MDPI

Article

Unsupervised Contrastive Graph Kolmogorov-Arnold Networks Enhanced Cross-modal Retrieval Hashing

Hongyu Lin [1]*0009-0008-0312-3256, Shaofeng Shen [2], Yuchen Zhang [2] and Renwei Xia [2]

[1] Dundee International Institute of Central South University, Central South University, Changsha 410083, China
[2] School of Computer Science and Engineering, Central South University, Changsha 410083, China
* Correspondence: 7802220129@csu.edu.cn

Abstract: To address modality heterogeneity and accelerate large-scale retrieval, cross-modal hashing strategies generate compact binary codes that enhance computational efficiency. Existing approaches often struggle with suboptimal feature learning due to fixed activation functions and limited cross-modal interaction. We propose Unsupervised Contrastive Graph Kolmogorov-Arnold Networks (GraphKAN) Enhanced Cross-modal Retrieval Hashing (UCGKANH), integrating GraphKAN with contrastive learning and hypergraph-based enhancement. GraphKAN enables more flexible cross-modal representation through enhanced nonlinear expression of features. We introduce contrastive learning that captures modality-invariant structures through sample pairs. To preserve high-order semantic relations, we construct a hypergraph-based information propagation mechanism, refining hash codes by enforcing global consistency. The efficacy of our UCGKANH approach is validated by thorough tests on the MIR-FLICKR, NUS-WIDE, and MS COCO datasets, which show significant gains in retrieval accuracy coupled with strong computational efficiency.

Keywords: Cross-modal Retrieval; Hashing; Graph Kolmogorov-Arnold Networks; Contrastive Learning; Hypergraph Neural Networks.

1. Introduction

With the growing availability of multimodal data across diverse media platforms, there has been increasing academic attention on cross-modal retrieval in recent years [1]. Among various modalities, image and text data are the most prevalent and voluminous, where cross-modal hashing demonstrates superior retrieval efficiency [2,3]. Cross-modal hashing encodes data into unified binary hash representations and measures the Hamming distance over the binary codes derived from matched image and text features [4,5]. Using the great computational efficiency of the Hamming distance algorithm and the compactness of hash codes for less storage, this method improves large-scale cross-modal retrieval [6,7]. In contrast, cross-modal retrieval methods that project heterogeneous data into a shared embedding space often suffer from significant storage and computational burdens when handling large-scale datasets [8]. Therefore, cross-modal hashing approaches has shown its effectiveness in the era of multi-modal data, and a variety of new approaches building upon these methods have emerged in recent years [9].

The two main categories of cross-modal hashing techniques are shallow and deep learning-based models. Shallow methods rely on manually crafted features, which often fail to generate sufficiently representative and discriminative hash codes [10]. Deep cross-modal hashing techniques, on the other hand, combine hash function creation and feature representation learning into a single architecture, enhancing their capacity to capture

Received:
Revised:
Accepted:
Published:

Citation: Lastname, F.; Lastname, F.; Lastname, F. Title. *Mathematics* **2025**, *1*, 0. https://doi.org/

complex latent semantics [11]. This typically leads to enhanced retrieval accuracy and better scalability [12].Our study focuses on deep cross-modal hashing because of these benefits. Additionally, there are two main classifications for deep cross-modal hashing techniques, including supervised and unsupervised methods, depending on the presence or absence of explicit semantic supervision [13]. Unsupervised deep hashing leverages statistical correlations across modalities to learn hash functions, aiming to extract modality-invariant semantic representations without requiring labeled data [14]. On the other hand, supervised deep hashing relies on predefined image-text similarity labels, which generally contribute to higher retrieval precision and stronger semantic consistency [15]. However, because large-scale labeled datasets are expensive and scarce, this work focuses on unsupervised deep hashing, exploring its potential in learning effective cross-modal hash representations without manual annotations [16,17].

Although deep cross-modal hashing approaches have achieved notable progress, they continue to encounter difficulties in feature representation, aligning global semantic embeddings, and addressing the brevity of text in image-oriented retrieval tasks [18,19]. First, the majority of models use convolutional neural networks (CNNs) for images and Bag-of-Words (BoW) for text as upstream feature extractors. These models use convolutional kernels to extract local features, but they are unable to depend on the total data instance over long distances [20]. While pre-trained transformer networks utilize the attention mechanism to model global correlation of data patches with built-in overall receptive field. Secondly, many cross-modal hashing techniques struggle to achieve precise alignment between the embedding spaces of heterogeneous modalities [21,22]. Many approaches only use pairwise similarity constraints, neglecting high-order relationships between samples, which can lead to suboptimal hash code learning [23]. Additionally, existing methods typically assume a one-to-one correspondence between image-text pairs, whereas real-world multimodal data often exhibit many-to-many relationships [24,25]. Without explicit modeling of such structures, the learned hash codes may not generalize well [26]. Finally, the effectiveness of text-to-image retrieval remains a key challenge. Because visual and textual information are represented differently, many models perform well in image-to-text retrieval, while text-to-image retrieval frequently suffers from semantic inconsistency [27]. Most deep hashing methods struggle to bridge the modality gap, particularly in cases where textual descriptions are abstract or contain high-level concepts that do not have direct visual counterparts [28].

To address the aforementioned limitations, we propose an Unsupervised Contrastive Graph Kolmogorov-Arnold Networks Enhanced Cross-modal Retrieval Hashing model, named UCGKANH. This framework leverages Graph Kolmogorov-Arnold Networks (GraphKAN) [29] to improve cross-modal feature learning and integrates unsupervised contrastive learning to optimize hash code generation. Additionally, a hypergraph-based structure is introduced to model high-order semantic relations between modalities, which makes it easier to create hash codes that are more resilient and discriminative. Here are our contributions:

- We present UCGKANH, an unsupervised cross-modal hashing framework that leverages contrastive learning and is further enhanced by GraphKAN and hypergraph-based modeling. By integrating Kolmogorov–Arnold Networks into the retrieval process, the model achieves more expressive and discriminative feature representations.
- We design an unsupervised contrastive learning strategy tailored for cross-modal hashing. By leveraging instance-level contrastive learning without requiring explicit labels, our method significantly enhances the discrimination and consistency of hash codes across different modalities.

- We incorporate hypergraph-based semantic structure modeling to capture high-order relationships across image-text pairs. This mitigates the shortcomings of traditional graph-based methods and enhances the generalization of the generated hash codes in challenging cross-modal retrieval environments.

2. Related Work

Recent developments in deep cross-modal hashing are reviewed in this section, covering both supervised and unsupervised techniques, along with the integration of Kolmogorov–Arnold Networks.

2.1. Deep Cross-modal Hashing

In order to close the semantic gap between various modalities, prior research on deep cross-modal hashing focuses on projecting multimodal data into a common representation space. Jiang et al. proposed the DCMH originating from CMH that firstly introducing hashing retrieval into cross-modal retrieval [30,31]. Deep learning networks are capable of capturing complex data relationships and handling high-dimensional information. In recent years, the performance of deep cross-modal hashing has been substantially improved through the incorporation of advanced neural modules and enhanced semantic alignment strategies. Deep cross-modal hashing techniques are generally categorized into supervised and unsupervised approaches. For supervised methods, the usage of semantic label works as a guidance of the hashing model to generate the semantically consistent hash codes. Supervised methods typically utilize the semantic label as similarity metric, and EMCHL method indicatively constructs cross-modal similarity relationships on different view [32]. Moreover, DSFSH innovates the similarity matrix construction that preserves both intrinsic inter-modality and intro-modality simultaneously [33].The semantic distribution of different modalities is also heterogeneous, and FedSCMR dealt with this problem by integrating federal learning as overall optimization module by extracting shared cross-modal semantic representation [34]. While supervised methods take the advantage of label to generate rich and accurate semantic hash codes, unsupervised approaches utilize the deep neural network to mine common features and implicit relationships between modalities with better generalization capabilities, for supervised methods may over-fitting in the distribution of data label that contain bias. Contrastive learning is integrated into deep hashing model, works like URWMCH, SACH utilized contrastive learning to improve modality consistency using contrastive loss [35,36]. Graph neural network is also powerful in building complex cross-modal cross-modal relationships, and SGRN builds cross-modal graph relationships both in local and global [37]. Other works like JMGCH and UGRDH utilizing adaptive weight assignment on graph representing networks and combining teacher-student knowledge disillusion [38,39]. Besides, works including HEH and UAHCH integrates hypergraph neural networks that captures the higher-order relationships across cross-modal data [40,41].

Although contrastive learning and graph-based approaches have shown progress, current cross-modal hashing methods continue to encounter difficulties in acquiring rich feature representations and effectively maintaining semantic structures. To address these issues, we propose an unsupervised contrastive graph Kolmogorov-Arnold networks enhanced cross-modal retrieval hashing framework. Unlike prior works, our method integrates Graph Kolmogorov-Arnold Networks (GraphKAN) to improve cross-modal feature learning, employs contrastive learning for robust hash optimization, and utilizes a hypergraph structure to capture high-order semantic relationships.

2.2. Kolmogorov-Arnold Networks

Kolmogorov–Arnold Networks (KANs) show as a promising alternative to replace Multi-Layer Perceptrons (MLPs), which is developed under the inspiration of the Kolmogorov–Arnold representation theorem [42]. Different to MLPs that employ fixed activation mechanisms at each node, KANs utilize adaptive, edge-specific activation functions modeled via univariate spline transformations. This design enables KANs to maintain competitive or even superior performance with fewer parameters, while also improving interpretability through more intuitive visualization of learned behaviors. KANs has demonstrated its potential potential in many areas, showing as an innovation in the deep learning area. GraphKAN is proposed to integrate KANs with graph neural networks to improve its feature extraction ability [29]. The integration of KANs into cross-modal retrieval systems is an emerging research area. By leveraging KANs' ability to learn flexible and interpretable activation functions, this approach holds promise for improving retrieval accuracy and providing more explainable results, addressing some of the shortcomings of the current cross-modal hashing techniques.

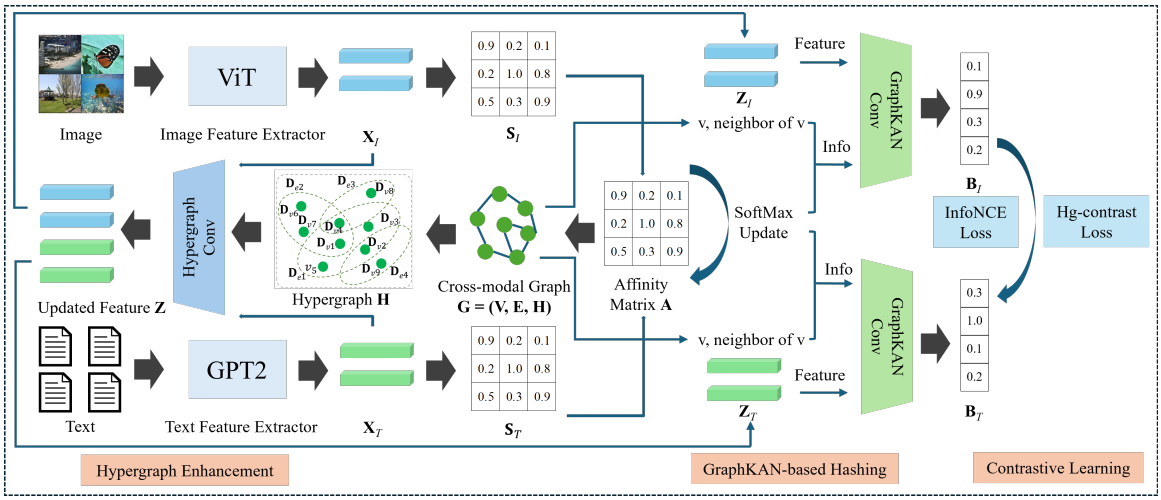


Figure 1. The overall framework of the proposed UCGKANH method.

3. Methodology

We present our designed framework for unsupervised cross-modal hashing in this section. Our approach is designed to address the limitations of existing cross-modal hashing methods in high-order structure modeling, semantic alignment, and feature representation learning. Fig. 1 illustrates the workflow of our designed approach.

3.1. Notation

Throughout this paper, bold uppercase letters (e.g., $\mathbf{M}$) are used to represent matrices, and bold lowercase letters (e.g., $\mathbf{a}$) denotes vectors. Besides, the Frobenius norm of a matrix is expressed as $|\cdot|_F$, while the ReLU activation is denoted by $\sigma(\cdot)$ in the hypergraph convolutional neural network (HGCN). We define the dataset as $\mathcal{D} = (\mathbf{x}_I^i, \mathbf{x}T^i)i = 1^n$, representing a set of paired image and text samples. To be specific, the image and text features as $\mathbf{X}_I \in \mathbb{R}^{n \times d_I}$ and $\mathbf{X}_T \in \mathbb{R}^{n \times d_T}$. Here, d_I, d_T indicate the dimensionality of image and text features respectively, while n denotes the total number of image-text sample pairs. The primary objective is for $f_I(\mathbf{X}_I; \theta_i)$ and $f_T(\mathbf{X}_T; \theta_t)$ these 2 hash functions' training, used to produce a unified binary code $\mathbf{B} \in \{-1, 1\}^{n \times r}$, where r is the the hash codes bit length.

3.2. Model Architecture

3.2.1. Similarity Matrix and Graph Relation Construction

Given the image and text features $\mathbf{X}_I \in \mathbb{R}^{n \times d_I}$ and $\mathbf{X}_T \in \mathbb{R}^{n \times d_T}$, we compute the inner-modal similarity matrices $\mathbf{S}_I = \hat{\mathbf{X}}_I \hat{\mathbf{X}}_I^\top \in [-1, +1]^{n \times n}$ and $\mathbf{S}_T = \hat{\mathbf{X}}_T \hat{\mathbf{X}}_T^\top \in [-1, +1]^{n \times n}$, where $\hat{\mathbf{X}}_I$ and $\hat{\mathbf{X}}_T$ are the normalized features. Then, we linearly combine these self-similarity matrices using a weighting factor α_1:

$$\mathbf{S}_1 = \alpha_1 \mathbf{S}_I + (1 - \alpha_1)\mathbf{S}_T. \tag{1}$$

To enhance the discrimination of similarity scores, we apply a Gaussian kernel weighting followed by an exponential transformation and normalization:

$$\begin{aligned} \mathbf{S}_{\text{weighted}} &= \exp\left(-\frac{(1 - \mathbf{S}_1)^2}{2\sigma^2}\right), \\ \mathbf{S}_{\exp} &= \exp(\mathbf{S}_{\text{weighted}}), \\ \mathbf{S} &= (1 - \alpha_2)\mathbf{S}_{\text{weighted}} + \frac{\alpha_2 \mathbf{S}_{\exp}}{n \times n}, \end{aligned} \tag{2}$$

where σ is a bandwidth parameter estimated as the median of pairwise distances in $\mathbf{S}_1$ with a regularization term $\sigma = \max(\text{median}, \epsilon)$, $\epsilon = 0.01$ to prevent degeneration, and $\alpha_2 \in [0, 1]$ controls the influence of the exponential transformation. The parameters $\alpha_1 \in [0, 1]$ and α_2 are tuned via grid search.

To leverage the computed similarity measures for cross-modal learning, we construct a comprehensive graph structure $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$. The vertex set $\mathcal{V}$ comprises image and text instances: $\mathcal{V} = \{v_1, v_2, \ldots, v_{2n}\}$, where v_i corresponds to the i-th image feature for $i \leq n$ and to the $(i - n)$-th text feature for $i > n$. The edge set $\mathcal{E}$ is established based on the similarity matrix $\mathbf{S}$, forming connections between both inter-modal and intra-modal nodes. Specifically, we create an edge $e_{ij} \in \mathcal{E}$ between vertices v_i and v_j if their similarity $\mathbf{S}_{ij}$ exceeds a threshold θ:

$$\mathcal{E} = \{(v_i, v_j) \mid \mathbf{S}_{ij} > \theta, v_i, v_j \in \mathcal{V}\}, \tag{3}$$

where θ is dynamically determined as the median value of all similarities in $\mathbf{S}$. This threshold-based edge construction ensures that the graph captures meaningful semantic relationships while filtering out weak or potentially noisy connections.

The edge weight set $\mathcal{W}$ directly adopts the corresponding similarity values from $\mathbf{S}$:

$$\mathbf{W}_{ij} = \mathbf{S}_{ij}, \quad \forall (v_i, v_j) \in \mathcal{E}, \tag{4}$$

which measures how strongly the connected nodes are related. The resulting graph $\mathcal{G}$ serves as the foundation for our GraphKAN layers and subsequent hypergraph enhancement. The adjacency matrix $\mathbf{A} \in \mathbb{R}^{2n \times 2n}$ of graph $\mathcal{G}$ is initially set as:

$$\mathbf{A}_{\text{init},ij} = \begin{cases} \mathbf{S}_{ij}, & \text{if } (v_i, v_j) \in \mathcal{E} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

This graph structure effectively encodes both the local pairwise relationships and the global semantic structure across modalities, offering abundant relational information to support downstream feature extraction and hash code construction. In the subsequent

GraphKAN layers, $\mathbf{A}$ is further refined using a self-attention mechanism to dynamically learn the affinity scores. The updated affinity matrix is computed as:

$$\mathbf{A}_{ij} = \text{softmax}\left(\frac{\mathbf{Q}_i \mathbf{K}_j^\top}{\sqrt{d_k}}\right) \cdot \mathbf{V}_{ij}, \tag{6}$$

where the dimension of the key vectors is represented as d_k. Here, $\mathbf{Q}_i$, $\mathbf{K}_j$, and $\mathbf{V}_{ij}$ are query, key, and value vectors derived from the node features $\mathbf{Z}_i$ and $\mathbf{Z}_j$ in the GraphKAN layer. By enabling $\mathbf{A}$ to adaptively focus on the most pertinent cross-modal links, this self-attention mechanism enhances the graph structure's resilience for hashing.

3.2.2. Hypergraph Enhancement

To further improve the modeling of complex relationships across modalities, we introduce a hypergraph-based enhancement that shares the same vertex set $\mathcal{V}$ with the previously constructed graph $\mathcal{G}$, but captures different relationship patterns. The hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E}_H, \mathcal{W}_H)$ is defined, where $\mathcal{E}_H$ denotes the set of hyperedges that can connect multiple nodes simultaneously, and $\mathcal{W}_H$ specifies the corresponding hyperedge weights.

Hyperedges in $\mathcal{E}_H$ are designed to link groups of nodes that exhibit strong cross-modal semantic coherence. We achieve this by performing clustering on the similarity matrix $\mathbf{S}$ using a spectral clustering technique, which partitions the nodes into K clusters (where K is a tunable parameter). Each cluster is treated as a hyperedge $e_k \in \mathcal{E}_H$, connecting all nodes within that group. The weight of each hyperedge $w(e_k) \in \mathcal{W}_H$ is calculated as the mean similarity among the nodes it connects:

$$w(e_k) = \frac{1}{|e_k|^2} \sum_{i,j \in e_k} \mathbf{S}_{ij}, \tag{7}$$

where the number of nodes in the hyperedge e_k is represented by $|e_k|$. The hypergraph's structure is encoded using the incidence matrix $\mathbf{H} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{E}_H|}$, with entries defined as:

$$\mathbf{H}_{v,e} = \begin{cases} 1, & \text{if node } v \in e, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

Here, $v \in \mathcal{V}$ and $e \in \mathcal{E}_H$. The following formula is used to calculate the node degree $d(v)$ and hyperedge degree $\delta(e)$:

$$\begin{aligned} d(v) &= \sum_{e \in \mathcal{E}_H} \mathbf{H}_{v,e}, \\ \delta(e) &= \sum_{v \in \mathcal{V}} \mathbf{H}_{v,e}. \end{aligned} \tag{9}$$

The diagonal degree matrices for nodes and hyperedges are denoted as $\mathbf{D}_v$ and $\mathbf{D}_e$, respectively, where $\mathbf{D}_e(e,e) = \delta(e)$ and $\mathbf{D}_v(v,v) = d(v)$.

To incorporate the hypergraph into the hashing process, we apply a hypergraph convolutional layer that aggregates information across hyperedges. The updated feature representation $\mathbf{Z}$ for nodes is obtained by:

$$\mathbf{Z} = \sigma\left(\mathbf{D}_v^{-\frac{1}{2}} \mathbf{H} \mathbf{W}_H \mathbf{D}_e^{-1} \mathbf{H}^\top \mathbf{D}_v^{-\frac{1}{2}} \mathbf{X}\right), \tag{10}$$

where the concatenated feature matrix is represented by $\mathbf{X} = [\mathbf{X}_I; \mathbf{X}_T]$, the learnable weight matrix is represented by $\mathbf{W}_H$, and the ReLU activation function is represented by $\sigma(\cdot)$. This operation normalizes the hypergraph structure while propagating high-order relational information to refine the feature representations, which are then used to create the final

hash codes $\mathbf{B}$ by putting them into the hash function $f(\cdot)$. This hypergraph enhancement improves the quality of the unified hash code $\mathbf{B} \in \{-1, 1\}^{n \times r}$ for retrieval tasks by allowing the model to capture complex cross-modal dependencies.

3.2.3. GraphKAN-based Hashing

To further enhance the cross-modal hashing process by leveraging the expressive power of Kolmogorov-Arnold Networks (KANs) on graph structures, we introduce GraphKAN, a novel graph-based neural network tailored for unsupervised hash code learning. GraphKAN integrates the hypergraph-enhanced features with a KAN-based architecture to capture intricate non-linear relationships across modalities, thereby the unified hash codes $\mathbf{B}$ are of higher quality.

The GraphKAN model operates on the cross-modal graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$ and the hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E}_H, \mathcal{W}_H)$ constructed earlier. The input feature matrix $\mathbf{X} = [\mathbf{X}_I; \mathbf{X}_T] \in \mathbb{R}^{2n \times d}$ (where $d = d_I + d_T$) is first processed by the hypergraph convolution to obtain the enhanced feature matrix $\mathbf{Z}$. This $\mathbf{Z}$ serves as the input to the GraphKAN layer. The GraphKAN layer replaces traditional graph convolution operations with KAN-based transformations. For each node $v \in \mathcal{V}$, the feature update is defined as:

$$\mathbf{z}_v^{(l+1)} = \sum_{u \in \mathcal{N}(v)} \mathbf{A}_{vu} \cdot \phi_{vu}\left(\mathbf{z}_u^{(l)}, \mathbf{z}_v^{(l)}\right) \cdot \mathbf{W}^{(l)} + \mathbf{b}^{(l)}, \tag{11}$$

where $\phi_{vu}(\cdot, \cdot)$ is parameterized with $K = 5$ B-spline basis functions, and $B_k(\cdot)$ are learnable B-spline functions with coefficients c_k. The neighborhood of node v is indicated by $\mathcal{N}(v)$, which includes both graph edges $\mathcal{E}$ and hypergraph hyperedges $\mathcal{E}_H$; the feature vector of node v at layer l is $\mathbf{z}_v^{(l)}$; a learnable weight matrix is $\mathbf{W}^{(l)}$; and a bias term is $\mathbf{b}^{(l)}$. The KAN-based activation function $\phi_{vu}(\cdot, \cdot)$ models the pairwise interaction between nodes u and v. Specifically, ϕ_{vu} is parameterized as a sum of B-spline functions:

$$\phi_{vu}(\mathbf{z}_u, \mathbf{z}_v) = \sum_{k=1}^{K} c_k B_k(\mathbf{z}_u - \mathbf{z}_v), \tag{12}$$

where $B_k(\cdot)$ are B-spline basis functions, c_k are learnable coefficients, and K is the number of spline components.

The hash codes are generated from the output of the last GraphKAN layer $\mathbf{Z}^{(L)}$. For ensuring the $\mathbf{B}$'s binary nature, we apply a sign activation function followed by a quantization step:

$$\mathbf{B} = \text{sign}\left(\mathbf{Z}^{(L)}\mathbf{W}_h + \mathbf{b}_h\right), \tag{13}$$

where $\mathbf{W}_h$ and $\mathbf{b}_h$ are learnable parameters for the hashing layer, and the $\text{sign}(\cdot)$ function maps values to $\{-1, 1\}$.

3.2.4. Contrastive Learning for Cross-Modal Alignment

We integrate a contrastive learning technique into our system to guarantee semantic alignment across picture and text modalities while maintaining the discriminative features of the hash codes. This approach leverages the GraphKAN-enhanced features and the hypergraph structure to align cross-modal representations in the Hamming space.

For image and text modalities, let $\mathbf{Z}_I^{(L)} \in \mathbb{R}^{n \times d}$ and $\mathbf{Z}_T^{(L)} \in \mathbb{R}^{n \times d}$ be the enhanced features derived from the final GraphKAN layer, respectively. $\mathbf{B}_I = \text{sign}(\mathbf{Z}_I^{(L)}\mathbf{W}_h + \mathbf{b}_h)$ and $\mathbf{B}_T = \text{sign}(\mathbf{Z}_T^{(L)}\mathbf{W}_h + \mathbf{b}_h)$ are modality-specific hash codes generated by these features, where $\mathbf{W}_h$ and $\mathbf{b}_h$ are shared parameters to ensure consistency across modalities. We define the contrastive loss $\mathcal{L}_{\text{contrast}}$ based on the InfoNCE loss, where $\cos(\mathbf{b}_i, \mathbf{b}_j) = \frac{\mathbf{b}_i \cdot \mathbf{b}_j}{|\mathbf{b}_i||\mathbf{b}_j|}$ is the

cosine similarity. To reduce computational cost, negative samples in $\mathcal{L}_{\text{hg-contrast}}$ are selected using a random sampling strategy with a fixed size of 100 per node. For a given image-text pair $(\mathbf{x}_I^i, \mathbf{x}_T^i)$, the contrastive loss is formulated as:

$$\mathcal{L}_{\text{contrast}} = -\frac{1}{n}\sum_{i=1}^{n}\left[\log\frac{\exp(\cos(\mathbf{b}_I^i, \mathbf{b}_T^i)/\tau)}{\sum_{j=1}^{n}\exp(\cos(\mathbf{b}_I^i, \mathbf{b}_T^j)/\tau)} \log\frac{\exp(\cos(\mathbf{b}_T^i, \mathbf{b}_I^i)/\tau)}{\sum_{j=1}^{n}\exp(\cos(\mathbf{b}_T^i, \mathbf{b}_I^j)/\tau)}\right], \tag{14}$$

where $\cos(\cdot,\cdot)$ is the cosine similarity function, $\tau > 0$ is a temperature parameter regulating the concentration of the distribution, and $\mathbf{b}_I^i$ and $\mathbf{b}_T^i$ are the hash codes for the i-th image and text instance. The paired image-text instance $(\mathbf{b}_I^i, \mathbf{b}_T^i)$ is encouraged to have comparable hash codes by the first term, while the second term ensures symmetry by considering the text-to-image alignment.

To further enhance the contrastive learning process, we leverage the hypergraph $\mathcal{H}$ to guide the selection of positive and negative pairs. Specifically, hyperedges in $\mathcal{E}_H$ naturally define clusters of semantically related nodes across modalities. For each node $v_i \in \mathcal{V}$, we define its positive set $\mathcal{P}(v_i)$ as the set of nodes connected through the same hyperedge:

$$\mathcal{P}(v_i) = \{v_j \mid \exists e \in \mathcal{E}_H \text{ such that } \mathbf{H}_{v_i,e} = 1 \text{ and } \mathbf{H}_{v_j,e} = 1\}, \tag{15}$$

and the negative set $\mathcal{N}(v_i) = \mathcal{V} \setminus \mathcal{P}(v_i)$. The hypergraph-guided contrastive loss is then modified to focus on these sets:

$$\mathcal{L}_{\text{hg-contrast}} = -\frac{1}{n}\sum_{i=1}^{n}\left[\log\frac{\sum_{v_j\in\mathcal{P}(v_i)}\exp(\cos(\mathbf{b}_i, \mathbf{b}_j)/\tau)}{\sum_{v_k\in\mathcal{V}}\exp(\cos(\mathbf{b}_i, \mathbf{b}_k)/\tau)}\right], \tag{16}$$

where the hash code of node v_i is $\mathbf{b}_i$, and the summation over $\mathcal{P}(v_i)$ encourages similarity within the hyperedge cluster, while the denominator includes all nodes to contrast against negatives.

3.3. Overall Objective Function

The final objective function minimizes the difference between the binary hash codes and the continuous features by combining the contrastive loss with a quantization regularization term:

$$\mathcal{L} = \mathcal{L}_{\text{hg-contrast}} + \lambda\|\mathbf{Z}^{(L)} - \mathbf{B}\|_F^2, \tag{17}$$

where $\mathbf{Z}^{(L)} = [\mathbf{Z}_I^{(L)}; \mathbf{Z}_T^{(L)}]$ is the concatenated feature matrix from the GraphKAN layer, $\mathbf{B} = [\mathbf{B}_I; \mathbf{B}_T]$ is the matrix of unified hash codes, and $\lambda > 0$ is a hyperparameter balancing the two terms. The Frobenius norm $\|\cdot\|_F$ ensures that the continuous features $\mathbf{Z}^{(L)}$ are close to the binary hash codes $\mathbf{B}$, reducing quantization errors. This contrastive learning strategy, enhanced by hypergraph guidance, ensures that the hash codes capture both pairwise semantic alignment and higher-order structural relationships across modalities, leading to improved retrieval performance.

4. Experiment

The experimental setup and evaluation outcomes of the suggested model on two cross-modal retrieval tasks—text-to-image (T→I) and image-to-text (I→T)—are detailed in this section.

Algorithm 1 UCGKANH Algorithm

Input: Image features $\mathbf{X}_I \in \mathbb{R}^{n \times d_I}$, text features $\mathbf{X}_T \in \mathbb{R}^{n \times d_T}$, hash length r, layers L, λ, τ, clusters K
Output: Trained hash model parameters $\theta = \{\mathbf{W}_h, \mathbf{b}_h, \mathbf{W}^{(l)}, \mathbf{b}^{(l)}, c_k\}$ and hash codes $\mathbf{B} \in \{-1, 1\}^{n \times r}$
Normalize $\mathbf{X}_I$, $\mathbf{X}_T$ to $\hat{\mathbf{X}}_I$, $\hat{\mathbf{X}}_T$
Compute $\mathbf{S}_I = \hat{\mathbf{X}}_I \hat{\mathbf{X}}_I^\top$, $\mathbf{S}_T = \hat{\mathbf{X}}_T \hat{\mathbf{X}}_T^\top$
Compute $\mathbf{S}$ using Eq. (1) and Eq. (2)
Build $\mathbf{A}_{\text{init}} = \mathbf{S}$ and $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$ using Eq. (3), Eq. (4), and Eq. (5)
Cluster $\mathbf{S}$ into K groups form $\mathcal{E}_H$ and set $w(e_k)$ using Eq. (7)
Build hypergraph $\mathbf{H}$ using Eq. (8)
Compute $d(v)$ and $\delta(e)$ using Eq. (9), form $\mathbf{D}_v$, $\mathbf{D}_e$
Initialize model parameters $\theta = \{\mathbf{W}_h, \mathbf{b}_h, \mathbf{W}^{(l)}, \mathbf{b}^{(l)}, c_k\}$
Update $\mathbf{Z} \in [\mathbf{X}_I, \mathbf{X}_T]$ from hypergraph convolution layer using Eq. (10)
for $l = 1$ to L **do**
    Compute $\mathbf{Q}_i = \mathbf{W}_Q \mathbf{z}_i$, $\mathbf{K}_j = \mathbf{W}_K \mathbf{z}_j$, $\mathbf{V}_{ij} = \mathbf{W}_V \mathbf{z}_j$
    Update $\mathbf{A}_{ij} = \text{softmax}(\frac{\mathbf{Q}_i \mathbf{K}_j^\top}{\sqrt{d_k}}) \cdot \mathbf{V}_{ij}$ using Eq. (6)
    for each $v \in \mathcal{V}$ **do**
        Compute $\mathbf{z}_v^{(l+1)}$ using Eq. (11) and Eq. (12)
    end for
end for
Set $\mathbf{B} = \text{sign}(\mathbf{Z}^{(L)} \mathbf{W}_h + \mathbf{b}_h)$ using Eq. (13), $\mathbf{Z}^{(L)} = [\mathbf{Z}_I^{(L)}; \mathbf{Z}_T^{(L)}]$
Split $\mathbf{B}$ into $\mathbf{B}_I$, $\mathbf{B}_T$
Define $\mathcal{P}(v_i)$ using Eq. (15)
Compute $\mathcal{L}_{\text{hg-contrast}}$ using Eq. (16)
Compute $\mathcal{L} = \mathcal{L}_{\text{hg-contrast}} + \lambda \|\mathbf{Z}^{(L)} - \mathbf{B}\|_F^2$ using Eq. (17)
Optimize $\mathcal{L}$ to update parameters θ
return Trained model parameters θ and hash codes $\mathbf{B}$

4.1. Datasets Description

As common benchmarks in cross-modal retrieval studies, MIRFlickr-25K, NUS-WIDE, and MS COCO are three well-known datasets that we experimented with to assess the performance of our proposed model. MIRFlickr-25K comprises 25,000 image-tag pairs spanning 24 different concepts. After filtering out instances with fewer than 20 tags, we retained 20,015 pairs, with subsets of 2,000 for query, 5,000 for training, and 18,015 for retrieval. NUS-WIDE contains a total of 269,648 image-text pairs categorized under 81 distinct concepts. We focused on 186,577 pairs belonging to the top 10 categories, partitioning them into 2,000 query pairs, 5,000 training pairs, and 184,577 retrieval pairs. MS COCO features 123,287 image-text pairs distributed across 80 categories. For evaluation, we designated 2,000 pairs for querying, 5,000 for training, and the remaining 121,287 for retrieval.

4.2. Baselines and Evaluation Criteria

We compare our unsupervised cross-modal retrieval hashing method with several of the top unsupervised hashing techniques in cross-modal retrieval to assess its performance. The selected baselines include Deep Binary Reconstruction (DBRC) [43], Correlation Identity Representation Hashing (CIRH) [44], Collective Matrix Factorization Hashing (CMFH) [45], Cross-View Hashing (CVH) [46], Aggregation-based Graph Convolutional Hashing (AGCH) [47], CLIP-based Fusion-modal Reconstructing Hashing (CFRH) [48], Unsupervised Deep Cross-Modal Hashing (UDCMH) [49], Inter-Media Hashing (IMH) [50], Deep Joint-Semantics Reconstructing Hashing (DJSRH) [51], Joint-modal Distribution-based Similarity Hashing (JDSH) [44], rapid image-text cross-modal hash retrieval (RICH) [52],

Table 1. The mAP results across three datasets of each method.

Task	Method	MIRFlickr-25K				NUS-WIDE				MS COCO			
		16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
I → T	CVH	0.606	0.599	0.596	0.598	0.372	0.362	0.406	0.390	0.505	0.509	0.519	0.510
	IMH	0.612	0.601	0.592	0.579	0.470	0.473	0.476	0.459	0.570	0.615	0.613	0.587
	LCMH	0.559	0.569	0.585	0.593	0.354	0.361	0.389	0.383	—	—	—	—
	CMFH	0.621	0.624	0.625	0.627	0.455	0.459	0.465	0.467	0.621	0.669	0.525	0.562
	LSSH	0.584	0.599	0.602	0.614	0.481	0.489	0.507	0.507	0.652	0.707	0.746	0.773
	DBRC	0.617	0.619	0.620	0.621	0.424	0.459	0.447	0.447	0.567	0.591	0.617	0.627
	RFDH	0.632	0.636	0.641	0.652	0.488	0.492	0.494	0.508	—	—	—	—
	UDCMH	0.689	0.698	0.714	0.717	0.511	0.519	0.524	0.558	—	—	—	—
	DJSRH	0.810	0.843	0.862	0.876	0.724	0.773	0.798	0.817	0.678	0.724	0.743	0.768
	AGCH	0.865	0.887	0.892	0.912	0.809	0.830	0.831	0.852	0.741	0.772	0.789	0.806
	CIRH	0.901	0.913	0.929	0.937	0.815	0.836	0.854	0.862	0.797	0.819	0.830	0.849
	RICH	0.869	0.875	0.908	0.925	0.790	0.806	0.842	0.852	—	—	—	—
	CFRH	0.902	0.914	0.936	0.945	0.807	0.824	0.854	0.859	0.845	0.895	0.916	0.928
	UCGKANH	**0.908**	**0.922**	**0.940**	**0.948**	**0.818**	**0.837**	**0.857**	**0.865**	**0.860**	**0.919**	**0.929**	**0.946**
T → I	CVH	0.591	0.583	0.576	0.576	0.401	0.384	0.442	0.432	0.543	0.553	0.560	0.542
	IMH	0.603	0.595	0.589	0.580	0.478	0.483	0.472	0.462	0.641	0.709	0.705	0.652
	LCMH	0.561	0.569	0.582	0.582	0.376	0.387	0.408	0.419	—	—	—	—
	CMFH	0.642	0.662	0.676	0.685	0.529	0.577	0.614	0.645	0.627	0.667	0.554	0.595
	LSSH	0.637	0.659	0.659	0.672	0.577	0.617	0.642	0.663	0.612	0.682	0.742	0.795
	DBRC	0.618	0.626	0.626	0.628	0.455	0.459	0.468	0.473	0.635	0.671	0.697	0.735
	RFDH	0.681	0.693	0.698	0.702	0.612	0.641	0.658	0.680	—	—	—	—
	UDCMH	0.692	0.704	0.718	0.733	0.637	0.653	0.695	0.716	—	—	—	—
	DJSRH	0.786	0.822	0.835	0.847	0.712	0.744	0.771	0.789	0.650	0.753	0.805	0.823
	AGCH	0.829	0.849	0.852	0.880	0.769	0.780	0.798	0.802	0.746	0.774	0.797	0.817
	CIRH	0.867	0.885	0.900	0.901	0.774	0.803	0.810	0.817	0.811	0.847	0.872	0.895
	RICH	0.830	0.843	0.885	0.902	0.771	0.777	0.802	0.822	—	—	—	—
	CFRH	0.874	0.885	0.896	0.910	0.780	0.791	0.798	0.817	0.852	0.903	0.920	0.937
	UCGKANH	**0.879**	**0.896**	**0.907**	**0.914**	**0.792**	**0.807**	**0.815**	**0.826**	**0.861**	**0.917**	**0.923**	**0.949**

and Latent Semantic Sparse Hashing (LSSH) [53]. For performance comparison, we adopt commonly used cross-modal retrieval metrics that rely on Hamming distance. The main evaluation measures include top-50 Mean Average Precision (mAP@50) and top-K retrieval accuracy.

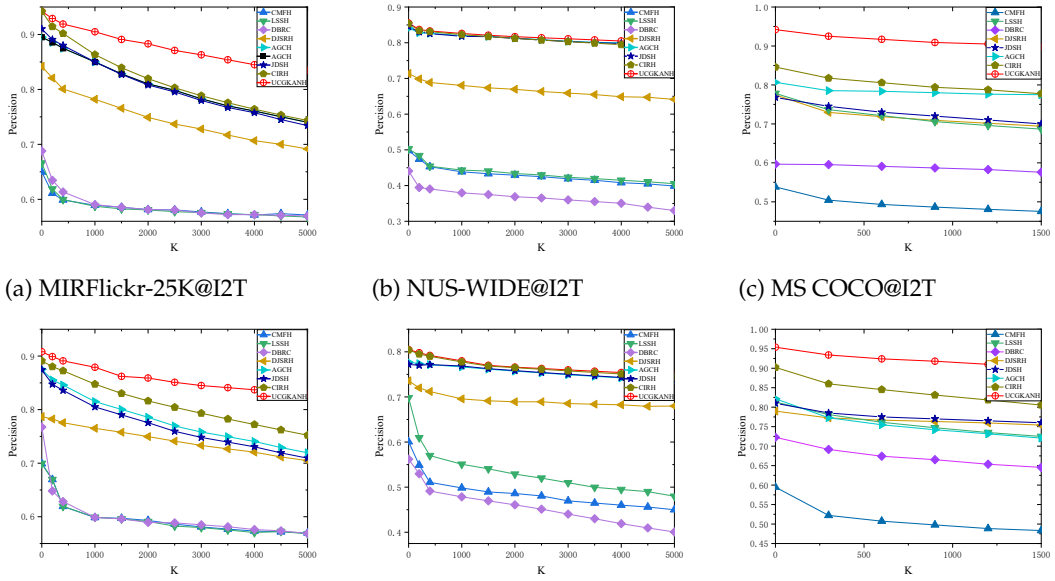
4.3. Comparing with Baseline Methods

4.3.1. mAP Analysis

Our UCGKANH approach outperforms the top-listed methods in terms of retrieval accuracy across a range of hash bit lengths, as evidenced by the mAP results in Table 1. For the I→T task on the MIRFlickr-25K, NUS-WIDE, and MS COCO datasets, our UCGKANH approach improves retrieval accuracy by 0.7% $\sim$ 1.4%, 0.7% $\sim$ 2.6%, and 1.3% $\sim$ 2.4%, respectively. The most notable improvement is on MS COCO at 32 bits, with a 2.4% gain. Similarly, for the T→I task on these datasets, UCGKANH achieves gains of 0.4% $\sim$ 1.1%, 0.4% $\sim$ 3%, and 0.3% $\sim$ 1.4%, respectively, with the largest gain of 3% on NUS-WIDE at 64 bits. These consistent improvements across datasets and hash bit lengths highlight the robustness of UCGKANH. The results underscore the effectiveness of our hypergraph-based approach, which excels at capturing high-order relationships and preserving cross-modal similarities. The significant mAP gains confirm that UCGKANH enhances retrieval performance reliably, making it a strong solution for cross-modal retrieval tasks.

4.3.2. Tok-K Analysis

To assess the ranking quality of UCGKANH in retrieval tasks, we investigate the top-K precision curves in the context of a 128-bit hash code. UCGKANH consistently beats the baseline approaches on all datasets and tasks, as shown by the top-K precision curves. On MIRFlickr-25K (Fig. 2(a) and (d)), UCGKANH exhibits a steep initial decline in precision, indicating high accuracy for the top-ranked results, and maintains a significant lead over other methods as K increases, especially in the I→T task. On NUS-WIDE (Fig. 2(b) and (e)), UCGKANH shows a more gradual decline in precision, reflecting the dataset's diverse categories, yet it consistently achieves higher precision than baselines, particularly in the T→I task where its curve remains above others across all K values. On MS COCO



(a) MIRFlickr-25K@I2T (b) NUS-WIDE@I2T (c) MS COCO@I2T

(d) MIRFlickr-25K@T2I (e) NUS-WIDE@T2I (f) MS COCO@T2I

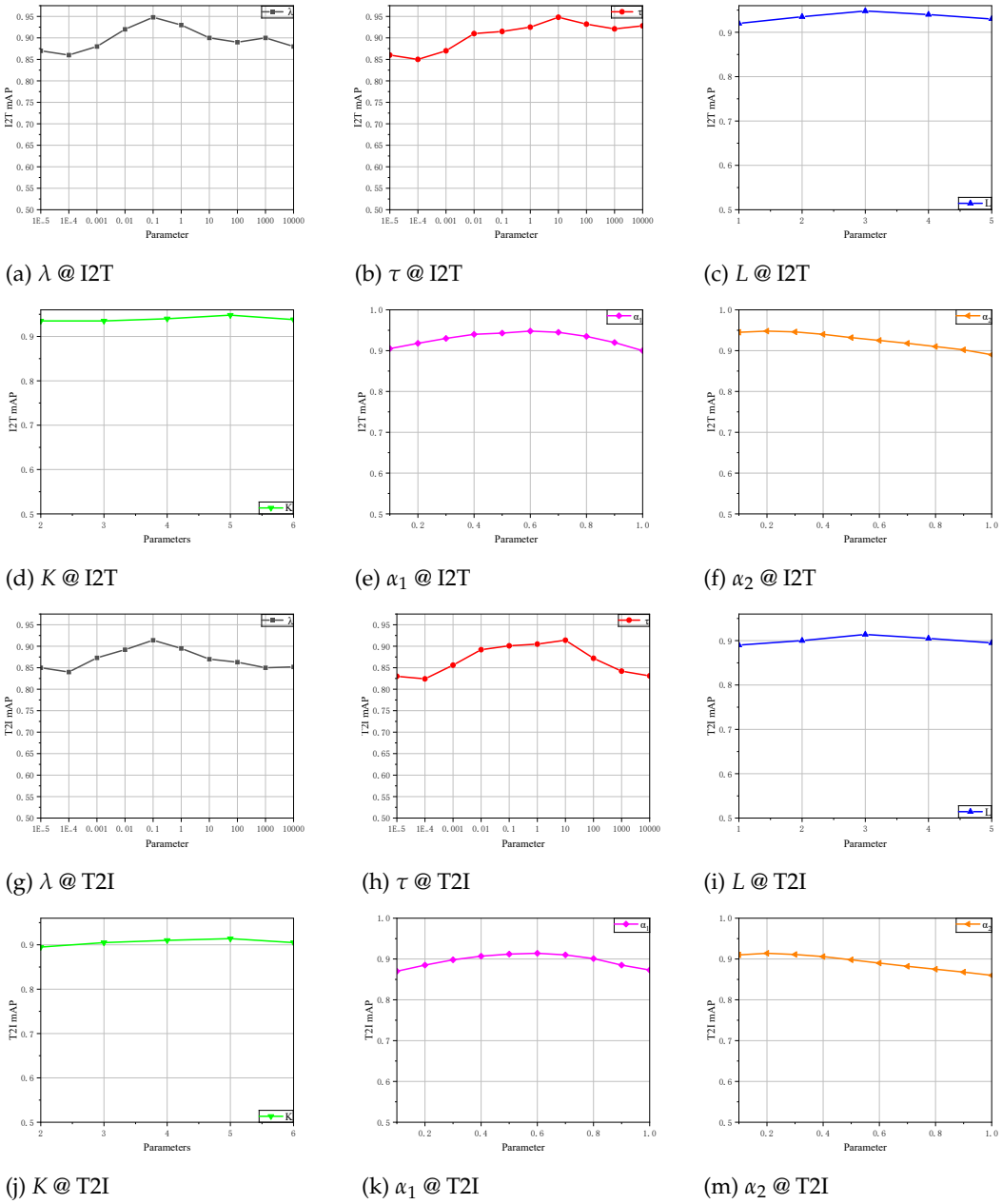
Figure 2. The performance of UCGKANH method with 128-bits in terms of top-K precision on MIRFlickr and NUS-WIDE datasets.

(Fig. 2(c) and (f)), UCGKANH demonstrates a stable and superior performance, with a slower precision drop compared to baselines, highlighting its robustness in handling complex semantic structures for both tasks. Overall, the trends in the top-K precision curves underscore the effectiveness of UCGKANH in delivering high-quality retrieval results across varying dataset complexities. The method's ability to maintain higher precision at larger K values, especially on challenging datasets like NUS-WIDE and MS COCO, suggests the integration including hypergraph enhancement, GraphKAN, and contrastive learning enables UCGKANH to better capture cross-modal semantic relationships and produce more discriminative hash codes compared to existing approaches.

4.4. Parameter Sensitivity Analysis

To evaluate the robustness of UCGKANH, a parameter sensitivity analysis is conducted on the hyperparameters λ (quantization loss weight) and τ (temperature parameter) using 128-bit hash codes on the MIRFlickr-25K dataset. We vary each parameter across a wide range, while keeping other parameters fixed ($\lambda = 0.1$, $\tau = 1$, $L = 3$, $K = 5$, $\alpha_1 = 0.6$, and $\alpha_2 = 0.2$). The impact is illustrated in Fig. 3.

For λ (Fig. 3(a) and (g)), the mAP increases with larger values, peaking at a moderate level, then gradually declines, indicating that a small λ fails to enforce quantization, while a large λ overemphasizes quantization at the expense of semantic alignment. Similarly, τ (Fig. 3(b) and (h)) shows optimal mAP at a medium value, with performance dropping at both extremes, as a small τ sharpens the contrastive loss excessively, risking overfitting, and a large τ flattens the loss, reducing discriminative power. The trends for L (Fig. 3(c) and (i)) and K (Fig. 3(d) and (j)) reveal that mAP improves as these parameters increase, reaching a peak before stabilizing or slightly decreasing, suggesting that moderate values enhance feature learning and high-order relationship modeling, while excessive values may introduce noise or redundancy. In contrast, α_1 (Fig. 3(e) and (k)) and α_2 (Fig. 3(f) and (l)) exhibit minimal impact on mAP, with relatively flat curves across the tested range, indicating that the model is less sensitive to these weights for similarity combination and exponential transformation. Overall, UCGKANH shows greater sensitivity to λ, τ, L, and K, requiring careful tuning to balance quantization, semantic alignment, feature depth, and high-order relationships, while demonstrating robustness to variations in α_1 and α_2.



(a) λ @ I2T (b) τ @ I2T (c) L @ I2T

(d) K @ I2T (e) α_1 @ I2T (f) α_2 @ I2T

(g) λ @ T2I (h) τ @ T2I (i) L @ T2I

(j) K @ T2I (k) α_1 @ T2I (m) α_2 @ T2I

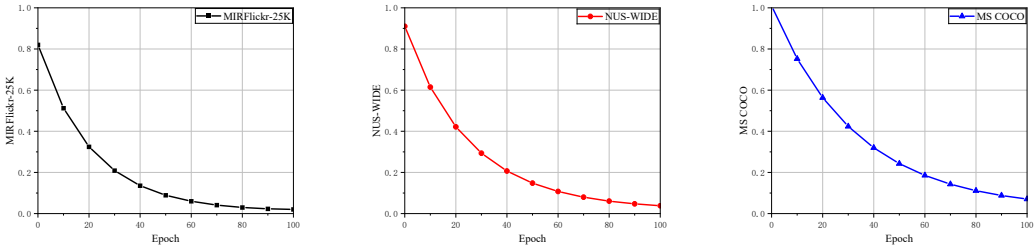
Figure 3. The effects of the parameters with 128-bit hash code length on MIRFlickr-25K.

Table 2. The mAP results across three dataset.

Task	Method	MIRFlickr-25K				NUS-WIDE				MS COCO			
		16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
I → T	**UCGKANH** w/o **H**	0.859	0.898	0.932	0.939	0.783	0.817	0.829	0.845	0.845	0.893	0.913	0.927
	UCGKANH w/o **K**	0.856	0.893	0.928	0.932	0.784	0.812	0.834	0.843	0.831	0.889	0.914	0.928
	UCGKANH w/o **CL**	0.869	0.896	0.917	0.928	0.782	0.811	0.829	0.837	0.809	0.872	0.916	0.923
	UCGKANH	**0.908**	**0.922**	**0.940**	**0.948**	**0.818**	**0.837**	**0.857**	**0.865**	**0.860**	**0.919**	**0.929**	**0.946**
T → I	**UCGKANH** w/o **H**	0.837	0.869	0.876	0.895	0.782	0.793	0.807	0.819	0.839	0.915	0.927	0.932
	UCGKANH w/o **K**	0.817	0.862	0.883	0.891	0.781	0.793	0.813	0.818	0.841	0.896	0.927	0.935
	UCGKANH w/o **CL**	0.842	0.865	0.883	0.891	0.771	0.796	0.805	0.811	0.836	0.871	0.917	0.929
	UCGKANH	**0.879**	**0.896**	**0.907**	**0.914**	**0.792**	**0.807**	**0.815**	**0.826**	**0.861**	**0.917**	**0.923**	**0.949**

4.5. Ablation Study

Through a selective exclusion of key modules including hypergraph enhancement, GraphKAN, and contrastive learning, we conduct an ablation research to examine the individual effects of each component in our system. Three benchmark datasets—MIRFlickr-25K, NUS-WIDE, and MS COCO—are utilized for the experiments, and the hash code



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

Figure 4. The convergence curves of UCGKANH in the case with 128-bit code length on the three datasets.

length is set to 128 bits, while results are demonstrated in Table 2. We define the following variants of our model for the ablation study:

- w/o **H**: We remove the hypergraph enhancement module, directly using the concatenated features $\mathbf{Z} = [\mathbf{X}_I; \mathbf{X}_T]$ as input to the GraphKAN layer. The contrastive learning step does not use hypergraph-guided positive pairs.
- w/o **K**: We replace the GraphKAN module with a standard graph convolutional network (GCN), where the feature update is simplified to $\mathbf{z}_v^{(l+1)} = \sigma\left(\sum_{u \in \mathcal{N}(v)} \mathbf{A}_{vu} \mathbf{z}_u^{(l)} \mathbf{W}^{(l)}\right)$, with σ being the ReLU activation.
- w/o **CL**: We remove the contrastive learning loss, optimizing the model solely with the quantization loss $\mathcal{L} = \lambda \|\mathbf{Z}^{(L)} - \mathbf{B}\|_F^2$.

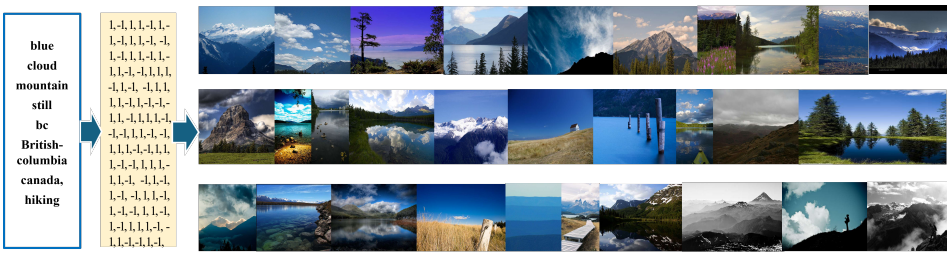
In summary, the ablation study demonstrates that all three components, including hypergraph, GraphKAN, and contrastive learning, are essential for achieving high retrieval performance in our UCGKANH framework. Among them, contrastive learning has the most significant impact, followed by hypergraph enhancement and GraphKAN, highlighting their complementary roles in capturing high-order relationships, enhancing non-linear feature representation, and ensuring cross-modal semantic alignment.

4.6. Convergence Analysis

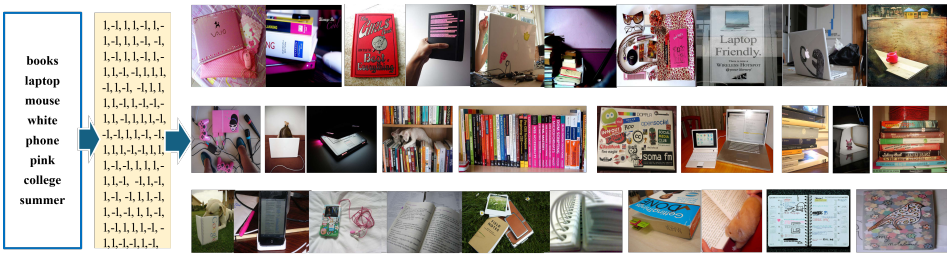
As seen in Fig. 4, we examine the convergence curves of the normalization loss $\mathcal{L}$ over 100 epochs for 128-bit hash codes on MIRFlickr-25K, NUS-WIDE, and MS COCO in order to look at the training dynamics of our model. On MIRFlickr-25K (Fig. 4(a)), the loss decreases rapidly from 0.8 to below 0.05 within 100 epochs, reflecting the dataset's simpler semantic structure. On NUS-WIDE (Fig. 4(b)), the loss starts at 0.9 and converges to around 0.05, showing a slightly slower decline due to its diverse categories. On MS COCO (Fig. 4(c)), the loss decreases more gradually from 1.0 to above 0.1, indicating higher complexity. These trends demonstrate that UCGKANH adapts effectively to varying dataset complexities, achieving stable convergence across all datasets.

4.7. Case Study

We conduct a case study on the T→I task using the MIRFlickr-25K dataset with 128-bit hash codes to further demonstrate the efficacy of UCGKANH in cross-modal retrieval. Two example queries are evaluated, and the top-30 images that were recovered are displayed in Fig. 5. The results demonstrate the model's ability to capture semantic relationships between text queries and images, reflecting its robustness in handling diverse concepts. For the first query, "blue cloud mountain still british columbia hiking" (Fig. 5(a)), UCGKANH successfully retrieves images that align with the described scene, predominantly featuring mountainous landscapes with blue skies and clouds, consistent with the British Columbia setting. The retrieved images capture the essence of hiking environments, showcasing natural scenery with clear skies, lakes, and peaks, indicating



(a)



(b)

Figure 5. The multi-media retrieval results on text to image task with hash code length of 128 bit.

that the model effectively bridges the modality gap by mapping the textual description to visually relevant content. This highlights the strength of the hypergraph enhancement and contrastive learning components in preserving high-order semantic relationships. For the second query, "books laptop mouse white pink phone college summer" (Fig. 5(b)), UCGKANH retrieves images that reflect a college or study environment, including books, laptops, and related items, with some images incorporating white and pink elements. The results capture the summer college context by including bright, casual settings with study materials, demonstrating the model's capability to handle complex, multi-concept queries. Overall, these case studies underscore the effectiveness of UCGKANH in achieving semantically consistent cross-modal retrieval, leveraging its integrated components to align diverse textual and visual representations.

5. Conclusion and Future Improvements

We propose UCGKANH, a novel unsupervised cross-modal hashing framework that integrates hypergraph enhancement, GraphKAN, and contrastive learning to successfully close the modality gap between images and texts. Our approach leverages hypergraph structures to capture high-order semantic relationships, employs GraphKAN for non-linear feature representation, and aligns cross-modal representations in the Hamming space using contrastive learning. Numerous tests on the MIRFlickr-25K, NUS-WIDE, and MS COCO benchmark datasets show that UCGKANH achieves superior retrieval performance, with significant improvements in mAP across various hash code lengths, as well as stable convergence behavior as shown in the normalization loss curves.

For future improvements, we plan to explore the integration of adaptive hyperparameter tuning to further enhance the model's performance on diverse datasets. Additionally, incorporating pre-trained multimodal models, such as CLIP, could improve the initial feature representations, potentially leading to better hash code quality. Last but not least, expanding UCGKANH to manage dynamic datasets with changing data distributions may increase its relevance in practical situations like online cross-modal

Author Contributions: Hongyu Lin: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – Original Draft. Shaofeng Shen: Methodology, Software, Validation,

Writing – Review & Editing. Yuchen Zhang: Data Curation, Visualization, Writing – Review & Editing. Renwei Xia: Data Collection, Experimental Support, Technical Assistance. All authors read and approved the final manuscript.

Funding: This work was supported in part by the Hunan Provincial Natural Science Foundation of China (2024JJ6533), in part by the High Performance Computing Center of Central South University, and in part by Professor Jizhong Zhao from Xi'an Jiaotong University

Data Availability Statement: The data will be made available upon reasonable request to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wang, T.; Li, F.; Zhu, L.; Li, J.; Zhang, Z.; Shen, H.T. Cross-Modal Retrieval: A Systematic Review of Methods and Future Directions. *Proceedings of the IEEE* **2024**, *112*, 1716–1754. https://doi.org/10.1109/JPROC.2024.3525147.
2. Bin, Y.; Li, H.; Xu, Y.; Xu, X.; Yang, Y.; Shen, H.T. Unifying Two-Stream Encoders with Transformers for Cross-Modal Retrieval, New York, NY, USA, 2023; MM '23. https://doi.org/10.1145/3581783.3612427.
3. Huang, H.; Nie, Z.; Wang, Z.; Shang, Z. Cross-modal and uni-modal soft-label alignment for image-text retrieval. AAAI Press, 2024, AAAI'24/IAAI'24/EAAI'24. https://doi.org/10.1609/aaai.v38i16.29789.
4. Liu, K.; Gong, Y.; Cao, Y.; Ren, Z.; Peng, D.; Sun, Y. Dual semantic fusion hashing for multi-label cross-modal retrieval. In Proceedings of the Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, 2024, IJCAI '24. https://doi.org/10.24963/ijcai.2024/505.
5. Hu, Z.; Cheung, Y.M.; Li, M.; Lan, W. Cross-Modal Hashing Method With Properties of Hamming Space: A New Perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2024**, *46*, 7636–7650. https://doi.org/10.1109/TPAMI.2024.3392763.
6. Sun, Y.; Dai, J.; Ren, Z.; Chen, Y.; Peng, D.; Hu, P. Dual Self-Paced Cross-Modal Hashing. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 15184–15192. https://doi.org/10.1609/aaai.v38i14.29441.
7. Li, F.; Wang, B.; Zhu, L.; Li, J.; Zhang, Z.; Chang, X. Cross-Domain Transfer Hashing for Efficient Cross-Modal Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology* **2024**, *34*, 9664–9677. https://doi.org/10.1109/TCSVT.2024.3374791.
8. Li, B.; Li, Z. Large-Scale Cross-Modal Hashing with Unified Learning and Multi-Object Regional Correlation Reasoning. *Neural Networks* **2024**, *171*, 276–292. https://doi.org/10.1016/j.neunet.2023.12.018.
9. Chen, H.; Zou, Z.; Liu, Y.; Zhu, X. Deep Class-Guided Hashing for Multi-Label Cross-Modal Retrieval. *Applied Sciences* **2025**, *15*. https://doi.org/10.3390/app15063068.
10. Wu, Y.; Li, B.; Li, Z. Revising similarity relationship hashing for unsupervised cross-modal retrieval. *Neurocomputing* **2025**, *614*, 128844. https://doi.org/10.1016/j.neucom.2024.128844.
11. Liu, H.; Xiong, J.; Zhang, N.; Liu, F.; Zou, X.; Köker, R. Quadruplet-Based Deep Cross-Modal Hashing. *Intell. Neuroscience* **2021**, *2021*. https://doi.org/10.1155/2021/9968716.
12. Shen, X.; Huang, Q.; Lan, L.; Zheng, Y. Contrastive transformer cross-modal hashing for video-text retrieval. In Proceedings of the Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, 2024, IJCAI '24. https://doi.org/10.24963/ijcai.2024/136.
13. Zhang, M.; Li, J.; Zheng, X. Semantic embedding based online cross-modal hashing method. *Scientific Reports* **2024**, *14*, 736. https://doi.org/10.1038/s41598-023-50242-w.
14. Wu, R.; Zhu, X.; Yi, Z.; Zou, Z.; Liu, Y.; Zhu, L. Multi-Grained Similarity Preserving and Updating for Unsupervised Cross-Modal Hashing. *Applied Sciences* **2024**, *14*. https://doi.org/10.3390/app14020870.
15. Su, H.; Han, M.; Liang, J.; Liang, J.; Yu, S. Deep supervised hashing with hard example pairs optimization for image retrieval. *The Visual Computer* **2023**, *39*, 5405–5420. https://doi.org/10.1007/s00371-022-02668-y.
16. Qin, Q.; Huo, Y.; Huang, L.; Dai, J.; Zhang, H.; Zhang, W. Deep Neighborhood-Preserving Hashing With Quadratic Spherical Mutual Information for Cross-Modal Retrieval. *IEEE Transactions on Multimedia* **2024**, *26*, 6361–6374. https://doi.org/10.1109/TMM.2023.3349075.
17. Kang, X.; Liu, X.; Zhang, X.; Xue, W.; Nie, X.; Yin, Y. Semi-Supervised Online Cross-Modal Hashing. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence. AAAI Press, 2025, Vol. 39, pp. 17770–17778. https://doi.org/10.1609/aaai.v39i17.33954.
18. Jiang, D.; Ye, M. Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2787–2797. https://doi.org/10.1109/CVPR52729.2023.00273.

19. Luo, H.; Zhang, Z.; Nie, L. Contrastive Incomplete Cross-Modal Hashing. *IEEE Transactions on Knowledge and Data Engineering* **2024**, *36*, 5823–5834. https://doi.org/10.1109/TKDE.2024.3410388.
20. Chen, B.; Wu, Z.; Liu, Y.; Zeng, B.; Lu, G.; Zhang, Z. Enhancing cross-modal retrieval via visual-textual prompt hashing. In Proceedings of the Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, 2024, IJCAI '24. https://doi.org/10.24963/ijcai.2024/69.
21. Zhu, J.; Ruan, X.; Cheng, Y.; Huang, Z.; Cui, Y.; Zeng, L. Deep Metric Multi-View Hashing for Multimedia Retrieval. In Proceedings of the 2023 IEEE International Conference on Multimedia and Expo (ICME), 2023, pp. 1955–1960. https://doi.org/10.1109/ICME55011.2023.00335.
22. Xie, X.; Li, Z.; Li, B.; Zhang, C.; Ma, H. Unsupervised cross-modal hashing retrieval via Dynamic Contrast and Optimization. *Engineering Applications of Artificial Intelligence* **2024**, *136*, 108969. https://doi.org/10.1016/j.engappai.2024.108969.
23. Wang, J.; Shi, H.; Luo, K.; Zhang, X.; Cheng, N.; Xiao, J. RREH: Reconstruction Relations Embedded Hashing for Semi-paired Cross-Modal Retrieval. In Proceedings of the Advanced Intelligent Computing Technology and Applications. Springer, Singapore, 2024, Vol. 14879, *Lecture Notes in Computer Science*. https://doi.org/10.1007/978-981-97-5675-9{_}32.
24. Jiang, X.; Hu, F. Multi-scale Adaptive Feature Fusion Hashing for Image Retrieval. *Arabian Journal for Science and Engineering* **2024**. https://doi.org/https://doi.org/10.1007/s13369-024-09627-w.
25. Li, Y.; Zhen, L.; Sun, Y.; Peng, D.; Peng, X.; Hu, P. Deep Evidential Hashing for Trustworthy Cross-Modal Retrieval. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence. AAAI Press, 2025, Vol. 39, pp. 18566–18574. https://doi.org/10.1609/aaai.v39i17.34043.
26. Li, Y.; Long, J.; Huang, Y.; Yang, Z. Adaptive Asymmetric Supervised Cross-Modal Hashing with consensus matrix. *Information Processing & Management* **2025**, *62*, 104037. https://doi.org/10.1016/j.ipm.2024.104037.
27. Yang, Y.; Wang, Y.; Wang, Y. SDA: Semantic Discrepancy Alignment for Text-conditioned Image Retrieval. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024; Ku, L.W.; Martins, A.; Srikumar, V., Eds., Bangkok, Thailand, 2024; pp. 5250–5261. https://doi.org/10.18653/v1/2024.findings-acl.311.
28. Liu, N.; Wu, G.; Huang, Y.; Chen, X.; Li, Q.; Wan, L. Unsupervised Contrastive Hashing With Autoencoder Semantic Similarity for Cross-Modal Retrieval in Remote Sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2025**, *18*, 6047–6059. https://doi.org/10.1109/JSTARS.2025.3538701.
29. Zhang, F.; Zhang, X. GraphKAN: Enhancing Feature Extraction with Graph Kolmogorov Arnold Networks, 2024, [arXiv:cs.LG/2406.13597].
30. Jiang, Q.Y.; Li, W.J. Deep Cross-Modal Hashing. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3270–3278. https://doi.org/10.1109/CVPR.2017.348.
31. Cao, Y.; Liu, B.; Long, M.; Wang, J. Cross-Modal Hamming Hashing. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), September 2018.
32. Tan, J.; Yang, Z.; Ye, J.; Chen, R.; Cheng, Y.; Qin, J.; Chen, Y. Cross-modal hash retrieval based on semantic multiple similarity learning and interactive projection matrix learning. *Information Sciences* **2023**, *648*, 119571. https://doi.org/10.1016/j.ins.2023.119571.
33. Ng, W.W.Y.; Xu, Y.; Tian, X.; et al.. Deep supervised fused similarity hashing for cross-modal retrieval. *Multimedia Tools and Applications* **2024**, *83*, 86537–86555. https://doi.org/10.1007/s11042-024-19581-2.
34. Li, A.; Li, Y.; Shao, Y. Federated learning for supervised cross-modal retrieval. *World Wide Web* **2024**, *27*, 41.
35. Chen, Y.; Long, Y.; Yang, Z.; et al. Unsupervised random walk manifold contrastive hashing for multimedia retrieval. *Complex Intell. Syst.* **2025**, *11*, 193. https://doi.org/10.1007/s40747-025-01814-y.
36. Cui, J.; He, Z.; Huang, Q.; Fu, Y.; Li, Y.; Wen, J. Structure-aware contrastive hashing for unsupervised cross-modal retrieval. *Neural Networks* **2024**, *174*, 106211. https://doi.org/10.1016/j.neunet.2024.106211.
37. Yao, D.; Li, Z.; Li, B.; Zhang, C.; Ma, H. Similarity Graph-correlation Reconstruction Network for unsupervised cross-modal hashing. *Expert Systems with Applications* **2024**, *237*, 121516. https://doi.org/10.1016/j.eswa.2023.121516.
38. Meng, H.; Zhang, H.; Liu, L.; Liu, D.; Lu, X.; Guo, X. Joint-Modal Graph Convolutional Hashing for unsupervised cross-modal retrieval. *Neurocomputing* **2024**, *595*, 127911. https://doi.org/10.1016/j.neucom.2024.127911.
39. Sun, L.; Dong, Y. Unsupervised graph reasoning distillation hashing for multimodal hamming space search with vision-language model. *Int J Multimed Info Retr* **2024**, *13*, 16. https://doi.org/10.1007/s13735-024-00326-8.
40. Chen, Y.; Long, Y.; Yang, Z.; Long, J. Unsupervised Adaptive Hypergraph Correlation Hashing for multimedia retrieval. *Information Processing & Management* **2025**, *62*, 103958. https://doi.org/10.1016/j.ipm.2024.103958.
41. Zhong, F.; Chu, C.; Zhu, Z.; Chen, Z. Hypergraph-Enhanced Hashing for Unsupervised Cross-Modal Retrieval via Robust Similarity Guidance, New York, NY, USA, 2023; MM '23, p. 3517–3527. https://doi.org/10.1145/3581783.3612116.
42. Liu, Z.; Wang, Y.; Vaidya, S.; Ruehle, F.; Halverson, J.; Soljačić, M.; Hou, T.Y.; Tegmark, M. KAN: Kolmogorov-Arnold Networks, 2025, [arXiv:cs.LG/2404.19756].

43. Hu, D.; Nie, F.; Li, X. Deep Binary Reconstruction for Cross-Modal Hashing. *IEEE Transactions on Multimedia* **2019**, *21*, 973–985. https://doi.org/10.1109/TMM.2018.2866771.
44. Zhu, L.; Wu, X.; Li, J.; Zhang, Z.; Guan, W.; Shen, H.T. Work Together: Correlation-Identity Reconstruction Hashing for Unsupervised Cross-Modal Retrieval. *IEEE Transactions on Knowledge and Data Engineering* **2023**, *35*, 8838–8851. https://doi.org/10.1109/TKDE.2022.3218656.
45. Ding, G.; Guo, Y.; Zhou, J. Collective Matrix Factorization Hashing for Multimodal Data **2014**. pp. 2083–2090. https://doi.org/10.1109/CVPR.2014.267.
46. Kumar, S.; Udupa, R. Learning hash functions for cross-view similarity search **2011**. p. 1360–1365.
47. Zhang, P.F.; Li, Y.; Huang, Z.; Xu, X.S. Aggregation-Based Graph Convolutional Hashing for Unsupervised Cross-Modal Retrieval. *IEEE Transactions on Multimedia* **2022**, *24*, 466–479. https://doi.org/10.1109/TMM.2021.3053766.
48. Liu, M.; Liu, Y.; Guo, M.; et al. CLIP-based Fusion-modal Reconstructing Hashing for Large-scale Unsupervised Cross-modal Retrieval. *International Journal of Multimedia Information Retrieval* **2023**, *12*, 139–149. https://doi.org/https://doi.org/10.1007/s13735-023-00268-7.
49. Wu, G.; Lin, Z.; Han, J.; Liu, L.; Ding, G.; Zhang, B.; Shen, J. Unsupervised deep hashing via binary latent factor models for large-scale cross-modal retrieval **2018**. p. 2854–2860.
50. Song, J.; Yang, Y.; Yang, Y.; Huang, Z.; Shen, H.T. Inter-media hashing for large-scale retrieval from heterogeneous data sources **2013**. p. 785–796. https://doi.org/10.1145/2463676.2465274.
51. Su, S.; Zhong, Z.; Zhang, C. Deep Joint-Semantics Reconstructing Hashing for Large-Scale Unsupervised Cross-Modal Retrieval **2019**. pp. 3027–3035. https://doi.org/10.1109/ICCV.2019.00312.
52. Li, B.; Yao, D.; Li, Z. RICH: A rapid method for image-text cross-modal hash retrieval. *Displays* **2023**, *79*, 102489. https://doi.org/10.1016/j.displa.2023.102489.
53. Zhou, J.; Ding, G.; Guo, Y. Latent semantic sparse hashing for cross-modal similarity search **2014**. p. 415–424. https://doi.org/10.1145/2600428.2609610.