ADAPTIVE RELATION WEIGHTING AND VIEW CURRICULUM FOR SUPERVISED MULTI-LABEL CROSS-MODAL HASHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-modal hashing maps heterogeneous data into compact binary codes for efficient retrieval. Existing methods mainly assume fixed image/text input configurations, while practical systems may need to jointly handle paired, image-only, and text-only inputs. This unified setting introduces heterogeneous and dynamically changing optimization difficulty across the nine directed relations among these input conditions. To address this issue, we propose a unified supervised multi-label cross-modal hashing framework that couples adaptive directed relation weighting with an adaptive view curriculum. Relation weights are estimated from raw contrastive statistics to calibrate multi-label contrastive learning, while the learned relation states further guide the sampling of paired, image-only, and text-only views for subsequent training. Experiments on COCO, NUS-WIDE, and MIRFlickr25K demonstrate the effectiveness of the proposed framework across multiple hash-code lengths and eight directed retrieval protocols.

1 INTRODUCTION

Cross-modal hashing has received increasing attention for large-scale multimedia retrieval because it maps heterogeneous data into compact binary codes, enabling efficient storage and similarity search. Early approaches learned modality-specific hash functions under semantic supervision, while deep hashing methods further integrated representation learning and hash-code optimization Mandal et al. (2017); Jiang & Li (2017). More recently, vision–language pretraining has provided stronger cross-modal representations, and pretrained models such as CLIP have been introduced into cross-modal hashing to improve semantic alignment Radford et al. (2021); Xia et al. (2023); Zou et al. (2025).

Despite these advances, existing cross-modal hashing methods are generally developed under fixed image/text input configurations. In practical retrieval systems, however, both modalities may be available for some inputs, while only an image or only text may be available for others. We therefore consider three modality-conditioned input states: paired (P), image-only (I), and text-only (T), and aim to map them into a common hash space using a shared hashing function. This unified setting supports diverse query–database combinations rather than restricting retrieval to conventional image-to-text and text-to-image protocols.

Joint learning of these heterogeneous input states introduces two coupled optimization challenges. First, P, I, and T induce nine directed relations whose learning difficulties can differ because paired and unimodal inputs contain different amounts and types of modality information. Treating all relations uniformly can therefore lead to imbalanced relation optimization. Second, the relative difficulty of the three input states may change during training, making a fixed sampling distribution unable to adapt when the optimization bottleneck shifts. Moreover, these two problems are closely related: view sampling determines which directed relations are observed during training, while the learned relation states provide useful signals for determining which views should receive more attention.

To address these challenges, we propose a unified supervised multi-label cross-modal hashing framework with adaptive directed relation weighting and an adaptive view curriculum. Multi-label supervision provides semantic positive relations beyond original image–text pairs Khosla et al. (2020); Zhang & Wu (2024), allowing us to estimate the learning states of the nine directed relations from

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.

in-batch contrastive statistics. Specifically, we maintain smoothed relation difficulties from detached raw relation losses and use them to dynamically calibrate relation-conditioned multi-label contrastive learning. The resulting relation states are further aggregated into modality-level difficulty signals, which adapt the sampling distribution over $P/I/T$ views for subsequent training. Thus, relation weighting and view sampling form a closed loop in which the current relation state affects future view construction, and the updated view distribution in turn determines subsequent relation learning.

Our contributions are summarized as follows:

- We establish a unified supervised multi-label cross-modal hashing framework for heterogeneous input conditions, in which paired, image-only, and text-only inputs are accommodated within a single retrieval setting. The resulting formulation unifies eight directed retrieval tasks that would otherwise be treated under different input conditions, providing a unified way to address diverse cross-modal retrieval scenarios.
- To solve the resulting unified retrieval problem, we develop an adaptive hash learning framework that accounts for the heterogeneous and dynamically changing learning states among different input relations. A directed relation weighting mechanism adaptively regulates relation optimization, while an adaptive view curriculum uses the learned relation state to adjust the sampling of paired, image-only, and text-only views. Together, they form a coupled adaptive training process.
- Extensive experiments on COCO, NUS-WIDE, and MIRFlickr25K across multiple hash-code lengths and eight directed retrieval tasks demonstrate the effectiveness of the proposed framework. Ablation and analysis results further validate the proposed adaptive learning mechanism.

2 Related Work

2.1 Deep Hashing

With the development of deep neural networks, deep hashing methods jointly optimize feature representation and binary-code learning in an end-to-end manner. DCMH jointly learns modality-specific deep representations and hash codes under pairwise semantic supervision Jiang & Li (2017), while DJSRH reconstructs joint semantic relations from multimodal data to guide unsupervised cross-modal hash learning Su et al. (2019).

More recent deep hashing methods further exploit contrastive learning and Transformer architectures. HARR improves unsupervised image hashing by reconstructing reliable semantic similarity structures and reducing redundancy among hash bits Ma et al. (2024). FSCH introduces fine-grained similarity-preserving contrastive learning to better preserve semantic structures Cao et al. (2024), while CTMIH combines Vision Transformer representations, cross-view contrastive learning, and masked reconstruction for robust image hashing Shen et al. (2024). These deep approaches substantially improve representation learning over shallow hashing, but they generally optimize hash representations under predefined input organizations or retrieval settings.

2.2 Pretrained and Foundation-Model-Based Hashing

Recent advances in large-scale pretraining have introduced stronger semantic representations into hashing-based retrieval. IPHash leverages a pretrained Vision Transformer and information-preserving objectives to reduce semantic information loss during binary compression Gong et al. (2022). For cross-modal hashing, UCMFH exploits pretrained CLIP visual and textual representations together with multimodal fusion and contrastive hash learning Xia et al. (2023). PromptHash introduces affinity-prompted collaborative learning to enhance semantic preservation and cross-modal interaction Zou et al. (2025). More recently, CroVCA investigates hash learning from foundation-model embeddings through cross-view code alignment, coding-rate regularization, and a lightweight HashCoder Moummad et al. (2026).

These studies demonstrate the potential of pretrained and foundation models for learning compact binary representations. However, existing methods mainly focus on fixed input configurations or

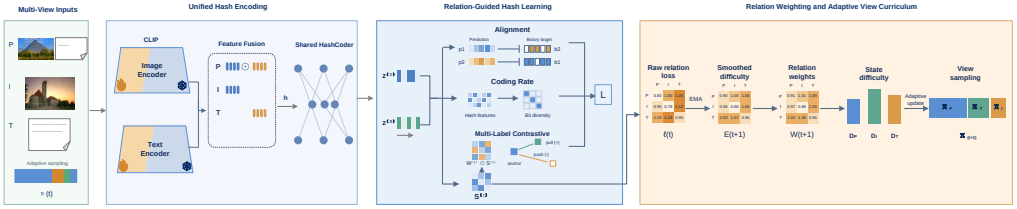


Figure 1: Overview of the proposed framework.

conventional retrieval relations. In contrast, we consider paired, image-only, and text-only inputs within a unified supervised multi-label hashing setting. Their joint learning induces nine directed relations with heterogeneous and dynamically changing learning states, which motivates our adaptive relation weighting and view curriculum.

3 METHOD

3.1 NOTATIONS AND PROBLEM DEFINITION

Let $\mathcal{D} = \{(x_i^I, x_i^T, \mathbf{y}_i)\}_{i=1}^N$ be a supervised multi-label dataset of N complete image–text records, where x_i^I and x_i^T denote the image and text of record i, respectively, and $\mathbf{y}_i \in \{0,1\}^C$ is its label vector over C semantic categories.

We consider three modality-conditioned input states: paired (P), image-only (I), and text-only (T). Our goal is to learn a common hashing function that maps all three input states into a shared b-bit hash space while preserving multi-label semantic relevance. The joint learning of these input states induces nine directed relations $u \to v$, where $u, v \in \mathcal{M} = \{P, I, T\}$. Since different input states contain different amounts and types of modality information, these directed relations may exhibit heterogeneous and dynamically changing learning difficulty.

3.2 MODEL FORMULATION

To address supervised multi-label cross-modal hashing under heterogeneous input conditions, we propose a unified adaptive hashing framework that jointly handles paired, image-only, and text-only inputs within a shared hash space. The framework employs CLIP to encode visual and textual information and a shared HashCoder to generate compact binary representations for different modality-conditioned inputs. We present the proposed framework from four aspects: modality-conditioned view encoding, cross-view multi-label hash learning, adaptive directed relation weighting, and adaptive view curriculum. An overview of the proposed framework is illustrated in Figure 1.

3.2.1 MODALITY-CONDITIONED VIEW ENCODING

For each record i, we construct two stochastic branches indexed by $k \in \{1, 2\}$. Let t index the interval between two consecutive controller updates. During interval t, each branch independently samples a paired, image-only, or text-only view according to $\pi_P^{(t)}$, $\pi_I^{(t)}$, and $\pi_T^{(t)}$, respectively, where $\pi_P^{(t)} + \pi_I^{(t)} + \pi_T^{(t)} = 1$. Whenever a branch contains an image, it uses a branch-specific augmentation of the same source image. Let $\mathbf{v}_i^{(k)}$ and $\mathbf{t}_i$ denote the CLIP visual and textual embeddings, respectively. The branch representation is defined as

$$\mathbf{h}_i^{(k)} = \begin{cases} \mathbf{v}_i^{(k)} \odot \mathbf{t}_i, & \text{paired input,} \\ \mathbf{v}_i^{(k)}, & \text{image-only input,} \\ \mathbf{t}_i, & \text{text-only input,} \end{cases} \tag{1}$$

where $\odot$ denotes element-wise multiplication. We use this parameter-free fusion because it preserves the original embedding dimension, allowing paired, image-only, and text-only representations to share the same HashCoder while still introducing multiplicative cross-modal interactions. In contrast, concatenation changes the feature dimension and requires an additional projection layer.

Since two text-only branches would contain identical textual embeddings, text–text conflicts within a mini-batch are reassigned as evenly as possible to text-only/image-only and image-only/text-only branch pairs. This avoids trivial identical-text positives while preserving directional balance.

A shared HashCoder ψ_θ maps each branch representation to a b-dimensional hash logit. The corresponding bit probabilities and binary codes are

$$\begin{aligned} \mathbf{z}_i^{(k)} &= \psi_\theta\left(\mathbf{h}_i^{(k)}\right) \in \mathbb{R}^b, \\ \mathbf{p}_i^{(k)} &= \sigma(\mathbf{z}_i^{(k)}) \in [0,1]^b, \\ \mathbf{b}_i^{(k)} &= \mathbf{1}\left[\mathbf{p}_i^{(k)} \geq 0.5\right] \in \{0,1\}^b. \end{aligned} \tag{2}$$

Here, $\mathbf{z}_i^{(k)}$, $\mathbf{p}_i^{(k)}$, and $\mathbf{b}_i^{(k)}$ denote the hash logits, bit probabilities, and binary code of record i in branch k, respectively. The two branches and all three input states share the same HashCoder.

3.2.2 Cross-View Multi-Label Hash Learning

This subsection defines the three differentiable losses used to optimize the trainable CLIP LoRA parameters and the shared HashCoder.

Cross-view code alignment. The binary code produced by one branch serves as a fixed target for the bit probabilities predicted by the other branch. The two directions are evaluated symmetrically:

$$\mathcal{L}_{\text{align}} = \frac{1}{2B}\sum_{i=1}^{B}\left[\text{BCE}\left(\mathbf{b}_i^{(1)}, \mathbf{p}_i^{(2)}\right) + \text{BCE}\left(\mathbf{b}_i^{(2)}, \mathbf{p}_i^{(1)}\right)\right]. \tag{3}$$

Here, $\text{BCE}(\mathbf{b}, \mathbf{p})$ denotes binary cross-entropy between a fixed binary target $\mathbf{b}$ and predicted bit probabilities $\mathbf{p}$, averaged over the b hash bits.

Coding-rate regularization. A diverse hash space requires the marginal distribution of binary codes to have high entropy, encouraging balanced and decorrelated bits. Since directly computing the entropy of high-dimensional discrete codes is intractable, we use a continuous surrogate based on the pre-threshold hash logits.

For notational simplicity, consider either branch and omit the branch superscript. Given a batch of B records, we normalize the hash logits and construct their second-moment matrix as

$$\overline{\mathbf{z}}_i = \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|_2}, \qquad \mathbf{C} = \frac{1}{B}\sum_{i=1}^{B}\overline{\mathbf{z}}_i\overline{\mathbf{z}}_i^\top. \tag{4}$$

Here, $\mathbf{z}_i \in \mathbb{R}^b$, where b is both the hash-logit dimension and the target hash-code length, and $\mathbf{C} \in \mathbb{R}^{b\times b}$ is the batch second-moment matrix.

Modeling the normalized logits as samples from a zero-mean Gaussian distribution, maximizing the log-determinant of their second-order statistics spreads the representations along independent directions. The coding-rate surrogate and the corresponding diversity loss are

$$\mathcal{R}(\mathbf{C}) = \frac{1}{2}\log\det\left(\mathbf{I}_b + \frac{b}{B}\mathbf{C}\right), \qquad \mathcal{L}_{\text{cr}} = -\mathcal{R}(\mathbf{C}). \tag{5}$$

The same regularizer is applied to both branches during training. This coding-rate formulation is inherited from CroVCA (Moummad et al., 2026).

Relation-conditioned multi-label contrastive learning. Following MulSupCon Zhang & Wu (2024), each active label of an anchor view defines a separate contrastive row. After concatenating the two branches, let $\mathbf{z}_a$, $\mathbf{y}_a$, and $\xi_a \in \{P, I, T\}$ denote the hash logit, inherited multi-label vector, and input type of view $a \in \{1, \ldots, 2B\}$, respectively. We define

$$\overline{\mathbf{z}}_a = \frac{\mathbf{z}_a}{\|\mathbf{z}_a\|_2}, \qquad s_{ar} = \overline{\mathbf{z}}_a^\top\overline{\mathbf{z}}_r, \tag{6}$$

where r indexes a generic reference view in the contrastive softmax.

For each active anchor label c satisfying $y_{a,c} = 1$, we define

$$\mathcal{J}^+_{a,c} = \{j \neq a : y_{j,c} = 1\}. \tag{7}$$

The original MulSupCon objective does not distinguish anchor–reference pairs according to their input types. In our setting, paired, image-only, and text-only views induce nine ordered relations whose learning behavior can differ across both relation types and directions. We therefore introduce $\mathbf{W}^{(t)} \in \mathbb{R}^{3\times3}_+$ and replace each similarity s_{ar} with the directed relation-calibrated similarity $W^{(t)}_{\xi_a,\xi_r} s_{ar}$. The resulting loss is

$$\mathcal{L}_{\text{mlc}} = -\frac{1}{\sum_{a=1}^{2B} \|\mathbf{y}_a\|_0} \sum_{a=1}^{2B} \sum_{\substack{c: y_{a,c}=1 \\ |\mathcal{J}^+_{a,c}|>0}} \frac{1}{|\mathcal{J}^+_{a,c}|} \sum_{j\in\mathcal{J}^+_{a,c}} \log \frac{\exp(W^{(t)}_{\xi_a,\xi_j} s_{aj}/\tau)}{\sum_{r\neq a} \exp(W^{(t)}_{\xi_a,\xi_r} s_{ar}/\tau)}. \tag{8}$$

Compared with the original MulSupCon objective, the change is confined to this similarity calibration inside the softmax. The adopted configuration does not use the additional per-anchor factor $1/\|\mathbf{y}_a\|_0$. The calibrated similarities optimize the neural network, whereas detached raw similarities estimate the relation matrix for the next relation-update interval.

Overall differentiable objective. The neural parameters are optimized with

$$\mathcal{L} = \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{cr}}\mathcal{L}_{\text{cr}} + \lambda_{\text{mlc}}\mathcal{L}_{\text{mlc}}. \tag{9}$$

3.2.3 Adaptive Directed Relation Weighting

Paired, image-only, and text-only views form nine directed relations in the shared hash space, whose optimization difficulties may differ. Let $\mathcal{M} = \{P, I, T\}$ denote the three input types. For each relation $u \to v$, we use $\ell^{(t)}_{uv}$ to denote its mean positive contrastive loss observed in relation-update interval t, computed from the raw similarities s_{aj} before relation weighting. Following MulSupCon, positive terms induced by different shared labels are counted independently. We use these relation-specific losses to adaptively calibrate the nine directed relations.

To reduce short-term fluctuations, we maintain a smoothed historical loss $E^{(t)}_{uv}$ for each directed relation and update it through an exponential moving average:

$$E^{(t+1)}_{uv} = \mu_W E^{(t)}_{uv} + (1-\mu_W)\ell^{(t)}_{uv}, \tag{10}$$

where $\mu_W \in [0, 1)$ controls temporal smoothing. For a newly observed relation, E_{uv} is initialized using its current interval loss. If an initialized relation is absent from the current interval, its previous state is retained.

After updating the historical relation losses, let $E^{(t+1)}_{\text{avg}}$ denote their average over all initialized directed relations. We measure the relative difficulty of relation $u \to v$ by its normalized deviation from this average:

$$\delta^{(t+1)}_{uv} = \frac{E^{(t+1)}_{uv} - E^{(t+1)}_{\text{avg}}}{E^{(t+1)}_{\text{avg}}}. \tag{11}$$

A positive δ_{uv} indicates that the relation is currently more difficult than the average, whereas a negative value indicates that it is relatively easier.

Accordingly, the relation weight used in the next interval is updated as

$$W^{(t+1)}_{uv} = 1 + \beta\delta^{(t+1)}_{uv}. \tag{12}$$

where 1 denotes the neutral weight and β controls the adjustment strength. The resulting weight is bounded within $[w_{\min}, w_{\max}]$ to avoid excessively large or small values.

3.2.4 ADAPTIVE VIEW CURRICULUM

Curriculum learning organizes training according to a predefined or dynamically adjusted learning progression Bengio et al. (2009), while self-paced learning adapts sample selection according to the current learning state of the model Kumar et al. (2010). Inspired by this adaptive training principle, we use the learned relation state to dynamically adjust the distribution of modality-conditioned views.Since each input type participates in the contrastive objective both as an anchor and as a reference, we characterize the difficulty of input type u by jointly considering its outgoing and incoming relations:

$$D_u^{(t+1)} = \frac{1}{2}\left(\frac{\sum_{v\in\mathcal{M}} W_{uv}^{(t+1)} E_{uv}^{(t+1)}}{\sum_{v\in\mathcal{M}} W_{uv}^{(t+1)}} + \frac{\sum_{v\in\mathcal{M}} W_{vu}^{(t+1)} E_{vu}^{(t+1)}}{\sum_{v\in\mathcal{M}} W_{vu}^{(t+1)}}\right). \tag{13}$$

The normalization in both directions prevents the absolute scale of $\mathbf{W}$ from directly inflating the estimated difficulty.

To capture the historical difficulty of each input type, we maintain an EMA baseline $B_u^{(t)}$, which is updated as

$$B_u^{(t+1)} = \mu_D B_u^{(t)} + (1-\mu_D) D_u^{(t+1)}, \tag{14}$$

where $\mu_D \in [0,1)$ controls temporal smoothing. At the first valid curriculum update, B_u is initialized with the current difficulty D_u.

Based on the historical baseline, the relative difficulty change of state u is defined as

$$R_u^{(t+1)} = \frac{D_u^{(t+1)} - B_u^{(t)}}{|B_u^{(t)}| + \varepsilon}. \tag{15}$$

A positive $R_u^{(t+1)}$ indicates that state u becomes harder relative to its historical level, whereas a negative value indicates that it becomes relatively easier.

To compare the relative difficulty changes between paired and unimodal views while suppressing short-term fluctuations, we update the difficulty gap by EMA:

$$g^{(t+1)} = \mu_g g^{(t)} + (1-\mu_g)\left[R_P^{(t+1)} - \frac{1}{2}\left(R_I^{(t+1)} + R_T^{(t+1)}\right)\right], \tag{16}$$

where $\mu_g \in [0,1)$ controls temporal smoothing. A positive $g^{(t+1)}$ indicates a larger relative difficulty increase for paired views, whereas a negative value indicates a larger increase for unimodal views.

To avoid unnecessary adjustments caused by small fluctuations, we further introduce a dead zone ζ:

$$g_{\text{eff}}^{(t+1)} = \operatorname{sign}\left(g^{(t+1)}\right)\max\left(\left|g^{(t+1)}\right| - \zeta, 0\right). \tag{17}$$

The paired-view sampling probability is then updated as

$$\pi_P^{(t+1)} = \pi_P^{(t)} + \eta \tanh\left(\frac{g_{\text{eff}}^{(t+1)}}{T_\pi}\right), \tag{18}$$

where η controls the adjustment magnitude and T_π controls the sensitivity to the difficulty gap. The resulting probability is bounded within $[\pi_P^{\min}, \pi_P^{\max}]$. Thus, a positive difficulty gap increases paired-view sampling, whereas a negative gap increases the probability assigned to unimodal views.

The remaining probability is not equally divided between image-only and text-only views, but is adaptively allocated according to their relative difficulties:

$$\left(\pi_I^{(t+1)}, \pi_T^{(t+1)}\right) = \left(1-\pi_P^{(t+1)}\right)\operatorname{softmax}\left(\frac{\left(R_I^{(t+1)}, R_T^{(t+1)}\right)}{T_S}\right), \tag{19}$$

where T_S controls the sensitivity of the allocation to the difficulty difference between the two unimodal views. Therefore, the unimodal view with a larger relative difficulty receives a higher sampling probability.

The resulting probabilities satisfy $\pi_P^{(t+1)} + \pi_I^{(t+1)} + \pi_T^{(t+1)} = 1$ and are used to determine the modality-conditioned view distribution in subsequent training batches.

Table 1: Image-to-image retrieval performance in terms of mAP@5000 on COCO, NUS-WIDE, and MIRFlickr25K. The best result under each setting is highlighted in bold, and the second-best result is underlined.

Method	COCO			NUS-WIDE			MIRFlickr25K		
	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
IPHASH	82.60	87.50	89.40	79.70	81.60	82.60	–	–	–
HARR	74.80	78.90	81.60	80.70	82.60	84.10	81.80	83.00	83.80
FSCH	76.00	78.70	79.90	81.20	83.20	84.40	81.50	83.80	84.90
CTMIH	80.90	83.40	84.60	79.50	81.60	82.60	–	–	–
CroVCA	87.50	89.20	90.10	81.80	83.10	83.00	83.30	82.40	81.90
Ours	**88.30**	**90.77**	**91.81**	**86.62**	**87.05**	**87.20**	**87.47**	**88.08**	**88.32**

Note. "–" indicates that the corresponding result was not reported in the source paper.

4 EXPERIMENTS

We conduct experiments on three widely used multi-label retrieval datasets, COCO Lin et al. (2014), NUS-WIDE Chua et al. (2009), and MIRFlickr25K Huiskes & Lew (2008). We present the experimental settings and evaluate the proposed method through performance comparisons, ablation studies, and parameter-sensitivity analyses.

4.1 EXPERIMENT SETTINGS

Baselines. Our setting jointly considers paired (P), image-only (I), and text-only (T) inputs within a unified hash space, giving rise to multiple directed query–database retrieval protocols. Existing cross-modal hashing methods typically assume fixed image and text modalities and are mainly evaluated on conventional image-to-text or text-to-image retrieval, making them not directly comparable under our unified $P/I/T$ setting. We therefore compare with representative recent hashing methods, including IPHash, HARR, FSCH, CTMIH, and CroVCA, under the common Image-to-Image (I2I) protocol, where all methods can be evaluated using the same query and database modality. This comparison examines the quality of the learned binary image representations under a shared evaluation setting. We additionally evaluate our full model on eight directed retrieval protocols, including $P2P$, $I2I$, $T2I$, $I2T$, $P2I$, $P2T$, $T2P$, and $I2P$, to assess its retrieval capability under different modality-availability conditions.

Implementation Details. We adopt CLIP as the vision–language backbone and apply LoRA to both the vision and text encoders for parameter-efficient adaptation. The LoRA rank and dropout rate are set to 16 and 0.1, respectively. Representations from different modality conditions are mapped into a unified hash space through a shared HashCoder. All models are trained for 12 epochs with a batch size of 256. The learning rate and weight decay are set to 10^{-3} and 10^{-2}, respectively. For the three training objectives, we set $\lambda_{\text{align}} = 1.0$, $\lambda_{\text{cr}} = 0.1$, and $\lambda_{\text{mlc}} = 0.05$. The initial sampling probabilities are set to $(\pi_P, \pi_I, \pi_T) = (0.8, 0.1, 0.1)$. The relation weights are bounded within $[0.7, 1.5]$, with $\mu_W = 0.8$ and $\beta = 1.0$. For the adaptive view curriculum, we set $\mu_D = 0.9$, $\mu_g = 0.8$, $\eta = 0.01$, $T_\pi = 0.02$, and $\zeta = 0.001$, and constrain π_P to $[0.4, 0.8]$. For COCO and MIRFlickr25K, we use the adaptive directed relation weighting introduced in the Method section, while for NUS-WIDE we use the progress-aware variant of the relation-weight estimator. This adjustment only changes the estimation of $\mathbf{W}$, while the remaining training pipeline is unchanged.

Evaluation Metrics. We evaluate hash codes of 16, 32, and 64 bits. Retrieval performance is measured using Asymmetric Hamming mAP@5000, where continuous query outputs are matched against binary database hash codes. For the I2I baseline comparison, we report mAP@5000 under the common image retrieval protocol. For our unified $P/I/T$ setting, we report the performance of all eight directed retrieval protocols individually and further compute their average to summarize the overall retrieval capability across different modality conditions.

Table 2: AsymHamming mAP@5000 of the full model under eight directed retrieval protocols on three datasets.

Dataset	bits	P2P	I2I	T2I	I2T	P2I	P2T	T2P	I2P	Avg.
COCO	16	89.85	88.30	88.28	88.62	88.93	89.43	88.95	88.98	88.92
	32	92.22	90.77	90.79	90.80	91.44	91.65	91.09	91.34	91.26
	64	92.73	91.81	91.77	91.64	92.48	92.19	91.75	92.18	92.07
NUS-WIDE	16	85.90	86.62	80.87	82.02	85.92	81.23	80.64	86.35	83.69
	32	86.01	87.05	81.62	83.12	86.39	82.00	81.46	86.76	84.30
	64	85.34	87.20	81.45	83.49	86.29	81.66	80.84	86.79	84.13
MIRFlickr25K	16	85.10	87.47	80.99	82.48	86.34	80.66	79.76	86.62	83.68
	32	85.96	88.08	81.48	83.22	87.19	81.36	80.31	87.41	84.38
	64	85.87	88.32	81.47	83.33	87.17	81.03	80.19	87.71	84.39

4.2 Performance Analyses

Table 1 compares our method with representative hashing methods under the common Image-to-Image (I2I) retrieval protocol using 16-,32-, and 64-bit hash codes. Our method achieves the best performance under all reported settings on COCO, NUS-WIDE, and MIRFlickr25K. In particular, the improvements over the strongest baseline are more pronounced on NUS-WIDE and MIRFlickr25K, while consistent gains are also observed on COCO. These results indicate that the proposed unified training framework does not sacrifice the quality of image hash representations when accommodating heterogeneous input conditions. Moreover, the performance generally improves with longer hash codes, although the gains gradually become smaller, suggesting that the benefit of additional hash capacity tends to saturate.

Table 2 further reports the performance of the full model under eight directed retrieval protocols involving paired, image-only, and text-only inputs. On COCO, the performances of different retrieval directions are relatively close, with paired-input protocols generally achieving strong results. In contrast, on NUS-WIDE and MIRFlickr25K, protocols using text-only queries, such as T2I and T2P, are consistently weaker than image-based or paired-input protocols. This difference is likely related to the textual inputs used by the datasets: COCO provides relatively rich natural-language captions, whereas NUS-WIDE and MIRFlickr25K rely on sparse user-provided tags, which provide less complete semantic information for text-only retrieval. The results also show that performance on NUS-WIDE and MIRFlickr25K largely stabilizes after 32 bits, while COCO continues to benefit from longer hash codes.

4.3 Ablation Study

Table 3 reports the ablation results on three datasets.

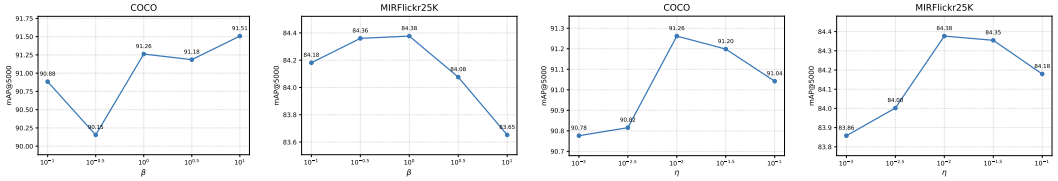
w/o Relation Weighting. This variant fixes all nine relation weights to 1 while retaining the adaptive view curriculum. Full generally performs better, with clearer gains on NUS-WIDE and MIRFlickr25K, indicating the benefit of adaptively calibrating directed relations according to their learning states.

w/o View Curriculum. This variant retains relation weighting but fixes the $P/I/T$ sampling probabilities to those reached by the full model at the final training stage. Full performs better on COCO and NUS-WIDE, while the effect is less consistent on MIRFlickr25K, showing that dynamic view allocation provides complementary benefits beyond relation weighting.

w/o Both. This variant fixes all relation weights to 1 and uses a fixed $P/I/T$ sampling distribution throughout training. Removing both adaptive components generally causes larger degradation, especially on COCO and NUS-WIDE, demonstrating the complementarity between relation weighting and adaptive view sampling.

Table 3: Ablation results in terms of the average AsymHamming mAP@5000 over the eight retrieval protocols.

Method	COCO			NUS-WIDE			MIRFlickr25K		
	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
w/o Relation Weighting	89.00	91.05	91.99	83.40	84.02	83.85	83.50	84.02	84.36
w/o Both	87.30	88.73	90.79	82.92	83.56	83.82	83.36	84.24	84.36
w/o View Curriculum	87.84	89.57	90.72	83.37	83.97	84.04	83.69	84.73	84.20
Full	88.92	91.26	92.07	83.69	84.30	84.13	83.68	84.38	84.39



(a) COCO (β) (b) MIRFlickr25K (β) (c) COCO (η) (d) MIRFlickr25K (η)

Figure 2: Parameter sensitivity of β and η with 32-bit hash codes.

4.4 Parameter Sensitivity

We first investigate the sensitivity to the relation-weight adjustment strength β using 32-bit hash codes while fixing $\eta = 10^{-2}$. Specifically, β is varied over $\{10^{-1}, 10^{-0.5}, 10^{0}, 10^{0.5}, 10^{1}\}$. As shown in Fig. 2(a)–(b), the performance varies moderately with β on both COCO and MIRFlickr25K. MIRFlickr25K achieves its best performance around $\beta = 1$, while COCO maintains competitive performance over a relatively broad range of values. This indicates that β controls the strength of relation-level calibration and that an appropriate moderate value provides a good balance across datasets. We therefore set $\beta = 1$ in all experiments.

We further evaluate the sampling update step η by fixing $\beta = 1$ and varying η over $\{10^{-3}, 10^{-2.5}, 10^{-2}, 10^{-1.5}, 10^{-1}\}$. As shown in Fig. 2(c)–(d), very small values of η lead to weaker performance, whereas moderate values around 10^{-2} achieve better results on both datasets. This suggests that an appropriate update step allows the adaptive curriculum to respond to changes in learning difficulty without causing overly aggressive variations in the sampling distribution. We therefore set $\eta = 10^{-2}$ in all experiments.

5 Conclusion

We presented a unified supervised multi-label cross-modal hashing framework for paired, image-only, and text-only inputs. To address the heterogeneous and dynamically changing optimization difficulties induced by these input conditions, we coupled adaptive directed relation weighting with an adaptive view curriculum. The former calibrates relation-conditioned multi-label contrastive learning, while the latter uses the learned relation states to dynamically adjust subsequent view sampling, forming a closed-loop training process. Experiments on COCO, NUS-WIDE, and MIRFlickr25K across multiple hash-code lengths and eight directed retrieval protocols demonstrate the effectiveness of the proposed framework, with ablation and parameter-sensitivity analyses further validating the two adaptive components.

AI use statement

Generative AI tools were used to assist with manuscript writing and editing, methodological and experimental discussions, result interpretation, and figure refinement. All AI-assisted content was reviewed and verified by the authors. The authors take full responsibility for the methodology, experiments, results, claims, and final content of this work.

REFERENCES

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 41–48, 2009. doi: 10.1145/1553374.1553380. URL https://icml.cc/Conferences/2009/papers/119.pdf.

Hu Cao, Lei Huang, Jie Nie, and Zhiqiang Wei. Unsupervised deep hashing with fine-grained similarity-preserving contrastive learning for image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(5):4095–4108, 2024. doi: 10.1109/TCSVT.2023.3320444.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proceedings of the 8th ACM International Conference on Image and Video Retrieval*. ACM, 2009. doi: 10.1145/1646396.1646452.

Qinkang Gong, Liangdao Wang, Hanjiang Lai, Yan Pan, and Jian Yin. Vit2hash: Unsupervised information-preserving hashing. *arXiv preprint arXiv:2201.05541*, 2022.

Mark J. Huiskes and Michael S. Lew. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM SIGMM International Conference on Multimedia Information Retrieval*, pp. 39–43. ACM, 2008. doi: 10.1145/1460096.1460104.

Qing-Yuan Jiang and Wu-Jun Li. Deep cross-modal hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3232–3240, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Jiang_Deep_Cross-Modal_Hashing_CVPR_2017_paper.html.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.

M. Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, volume 23, 2010. URL https://proceedings.neurips.cc/paper/2010/hash/e57c6b956a6521b28495f2886ca0977a-Abstract.html.

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 *European Conference on Computer Vision*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48.

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):134:1–134:23, 2024. doi: 10.1145/3627162.

Devraj Mandal, Kunal N. Chaudhury, and Soma Biswas. Generalized semantic preserving hashing for n-label cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4076–4084, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Mandal_Generalized_Semantic_Preserving_CVPR_2017_paper.html.

Ilyass Moummad, Kawtar Zaher, Hervé Goëau, and Alexis Joly. Image hashing via cross-view code alignment in the age of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 3589–3598, 2026. URL https://openaccess.thecvf.com/content/CVPR2026W/ECV/html/Moummad_Image_Hashing_via_Cross-View_Code_Alignment_in_the_Age_of_CVPRW_2026_paper.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Xiaobo Shen, Haoyu Cai, Xiuwen Gong, and Yuhui Zheng. Contrastive transformer masked image hashing for degraded image retrieval. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 1218–1226, 2024. doi: 10.24963/ijcai.2024/135.

Shupeng Su, Zhisheng Zhong, and Chao Zhang. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3027–3035, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Su_Deep_Joint-Semantics_Reconstructing_Hashing_for_Large-Scale_Unsupervised_Cross-Modal_Retrieval_ICCV_2019_paper.html.

Xinyu Xia, Guohua Dong, Fengling Li, Lei Zhu, and Xiaomin Ying. When CLIP meets cross-modal hashing retrieval: A new strong baseline. *Information Fusion*, 100:101968, 2023. doi: 10.1016/j.inffus.2023.101968. URL https://doi.org/10.1016/j.inffus.2023.101968.

Pingyue Zhang and Mengyue Wu. Multi-label supervised contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16786–16793, 2024. doi: 10.1609/aaai.v38i15.29619. URL https://ojs.aaai.org/index.php/AAAI/article/view/29619.

Qiang Zou, Shuli Cheng, and Jiayi Chen. PromptHash: Affinity-prompted collaborative cross-modal learning for adaptive hashing retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19649–19658, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zou_PromptHash_Affinity-Prompted_Collaborative_Cross-Modal_Learning_for_Adaptive_Hashing_Retrieval_CVPR_2025_paper.html.