HIERARCHICAL ENCODING TREE WITH MODALITY MIXUP FOR CROSS-MODAL HASHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-modal retrieval is a significant task that aims to learn the semantic correspondence between visual and textual modalities. Unsupervised hashing methods can efficiently manage large-scale data and can be effectively applied to cross-modal retrieval studies. However, existing methods typically fail to fully exploit the hierarchical structure between text and image data. Moreover, the commonly used direct modal alignment cannot effectively bridge the semantic gap between these two modalities. To address these issues, we introduce a novel **Hi**erarchical Encodi**n**g **T**ree with Modality Mixup (**HINT**) method, which achieves effective cross-modal retrieval by extracting hierarchical cross-modal relations. HINT constructs a cross-modal encoding tree guided by hierarchical structural entropy and generates *proxy* samples of text and image modalities for each instance from the encoding tree. Through the curriculum-based mixup of proxy samples, HINT achieves progressive modal alignment and effective cross-modal retrieval. Furthermore, we conduct cross-modal consistency learning to achieve global-view semantic alignment between text and image representations. Extensive experiments on a range of cross-modal retrieval datasets demonstrate the superiority of HINT over state-of-the-art methods. Our source codes are available at this link.

1 INTRODUCTION

Cross-modal retrieval aims to measure the semantic similarity between different modalities, using retrieval methods such as approximate nearest neighbors (ANNs) search (Zhu et al., 2023; Zhen et al., 2019). Cross-modal retrieval has significant application value, such as in retrieval-augmented generation (RAG) (Li et al., 2024b; Cui et al., 2024) and search engines (Song et al., 2024; Chen et al., 2017). With the rapid development of large-scale vision-language datasets, cross-modal retrieval has attracted increasing attention. Therefore, researchers have turned to hashing-based cross-modal retrieval (Cao et al., 2017; Yan et al., 2020), which achieves efficient storage and indexing by replacing computationally expensive pairwise distance comparisons with bit-wise operations. Hashing-based cross-modal retrieval methods map high-dimensional semantic vectors from different modalities into a unified Hamming space (binary space) (Luo et al., 2023; Huang et al., 2024), enabling similarity comparison and ANNs.

Cross-modal hashing has garnered significant attention from the community (Zhang et al., 2024b; Sun et al., 2024; Tu et al., 2023). It includes supervised and unsupervised methods. Supervised cross-modal hashing (Shen et al., 2024; Ma et al., 2024; Liu et al., 2019a; Lu et al., 2019) learn hash codes using labeled data. However, due to the expensiveness and scarcity of cross-modal annotations in real-world scenarios (Wang et al., 2023), researchers have shifted their focus to unsupervised methods that do not rely on annotations. Unsupervised cross-modal hashing (Liang et al., 2024; Li et al., 2024a; Zhang et al., 2018) leverage pair-wise cross-modal data, exploiting the similarity between samples and employing mechanisms such as contrastive learning (Hu et al., 2022) and adversarial learning (Li et al., 2019) to guide hash learning.

Unsupervised cross-modal hashing has made promising progress (Liang et al., 2024; Li et al., 2024a; Zhang et al., 2024b; Tu et al., 2023), but still suffers from the following issues: The first challenge is the ***lack of hierarchical semantic structure***. Due to the absence of annotations, previous methods mainly rely on paired data (Hu et al., 2022), such as image-text pairs, which provide flat (*i.e.*, no hierarchical structure, all data points are on the same level) and sparse signals. However, real-world data

exhibits a hierarchical semantic structure, containing numerous local communities. The instances within each community have similar semantics, while the semantic differences across communities are substantial. The absence of hierarchical-information-mining leads to insufficient exploration of community relationships, hindering the learning of generalizable hash codes. Furthermore, this challenge is compounded by the ***ineffective alignment of heterogeneous modalities***. Existing methods (Hu et al., 2022; Zhang et al., 2024b; Tu et al., 2023) employ different encoders to project data from various modalities and optimize towards a common objective. Nevertheless, due to the inherent heterogeneity across modalities (*e.g.*, manifested in structure and semantics), direct alignment can pose a high learning difficulty and lead to suboptimal performance. Therefore, it is necessary to conduct hierarchical cross-modal learning and in a progressive manner.

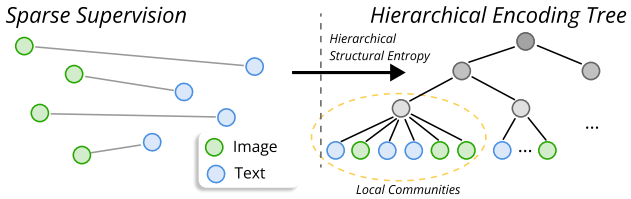


Figure 1: HINT transforms sparse cross-modal supervision (left) into a meaningful hierarchical encoding tree (right), which reveals local semantic communities for robust cross-modal alignment.

To address the aforementioned issues, we propose a novel approach called **Hi**erarchical Codi**n**g **T**ree with Modality Mixup (**HINT**) for hashing-based cross-modal retrieval. The core idea of HINT lies in constructing a cross-modal encoding tree that recovers hierarchical semantic structures and mines local semantic communities. Specifically, guided by the hierarchical structure entropy, we construct the encoding tree from the enhanced cross-modal relationship graph. The encoding tree has dense connections within local communities and sparse connections between communities. By utilizing the cross-modal encoding tree, we avoid the performance degradation caused by flat and sparse cross-modal connections. Next, we synthesize proxy samples in different modalities for each sample based on the encoding tree. Through curriculum-based mixup on these proxy samples, we achieve progressive modality alignment, circumventing the challenging task of directly aligning heterogeneous modalities. Furthermore, we achieve semantic alignment from a global perspective by optimizing the consistency of the semantic distributions of the proxy samples in different modalities. Extensive experiments on benchmark datasets demonstrate the superior performance of HINT.

The main contributions of this paper are: ❶ ***New Perspective.*** We connect the encoding tree with cross-modal hashing problems. Specifically, we construct a cross-modal encoding tree to explore the cross-modal relationships and uncover local semantic communities hierarchically. ❷ ***Coherent Framework.*** Based on the hierarchical encoding tree, we extract cross-modal proxy samples. Leveraging the proxy samples, we design a curriculum-based modality mixup mechanism for effective cross-modal hash learning. On the other hand, we achieve global-view consistency learning through the distribution alignment. ❸ ***Outstanding Performance.*** Comprehensive experiments demonstrate that HINT achieves state-of-the-art performance on benchmark datasets.

2 PRELIMINARY

In this work, we consider cross-modal hash retrieval problems. The objective is to map samples from both modalities into the shared Hamming space, enabling efficient cross-modal retrieval. Specifically, let the visual vector space be $\mathcal{D}^v = \{\boldsymbol{f}_i^v\}_{i=1}^N$ and the text vector space be $\mathcal{D}^t = \{\boldsymbol{f}_i^t\}_{i=1}^N$, which is encoded by common visual and text encoders. We have N image-text pairs without label information. We employ neural networks $\phi^v(\cdot)$ and $\phi^t(\cdot)$ to map each visual and text feature vector $\boldsymbol{f}_i^v$ and $\boldsymbol{f}_i^t$ into the Hamming space as:

$$\boldsymbol{b}_i^v = \operatorname{sign}\left(\phi^v\left(\boldsymbol{f}_i^v\right)\right), \quad \boldsymbol{b}_i^t = \operatorname{sign}\left(\phi^t\left(\boldsymbol{f}_i^t\right)\right), \tag{1}$$

where $\boldsymbol{b}_i^v$ and $\boldsymbol{b}_i^t$ are L-length hash codes, *i.e.*, $\boldsymbol{b}_i^* \in \{-1, +1\}^L$, $* \in \{v, t\}$, and $\operatorname{sign}(\cdot)$ is the sign function. The hash codes could be used for subsequent efficient retrieval. Therefore, we need to minimize the Hamming distances between semantically similar samples across modalities while maximizing the distances between dissimilar ones. The Hamming distance is calculated as $d(\boldsymbol{b}_i^*, \boldsymbol{b}_j^*) = \frac{1}{2}(L - \langle \boldsymbol{b}_i^*, \boldsymbol{b}_j^* \rangle)$. where L is the code length and $\langle \cdot, \cdot \rangle$ denotes the inner product.

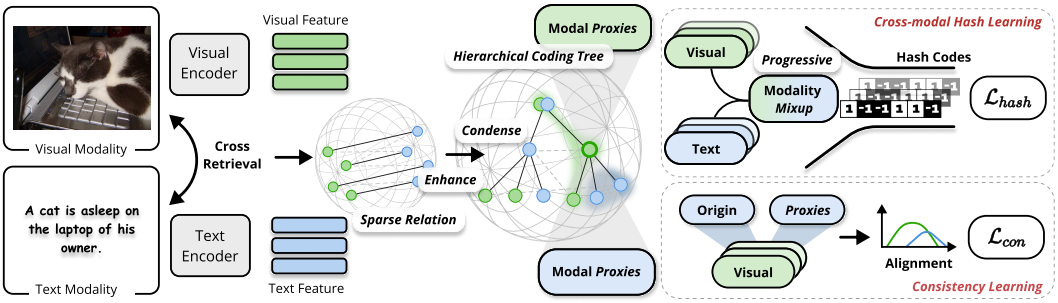


Figure 2: Overview of HINT, which constructs a hierarchical encoding tree from sparse cross-modal relationships. It then synthesizes modality proxies and performs progressive modality mixup and global-view consistency learning.

3 METHODOLOGY

3.1 FRAMEWORK OVERVIEW

Sparse cross-modal connections in unsupervised scenarios pose challenges for modality alignment and cross-modal retrieval. The core idea of HINT is to establish a cross-modal encoding tree to recover the hierarchical structure across modalities and enhance the connections among local semantic clusters. Specifically, as illustrated in Figure 2, our method comprises three main components: ❶ *Hierarchical encoding tree construction.* Guided by the hierarchical structural entropy, we optimize the enhanced cross-modal relationship graph to obtain the encoding tree. ❷ *Cross-modal hash learning with modality mixup.* To bridge the heterogeneous gap between modalities, we construct proxy samples for different modalities and progressively align them through a curriculum-based modality mixup mechanism. ❸ *Proxy-based consistency learning.* We optimize the distribution of cross-modal proxy samples, achieving a global-level alignment.

Key novelty of HINT: ❶ introducing the first hierarchical encoding tree for unsupervised cross-modal hashing that enables adaptive semantic partitioning, ❷ designing a curriculum-based modality mixup strategy that progressively bridges the heterogeneous gap, and ❸ unifying these components in a coherent framework that achieves effective cross-modal alignment through curriculum learning.

3.2 HIERARCHICAL CODING TREE CONSTRUCTION

To address the challenges posed by flat and sparse cross-modal connections, we construct a hierarchical encoding tree in an *enhance*-and-*condense* manner, as shown in Figure 3. First, we enhance the intra-modal connections within the relation graph. Then, guided by the structure entropy (Li et al., 2018; Zou et al., 2023), we condense the relation graph to obtain the hierarchical encoding tree. The encoding tree exhibits a hierarchical community structure, facilitating the hash learning.

Enhance. In unsupervised cross-modal retrieval, we primarily rely on cross-modal pairwise relations. We first construct a inter-modal relation graph $\mathcal{G}_{inter} = \{\mathcal{V}, \mathcal{E}_{inter}\}$, where $\mathcal{V} = \mathcal{D}^v \cup \mathcal{D}^t$ and $\mathcal{E}_{inter} = \{\boldsymbol{f}_i^v, \boldsymbol{f}_i^t\}_{i=1}^N$. Since cross-modal pairs only provide sparse supervision signals in unsupervised scenarios, we strategically employ KNN to enhance intra-modal relationships. This enables us to capture fine-grained local similarities and form cohesive bottom-level semantic communities, which serve as a robust foundation for subsequent hierarchical modeling. Since $\mathcal{G}_{inter}$ is sparse and inadequate for cross-modal learning, we enrich the intra-modal relationships to construct tightly-knit low-level communities. Specifically, we turn to the cosine similarity within modality by $S^*_{(i,j)} = \cos(\boldsymbol{f}_i^*, \boldsymbol{f}_j^*), * \in \{v, t\}$. We choose cosine similarity as it focuses on semantic directional alignment by normalizing vector magnitudes, which is crucial for cross-modal feature comparison. We then construct the intra-modal relationship graph $\mathcal{G}_{intra} = \{\mathcal{V}, \mathcal{E}_{intra}\}$ based on the similarity matrix by KNN manner:

$$\mathcal{E}_{intra} = \left\{ \left\{ \boldsymbol{f}_i^*, \boldsymbol{f}_j^* \right\} \mid j \in \text{topk}\left(\boldsymbol{f}_i^*, S, k\right) \right\}_{i=1}^N , \tag{2}$$

where k is set to 3 according to the hyperparameter study in Section 4.2. Then, we merge the intra-modal relationships $\mathcal{G}_{intra}$ and the inter-modal relationships $\mathcal{G}_{inter}$ to obtain the cross-modal relationship graph $\mathcal{G}_{cross} = \{\mathcal{V}, \mathcal{E}_{intra} \cup \mathcal{E}_{inter}\}$.

Condense. To simplify the cross-modal semantics and construct the hierarchical relationship, we extract the encoding tree $\mathcal{T}$ from $\mathcal{G}_{cross}$ with the guidance of structural entropy, as illustrated in the center part of Figure 2. We first introduce the 1-D structural entropy as:

$$E^1(\mathcal{G}) = -\sum_{\boldsymbol{v}\in V} \frac{\boldsymbol{d}_v}{vol\,(\mathcal{G})} \log \frac{\boldsymbol{d}_v}{vol\,(\mathcal{G})}\,, \quad (3)$$

where $\boldsymbol{d}_v$ denotes the degree for node v and $vol(\cdot)$ denotes the sum of degree for nodes in $\mathcal{G}$. Then, we introduce the hierarchical structural entropy for the $\mathcal{T}$ as:

$$E^{\mathcal{T}}(\mathcal{G}) = -\sum_{\alpha\in\mathcal{T}} \underbrace{\frac{g_\alpha}{vol\,(\mathcal{G})}}_{\text{information leakage}} \underbrace{\log \frac{\mathcal{V}_\alpha}{\mathcal{V}_{\alpha^-}}}_{\text{encoding efficiency}}. \quad (4)$$

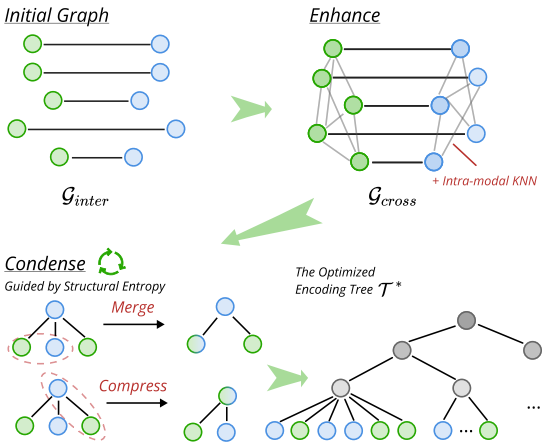


Figure 3: The construction pipeline of the encoding tree, involving graph enhancement and entropy-guided condensing operations.

Intuitively, the factor $\frac{g_\alpha}{vol(\mathcal{G})}$ captures the information leakage of community α, while the logarithmic term $\log \frac{\mathcal{V}_\alpha}{\mathcal{V}_{\alpha^-}}$ reflects the encoding efficiency of the hierarchy. In Eq. 4, $\mathcal{G}$ is $\mathcal{G}_{cross}$, α is a node in $\mathcal{T}$, $\mathcal{T}_\alpha$ is the subtree rooted at α, $\mathcal{T}_{\alpha^-}$ is the subtree rooted at α's parent node, g_α is the number of intra-modal relation links originating from the subtree $\mathcal{T}_\alpha$, and $\mathcal{V}_\alpha$ is the sum of degrees in $\mathcal{T}_\alpha$, and $\mathcal{V}_{\alpha^-}$ is the sum of degrees in $\mathcal{T}_\alpha$ excluding α. Eq. 4 generalizes Eq. 3 to hierarchical communities and reduces to Eq. 3 when the hierarchy collapses to a flat partition.

Guided by structural entropy, we convert the cross-modal relation graph into an encoding tree through *Merge* and *Compress*, as shown in Figure 3. Firstly, we perform node *Merge* operation to generate a binary encoding tree. The node *Merge* operation merges nodes that belong to the same parent node. These nodes may be highly semantically similar, and merging them can reduce the overall structural entropy of the cross-modal encoding tree. The node *Merge* operation is defined as: $\mathcal{T}' = Merge_{\mathcal{T}}(\alpha, \beta)$, we check all nodes if it can introduce decrease in $E^{\mathcal{T}}(\mathcal{G})$.

Secondly, *Compress* operation is performed to optimize the encoding tree, mainly targeting adjacent nodes at different levels, and constructing local clusters. This is achieved by attempting to move the child encoding tree with α as the root to its parent node's parent node, thereby enabling compression of the cross-modal semantic graph. After encoding tree compression, if there are no child encoding trees connected to a parent node, this parent node can be contracted to its parent node. The *Compress* operation is defined as: $\mathcal{T}' = Compress_{\mathcal{T}}(\alpha, \beta)$. Similarly, we check the nodes and conduct the *Compress* operation if it can introduce entropy decrease. Overall, we optimize the cross-modal encoding tree following a greedy principle. Specifically, we traverse the tree nodes in a breadth-first search manner. We attempt the aforementioned operations, and if they can decrease the structural entropy, we execute the operation. The cross-modal encoding tree after optimization is defined as:

$$\mathcal{T}^* = \arg\min\left(E^{\mathcal{T}}(\mathcal{G})\right)\,, \quad (5)$$

where $\mathcal{T}^*$ is the optimized encoding tree. The resulting tree naturally captures semantic granularity transitions, with upper nodes representing broad categories (e.g., "animals"), mid-level nodes capturing fine-grained concepts (e.g., "dogs", "cats"), and leaf nodes corresponding to specific instances. $\mathcal{T}^*$ exhibits better cross-modal semantic properties. Specifically, they encompass more comprehensive local semantic motifs while mitigating the connections within high-density communities. These characteristics facilitate subsequent discriminative hash code learning and consistency learning.

3.3 Structure-guided Cross-modal Hash Learning

After obtaining the optimized cross-modal encoding tree $\mathcal{T}^*$, we use it for unsupervised hash learning. Compared to existing unsupervised cross-modal hashing works (Hu et al., 2022; Liu et al., 2017; Zhang et al., 2018; 2024b; Tu et al., 2023), our method can exploit local semantic communities, avoiding the bias caused by individual samples. Simultaneously, we jointly model both image and text modalities, mapping the vectors from different modalities into a unified Hamming space.

Proxy Construction. For each sample $\boldsymbol{f}_i^*$, we sample its neighboring nodes $\mathcal{N}^+(\boldsymbol{f}_i^*)$ with the same modality and $\mathcal{N}^-(\boldsymbol{f}_i^*)$ with the opposite modality on the cross-modal encoding tree $\mathcal{T}^*$, and obtain the *proxy* samples via:

$$\boldsymbol{m}_i^{same} = \frac{1}{|\mathcal{N}^+(\boldsymbol{f}_i^*)|} \sum_{j \in \mathcal{N}^+(\boldsymbol{f}_i^*)} \phi^*\left(\boldsymbol{f}_j^*\right),$$
$$\boldsymbol{m}_i^{cross} = \frac{1}{|\mathcal{N}^-(\boldsymbol{f}_i^*)|} \sum_{j \in \mathcal{N}^-(\boldsymbol{f}_i^*)} \phi^*\left(\boldsymbol{f}_j^*\right), \tag{6}$$

where $\boldsymbol{f}_i^*, * \in \{v, t\}$ is the vector, $\phi^*(\cdot)$ is the hash model we introduced in Eq. 1. Therefore, $\boldsymbol{m}_i^{same}$ for a text sample $\boldsymbol{f}_i^t$ (aggregating text neighbors) is distinct from $\boldsymbol{m}_i^{same}$ for a visual sample $\boldsymbol{f}_i^v$ (aggregating visual neighbors), as they are derived from different sets of modality-specific neighbors. A similar distinction applies to $\boldsymbol{m}_i^{cross}$. The modal *proxy* samples can be viewed as a mediator between the two modalities, consisting of semantically similar nodes from the opposite modality, exhibiting better semantic robustness.

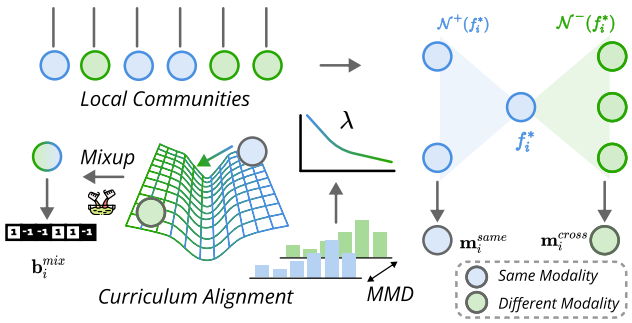


Figure 4: The Modality Mixup pipeline: generating modality-specific proxies, using their MMD to guide a curriculum-based mixup, and producing $\boldsymbol{b}_i^{mix}$.

Modality Mixup. Leveraging *proxy* samples, we introduce a mixup mechanism for progressively learning. Specifically, as Figure 4, $\boldsymbol{m}_i^{same}$ and $\boldsymbol{m}_i^{cross}$ are mixup to generate the hash codes:

$$\boldsymbol{b}_i^{mix} = \operatorname{sign}\left(\frac{\lambda}{1+\lambda}\boldsymbol{m}_i^{same} + \frac{1}{1+\lambda}\boldsymbol{m}_i^{cross}\right),$$
$$\lambda = \widehat{\mathrm{MMD}}\left(\rho\left(\boldsymbol{m}^{same}, \mathcal{B}\right), \rho\left(\boldsymbol{m}^{cross}, \mathcal{B}\right)\right), \tag{7}$$

where λ is to measure the distribution difference of different modalities, and ρ is the cosine distance metrics, we sample in mini-batch $\mathcal{B}$ and calculate $\widehat{\mathrm{MMD}}$ using the Gaussian kernel (Long et al., 2015; Tolstikhin et al., 2016). The equation shows the modal alignment process with a progressive modality mixup. As shown in Figure 5, initially, there exists a large modal discrepancy, *i.e.*, larger λ indicates $\boldsymbol{b}_i^{mix}$ predominantly leverages the same-modal features $\boldsymbol{m}_i^{same}$. As the modality alignment progresses, the cross-modal features $\boldsymbol{m}_i^{cross}$ become more prominent.

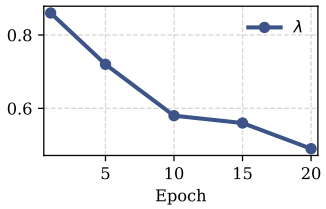


Figure 5: Evolution of λ during learning on MIRFlickr-25K.

Cross-modal Hash Learning. We employ the mixed hash code $\boldsymbol{b}_i^{mix}$ for hash learning. In the learning process, we sample in batches, and the other samples in the batch can serve as negative samples, enhancing the discriminative power of the hash codes. The objective function for cross-modal hash learning is:

$$\mathcal{L}_{hash} = -\sum_{i=1}^{N}\left(\log \frac{\exp\left(\langle \boldsymbol{f}_i^*, \boldsymbol{b}_i^{mix}\rangle/\tau\right)}{\sum_{j=1}^{|\mathcal{B}|}\exp\left(\langle \boldsymbol{f}_i^*, \boldsymbol{b}_j^{mix}\rangle/\tau\right)}\right), \tag{8}$$

where $* \in \{v, t\}$ and τ is the temperature parameter, which is set to 0.3 according to Section 4.2.

In HINT, we utilize the cross-modal encoding tree to guide hash learning. Since our cross-modal encoding tree has more comprehensive connections on local base groups, it can help align hash codes to more robust semantics. Meanwhile, the cross-modal encoding tree is sparser between communities, facilitating discriminability between groups and achieving discriminative hash codes.

Theoretical Discussion. This paragraph will discuss the limiting behavior of the cross-modal hashing loss $\mathcal{L}_{hash}$ and demonstrate how it enables effective retrieval. In Eq. 8, the inner product $\langle \boldsymbol{f}_i^*, \boldsymbol{b}_i^{mix}\rangle$ serves as a similarity measure between the feature $\boldsymbol{f}_i^*$ and the hash code $\boldsymbol{b}_i^{mix}$, and the temperature τ controls the scale of the measured similarity.Since τ is relatively small in implementation, the following theorem shows that $\mathcal{L}_{hash}$ converges to the triplet loss with zero margin.

Theorem 1 (Limiting behavior of $\mathcal{L}_{hash}$)**.** *For sufficiently large N and batch size $|\mathcal{B}|$, the $\mathcal{L}_{hash}$ converges to the triplet loss with zero-margin, that is*

$$\lim_{\tau \to 0^+} \frac{1}{N}\mathcal{L}_{hash} \simeq \mathbb{E}[\|\boldsymbol{f}_i^v - \boldsymbol{b}_i^{mix}\|_2^2 + \|\boldsymbol{f}_i^t - \boldsymbol{b}_i^{mix}\|_2^2 - \min_{k \neq i}\|\boldsymbol{f}_i^v - \boldsymbol{b}_k^{mix}\|_2^2 - \min_{k \neq i}\|\boldsymbol{f}_i^t - \boldsymbol{b}_k^{mix}\|_2^2]. \tag{9}$$

The right-hand side of Eq. 9, which is also called the *alignment* of the model, evaluates the difference between the distances of positive pairs compared with the hardest negative pairs that are closest to the positive pair. Theorem 1 implies that minimizing $\mathcal{L}_{hash}$ is equivalent to minimizing the triplet loss, and the smaller triplet loss implies more transferable representation and more effective retrieval. This theorem shows how $\mathcal{L}_{hash}$ contributes to the cross-modal retrieval performance. The detailed proof is available in the Appendix B.

3.4 Proxy-based Consistency Learning

Due to the heterogeneous gap between modalities, it is necessary to introduce an additional global-view modal alignment mechanism to achieve better alignment and enhance the generalization ability.

Semantic Consistency Learning. We leverage the modal *proxy* samples for modality alignment learning, assuming that the original samples and their *proxy* samples should have similar semantic positions and similar distribution. Specifically, the modality semantics in a global view can be represented as the similarity distribution between samples and other samples within the same batch:

$$p(\boldsymbol{f}_i^*) = \left[\rho\left(\boldsymbol{f}_i^*, \boldsymbol{f}_j^*\right) \mid \boldsymbol{f}_j^* \in \mathcal{B}^-\right], \tag{10}$$

where $\mathcal{B}^-$ includes the opposite modality instances within the same mini-batch, and $\rho(\cdot)$ is the cosine similarity function. Our objective is achieved by optimizing the KL divergence:

$$\mathcal{L}_{con} = \sum_{i=1}^{|\mathcal{B}|} \left(D_{KL}\left(p\left(\boldsymbol{f}_i^*\right) \parallel p\left(\boldsymbol{m}_i^{cross}\right)\right)\right), \tag{11}$$

where $|\mathcal{B}|$ is the batch size, $p\left(\boldsymbol{f}_i^*\right)$ and $p\left(\boldsymbol{m}_i^{cross}\right)$ are the semantic distributions of the i-th sample and its cross-modal *proxy* sample, respectively. By optimizing the consistency learning $\mathcal{L}_{con}$, we achieve modality alignment learning at a high level by leveraging the semantic-stable *proxy* samples.

Summary. Our method constructs a hierarchical encoding tree by unsupervised cross-modal mining and simultaneously leverages the encoding tree for cross-modal hash learning and semantic consistency learning. During the testing phase, the hierarchical encoding tree and proxy samples are not used. We directly employ the corresponding hash model to generate its hash code. This design ensures efficient retrieval with minimal computational overhead during inference time. Due to the non-differentiability of the $\mathrm{sign}(\cdot)$ function, it is challenging to optimize the overall objective. Therefore, we adopt $\tanh(\cdot)$ as a surrogate during the training process. The whole algorithm is summarized in Algorithm 1 and Appendix A. The computational complexity and time efficiency are discussed in the Appendix C.

4 Experiment

4.1 Experimental Settings

Datasets. To comprehensively evaluate the performance of HINT, we conduct experiments on three widely used public datasets: MIRFlickr-25K (Huiskes & Lew, 2008), NUS-WIDE (Rasiwasia et al., 2010), and MS-COCO (Lin et al., 2014). The detailed information is available in Appendix E.2.

Baselines. We compare our method HINT with 13 baselines from related fields, including the latest state-of-the-art works. Details in Appendix E.1.

Implementation Details. To ensure a fair comparison, we implement our method based on the latest SOTA works (Zhang et al., 2024b; Hu et al., 2022). Details in Appendix E.3.

4.2 Results

Hamming Ranking. The experiments on cross-modal retrieval benchmarks demonstrate that the proposed method consistently outperforms baseline approaches across different code lengths (16-128 bits). Key findings show that: ❶ Deep unsupervised hashing methods generally perform better than traditional approaches, ❷ Supervised methods struggle when labeled data is limited, ❸ The method shows improved performance on challenging sub-tasks like Text→Image retrieval, and ❹ The hierarchical modeling approach proves more effective than other deep cross-modal methods for

Table 1: Comparison of MAP performance (%) across various cross-modal hashing methods.

Methods	MIRFlickr-25K				NUS-WIDE				MS-COCO			
	16	32	64	128	16	32	64	128	16	32	64	128
Image → Text												
CVH	62.0	60.8	59.4	58.3	48.7	49.5	45.6	41.9	50.3	50.4	47.1	42.5
LSSH	59.7	60.9	60.6	60.5	44.2	45.7	45.0	45.1	48.4	52.5	54.2	55.1
CMFH	55.7	55.7	55.6	55.7	33.9	33.8	34.3	33.9	36.6	36.9	37.0	36.5
FSH	58.1	61.2	63.5	66.2	55.7	56.5	59.8	63.5	53.9	54.9	57.6	58.7
MTFH	50.7	51.2	55.8	55.4	29.7	29.7	27.2	32.8	39.9	29.3	29.5	39.5
FOMH	57.5	64.0	69.1	65.9	30.5	30.5	30.6	31.4	37.8	51.4	57.1	60.1
DCH	59.6	60.2	62.6	63.6	39.2	42.2	43.0	43.6	42.2	42.0	44.6	46.8
DGCPN	65.1	68.3	71.8	72.4	60.1	61.8	63.1	64.0	55.6	56.9	57.8	58.0
UCHSTM	70.1	71.5	72.4	72.3	62.5	63.5	64.6	64.4	55.8	57.2	57.6	57.3
UCCH	71.6	72.6	72.8	73.2	62.1	62.3	64.0	64.5	56.0	56.2	56.6	57.4
UDDH	71.4	72.9	74.0	74.6	63.7	64.2	65.1	65.9	56.8	57.8	59.0	59.9
HuggingHash+	71.6	73.2	74.3	74.5	63.9	64.8	65.6	66.4	57.1	58.3	59.4	60.5
DEMO	71.8	73.3	73.4	74.3	64.6	64.8	66.2	66.4	57.5	57.8	58.6	60.5
HINT	**72.9**	**74.4**	**75.1**	**75.5**	**65.1**	**65.5**	**66.5**	**67.3**	**58.5**	**59.5**	**60.4**	**61.1**
Text → Image												
CVH	62.9	61.5	59.9	58.7	47.0	47.5	44.4	41.2	50.6	50.8	48.6	42.9
LSSH	60.2	59.8	59.8	59.7	47.3	48.2	47.1	45.7	49.0	52.2	54.7	56.0
CMFH	55.3	55.3	55.3	55.3	30.6	30.6	30.6	30.6	34.6	34.6	34.6	34.6
FSH	57.6	60.7	63.5	66.0	56.9	60.4	65.1	66.6	53.7	52.4	56.4	57.3
MTFH	51.4	52.4	51.8	58.1	35.3	31.4	39.9	41.0	33.5	37.4	30.0	33.4
FOMH	58.5	64.8	71.9	68.8	30.2	30.4	30.0	30.6	36.8	48.4	55.9	59.5
DCH	61.2	62.3	65.3	66.5	37.9	43.2	44.4	45.9	42.1	42.8	45.4	47.1
DGCPN	65.3	68.2	71.2	71.5	60.5	62.6	63.7	64.4	55.0	56.6	57.8	57.7
UCHSTM	69.5	71.1	71.3	72.3	63.2	64.3	65.1	65.2	55.5	56.7	57.8	57.3
UCCH	70.3	71.2	72.0	72.1	62.5	63.7	65.0	65.2	56.4	57.3	57.2	58.1
UDDH	70.5	71.6	72.8	73.5	64.5	65.2	66.0	66.6	56.6	57.5	58.5	59.4
HuggingHash+	70.7	72.0	73.2	73.8	64.8	65.7	66.5	67.0	56.9	57.9	59.0	60.1
DEMO	70.8	71.9	72.2	72.8	65.4	65.5	66.9	67.1	57.2	57.9	58.3	59.7
HINT	**72.0**	**73.1**	**74.0**	**74.6**	**66.0**	**66.6**	**67.3**	**67.8**	**58.2**	**59.0**	**59.8**	**60.8**

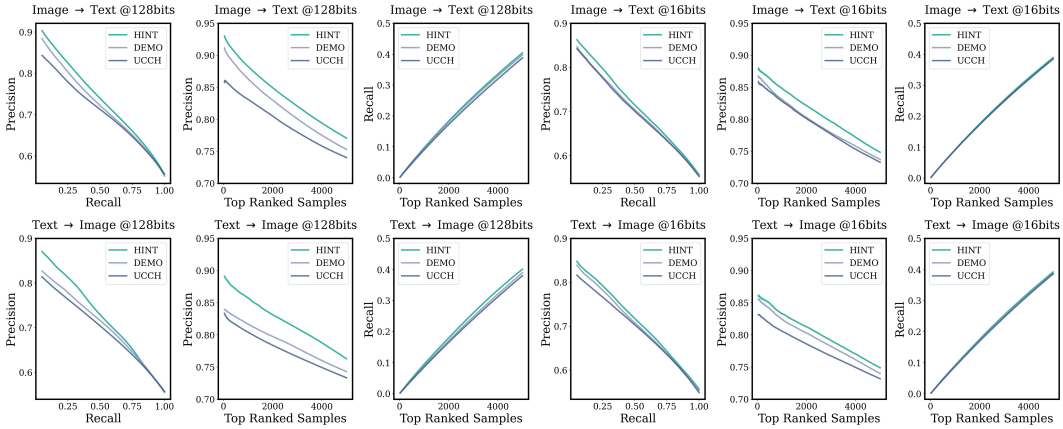


Figure 6: Hash lookup performance with 128-bit and 16-bit codes on the MIRFlickr-25K dataset. The Precision-Recall curves, Precision-N curves, and Recall-N curves are shown from left to right.

achieving progressive cross-modal alignment. Additionally, HINT demonstrates strong robustness against noisy data, maintaining superior performance even with 10% corrupted pairs. Detailed noise robustness analysis and results are provided in Appendix C.6. As shown in Table 1, performance generally improves with increasing bit length as longer hash codes provide larger Hamming space for encoding more information, though with diminishing returns at higher lengths.

Hash Lookup. To comprehensively analyze HINT's performance, we evaluate Precision-Recall, Precision-N, and Recall-N curves with 128-bit and 64-bit codes on MirFlickr-25K. As shown in Figure 6, HINT consistently outperforms baselines across all metrics, aligning with the MAP scores

Table 2: Ablation studies. The component columns "KNN, Tree, Curr, Con" respectively denote intra-modal KNN, hierarchical encoding tree, curriculum-based mixup, and proxy-based learning.

Methods	Components				MIRF-25K		NUS-WIDE		MS-COCO	
	KNN	Tree	Curr	Con	I→T	T→I	I→T	T→I	I→T	T→I
HINT *V1*					73.2	72.0	64.0	65.1	57.9	58.5
HINT *V2*	✓				73.8	72.8	65.2	65.7	59.1	58.9
HINT *V3*	✓	✓			74.2	73.6	65.9	66.4	60.0	59.5
HINT *V4*	✓	✓	✓		75.1	74.1	67.0	67.3	60.7	60.2
HINT	✓	✓	✓	✓	**75.5**	**74.6**	**67.3**	**67.8**	**61.1**	**60.8**

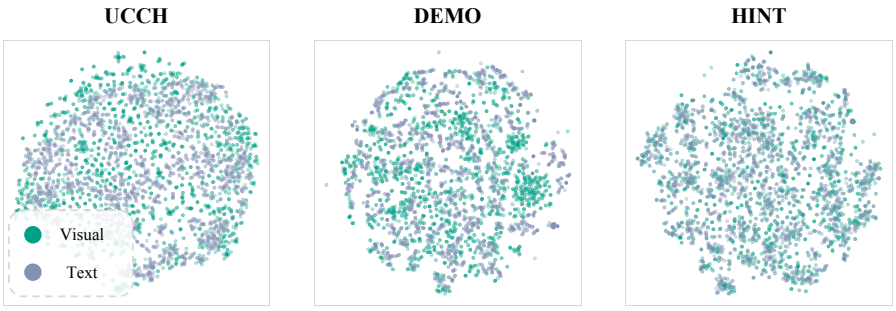


Figure 7: The t-SNE projection of hash codes from different modalities. Among competing methods, HINT shows the best ability of modal alignment.

from Hamming ranking. In summary, HINT exhibited optimal performance in cross-modal hash retrieval. Experiments for other code lengths is in Appendix C.1.

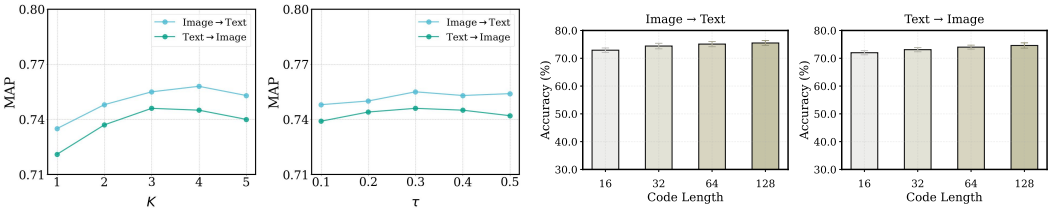
Visualization. We provide a t-SNE visualization analysis of HINT's performance using 128-bit hash codes on the MirFlickr-25K dataset, comparing with DEMO and UCCH, distinguishing different modalities with distinct colors. As shown in Figure 7, HINT demonstrates a superior ability to map representations from different modalities into a unified hash space, exhibiting higher alignment between text and visual modalities. The visualization suggests that HINT effectively aligns modalities and learns hash codes with generalization capabilities.

Ablation Study. We compare the following variants of HINT: *V1,* which only uses text-image pairs without consistency learning ($\mathcal{L}_{con}$), where $\boldsymbol{b}^*_{mix}$ is obtained solely from the opposite modality; *V2,* which uses both text-image pairs and intra-modal KNN without $\mathcal{L}_{con}$, where $\boldsymbol{b}^*_{mix}$ is obtained by averaging opposite modality and KNN samples; *V3,* which constructs the hierarchical encoding tree with sample selection but without curriculum-based progressive alignment and without consistency loss L_{con}; *V4,* which uses $\mathcal{L}_{hash}$ consistent with the full model but excludes $\mathcal{L}_{con}$. As shown in Table 2, the full model achieves optimal performance, with hierarchical encoding tree and progressive alignment (*V3* and *V4*) yielding the most improvements. We also explored iterative tree updates but found static construction provides better efficiency-performance trade-off, with detailed analysis in Appendix C.8. Additional experiments comparing different similarity metrics for tree construction demonstrate cosine similarity's superiority over L1/L2 distances, with detailed analysis in Appendix C.7. Additional ablation studies are available in Appendix C.2.

Sensitivity Analysis. We analyze the hyperparameters K and τ. As shown in Figure 8a, increasing K from 1 to 3 improves performance on both retrieval tasks, validating the benefits of enhanced cross-modal relationships. However, further increasing K to 5 introduces noisy relationships and decreases performance. Our experiments demonstrate HINT's remarkable stability, with performance fluctuation remaining within a narrow 2% margin when varying τ from 0.1 to 0.5. The model consistently outperforms baselines across most parameter settings. Based on these findings, we set $K = 3$ and $\tau = 0.3$ as the default values in our experiments.

Stability Analysis. We conducted 5 independent runs with different seeds. As shown in Figure 8b, the results show that HINT exhibits remarkable stability, with performance variations consistently remaining below 1% standard deviation across different code lengths. Details in Appendix C.9.

Time Efficiency. HINT maintains competitive efficiency despite its additional tree construction step, requiring only 3 minutes for tree building (5% of total training). It achieves better MAP scores than DEMO (75.5% vs 74.3%) with comparable training time. Details in Appendix C.5. Empirical speed



(a) Sensitive analysis of K and τ. HINT demonstrates robustness to hyperparameters.

(b) Stability analysis of HINT. Error bars indicate standard deviation.

Figure 8: Sensitivity and stability analyses on the MIRFlickr-25K dataset.

tests on MIRFlicker-25K demonstrate HINT's significant advantage in retrieval efficiency compared to dense vector approaches. Details in Appendix C.4.

5 Related Works

Cross-modal Retrieval is a fundamental task for bridging data of different modalities (Lee et al., 2024; Chen et al., 2024; Li et al., 2023; Krojer et al., 2022; Radford et al., 2021; Ge et al., 2024). Due to the diverse distributions and structures of texts and images, it is necessary to map them effectively into a unified representation space to calculate the semantic similarity between samples (Ding et al., 2016b; Liu et al., 2019b). With this unified representation, we can employ the approximate nearest neighbors (ANNs) (Zhu et al., 2023; Zhen et al., 2019) methods for similarity search. Recently, researchers turn to cross-modal hashing methods to enhance efficiency in terms of storage costs and large-scale retrieval processes (Xu et al., 2017; Jiang & Li, 2019; Wang et al., 2024b).

Unsupervised Cross-modal Hashing (Liang et al., 2024; Li et al., 2024a; Zhang et al., 2024b; Wang et al., 2024b) utilizes data correlation information to map cross-modal data into a unified Hamming space (Huang et al., 2024; Wu et al., 2018). Due to the expensive and difficult acquisition of labeled cross-modal data, supervised methods sometimes face challenges in the real-world (Hu et al., 2022). Therefore, unsupervised cross-modal hashing has attracted widespread attention (Zhou et al., 2014; Gao et al., 2023; Mikriukov et al., 2022). Researchers also use adversarial networks (Li et al., 2019) and contrastive learning (Hu et al., 2022) to handle cross-modal hash learning.

Advantages of HINT: These methods generally rely on sparse text-image relationships, lacking local community mining. We explore the hierarchical cross-modal relationship and learn more generalizable hash representations through modality mixup and cross-modal consistency learning.

Cross-modal Relationship Modeling is an essential topic in multi-modal research (Oh et al., 2024; Wang et al., 2024a; Li et al., 2024c; Liang et al., 2022; Huang et al., 2021). Some methods employ similarity modeling within modalities (Zhang et al., 2018), such as constructing graph structures for images and texts separately using the Wasserstein metric. The tree-based methods (Ge et al., 2021; Chen et al., 2022) are also introduced for cross-modal relationship modeling.

Advantages of HINT: Existing methods fail to recover the hierarchical cross-modal relationships effectively. Furthermore, they do not effectively align the heterogeneous gaps across modalities. In this paper, we combine cross-modal encoding trees and modality mixup to address these challenges.

6 Conclusion

This paper investigates the problem of efficient cross-modal retrieval through unsupervised cross-modal hashing. We propose HINT, a novel unsupervised cross-modal hashing method that leverages hierarchical structural entropy to guide the construction of a cross-modal encoding tree, which has tightly connected local clusters. By incorporating progressive mixup for proxy-based alignment and consistency learning from a global perspective, we enhance the generalization capability of the learned hash codes. Through comprehensive experiments on benchmark datasets, we demonstrate the effectiveness of HINT. HINT still has certain limitations. In real-world applications, there may be domain shifts or more modal data that need to be jointly retrieved, such as audio and video. In the future, we will explore extending HINT to more generalized scenarios. We will also investigate strategies such as partial labeling or active learning to further improve retrieval performance.

ETHICS STATEMENT

Our research adheres to the ICLR Code of Ethics. All datasets used in this study are publicly available. The code and related materials will be appropriately released to ensure transparency and reproducibility of our work.

REPRODUCIBILITY STATEMENT

For reproducibility purposes, we have made our code available at https://anonymous.4open.science/r/HINT. Also, we provided the detailed implementation details in Appendix E.3 and Appendix E.1.

REFERENCES

Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Qiang Yang. Transitive hashing network for heterogeneous multimedia retrieval. In *Proc. of Association for the Advancement of Artificial Intelligence*, volume 31, 2017.

Dapeng Chen, Min Wang, Haobin Chen, Lin Wu, Jing Qin, and Wei Peng. Cross-modal retrieval with heterogeneous graph embedding. In *Proc. of ACM International Conference on Multimedia*, pp. 3291–3300, 2022.

Haonan Chen, Zhicheng Dou, Kelong Mao, Jiongnan Liu, and Ziliang Zhao. Generalizing conversational dense retrieval via LLM-cognition data augmentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.

Kan Chen, Trung Bui, Chen Fang, Zhaowen Wang, and Ram Nevatia. Amc: Attention guided multi-modal correlation learning for image search. In *Proc. of Computer Vision and Pattern Recognition*, 2017.

Yu Chen, Lingfei Wu, and Mohammed Zaki. Iterative deep graph learning for graph neural networks: Better and robust node embeddings. *Advances in neural information processing systems*, 33:19314–19326, 2020.

Wanqing Cui, Keping Bi, Jiafeng Guo, and Xueqi Cheng. MORE: Multi-mOdal REtrieval augmented generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Guiguang Ding, Yuchen Guo, Jile Zhou, and Yue Gao. Large-scale cross-modality search via collective matrix factorization hashing. *IEEE Transactions on Image Processing*, 2016a.

Kun Ding, Bin Fan, Chunlei Huo, Shiming Xiang, and Chunhong Pan. Cross-modal hashing via rank-order preserving. *IEEE Transactions on Multimedia*, 2016b.

Zijun Gao, Jun Wang, Guoxian Yu, Zhongmin Yan, Carlotta Domeniconi, and Jinglin Zhang. Long-tail cross modal hashing. In *Proc. of Association for the Advancement of Artificial Intelligence*, 2023.

Xuri Ge, Fuhai Chen, Joemon M Jose, Zhilong Ji, Zhongqin Wu, and Xiao Liu. Structured multi-modal feature embedding and alignment for image-sentence retrieval. In *Proc. of ACM International Conference on Multimedia*, pp. 5185–5193, 2021.

Xuri Ge, Songpei Xu, Fuhai Chen, Jie Wang, Guoxin Wang, Shan An, and Joemon M Jose. 3shnet: Boosting image–sentence retrieval via visual semantic–spatial self-highlighting. *Information Processing & Management*, 61(4):103716, 2024.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

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*, 2022.

Hailang Huang, Zhijie Nie, Ziqiao Wang, and Ziyu Shang. Cross-modal and uni-modal soft-label alignment for image-text retrieval. In *Proc. of Association for the Advancement of Artificial Intelligence*, volume 38, pp. 18298–18306, 2024.

Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. *Proc. of Advances in Neural Information Processing Systems*, 2021.

Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *Proc. of ACM international conference on Multimedia information retrieval*, 2008.

Qing-Yuan Jiang and Wu-Jun Li. Discrete latent factor model for cross-modal hashing. *IEEE Transactions on Image Processing*, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Benno Krojer, Vaibhav Adlakha, Vibhav Vineet, Yash Goyal, Edoardo Ponti, and Siva Reddy. Image retrieval from contextual descriptions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland, May 2022. Association for Computational Linguistics.

Shaishav Kumar and Raghavendra Udupa. Learning hash functions for cross-view similarity search. In *Proc. of International Joint Conference on Artificial Intelligence*, 2011.

Saehyung Lee, Sangwon Yu, Junsung Park, Jihun Yi, and Sungroh Yoon. Interactive text-to-image retrieval with large language models: A plug-and-play approach. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.

Angsheng Li and Yicheng Pan. Structural information and dynamical complexity of networks. *IEEE Transactions on Information Theory*, 2016.

Angsheng Li, Xianchen Yin, Bingxiang Xu, Danyang Wang, Jimin Han, Yi Wei, Yun Deng, Ying Xiong, and Zhihua Zhang. Decoding topologically associating domains with ultra-low resolution hi-c data by graph structural entropy. *Nature communications*, 2018.

Chao Li, Cheng Deng, Lei Wang, De Xie, and Xianglong Liu. Coupled cyclegan: Unsupervised hashing network for cross-modal retrieval. In *Proc. of Association for the Advancement of Artificial Intelligence*, 2019.

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*, 2024a.

Wenyan Li, Jiaang Li, Rita Ramos, Raphael Tang, and Desmond Elliott. Understanding retrieval robustness for retrieval-augmented image captioning. *arXiv preprint arXiv:2406.02265*, 2024b.

Yunxin Li, Baotian Hu, Yuxin Ding, Lin Ma, and Min Zhang. A neural divide-and-conquer reasoning framework for image retrieval from linguistically complex text. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023.

Zhi Li, Yifan Liu, and Yin Zhang. Back-modality: Leveraging modal transformation for data augmentation. *Proc. of Advances in Neural Information Processing Systems*, 2024c.

Meiyu Liang, Junping Du, Zhengyang Liang, Yongwang Xing, Wei Huang, and Zhe Xue. Self-supervised multi-modal knowledge graph contrastive hashing for cross-modal search. In *Proc. of Association for the Advancement of Artificial Intelligence*, volume 38, pp. 13744–13753, 2024.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Proc. of Advances in Neural Information Processing Systems*, 2022.

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 *Proc. of European Conference on Computer Vision*, 2014.

Hong Liu, Rongrong Ji, Yongjian Wu, Feiyue Huang, and Baochang Zhang. Cross-modality binary code learning via fusion similarity hashing. In *Proc. of Computer Vision and Pattern Recognition*, 2017.

Wei Liu, Jun Wang, Rongrong Ji, Yu-Gang Jiang, and Shih-Fu Chang. Supervised hashing with kernels. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 2074–2081. IEEE, 2012.

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*, 2019a.

Xuanwu Liu, Guoxian Yu, Carlotta Domeniconi, Jun Wang, Yazhou Ren, and Maozu Guo. Ranking-based deep cross-modal hashing. In *Proc. of Association for the Advancement of Artificial Intelligence*, 2019b.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *Proc. of International Conference on Machine Learning*. PMLR, 2015.

Xu Lu, Lei Zhu, Zhiyong Cheng, Jingjing Li, Xiushan Nie, and Huaxiang Zhang. Flexible online multi-modal hashing for large-scale multimedia retrieval. In *Proc. of ACM International Conference on Multimedia*, 2019.

Xiao Luo, Haixin Wang, Daqing Wu, Chong Chen, Minghua Deng, Jianqiang Huang, and Xian-Sheng Hua. A survey on deep hashing methods. *ACM Transactions on Knowledge Discovery from Data*, 2023.

Zeyu Ma, Siwei Wang, Xiao Luo, Zhonghui Gu, Chong Chen, Jinxing Li, Xian-Sheng Hua, and Guangming Lu. Harr: Learning discriminative and high-quality hash codes for image retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(5):1–23, 2024.

Georgii Mikriukov, Mahdyar Ravanbakhsh, and Begüm Demir. Unsupervised contrastive hashing for cross-modal retrieval in remote sensing. In *Proc. of International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2022.

Changdae Oh, Junhyuk So, Hoyoon Byun, YongTaek Lim, Minchul Shin, Jong-June Jeon, and Kyungwoo Song. Geodesic multi-modal mixup for robust fine-tuning. *Proc. of Advances in Neural Information Processing Systems*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. of International Conference on Machine Learning*, 2021.

Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proc. of ACM International Conference on Multimedia*, 2010.

Fumin Shen, Chunhua Shen, Wei Liu, and Heng Tao Shen. Supervised discrete hashing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 37–45, 2015.

Xiaobo Shen, Yinfan Chen, Weiwei Liu, Yuhui Zheng, Quan-Sen Sun, and Shirui Pan. Graph convolutional multi-label hashing for cross-modal retrieval. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Mingyang Song, Liping Jing, and Yi Feng. Match more, extract better! hybrid matching model for open domain web keyphrase extraction. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 17–27, 2024.

Yuan Sun, Jian Dai, Zhenwen Ren, Yingke Chen, Dezhong Peng, and Peng Hu. Dual self-paced cross-modal hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15184–15192, 2024.

Ilya O Tolstikhin, Bharath K Sriperumbudur, and Bernhard Schölkopf. Minimax estimation of maximum mean discrepancy with radial kernels. *Proc. of Advances in Neural Information Processing Systems*, 2016.

Rong-Cheng Tu, Xian-Ling Mao, Qinghong Lin, Wenjin Ji, Weize Qin, Wei Wei, and Heyan Huang. Unsupervised cross-modal hashing via semantic text mining. *IEEE Transactions on Multimedia*, 2023.

Dongsheng Wang, Miaoge Li, Xinyang Liu, MingSheng Xu, Bo Chen, and Hanwang Zhang. Tuning multi-mode token-level prompt alignment across modalities. *Proc. of Advances in Neural Information Processing Systems*, 2024a.

Haixin Wang, Hao Wu, Jinan Sun, Shikun Zhang, Chong Chen, Xian-Sheng Hua, and Xiao Luo. Idea: An invariant perspective for efficient domain adaptive image retrieval. *Advances in Neural Information Processing Systems*, 36:57256–57275, 2023.

Jinpeng Wang, Ziyun Zeng, Bin Chen, Yuting Wang, Dongliang Liao, Gongfu Li, Yiru Wang, and Shu-Tao Xia. Hugs bring double benefits: Unsupervised cross-modal hashing with multi-granularity aligned transformers. *International Journal of Computer Vision*, pp. 1–33, 2024b.

Jun Wang, Sanjiv Kumar, and Shih-Fu Chang. Semi-supervised hashing for scalable image retrieval. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3424–3431. IEEE, 2010.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proc. of Computer Vision and Pattern Recognition*, 2018.

Xing Xu, Fumin Shen, Yang Yang, Heng Tao Shen, and Xuelong Li. Learning discriminative binary codes for large-scale cross-modal retrieval. *IEEE Transactions on Image Processing*, 2017.

Chenggang Yan, Biao Gong, Yuxuan Wei, and Yue Gao. Deep multi-view enhancement hashing for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Jun Yu, Hao Zhou, Yibing Zhan, and Dacheng Tao. Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing. In *Proc. of Association for the Advancement of Artificial Intelligence*, 2021.

Bin Zhang, Yue Zhang, Junyu Li, Jiazhou Chen, Tatsuya Akutsu, Yiu-Ming Cheung, and Hongmin Cai. Unsupervised dual deep hashing with semantic-index and content-code for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024a.

Fan Zhang, Xian-Sheng Hua, Chong Chen, and Xiao Luo. DEMO: A Statistical Perspective for Efficient Image-Text Matching. *arXiv preprint arXiv:2405.11496*, 2024b.

Jian Zhang, Yuxin Peng, and Mingkuan Yuan. Unsupervised generative adversarial cross-modal hashing. In *Proc. of Association for the Advancement of Artificial Intelligence*, 2018.

Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *Proc. of Computer Vision and Pattern Recognition*, 2019.

Jile Zhou, Guiguang Ding, and Yuchen Guo. Latent semantic sparse hashing for cross-modal similarity search. In *Proc. of ACM SIGIR conference on Research & development in information retrieval*, 2014.

Lei Zhu, Chaoqun Zheng, Weili Guan, Jingjing Li, Yang Yang, and Heng Tao Shen. Multi-modal hashing for efficient multimedia retrieval: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

Dongcheng Zou, Hao Peng, Xiang Huang, Renyu Yang, Jianxin Li, Jia Wu, Chunyang Liu, and Philip S Yu. Se-gsl: A general and effective graph structure learning framework through structural entropy optimization. In *Proc. of ACM Web Conference*, 2023.

A ALGORITHM

We present the optimization algorithm of our method in Algorithm 1. The algorithm first constructs a cross-modal relationship graph and optimizes the hierarchical encoding tree. Then during training, it performs modality mixup and consistency learning to generate effective hash codes through back-propagation.

Algorithm 1 Optimization Algorithm of HINT

Require: Visual modality $\mathcal{D}^v$; Text modality $\mathcal{D}^t$;
Ensure: Hashing model $\phi^v(\cdot)$ and $\phi^t(\cdot)$;

```
Construct the cross-modal relationship graph G_cross;
Optimize the hierarchical encoding tree T*;
for each epoch do
    for each batch do
        Sample B^v, B^t from D^v, D^t;
        Construct proxy samples m^same, m^cross using Eq. 6;
        Perform modality mixup and generate hash code b^mix with T* by Eq. 7;
        Calculate the L_hash and L_con;
        Update parameters through back-propagation;
    end for
end for
```

B PROOF OF LIMITING BEHAVIOR OF CROSS-MODAL HASH LOSS

We mainly discuss the limiting behavior of the cross-modal hashing loss $\mathcal{L}_{hash}$ in Equation 8. We seek to prove how it enables effective retrieval capabilities.

In Equation 8, $\langle \boldsymbol{f}_i^*, \boldsymbol{b}_i^{mix} \rangle$ serves as a similarity measure between the feature $\boldsymbol{f}_i^*$ and its corresponding hash code $\boldsymbol{b}_i^{mix}$. Since τ is relatively small, we have the following theorem, which shows that $\mathcal{L}_{hash}$ converges to the triplet loss with zero margin.

Theorem (Limiting behavior of $\mathcal{L}_{hash}$)**.** *For sufficiently large N and batch size $|\mathcal{B}|$, the $\mathcal{L}_{hash}$ converges to the triplet loss with zero-margin, that is*

$$\lim_{\tau \to 0^+} \frac{1}{N} \mathcal{L}_{hash} \simeq \mathbb{E}[\|\boldsymbol{f}_i^v - \boldsymbol{b}_i^{mix}\|_2^2 + \|\boldsymbol{f}_i^t - \boldsymbol{b}_i^{mix}\|_2^2 - \min_{k \neq i} \|\boldsymbol{f}_i^v - \boldsymbol{b}_k^{mix}\|_2^2 - \min_{k \neq i} \|\boldsymbol{f}_i^t - \boldsymbol{b}_k^{mix}\|_2^2]. \tag{12}$$

The right-hand side of Equation 12 is for *alignment* of the model, and evaluates the difference between the distances (or similarities) of positive pairs compared with the hardest negative pairs that are closest to the positive pair. Theorem B implies that minimizing $\mathcal{L}_{hash}$ is equivalent to minimizing the triplet loss (alignment), and smaller alignment implies more transferable representation and more efficient retrieval. It shows how $\mathcal{L}_{hash}$ contributes to the retrieval performance.

Proof of Theorem 1. Denote $\mathcal{N}_v$ and $\mathcal{N}_t$ denote the index sets of visual vector and text vector, correspondingly. We assume the batch number $|\mathcal{B}|$ is equal to N to reduce the notation burden and by definition,

$$\mathcal{L}_{hash} = -\sum_{i \in \mathcal{N}_v} \left(\log \frac{\exp\left(\langle \boldsymbol{f}_i^*, \boldsymbol{b}_i^{mix} \rangle / \tau\right)}{\sum_{j=1}^{|N|} \exp\left(\langle \boldsymbol{f}_i^*, \boldsymbol{b}_j^{mix} \rangle / \tau\right)} \right) - \sum_{i \in \mathcal{N}_t} \left(\log \frac{\exp\left(\langle \boldsymbol{f}_i^*, \boldsymbol{b}_i^{mix} \rangle / \tau\right)}{\sum_{j=1}^{|N|} \exp\left(\langle \boldsymbol{f}_i^*, \boldsymbol{b}_j^{mix} \rangle / \tau\right)} \right). \tag{13}$$

We focus on the first term of the right-hand side of the above equation. It is easily shown the derivation of the other term by changing the index,

$$
\begin{aligned}
&\lim_{\tau\to 0^+} -\frac{1}{N}\sum_{i\in\mathcal{N}_v}\left(\log\frac{\exp\left(\langle \boldsymbol{f}_i^*, \boldsymbol{b}_i^{mix}\rangle/\tau\right)}{\sum_{j=1}^{|N|}\exp\left(\langle \boldsymbol{f}_i^*, \boldsymbol{b}_j^{mix}\rangle/\tau\right)}\right)\\
&=\lim_{\tau\to 0^+}\frac{1}{N}\left[\sum_{i\in\mathcal{N}_v}-\frac{\langle \boldsymbol{f}_i^*, \boldsymbol{b}_i^{mix}\rangle}{\tau}\right]+\left[\sum_{i\in\mathcal{N}_v}\log\left\{\sum_{j=1}^{N}\exp\left(\langle \boldsymbol{f}_i^*, \boldsymbol{b}_j^{mix}\rangle/\tau\right)\right\}\right]\\
&=\lim_{\tau\to 0^+}\frac{1}{N}\left[\sum_{i\in\mathcal{N}_v}-\frac{\langle \boldsymbol{f}_i^*, \boldsymbol{b}_i^{mix}\rangle}{\tau}\right]+\left[\sum_{i\in\mathcal{N}_v}\log\left\{\exp\left(\langle \boldsymbol{f}_i^*, \boldsymbol{b}_i^{mix}\rangle/\tau\right)+\sum_{j\neq i}^{N}\exp\left(\langle \boldsymbol{f}_i^*, \boldsymbol{b}_j^{mix}\rangle/\tau\right)\right\}\right]\\
&=\lim_{\tau\to 0^+}\frac{1}{N}\sum_{i\in\mathcal{N}_v}\log\left\{1+\sum_{j\neq i}^{N}\exp\left(\langle \boldsymbol{f}_i^*, \boldsymbol{b}_j^{mix}-\boldsymbol{b}_i^{mix}\rangle/\tau\right)\right\}\\
&=\lim_{\tau\to 0^+}\frac{1}{N}\sum_{i\in\mathcal{N}_v}\frac{1}{\tau}\max\{\max_j\{\langle \boldsymbol{f}_i^*, \boldsymbol{b}_j^{mix}-\boldsymbol{b}_i^{mix}\rangle\},0\}.
\end{aligned}
\tag{14}
$$

Since $\boldsymbol{m}_i^{\text{same}}$ and $\boldsymbol{m}_i^{\text{cross}}$ are continuous random variables, then

$$
\mathbb{P}\Big(\|\boldsymbol{b}_i^{mix}\|^2=L\Big)=\mathbb{P}\Big(\Big\|\operatorname{sign}\Big(\frac{\lambda}{1+\lambda}\boldsymbol{m}_i^{\text{same}}+\frac{1}{1+\lambda}\boldsymbol{m}_i^{\text{cross}}\Big)\Big\|^2=L\Big)\quad=1. \tag{15}
$$

and

$$
\begin{aligned}
&\|\boldsymbol{f}_i^v-\boldsymbol{b}_i^{mix}\|_2^2-\min_{j\neq i}\|\boldsymbol{f}_i^v-\boldsymbol{b}_k^{mix}\|_2^2\\
=&\max_{j\neq i}\|\boldsymbol{f}_i^v\|_2^2+\|\boldsymbol{b}_i^{mix}\|_2^2-2\langle \boldsymbol{f}_i^v, \boldsymbol{b}_i^{mix}\rangle-\|\boldsymbol{f}_i^v\|_2^2-\|\boldsymbol{b}_j^{mix}\|_2^2+2\langle \boldsymbol{f}_i^v, \boldsymbol{b}_j^{mix}\rangle\\
=&2\max_{j\neq i}\langle \boldsymbol{f}_i^v, \boldsymbol{b}_j^{mix}-\boldsymbol{b}_i^{mix}\rangle,
\end{aligned}
\tag{16}
$$

which implies that

$$
\lim_{\tau\to 0^+} -\frac{1}{N}\sum_{i\in\mathcal{N}_v}\left(\log\frac{\exp\left(\langle \boldsymbol{f}_i^*, \boldsymbol{b}_i^{mix}\rangle/\tau\right)}{\sum_{j=1}^{|N|}\exp\left(\langle \boldsymbol{f}_i^*, \boldsymbol{b}_j^{mix}\rangle/\tau\right)}\right)\simeq\|\boldsymbol{f}_i^v-\boldsymbol{b}_i^{mix}\|_2^2-\min_{j\neq i}\|\boldsymbol{f}_i^v-\boldsymbol{b}_k^{mix}\|_2^2. \tag{17}
$$

□

C Additional Experimental Results

C.1 Additional Experiments for Hash Lookup

To comprehensively evaluate the performance of our HINT, we present precision-recall curves, precision-N curves, and recall-N curves on the MirFlickr-25K dataset with code lengths of 32 and 64 bits. As shown in Figure 10, 9, our method consistently outperforms other approaches in terms of precision and recall, which is consistent with the corresponding MAP score based on Hamming ranking. Additionally, we compute the precision and recall rates of the top-N retrieved results, demonstrating HINT's persistent advantage. Our method achieves superior performance in cross-modal hash retrieval.

C.2 Additional Ablation Experiments

To evaluate the effectiveness of these components, we introduce four variants:

- HINT V1, which only employs image-text pairs for cross-modal hash learning without the global-view consistency learning module;

Table 3: Additional ablation studies with different code lengths.

Methods	MIRFlickr-25K							
	Image→Text				Text→Image			
	16bit	32bit	64bit	128bit	16bit	32bit	64bit	128bit
HINT *V1*	71.5	72.4	72.6	73.2	70.5	71.1	71.7	72.0
HINT *V2*	71.8	72.8	73.0	73.7	70.8	71.4	72.2	72.8
HINT *V3*	72.0	73.4	73.8	74.2	71.5	72.4	73.2	73.6
HINT *V4*	72.6	73.9	74.8	75.2	71.5	72.8	73.9	74.3
Full Model	72.9	74.4	75.1	75.5	72.0	73.1	74.0	74.9

Table 4: Retrieval time cost (ms) varies with code length.

	16 Bit	32 Bit	48 Bit	64 Bit	96 Bit	128 Bit
Hash Code	16.7	18.0	19.4	19.9	21.8	22.2
Dense Vector	441.4	491.0	543.0	602.3	657.7	696.6
Speed Up	26.5×	27.2×	28.0×	30.2×	30.1×	31.4×

- HINT V2, which combines image-text pairs with intra-modal KNN for cross-modal hash learning, also without the consistency learning module;
- HINT V3, which uses the average of two modalities as the learning target and builds a hierarchical encoding tree with sample selection, but without curriculum-based progressive modal alignment (Equation 7), also without the global-view consistency learning module;
- HINT V4, which adopts the full model's Lhash but excludes consistency learning.

As shown in Table 3, our results demonstrate that our complete method achieves optimal performance, confirming the importance of each component. Furthermore, our ablation study reveals that hierarchical encoding trees and progressive alignment yield significant improvements (*V2* and *V3*), validating our motivation.

C.3 Computational Complexity

We mainly discuss the computational complexity of the additionally introduced encoding tree construction. Assuming the number of data points is N (including samples from different modalities), the time cost for calculating the similarity matrix is $O(N^2)$, the cost for constructing the KNN graph is $O(N)$, and the cost for optimizing the cross-modal encoding tree is $O(N \log^2 N)$ (Li & Pan, 2016). It is worth noting that the most time-consuming similarity matrix calculation can be accelerated by parallel computing. Moreover, the encoding tree construction is only performed once at the beginning of the training, so the additional computational complexity and time consumption brought to the overall training process are negligible.

C.4 Time Efficiency

We conducted a speed test between HINT and dense vector retrieval on an Intel Xeon CPU E5-2697 v4 (2.30GHz), as illustrated in Table 4. The speed test is conducted on the MIRFlicker-25K dataset with a retrieval database of 10^5 items. We perform 10^3 runs and report the average retrieval speed cost (ms). Hash methods excel in enabling efficient and scalable image retrieval, especially for large-scale datasets, due to fast Hamming distance computation. In contrast, existing pre-trained models only output dense vectors, resulting in slower computation. Table 4 compares the efficiency of hash codes and dense codes generated by our model and a pre-trained model at various bit lengths. The results clearly demonstrate that hash codes achieve significantly faster retrieval speeds than deep feature codes, confirming their superiority, particularly in large-scale image retrieval scenarios.

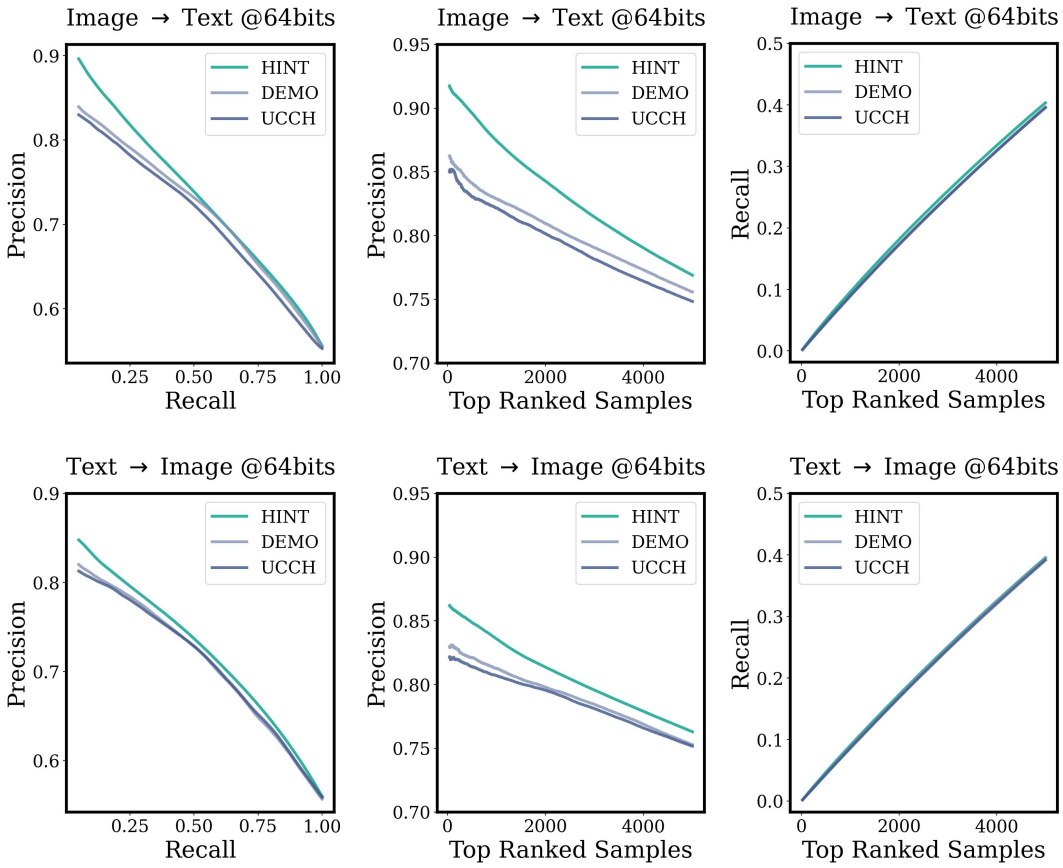


Figure 9: Hash lookup performance with 64 bits codes on the MIRFlickr-25K dataset. The precision-recall curves, precision-N curves, and recall-N curves are shown from left to right.

C.5 Computational Efficiency Analysis

While HINT introduces additional computational steps through the hierarchical encoding tree, we have implemented several optimizations to ensure practical efficiency:

- **Controlled Time Complexity**: The hierarchical encoding tree construction has a complexity of $O(N \log^2 N)$ and is executed only once during the initial training phase. Our experiments on MIRFlickr-25K show that tree construction takes <3 minutes (single GPU), representing <5% of total training time. Compared to existing methods like DEMO and UCCH, HINT does not significantly increase the overall training duration.

- **Memory-Efficient Design**: The encoding tree is stored using compressed relation triplets (parent-child-edge weight) instead of maintaining complete similarity matrices. Furthermore, the hash code generation phase is completely decoupled from the encoding tree, eliminating the need to load tree structures during inference and conserving deployment resources.

- **Practical Scalability**: The encoding tree's one-time construction and reusability make it particularly suitable for large-scale applications. This design choice significantly amortizes the initial computational investment across multiple training sessions.

These results demonstrate that HINT achieves superior performance while maintaining competitive training efficiency through its optimized design. The empirical training time and performance comparison with existing methods is shown in Table 5.

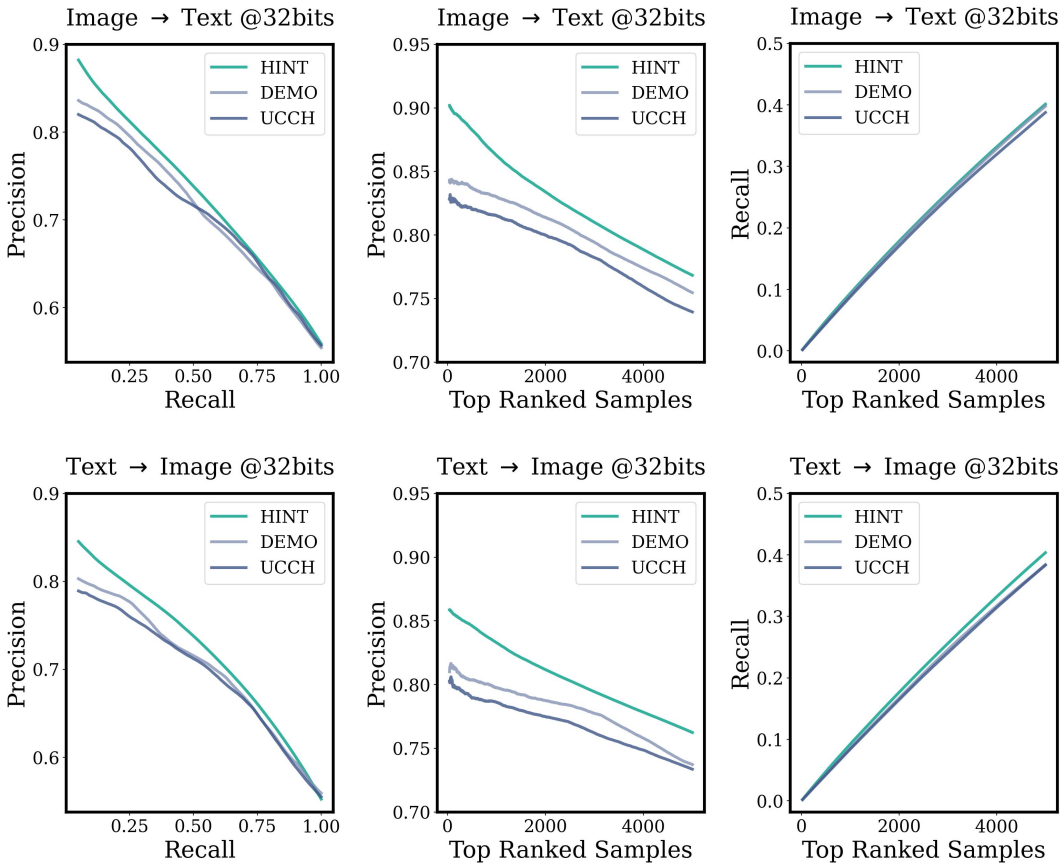


Figure 10: Hash lookup performance with 32 bits codes on the MIRFlickr-25K dataset. The precision-recall curves, precision-N curves, and recall-N curves are shown from left to right.

Table 5: Training time and performance comparison on MIRFlickr-25K.

Method	Training Time (h)	MAP (I→T,128bit)
UCCH	2.0	73.2
DEMO	2.5	74.3
HINT	2.1	75.5

C.6 Noise Robustness Analysis

To evaluate HINT's robustness against noisy data, we conducted experiments by randomly corrupting 10% of text-image pairs in the MIRFlickr-25K dataset. The results demonstrate HINT's superior noise resilience through both architectural design and experimental validation:

Architectural Robustness: HINT's hierarchical encoding tree provides two-level noise adaptation:

- The tree construction process inherently suppresses individual outliers by aggregating local semantic communities. Proxy samples, generated through neighbor feature averaging, effectively smooth out the impact of noisy samples within local communities.
- The cross-modal consistency learning module (L_{con}) constrains the influence of outliers on hash space mapping by enforcing semantic distribution alignment between proxy and original samples from a global perspective.

As shown in Table 6, HINT consistently maintains higher performance under noisy conditions, with minimal degradation compared to baseline methods. This demonstrates that the encoding tree's

Table 6: Noise robustness comparison on MIRFlickr-25K (128 bit code length) with 10% corrupted pairs.

Method	I→T	T→I	I→T($10\%n$)	T→I($10\%n$)
UCCH	73.2	73.2	70.7	72.6
DEMO	74.3	74.3	71.4	73.3
HINT	75.5	73.8	74.4	72.9

Table 7: Performance comparison of different similarity metrics on MIRFlickr-25K.

Metric	I→T	T→I
Cosine	75.5	74.6
L2	74.8	74.2
L1	73.5	72.8

hierarchical structure effectively identifies and mitigates the interference of mismatched pairs in cross-modal alignment.

C.7 Additional Analysis of Similarity Metrics

We conducted experiments comparing different similarity metrics for tree construction on MIRFlickr-25K. The results demonstrate that cosine similarity achieves optimal performance due to three key advantages:

- **Directional Consistency**: Cosine similarity focuses on semantic directional alignment by normalizing vector magnitudes
- **Loss Function Alignment**: The training objective relies on inner product similarity, which aligns with cosine similarity computation
- **Feature Space Compatibility**: Hamming distance is not suitable since features are not yet binarized during tree construction

The results show cosine similarity's 1.5-2.0% performance advantage over alternative metrics, confirming that feature directional alignment is more crucial than absolute distance for tree construction.

C.8 Iterative Tree Construction Analysis

Iterative structural optimization is a promising direction (Chen et al., 2020). Our experiments reveal that a static tree construction strategy achieves better balance between efficiency and performance:

- **Empirical Results**: We compared HINT with HINT-ITER (dynamic tree updates every 5 epochs) across multiple datasets:

The results show that HINT-ITER achieves comparable but slightly lower performance while requiring more training time. Two fundamental reasons explain this phenomenon:

- **Robust Initial Structure**: Our one-time tree construction leverages hierarchical structural entropy to recover semantically coherent communities, providing a stable foundation for proxy sample generation. Iterative refinement struggles to further improve this already optimized structure.
- **Stability-Aware Alignment**: The curriculum-based mixup mechanism and consistency learning rely on stable neighborhood relationships to progressively align modalities. Frequent tree updates disrupt this process, similar to how unstable negative samples degrade contrastive learning (He et al., 2020).

While our current approach suits existing tasks, we acknowledge potential benefits of dynamic structures for specific scenarios (e.g., evolving data streams), which we leave for future work.

Table 8: Performance comparison between static and iterative tree construction.

Method	MIRFlickr-25K		MS-COCO		Training
	I→T	T→I	I→T	T→I	Time (h)
HINT	75.5	74.6	61.1	60.8	2.1
HINT-ITER	75.0	74.7	60.5	60.4	2.9

Table 9: Stability analysis of HINT across different code lengths on MIRFlickr-25K dataset. Results show mean MAP scores ± standard deviation over five runs.

Task	32 bits	64 bits	96 bits	128 bits
I→T	72.9±0.81	74.4±0.99	75.1±0.92	75.5±0.88
T→I	72.0±0.75	73.1±0.75	74.0±0.72	74.6±0.96

C.9 Stability Analysis

To rigorously assess the stability of HINT, we conducted extensive experiments with five independent runs using different random seeds. Table 9 presents the mean performance and standard deviations across different code lengths for both Image-to-Text (I→T) and Text-to-Image (T→I) retrieval tasks on MIRFlickr-25K dataset.

The results demonstrate that HINT maintains consistent performance with remarkably low variance across different code lengths. The standard deviations consistently remain below 1% for both retrieval directions, indicating strong robustness to random initialization. This stability can be attributed to our hierarchical encoding tree structure and curriculum-based progressive alignment strategy, which provide reliable guidance for hash code learning regardless of initialization conditions.

D Additional Discussion

D.1 Rationale for Hierarchical Mining of Relationships

Hierarchical semantic structures are inherent in real-world data. Visual and textual content naturally form multi-level conceptual taxonomies. For instance, a general category like "Objects" can be decomposed into "Animals" and "Vehicles". "Animals" can be further subdivided into "Domestic" and "Wild", with "Domestic" containing specific instances like "Cats" and "Dogs" (e.g., "Maine Coon", "Siamese"). Similarly, "Vehicles" might branch into "Ground" and "Air" transport, with "Ground" including "Cars" and "Trucks".

Prevailing unsupervised cross-modal hashing methods often rely on flat representations of image-text pair relationships. Such flat structures exhibit limitations in capturing these intrinsic hierarchical dependencies. Specifically, they may:

- Treat the semantic dissimilarity between disparate pairs (e.g., "Cat-Dog" vs. "Cat-Car") undifferentiatedly, failing to recognize varying degrees of relatedness based on hierarchical proximity.
- Implicitly assume transitive relationships (i.e., if A is similar to B, and B is similar to C, then A is similar to C), an assumption that does not consistently hold for complex semantic relationships across different levels of abstraction.
- Struggle to ensure that instances within a sub-category (e.g., "Maine Coon" and "Siamese" under "Cats") are represented as being semantically closer to each other than to instances from distant categories (e.g., "Cars").

Our proposed HINT addresses these limitations through the Hierarchical Encoding Tree, which explicitly discovers and models inherent semantic hierarchies by optimizing structural entropy. This hierarchical approach offers several advantages:

- It facilitates a deeper exploration of semantic relationships that extend beyond direct, observed pairings, uncovering latent community structures.
- It enables the generation of more generalizable hash codes that are grounded in these discovered semantic communities, rather than isolated instances.
- It promotes smoother and more effective cross-modal alignment through the use of proxy samples derived from semantic neighborhoods and a curriculum learning strategy, thereby more effectively bridging the semantic gap between modalities.

By capturing these multi-level containment relationships, HINT can learn hash codes that better reflect real-world semantic structures, leading to improved retrieval performance. Current flat modeling approaches, by contrast, which treat "Cat-Dog" and "Cat-Car" similarity differences with equal weight, miss this crucial hierarchical semantic information, significantly impeding their ability to learn generalizable hash codes that align with complex real-world semantic organizations.

E Additional Implementation Details

E.1 Baseline Details

We evaluate our method against a range of state-of-the-art cross-modal hashing techniques: three supervised cross-modal hashing retrieval methods: MTFH (Liu et al., 2019a), FOMH (Lu et al., 2019), and DCH (Xu et al., 2017); four shallow unsupervised cross-modal hashing retrieval methods: CVH (Kumar & Udupa, 2011), LSSH (Zhou et al., 2014), CMFH (Ding et al., 2016a), and FSH (Liu et al., 2017); and four deep unsupervised cross-modal hashing retrieval methods: DGCPN (Yu et al., 2021), UCHSTM (Tu et al., 2023), UCCH (Hu et al., 2022), UDDH (Zhang et al., 2024a), HuggingHash+ (Wang et al., 2024b), and DEMO (Zhang et al., 2024b). Comparisons are conducted across different datasets, different cross-modal directions, and different hash code lengths.

The introduction of representative methods is as follows:

- ***CVH*** (Kumar & Udupa, 2011) introduced a relaxation technique, addressing the dimensionality reduction problem by leveraging techniques such as local sensitive hashing and canonical correlation analysis. This approach transforms the learning process into a manageable feature-based hashing problem.
- ***LSSH*** (Zhou et al., 2014) presented an effective iterative strategy, which explored the correlations between multi-modal representations and bridges the semantic gaps in the latent semantic space. By leveraging sparse coding to capture high-level salient structures in images, and matrix decomposition to extract latent concepts from texts, LSSH consolidated the heterogeneous modalities.
- ***CMFH*** (Ding et al., 2016a) exploited cross-modal decomposition to establish strong connections. It integrated linear embedding to preserve the Euclidean structure and a classifier-inspired loss function that leverages semantic label information.
- ***FSH*** (Liu et al., 2017) proposed a graph hashing architecture, constructing a unified graph to define the similarity between multi-modal instances. This framework alternated optimization to learn consistent binary codes and hash functions.
- ***MTFH*** (Liu et al., 2019a) employed an efficient objective function to jointly learn modality-specific hash codes with varying lengths while simultaneously learning semantic relevance matrices, thereby ensuring the comparability of heterogeneous data.
- ***FOMH*** (Lu et al., 2019) integrated a self-weighted and flexible multi-modal fusion strategy, enabling robust fusion even when missing modalities. Moreover, FOMH employed semantic supervision to learn shared hash codes.
- ***DCH*** (Xu et al., 2017) jointly optimized modality-specific hash functions and unified binary codes. Furthermore, it proposed an efficient optimization algorithm that iteratively obtained the optimized binary codes bit by bit.
- ***DGCPN*** (Yu et al., 2021) introduced a graph neighborhood approach to explore the relationships between data points and their neighbors through a graph neighborhood approach, thereby enhancing the accuracy of data similarity measurement. By leveraging semi-integer

and semi-binary optimization strategies, the gap between real-valued space and Hamming space was reduced in terms of value and similarity differences.

- ***UCHSTM*** (Tu et al., 2023) exploited the correlations between text data points, thereby constructing a modality-specific similarity matrix based on these correlations. Furthermore, it employed a custom-designed similarity loss to rectify any ill-defined similarities in the instance similarity matrix.
- ***UCCH*** (Hu et al., 2022) used contrastive learning, which enforced alignment between different modalities and unified binary representations, focusing on leveraging discriminative information from all pairs rather than just the hardest negative ones.
- ***UDDH*** (Zhang et al., 2024a) proposes a dual deep hashing architecture that combines semantic indexing with content codes for cross-modal retrieval. It employs deep hashing networks to extract features and jointly encode dual hashing codes, using K-means clustering for semantic indexing.
- ***HuggingHash+*** (Wang et al., 2024b) introduces a transformer-based multi-granularity learning framework for unsupervised cross-modal hashing. It constructs a fine-grained semantic space using aggregated local embeddings and incorporates an optimized quantization re-ranking strategy to enhance retrieval performance.
- ***DEMO*** (Zhang et al., 2024b) utilized multi-view augmentation to represent each image, followed by parameterized distribution divergence to ensure robust similarity structures. Meanwhile, it encouraged self-supervised consistency across retrieval distributions from different directions.

E.2 Dataset Details

We conduct experiments on three widely used public datasets:

- *MIRFlickr-25K* (Huiskes & Lew, 2008) contains $25,000$ text-image pairs. Each text data is represented by a 1386-dimensional Bag-of-Words (BoW) vector.
- *NUS-WIDE* (Rasiwasia et al., 2010) comprises $269,498$ text-image pairs with multiple labels from 81 categories, where each text data involves a 1000-dimensional BoW vector.
- *MS-COCO* (Lin et al., 2014) includes $123,287$ text-image pairs with multiple labels from 80 categories. Each text data is represented by a 2026-dimensional BoW vector.

Following the problem settings of the latest baseline (Zhang et al., 2024b), each dataset is divided into a query set and a retrieval set. During the training process, only text-image pair information is accessible without label information.

E.3 Implementation Details

Our method is implemented based on the latest baseline (Zhang et al., 2024b). For reproducibility purposes, we have made our code and model checkpoints publicly available at https://anonymous.4open.science/r/HINT.

For training, we maintain consistency with baseline approaches by selecting 10,000 samples as our training set. We utilize pre-extracted visual and text feature vectors, which are mapped to the Hamming space through two-layer MLPs with a dimension of 512. The implementation is done using PyTorch framework, with all experiments conducted on a single NVIDIA A40 GPU.

For optimization, we employ the Adam optimizer (Kingma & Ba, 2014) with a batch size of 128 and a learning rate of 1e-3. The model is trained for 20 epochs. Performance evaluation is conducted using Mean Average Precision (MAP) and hamming lookup curves, specifically the Precision-Recall curve, Precision-N curve, and Recall-N curve.

E.4 Evaluation Metrics

Mean Average Precision (MAP) is a comprehensive metric widely used to evaluate retrieval performance in cross-modal hashing research (Wang et al., 2010; Liu et al., 2012; Shen et al., 2015). The

MAP score has a range of 0 to 1, where higher values indicate better retrieval performance. It works by calculating the average precision for each query and then taking the mean across all queries in the test set. This provides a single-figure measure that reflects system performance across all relevant documents and all recall levels. MAP considers both precision and recall aspects of the retrieval system, making it particularly suitable for evaluating hashing-based retrieval systems where we are concerned with the overall ranking quality of results.

F THE USE OF LARGE LANGUAGE MODELS

In this article, we use LLM for language polishing and retrieval of the latest research works. We confirm that we take full responsibility for the contents written in this paper.