

Universiteit Leiden

Master Computer Science

Multi-Level Hypergraph Enhanced Hashing for Unsupervised Cross-Modal Retrieval

Name:	Yesmina el Arkoubi
Student ID:	S2989018
Date:	10/07/2026
Specialisation:	Artificial Intelligence
1st supervisor:	Prof. Dr. Michael S. Lew
2nd supervisor:	Dr. Erwin M. Bakker

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Unsupervised cross-modal hashing has gained considerable attention in large-scale cross-modal retrieval. Although recent advances in cross-modal hashing have improved retrieval efficiency, learning semantically consistent representations across heterogeneous modalities remains challenging. Existing unsupervised cross-modal hashing methods often rely on global representations and pairwise similarity modelling, limiting their ability to capture fine-grained correspondences and high-order semantic relationships. In this study, a novel *Multi-Level Hypergraph-Enhanced* (MLHE) framework is proposed for unsupervised cross-modal hashing. The proposed framework integrates multi-level semantic representation learning with hypergraph-based learning to learn discriminative binary hash codes for efficient cross-modal retrieval. Specifically, global semantic representations are combined with fine-grained region-word correspondences to construct the similarity matrix, while a hypergraph is employed to model high-order semantic relationships among samples. Together, these components enable more effective preservation of cross-modal semantic structure within the learned Hamming space. The effectiveness of the proposed framework is evaluated on three widely used benchmark datasets, MIRFlickr-25K, NUS-WIDE, and MS COCO, on both Image-to-Text and Text-to-Image retrieval tasks. Experimental results demonstrate that MLHE outperforms the baseline hashing methods on the MIRFlickr-25K and NUS-WIDE datasets while achieving competitive performance on MS COCO.

Contents

1 Introduction **5**

2 Related Work **5**
2.1 Supervised Cross-modal Hashing . . . 5
2.2 Unsupervised Cross-modal Hashing . . . 6
2.3 Graph learning . . . 7
2.4 Fine-grained Semantic Alignment . . . 7

3 Background **8**
3.1 Cross-Modal Retrieval . . . 8
3.2 Cross-Modal Hashing . . . 9
3.3 Graph-Based Representation Learning . . . 9

4 Methodology **10**
4.1 Notation and Problem Definition . . . 11
4.2 Framework overview . . . 11
4.3 Multi-Level Representation Learning . . . 13
4.3.1 Image Feature Extraction . . . 13
4.3.2 Text Feature Extraction . . . 14
4.4 Fine-Grained Region-Word Alignment . . . 14
4.5 Cross-Modal Feature Fusion . . . 15
4.5.1 Local Fusion . . . 15
4.5.2 Cross-Weighted Global Fusion . . . 15
4.6 Similarity Matrix Construction . . . 16
4.7 Hypergraph Enhancement . . . 17
4.8 Neighbour-Enhanced Semantic Aggregation . . . 18
4.9 Shared Representation Learning . . . 19
4.10 Loss Function . . . 20
4.10.1 Similarity Preservation Loss . . . 20
4.10.2 Quantization Loss . . . 21
4.10.3 Reconstruction Loss . . . 21
4.10.4 Contrastive Loss . . . 21
4.10.5 Overall Objective . . . 22
4.11 Hash Function Learning . . . 22

5 Experiments **23**
5.1 Datasets and Preprocessing . . . 23
5.1.1 MIRFlickr-25K . . . 23
5.1.2 NUS-WIDE . . . 24
5.1.3 MS COCO . . . 24
5.2 Baseline Methods and Evaluation Metrics . . . 25
5.2.1 Mean Average Precision . . . 25
5.2.2 Efficiency Metrics . . . 26
5.2.3 Curves . . . 26
5.3 Ablation Experiment . . . 26
5.4 Backbone Comparison . . . 26

5.5 Implementation Details 27

6 Results **28**

7 Discussion **32**

8 Conclusion **33**

References **37**

1 Introduction

Cross-modal retrieval has emerged as a significant research area within the field of information retrieval, aiming at identifying relationships between data from different modalities. Cross-modal retrieval methods can be categorized into two types based on their data encoding form: real-value retrieval and hashing retrieval. The real-value based methods provide more semantic information by mapping each modality to a continuous vector space, but elevate storage costs and computational demands. In contrast, hashing-based retrieval allows for more efficient retrieval as it encodes data into compressed binary representations in a Hamming space. Furthermore, the two methods can be subdivided into two categories based on the availability of supervisory information: supervised and unsupervised methods.

In the field of unsupervised cross-modal hashing, three main components affect retrieval performance: multi-modal semantic modeling, cross-modal interaction, and hash function learning [1]. A key challenge in this field lies in the inherent high dimensionality and heterogeneity of multi-modal data. Hence, it is essential for multi-modal semantic modeling to effectively capture and model heterogeneous multi-modal semantics with deep network models. Moreover, it is essential to strengthen the interaction between the modalities and reduce the heterogeneous modality gap. Improving the semantic modeling and alignment by mapping data from different modalities into a shared space where similar items are close to each other leads to improved hash function learning, which directly determines the quality of the hash codes used for retrieval. While unsupervised cross-modal hashing methods[2, 3, 4, 5] have made promising progress, they still suffer from limited capability in modeling and preserving the multi-modal semantics[1]. Most methods construct an affinity matrix by mining semantic correlations among data to guide the learning of hashing functions [6, 7]. However, they overlook the multidimensionality of information, resulting in a similarity matrix that fails to capture rich semantic associations[8]. Specifically, most methods construct similarity relationships using only global feature representations and pairwise similarity modeling. Therefore, they often fail to exploit complementary semantic information available at different levels[8]. Moreover, these methods explore ordinary graphs to construct the similarity guidance, which can only extract pairwise first-order semantic information and fails to effectively obtain information on the correlation between samples [9]. Consequently, the generated hash codes may not fully preserve the underlying semantic structure across modalities.

To address these limitations, this study proposes a novel *Multi-Level Hypergraph-Enhanced* (MLHE) framework for unsupervised cross-modal hashing. The proposed framework integrates multi-level semantic representation learning with hypergraph-based semantic modeling to improve cross-modal similarity learning. Specifically, global semantic representations are combined with fine-grained region-word correspondences to construct a richer semantic similarity matrix, while a hypergraph is employed to capture high-order semantic relationships among samples. The resulting semantic representations are then used to learn discriminative binary hash codes for efficient cross-modal retrieval.

2 Related Work

2.1 Supervised Cross-modal Hashing

Supervised cross-modal hashing methods[10, 11, 12, 13, 14] aim to learn hash functions using label supervision. Early cross-modal methods primarily rely on statistical correlation analysis

and matrix factorization techniques to semantically align multi-modal data in a shared latent space. Canonical Correlation Analysis(CCA) [15] maximizes the correlation between different modals while graph-based methods were introduced to preserve the local geometric structure of data after hashing. Matrix Tri-Factorization Hashing(MFTH)[16] proposed an objective function to jointly learn hashing codes tailored to specific modalities, which allows various length configurations. This method learns two semantic correlation matrices to achieve meaningful associations between two types of unequal hash codes.
The success of deep neural networks inspired the development of numerous supervised cross-modal hashing methods that rely on a deep learning framework. Deep Cross-Modal Hashing(DCMH)[17] proposed an end-to-end deep learning framework to address the limitations of early cross-modal hashing methods that are based on hand-crafted feature representations. Deep Normalized Cross-Modal Hashing with Bi-Directional Relation Reasoning (Bi_NCMH)[12] proposes a bi-directional reasoning mechanism that constructs a multi-level semantic similarity matrix to model correlations among instances. Graph Convolutional Hashing(GCH)[11] uses Graph Convolutional Network(GCN) to embed the inherent similarities between image-text pairs into the hash codes. GCN is proposed to update the feature of each node by aggregating information from its neighbouring nodes based on the adjacency matrix.

2.2 Unsupervised Cross-modal Hashing

In contrast to supervised methods[17, 11], unsupervised cross-modal hashing methods[18, 2, 1, 5] aim to learn hash functions for multi-modal data without relying on labels. Unlike supervised methods, these methods mainly exploit the inherent data properties among multi-modal data, such as the intrinsic structure, distribution, correlation, and geometric relationships. Similar to the supervised methods, the early unsupervised methods primarily rely on statistical correlation analysis and matrix factorization techniques. Recent unsupervised methods[19, 2] combine deep neural networks with reconstruction, adversarial learning, or contrastive learning objectives. Deep Joint-Semantics Reconstructing Hashing(DJSRH)[2] proposes a reconstruction-based unsupervised cross-modal hashing framework. This method constructs a joint semantic affinity matrix by integrating the original neighborhood information from different modalities. Then it learns hash codes by reconstructing this joint-semantics structure, which helps preserve the semantic similarity across modalities. Correlation-Identity Reconstruction Hashing (CIRH)[1] conducts cross-modal hash code and function learning within a semantic reconstruction network by using a collaborated graph to model the intrinsic multi-modal correlations and perform aggregation across modalities to strengthen cross-modal interactions. Self-supervised Adversarial Hashing(SSAH)[20] is one of the early approaches to incorporate adversarial hashing into self-supervised cross-modal hashing. This method uses two adversarial networks on bi-modal data to preserve semantic relevance and representation consistency across modalities. Thereby, it allows the model to learn modality-invariant embeddings. Motivated by the success of contrastive learning, studies have proposed contrastive cross-modal hashing methods, including Unsupervised Contrastive Cross-Modal Hashing(UCCH)[3] and Structure-aware Contrastive Hashing(SACH)[4]. Contrastive learning aims to learn discriminative representations by minimizing the distance between positive samples(similar image-text pairs) in the shared embedding space while maximizing the distance of negative samples(dissimilar image-text pairs). Graph Contrastive-and-Reconstructive Hashing(GCRH)[5] integrates cross-modal contrastive learning and cross-modal reconstruction to jointly reduce the modality gap in unsupervised hashing. This method introduces both global and local graphs into contrastive and reconstructive learning to

exploit the graph structural information, enabling the generation of more discriminative hash codes. Unsupervised Adversarial Contrastive Hashing(UACH)[21] is among the first hashing methods to integrate adversarial learning and contrastive learning within a unified framework. This method introduces a cycle generative adversarial network to learn the transformation relationships between different modality feature domains, thereby promoting semantic alignment across modalities. Furthermore, it employs dual contrastive learning to optimize representation and hashing learning components for each modality, enabling the learning of unified hash codes, which mitigates the impact of modality discrepancies.

2.3 Graph learning

Graph Neural Networks(GNNs) have attracted research interests due to their ability to effectively learn representations from graph-structured data. The first proposed GNN framework[22] leverages a recursive neighborhood aggregation strategy to calculate the feature for each node. Through iterative feature aggregation, it can capture structural information from the k-hop neighborhood, thereby learning more informative node representations and improving model performance.

The ability of GNNs to model relationships between connected samples has made them attractive for cross-modal retrieval. Unlike standard methods that learn representations independently for each sample, graph-based methods can encode semantic relationships among image-text pairs and propagate information between neighbouring samples. This allows the learned representations to better preserve the underlying data structure, improving cross-modal alignment and ultimately retrieval performance. Hence, many cross-modal retrieval methods[23, 9, 24, 25, 26] that incorporate graphs have been proposed. Aggregation-based Graph Convolutional Hashing (AGCH) [23] is among the first unsupervised cross-modal hashing methods to incorporate Graph Convolutional Networks (GCNs) for hash learning. The method preserves both intra-modal and inter-modal consistency while modeling cross-modal interactions and neighborhood structures, thereby learning more discriminative and representative binary embeddings. Hypergraph-Enhanced Hashing(HEH)[9] is is among the first unsupervised cross-modal hashing methods to incorporate hypergraph convolutional networks for hash learning. The Hypergraph-based Enhancement(HE) component uses hypergraph convolution and attention mechanism to explore the underlying neighbourhood structure, thereby enabling full exploration and utilization of high-order semantic information in input samples.

2.4 Fine-grained Semantic Alignment

Traditional image-text retrieval methods map entire images and sentences to the shared semantic space for image-text matching. Although these global representations[27, 28, 29] effectively capture high-level semantics, they often overlook fine-grained correspondences between image regions and textual words, which limits retrieval performance. To address this limitation, fine-grained semantic alignment methods[30, 31, 32, 33, 34] are proposed to infer latent correspondences between image regions and words, enabling improved cross-modal matching. One of the early fine-grained alignment methods, Stacked Cross Attention Network(SCAN)[30], aligns image regions with sentence words through local semantic associations and aggregates the similarities of relevant region–word pairs to measure the overall image–text relevance. The method employs a cross-attention mechanism that attends to fragments in one modality based on query fragments from the other modality. As a result, semantically aligned fine-

grained pairs contribute more strongly to the final image–text similarity score, while the influence of inconsistent pairs is reduced. Cross-modal Hard Aligning Network (CHAN)[31] addresses the limitations of soft cross-attention methods such as SCAN[30], which often generate redundant region-word alignments that can reduce retrieval accuracy while increasing memory and computational overhead. CHAN introduces a hard assignment coding scheme which obtains shared image-text semantics by mining informative region-word pairs while discarding redundant alignments. In contrast to soft cross-attention approaches, the method retains only the most strongly matching image region for each word. Cross-Dimensional Coarse-Fine-Grained complementary network(CDGCN)[32] combines fine-grained alignment with coarse-grained semantic aggregation for image–text retrieval. The method performs fine-grained semantic alignment between image regions and sentence words by modeling cross-dimensional dependencies. In addition, a coarse-grained cross-dimensional semantic aggregation branch captures global image–text semantic matching, which complements the local semantic alignment to ensure consistency between global and local semantic information. Region-feature enhancement-based image-text similarity inference network (RFE- SRN)[33] proposes a similarity-based graph inference network that propagates the relationship between global and local similarity information. Moreover, RFE-SRN employs a vector-based similarity method, instead of a scalar-based method, to improve interaction between feature vectors. This method can better describe the global alignment between images and sentences, and the local alignment between regions and words. Multi-level Semantic Fusion-Driven Hashing(MFSH)[8] employs multi-level feature decomposition and fusion to mine semantic associations from both global and local image-text representations. Specifically, it obtains the fine-grained salient representations by selecting the highest-scoring image patches and textual tokens based on their alignment scores, thereby suppressing irrelevant features. Furthermore, MFSH employs quality-filtered contrastive learning, which excludes low-quality, hard-to-learn samples from the contrastive learning process in the hash layer. Unsupervised Cross-Modal Semantic Invariance Hashing (USIH)[35] addresses the limitations of cross-modal similarity matrices, which are typically constructed from shallow statistical co-occurrence patterns and are therefore prone to noise and spurious correlations. USIH proposes an object-aware attention mechanism that focuses on the key semantic objects in images while suppressing background interference. Subsequently, the resulting object clusters are leveraged to guide textual semantic modeling, thereby obtaining more reliable cross-modal similarity matrices. Additionally, USIH introduces a Semantic Invariance Enhancement(SIE) module that applies controlled perturbations to multimodal features in the frequency domain and enforces consistency constraints, encouraging the model to focus on deep semantic representations shared across modalities.

3 Background

3.1 Cross-Modal Retrieval

Cross-modal retrieval aims to retrieve data from one modality using a query from another modality, such as retrieving images from a textual query. The rapid growth of multimodal data on the web has made cross-modal retrieval an important research area within information retrieval. A fundamental challenge in cross-modal retrieval is the heterogeneity gap, which arises from the distinct characteristics and feature distributions of different modalities, making direct comparison across different modalities infeasible. To bridge this gap, cross-modal retrieval methods learn a shared latent representation space where semantically related instances from

different modalities are aligned.
Early approaches in the field of cross-modal retrieval used linear projection techniques to construct a shared latent space[15]. In the past years, with the rapid development of deep learning, many deep learning-based methods have been proposed. Such methods often use a text encoder to transform the text into a fixed-length embedding and an image encoder to transform the image into a visual embedding. The embeddings from both modalities are then projected into the shared space and typically normalized to facilitate similarity computation. Furthermore, different optimization objectives are incorporated, such as contrastive loss, which is used to optimize the embedding space by pulling paired samples closer while pushing negative samples further apart.
Cross-modal retrieval methods can be categorized into two types based on their data encoding form: real-value retrieval and hashing retrieval. Real-valued retrieval methods aim to learn low-dimensional continuous vector representations for multimodal data. While these representations preserve richer semantic information, it incurs elevated memory costs and computational resources. In contrast, cross-modal hashing aims to learn compact binary hash codes. This representation significantly improves retrieval efficiency in terms of both storage and computation, but may lead to loss of (fine-grained) semantic information.
Alongside the data coding form, cross-modal retrieval methods can be categorized into supervised and unsupervised methods, which are distinguished by the utilization of label information. The supervised methods can exploit the label information to guide the learning process, which may enable the model to better preserve semantic relationships across modalities. However, collecting labeled data requires massive labor, resulting in high labor costs in real world application. In contrast, unsupervised methods capture the underlying correlation patterns between different modalities without relying on manual annotations. Although these methods are more flexible and applicable for real-world applications, the absence of explicit semantic supervision makes it more challenging to learn highly discriminative cross-modal representations.

3.2 Cross-Modal Hashing

Hashing-based methods[36] have attracted considerable attention in recent years on large-scale retrieval tasks, offering improvements in storage and computation efficiency compared to real-valued retrieval methods [37]. The objective of hash methods is to learn hash functions to map heterogeneous data into a common Hamming space while preserving their semantic relationships. In these methods, each sample is represented by a binary hash code consisting of a fixed number of bits. Similarity between hash codes is measured using the Hamming distance, which counts the number of differing bits between two codes. Semantically similar samples should have small Hamming distances, whereas dissimilar samples should be assigned different codes. Compared to real-valued representations, binary hash codes require significantly less storage and enable highly efficient retrieval through these bitwise operations [38].
The key challenge in hashing-based methods is the inevitable information loss caused by compressing high-dimensional real-valued features into compact binary codes. The information loss may lead to a decrease in feature discriminability, thus affecting retrieval accuracy[39].

3.3 Graph-Based Representation Learning

Graph-based methods[40, 41, 11, 23, 5, 42, 43] have achieved remarkable success in cross-modal retrieval as they can effectively capture the inter-modal and intra-modal relationships and

leverage the underlying data structure to improve the representation learning. The graphs are constructed based on the similarity relationships among the image and text samples, where the nodes represent samples and the edges represent similarities. The similarity relationships are typically computed using similarity measures such as cosine similarity or k-nearest-neighbour strategies. Graph Convolutional Networks(GCNs) are often used to update each node's representation by aggregating information from its neighbouring nodes. The learned representations incorporate both individual feature information and the neighbourhood structure.
Graph-based methods effectively capture semantic relationships among samples by exploiting the underlying neighbourhood structure. This helps reduce the heterogeneity gap and improves the discriminative ability of the learned hash codes and retrieval performance. Nevertheless, most existing graph-based methods[1, 2, 44] focus on pairwise relationships among samples, overlooking the higher-order relationships in real-world scenarios, and consequently failing to fully exploit the underlying structure of the data. Therefore, hypergraph-based methods[24, 9] are proposed to model the higher-order relationships among samples that share similar semantics, enabling semantic alignment and accurate shared semantic representations. In hypergraph-based methods, the nodes still represent samples, but the hyperedges can connect to multiple nodes rather than just two. Hyperedges are constructed by connecting each sample with its top-K nearest neighbors according to a similarity measure, thereby grouping semantically similar samples into a shared hyperedge.

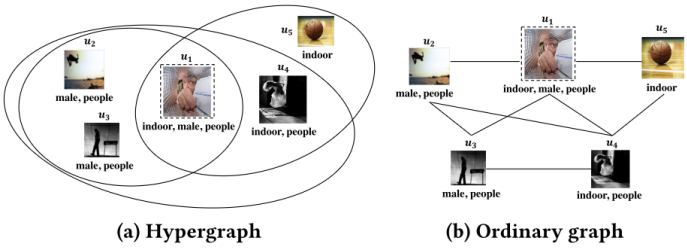


Figure 1: Comparison of a hypergraph and a ordinary graph[9]

4 Methodology

In this paper, a novel unsupervised cross-modal hashing framework, named Multi-Level Hypergraph Enhanced Hashing (MLHE), is proposed for cross-modal retrieval. Existing unsupervised cross-modal hashing methods typically construct semantic relationships using global image and text representations. These methods overlook the fine-grained correspondences between image regions and textual words, limiting their ability to capture complementary semantic information across modalities. Furthermore, most methods use graphs that only model pairwise relationships, which limits their ability to capture high-order relationships between samples. To address these limitations, the proposed MLHE framework learns discriminative binary hash codes by integrating global image-text semantics with region-word alignment and hypergraph-based semantic modeling. Specifically, MLHE uses both global semantic information and fine-grained region-word correspondences to construct semantically rich representations. These multi-level representations are used to construct a similarity matrix, which is further enhanced using hypergraphs to capture higher-order semantic relationships among the neighbouring samples.

MLHE builds upon the two-stage learning framework of CIRH [1], adopting its shared representation learning strategy, correlation-identity reconstruction network, and hash function learning scheme. The MLHE framework introduces a multi-level cross-modal module that models global semantic features and fine-grained region-word correspondences to enhance semantic representation learning. The multi-level cross-modal module is inspired by CHAN[31], which models fine-grained cross-modal semantic correspondences by exploiting the most relevant region-word pairs and eliminating all other alignments. Furthermore, MLHE incorporates a hypergraph construction method inspired by HEH [9] to capture higher-order semantic relationships among the samples and contrastive learning to further improve cross-modal semantic alignment. These extensions provide richer semantic supervision and lead to more discriminative multimodal hash representations for unsupervised cross-modal retrieval.

4.1 Notation and Problem Definition

Given a dataset

$$\mathcal{D} = \{(I_i, T_i)\}_{i=1}^{n}, \tag{1}$$

where I_i denotes the i-th image, T_i denotes its corresponding text, and n is the number of image-text pairs.
The objective of the proposed method is to learn two modality-specific hash functions

$$h_I : \mathbb{R}^D \to \{-1, +1\}^r, \tag{2}$$

$$h_T : \mathbb{R}^D \to \{-1, +1\}^r, \tag{3}$$

which map image and text features into a shared r-bit Hamming space while preserving the semantic correspondence between paired samples.
To achieve this, the framework learns the unified compact binary codes

$$B \in \{-1, +1\}^{n \times r}, \tag{4}$$

which serve as shared semantic representations for the image-text pairs. These binary codes are subsequently used as pseudo-labels for learning of the modality-specific hash functions.

4.2 Framework overview

As illustrated in Figure 2, MLHE consists of two learning stages: hash code learning and hash function learning. In the first stage, a multi-level encoder obtains global and fine-grained semantic representations from the image and text modalities. The image encoder obtains global image features and region-level features, while the text encoder obtains global textual features and word-level features. In the local branch, the region-word alignment module assigns each textual word to its most relevant image region. Then, the semantic weighting module emphasizes the most semantically relevant region-word correspondences. These resulting local representations are subsequently passed through a Transformer-based fusion module, allowing for cross-modal interaction to produce enhanced local image and text representations. The final local representations are used to compute modality-specific local similarity matrices, which

are combined to form the local similarity matrix. Simultaneously, in the global branch, the semantic weighting module refines the image features according to the text representation, while the text features are refined according to the image features based on their semantic relevance. The resulting global representations are further refined through a Transformer-based fusion module and then fused with the original global image and text features through weighted feature fusion. The resulting final global image and text representations are used to compute modality-specific global similarity matrices, which are then combined to obtain the global semantic similarity matrix. Finally, the local and global similarity matrices are combined into one similarity matrix. The similarity matrix and the fused global representations are used to construct the hypergraph that captures higher-order semantic relationships among neighboring samples. More specifically, the similarity matrix is used to select the top k similar neighbours for each sample and the fused global representations are used to measure the similarity of each neighbouring sample to the corresponding sample node, allowing more representative neighbours to contribute more strongly within the hyperedge. The hypergraph is used to perform shared representation learning, through convolution, to obtain continuous latent representations that preserve the semantic neighborhood structures across both modalities. Subsequently, the correlation identity reconstruction module reconstructs the modality-specific latent image and text features from the shared latent representation. This encourages the shared representation to preserve complementary semantic information from both modalities while retaining modality-specific semantics. Furthermore, an image-text contrastive objective enhances cross-modal alignment by pulling semantically similar image-text pairs closer together while pushing unmatched pairs farther apart in the shared embedding space. Finally, the discrete hash objective transforms the shared representations into discriminative binary hash codes.
In the second stage, the learned binary hash codes supervise the training of modality-specific image and text hash functions. The trained hash functions can then generate binary hash codes for previously unseen image and text samples, projecting them into a common Hamming space for efficient cross-modal retrieval.

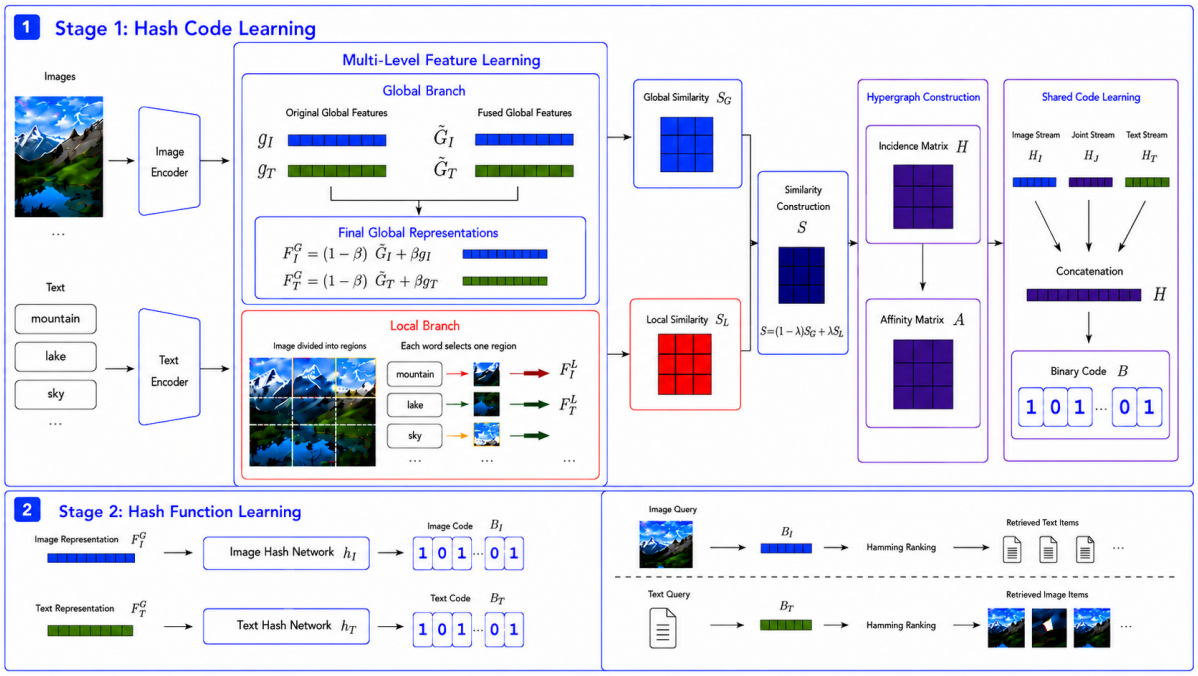


Figure 2: Overview of the proposed MLHE framework. Stage 1 performs hash code learning by extracting global and local semantic information and constructing similarity matrix, enhanced by hypergraphs, to learn binary pseudo-labels. Stage 2 trains modality-specific hash functions to generate binary codes for unseen image and text samples, enabling efficient and fast cross-modal retrieval

4.3 Multi-Level Representation Learning

The primary goal of cross-modal hashing is to learn hash functions that preserve the semantic relationships across different modalities within compact binary codes To achieve this, it is essential to construct a comprehensive similarity matrix that captures the semantic relationships between the data points. The proposed MLHE framework constructs the similarity matrix by learning global and local representations from images and texts. The global representations capture the overall semantics of each modality, while the local representations preserve fine-grained correspondences between image regions and text words that provide complementary semantic information. By integrating both representation levels, the resulting similarity matrix captures complementary semantic information.

4.3.1 Image Feature Extraction

Given an input image I, a pretrained ResNet-50 network is used as the visual backbone to extract convolutional feature maps,

$$E_I \in \mathbb{R}^{2048\times7\times7},$$

where E_I denotes the output of the last convolutional layer.
A local image representation is obtained by projecting the feature map into a 512-dimensional shared embedding space using a 1×1 convolution:

$$L_I \in \mathbb{R}^{512\times7\times7},$$

The projected feature map is then flattened into a sequence of 49 image regions:

$$\mathcal{L}_I = \left\{\ell_1^I, \ell_2^I, \ldots, \ell_{49}^I\right\}, \qquad \ell_i^I \in \mathbb{R}^{512}.$$

Simultaneously, a global image representation is obtained by aggregating the feature maps into compact semantic representations through adaptive average pooling, followed by a linear projection into the shared 512-dimensional embedding space,

$$g_I = \mathrm{Pool}(E_I), \qquad g_I \in \mathbb{R}^{512}.$$

4.3.2 Text Feature Extraction

Given an input text that consists of L tokens, each token is embedded into the shared 512-dimensional embedding space using a learnable embedding layer:

$$\mathcal{L}_T = \left\{\ell_1^T, \ell_2^T, \ldots, \ell_L^T\right\}, \qquad \ell_j^T \in \mathbb{R}^{512}.$$

The global text representations are obtained through masked average pooling over all tokens, followed by linear projection into the shared embedding space:

$$g_T = \frac{1}{L_v}\sum_{j=1}^{L_v} \ell_j^T, \qquad g_T \in \mathbb{R}^{512},$$

where L_v denotes the number of non-padding tokens.

4.4 Fine-Grained Region-Word Alignment

Although global representations effectively capture the overall semantic content of each modality, they often overlook fine-grained correspondences between image regions and text words that provide complementary semantic information for cross-modal matching. Therefore, MLHE employs a fine-grained region-word alignment module, inspired by the hard alignment strategy proposed in CHAN [31], to exploit complementary local semantics. Unlike soft attention mechanisms that assign weights to all image regions, the hard alignment strategy aligns each text word with only the most relevant region, which reduces redundant correspondences and emphasizes the most informative local semantic relationships.
Given the local image tokens $\mathcal{L}_I$ and local text tokens $\mathcal{L}_T$, the region-word alignment matrix is computed as

$$M = \mathcal{L}_I \mathcal{L}_T^\top,$$

where

$$M \in \mathbb{R}^{49\times L}.$$

For each text token ℓ_j^T, only the image region with the highest alignment score is selected,

$$i^*(j) = \arg\max_i M_{ij}.$$

The corresponding alignment score is

$$s_j = M_{i^*(j),j}.$$

To emphasize the most informative region-word correspondences, a softmax weighting strategy is applied to the selected alignment scores.

$$w_j = \frac{\exp(\gamma s_j)}{\sum_{k=1}^{L_v} \exp(\gamma s_k)},$$

where γ is the temperature parameter that controls the sharpness of the weighting distribution. The final aligned local image representation is obtained by aggregating the selected image regions,

$$\hat{L}_I = \sum_{j=1}^{L_v} w_j \, \ell^I_{i^*(j)},$$

while the final aligned local text representation is computed as

$$\hat{L}_T = \sum_{j=1}^{L_v} w_j \, \ell^T_j.$$

4.5 Cross-Modal Feature Fusion

Transformer-based feature fusion modules are employed to enable semantic interactions between the different modalities, at both local and global representation levels. The local fusion module enhances the fine-grained region-word correspondences, while the global fusion module enhances the global semantic representations through cross-modal interactions.

4.5.1 Local Fusion

Following region-word alignment, the aligned local image and text representations are fused using a transformer encoder:

$$\left[F^L_I, F^L_T\right] = \text{Transformer}\left(\hat{L}_I, \hat{L}_T\right).$$

4.5.2 Cross-Weighted Global Fusion

To incorporate complementary semantic information across modalities, each modality is weighted according to its relevance to the other modality using cross-modal attention.
The image attention weights are computed as

$$w^I_i = \text{softmax}\left((\ell^I_i)^\top \bar{\ell}^T\right),$$

where

$$\bar{\ell}^T = \frac{1}{L_v} \sum_{j=1}^{L_v} \ell^T_j$$

denotes the mean textual token representation.
Similarly, the textual attention weights are computed as

$$w_j^T = \mathrm{softmax}\left((\ell_j^T)^\top \bar{\ell}^I\right),$$

where

$$\bar{\ell}^I = \frac{1}{49}\sum_{i=1}^{49} \ell_i^I$$

denotes the mean visual region representation.
The cross-weighted global representations are then computed as

$$\hat{G}_I = \sum_{i=1}^{49} w_i^I \, \ell_i^I,$$

and

$$\hat{G}_T = \sum_{j=1}^{L_v} w_j^T \, \ell_j^T.$$

The cross-weighted global representations are subsequently fused using a Transformer encoder,

$$\left[\tilde{G}_I, \tilde{G}_T\right] = \mathrm{Transformer}\left(\hat{G}_I, \hat{G}_T\right).$$

Finally, the Transformer-enhanced global representations are combined through residual fusion with the original global representations,

$$F_I^G = (1-\beta)\tilde{G}_I + \beta g_I,$$

$$F_T^G = (1-\beta)\tilde{G}_T + \beta g_T,$$

where β controls the trade-off between preserving the original global semantics and integrating the Transformer-enhanced cross-modal semantics.
The resulting representations F_I^G and F_T^G denote the final fused global semantic representations, which are subsequently used to construct the global similarity matrix

4.6 Similarity Matrix Construction

The similarity matrix is constructed by integrating both local and global semantic information. First, pairwise intra-modal cosine similarities are calculated independently from the global image and text fused representations and then combined to obtain the global similarity matrix. The same procedure is applied to the local image and text representations. Finally, the global and local similarity matrices are combined to produce the final semantic similarity matrix
More specifically, the global image similarity matrix is calculated as:

$$(S_I)_{ij} = \cos\left(F_{I,i}^G, F_{I,j}^G\right),$$

while the global text similarity matrix is computed as:

$$(S_T)_{ij} = \cos\left(F_{T,i}^G, F_{T,j}^G\right).$$

These two matrices are combined to obtain the global semantic similarity matrix:

$$S_G = \alpha S_I + (1 - \alpha) S_T,$$

where α balances the importance of the image and textual modalities.
Similarly, the local image and text similarity matrices are computed as:

$$(S_{LI})_{ij} = \cos\left(F_{I,i}^L, F_{I,j}^L\right),$$

and

$$(S_{LT})_{ij} = \cos\left(F_{T,i}^L, F_{T,j}^L\right).$$

The local similarity matrix is defined as:

$$S_L = \alpha S_{LI} + (1 - \alpha) S_{LT}.$$

The global and local similarity matrices are subsequently combined to obtain the initial similarity matrix:

$$S_1 = (1 - \lambda) S_G + \lambda S_L.$$

Subsequently, the initial similarity matrix S_1 is enhanced by nonlinearly strengthening the adjacent correlation between the more similar pairs:

$$S_2 = 2\ \mathrm{Sigmoid}(S_1) - 1 + I,$$

where I denotes the identity matrix.
Finally, the initial and enhanced similarity matrices are combined to obtain the final semantic similarity matrix:

$$S = \alpha_2 S_1 + (1 - \alpha_2) S_2.$$

where α_2 balances the importance of the initial and enhanced similarity matrices.

4.7 Hypergraph Enhancement

Most graph-based methods model only pairwise relationships between samples, which limits their ability to capture higher-order semantic relationships. To address this limitation, MLHE employs a hypergraph enhancement module to capture higher-order semantic information among neighbouring samples.
The hypergraph module takes the similarity matrix

$$S \in \mathbb{R}^{N \times N},$$

and the final fused global features F_I^G and F_T^G as input. The similarity matrix is used to determine the hypergraph connectivity, while the fused global features are used to estimate the semantic importance of the connected samples within each hyperedge.
Given the similarity matrix S, each sample serves as the center of a hyperedge. For every sample, the top-k nearest neighbours are selected according to S and connected to the corresponding hyperedge.

A unified cross-modal representation is first obtained to capture the shared semantic information across the modalities:

$$X_{hg} = \frac{1}{2}\left(F_I^G + F_T^G\right),$$

For each hyperedge, the node features of all samples belonging to that hyperedge are averaged to obtain the hyperedge representation,

$$h_j = \frac{1}{|e_j|}\sum_{i \in e_j} X_{hg}^{(i)},$$

where h_j denotes the representation of the j-the hyperedge.
Unlike most methods, the proposed weighted incidence matrix assigns different weights to neighbouring samples, according to their semantic similarity within the corresponding hyperedge. Specifically, it applies softmax normalization to cosine similarity between the feature of sample i and the representation of the hyperedge centered at sample j. This weighting allows semantically relevant samples to contribute more strongly to the corresponding hyperedge than less related samples.
The weighted incidence matrix is defined as:

$$\mathcal{H} \in \mathbb{R}^{N \times N},$$

where each element is computed as:

$$\mathcal{H}_{ij} = \frac{\exp\Big(\cos\Big(X_{hg}^{(i)},\, h_j\Big)\Big)}{\sum\limits_{k \in e_j} \exp\Big(\cos\Big(X_{hg}^{(k)},\, h_j\Big)\Big)}.$$

The resulting weighted incidence matrix is subsequently normalized using the hyperedge-degree and vertex-degree matrices to construct the adjacency matrix, thereby ensuring balanced information propagation among samples. The hyperedge-degree and vertex-degree matrices are defined as:

$$D_e(j,j) = \sum_{i=1}^{N} \mathcal{H}_{ij},$$

and

$$D_v(i,i) = \sum_{j=1}^{N} \mathcal{H}_{ij}.$$

Finally, the normalized hypergraph adjacency matrix is obtained as:

$$A = D_v^{-1/2}\mathcal{H}D_e^{-1}\mathcal{H}^\top D_v^{-1/2}.$$

4.8 Neighbour-Enhanced Semantic Aggregation

The neighbour aggregation module enriches each sample with the semantic information of its most similar neighbours while preserving its own global representation.
For each image-text pair, the two most similar semantic neighbours are selected according to the affinity matrix A. The original global semantic representations F_I^G and F_T^G are subsequently enriched using semantics from their neighbouring samples.

For the image modality, the neighbour-enhanced global representation is computed as

$$\tilde{F}_I^G = 0.7F_I^G + 0.1F_{I,NN_1}^G + 0.2F_{I,NN_2}^G,$$

where F_{I,NN_1}^G and F_{I,NN_2}^G denote the first and second most similar image neighbours, respectively. The original representation receives the largest weight to preserve its semantic identity, while the neighbouring samples provide complementary semantics.
Similarly, the neighbour-enhanced text representation is computed as

$$\tilde{F}_T^G = 0.7F_T^G + 0.1F_{T,NN_1}^G + 0.2F_{T,NN_2}^G,$$

where F_{T,NN_1}^G and F_{T,NN_2}^G denote the corresponding neighbouring text representations.

4.9 Shared Representation Learning

The shared representation learning module propagates the neighbour-enhanced semantic representations through the hypergraph affinity matrix to learn modality-specific and shared latent representations while preserving the higher-order semantic relationships captured by the hypergraph.
Let

$$X_I = \tilde{F}_I^G, \qquad X_T = \tilde{F}_T^G,$$

denote the neighbour-enhanced image and text feature matrices, respectively.
The image and text representations are first propagated independently over the hypergraph adjacency matrix A to learn modality-specific latent representations. Specifically, during graph propagation, each sample is updated by aggregating transformed features from its semantically related neighbours according to the hypergraph adjacency matrix.
The image-specific latent representation is computed as:

$$Z_I = \sigma\left(AX_IW_I\right),$$

while the text-specific latent representation is computed as:

$$Z_T = \sigma\left(AX_TW_T\right),$$

where W_I and W_T denote learnable graph transformation matrices and $\sigma(\cdot)$ denotes the ReLu activation function.
To facilitate cross-modal information interaction, the neighbour-enhanced image and text representations are concatenated to form a joint feature matrix:

$$X_C = [X_I; X_T],$$

The joint feature matrix X_C is propagated over the cross-modal adjacency matrix,

$$Z_J = \sigma\left(A_CX_CW_J\right),$$

where W_J is a learnable graph transformation matrix and the cross-modal adjacency matrix A_C is defined as

$$A_C = \begin{bmatrix} A & I \\ I & A \end{bmatrix},$$

where I denotes the identity matrix, which connects each image with its corresponding text representation, to facilitate cross-modal information propagation
The modality-specific and shared latent representations are subsequently transformed into continuous hash representations,

$$H_I = W_I^h \left[Z_I, Z_J \right],$$

$$H_T = W_T^h \left[Z_J, Z_T \right],$$

while the shared hash representation is computed as

$$H_J = W_J^h Z_J.$$

Finally, the three hash representations are concatenated to obtain the continuous common hash representation H:

$$H = \tanh \left(W_H \left[H_I, H_J, H_T \right] \right).$$

Binary hash codes are generated through sign quantization:

$$B = \operatorname{sign}(H).$$

The resulting binary hash codes preserve both modality-specific semantics and shared cross-modal semantic relationships. These binary hash codes are used as pseudo-labels in the second stage of the MLHE framework to learn the modality-specific hash functions.

4.10 Loss Function

The proposed framework is optimized using four objectives: semantic similarity preservation, quantization, reconstruction, and contrastive learning.

4.10.1 Similarity Preservation Loss

The similarity preservation objective encourages the latent image, text, and shared hash representations to preserve the semantic neighbourhood structure encoded by the hypergraph adjacency matrix A. Moreover, the pairwise cosine similarities computed from the latent representations are encouraged to reflect semantic relationships represented by A, which allows the similarity structure to be preserved in the hash codes B.
The similarity preservation loss is defined as:

$$\mathcal{L}_{sim} = \left\| A - \cos\left(\bar{H}_I, \bar{H}_I \right) \right\|_F^2 + \left\| A - \cos\left(\bar{H}_T, \bar{H}_T \right) \right\|_F^2 + \left\| A - \cos\left(\bar{H}, \bar{H} \right) \right\|_F^2,$$

where A denotes the hypergraph affinity matrix, and $(\bar{H}_I)$, $(\bar{H}_T)$, and $(\bar{H})$ denote the normalized image, text, and shared hash representations, respectively.

4.10.2 Quantization Loss

The quantization objective is introduced to minimize the distance between the continuous hash representations H and the binary representations B, thereby reducing the quantization error introduced during binarization.
The binary hash codes are obtained as:

$$B = \operatorname{sign}(H),$$

and the quantization loss is defined as:

$$\mathcal{L}_{quan} = \|H - B\|_F^2 .$$

4.10.3 Reconstruction Loss

The reconstruction objective encourages the neighbour-enhanced representations to preserve the semantic information learned during graph propagation by reconstructing the graph-enhanced image and textual features through decoder networks.
The reconstruction loss is defined as:

$$\mathcal{L}_{rec} = \left\| D_I - \tilde{F}_I^G \right\|_F^2 + \left\| D_T - \tilde{F}_T^G \right\|_F^2 ,$$

where D_I and D_T denote the reconstructed image and text features, respectively.

4.10.4 Contrastive Loss

The symmetric InfoNCE objective encourages paired image-text representations to remain close in the embedding space, while pushing mismatched pairs apart.
The cosine similarity between the global image and text representations is defined as:

$$\operatorname{sim}\left(F_{I,i}^G, F_{T,j}^G\right) = \cos\left(F_{I,i}^G, F_{T,j}^G\right) .$$

The image-to-text contrastive objective is defined as:

$$\mathcal{L}_{I\to T} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\operatorname{sim}(F_{I,i}^G, F_{T,i}^G)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\operatorname{sim}(F_{I,i}^G, F_{T,j}^G)/\tau\right)},$$

where τ denotes the temperature parameter.
Similarly, the text-to-image contrastive objective is defined as

$$\mathcal{L}_{T\to I} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\operatorname{sim}(F_{T,i}^G, F_{I,i}^G)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\operatorname{sim}(F_{T,i}^G, F_{I,j}^G)/\tau\right)}.$$

The final symmetric contrastive loss is therefore

$$\mathcal{L}_{NCE} = \frac{1}{2}\left(\mathcal{L}_{I\to T} + \mathcal{L}_{T\to I}\right).$$

4.10.5 Overall Objective

The overall optimization objective for the hash code learning stage is

$$\mathcal{L} = \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{sim} + \mathcal{L}_{quan} + \alpha_{NCE} \mathcal{L}_{NCE}.$$

where λ_1, λ_2, and α_{NCE} control the importance of the reconstruction, similarity preservation, quantization, and contrastive objectives.

4.11 Hash Function Learning

The second stage of the proposed MLHE framework learns modality-specific hash functions using the fused global image and text representations of the training samples along with the binary pseudo codes obtained during the hash code learning stage as supervisory signals. During inference, the learned hash functions can directly map the fused global representations of unseen image and text samples into the common Hamming space.
Specifically, the image and text hash codes are produced by applying the corresponding hash functions to the fused global image and text representations:

$$B_I = h_I\left(F_I^G\right),$$

$$B_T = h_T\left(F_T^G\right),$$

where $h_I(\cdot)$ and $h_T(\cdot)$ denote the image and text hash functions, respectively.
The hash functions are optimized using two objectives: the code consistency and graph-preserving objectives. The code consistency objective encourages the generated hash codes to reproduce the binary pseudo codes while simultaneously encouraging the image and text hash codes of the same sample to be consistent. This objective is defined as:

$$\mathcal{L}_{code} = \|B_I - B\|_F^2 + \|B_T - B\|_F^2 + \|B_I - B_T\|_F^2.$$

To preserve the semantic neighbourhood structure learned during the first stage, the pairwise cosine similarities computed from the learned hash codes are constrained to approximate the hypergraph adjacency matrix A. The image-image, text-text, and image-text similarity matrices preserve the intra-modal semantic structure of each modality as well as the semantic consistency across modalities.
The pairwise similarity matrices are computed as:

$$S_{BI,BT} = \cos\left(B_I, B_T\right),$$

$$S_{BI,BI} = \cos\left(B_I, B_I\right),$$

and

$$S_{BT,BT} = \cos\left(B_T, B_T\right).$$

The graph-preserving objective is defined as:

$$\mathcal{L}_{graph} = \|S_{BI,BT} - A\|_F^2 + \|S_{BI,BI} - A\|_F^2 + \|S_{BT,BT} - A\|_F^2.$$

Finally, the overall hash function learning objective is defined as:

$$\mathcal{L}_{hash} = \mathcal{L}_{code} + \beta \mathcal{L}_{graph},$$

where β controls the importance of the code consistency and graph-preserving objectives.

5 Experiments

This section presents the experimental setup used to evaluate the proposed MLHE framework. The proposed framework is evaluated on the three widely used benchmark datasets for cross-modal retrieval: MIRFlickr-25K, NUS-WIDE, and MS COCO. The performance is evaluated on two cross-modal retrieval tasks: image $\rightarrow$ text (I$\rightarrow$T), which retrieves relevant text given an image query, and text $\rightarrow$ image (T$\rightarrow$I), which retrieves relevant images given a text query. Comparison experiments are conducted to evaluate the performance of the proposed MLHE method against five state-of-the-art cross-modal hashing methods. In addition, an ablation experiment is conducted to validate the impact of the contrastive learning objective and the two main learning modules: multi-level representation learning and hypergraph-enhancement. Furthermore, as the comparison baseline methods use AlexNet[45] as the visual backbone, the effect of the visual encoder on the performance of MLHE is investigated by replacing the ResNet-50[46] backbone with AlexNet.

5.1 Datasets and Preprocessing

The performance of the proposed MLHE framework is evaluated on three widely used benchmark datasets for cross-modal retrieval: MIRFlickr-25K[47], NUS-WIDE[48], and MS COCO[49]. To ensure a fair comparison with existing methods, the preprocessing and dataset splits follow the experimental protocol adopted by the compared cross-modal hashing methods [1, 2, 44, 23, 50].

Table 1: Statistics of the datasets

Datasets	Categories	Training	Query	Retrieval
MIRFlickr-25K[47]	24	5,000	2,000	18,015
NUS-WIDE[48]	10	5,000	2,000	135,457
MS COCO[49]	80	5,000	2,000	120,218

5.1.1 MIRFlickr-25K

The MIRFlickr-25K dataset [47] contains 25,000 image-text pairs annotated with multiple labels from the 24 categories. To maintain consistency with the experiments of other cross-modal hashing methods, textual tags appearing fewer than 20 times are removed and subsequently, image-text pairs without textual tags or annotated labels are removed. This preprocessing yields 20,015 valid image-text pairs. The dataset is partitioned using class-based sampling: 160 image-text pairs are randomly selected from the first category and 80 image-text pairs from each of the remaining categories, resulting in a query set of 2,000 image-text pairs. The remaining 18,015 samples form the retrieval database, from which 5,000 image-text pairs are randomly sampled to create the training set.

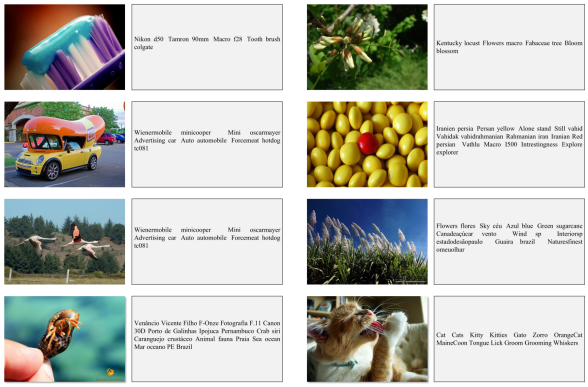


Figure 3: Examples from the MIRFlickr-25k dataset(Source: [51])

5.1.2 NUS-WIDE

The NUS-WIDE[48] dataset consists of 269,648 image-text pairs annotated with 81 concepts. A portion of the original images is no longer publicly available, 197,891 image-text pairs are accessible and used in this study. The data is preprocessed by selecting only the image-text pairs corresponding to the 10 most frequent categories, resulting in 137,457 pairs. Furthermore, only the 1000 most frequent textual tags are retained for building the text vocabulary. Similar to MIRFlickr-25k, the dataset is partitioned using class-based sampling, where 2,000 image-text pairs(200 per concept) are selected for the query set, while the remaining 135,457 samples form the retrieval database. Subsequently, 5,000(500 per concept) image-text pairs are randomly sampled from the retrieval database to create the training set.

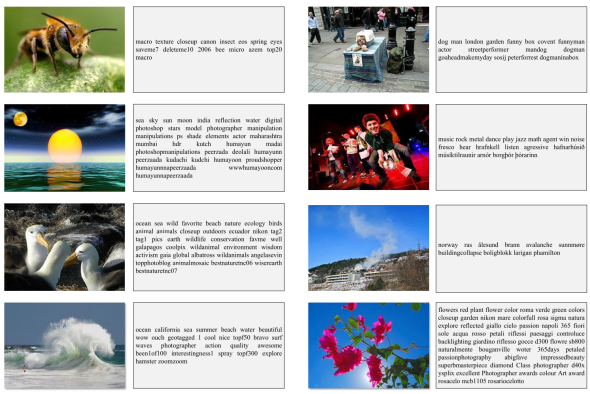


Figure 4: Examples from the NUS-WIDE dataset(Source: [51])

5.1.3 MS COCO

The MS COCO[49] dataset contains 123,287 image-text pairs across 80 categories, where each image is annotated by 5 captions. The pairs without category, caption, or valid image path are removed, which results in 122,218 pairs. Following the same class-based partition strategy used for MIRFlickr-25K and NUS-WIDE, 2,000(25 per category) pairs are randomly selected for the query set, the remaining 120,218 pairs for the retrieval set, and 5,000 pairs are randomly sampled from the retrieval set for training.

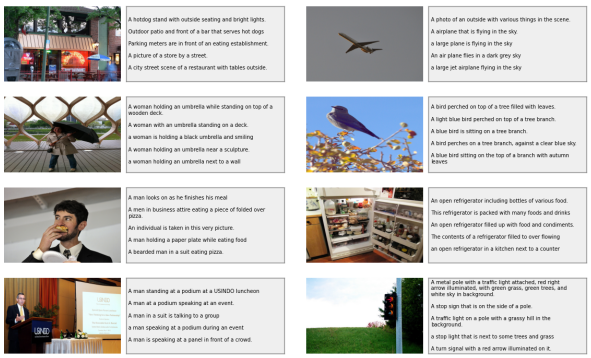


Figure 5: Examples from the MS COCO dataset

5.2 Baseline Methods and Evaluation Metrics

The proposed MLHE method is compared with five state-of-the-art unsupervised baseline methods: Unsupervised Deep Cross-Modal Hashing(UDCMH)[50], Joint-modal Distribution-based Similarity Hashing(JDSH)[44], Aggregation-based Graph Convolutional Hashing (AGCH)[23], Deep Joint Semantics Reconstructing Hashing(DJSRH)[2], Correlation-Identity Reconstruction Hashing(CIRH)[1]. In addition to the originally reported results, reproduced implementations of DJSRH and CIRH, denoted as DJSRH† and CIRH†, are evaluated using the same data preprocessing, class-based data splits, and evaluation protocol to ensure fair comparison. Additionally, as the baseline methods use AlexNet[45] as the visual backbone, an additional variant of MLHE, denoted as MLHE*, is evaluated using AlexNet instead of ResNet-50[46]. The retrieval performance of all methods is evaluated using three evaluation metrics: mean Average Precision (mAP), Precision-Recall curves, and Precision@N curves. An image data point x_i and a text data point x_t are considered a similar (relevant) pair if they share at least one common label. Otherwise, they are treated as a different (non-relevant) pair. This definition is used to determine relevance in all evaluation metrics. In addition, the efficiency of the proposed MLHE method is assessed by comparing the training and hash code generation time with the two baselines DJSRH† [2] and CIRH† [1].

5.2.1 Mean Average Precision

The retrieval performance of the methods is evaluated using mean Average Precision (mAP). Specifically, for each query, the Average Precision (AP) is computed as:

$$\mathrm{AP} = \frac{1}{N} \sum_{k=1}^{n} P(k) \cdot \mathrm{rel}(k),$$

where n is the number of retrieved items, N is the total number of relevant items for the query, (P(k)) denotes the precision at rank k, and $\mathrm{rel}(k)$ is an indicator function that equals 1 if the retrieved item at rank k is relevant, and 0 otherwise. The mean Average Precision (mAP) is then obtained by averaging the AP values over all queries:

$$\mathrm{mAP} = \frac{1}{Q} \sum_{q=1}^{Q} AP_q$$

where Q is the total number of queries and AP_q represents the Average Precision for the q-th query.

5.2.2 Efficiency Metrics

The computational efficiency of the methods is evaluated with the training and hash generation time. The training time is measured as the total time required to complete the entire training process over 60 epochs. The hash code generation time is measured during the evaluation stage, which is the time required to generate binary hash codes, using the learned modality-specific hash functions, for the query and retrieval samples.

5.2.3 Curves

The precision-recall and precision@N curves are plotted on the three benchmark datasets with the hash code length set to 128 bits.

Precision-Recall Curves Precision-recall (PR) curves illustrate the trade-off between precision and recall at different retrieval depths. For each query, the retrieved samples are ranked according to their Hamming distance. Subsequently, precision and recall are computed at 20 uniformly sampled retrieval depths and averaged over all queries. These curves provide insights into the effectiveness of a method in increasing recall, by retrieving a larger proportion of the relevant samples, while maintaining high precision.

Precision@N Curves Precision@N curves measure the precision within the top-N retrieved samples, where N varies over a predefined range ($N = 1000, 2000, 3000, 4000, 5000$). For each query, precision@N is defined as the proportion of relevant samples among the top-N retrieved samples. In cross-modal retrieval, these curves provide insights into the effectiveness of the method at different retrieval depths, High precision at small N indicates that the model retrieves relevant samples at the top of the ranking, which is critical for practical applications. However, to remain consistent with the evaluation strategy in existing work[2, 44, 23, 50], larger retrieval depths are used in this study.

5.3 Ablation Experiment

The impact of the two main components, multi-level representation learning and hypergraph enhancement, on the MLHE framework are evaluated by conducting ablation experiments. The ablation experiments are conducted on the MIRFlickr-25K dataset using 128-bit hash codes. Three MLHE variants are used to evaluate the impact of those components on the retrieval performance: MLHE w/o local features, MLHE w/o hypergraph enhancement, and MLHE w/o contrastive learning.

5.4 Backbone Comparison

The effect of the visual backbone on the retrieval performance of MLHE is evaluated by implementing an additional MLHE variant, denoted as MLHE*, using AlexNet while keeping all other components of the framework unchanged. The learning rate of the main MLHE network was reduced from5×10^{-5} to 1×10^{-4}, while the other hyperparameters are kept consistent with the proposed MLHE framework.

5.5 Implementation Details

The MLHE framework is implemented using PyTorch and is executed on a workstation containing NVIDIA A100 GPUs. Unlike most existing cross-modal hashing methods[1, 2, 44], which employ AlexNet and bag-of-words (BoW), MLHE adopts ResNet-50 as the visual backbone and trainable word embeddings for text representation, enabling richer semantic feature learning. Images are resized to 224×224 pixels before feature extraction, whereas text sequences are truncated or padded to a maximum length of 32 words. The final convolutional block of ResNet-50 is used to obtain the feature maps($2048 \times 7 \times 7$), which are used to derive the global and local image representations. The global image representation is obtained by applying global average pooling, followed by a linear projection that maps the 2048-dimensional feature vector to a 512-dimensional embedding. Simultaneously, the local representations are obtained by applying (1×1) convolution to project the feature maps from 2048 to 512 dimensions, producing 49 region tokens. The text encoder is a trainable embedding network that maps each text token to a 512-dimensional embedding vector. The global text representation is obtained by applying masked mean pooling over the token embeddings, followed by linear projection and normalization. MLHE employs two Transformer encoders, each consisting of two encoder layers with four attention heads, to model cross-modal fusion. The first Transformer models interactions between the aggregated image-region and text-region representations, while the second models interactions between cross-weighted global image and text representations. Specifically, in the global fusion module, image features are weighted using the global text representation, whereas text features are weighted using the global image representation before cross-modal fusion. The resulting Transformer-enhanced global representations are then combined with the respective original global representations through weighted feature fusion, where $\beta_f = 0.4$ controls the importance of the original and Transformer-enhanced features. A frozen copy of the multi-level encoder is used to extract image and text representations for similarity graph construction, while the trainable encoder is optimized for representation and hash-code learning. The hypergraph-enhanced learning module consists of three graph convolution layers, corresponding to the image, text, and joint graph representations. Each layer is implemented as a $512 \rightarrow 512$ linear projection followed by Batch Normalization and ReLU activation. The image-specific, text-specific, and joint representations are projected into the hash space using fully connected layers, while the shared hash representation is obtained by a fully connected layer of size $(3r \rightarrow r)$, where r denotes the hash code length. The image and text hash functions each consist of a two-layer multi-layer perceptron followed by a `tanh` activation. The learning rates of the hash code and hash function learning are set to 5×10^{-5} and 1×10^{-4}, respectively, on MIRFlickr-25K and NUS-WIDE, and 5×10^{-5} and 5×10^{-4} respectively on MS COCO. The graph weighting parameters are set to $a_1 = 0.6$ and $a_2 = 0.6$ for MIRFlickr and NUS-WIDE, and to $a_1 = 0.1$ and $a_2 = 0.3$ for MS COCO, where a_1 balances the image- and text-based similarity graphs and a_2 controls the fusion between the collaborative similarity graph S_1 and its nonlinearly enhanced semantic similarity matrix S_2. Unless otherwise stated, the remaining hyperparameters are shared across all benchmark datasets. The hypergraph is constructed using $K_h = 12$ nearest neighbours with hyperedge weighting parameter $\rho = 0.2$, while the graph pruning parameter is fixed at $K = 3000$. The importance assigned to the region-level similarity during similarity graph construction is fixed to $\alpha_r = 0.2$. The region alignment parameter is fixed to $\gamma = 10$, which controls the weighting of the selected word-region correspondences before region-level feature aggregation. The semantic reconstruction and adjacency reconstruction are weighted by $\lambda_1 = 10$ and $\lambda_2 = 1$, respectively,

whereas the importance of the hash function learning objective is controlled by $\beta = 0.01$. Finally, the contrastive learning weight is fixed to $\alpha_{\text{nce}} = 0.03$, and the InfoNCE temperature parameter is set to $\tau = 0.07$. Training is performed for 60 epochs using a mini-batch size of 512 image-text pairs. The model is evaluated every ten epochs using mean Average Precision (mAP@50) for both image-to-text and text-to-image retrieval.

6 Results

The proposed Multi-Level Hypergraph-Enhanced (MLHE) framework is compared with state-of-the-art unsupervised cross-modal hashing methods on the MIRFlickr-25K, NUS-WIDE, and MS COCO datasets. The performance of the methods is evaluated using mAP@50 for both Image-to-Text and Text-to-Image retrieval across four hash code lengths (16, 32, 64, and 128 bits). In addition, Precision-Recall curves, Precision@N curves, training time, and hash code generation time are reported, for MLHE, MLHE*, DJSRH†, and CIRH†, to provide insights into both the retrieval effectiveness and computational efficiency.

Table 2: Comparison of mAP@50 for Image-to-Text (I→T) and Text-to-Image (T→I) retrieval on MIRFlickr, NUS-WIDE, and MS COCO at different hash code lengths.

Task	Method	Year	Image Encoder	Text Encoder	MIRFlickr				NUS-WIDE				MS COCO			
					16	32	64	128	16	32	64	128	16	32	64	128
I→T	UDCMH[50]	2018	AlexNet	BoW	0.689	0.698	0.714	0.717	0.511	0.519	0.524	0.558	–	–	–	–
	JDSH[44]	2020	AlexNet	BoW	0.832	0.853	0.882	0.892	0.736	0.793	0.832	0.835	0.694	0.738	0.769	0.788
	AGCH[23]	2021	AlexNet	BoW	0.865	0.887	0.892	0.912	0.809	0.830	0.831	0.852	0.741	0.772	0.789	0.806
	DJSRH[2]	2019	AlexNet	BoW	0.810	0.843	0.862	0.876	0.724	0.773	0.798	0.817	0.678	0.724	0.743	0.768
	DJSRH†[2]	2019	AlexNet	BoW	0.829	0.857	0.882	0.893	0.719	0.802	0.810	0.831	–	–	–	–
	CIRH[1]	2023	AlexNet	BoW	0.901	0.913	0.929	0.937	0.815	0.836	0.854	0.862	0.797	0.819	0.830	0.849
	CIRH†[1]	2023	AlexNet	BoW	0.857	0.880	0.893	0.904	0.777	0.808	0.828	0.837	0.650	0.746	0.806	0.819
	MLHE*	2026	AlexNet	Word Embeddings	0.827	0.852	0.862	0.869	0.777	0.799	0.815	0.817	0.675	0.719	0.738	0.748
	MLHE	2026	ResNet50	Word Embeddings	**0.928**	**0.939**	**0.948**	**0.948**	**0.839**	**0.871**	**0.864**	**0.880**	**0.808**	**0.846**	**0.854**	**0.875**
T→I	UDCMH[50]	2018	AlexNet	BoW	0.692	0.704	0.718	0.733	0.637	0.653	0.695	0.716	–	–	–	–
	JDSH[44]	2020	AlexNet	BoW	0.825	0.864	0.878	0.880	0.721	0.785	0.794	0.804	0.703	0.759	0.793	0.825
	AGCH[23]	2021	AlexNet	BoW	0.829	0.849	0.852	0.880	0.769	0.780	0.798	0.802	0.746	0.774	0.797	0.817
	DJSRH[2]	2019	AlexNet	BoW	0.786	0.822	0.835	0.847	0.712	0.744	0.771	0.789	0.650	0.753	0.805	0.823
	DJSRH†[2]	2019	AlexNet	BoW	0.815	0.843	0.863	0.867	0.733	0.744	0.802	0.808	–	–	–	–
	CIRH[1]	2023	AlexNet	BoW	0.867	0.885	0.900	0.901	0.774	0.803	0.810	0.817	0.811	**0.847**	**0.872**	**0.895**
	CIRH†[1]	2023	AlexNet	BoW	0.845	0.863	0.876	0.875	0.760	0.783	0.803	0.806	0.638	0.737	0.850	0.861
	MLHE*	2026	AlexNet	Word Embeddings	0.833	0.857	0.859	0.868	0.783	0.796	0.817	0.814	0.693	0.710	0.730	0.738
	MLHE	2026	ResNet50	Word Embeddings	**0.924**	**0.939**	**0.947**	**0.942**	**0.848**	**0.867**	**0.865**	**0.861**	**0.813**	0.831	0.847	0.865

† Reproduced baselines using the proposed data preprocessing.

* MLHE implemented with AlexNet as the visual backbone for fair comparison with baselines

Table 2 compares the retrieval performance of the hashing methods on the three benchmark datasets. The proposed MLHE method achieves the best retrieval accuracy on MIRFlickr-25k and NUS-WIDE across the hash bit codes in both I→T and T→I retrieval. On the MS COCO dataset, MLHE achieves the highest retrieval performance for I→T retrieval across all hash code lengths, while achieving the highest T→I retrieval performance only at the 16-bit hash length. More specifically, MLHE improves the best baseline CIRH by 1.1%–2.7% for Image-to-Text retrieval and 4.1%–5.7% for Text-to-Image retrieval on the MIRFlickr-25K dataset. On the NUS-WIDE dataset, MLHE improves the best baseline by 1.0%–3.5% for I→T retrieval and 4.4%–7.4% for T→I retrieval. On the MS COCO dataset, MLHE improves I→T retrieval by 1.4%–3.3% across all hash code lengths. For T→I retrieval, MLHE achieves a small improvement of 0.2% at 16 bits, while CIRH achieves higher retrieval performance for the remaining hash code lengths.

The results of MLHE and the AlexNet-based variant, MLHE*, show the impact of the visual backbone on the retrieval performance. By replacing ResNet-50 with AlexNet, the model

achieved significantly lower retrieval performance across all datasets, retrieval tasks, and hash code lengths. Compared with MLHE, MLHE* shows performance reductions of 5.9%–10.1% for I→T retrieval and 5.6%–9.4% for T→I retrieval on MIRFlickr-25K, 4.0%–6.3% for I→T retrieval and 4.8%–6.5% for T→I retrieval on NUS-WIDE, and 6.0%–12.7% for I→T retrieval and 5.2%–15.7% for T→I retrieval on MS COCO.

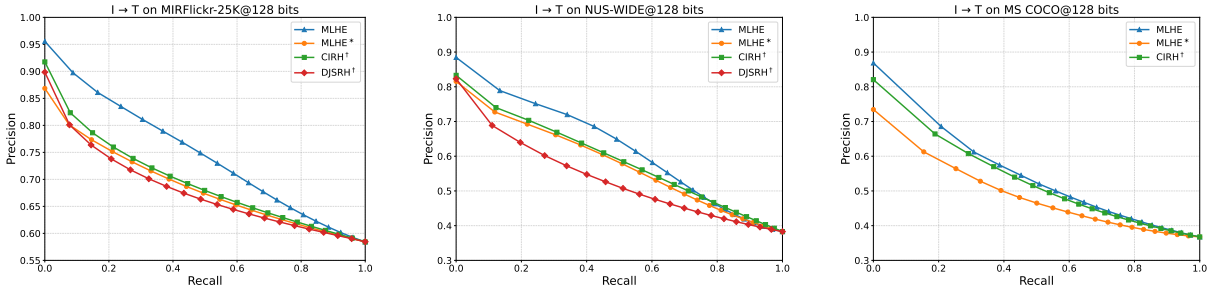
Table 3: Comparison of Training Time (s) on Three Benchmark Datasets with 128-bit Hash Codes.

Method	MIRFlickr	NUS-WIDE	MS COCO
DJSRH†[2]	824.73	3807.83	–
CIRH†[1]	**154.79**	**354.36**	811.29
MLHE*	1599.51	3747.38	1830.48
MLHE	1558.99	4383.66	2664.89

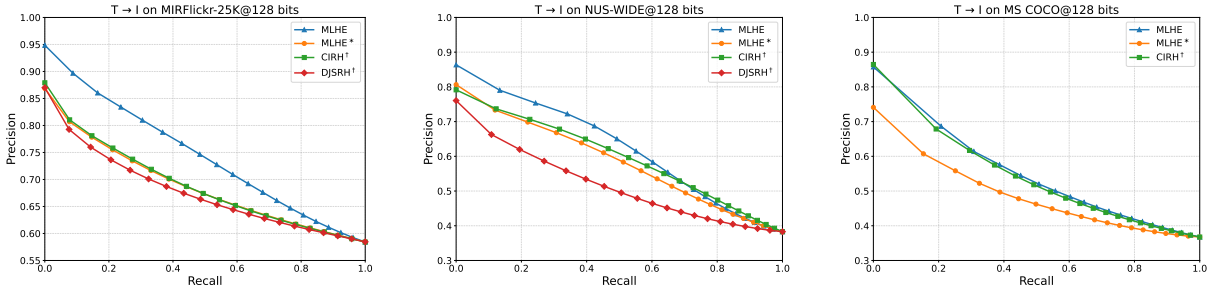
Table 4: Comparison of Hash Code Generation Time (s) on Three Benchmark Datasets with 128-bit Hash Codes.

Method	MIRFlickr	NUS-WIDE	MS COCO
DJSRH†[2]	34.89	392.69	–
CIRH†[1]	**3.76**	**7.11**	25.38
MLHE*	56.04	468.32	160.92
MLHE	43.66	442.31	205.37

Tables 3 and 4 compare the computational efficiency of MLHE and the reproduced baselines using 128-bit hash codes. The results show MLHE requires higher training and hash code generation times than CIRH† and DJSRH† on the datasets. The increased computational cost is expected, as MLHE incorporates several components that are absent in the baseline methods, including the region-word alignment module, Transformer-based feature fusion, and hypergraph construction.



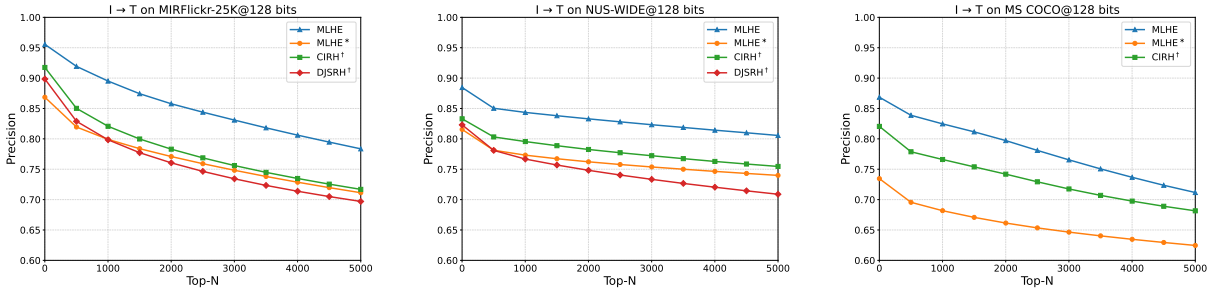
(a) MIRFlickr-25K (b) NUS-WIDE (c) MS COCO



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

Figure 6: Precision–Recall curves for image-to-text (top row) and text-to-image (bottom row) retrieval on the MIRFlickr-25K, NUS-WIDE, and MS COCO datasets using 128-bit hash codes.

Figure 6 presents the precision-recall curves on the benchmark datasets using 128-bit hash codes. On MIRFlickr-25K, MLHE consistently achieves higher precision than DJSRH† and CIRH† across the entire recall range for both retrieval directions. This indicates that MLHE can retrieve a larger proportion of relevant samples while maintaining fewer false positives. On NUS-WIDE, MLHE achieves higher precision than DJSRH† and CIRH† at low recall values. As recall increases, the performance gap narrows, and the precision of the three methods converges at higher recall levels. On MS COCO, the precision–recall curves of MLHE and CIRH† are nearly identical across the recall range, which indicates both methods achieve similar retrieval performance on this more challenging dataset.



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

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

Figure 7: Precision@N curves for image-to-text (top row) and text-to-image (bottom row) retrieval on the MIRFlickr-25K, NUS-WIDE, and MS COCO datasets using 128-bit hash codes.

Figure 7 presents the Precision@N curves on the benchmark datasets using 128-bit hash codes. On all datasets, MLHE achieves higher precision than DJSRH† and CIRH† across all retrieval depths, which indicates MLHE ranks relevant cross-modal samples more accurately.

Table 5: Ablation study on the MIRFlickr-25K dataset using different hash code lengths.

Task	**Method**	**MIRFlickr-25K**			
		16	**32**	**64**	**128**
I→T	MLHE w/o Local Features	0.926	0.937	0.946	0.947
	MLHE w/o Hypergraph	0.919	0.931	0.947	**0.948**
	MLHE w/o Contrastive Learning	0.918	0.920	0.931	0.923
	MLHE	**0.928**	**0.939**	**0.948**	**0.948**
T→I	MLHE w/o Local Features	0.923	**0.942**	0.941	0.941
	MLHE w/o Hypergraph	0.921	0.934	0.946	**0.945**
	MLHE w/o Contrastive Learning	0.916	0.927	0.926	0.909
	MLHE	**0.924**	0.939	**0.947**	0.942

Table 5 presents the results from the ablation study on the MIRFlickr-25K dataset across different hash code lengths. Removing the local feature module or the hypergraph enhancement module results in small changes in retrieval performance, with mAP@50 differences below 1% across both retrieval tasks. In contrast, removing the contrastive learning objective results in a significantly performance decrease, reducing mAP@50 by 1.0%–2.5% for the I→T task and 0.8%–3.3% for the T→I task across the hash code lengths.

7 Discussion

The experimental results demonstrate the superiority of the proposed Multi-Level Hypergraph-Enhanced (MLHE) framework in unsupervised cross-modal hashing. MLHE achieves significantly higher retrieval performance on MIRFlickr-25K and NUS-WIDE compared to the baseline methods across the I→T and T→I retrieval tasks and evaluated hash code lengths(16,32,64,128). These results indicate that combining global semantic representations with fine-grained region-word correspondences enables the learned hash codes to preserve cross-modal semantic relationships more effectively than methods relying only on global representations. Moreover, the consistent improvements across different hash code lengths suggest the method remains effective under stronger compression. On the MS COCO dataset, MLHE achieves superior performance on the I→T task across all hash codes. However, on the T→I task, MLHE is outperformed by CIRH at the 32,64, and 128 hash code lengths. The relatively small performance differences and the nearly identical Precision–Recall curves indicate that both methods achieve comparable retrieval performance on this dataset. These results suggest that on datasets with dense multi-object images, the standard global semantic modeling can effectively capture rich global semantics, thereby narrowing the performance gap with fine-grained alignment methods such as MLHE. The results also provide important insights into the effect of the visual backbone on the retrieval performance. Replacing the ResNet-50 backbone with AlexNet in MLHE* results in a significant performance decrease across all datasets. Since the remaining components of the framework are kept unchanged, these results indicate that stronger visual representations significantly improve the quality of the learned similarity matrix and the resulting hash codes. However, the hyperparameters were optimized for ResNet-50, therefore, the MLHE* baseline might achieve lower performance due to unoptimized hyperparameters for the AlexNet backbone. More specifically, due to time constraints, only the learning rate of MLHE* was adjusted, while all other hyperparameters remained unchanged. The Precision–Recall and Precision@N curves provide additional insight into the retrieval performance of MLHE. The curves also demonstrate the superior performance of MLHE, compared to CIRH† and DJSRH†, where MLHE maintains higher precision across a range of recall values and different retrieval depths on all benchmark datasets. Specifically, on MIRFlickr-25K, MLHE maintains higher precision over the complete recall range and consistently ranks relevant samples ahead compared to CIRH and DJSRH. On NUS-WIDE, the superior performance of MLHE mostly occurs at lower recall values, as it converges with the baseline methods at higher recall levels. This indicates that MLHE retrieves the most relevant cross-modal pairs earlier in the ranking. On MS COCO, the Precision–Recall curves of MLHE and CIRH are consistent with the mAP results, confirming that both methods achieve comparable retrieval performance. Despite the performance improvements, the computational efficiency metrics show MLHE requires significantly longer training and hash code generation times. The increase in computational resources can be attributed to the Transformer-based feature fusion, region-word alignment, and hypergraph construction modules. The ablation study shows small performance differences after removing the local feature module or the hypergraph enhancement module. This suggests that the performance gains of MLHE may arise from the complementary interaction of these components rather than any single module. Nevertheless, more comprehensive experiments are required to better quantify the contribution of those components. In contrast, removing the contrastive learning objective results in a significant decrease in retrieval performance. This indicates that contrastive learning plays a crucial role in aligning image and text representations within the shared embedding space and in learning more discriminative hash codes.

The MLHE framework contains several limitations. First, the framework introduces a large number of hyperparameters for the different modules; an extensive hyperparameter sensitivity analysis was not conducted due to time constraints. Therefore, the selected hyperparameter values may not be optimal, and additional hyperparameter tuning may further improve performance. Second, while MLHE consistently outperforms the baseline methods on MIRFlickr-25K and NUS-WIDE, its performance gains are less pronounced on the MS COCO dataset. This may result from the high-quality annotations in MS COCO, which allow methods that only use global semantic representations to capture most cross-modal relationships effectively. This suggests MLHE is particularly effective for complex, weakly-annotated, and noisy datasets. Finally, the proposed framework is only evaluated on three image-text retrieval benchmark datasets. Therefore, its generalizability to other retrieval scenarios, such as video-text retrieval, larger-scale multimodal datasets, or fine-grained retrieval benchmarks, remains to be investigated.

8 Conclusion

This study proposed a novel framework, denoted as Multi-Level Hypergraph-Enhanced(MLHE), for unsupervised cross-modal hashing. The objective was to improve cross-modal retrieval by learning more discriminative binary hash codes that preserve semantic relationships between image and text modalities. The proposed MLHE framework integrates multi-level representation learning with hypergraph-based modelling. Specifically, the framework combines global semantic representations with fine-grained region-word correspondences to construct a richer semantic similarity matrix, while the hypergraph module captures higher-order semantic relationships that cannot be represented by pairwise graphs. The experimental results on the MIRFlickr-25K, NUS-WIDE, and MS COCO benchmark datasets demonstrate the effectiveness of the proposed framework. MLHE outperforms the hashing baseline methods on MIRFlickr-25K and NUS-WIDE across both Image-to-Text and Text-to-Image retrieval tasks across different hash code lengths. Furthermore, the Precision–Recall and Precision@N curves show that MLHE retrieves relevant cross-modal samples more effectively than the baseline methods across different retrieval depths. The performance gains are less pronounced on the MS COCO dataset, containing relatively high-quality annotations compared to the other datasets, which suggests MLHE is particularly effective for complex, weakly-annotated, and noisy datasets. Additionally, experiments replacing ResNet-50 with AlexNet in the MLHE framework shows the stronger visual representations produced by ResNet-50 have a significant impact on the retrieval performance of the proposed framework. The computational efficiency experiments highlight the trade-off between retrieval accuracy and computational efficiency introduced by the MLHE modules, as it requires significantly higher computational costs compared to the baselines. Finally, the ablation study shows small performance differences after removing the local feature module or the hypergraph enhancement module. Since the observed differences are small, the results indicate that further investigation into the local feature module and hypergraph module may be needed.

Future work could incorporate modern vision foundation models and pretrained language models to investigate the effects on improving the retrieval performance. Additionally, experiments could be conducted on reducing the computational overhead of the proposed framework. Finally, additional experiments could be conducted to provide insights into the generalizability of the proposed framework by extending it to other multimodal retrieval scenarios, such as image-video or fine-grained retrieval

References

[1] Lei Zhu, Xize Wu, Jingjing Li, Zheng Zhang, Weili Guan, and Heng Tao Shen. Work together: Correlation-identity reconstruction hashing for unsupervised cross-modal retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):8838–8851, 2023.

[2] Shupeng Su, Zhisheng Zhong, and Chao Zhang. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3027–3035, 2019.

[3] Peng Hu, Hongyuan Zhu, Jie Lin, Dezhong Peng, Yin-Ping Zhao, and Xi Peng. Unsupervised contrastive cross-modal hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3877–3889, 2023.

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

[5] R. Wei, Y. Liu, H. Cui, Y. Xie, and K. Zhou. Graph contrastive-and-reconstructive hashing for unsupervised cross-modal retrieval. *Data Science and Engineering*, 10(3):411–427, 2025.

[6] Y. Shi et al. Deep adaptively-enhanced hashing with discriminative similarity guidance for unsupervised cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):7255–7268, 2022.

[7] Fengling Li, Bowen Wang, Lei Zhu, Jingjing Li, Zheng Zhang, and Xiaojun Chang. Cross-domain transfer hashing for efficient cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10):9664–9677, 2024.

[8] Hao Du, Huaxiang Zhang, Bin Yin, Li Liu, and Hengchang Wang. Unsupervised multi-level semantic fusion-driven hashing for cross-modal retrieval. In *Pattern Recognition and Computer Vision: 8th Chinese Conference (PRCV)*, pages 329–344, 2025.

[9] Fangming Zhong, Chenglong Chu, Zijie Zhu, and Zhikui Chen. Hypergraph-enhanced hashing for unsupervised cross-modal retrieval via robust similarity guidance. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3517–3527, 2023.

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

[11] Ruiqing Xu, Chao Li, Junchi Yan, Cheng Deng, and Xianglong Liu. Graph convolutional network hashing for cross-modal retrieval. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 982–988, 2019.

[12] Changchang Sun, Hugo Latapie, Gaowen Liu, and Yan Yan. Deep normalized cross-modal hashing with bi-direction relation reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4937–4945, 2022.

[13] Kai Han, Yu Liu, Rukai Wei, Ke Zhou, Jinhui Xu, and Kun Long. Supervised hierarchical online hashing for cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 20(4), 2024.

[14] Donglin Zhang, Xiao-Jun Wu, Tianyang Xu, and Josef Kittler. Two-stage supervised discrete hashing for cross-modal retrieval. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(11):7014–7026, 2022.

[15] David R. Hardoon, Sandor R. Szedmak, and John R. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[16] Xin Liu, Zhikai Hu, Haibin Ling, and Yiu-Ming Cheung. Mtfh: A matrix tri-factorization hashing framework for efficient cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):964–981, 2021.

[17] Qing-Yuan Jiang and Wu-Jun Li. Deep cross-modal hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3270–3278, 2017.

[18] Di Hu, Feiping Nie, and Xuelong Li. Deep binary reconstruction for cross-modal hashing. *IEEE Transactions on Multimedia*, 21(4):973–985, 2019.

[19] Shiping Ge, Zhiwei Jiang, Yafeng Yin, Cong Wang, Zifeng Cheng, and Qing Gu. Fine-grained alignment network for zero-shot cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 21(10), 2025.

[20] Yangchao Wang, Shiyuan He, Xing Xu, Yang Yang, Jingjing Li, and Heng Tao Shen. Self-supervised adversarial learning for cross-modal retrieval. In *Proceedings of the 2nd ACM International Conference on Multimedia in Asia*, 2021.

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

[22] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

[23] Peng-Fei Zhang, Yang Li, Zi Huang, and Xin-Shun Xu. Aggregation-based graph convolutional hashing for unsupervised cross-modal retrieval. *IEEE Transactions on Multimedia*, 24:466–479, 2022.

[24] Guohui Ding, Jing Li, Yimin Xu, and Rui Zhou. Hl-cmr: Hypergraph learning for cross-modal retrieval. In *Proceedings of the ACM Web Conference 2026*, pages 2263–2272, 2026.

[25] Shengsheng Qian, Dizhan Xue, Quan Fang, and Changsheng Xu. Integrating multi-label contrastive learning with dual adversarial graph neural networks for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4794–4811, 2023.

[26] Cong Bai, Chao Zeng, Qing Ma, and Jinglin Zhang. Graph convolutional network discrete hashing for cross-modal retrieval. *IEEE Transactions on Neural Networks and Learning Systems*, 35(4):4756–4767, 2024.

[27] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Hierarchical multimodal lstm for dense visual-semantic embedding. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1899–1907, 2017.

[28] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676, 2017.

[29] Jiuxiang Gu, Jianfei Cai, Shafiq Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7181–7189, 2018.

[30] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Lecture Notes in Computer Science*, pages 212–228, 2018.

[31] Zhengxin Pan, Fangyu Wu, and Bailing Zhang. Fine-grained image-text matching by cross-modal hard aligning network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19275–19284, 2023.

[32] M. Liu, A. S. M. Khairuddin, K. Hasikin, and W. Liu. Novel cross-dimensional coarse-fine-grained complementary network for image-text matching. *PeerJ Computer Science*, 11:e2725, 2025.

[33] Xiaoyu Yang, Chao Li, Dongzhong Zheng, Peng Wen, and Guangqiang Yin. Rfe-srn: Image-text similarity reasoning network based on regional feature enhancement. *Neurocomputing*, 518:593–601, 2023.

[34] Yuxin Peng, Jinwei Qi, and Yunkan Zhuo. Mava: Multi-level adaptive visual-textual alignment by cross-media bi-attention mechanism. *IEEE Transactions on Image Processing*, 29:2728–2741, 2020.

[35] Qiyou Huang and Meng Wang. Unsupervised cross-modal semantic invariance hashing. In *Proceedings of the 2026 International Conference on Multimedia Retrieval*, pages 157–166, 2026.

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

[37] Di Wang, Quan Wang, and Xinbo Gao. Robust and flexible discrete hashing for cross-modal similarity search. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2703–2715, 2018.

[38] Christian Hansen, Casper Hansen, Jakob Grue Simonsen, and Christina Lioma. Projected hamming dissimilarity for bit-level importance coding in collaborative filtering. In *Proceedings of the Web Conference 2021*, pages 261–269, 2021.

[39] Ying Ma, Meng Wang, Guangyun Lu, and Yajun Sun. Deep hashing similarity learning for cross-modal retrieval. *IEEE Access*, 12:8609–8618, 2024.

[40] Xinfeng Dong, Li Liu, Lei Zhu, Liqiang Nie, and Huaxiang Zhang. Adversarial graph convolutional network for cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1634–1645, 2022.

[41] J. Yu, H. Zhou, Y. Zhan, and D. Tao. Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4626–4634, 2021.

[42] Lei Zhang, Leiting Chen, Chuan Zhou, Xin Li, Fan Yang, and Zhang Yi. Weighted graph-structured semantics constraint network for cross-modal retrieval. *IEEE Transactions on Multimedia*, 26:1551–1564, 2024.

[43] J. Yu, H. Zhou, Y. Zhan, and D. Tao. Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4626–4634, 2021.

[44] Song Liu, Shengsheng Qian, Yang Guan, Jiawei Zhan, and Long Ying. Joint-modal distribution-based similarity hashing for large-scale unsupervised deep cross-modal retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1379–1388, 2020.

[45] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pages 1097–1105, 2012.

[46] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[47] Mark J. Huiskes and Michael S. Lew. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, pages 39–43, 2008.

[48] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, 2009.

[49] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755, 2014.

[50] Gengshen Wu, Zijia Lin, Jungong Han, Li Liu, Guiguang Ding, Baochang Zhang, and Jialie Shen. Unsupervised deep hashing via binary latent factor models for large-scale cross-modal retrieval. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2854–2860, 2018.

[51] Xueying Zhu, Lei Cai, Ziqiang Zou, and Lei Zhu. Deep multi-semantic fusion-based cross-modal hashing. *Mathematics*, 10(3):430, 2022.