Stationary and Clustering Transformer Hashing for Cross-modal Retrieval

Zhan Yang[1], **Yiran Liu**[1], **Youyuan Huang**[2], **Yinan Li**[1*]

[1]Big Data Institute, Central South University, Changsha 410083, China
[2]School of Computer Science and Engineering, Central South University, Changsha 410083, China
{zyang22, liuyiran, huangyouyuan, liyinan}@csu.edu.cn

Abstract

Unsupervised cross-modal hashing has gained significant attention for efficient retrieval between heterogeneous modalities through encoding data into the unified binary representations, offering low storage cost and fast response. However, the constraints of existing methods persist in bridging the cross-modal semantic gap and capturing fine-grained global semantic structures without explicit labels. In this paper, we propose an innovative unsupervised **S**tationary distribution and soft **C**lustering **T**ransformer **H**ashing approach for cross-modal retrieval, denoted as **SCTH**. Initially, a Transformer-based modality fusion encoder is employed to extract abundant cross-modal semantic representations, further integrated with contrastive hashing to minimize the semantic gap. To enhance the inter-modal alignment, a pseudo-classifier clustering module with entropy-regularized contrastive loss is presented, ensuring balanced and diverse cluster assignments in unsupervised settings. Additionally, a Markovian stationary distribution strategy stabilizes the feature representations through mitigating the interference of noise and outliers. Comprehensive experiments on MIRFlickr, NUS-WIDE, and IAPR-TC12 datasets validate that **SCTH** outperforms state-of-the-art hashing methods in cross-modal retrieval tasks, demonstrating superior generalization performance.

Introduction

With the swift growth of multimedia data, cross-modal retrieval methodologies have become a research hotspot (Zhang et al. 2025b; Duan et al. 2025), which intends to query the relevant content via another modality, realizing efficient information integration. In recent years, hash learning (Huo et al. 2024; Liu et al. 2024a; Yang et al. 2025) has been widely utilized in cross-modal retrieval tasks attribute to the low storage cost, and fast response speed.

Cross-modal hashing (CMH) is designed to encode heterogeneous modalities data into a set of binary codes, enabling efficient retrieval through calculating the Hamming distance in terms of hash codes (Yang et al. 2022; Jin et al. 2024; Pu et al. 2025; Li, Long, and Yang 2025). Further, hashing methods can be partitioned into supervised and unsupervised variants. Contemporary supervised approaches leverage identifier tags to explore semantic relations within the Hamming space (Yang et al. 2024; Sun et al. 2024; Li et al. 2025), ensuring the preservation of semantic relevance. In contrast, unsupervised CMH methods maintain the semantic coherence between diverse modality data in the absence of labels (Qin et al. 2024; Wang et al. 2024b; Zhang et al. 2025a; Zhu et al. 2025), which are insufficient to deal with the issues of feature alignment and accumulation of quantization errors, consequently struggling to exceed supervised methods in terms of retrieval efficacy. However, most data is unlabeled in practical applications, and it is time-consuming to annotate data via human analysts. Thus, there exists an urgent necessity to introduce novel modeling mechanism and integration strategy to improve the comprehensive performance of hash learning.

Transformer architecture demonstrates superior capability in modeling long-range dependencies and global semantic context (Vaswani et al. 2017), and has been extended to multimodal tasks (Srivastava and Sharma 2024; Zheng et al. 2024), which enables each element in a sequence to attend to others, facilitating deep semantic interaction between heterogeneous modalities in a context-aware manner. Particularly, it is appropriate for hash learning (Liu et al. 2023; Shen et al. 2024b), where accurate feature fusion is critical to generate modality-invariant and semantically aligned binary codes. Moreover, the self-attention mechanism ensures that models automatically identify important semantic patterns, contributing to suppressing the irrelevant noise and enhancing retrieval precision.

Pseudo-cluster learning that rely on neural network (Liu et al. 2024b; Yu et al. 2025) enables the discovery of latent semantic structures without annotations. Other than hard clustering that forces a single label assignment, soft clustering provides probabilistic membership distributions, realizing the flexible and nuanced representation of semantic similarity. For unsupervised cross-modal retrieval, soft clustering serves as a form of pseudo-supervision (Zeng, Yu, and Oyama 2020; Zeng, Sun, and Mao 2021), guiding the model to preserve consistent cluster-level semantics between heterogeneous modalities. Furthermore, soft clustering promotes compact intra-cluster distances and separation between diverse semantic groups, which is compatible with the objective of hash learning (Wang et al. 2021, 2024a), *i.e.*, maps semantically similar instances to adjacent binary codes, and dissimilar ones to distant codes, enhancing the

*Corresponding author

semantic consistency of hash codes between multi-views.

Markovian stationary (Meyer 1994) provides a principled solution in large-scale hashing retrieval, where the distribution represents the long-term importance of each data point within a relation graph, enabling the model to capture global structure of multimodal data. Besides, when employed as a supervisory signal, the stationary distribution can prioritize semantically central instances and reduce the interference of local noise, leading to compact and consistent hash codes. Based on the above, leveraging the sophisticated representation capability of deep neural networks, heterogeneous modalities can be potentially aligned, realizing precisely semantic correlation and generalization performance. However, current unsupervised cross-modal retrieval faces the following core challenges.

- **Discrepant Cross-modal Semantic Fusion**: the inter-modal semantic gap is hard to bridge, and there exists an innate difference between heterogeneous modalities in perceptual structure and feature distribution. Besides, conventional unsupervised methods lack the effective alignment mechanism, and it is difficult to learn the consistent semantic representation.
- **Inaccurate Sample Clustering**: in unsupervised learning environments, hard clustering and other approaches for estimating semantic similarity metrics produce inaccurate assignments, where semantically unrelated samples are clustered, compromising the integrity of semantic representations and degrading retrieval performance.
- **Limited Global Semantic Modeling**: most approaches focus on local sample relations, neglecting the global semantic structure. The limitation hinders the model from capturing comprehensive distribution and aggregation characteristics of data in the latent semantic space. Meanwhile, local noise can severely degrade performance, reducing the availability of cross-modal retrieval.

In summary, constructing the cross-modal global semantic relations in an unsupervised environment and improving the clustering consistency between heterogeneous modalities becomes a key issue in the current research. To address the challenges, this paper proposes a **S**tationary distribution and soft **C**lustering **T**ransformer **H**ashing (**SCTH**) method for cross-modal retrieval, which integrates the steady-state probability transfer mechanism with cross-modal contrastive clustering framework to jointly optimize semantic structure perception and feature alignment, enabling the generation of consistent cross-modal hash codes without label supervision. The key contributions are threefold:

- A multimodal fusion Transformer encoder is introduced to integrate features from image and text modalities, capturing abundant cross-modal semantics and enhancing the information interactions. Further, the fusion representations are transformed into hash codes combined with cross-modal contrastive learning, effectively reducing the inter-modal semantic gap and quantization error.
- An advanced multimodal clustering mechanism based on pseudo-classifiers is proposed to solve the issue of semantic modeling in unsupervised settings, which employs the pseudo-cluster to generate soft cluster assignments and incorporates the cross-modal contrastive loss with adaptive entropy regularization, ensuring balanced and diverse cluster assignments, promoting the inter-modal semantic alignment and representation.
- Markovian stationary distribution is leveraged to enhance the feature stability. Specifically, through cross-modal representations to construct the transfer probability matrix, guiding the model to converge to a stable feature distribution. The process utilizes iterative diffusion to suppress the influence of outliers and employs stationary probabilities to ensure the consistency between integrated features and hash codes, enhancing the generalization of representations against noise and complex semantics.

Method

Notations

Consider the dataset $\mathbf{X} = \{\mathbf{X}^1, ..., \mathbf{X}^m\}$ that encompasses m modalities, where subspace $\mathbf{X}^m = \{\mathbf{x}_1, ..., \mathbf{x}_n\}^\top \in \mathbb{R}^{n \times d_m}$, n is the count of instances, and d_m is the dimension of modality $\mathbf{X}^m$. The objective of this paper is to learn compact binary codes $\mathbf{B} \in \{-1, 1\}^{n \times k}$, which is derived by the **sign** function from consensus representations $\mathbf{H}$, where k is the length of hash code. Besides, matrices and vectors are represented by bold uppercase and lowercase letters, respectively, and the transpose of matrix $\mathbf{N}$ is denoted by $\mathbf{N}^\top$.

SCTH Framework

As demonstrated in Fig.1, the proposed **SCTH** mainly contains three modules, *i.e.*, multimodal representation fusion based on Transformer, individual modality-specific soft clustering, and global supervision based on Markovian stationary. The fine-grained explications are as follows.

Multimodal representation fusion based on Transformer To reinforce the capacity of multimodal information interaction, the fusion module is designed to integrate heterogeneous features based on a standard Transformer encoder. For clear illustration, this paper mainly deals with image $\mathbf{X}^I \in \mathbb{R}^{n \times d_I}$ and text $\mathbf{X}^T \in \mathbb{R}^{n \times d_T}$ modalities, the representations are spliced with linear transformation as,

$$\begin{aligned} &\mathbf{X}_L^I = \mathbf{X}^I \mathbf{W}_I + \mathbf{1}\mathbf{b}_I,\ \mathbf{X}_L^T = \mathbf{X}^T \mathbf{W}_T + \mathbf{1}\mathbf{b}_T, \\ &\mathbf{X}^F = \text{concat}[\mathbf{X}_L^I, \mathbf{X}_L^T] \in \mathbb{R}^{n \times d}, \end{aligned} \tag{1}$$

where $\mathbf{b}_I, \mathbf{b}_T \in \mathbb{R}^{1 \times \frac{d}{2}}$ denote the bias terms, $\mathbf{W}_I \in \mathbb{R}^{d_I \times \frac{d}{2}}$, $\mathbf{W}_T \in \mathbb{R}^{d_T \times \frac{d}{2}}$, d is the hidden size of Transformer encoder, d_I and d_T are the dimensions of image and text features, respectively. Sequentially, the modality fusion Transformer encoder takes the multimodal representation $\mathbf{X}^F$ as input, which is utilized to construct queries, keys, and values as,

$$\mathbf{Q} = \mathbf{X}^F \mathbf{W}^Q,\ \mathbf{K} = \mathbf{X}^F \mathbf{W}^K,\ \mathbf{V} = \mathbf{X}^F \mathbf{W}^V, \tag{2}$$

where $\mathbf{W}^Q \in \mathbb{R}^{d \times d_K}$, $\mathbf{W}^K \in \mathbb{R}^{d \times d_K}$, $\mathbf{W}^V \in \mathbb{R}^{d \times d_V}$ are learnable parameters, d_K and d_V are the dimensions of keys

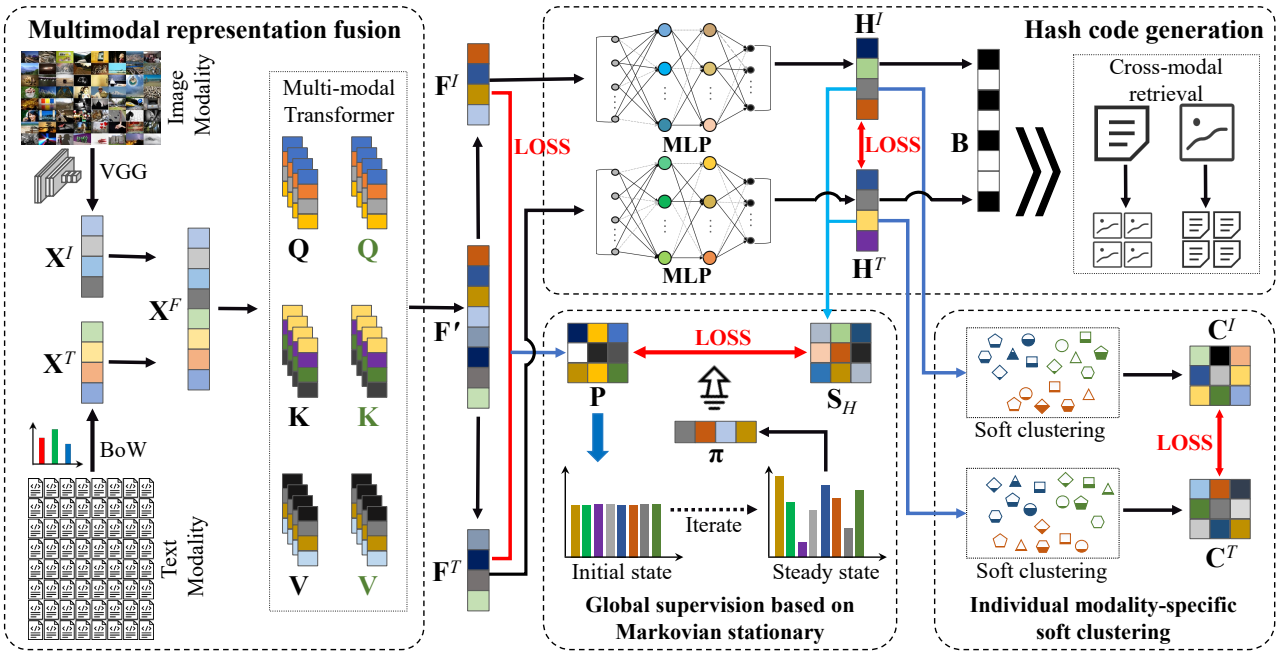


Figure 1: The framework of **S**tationary distribution and soft **C**lustering **T**ransformer **H**ashing for cross-modal retrieval (**SCTH**).

and values, respectively. For any image-text pair, the Transformer encoder generates a fusion representation as,

$$\mathbf{F}' = \mathrm{SAtt}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_K}})\mathbf{V}. \tag{3}$$

Furthermore, the self-attention mechanism captures inter-modal dependencies and semantic correlations, and the cross-modal semantic representation $\mathbf{F}'_i$ can be decomposed into individual modality-specific representations $\mathbf{F}_i^I$ and $\mathbf{F}_i^T$, which represent the respective image and text modalities of i-th sample. To ensure data representations of the same category between diverse modalities possess consistent semantics, the modality fusion representation contrast loss can be formulated as,

$$\begin{aligned}\mathcal{L}_{\text{trans}}^I &= -\sum_{i=1}^{n} \log \frac{\exp(\langle \mathbf{F}_i^I, \mathbf{F}_i^T \rangle / \tau)}{\sum_{j=1}^{n} \exp\left(\langle \mathbf{F}_i^I, \mathbf{F}_j^T \rangle / \tau\right)}, \\ \mathcal{L}_{\text{trans}}^T &= -\sum_{i=1}^{n} \log \frac{\exp(\langle \mathbf{F}_i^T, \mathbf{F}_i^I \rangle / \tau)}{\sum_{j=1}^{n} \exp\left(\langle \mathbf{F}_i^T, \mathbf{F}_j^I \rangle / \tau\right)},\end{aligned} \tag{4}$$

where τ is the temperature parameter. In addition, the hash features can be obtained from fusion representations through hash networks of the multi-layer perceptron (MLP) as,

$$\mathbf{H}^I = \mathrm{MLP}(\mathbf{F}^I), \ \mathbf{H}^T = \mathrm{MLP}(\mathbf{F}^T), \tag{5}$$

where $\mathbf{H}_i^I$ and $\mathbf{H}_i^T$ represent the respective hash features of i-th sample. Similar to fusion loss, *i.e.*, Eq. (4), the hash feature contrast loss for hash features can be formulated as,

$$\begin{aligned}\mathcal{L}_{\text{mlp}}^I &= -\sum_{i=1}^{n} \log \frac{\exp(\langle \mathbf{H}_i^I, \mathbf{H}_i^T \rangle / \tau)}{\sum_{j=1}^{n} \exp\left(\langle \mathbf{H}_i^I, \mathbf{H}_j^T \rangle / \tau\right)}, \\ \mathcal{L}_{\text{mlp}}^T &= -\sum_{i=1}^{n} \log \frac{\exp(\langle \mathbf{H}_i^T, \mathbf{H}_i^I \rangle / \tau)}{\sum_{j=1}^{n} \exp\left(\langle \mathbf{H}_i^T, \mathbf{H}_j^I \rangle / \tau\right)}.\end{aligned} \tag{6}$$

Therefore, the cross-modal fusion loss is defined as the sum of modality fusion representation contrastive loss and weighted hash feature contrastive loss,

$$\mathcal{L}_{\text{fusion}} = \mathcal{L}_{\text{trans}}^I + \mathcal{L}_{\text{trans}}^T + \frac{1}{2}\left(\mathcal{L}_{\text{mlp}}^I + \mathcal{L}_{\text{mlp}}^T\right). \tag{7}$$

Individual modality-specific soft clustering To further optimize the cross-modal hashing representations, a pseudo-classifier clustering method is proposed, which enables heterogeneous modalities to align the soft cluster assignments, facilitating the generation of compact hash codes that preserve semantic consistency.

Specifically, instances from each modality are processed through a pseudo-classifier to obtain soft cluster assignments, represented as $\mathbf{C}^m = h(\mathbf{H}^m) \in \mathbb{R}^{n \times c}$, where c is the number of clusters. The j-th column of matrices $\mathbf{C}^I$ and $\mathbf{C}^T$ are defined as,

$$\mathbf{C}_j^I = \left[\mathbf{C}_{1j}^I, \ldots, \mathbf{C}_{nj}^I\right]^\top, \ \mathbf{C}_j^T = \left[\mathbf{C}_{1j}^T, \ldots, \mathbf{C}_{nj}^T\right]^\top, \tag{8}$$

where $\mathbf{C}_j^m$ represents the probability distribution of samples assigned to the j-th cluster, satisfying $\sum_{j=1}^{c} \mathbf{C}_{i,j}^m = 1$.

Note that soft labels capture the multi-semantic property of instances, enabling an image or text to belong to multiple clusters. Besides, soft cluster assignments within the same modality encourage clear distinctions between clusters, ensuring each cluster can capture the unique semantic pattern. At the same time, clusters corresponding to the same semantic concept align well between heterogeneous modalities to enhance cross-modal matching. Additionally, the soft label distribution remains diverse to avoid over-concentration in a few dominant clusters, which could otherwise reduce the retrieval diversity.

To achieve the above goals, the clustering module optimizes cross-modal semantic alignment through maximizing the probability of correctly matching image-text pairs, and coupled with entropy regularization to ensure the diversity in soft cluster assignments. Thus, the loss for image modality can be formulated as,

$$\mathcal{L}_{\mathrm{c}}^{I} = -\sum_{j=1}^{c} \log \frac{\varepsilon\left(\mathbf{C}_j^I, \mathbf{C}_j^T\right)}{\sum_{l=1}^{c}\left[\varepsilon\left(\mathbf{C}_j^I, \mathbf{C}_l^I\right) + \varepsilon\left(\mathbf{C}_j^I, \mathbf{C}_l^T\right)\right]}, \quad (9)$$

where $\varepsilon(\mathbf{a}, \mathbf{b}) = \exp(\langle \mathbf{a}, \mathbf{b} \rangle / \tau)$, τ is the cluster-level temperature parameter. Similarly, the loss for text modality can be formulated as,

$$\mathcal{L}_{\mathrm{c}}^{T} = -\sum_{j=1}^{c} \log \frac{\varepsilon\left(\mathbf{C}_j^I, \mathbf{C}_j^T\right)}{\sum_{l=1}^{c}\left[\varepsilon\left(\mathbf{C}_j^T, \mathbf{C}_l^T\right) + \varepsilon\left(\mathbf{C}_j^T, \mathbf{C}_l^I\right)\right]}. \quad (10)$$

Additionally, the information entropy is represented as,

$$H(\mathbf{C}^m) = -\sum_{j=1}^{c} P(\mathbf{C}_j^m) \log P(\mathbf{C}_j^m),\ P\left(\mathbf{C}_j^m\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{C}_{ij}^m, \quad (11)$$

which encourages the average distribution of soft cluster assignments to approach uniformity, preventing the collapse of hash codes into a few dominant clusters, ensuring comprehensive semantic coverage. Therefore, the cluster-level loss is defined as,

$$\mathcal{L}_{\text{cluster}} = \frac{1}{2c}\left(\mathcal{L}_{\mathrm{c}}^{I} + \mathcal{L}_{\mathrm{c}}^{T}\right) - H(\mathbf{C}^I) - H(\mathbf{C}^T). \quad (12)$$

Global supervision based on Markovian stationary Mainstream unsupervised CMH methods rely on local similarity metric or reconstruction loss, ignoring the global structural properties of multimodal data distributions. The limitation frequently results in unstable hash codes generation when dealing with outliers or intricate semantic relations. To address the issues, we introduce a global supervision architecture, which utilizes Markovian stationary distribution to capture the topological and statistical properties of heterogeneous modality data, improving the alignment and generalization performance.

The stationary framework models global inter-modal relations through a discrete-time Markov chain, with each component represents a data sample. Particularly, matrix $\mathbf{P} \in \mathbb{R}^{n \times n}$ encodes transition probabilities between states, and the Markovian stationary distribution $\pi \in \mathbb{R}^{1 \times n}$ represents the long-term equilibrium probabilities of each instance, which enhances feature stability via reducing the impact of outliers through iterative diffusion and optimizing global structural alignment. To construct the transition probability matrix $\mathbf{P}$, the fusion image and text representations $\mathbf{F}^I$ and $\mathbf{F}^T$ is employed, which can be defined as,

$$\mathbf{P} = \operatorname{softmax}\left(\cos\left(\mathbf{F}^I, \mathbf{F}^T\right)\right), \quad (13)$$

where $\mathbf{P}$ is row-stochastic, satisfying $\sum_{j=1}^{n} \mathbf{P}_{ij} = 1$.

Moreover, we provide a rigorous foundation through the Perron-Frobenius theorem for non-negative matrices, ensuring the existence and uniqueness of stationary distribution.

Theorem 1. (Caswell 2001) Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a non-negative, irreducible matrix. Then:

- There exists a positive real eigenvalue $\lambda_{\max}$ such that $|\lambda| \leq \lambda_{\max}$ for other eigenvalues.
- $\lambda_{\max}$ has a unique (up to scaling) eigenvector $\mathbf{v} > 0$.
- If $\mathbf{A}$ is primitive, then $\lambda_{\max}$ is simple, and the power iteration converges to $\mathbf{v}$.

Accordingly, the **Proposition 1** is presented, which demonstrates that the transition probability matrix constructed via softmax of cosine similarities satisfies the conditions of Perron-Frobenius theorem.

Proposition 1. $\mathbf{P} \in \mathbb{R}^{n \times n}$ is a row-stochastic matrix, satisfying $\sum_{j=1}^{n} \mathbf{P}_{ij} = 1$, and $\mathbf{P}$ is irreducible and aperiodic. Then, there exists a unique positive stationary distribution $\pi > 0$ such that,

$$\pi\mathbf{P} = \pi, \quad \sum_{i=1}^{n} \pi_i = 1, \quad (14)$$

and the power iteration $\pi^{(t)}\mathbf{P}^t \to \pi$ converges geometrically as $t \to \infty$.

Proof. Since $\mathbf{P}$ in Eq.(13) is a row-stochastic matrix, satisfying $\sum_{j=1}^{n} \mathbf{P}_{ij} = 1$, so $\lambda = 1$ is an eigenvalue with right eigenvector $\mathbf{1} = (1, \ldots, 1)^\top$, *i.e.*, $\mathbf{P1} = \mathbf{1}$. Besides, the softmax operation ensures that components $\forall\ \mathbf{P}_{ij} > 0$, which implies that $\mathbf{P}$ is irreducible and aperiodic.

And since $\mathbf{P}$ is irreducible and aperiodic, according to **Theorem 1**, there is a unique positive left eigenvector $\pi > 0$ for eigenvalue $\lambda = 1$, satisfying Eq.(14). Furthermore, the eigenvalue 1 is simple, and other eigenvalues satisfy $|\lambda_i| < 1$. Thus, for any initial distribution $\pi^{(0)}$, the power iteration can be formulated as,

$$\pi^{(t+1)} = \pi^{(t)}\mathbf{P},\ \pi^{(t)} = \pi^{(0)}\mathbf{P}^t. \quad (15)$$

Moreover, the spectral gap ensures geometric convergence,

$$\|\pi^{(0)}\mathbf{P}^t - \pi\| \leq C\rho^t,\ \rho < 1, \quad (16)$$

where $\rho = \max_{i \neq 1} |\lambda_i|$ and $C > 0$ is a constant. Thus, $\pi^{(t)}\mathbf{P}^t \to \pi$ as $t \to \infty$. The existence, uniqueness, and convergence are proved. □

To ensure numerical stability, the Markovian stationary distribution is calculated iteratively. Specifically, we initialize the state vector with a uniform distribution as,

$$\pi^{(0)} = \frac{1}{n} \cdot \mathbf{1}^\top = \left[\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\right], \quad (17)$$

which maximizes entropy as,

$$H(\pi^{(0)}) = -\sum_{i=1}^{n} \pi_i^{(0)} \log \pi_i^{(0)}, \quad (18)$$

and assuming no prior sample preference,

$$E\left[\pi^{(0)}\right] = \frac{1}{n}. \quad (19)$$

For an irreducible and aperiodic Markov chain, $\lambda(\mathbf{P}) < 1$ holds, where $\lambda(\mathbf{P})$ is the contraction coefficient of $\mathbf{P}$. Thus,

Task	Method	Source	MIRFlickr				NUS-WIDE				IAPR-TC12			
			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	DGCPN	AAAI 21	0.881	0.900	0.914	0.920	0.785	0.824	0.840	0.856	0.594	0.638	0.664	0.674
	CIRH	TKDE 23	0.887	0.902	0.919	0.929	0.802	0.829	0.842	0.851	0.611	0.656	0.681	0.697
	UCCH	TPAMI 23	0.867	0.895	0.916	0.930	0.758	0.802	0.837	0.857	0.528	0.622	0.651	0.700
	PT-FUCH	ACM MM 23	0.801	0.814	0.829	0.834	0.683	0.717	0.751	0.765	0.518	0.563	0.593	0.611
	CMCL	TKDE 24	0.860	0.898	0.912	0.905	0.753	0.808	0.820	0.797	0.607	0.645	0.630	0.605
	SCH	TPAMI 24	0.800	0.871	0.892	0.898	0.775	0.813	0.817	0.841	0.521	0.621	0.654	0.671
	GCMLH	TNNLS 24	0.863	0.881	0.889	0.890	0.764	0.789	0.801	0.796	0.544	0.578	0.593	0.600
	SMSH	AAAI 25	0.876	0.898	0.907	0.927	0.783	0.818	0.825	0.834	0.619	0.656	0.670	0.692
	VTM-UCH	AAAI 25	0.884	0.894	0.898	0.895	0.735	0.733	0.728	0.706	0.616	0.657	0.691	0.695
	SCTH	**Ours**	**0.894**	**0.916**	**0.933**	**0.938**	**0.811**	**0.838**	**0.857**	**0.863**	**0.620**	**0.661**	**0.699**	**0.705**
T → I	DGCPN	AAAI 21	0.856	0.877	0.873	0.876	0.765	0.794	0.805	0.811	0.600	0.655	0.672	0.676
	CIRH	TKDE 23	0.855	0.882	0.894	0.898	0.778	0.798	0.804	0.819	0.611	0.659	0.679	0.699
	UCCH	TPAMI 23	0.830	0.844	0.868	0.884	0.728	0.760	0.777	0.799	0.554	0.616	0.659	0.694
	PT-FUCH	ACM MM 23	0.786	0.798	0.805	0.815	0.676	0.705	0.726	0.732	0.519	0.567	0.589	0.608
	CMCL	TKDE 24	0.849	0.878	0.889	0.886	0.721	0.749	0.775	0.773	0.599	0.655	0.634	0.608
	SCH	TPAMI 24	0.792	0.852	0.873	0.889	0.775	0.800	0.821	0.826	0.524	0.619	0.637	0.662
	GCMLH	TNNLS 24	0.822	0.833	0.833	0.837	0.693	0.721	0.724	0.719	0.510	0.551	0.563	0.583
	SMSH	AAAI 25	0.852	0.868	0.885	0.896	0.761	0.781	0.803	0.803	0.601	0.655	0.668	0.671
	VTM-UCH	AAAI 25	0.842	0.858	0.846	0.871	0.707	0.692	0.723	0.718	0.626	0.657	0.645	0.692
	SCTH	**Ours**	**0.897**	**0.915**	**0.935**	**0.937**	**0.810**	**0.839**	**0.857**	**0.862**	**0.628**	**0.663**	**0.699**	**0.706**

Table 1: Performance comparison of **SCTH** and baselines on the three datasets with various code lengths.

the iterative update via Eq.(15) is a contraction mapping with respect to the L_1 norm, guaranteeing the convergence,

$$\|\pi^{(t+1)} - \pi\| \leq \lambda(\mathbf{P})\|\pi^{(t)} - \pi\|. \tag{20}$$

Accordingly, iterate the state vector through transfer probability matrix, which reaches steady state when the difference between neighboring vectors is less than a preset threshold,

$$\begin{aligned} &\pi = \pi^{(t+1)}, \\ &s.t.\ \pi^{(t+1)} = \pi^{(t)}\mathbf{P},\ \|\pi^{(t+1)} - \pi^{(t)}\|_\infty < \epsilon, \end{aligned} \tag{21}$$

where ϵ is a very small value.

Besides, the process of constructing hash features similarity matrices involves the same calculation method as fusion representation, ensuring that the neighborhood structure of original features is preserved in the hash space, which is essential for maintaining the semantic integrity of both image and text features,

$$\mathbf{S}_H = \text{softmax}\left(\cos\left(\mathbf{H}^I, \mathbf{H}^T\right)\right). \tag{22}$$

Particularly, Markovian stationary properties automatically identify key pivot points in the multimodal similarity graph, which forces the loss function to maintain consistency of the inter-modal similarity ordering in the crucial regions, while suppressing the interference of noisy samples,

$$\mathcal{L}_{\text{steady}} = \sum_{i=1}^{n} \pi_i \|\mathbf{P}_i - \mathbf{S}_{H_i}\|_2^2. \tag{23}$$

Overall Loss Function

Through minimization of the contrast loss, the embedding distance of positive sample pairs is brought closer while the negative sample pairs are pushed away in the unified metric space, eliminating the inter-modal representation difference and enhancing the cross-modal discriminative properties of generated hash codes. Specifically, the global semantic relevance is reinforced via performing dynamic weights to reduce the reconstruction error between semantic fusion representations and hashing feature. Integrating Eqs.(7), (12), and (23) into a unified learning framework, the overall loss function for hash codes learning can be derived as,

$$\mathcal{L} = \alpha\mathcal{L}_{\text{fusion}} + \beta\mathcal{L}_{\text{cluster}} + \gamma\mathcal{L}_{\text{steady}}, \tag{24}$$

where α, β, γ are balance parameters.

Experiment

Experiment Settings

Datasets In this paper, we employ **MIRFlickr** (Huiskes and Lew 2008), **NUS-WIDE** (Chua et al. 2009), and **IAPR-TC12** (Escalante et al. 2010) datasets for evaluation, where 5,000/2,000 image-text pairs are identified as the training/query sets, respectively. For MIRFlickr, the residual 18,015 image-text pairs as the retrieval set from 24 categories, where the text utilizes 1,386-D BoW features. For NUS-WIDE, 10 widely accepted concepts with the residual 184,577 image-text pairs as the retrieval set, where the text uses 1,000-D BoW features. For IAPR-TC12, the residual 18,000 image-text pairs as the retrieval set with 255 labels, where the text utilizes 2,912-D BoW features. Particularly, we employ the 4096-D image feature, which is extracted from a 19-layer VGG network (Simonyan and Zisserman 2015) on ImageNet dataset (Deng et al. 2009).

Baselines and evaluation metrics In the experiments, DGCPN (Yu et al. 2021), CIRH (Zhu et al. 2023), UCCH

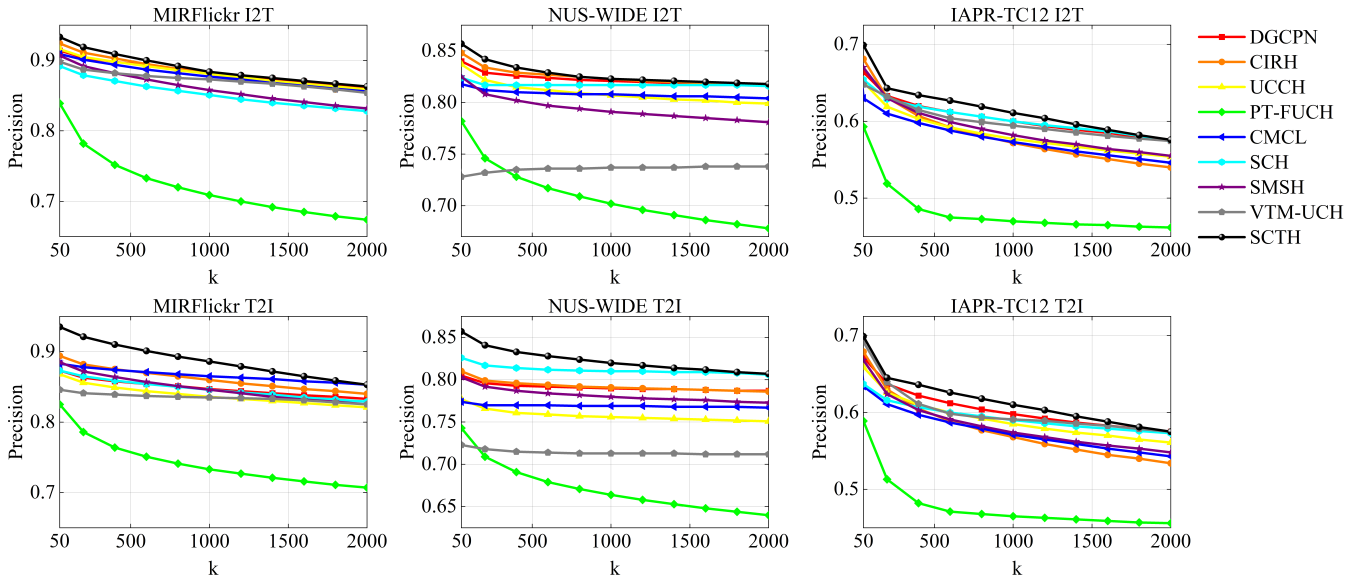


Figure 2: Top-k precision curves of **SCTH** and baselines on the three datasets with code length of 64 bits.

Task	Method	MIRFlickr				NUS-WIDE				IAPR-TC12			
		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	**SCTH-$\mathcal{T}$**	0.849	0.846	0.856	0.854	0.718	0.757	0.771	0.764	0.616	0.631	0.634	0.637
	SCTH-$\mathcal{C}$	0.879	0.905	0.923	0.926	0.790	0.830	0.850	0.856	0.596	0.646	0.685	0.694
	SCTH-$\mathcal{S}$	0.885	0.907	0.924	0.928	0.790	0.826	0.848	0.852	0.610	0.648	0.663	0.669
	SCTH	**0.894**	**0.916**	**0.933**	**0.938**	**0.811**	**0.838**	**0.857**	**0.863**	**0.620**	**0.661**	**0.699**	**0.705**
T → I	**SCTH-$\mathcal{T}$**	0.834	0.845	0.843	0.848	0.725	0.742	0.761	0.760	0.621	0.643	0.644	0.645
	SCTH-$\mathcal{C}$	0.873	0.905	0.922	0.928	0.792	0.832	0.850	0.855	0.590	0.649	0.686	0.696
	SCTH-$\mathcal{S}$	0.884	0.908	0.925	0.926	0.794	0.831	0.844	0.852	0.614	0.647	0.664	0.669
	SCTH	**0.897**	**0.915**	**0.935**	**0.937**	**0.810**	**0.839**	**0.857**	**0.862**	**0.628**	**0.663**	**0.699**	**0.706**

Table 2: The mAP results of **SCTH** and variants on MIRFlickr, NUS-WIDE, and IAPR-TC12 datasets.

(Hu et al. 2023), PT-FUCH (Li et al. 2023), CMCL (Wu et al. 2024), SCH (Hu et al. 2024), GCMLH (Shen et al. 2024a), SMSH (Kuai, Long, and Yang 2025), and VTM-UCH (Fan and Cao 2025) are compared with **SCTH** in retrieving textual data by visual query (I2T) and retrieving visual data by text query (T2I).

To access the retrieval performance of **SCTH**, mean average precision (**mAP**) and top-k precision (**P**@k) curves operate as gauges for assessment. Specifically, the number of retrieved instances is set to 50.

Implementation details The hyper-parameters α, β, γ remain uniform on the three datasets with the settings of α = 1.0, β = 0.1, γ = 1000. Besides, we adopt Adam optimizer with parameters betas = (0.5, 0.999) and set the learning rate of fusion, image, text, and clustering networks to 0.0001, 0.001, 0.001, 0.0001, respectively. Moreover, for the cluster number, we set 25, 10, and 100 for MIRFlickr, NUS-WIDE, and IAPR-TC12 datasets, respectively.

For the baselines, we realized their works via open-source codes. Experiments were conducted on a server with Intel Xeon Silver 4210 CPU @2.20 GHz, NVIDIA GeForce RTX 4060, 128 GB RAM.

Results

Table 1 presents the mAP scores for each method on the three datasets, utilizing hash code lengths from 16 to 128 bits. The top-k precision curves are illustrated in Fig.2. The analysis of results yields the following key insights:

- **SCTH** surpassed all the baselines. Specifically, on the I2T task, the average mAP scores of **SCTH** exceed the best baselines by 1.2%, 1.3%, and 1.5% on MIRFlickr, NUS-WIDE, and IAPR-TC12 datasets. Especially on the T2I task, **SCTH** exhibits improvements of at least 4.4%, 5.3%, and 1.8% on the three datasets, respectively.
- The mAP scores of most methods increase with longer hash code lengths. Additionally, the top-k precision curves exhibit a parallel trend to the mAP scores, demonstrating an downward trajectory alongside the increase in the number of instances retrieved.
- Although CIRH and VTM-UCH are efficient hashing methods, **SCTH** achieves higher mAP scores. The

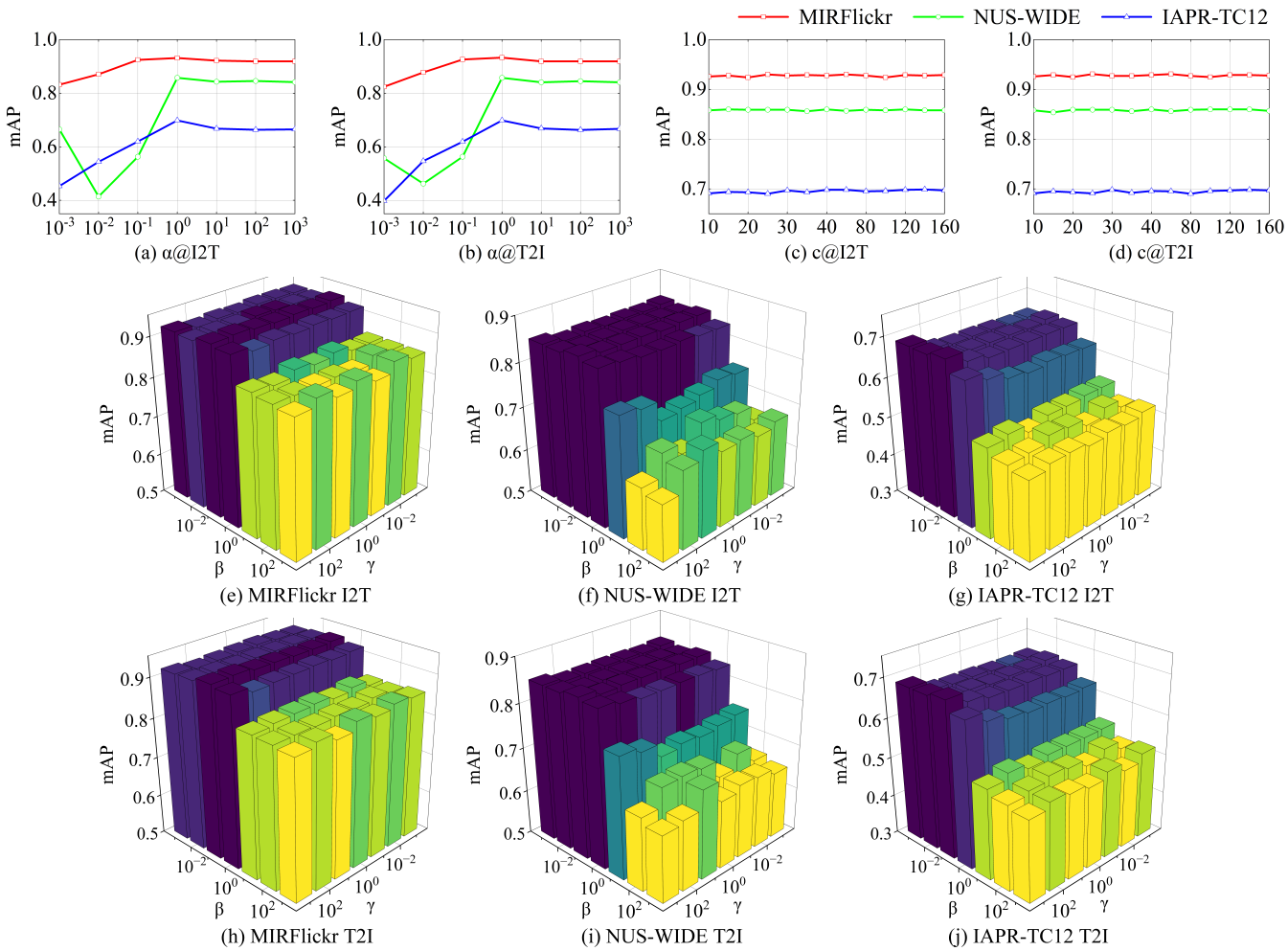


Figure 3: Parameter sensitivity analysis of α, β, γ, c on MIRFlickr, NUS-WIDE, and IAPR-TC12 datasets.

improvement is attributed to employing the pseudo-clustering, which generates assignment statistics vectors and incorporates the cross-modal contrastive loss with adaptive entropy regularization, ensuring balanced and diverse cluster assignments, promoting the inter-modal semantic representation.

Ablation Experiments

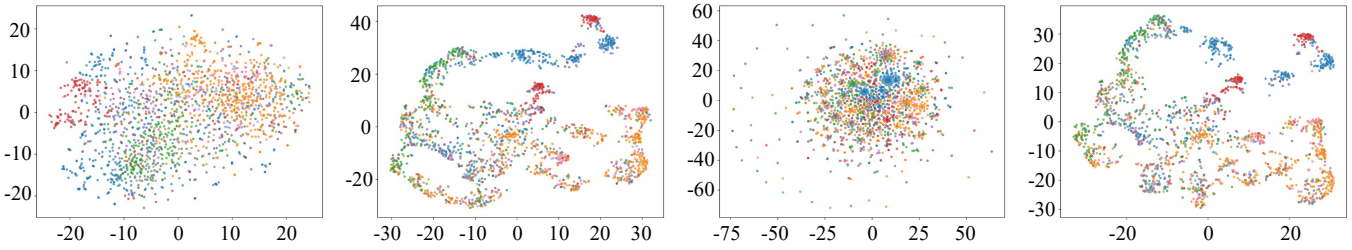
Impacts of modality fusion Transformer To verify the influence of modality fusion Transformer encoder, variant **SCTH**-$\mathcal{T}$ replaced the Transformer encoder with image and text MLPs, respectively, and accordingly removed $\mathcal{L}^{I}_{\text{trans}}$ and $\mathcal{L}^{T}_{\text{trans}}$ from $\mathcal{L}_{\text{fusion}}$. Results in Table.3 reveals that the modality fusion Transformer encoder is superior to directly using MLP to generate hash codes, which indicates that the proposed multimodal representation fusion based on Transformer can effectively improve the quality of hash codes.

Impacts of soft-clustering To verify the impact of soft-clustering module, variant **SCTH**-$\mathcal{C}$ abandoned the pseudo-classifier clustering and removed $\mathcal{L}_{\text{cluster}}$ from $\mathcal{L}$. Results in Table.3 reveals that imposing constraints on the cluster-level loss is more available than removing the loss, which indicates that the proposed individual modality-specific soft clustering can efficiently improve the retrieval performance, especially for low-bit hash codes.

Impacts of Markovian stationary To verify the effect of Markovian stationary distribution module, variant **SCTH**-$\mathcal{S}$ removed $\mathcal{L}_{\text{steady}}$ from $\mathcal{L}$. Results in Table.3 reveals that utilizing Markovian stationary to force the hash codes to maintain consistency in the inter-modal similarity ordering within crucial regions can improve the retrieval precision.

Parameters Sensitivity Analysis

To ascertain the influence of diverse parameters on the three datasets, we conduct the grid search for α, β, γ, c, preserving the remaining parameters constant with the hash code length of 64 bits. The mAP scores are presented in Fig.3, which demonstrate stable retrieval accuracy for a range of parameters, that is, $\alpha \in [10^0, 10^3]$, $\beta \in [10^{-3}, 10^0]$, $\gamma \in [10^{-3}, 10^3]$. Primarily, it is evident that **SCTH** achieves enhanced retrieval precision when α, β, γ are set to moderate values, *i.e.*, 10^0, and there exists a non-significant impact for



(a) Original image embeddings (b) Image based on pseudo-cluster (c) Original text embeddings (d) Text based on pseudo-cluster

Figure 4: t-SNE visualization on MIRFlickr dataset with code length of 64 bits.

Figure 5: The top-30 retrieval results on MIRFlickr dataset.

cluster number c on the results. Consequently, the parameters exhibit insensitivity to cross-modal tasks across a wide value spectrum, highlighting the generalization of **SCTH**.

Embedding Visualization

The t-distributed stochastic neighbor embedding (t-SNE) is utilized to visualize diverse embeddings on MIRFlickr dataset in Fig.4. Evidently, compared with the original embeddings, obvious separation is achieved after performing the proposed individual modality-specific soft clustering, which demonstrates that **SCTH** exhibits superior discriminability in a common latent space.

Case Study

This section exhibits the practical case of top 30 image retrievals from text query via **SCTH** with code length of 64 bits on MIRFlickr dataset. The instance with red border is marked as a retrieval error, which is visually distinctive from the other. Note that the query text semantically contains delicious food, such that vegetables, meat, eggs, and dairy products are involved, and the correct samples contain the corresponding elements. In consequence, Fig.5 demonstrates that **SCTH** performs superior in practical applications.

Conclusion

In this paper, we propose an unsupervised Transformer-based cross-modal hashing method, which integrates semantic clustering and stationary optimization to jointly enhance the multimedia retrieval. By utilizing the multimodal fusion Transformer encoder and contrastive hashing, **SCTH** effectively captures cross-modal association and minimizes the semantic gap. Furthermore, the modality-specific soft clustering with entropy-regularized contrastive loss ensures generalized inter-modal alignment, meanwhile, the Markovian stationary distribution enhances feature stability against noise and outliers. Extensive experiments on three benchmark datasets demonstrate that **SCTH** exceeds state-of-the-art hashing methods in retrieval performance.

Acknowledgments

This work was supported in part by the Natural Science Foundation of Hunan Province (2025JJ40057), and in part by the National Natural Science Foundation of China (62202501).

References

Caswell, H. 2001. *Matrix population models: construction, analysis, and interpretation, 2nd Edition*. Sunderland, UK: Sinauer Associates Inc.

Chua, T.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. NUS-WIDE: a real-world web image database from National University of Singapore. In *Proceedings of the 8th ACM International Conference on Image and Video Retrieval, CIVR 2009*. ACM.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, 248–255. IEEE Computer Society.

Duan, S.; Sun, Y.; Peng, D.; Liu, Z.; Song, X.; and Hu, P. 2025. Fuzzy Multimodal Learning for Trusted Cross-modal Retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025*, 20747–20756. Computer Vision Foundation / IEEE.

Escalante, H. J.; Hernández, C. A.; González, J. A.; López-López, A.; Montes-y-Gómez, M.; Morales, E. F.; Sucar,

L. E.; Pineda, L. V.; and Grubinger, M. 2010. The segmented and annotated IAPR TC-12 benchmark. *Comput. Vis. Image Underst.*, 114(4): 419–428.

Fan, H.; and Cao, Y. 2025. Vision-guided Text Mining for Unsupervised Cross-modal Hashing with Community Similarity Quantization. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence*, 2843–2851. AAAI Press.

Hu, P.; Zhu, H.; Lin, J.; Peng, D.; Zhao, Y.; and Peng, X. 2023. Unsupervised Contrastive Cross-Modal Hashing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(3): 3877–3889.

Hu, Z.; Cheung, Y.; Li, M.; and Lan, W. 2024. Cross-Modal Hashing Method With Properties of Hamming Space: A New Perspective. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12): 7636–7650.

Huiskes, M. J.; and Lew, M. S. 2008. The MIR flickr retrieval evaluation. In *Proceedings of the 1st ACM SIGMM International Conference on Multimedia Information Retrieval, MIR 2008*, 39–43. ACM.

Huo, Y.; Qin, Q.; Zhang, W.; Huang, L.; and Nie, J. 2024. Deep Hierarchy-Aware Proxy Hashing With Self-Paced Learning for Cross-Modal Retrieval. *IEEE Trans. Knowl. Data Eng.*, 36(11): 5926–5939.

Jin, H.; Zhang, Y.; Shi, L.; Zhang, S.; Kou, F.; Yang, J.; Zhu, C.; and Luo, J. 2024. An End-To-End Graph Attention Network Hashing for Cross-Modal Retrieval. In *Advances in Neural Information Processing Systems 38: 38th Annual Conference on Neural Information Processing Systems 2024*.

Kuai, M.; Long, J.; and Yang, Z. 2025. Statistical Model-driven Similarity Hashing: Bridging Modalities for Efficient Unsupervised Retrieval. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence*, 11977–11986. AAAI Press.

Li, J.; Li, F.; Zhu, L.; Cui, H.; and Li, J. 2023. Prototype-guided Knowledge Transfer for Federated Unsupervised Cross-modal Hashing. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023*, 1013–1022. ACM.

Li, Y.; Long, J.; Huang, Y.; and Yang, Z. 2025. Adaptive Asymmetric Supervised Cross-Modal Hashing with consensus matrix. *Inf. Process. Manag.*, 62(3): 104037.

Li, Y.; Long, J.; and Yang, Z. 2025. Asymmetric Cross-Modal Hashing Based on Formal Concept Analysis. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence*, 1392–1401. AAAI Press.

Liu, K.; Gong, Y.; Cao, Y.; Ren, Z.; Peng, D.; and Sun, Y. 2024a. Dual Semantic Fusion Hashing for Multi-Label Cross-Modal Retrieval. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024*, 4569–4577. ijcai.org.

Liu, S.; Cao, W.; Fu, R.; Yang, K.; and Yu, Z. 2024b. RPSC: Robust Pseudo-Labeling for Semantic Clustering. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024*, 14008–14016. AAAI Press.

Liu, Y.; Wu, Q.; Zhang, Z.; Zhang, J.; and Lu, G. 2023. Multi-Granularity Interactive Transformer Hashing for Cross-modal Retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023*, 893–902. ACM.

Meyer, C. D. 1994. Sensitivity of the Stationary Distribution of a Markov Chain. *SIAM J. Matrix Anal. Appl.*, 15(3): 715–728.

Pu, R.; Sun, Y.; Qin, Y.; Ren, Z.; Song, X.; Zheng, H.; and Peng, D. 2025. Robust Self-Paced Hashing for Cross-Modal Retrieval with Noisy Labels. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence*, 19969–19977. AAAI Press.

Qin, Q.; Huo, Y.; Huang, L.; Dai, J.; Zhang, H.; and Zhang, W. 2024. Deep Neighborhood-Preserving Hashing With Quadratic Spherical Mutual Information for Cross-Modal Retrieval. *IEEE Trans. Multim.*, 26: 6361–6374.

Shen, X.; Chen, Y.; Liu, W.; Zheng, Y.; Sun, Q.; and Pan, S. 2024a. Graph Convolutional Multi-Label Hashing for Cross-Modal Retrieval. *IEEE Trans. Neural Networks Learn. Syst.*, 1–13.

Shen, X.; Huang, Q.; Lan, L.; and Zheng, Y. 2024b. Contrastive Transformer Cross-Modal Hashing for Video-Text Retrieval. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024*, 1227–1235. ijcai.org.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations, ICLR 2015*.

Srivastava, S.; and Sharma, G. 2024. OmniVec2 - A Novel Transformer Based Network for Large Scale Multimodal and Multitask Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, 27402–27414. IEEE.

Sun, Y.; Liu, K.; Li, Y.; Ren, Z.; Dai, J.; and Peng, D. 2024. Distribution Consistency Guided Hashing for Cross-Modal Retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024*, 5623–5632. ACM.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: 31st Annual Conference on Neural Information Processing Systems 2017*, 5998–6008.

Wang, H.; Yao, M.; Jiang, G.; Mi, Z.; and Fu, X. 2024a. Graph-Collaborated Auto-Encoder Hashing for Multiview Binary Clustering. *IEEE Trans. Neural Networks Learn. Syst.*, 35(7): 10121–10133.

Wang, J.; Zeng, Z.; Chen, B.; Wang, Y.; Liao, D.; Li, G.; Wang, Y.; and Xia, S. 2024b. Hugs Bring Double Benefits: Unsupervised Cross-Modal Hashing with Multi-granularity Aligned Transformers. *Int. J. Comput. Vis.*, 132(8): 2765–2797.

Wang, L.; Yang, J.; Zareapoor, M.; and Zheng, Z. 2021. Cluster-wise unsupervised hashing for cross-modal similarity search. *Pattern Recognit.*, 111: 107732.

Wu, Q.; Zhang, Z.; Liu, Y.; Zhang, J.; and Nie, L. 2024. Contrastive Multi-Bit Collaborative Learning for Deep Cross-Modal Hashing. *IEEE Trans. Knowl. Data Eng.*, 36(11): 5835–5848.

Yang, F.; Ding, X.; Liu, Y.; Ma, F.; and Cao, J. 2022. Scalable semantic-enhanced supervised hashing for cross-modal retrieval. *Knowl. Based Syst.*, 251: 109176.

Yang, F.; Han, M.; Ma, F.; Liu, Y.; Ding, X.; and Tong, D. 2024. Disperse Asymmetric Subspace Relation Hashing for Cross-Modal Retrieval. *IEEE Trans. Circuits Syst. Video Technol.*, 34(1): 603–617.

Yang, F.; Liu, X.; Ma, F.; Ding, X.; and Wang, K. 2025. Online Asymmetric Supervised Discrete Cross-Modal Hashing for Streaming Multimedia Data. *Pattern Recognit.*, 165: 111604.

Yu, J.; Zhou, H.; Zhan, Y.; and Tao, D. 2021. Deep Graph-neighbor Coherence Preserving Network for Unsupervised Cross-modal Hashing. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, 4626–4634. AAAI Press.

Yu, Y.; Lu, Z.; Nie, F.; Yu, W.; Miao, Z.; and Li, X. 2025. Pseudo-Label Guided Bidirectional Discriminative Deep Multi-View Subspace Clustering. *IEEE Trans. Knowl. Data Eng.*, 37(7): 4213–4224.

Zeng, D.; Yu, Y.; and Oyama, K. 2020. Deep Triplet Neural Networks with Cluster-CCA for Audio-Visual Cross-Modal Retrieval. *ACM Trans. Multim. Comput. Commun. Appl.*, 16(3): 76:1–76:23.

Zeng, Z.; Sun, Y.; and Mao, W. 2021. MCCN: Multimodal Coordinated Clustering Network for Large-Scale Cross-modal Retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia, MM 2021*, 5427–5435. ACM.

Zhang, B.; Zhang, Y.; Li, J.; Chen, J.; Akutsu, T.; Cheung, Y.; and Cai, H. 2025a. Unsupervised Dual Deep Hashing With Semantic-Index and Content-Code for Cross-Modal Retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(1): 387–399.

Zhang, H.; Yang, Y.; Qi, F.; Qian, S.; and Xu, C. 2025b. Active Supervised Cross-Modal Retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(6): 5112–5126.

Zheng, J.; Liang, M.; Yu, Y.; Li, Y.; and Xue, Z. 2024. Knowledge Graph Enhanced Multimodal Transformer for Image-Text Retrieval. In *40th IEEE International Conference on Data Engineering, ICDE 2024*, 70–82. IEEE.

Zhu, C.; Hu, W.; Hou, J.; Qin, Q.; Zhang, W.; and Huang, L. 2025. Deep adaptive gradient-triplet hashing for cross-modal retrieval. *Expert Syst. Appl.*, 291: 128566.

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 Trans. Knowl. Data Eng.*, 35(9): 8838–8851.