HIGH-ORDER INTERACTION AND MODALITY-SPECIFIC DISTILLATION FOR UNSUPERVISED CROSS-MODAL HASHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised cross-modal hashing aims to encode heterogeneous data into a shared compact Hamming space without relying on semantic annotations. A key design challenge is to exploit interactions between paired image and text features during training while transferring the resulting multimodal knowledge to independently deployable unimodal hash encoders. To address this limitation, we propose a High-Order Cross-Modal Hashing framework, termed **HOCH**, which combines progressive cross-modal interaction with relational knowledge distillation. Its cascaded multi-head interaction blocks form products of normalized low-rank projections and use these interaction features to update the image and text representations through separate gated residual updates. Each subsequent block computes interactions from the updated representations, without explicitly constructing full outer products. An adaptive modality fusion module then combines the enhanced representations using sample-dependent weights to construct a multimodal teacher. Modality-specific relational distillation transfers both cross-modal and intra-modal similarity structures from the teacher to image and text student hash branches, enabling each branch to generate hash codes independently at inference. Experiments on MIRFlickr, NUS-WIDE, and MSCOCO with hash-code lengths from 16 to 128 bits show that **HOCH** achieves the best results in most evaluated settings.

1 INTRODUCTION

With the explosive growth of multimedia data on the Internet, cross-modal retrieval has garnered extensive attention from both academia and industry. Due to its significant advantages in storage cost and retrieval speed (Wang et al., 2024b; 2018), cross-modal hashing, which maps high-dimensional heterogeneous data into compact binary codes, has become an important solution for large-scale application scenarios (Wang et al., 2024a; Liu et al., 2025; Zhao et al., 2026). The fundamental challenge of cross-modal hashing lies in bridging the semantic gap between different modalities, such as vision and language, to construct a highly discriminative joint Hamming space. The key to achieving this goal is the deep interaction and alignment of multimodal features (Li et al., 2025b): since vision and language are heterogeneous and complementary, a discriminative joint Hamming space must exploit both their shared semantics and their cross-modal correlations at the feature level.

Recent unsupervised cross-modal hashing methods have progressed well beyond early linear fusion and have approached the semantic gap from several complementary directions. Reconstruction-based methods preserve joint affinities or correlation–identity semantics (Su et al., 2019; Zhu et al., 2023); methods based on graph modeling and contrastive learning model neighborhood structures within and across modalities (Liang et al., 2024); and methods based on clustering, prototypes, or similarity refinement construct increasingly reliable pseudo-semantic supervision from unlabeled data (Liu et al., 2025; Fan & Cao, 2025; Fu et al., 2026; Zhao et al., 2026). Meanwhile, pretrained vision–language models and Transformer-based architectures provide stronger semantic initialization and contextual alignment (Wang et al., 2024a; Fan & Cao, 2025). Some recent approaches also model higher-order sample structures or dynamically balance multiple learning objectives (Zhao et al., 2026). Therefore, contemporary unsupervised cross-modal hashing methods cannot be reduced to simple first-order feature aggregation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Alongside these advances in semantic relation learning and multimodal representation, the interaction between paired image and text features deserves further consideration. Similarity constraints guide which samples should be close in Hamming space, but do not by themselves determine how the two modalities interact at the feature level. In particular, the higher-order sample-structure matrix used by UDCH (Zhao et al., 2026) captures relationships among samples, which differ from multiplicative interactions between image and text feature channels. At the feature level, concatenation and element-wise addition do not explicitly form cross-modal feature products. Attention-based architectures can capture complex dependencies, while explicit multiplicative interactions offer a different modeling choice. Full bilinear or higher-order outer products can represent such feature interactions, but their dimensionality and computational cost can grow rapidly. Compact bilinear pooling and factorized bilinear pooling reduce this cost in multimodal learning (Fukui et al., 2016; Yu et al., 2017). Motivated by these approaches, we investigate whether recursively composed, compact multiplicative interactions can improve joint image-text representations for unsupervised cross-modal hashing. MFH constructs high-order pooled features through cascaded multiplicative interactions on the same input representations (Yu et al., 2018). This motivates examining whether compact multiplicative interactions can progressively refine both image and text representations for unsupervised cross-modal hashing.

Two further issues are coupled with interaction modeling. First, the relative informativeness of image and text varies across samples, so a fixed or globally shared fusion rule may overemphasize a weak or noisy modality. Although quality-aware dynamic fusion has been explored in broader multimodal learning (Zhang et al., 2023), it remains a separate design choice from explicit multiplicative interaction. Second, paired modalities can be jointly exploited during training, whereas an online query normally contains only an image or only text. Existing reconstruction and cross-modal knowledge-transfer methods have transferred multimodal semantics, prototypes, or pairwise similarity information to modality-specific or lightweight hashing networks (Zhu et al., 2023; Li et al., 2023a; Hu et al., 2020; Li et al., 2025a). Together, these observations point to a two-fold objective: capturing higher-order cross-modal interactions at low cost, and distilling the resulting relational structure to unimodal branches that incur no extra inference overhead.

Motivated by these challenges and related advances in multimodal learning, we propose High-Order Cross-Modal Hashing (**HOCH**), which integrates high-order cross-modal interaction (HOI), adaptive modality fusion (AMF), and modality-specific relational distillation (MSRD) to construct a sample-adaptive multimodal teacher and transfer its relational structure to independently deployable unimodal hash branches.

Contributions. The main contributions of this paper are: 1) We are the first to introduce a high-order cross-modal interaction module into hashing-based retrieval. Specifically, we propose a high-order cross-modal interaction module that captures complex semantic relationships between images and texts through multi-head low-rank factorized projections and multiplicative feature interactions, with gated residual updates and multi-level feature aggregation that progressively enhance cross-modal representations without explicitly constructing high-dimensional outer-product features; 2) We introduce a lightweight adaptive modality fusion mechanism that dynamically assigns fusion weights to image and text representations according to the content of each sample, avoiding a fixed modality ratio across all instances and constructing a sample-dependent multimodal teacher representation; 3) We develop a modality-specific relational distillation strategy that transfers both cross-modal and intra-modal relationships among samples learned by the multimodal teacher to the image and text student branches, enabling each branch to independently generate hash codes from a single modality at inference time without requiring paired multimodal inputs.

2 Related work

The core objective of cross-modal hashing is to project heterogeneous multimedia data into a common Hamming space, thereby enabling efficient and large-scale cross-modal retrieval. Based on whether human-annotated semantic labels are utilized during the training phase, existing cross-modal hashing methods can be broadly categorized into two main paradigms: supervised approaches (Jiang et al., 2025; Li et al., 2024; Yang et al., 2024) and unsupervised approaches (Zhu et al., 2023; Wang et al., 2024a; Zhao et al., 2026). Given the prohibitive cost of acquiring large-scale annotated data in real-world scenarios, this paper primarily focuses on the more practical unsupervised

cross-modal hashing paradigm. This section reviews the studies most relevant to **HOCH** from three perspectives: unsupervised cross-modal hashing, high-order feature interaction and adaptive fusion, and relational knowledge distillation.

Unsupervised Cross-Modal Hashing. A central challenge in unsupervised cross-modal hashing is to learn semantic relationships without human annotations. Reconstruction and graph learning offer two approaches: CIRH learns complementary representations through correlation and identity reconstruction (Zhu et al., 2023), while CMGCH models relationships within and across modalities through multimodal graph learning and contrastive learning (Liang et al., 2024). Other methods refine the information used to supervise hash learning. VTM-UCH uses CLIP-based visual guidance and community structure to improve similarity supervision (Fan & Cao, 2025), while UDCH combines instance-level alignment, clustering, and higher-order structural supervision (Zhao et al., 2026). ENHash explores error-guided optimization, using recorded hash prediction errors to guide fine-grained learning (Fu et al., 2026).

High-Order Interaction and Adaptive Fusion. Multimodal interaction and fusion determine how visual and textual information are combined. HuggingHash uses Transformer representations and multi-granularity alignment for unsupervised cross-modal hashing (Wang et al., 2024a). Explicit multiplicative interactions have also been studied in multimodal learning: MCB and MFB provide compact alternatives to full bilinear pooling (Fukui et al., 2016; Yu et al., 2017), while MFH extends factorized pooling to higher-order interactions (Yu et al., 2018). Beyond interaction modeling, adaptive fusion adjusts how information from different modalities contributes to the prediction. QMF studies quality-aware dynamic fusion using uncertainty estimation (Zhang et al., 2023). In visual grounding, VG-LAW uses language-conditioned weight generation to adapt visual feature extraction (Su et al., 2023).

Relational Knowledge Distillation. Knowledge distillation transfers knowledge from a teacher to a student model. Relational knowledge distillation (Park et al., 2019) extends the transfer target from individual representations to relationships between samples, allowing the student to preserve the structure of the teacher's representation space. In cross-modal hashing, UKD derives similarity information from an unsupervised teacher to supervise student hash learning without human annotations (Hu et al., 2020). In federated unsupervised cross-modal hashing, prototype-guided knowledge transfer uses global prototypes to guide local hash learning and feature alignment (Li et al., 2023a). For online cross-modal hashing, LCDH transfers knowledge from an offline teacher to a lightweight student through similarity matrix approximation (Li et al., 2025a).

3 Proposed Method

3.1 Overall Framework

Given an unsupervised training set $\mathcal{D} = \{(\mathbf{x}_i^n, \mathbf{x}_t^n)\}_{n=1}^N$, where $\mathbf{x}_i^n \in \mathbb{R}^{d_i}$ and $\mathbf{x}_t^n \in \mathbb{R}^{d_t}$ denote the image and text features of the n-th sample, our goal is to learn modality-specific hashing functions $F_m(\cdot)$ that generate $\mathbf{b}_m = \operatorname{sign}(F_m(\mathbf{x}_m)) \in \{-1, +1\}^K$ for $m \in \{i, t\}$. The modality-specific projection networks first produce the shared-space representations $\mathbf{z}_i^0 = P_i(\mathbf{x}_i)$ and $\mathbf{z}_t^0 = P_t(\mathbf{x}_t)$. All feature and hash vectors are represented as row vectors, and mini-batch matrices stack samples along their rows. The teacher and student branches share the modality-specific encoders P_i and P_t but use separate hash heads. Teacher similarity targets are detached only in the distillation loss; the teacher is jointly optimized by its own objective.

Starting from the initial representations, HOI produces the enhanced image and text features, which AMF combines to form the multimodal teacher representations. MSRD then aligns the cross-modal and intra-modal similarity structures of the teacher and student hash outputs. During testing, only the corresponding unimodal student branch is retained, allowing image or text queries to independently generate binary hash codes without requiring paired multimodal inputs. The architectures and learning mechanisms of the three modules are described in the following subsections. The overall framework of **HOCH** is illustrated in Figure 1.

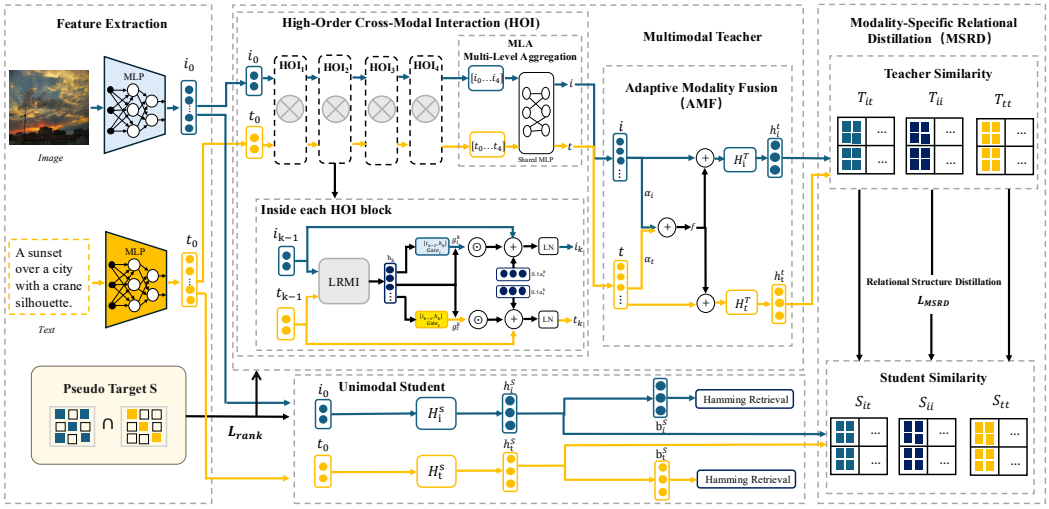


Figure 1: The overall framework of the proposed High-Order Cross-Modal Hashing (HOCH).

3.2 High-Order Cross-Modal Interaction

To capture complex semantic dependencies across modalities, the HOI module applies multi-head normalized low-rank multiplicative interactions to the image and text representations. Unlike simple feature alignment or element-wise addition, it explicitly models multiplicative dependencies between heterogeneous feature channels while avoiding the dimensional expansion of a full bilinear outer product. Within each interaction block, the two modality-specific low-rank projections use $2dr$ parameters in total and produce Hr multiplicative interaction features without constructing a full outer product. This parameter count covers only the low-rank projections; the complete HOI module also includes output projections, gated updates, and multi-level aggregation.

Given the image representation $\mathbf{z}_i^{l-1}$ and text representation $\mathbf{z}_t^{l-1}$ as the inputs of the l-th interaction block, we divide each representation into H heads, where $\mathbf{z}_{m,h}^{l-1} \in \mathbb{R}^{d/H}$ denotes the h-th sub-representation of modality $m \in \{i, t\}$. The modality-specific low-rank projection and multiplicative interaction are jointly formulated as:

$$\begin{aligned} \widetilde{\mathbf{p}}_{m,h}^{l} &= \mathbf{z}_{m,h}^{l-1}\mathbf{W}_{m,h}^{l}, \\ \mathbf{p}_{m,h}^{l} &= \frac{\widetilde{\mathbf{p}}_{m,h}^{l}}{\max\left(\left\|\widetilde{\mathbf{p}}_{m,h}^{l}\right\|_2, \varepsilon\right)}, \qquad \mathbf{W}_{m,h}^{l} \in \mathbb{R}^{\frac{d}{H}\times r}, \\ \mathbf{u}_h^{l} &= \mathbf{p}_{i,h}^{l} \odot \mathbf{p}_{t,h}^{l}, \end{aligned} \tag{1}$$

where $\odot$ denotes element-wise multiplication, r denotes the interaction rank, and ε is a small constant for numerical stability. Each scalar component of $\mathbf{u}_h^l$ is a normalized product of two linear projections. Its connection to a rank-one bilinear form is derived in Appendix A.1. By cascading multiple interaction blocks, **HOCH** recursively composes these interactions to capture increasingly complex nonlinear dependencies. Here, "high-order" refers to this recursive composition rather than a strict polynomial degree.

The interaction features from all heads are concatenated and projected into the shared hidden space. During training, DropPath is applied to the projected features, yielding $\mathbf{u}^l$ for gate computation and residual updates. After obtaining the high-order interaction feature, we introduce a gated interaction update mechanism to adaptively control how much high-order information should be injected into each modality branch. Specifically, the image and text branches use independent gating functions:

$$\mathbf{g}_m^l = \sigma\left([\mathbf{z}_{\bar{m}}^{l-1}; \mathbf{u}^l]\mathbf{W}_{g_m}^l + \mathbf{b}_{g_m}^l\right), \qquad m \in \{i, t\}, \tag{2}$$

where $\mathbf{g}_i^l$ and $\mathbf{g}_t^l$ denote the gating weights for the image and text branches, respectively. $\bar{m}$ denotes the opposite modality of m, namely $\bar{i} = t$ and $\bar{t} = i$. $\sigma(\cdot)$ denotes the sigmoid function, and $[\cdot;\cdot]$ denotes feature concatenation. The image gate is determined by the text representation and the high-order interaction feature, while the text gate is determined by the image representation and the high-order interaction feature. In this way, the two modalities can absorb cross-modal interaction information in different adaptive manners. The image and text representations are then updated as follows:

$$\mathbf{z}_m^l = \mathrm{LN}\left(\mathbf{z}_m^{l-1} + \mathbf{g}_m^l \odot \mathbf{u}^l + \alpha \mathbf{a}_m^l\right), \quad m \in \{i, t\}, \tag{3}$$

where $\mathbf{a}_m^l$ denotes the modality-specific residual anchor, $\mathrm{LN}(\cdot)$ denotes layer normalization, and α controls the anchor strength. The four interaction blocks use $\mathbf{z}_m^0$, $\mathbf{z}_m^0$, $\mathbf{z}_m^1$, and $\mathbf{z}_m^2$ as residual anchors, respectively, with $\alpha = 0.1$. This update preserves modality-specific semantics while injecting gated high-order interaction information.

After L interaction blocks, a shared aggregation network combines the initial and intermediate representations to obtain $\mathbf{z}_i^*$ and $\mathbf{z}_t^*$, which serve as inputs to the adaptive fusion module. Details of the output projection, DropPath, and multi-level aggregation are provided in Appendix A.2.

3.3 Adaptive Modality Fusion

Given the enhanced representations $\mathbf{z}_i^*$ and $\mathbf{z}_t^*$, AMF predicts sample-dependent modality weights and constructs the fused representation and teacher hash representations as:

$$\begin{aligned} [\omega_i, \omega_t] &= \mathrm{Softmax}\left((\mathbf{z}_i^* + \mathbf{z}_t^*)\mathbf{W}_f + \mathbf{b}_f\right), \\ \mathbf{z}_f &= \omega_i \mathbf{z}_i^* + \omega_t \mathbf{z}_t^*, \\ \mathbf{h}_m^T &= \tanh\left(\beta\left[(\mathbf{z}_m^* + \mathbf{z}_f)\mathbf{W}_m^T + \mathbf{c}_m^T\right]\right), \qquad m \in \{i, t\}. \end{aligned} \tag{4}$$

Here, the weight matrices have dimensions $\mathbf{W}_f \in \mathbb{R}^{d\times 2}$ and $\mathbf{W}_m^X \in \mathbb{R}^{d\times K}$ for $X \in \{T, S\}$. The corresponding bias vectors are $\mathbf{b}_f \in \mathbb{R}^{1\times 2}$ and $\mathbf{c}_m^X \in \mathbb{R}^{1\times K}$. The weights ω_i and ω_t sum to one and vary with the input pair, allowing the fusion ratio to adapt across samples. The fused representation $\mathbf{z}_f$ is added to each enhanced modality representation before separate teacher hash projections, so both teacher outputs incorporate information from the paired image and text. The positive scalar β controls the sharpness of the continuous hash representations.

3.4 Modality-Specific Relational Distillation

The multimodal teacher branch exploits paired image-text inputs to learn pairwise semantic relationships, whereas practical cross-modal retrieval typically involves a query from only one modality. To bridge this discrepancy, we introduce MSRD, which transfers the relational structure learned by the multimodal teacher to independent image and text student branches. Given two representation matrices $\mathbf{A}$ and $\mathbf{B}$, their cosine similarity matrix is defined as:

$$\mathrm{Cos}(\mathbf{A}, \mathbf{B})_{mn} = \frac{\mathbf{a}_m \mathbf{b}_n^\top}{\max(\|\mathbf{a}_m\|_2, \varepsilon)\max(\|\mathbf{b}_n\|_2, \varepsilon)}, \tag{5}$$

where $\mathbf{a}_m$ and $\mathbf{b}_n$ denote the m-th and n-th row vectors of $\mathbf{A}$ and $\mathbf{B}$, respectively. Specifically, we construct an image student hashing branch and a text student hashing branch. Given the initial image representation $\mathbf{z}_i^0$ and the initial text representation $\mathbf{z}_t^0$, the two student branches generate the corresponding student hash representations as follows:

$$\mathbf{h}_m^S = \tanh\left(\beta\left[\mathbf{z}_m^0 \mathbf{W}_m^S + \mathbf{c}_m^S\right]\right), \qquad m \in \{i, t\}, \tag{6}$$

where $\mathbf{W}_m^S$ denotes the modality-specific student hash projection. Each student branch depends only on its corresponding initial representation $\mathbf{z}_m^0$ and does not participate in high-order cross-modal interaction. Therefore, the image and text student branches can independently generate hash codes during inference.

Instead of directly matching teacher and student hash representations sample by sample, MSRD aligns their pairwise similarity structures. This relational constraint allows the student branches to inherit multimodal semantic knowledge while retaining their modality-specific characteristics.

Specifically, based on the image teacher hash representation $\mathbf{H}_i^T$ and text teacher hash representation $\mathbf{H}_t^T$ in a mini-batch, we construct the teacher similarity matrices as follows:

$$\mathcal{S}^T = \left\{\mathbf{S}_{it}^T, \mathbf{S}_{ii}^T, \mathbf{S}_{tt}^T\right\} = \left\{\mathrm{Cos}(\mathbf{H}_i^T, \mathbf{H}_t^T), \mathrm{Cos}(\mathbf{H}_i^T, \mathbf{H}_i^T), \mathrm{Cos}(\mathbf{H}_t^T, \mathbf{H}_t^T)\right\}. \tag{7}$$

Here, $\mathbf{S}_{it}^T$ captures image-text similarities, $\mathbf{S}_{ii}^T$ captures similarities among images, and $\mathbf{S}_{tt}^T$ captures similarities among texts. Similarly, the student similarity matrices are constructed from the image student hash representation $\mathbf{H}_i^S$ and text student hash representation $\mathbf{H}_t^S$:

$$\mathcal{S}^S = \left\{\mathbf{S}_{it}^S, \mathbf{S}_{ii}^S, \mathbf{S}_{tt}^S\right\} = \left\{\mathrm{Cos}(\mathbf{H}_i^S, \mathbf{H}_t^S), \mathrm{Cos}(\mathbf{H}_i^S, \mathbf{H}_i^S), \mathrm{Cos}(\mathbf{H}_t^S, \mathbf{H}_t^S)\right\}. \tag{8}$$

Then, the relational knowledge learned by the high-order interaction branch is transferred to the student branches by minimizing the discrepancy between teacher and student similarity matrices:

$$\begin{aligned}\mathcal{L}_{\mathrm{MSRD}} = \frac{1}{B^2}\Big(&\left\|\mathbf{S}_{it}^S - \mathrm{sg}(\mathbf{S}_{it}^T)\right\|_F^2 \\ &+ \left\|\mathbf{S}_{ii}^S - \mathrm{sg}(\mathbf{S}_{ii}^T)\right\|_F^2 \\ &+ \left\|\mathbf{S}_{tt}^S - \mathrm{sg}(\mathbf{S}_{tt}^T)\right\|_F^2 \Big),\end{aligned} \tag{9}$$

where B denotes the mini-batch size, $\|\cdot\|_F^2$ denotes the squared Frobenius norm, and $\mathrm{sg}(\cdot)$ represents the stop-gradient operation. The factor $1/B^2$ converts each matrix discrepancy into the mean squared error over all pairwise relations in the mini-batch. The teacher similarity matrices are treated as fixed distillation targets when optimizing the student branches. The first term aligns image-text similarities between the teacher and students, while the other two terms align similarities among images and among texts. These constraints transfer the teacher's relational structure to the student branches, each of which requires only its own modality at inference time.

3.5 Learning Objective

We jointly optimize the multimodal teacher and modality-specific student branches using pseudo-supervised ranking, relational distillation, and hash regularization. Since semantic labels are unavailable, the supervision is constructed from the agreement between the initial image and text similarities within each mini-batch. For distinct training pairs, a positive relation is assigned only when their similarities in both modalities exceed the corresponding mini-batch quantile thresholds. Diagonal entries are always set to one. The complete pseudo-similarity construction is provided in Appendix A.3.

Given the pseudo-similarity matrix, we employ a ranking loss to preserve the retrieval structure of the learned hash representations. For an arbitrary similarity matrix $\mathbf{M}$, the mean positive similarity of the m-th query is defined as $\overline{M}_m^+ = |\mathcal{P}_m|^{-1}\sum_{p\in\mathcal{P}_m} M_{mp}$, where $\mathcal{P}_m = \{p \mid S_{mp} = 1\}$. Let $\mathcal{N}_m = \{n \mid S_{mn} = 0,\ n \neq m\}$ denote the negative set, and let $\Omega = \{m \mid |\mathcal{P}_m| > 0\}$. For each $m \in \Omega$, the query-wise ranking loss is

$$\ell_m(\mathbf{M};\gamma,\tau) = \begin{cases} \max\left(0, \gamma - \overline{M}_m^+ + \tau\log\displaystyle\sum_{n\in\mathcal{N}_m}\exp\left(\frac{M_{mn}}{\tau}\right)\right), & |\mathcal{N}_m| > 0, \\ 0, & |\mathcal{N}_m| = 0. \end{cases} \tag{10}$$

The overall ranking loss is

$$\mathcal{L}_{\mathrm{rank}}(\mathbf{M};\gamma,\tau) = \begin{cases} \dfrac{1}{|\Omega|}\displaystyle\sum_{m\in\Omega}\ell_m(\mathbf{M};\gamma,\tau), & |\Omega| > 0, \\ 0, & |\Omega| = 0. \end{cases} \tag{11}$$

Here, γ is the margin and $\tau > 0$ is the temperature. Queries without negative samples contribute zero while remaining in the averaging set Ω. Since $S_{mm} = 1$, every query in a nonempty mini-batch has at least one positive sample. The loss encourages the mean positive similarity to exceed the log-sum-exp aggregation of negative similarities by a margin, with greater emphasis on hard negatives. We impose bidirectional cross-modal and intra-modal ranking constraints on the teacher similarity matrices. Their complete definitions are provided in Appendix A.4.

To improve cross-modal consistency and hash quality, we apply alignment, quantization, and bit-balance regularization to both the teacher and student branches:

$$\mathcal{R}^X = \lambda_a^X \mathcal{R}_a^X + \lambda_q^X \mathcal{R}_q^X + \lambda_b^X \mathcal{R}_b^X, \qquad X \in \{T, S\}, \tag{12}$$

where $\mathcal{R}_a^X$, $\mathcal{R}_q^X$, and $\mathcal{R}_b^X$ denote the alignment, quantization, and bit-balance terms, respectively. Their complete definitions are provided in Appendix A.5. Therefore, the teacher branch loss is formulated as:

$$\mathcal{L}_T = \lambda_c^T \mathcal{L}_{\text{cross}}^T + \lambda_r^T \mathcal{L}_{\text{intra}}^T + \mathcal{R}^T. \tag{13}$$

For the modality-specific student branches, relational distillation is combined with a bidirectional cross-modal ranking constraint, whose complete definition is provided in Appendix A.4. Similarly, the student branch loss is formulated as:

$$\mathcal{L}_S = \lambda_d \mathcal{L}_{\text{MSRD}} + \lambda_c^S \mathcal{L}_{\text{cross}}^S + \mathcal{R}^S. \tag{14}$$

In addition, we introduce a lightweight feature-preservation regularization term to prevent the enhanced representations from deviating excessively from their initial modality-specific representations. Let $\mathbf{Z}_i^0$ and $\mathbf{Z}_t^0$ denote the initial image and text representations in a mini-batch, and let $\mathbf{Z}_i^*$ and $\mathbf{Z}_t^*$ denote their corresponding representations after high-order interaction and multi-level aggregation. The feature-preservation loss is defined as:

$$\mathcal{L}_{\text{fp}} = \frac{1}{Bd} \left(\left\| \mathbf{Z}_i^* - \mathbf{Z}_i^0 \right\|_F^2 + \left\| \mathbf{Z}_t^* - \mathbf{Z}_t^0 \right\|_F^2 \right), \tag{15}$$

where B and d denote the mini-batch size and representation dimension, respectively. This regularization preserves the original modality-specific information while allowing the interaction modules to learn complementary cross-modal semantics. Finally, the overall optimization objective is defined as:

$$\mathcal{L} = \kappa_T \lambda_T(e) \mathcal{L}_T + \lambda_S(e) \mathcal{L}_S + \lambda_{\text{fp}} \mathcal{L}_{\text{fp}}, \tag{16}$$

where e denotes the current training epoch. The global scaling factor κ_T adjusts the teacher objective and is set to 1.0 by default. The epoch-dependent weights $\lambda_T(e)$ and $\lambda_S(e)$ control the contributions of the teacher and student objectives. The coefficient λ_{fp} controls the strength of feature-preservation regularization. Here, B is the mini-batch size and d is the representation dimension.

4 Experiments

4.1 Experimental Setup

We evaluate **HOCH** on MIRFlickr (Huiskes & Lew, 2008), NUS-WIDE (Chua et al., 2009), and MSCOCO (Lin et al., 2014) for image-to-text (I2T) and text-to-image (T2I) retrieval. We report mAP@50 using 16-, 32-, 64-, and 128-bit hash codes and additionally examine mAP@K. For **HOCH**, we use fixed 512-dimensional CLIP image and text features. HOI contains four interaction blocks with eight heads and rank $r = 16$, and the model is trained for 100 epochs using AdamW with a batch size of 128. Dataset statistics, relevance definitions, evaluation protocols, and additional implementation details are provided in Appendix B. Semantic labels are used only for evaluation and never for training. For a fair comparison, all methods use the same data partitions, fixed CLIP features, hash-code lengths, and evaluation protocol.

4.2 Comparison with State-of-the-Art Methods

Table 1 compares **HOCH** with nine state-of-the-art baselines in terms of mAP@50: DJSRH (Su et al., 2019), JDSH (Liu et al., 2020), DGCPN (Yu et al., 2021), DAEH (Shi et al., 2022), CA-GAN (Li et al., 2023b), MSRH (Wu & Li, 2024), SMSH (Kuai et al., 2025), VTM-UCH (Fan & Cao, 2025), and USFTH (Yang et al., 2025). On MIRFlickr, **HOCH** achieves average mAP@50 scores of 0.924 for I2T and 0.896 for T2I. On NUS-WIDE, **HOCH** achieves the best results in five settings and remains competitive in the others, although several baselines perform better for long-code T2I retrieval. The strongest advantage is observed on MSCOCO, with its performance increasing from 0.881 to 0.944 for I2T and from 0.883 to 0.956 for T2I as the code length increases from 16 to 128 bits.

Table 1: Performance comparison (mAP@50) of **HOCH** and baselines. Code lengths are measured in bits.

Method	MIRFlickr				NUS-WIDE				MSCOCO			
	16	32	64	128	16	32	64	128	16	32	64	128
Image-to-text (I2T)												
DJSRH	0.805	0.868	0.902	0.913	0.686	0.768	0.805	0.826	0.705	0.818	0.890	0.916
JDSH	0.643	0.886	0.915	0.929	0.629	0.720	0.820	**0.847**	0.531	0.747	0.922	0.934
DGCPN	0.806	0.891	0.917	0.920	0.719	0.776	0.817	0.822	0.719	0.756	0.775	0.815
DAEH	0.891	0.914	0.926	0.931	0.793	0.818	0.836	0.842	0.735	0.818	0.832	0.813
CAGAN	0.595	0.624	0.727	0.880	0.303	0.550	0.685	0.673	0.380	0.672	0.833	0.915
MSRH	0.869	0.892	0.900	0.903	0.774	0.804	0.819	0.826	0.760	0.790	0.808	0.824
SMSH	0.884	0.881	0.908	0.917	0.767	0.788	0.800	0.785	0.773	0.853	0.883	0.904
VTM-UCH	**0.915**	0.916	0.929	0.929	0.801	0.798	0.817	0.818	0.745	0.763	0.768	0.772
USFTH	0.897	0.912	0.928	0.936	0.795	0.814	0.824	0.841	0.799	0.846	0.873	0.892
HOCH	0.898	**0.922**	**0.933**	**0.942**	**0.806**	**0.831**	**0.837**	0.845	**0.881**	**0.910**	**0.936**	**0.944**
Text-to-image (T2I)												
DJSRH	0.819	0.841	0.859	0.873	0.711	0.743	0.785	0.808	0.700	0.851	0.914	0.934
JDSH	0.595	0.882	0.888	0.899	0.584	0.731	0.802	**0.824**	0.490	0.762	0.937	0.944
DGCPN	0.830	0.881	0.888	0.899	0.719	0.778	0.792	0.807	0.702	0.753	0.778	0.829
DAEH	0.872	0.888	0.899	0.903	0.781	0.802	**0.821**	0.822	0.734	0.820	0.850	0.832
CAGAN	0.595	0.668	0.744	0.860	0.303	0.589	0.584	0.657	0.440	0.678	0.838	0.924
MSRH	0.853	0.867	0.874	0.882	0.765	0.793	0.794	0.804	0.755	0.792	0.806	0.822
SMSH	0.801	0.847	0.848	0.867	0.727	0.771	0.784	0.794	0.816	0.861	0.910	0.935
VTM-UCH	0.880	0.877	0.879	0.882	0.770	0.783	0.785	0.782	0.761	0.729	0.749	0.743
USFTH	0.860	0.878	0.892	0.897	0.773	0.802	0.809	0.819	0.800	0.851	0.879	0.900
HOCH	**0.881**	**0.896**	**0.901**	**0.907**	**0.784**	**0.803**	0.811	0.815	**0.883**	**0.925**	**0.949**	**0.956**

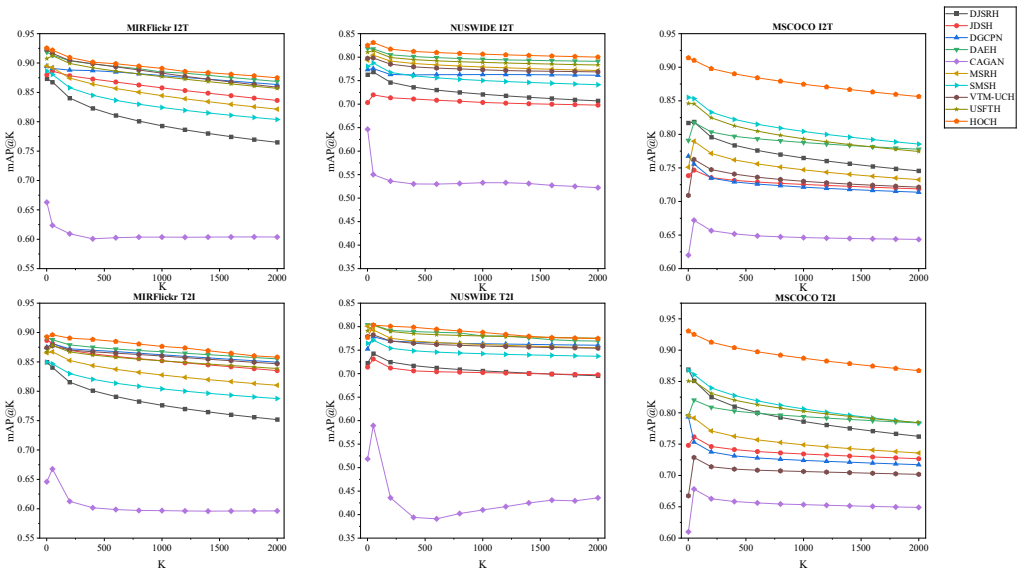


Figure 2: The mAP@K curves on the three datasets at 32 bits.

Figure 2 further evaluates retrieval performance at different depths using 32-bit hash codes. **HOCH** remains competitive across a broad range of K on MIRFlickr and NUS-WIDE and maintains a clear advantage on MSCOCO in both retrieval directions. This indicates that its performance is not limited to the mAP@50 setting. More detailed dataset-wise observations are provided in Appendix C.1. Qualitative t-SNE visualizations in Appendix C.2 further show that, under identical visualization settings, **HOCH** exhibits greater overlap between image and text query hash codes in the two-dimensional projections than the selected baselines.

Parameter Sensitivity. Across the tested ranges, **HOCH** shows moderate sensitivity to the interaction rank, teacher-loss scale, distillation weight, and feature-preservation scale. Detailed results and the default configuration are provided in Appendix C.3.

4.3 Ablation and Efficiency Analysis

Component Ablation. Table 2 reports ablation results using 64-bit hash codes. Compared with FirstOrder, the full model achieves modest improvements in bidirectional average mAP@50 across all three datasets, supporting the contribution of multiplicative interaction within HOCH. Replacing adaptive fusion with mean fusion or removing feature preservation also reduces performance. Removing MSRD yields the largest average decline, highlighting the importance of relational knowledge transfer. The following efficiency analysis compares the computational costs of HOCH and FirstOrder. Detailed dataset-wise results are provided in Appendix C.4.

Table 2: Ablation study in terms of bidirectional average mAP@50 with 64-bit hash codes.

Method	MIRFlickr	NUS-WIDE	MSCOCO	Average
HOCH-FirstOrder	0.9146	0.8206	0.9400	0.8917
HOCH-MeanFusion	0.9162	0.8236	0.9401	0.8933
HOCH-w/o MSRD	0.9020	0.8091	0.9406	0.8839
HOCH-w/o FP	0.9157	0.8203	0.9367	0.8909
HOCH	**0.9169**	**0.8243**	**0.9425**	**0.8945**

Efficiency and Deployment Analysis. Table 3 compares retrieval accuracy and computational cost on NUS-WIDE with 64-bit hash codes. Details of the efficiency measurements are provided in Appendix B.3. Compared with FirstOrder, **HOCH** improves bidirectional average mAP@50 from 0.8206 to 0.8243 while reducing active training parameters, forward MACs, training-step time, and peak memory by 5.4%, 4.4%, 7.4%, and 4.3%. Both variants use the same single-modal students at inference. HOI and AMF are used only during training, so the retrieval gain in this setting comes with lower measured training cost and no additional student inference overhead.

Table 3: Bidirectional retrieval accuracy and computational efficiency on NUS-WIDE with 64-bit hash codes.

Stage	Variant/Path	mAP@50	Active Params (M)	MACs (M)	Time (ms)	Memory (MiB)
Training	FirstOrder	0.8206	13.27	16.39	20.81 ± 0.59	255
Training	**HOCH**	**0.8243**	**12.55**	**15.66**	$\mathbf{19.27 \pm 0.30}$	**244**
Inference	Image Student	–	2.40	2.39	0.185 ± 0.005	–
Inference	Text Student	–	2.40	2.39	0.184 ± 0.008	–

MACs are estimated per sample for linear layers and low-rank projections. Latencies are reported as mean $\pm$ standard deviation.

5 Conclusions

In this paper, we proposed **HOCH**, an unsupervised cross-modal hashing framework that integrates high-order cross-modal interaction, adaptive modality fusion, and modality-specific relational distillation. The cascaded low-rank multiplicative interaction blocks model complex image–text dependencies without explicitly constructing full outer-product representations. Adaptive fusion constructs a sample-dependent multimodal teacher, whose cross-modal and intra-modal similarity structures are transferred to independently deployable unimodal student hash branches. Experiments on MIRFlickr, NUS-WIDE, and MSCOCO show that **HOCH** obtains the best results in most evaluated settings. At 64 bits, HOI provides modest but consistent gains over FirstOrder in bidirectional average mAP@50 across all three datasets. On NUS-WIDE, these gains are achieved with lower measured training cost and no additional student inference overhead.

AI USE STATEMENT

In this work, we did not use generative AI tools for any task requiring disclosure. We have not used generative AI tools for research ideation, method design, code generation, experiment execution, data analysis, result interpretation, or artifact creation, and the remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI for language polishing and grammar checking. We have reviewed all AI assisted work. Specifically, all AI suggested language edits were manually checked by the authors against the original technical content, and no AI generated research ideas, code, experimental results, or scientific claims were used. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The model and training objectives are described in Section 3. Dataset partitions, evaluation metrics, and implementation settings are provided in Appendix B. We plan to release the code upon publication.

REFERENCES

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 Stéphane Marchand-Maillet and Yiannis Kompatsiaris (eds.), *Proceedings of the 8th ACM International Conference on Image and Video Retrieval, CIVR 2009, Santorini Island, Greece, July 8-10, 2009*. ACM, 2009. doi: 10.1145/1646396.1646452. URL https://doi.org/10.1145/1646396.1646452.

Haozhi Fan and Yuan Cao. Vision-guided text mining for unsupervised cross-modal hashing with community similarity quantization. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 2843–2851. AAAI Press, 2025. doi: 10.1609/AAAI.V39I3.32290. URL https://doi.org/10.1609/aaai.v39i3.32290.

Hao Fu, Zebing Yao, Chuangchuang Tan, and Guanghua Gu. ENHash: Error notebook-guided fine-grained learning for unsupervised cross-modal hashing. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 21136–21144. AAAI Press, 2026. doi: 10.1609/AAAI.V40I25.39257. URL https://doi.org/10.1609/aaai.v40i25.39257.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In Jian Su, Xavier Carreras, and Kevin Duh (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 457–468. The Association for Computational Linguistics, 2016. doi: 10.18653/V1/D16-1044. URL https://doi.org/10.18653/v1/d16-1044.

Hengtong Hu, Lingxi Xie, Richang Hong, and Qi Tian. Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 3120–3129. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00319. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Hu_Creating_Something_From_Nothing_Unsupervised_Knowledge_Distillation_for_Cross-Modal_Hashing_CVPR_2020_paper.html.

Mark J. Huiskes and Michael S. Lew. The MIR flickr retrieval evaluation. In Michael S. Lew, Alberto Del Bimbo, and Erwin M. Bakker (eds.), *Proceedings of the 1st ACM SIGMM International Conference on Multimedia Information Retrieval, MIR 2008, Vancouver, British Columbia,*

Canada, October 30-31, 2008, pp. 39–43. ACM, 2008. doi: 10.1145/1460096.1460104. URL https://doi.org/10.1145/1460096.1460104.

Kaihang Jiang, Wai Keung Wong, Jianyang Qin, Xiaozhao Fang, Jie Wen, Bingzhi Chen, and Hongbo Gao. Label prediction inherited hashing for cross-modal retrieval: Applying supervised hashing to unsupervised tasks. In Cathal Gurrin, Klaus Schoeffmann, Min Zhang, Luca Rossetto, Stevan Rudinac, Duc-Tien Dang-Nguyen, Wen-Huang Cheng, Phoebe Chen, and Jenny Benois-Pineau (eds.), *Proceedings of the 33rd ACM International Conference on Multimedia, MM 2025, Dublin, Ireland, October 27-31, 2025*, pp. 6343–6352. ACM, 2025. doi: 10.1145/3746027.3755849. URL https://doi.org/10.1145/3746027.3755849.

Mingjin Kuai, Jun Long, and Zhan Yang. Statistical model-driven similarity hashing: Bridging modalities for efficient unsupervised retrieval. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 11977–11986. AAAI Press, 2025. doi: 10.1609/AAAI.V39I11.33304. URL https://doi.org/10.1609/aaai.v39i11.33304.

Jiaxing Li, Lin Jiang, Zeqi Ma, Kaihang Jiang, Xiaozhao Fang, and Jie Wen. Lightweight contrastive distilled hashing for online cross-modal retrieval. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 4779–4787. AAAI Press, 2025a. doi: 10.1609/AAAI.V39I5.32505. URL https://doi.org/10.1609/aaai.v39i5.32505.

Jiaxing Li, Wai Keung Wong, Lin Jiang, Kaihang Jiang, Xiaozhao Fang, Shengli Xie, and Jie Wen. Collaboratively semantic alignment and metric learning for cross-modal hashing. *IEEE Trans. Knowl. Data Eng.*, 37(5):2311–2328, 2025b. doi: 10.1109/TKDE.2025.3537704. URL https://doi.org/10.1109/TKDE.2025.3537704.

Jingzhi Li, Fengling Li, Lei Zhu, Hui Cui, and Jingjing Li. Prototype-guided knowledge transfer for federated unsupervised cross-modal hashing. In Abdulmotaleb El Saddik, Tao Mei, Rita Cucchiara, Marco Bertini, Diana Patricia Tobon Vallejo, Pradeep K. Atrey, and M. Shamim Hossain (eds.), *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pp. 1013–1022. ACM, 2023a. doi: 10.1145/3581783.3613837. URL https://doi.org/10.1145/3581783.3613837.

Yewen Li, Mingyuan Ge, Mingyong Li, Tiansong Li, and Sen Xiang. CLIP-based adaptive graph attention network for large-scale unsupervised multi-modal hashing retrieval. *Sensors*, 23(7): 3439, 2023b. doi: 10.3390/S23073439. URL https://doi.org/10.3390/s23073439.

Ze Li, Tao Yao, Lili Wang, Ying Li, and Gang Wang. Supervised contrastive discrete hashing for cross-modal retrieval. *Knowl. Based Syst.*, 295:111837, 2024. doi: 10.1016/J.KNOSYS.2024.111837. URL https://doi.org/10.1016/j.knosys.2024.111837.

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 Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pp. 13744–13753. AAAI Press, 2024. doi: 10.1609/AAAI.V38I12.29280. URL https://doi.org/10.1609/aaai.v38i12.29280.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. doi: 10.1007/

978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48.

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 Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (eds.), *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pp. 1379–1388. ACM, 2020. doi: 10.1145/3397271.3401086. URL https://doi.org/10.1145/3397271.3401086.

Xiaoqing Liu, Huanqiang Zeng, Yifan Shi, Jianqing Zhu, Kaixiang Yang, and Zhiwen Yu. Ensemble prototype networks for unsupervised cross-modal hashing with cross-task consistency. *IEEE Trans. Multim.*, 27:3476–3488, 2025. doi: 10.1109/TMM.2025.3535378. URL https://doi.org/10.1109/TMM.2025.3535378.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 3967–3976. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00409. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Park_Relational_Knowledge_Distillation_CVPR_2019_paper.html.

Yufeng Shi, Yue Zhao, Xin Liu, Feng Zheng, Weihua Ou, Xinge You, and Qinmu Peng. Deep adaptively-enhanced hashing with discriminative similarity guidance for unsupervised cross-modal retrieval. *IEEE Trans. Circuits Syst. Video Technol.*, 32(10):7255–7268, 2022. doi: 10.1109/TCSVT.2022.3172716. URL https://doi.org/10.1109/TCSVT.2022.3172716.

Shupeng Su, Zhisheng Zhong, and Chao Zhang. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 3027–3035. IEEE, 2019. doi: 10.1109/ICCV.2019.00312. URL https://doi.org/10.1109/ICCV.2019.00312.

Wei Su, Peihan Miao, Huanzhang Dou, Gaoang Wang, Liang Qiao, Zheyang Li, and Xi Li. Language adaptive weight generation for multi-task visual grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 10857–10866. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01045. URL https://doi.org/10.1109/CVPR52729.2023.01045.

Jingdong Wang, Ting Zhang, Jingkuan Song, Nicu Sebe, and Heng Tao Shen. A survey on learning to hash. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):769–790, 2018. doi: 10.1109/TPAMI.2017.2699960. URL https://doi.org/10.1109/TPAMI.2017.2699960.

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. *Int. J. Comput. Vis.*, 132(8):2765–2797, 2024a. doi: 10.1007/S11263-024-02009-7. URL https://doi.org/10.1007/s11263-024-02009-7.

Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. Cross-modal retrieval: A systematic review of methods and future directions. *Proc. IEEE*, 112(11):1716–1754, 2024b. doi: 10.1109/JPROC.2024.3525147. URL https://doi.org/10.1109/JPROC.2024.3525147.

You Wu and Zhixin Li. Mining similarity relationships for unsupervised cross-modal hashing. In *IEEE International Conference on Multimedia and Expo, ICME 2024, Niagara Falls, ON, Canada, July 15-19, 2024*, pp. 1–6. IEEE, 2024. doi: 10.1109/ICME57554.2024.10687927. URL https://doi.org/10.1109/ICME57554.2024.10687927.

Zhan Yang, Xiyin Deng, Lin Guo, and Jun Long. Asymmetric supervised fusion-oriented hashing for cross-modal retrieval. *IEEE Trans. Cybern.*, 54(2):851–864, 2024. doi: 10.1109/TCYB.2023.3241018. URL https://doi.org/10.1109/TCYB.2023.3241018.

Zhan Yang, Binghong Chen, Jiajun Tang, and Yinan Li. Unsupervised similarity-fusion transformer hashing for multimodal retrieval. In Cathal Gurrin, Klaus Schoeffmann, Min Zhang, Luca Rossetto, Stevan Rudinac, Duc-Tien Dang-Nguyen, Wen-Huang Cheng, Phoebe Chen, and Jenny Benois-Pineau (eds.), *Proceedings of the 33rd ACM International Conference on Multimedia, MM 2025, Dublin, Ireland, October 27-31, 2025*, pp. 5932–5941. ACM, 2025. doi: 10.1145/3746027.3754753. URL https://doi.org/10.1145/3746027.3754753.

Jun Yu, Hao Zhou, Yibing Zhan, and Dacheng Tao. Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 4626–4634. AAAI Press, 2021. doi: 10.1609/AAAI.V35I5.16592. URL https://doi.org/10.1609/aaai.v35i5.16592.

Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 1839–1848. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.202. URL https://doi.org/10.1109/ICCV.2017.202.

Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Trans. Neural Networks Learn. Syst.*, 29(12):5947–5959, 2018. doi: 10.1109/TNNLS.2018.2817340. URL https://doi.org/10.1109/TNNLS.2018.2817340.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 41753–41769. PMLR, 2023. URL https://proceedings.mlr.press/v202/zhang23ar.html.

Yuanzhi Zhao, Fan Yang, Yudong Zhao, and Xiaoyu Li. UDCH: unsupervised dynamic weighted cluster-cooperative hashing for cross-modal retrieval. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 13297–13304. AAAI Press, 2026. doi: 10.1609/AAAI.V40I16.38332. URL https://doi.org/10.1609/aaai.v40i16.38332.

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 Trans. Knowl. Data Eng.*, 35(9):8838–8851, 2023. doi: 10.1109/TKDE.2022.3218656. URL https://doi.org/10.1109/TKDE.2022.3218656.

A Additional Method Details

A.1 Derivation of the Normalized Low-Rank Interaction

For the k-th interaction component in Equation 1, the multiplicative interaction can be expanded as:

$$u_{h,k}^{l} = \frac{\mathbf{z}_{i,h}^{l-1}\left(\mathbf{w}_{i,h,k}^{l}(\mathbf{w}_{t,h,k}^{l})^{\top}\right)(\mathbf{z}_{t,h}^{l-1})^{\top}}{\alpha_{i,h}^{l}\alpha_{t,h}^{l}}, \tag{17}$$

where

$$\alpha_{m,h}^{l} = \max\left(\left\|\widetilde{\mathbf{p}}_{m,h}^{l}\right\|_{2}, \varepsilon\right). \tag{18}$$

Here, $\mathbf{w}_{m,h,k}^{l}$ denotes the k-th column of $\mathbf{W}_{m,h}^{l}$. The factor matrix $\mathbf{w}_{i,h,k}^{l}(\mathbf{w}_{t,h,k}^{l})^{\top}$ has rank at most one. Therefore, the numerator of Equation 17 is a rank-one bilinear form, while the sample-dependent normalization makes the complete operation not strictly bilinear.

A.2 Additional Architectural Details of HOI

The interaction features from all heads are concatenated and projected back to the shared hidden space:

$$\begin{aligned} \mathbf{v}^l &= \mathrm{GELU}\left(\mathrm{LN}\left([\mathbf{u}_1^l;\ldots;\mathbf{u}_H^l]\mathbf{W}_o^l + \mathbf{b}_o^l\right)\right), \\ \mathbf{u}^l &= \mathcal{D}_l(\mathbf{v}^l). \end{aligned} \tag{19}$$

Here, $\mathrm{LN}(\cdot)$ denotes layer normalization, and $\mathbf{W}_o^l$ and $\mathbf{b}_o^l$ are the output projection parameters. The operator $\mathcal{D}_l$ denotes DropPath. During training, it independently sets each sample's entire interaction vector to zero with probability p_l and scales retained vectors by $1/(1-p_l)$. The same resulting vector $\mathbf{u}^l$ is used for both modality gates and residual updates.

Stacking L interaction blocks produces the representation sequence $\{\mathbf{z}_m^0, \mathbf{z}_m^1, \ldots, \mathbf{z}_m^L\}$ for each modality $m \in \{i, t\}$. Here, $\mathbf{z}_m^0$ is the initial representation, and $\mathbf{z}_m^l$ is the output of the l-th interaction block. To retain access to both the initial features and the outputs of all interaction stages, we concatenate these representations and pass them through a shared aggregation network:

$$\mathbf{z}_m^* = \mathrm{Agg}\left([\mathbf{z}_m^0; \mathbf{z}_m^1; \ldots; \mathbf{z}_m^L]\right), \qquad m \in \{i, t\}. \tag{20}$$

The aggregation network shares parameters between the image and text branches and processes their representation sequences separately, producing $\mathbf{z}_i^*$ and $\mathbf{z}_t^*$ for adaptive modality fusion.

A.3 Pseudo-Similarity Construction

Given the initial image representation matrix $\mathbf{Z}_i^0$ and text representation matrix $\mathbf{Z}_t^0$ in a mini-batch, we compute the intra-modal similarity matrices as

$$\mathbf{A}_i = \mathrm{Cos}\left(\mathbf{Z}_i^0, \mathbf{Z}_i^0\right), \quad \mathbf{A}_t = \mathrm{Cos}\left(\mathbf{Z}_t^0, \mathbf{Z}_t^0\right), \tag{21}$$

where $\mathrm{Cos}(\cdot,\cdot)$ denotes the cosine similarity matrix defined in Section 3.4. Then, high-confidence positive relations are selected according to the quantile-based thresholds of the two similarity matrices:

$$S_{mn} = \begin{cases} 1, & m = n, \\ \mathbb{I}(A_{i,mn} > \delta_i)\mathbb{I}(A_{t,mn} > \delta_t), & m \neq n. \end{cases} \tag{22}$$

Here, $A_{i,mn}$ and $A_{t,mn}$ denote the (m,n)-th entries of $\mathbf{A}_i$ and $\mathbf{A}_t$, respectively, and $\mathbb{I}(\cdot)$ is the indicator function. The thresholds δ_i and δ_t are recomputed for each mini-batch as the q-quantiles of all B^2 entries of their respective similarity matrices, including the diagonal entries. The resulting binary matrix is used as a supervision target without propagating gradients through its construction. Setting $S_{mm} = 1$ retains the matched image-text pair as a positive cross-modal relation and the self-relation as a positive intra-modal relation.

A.4 Expanded Ranking Objectives

Using the teacher similarity matrices defined in Section 3.4, we impose ranking constraints on both cross-modal and intra-modal relationships:

$$\begin{aligned} \mathcal{L}_{\mathrm{cross}}^T &= \frac{1}{2}\left[\mathcal{L}_{\mathrm{rank}}(\mathbf{S}_{it}^T; \gamma_{\mathrm{cross}}^T, \tau_{\mathrm{cross}}) + \mathcal{L}_{\mathrm{rank}}((\mathbf{S}_{it}^T)^\top; \gamma_{\mathrm{cross}}^T, \tau_{\mathrm{cross}})\right], \\ \mathcal{L}_{\mathrm{intra}}^T &= \frac{1}{2}\left[\mathcal{L}_{\mathrm{rank}}(\mathbf{S}_{ii}^T; \gamma_{\mathrm{intra}}^T, \tau_{\mathrm{intra}}) + \mathcal{L}_{\mathrm{rank}}(\mathbf{S}_{tt}^T; \gamma_{\mathrm{intra}}^T, \tau_{\mathrm{intra}})\right]. \end{aligned} \tag{23}$$

For the modality-specific student branches, relational distillation is combined with their own cross-modal ranking constraint:

$$\mathcal{L}_{\mathrm{cross}}^S = \frac{1}{2}\left[\mathcal{L}_{\mathrm{rank}}(\mathbf{S}_{it}^S; \gamma_{\mathrm{cross}}^S, \tau_{\mathrm{cross}}) + \mathcal{L}_{\mathrm{rank}}((\mathbf{S}_{it}^S)^\top; \gamma_{\mathrm{cross}}^S, \tau_{\mathrm{cross}})\right]. \tag{24}$$

A.5 Complete Hash Regularization

For branch $X \in \{T, S\}$, the complete hash regularization is

$$\begin{aligned}\mathcal{R}^X = &\frac{\lambda_a^X}{BK}\left\|\mathbf{H}_i^X - \mathbf{H}_t^X\right\|_F^2 \\ &+ \frac{\lambda_q^X}{BK}\left(\left\|\mathbf{H}_i^X - \operatorname{sign}(\mathbf{H}_i^X)\right\|_F^2 + \left\|\mathbf{H}_t^X - \operatorname{sign}(\mathbf{H}_t^X)\right\|_F^2\right) \\ &+ \frac{\lambda_b^X}{K}\left(\left\|\frac{1}{B}\sum_{b=1}^{B}\mathbf{h}_{i,b}^X\right\|_2^2 + \left\|\frac{1}{B}\sum_{b=1}^{B}\mathbf{h}_{t,b}^X\right\|_2^2\right), \qquad X \in \{T, S\}.\end{aligned} \tag{25}$$

Here, B and K denote the mini-batch size and hash-code length. The three terms encourage paired cross-modal consistency, reduce quantization error, and prevent imbalanced hash bits, respectively.

B Experimental Details

B.1 Datasets and Protocol

We evaluate **HOCH** on three widely used multimodal benchmarks: MIRFlickr, NUS-WIDE, and MSCOCO. MIRFlickr contains Flickr images with textual tags, from which 24 semantic categories are used. For NUS-WIDE, we adopt the subset associated with the 10 most frequent categories. MSCOCO contains natural images and textual descriptions covering 80 semantic categories. The detailed data partitions and statistics are summarized in Table 4.

For all datasets, we use pre-extracted 512-dimensional CLIP image and text features, which remain fixed during **HOCH** training. Two modality-specific encoders subsequently map these features into a shared 512-dimensional latent space. The semantic labels are used only to determine retrieval relevance during evaluation and are not involved in training. Instead, pseudo-semantic relationships are constructed from the high-confidence agreement between the initial image and text similarities. For **HOCH**, all experiments use the data partitions summarized in Table 4 and the same pre-extracted CLIP features across different hash-code lengths.

Table 4: Dataset statistics.

Dataset	Train	Query	Retr.	Dim. (I/T)	Labels
MIRFlickr	5,000	2,000	18,015	512/512	24
NUS-WIDE	5,000	2,000	184,577	512/512	10
MSCOCO	5,000	2,000	121,287	512/512	80

B.2 Evaluation Metrics

This paper evaluates cross-modal retrieval performance in two directions: image-to-text (I2T) and text-to-image (T2I). For a given query sample, a database sample is considered relevant if it shares at least one semantic label with the query. We adopt mAP@50 as the primary evaluation metric and further report mAP@K curves at different retrieval depths. For each query, database samples are ranked according to the Hamming distances between their binary hash codes. AP@K is computed as the mean of the precision values at the ranks occupied by relevant samples within the top K results, and queries without a relevant sample in the top K contribute zero. mAP@K is obtained by averaging AP@K over all queries. Higher mAP@K indicates better retrieval performance.

B.3 Implementation Details

We implement **HOCH** based on PyTorch, and all experiments are conducted on a single NVIDIA GeForce RTX 4090 D GPU. The learning rates for the base modality encoder and the remaining network parameters are set to 1×10^{-5} and 1×10^{-4}, respectively. The weight decay coefficient

is set to 1×10^{-4}, and a cosine annealing strategy is used to adjust the learning rate. The feature-preservation coefficient λ_{fp} is set to 0.03, 0.10, and 0.03 for MIRFlickr, NUS-WIDE, and MSCOCO, respectively. All experiments follow a unified training protocol.

For the efficiency evaluation, training-step latency is measured over 100 steps after 20 warm-up steps with a batch size of 128; inference latency is measured over 500 steps with a batch size of 1. Measurements exclude CLIP feature extraction, data loading, evaluation, and Hamming ranking. Active parameter counts include only modules executed in the measured forward path.

C Additional Experimental Results

C.1 Detailed Retrieval Results

Table 1 presents the mAP@50 comparison between **HOCH** and the representative baselines. The detailed observations are as follows:

- On MIRFlickr, **HOCH** achieves consistently strong performance across different hash code lengths. Its average mAP@50 values over the four code lengths reach approximately 0.924 and 0.896 in the I2T and T2I directions, respectively, outperforming the strongest baseline averages of VTM-UCH in I2T (0.922) and DAEH in T2I (0.891). Except for the 16-bit I2T setting, **HOCH** obtains the best result in all remaining settings. In particular, its I2T performance increases from 0.922 at 32 bits to 0.942 at 128 bits, while its T2I performance improves from 0.881 to 0.907.
- On NUS-WIDE, **HOCH** achieves the best I2T results under the 16-, 32-, and 64-bit settings, with mAP@50 values of 0.806, 0.831, and 0.837, respectively. At 128 bits, it remains competitive with JDSH, obtaining 0.845 compared with 0.847. For T2I retrieval, **HOCH** performs best at 16 bits with a score of 0.784 and achieves a competitive result of 0.803 at 32 bits, although DAEH and JDSH perform better under some longer-code settings.
- On MSCOCO, **HOCH** achieves the best performance in all eight combinations of retrieval direction and code length. Specifically, its mAP@50 increases from 0.881 to 0.944 for I2T and from 0.883 to 0.956 for T2I, demonstrating a clear advantage over the compared methods on this dataset.
- As the hash code length increases from 16 to 128 bits, the mAP@50 of **HOCH** improves consistently across all three datasets and both retrieval directions. This trend indicates that longer hash codes enable the model to preserve more discriminative cross-modal information.

Figure 2 further compares the retrieval performance at different retrieval depths using 32-bit hash codes. **HOCH** maintains consistently competitive mAP@K values across a broad range of K on MIRFlickr and NUS-WIDE, indicating that its performance is not limited to the mAP@50 setting. On MSCOCO, **HOCH** exhibits a clear advantage over the compared methods in both retrieval directions as K increases. Although the mAP@K values generally decrease when more database samples are retrieved, **HOCH** shows relatively stable performance, demonstrating its effectiveness under different retrieval depths.

C.2 Visualization of Cross-Modal Alignment

To qualitatively examine the cross-modal structure of the final hash representations, we visualize 64-bit query hash codes on NUS-WIDE using t-SNE. All methods use the same 2,000 query pairs and identical visualization settings, with image and text codes embedded jointly for each method. As shown in Figure 3, SMSH and VTM-UCH exhibit regions dominated by one modality, while **HOCH** shows greater overlap between image and text codes across multiple local regions of the two-dimensional projection. For **HOCH**, these codes are generated independently by the image and text student branches. The observed overlap is consistent with the goal of learning compatible unimodal hash representations through multimodal training, providing a qualitative complement to the quantitative retrieval results.

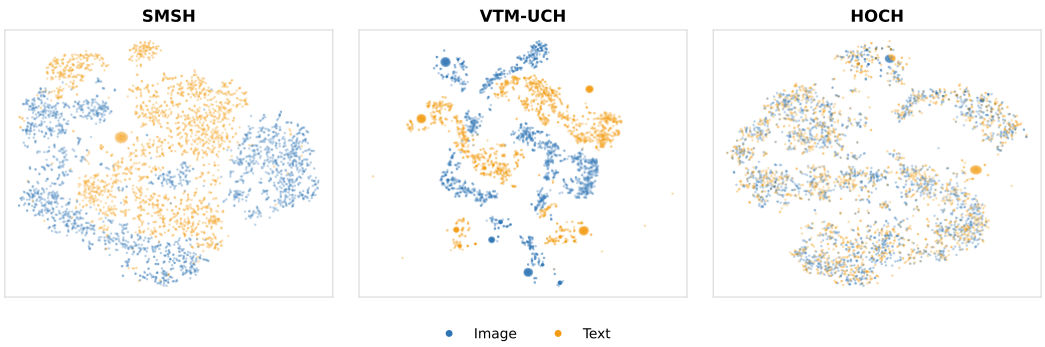


Figure 3: t-SNE visualization of 64-bit query hash codes on NUS-WIDE. All methods use the same 2,000 query pairs and identical visualization settings. **HOCH** exhibits greater overlap between image and text codes than the selected baselines.

C.3 Parameter Sensitivity Analysis

To investigate the influence of different hyperparameters, we vary the low-rank dimension r, teacher-loss scale κ_T, relational distillation weight λ_{d}, and feature-preservation scale κ_{FP} individually, while fixing the remaining parameters to their default values. All experiments are conducted on MIRFlickr, NUS-WIDE, and MSCOCO using 64-bit hash codes. The corresponding I2T and T2I $\mathrm{mAP@50}$ results are presented in Figure 4. The feature-preservation scale κ_{FP} multiplies the dataset-specific default value of λ_{fp} reported in Appendix B.3.

As shown in Figure 4, the performance varies only moderately over the tested parameter ranges, demonstrating the overall robustness of **HOCH**. Increasing r does not produce monotonic improvements. For κ_T, intermediate values provide relatively stable performance by balancing the optimization of the multimodal teacher and unimodal students. Regarding λ_{d}, moderate values, particularly within $[0.4, 0.8]$, generally yield competitive performance, whereas an excessively large distillation weight may overconstrain the student representations and degrade retrieval accuracy. In addition, using a positive κ_{FP} generally outperforms disabling feature preservation, while an overly strong constraint provides no consistent benefit.

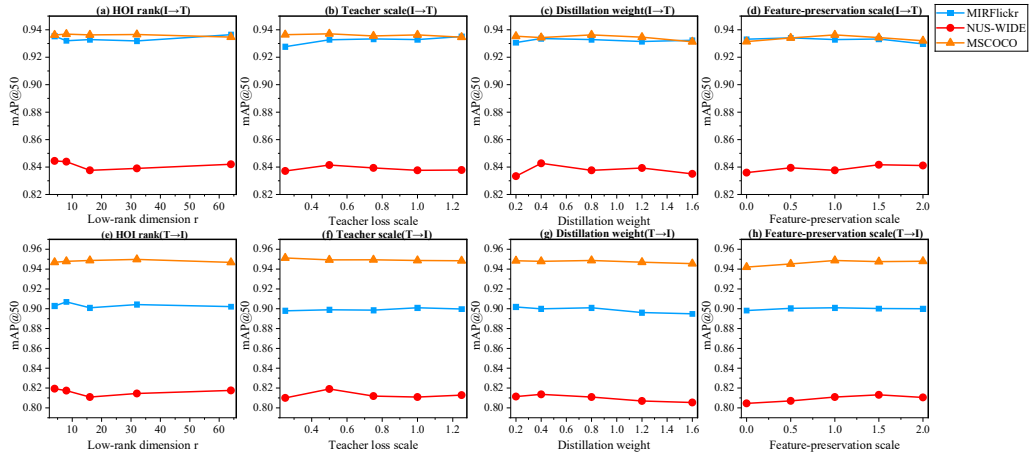
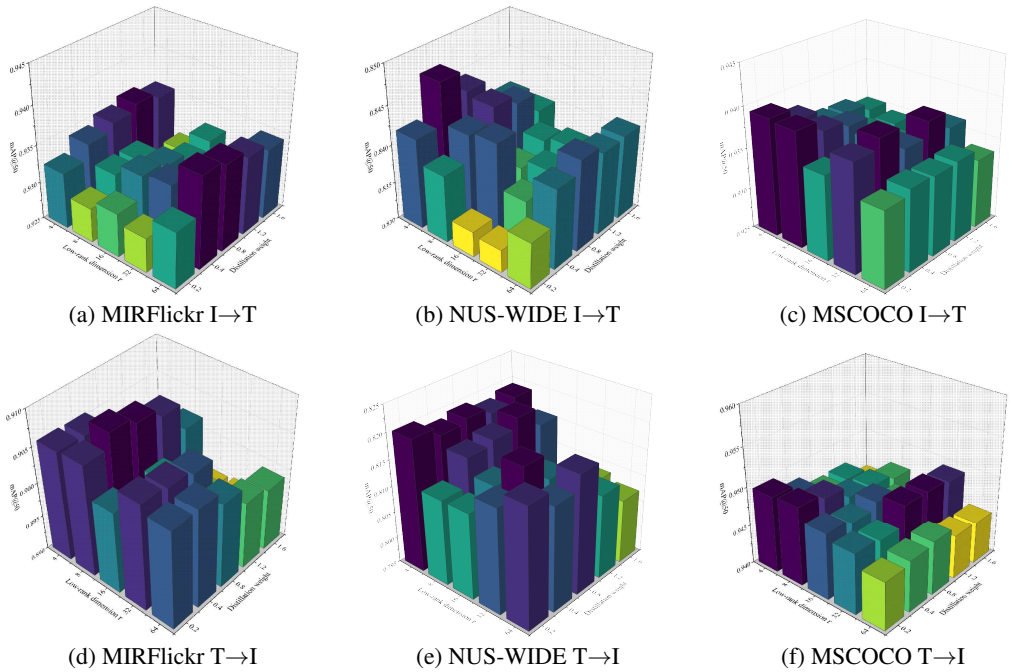


Figure 4: Sensitivity analysis of the low-rank dimension r, teacher-loss scale κ_T, relational distillation weight λ_{d}, and feature-preservation scale κ_{FP} on the three datasets at 64 bits.



(a) MIRFlickr I→T (b) NUS-WIDE I→T (c) MSCOCO I→T

(d) MIRFlickr T→I (e) NUS-WIDE T→I (f) MSCOCO T→I

Figure 5: Joint sensitivity analysis of the low-rank dimension r and the relational distillation weight λ_d with 64-bit hash codes. (a) MIRFlickr I2T; (b) NUS-WIDE I2T; (c) MSCOCO I2T; (d) MIRFlickr T2I; (e) NUS-WIDE T2I; and (f) MSCOCO T2I.

Overall, instead of selecting a separate optimum for each dataset or retrieval direction, we adopt the unified configuration $r = 16$, $\kappa_T = 1.0$, $\lambda_{\mathrm{d}} = 0.8$, and $\kappa_{\mathrm{FP}} = 1.0$ in all main experiments. To further investigate the interaction between the low-rank dimension r in HOI and the relational distillation weight λ_d, we conduct a joint parameter analysis on the three datasets with the hash-code length fixed to 64 bits.

As shown in Figure 5, the retrieval performance varies moderately over a broad range of parameter combinations, indicating that **HOCH** is generally robust to the joint variation of r and λ_d. The best-performing parameter settings differ across datasets, which may reflect differences in their semantic distributions and modality characteristics. An appropriate distillation weight facilitates the transfer of relational knowledge from the multimodal teacher to the unimodal students, whereas an overly weak or strong weight may limit the transfer effectiveness or over-constrain student learning. At $r = 16$ on NUS-WIDE, increasing λ_d from 0.8 to 1.6 reduces T2I mAP@50 from 0.8109 to 0.8053, while the I2T scores vary non-monotonically over the same range. This direction-dependent response shows why both retrieval directions should be considered when assessing a shared parameter setting.

C.4 Detailed Ablation Analysis

To analyze the contributions of the three core components and the feature-preservation regularization in **HOCH**, we construct the following variants: **HOCH**-FirstOrder replaces the multiplicative interaction in each HOI block with an additive first-order interaction; **HOCH**-MeanFusion replaces the sample-adaptive modality weights with fixed equal weights of 0.5; **HOCH**-w/o MSRD sets the weight λ_d of the relational distillation loss to zero; and **HOCH**-w/o FP removes the feature preservation regularization by setting λ_{fp} to zero. Except for the modified component, all other network structures and training settings remain identical to those of the full model. Table 2 reports the bidirectional average mAP@50 results on the three datasets using 64-bit hash codes. Based on the experimental results, we draw the following conclusions:

- **HOCH-FirstOrder**: Replacing the multiplicative interaction in *HOI* with an additive first-order interaction decreases the overall average mAP@50 from 0.8945 to 0.8917, corre-

sponding to a drop of approximately 0.0028. Specifically, the performance decreases by 0.0023, 0.0037, and 0.0025 on MIRFlickr, NUS-WIDE, and MSCOCO, respectively.

- **HOCH**-**MeanFusion**: When *AMF* is replaced with fixed modality weights of 0.5/0.5, the overall average mAP@50 decreases from 0.8945 to 0.8933, representing a drop of approximately 0.0012. Although the improvement brought by adaptive fusion is relatively modest, the full model outperforms MeanFusion on all three datasets. In particular, the improvement is more apparent on MSCOCO, where the mAP@50 increases from 0.9401 to 0.9425.
- **HOCH**-w/o **MSRD**: Removing *MSRD* lowers the overall average mAP@50 from 0.8945 to 0.8839, corresponding to the largest performance decrease of approximately 0.0106 among all variants. In particular, the performance drops by 0.0149 on MIRFlickr and 0.0152 on NUS-WIDE, while a smaller reduction of 0.0019 is observed on MSCOCO. This indicates that transferring relational structures learned by the multimodal teacher to the unimodal student branches is particularly beneficial on MIRFlickr and NUS-WIDE. Overall, *MSRD* provides the largest contribution among the evaluated components.