

[Cao *et al.*, 2024] Bing Cao, Yinan Xia, Yi Ding, Changqing Zhang, and Qinghua Hu. Predictive dynamic fusion. In *International Conference on Machine Learning*, pages 5608–5628, 2024.

[Cui *et al.*, 2024] Jinrong Cui, Zhipeng He, Qiong Huang, Yulu Fu, Yuting Li, and Jie Wen. Structure-aware contrastive hashing for unsupervised cross-modal retrieval. *Neural Networks*, 174:106211, 2024.

[Ding *et al.*, 2025] Guohui Ding, Zhonghua Li, Rui Zhou, and Qian Gao. Unsupervised adversarial contrastive hashing for cross-modal retrieval. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, page 575–584, 2025.

[Duan *et al.*, 2025] Siyuan Duan, Yuan Sun, Dezhong Peng, Zheng Liu, Xiaomin Song, and Peng Hu. Fuzzy multimodal learning for trusted cross-modal retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20747–20756, 2025.

[Dufumier *et al.*, 2025] Benoit Dufumier, Javiera Castillo-Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? In *International Conference on Learning Representations*, 2025.

[Fan and Cao, 2025] 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.

[Hu *et al.*, 2024] Zhikai Hu, Yiu-ming Cheung, Mengke Li, and Weichao Lan. Cross-modal hashing method with properties of hamming space: A new perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):7636–7650, 2024.

[Huo *et al.*, 2024] Yadong Huo, Qin Qibing, Jiangyan Dai, Wenfeng Zhang, Lei Huang, and Chengduan Wang. Deep neighborhood-aware proxy hashing with uniform distribution constraint for cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(6):1–23, 2024.

[Jiang and Li, 2017] Qing-Yuan Jiang and Wu-Jun Li. Deep cross-modal hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3232–3240, 2017.

[Kuai *et al.*, 2025] Mingjin Kuai, Jun Long, and Zhan Yang. Statistical model-driven similarity hashing: Bridging modalities for efficient unsupervised retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 11977–11986, 2025.

[Liang *et al.*, 2024] Xiao Liang, Erkun Yang, Yanhua Yang, and Cheng Deng. Multi-relational deep hashing for cross-modal search. *IEEE Transactions on Image Processing*, 33:3009–3020, 2024.

[Lin *et al.*, 2025] Zengrong Lin, Zheng Wang, Tianwen Qian, Pan Mu, Sixian Chan, and Cong Bai. Neighborretr: Balancing hub centrality in cross-modal retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9263–9273, 2025.

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

[Liu *et al.*, 2020] Song Liu, Shengsheng Qian, Yang Guan, Jiawei Zhan, and Long Ying. 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*, pages 1379–1388, 2020.

[Liu *et al.*, 2024] 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.

[Qin *et al.*, 2025a] 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.

[Qin *et al.*, 2025b] Yang Qin, Lifu Huang, Dezhong Peng, Bohan Jiang, Joey Tianyi Zhou, Xi Peng, and Peng Hu. Trustworthy visual-textual retrieval. *IEEE Transactions on Image Processing*, 34:4515–4526, 2025.

[Su *et al.*, 2019] Shupeng Su, Zhisheng Zhong, and Chao Zhang. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3027–3035, 2019.

[Sun *et al.*, 2023] Lina Sun, Yewen Li, and Yumin Dong. Learning from expert: Vision-language knowledge distillation for unsupervised cross-modal hashing retrieval. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 499–507, 2023.

[Tu *et al.*, 2025] Junfeng Tu, Xueliang Liu, Yanbin Hao, and Richang Hong. A unified generative hashing for cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(12), 2025.

[Wang *et al.*, 2024] Jinpeng Wang, Ziyun Zeng, Bin Chen, Yuting Wang, Dongliang Liao, Gongfu Li, Yiru Wang, and Shu-Tao Xia. Hugs bring double benefits: Unsupervised cross-modal hashing with multi-granularity aligned transformers. *International Journal of Computer Vision*, 132(8):2765–2797, 2024.

[Wen *et al.*, 2025] Liangjian Wen, Qun Dai, Jianzhuang Liu, Jiangtao Zheng, Yong Dai, Dongkai Wang, Zhao Kang,

Jun Wang, Zenglin Xu, and Jiang Duan. Infmasking: Unleashing synergistic information by contrastive multimodal interactions. In *Advances in Neural Information Processing Systems*, 2025. Spotlight.

[Xu *et al.*, 2025] Liming Xu, Hanqi Li, Jie Shao, Xianhua Zeng, and Weisheng Li. Multi-scale consistency deep lifelong cross-modal hashing. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(10), 2025.

[Yang *et al.*, 2024] Fan Yang, Meng Han, Fumin Ma, Yufeng Liu, Xiaojian Ding, and Deyu Tong. Disperse asymmetric subspace relation hashing for cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(1):603–617, 2024.

[Yang *et al.*, 2025a] Fan Yang, Xinqi Liu, Fumin Ma, Xiaojian Ding, and Kaixiang Wang. Online asymmetric supervised discrete cross-modal hashing for streaming multimedia data. *Pattern Recognition*, 165:111604, 2025.

[Yang *et al.*, 2025b] Zhan Yang, Binghong Chen, Jiajun Tang, and Yinan Li. Unsupervised similarity-fusion transformer hashing for multimodal retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 5932–5941, 2025.

[Zhang *et al.*, 2021] Peng-Fei Zhang, Yang Li, Zi Huang, and Xin-Shun Xu. Aggregation-based graph convolutional hashing for unsupervised cross-modal retrieval. *IEEE Transactions on Multimedia*, 24:466–479, 2021.

[Zhang *et al.*, 2024] 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, 2024.

[Zhang *et al.*, 2025] Huaiwen Zhang, Yang Yang, Fan Qi, Shengsheng Qian, and Changsheng Xu. Active supervised cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):5112–5126, 2025.

[Zhao *et al.*, 2026] Yuanzhi Zhao, Fan Yang, Yudong Zhao, and Xiaoyu Li. Udch: Unsupervised dynamic weighted cluster-cooperative hashing for cross-modal retreival. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(16):13297–13304, Mar. 2026.

[Zhu *et al.*, 2022] 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.

[Zhu *et al.*, 2023] Lei Zhu, Chaoqun Zheng, Weili Guan, Jingjing Li, Yang Yang, and Heng Tao Shen. Multimodal hashing for efficient multimedia retrieval: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(1):239–260, 2023.

[Zhu *et al.*, 2025] Sa Zhu, Dayan Wu, Chenming Wu, Pengwen Dai, and Bo Li. Endogenous recovery via within-modality prototypes for incomplete multimodal hashing. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 2521–2529, August 2025.

[Zou *et al.*, 2025] Qiang Zou, Shuli Cheng, and Jiayi Chen. Prompthash: Affinity-prompted collaborative cross-modal learning for adaptive hashing retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19649–19658, 2025.