RELEVANCE-RESOLUTION TRANSFER VIA SCALE-DECOMPOSABLE FRACTIONAL DIFFUSION FOR MULTI-LENGTH CROSS-MODAL HASH RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-modal hashing enables efficient retrieval by encoding heterogeneous data into compact binary codes. Recent methods exploit fine-grained relations encoded in multi-label training structure, yet none of them constrains how those relations survive as consistent candidate rankings in finite, multi-length Hamming spaces, which we term the relevance resolution bottleneck (RRB). To address the RRB, we propose MultiBit, which transfers relevance resolution from multi-label structure to multi-length Hamming spaces. MultiBit first constructs a scale-decomposable fractional relation teacher from dataset-level label co-occurrence and label specificity, and models dependencies from local to long-range over continuous diffusion scales. It then maps the discretized diffusion scales and their quadrature weights to scale-aware bit subblocks of the maximum-length code, organizes the target code lengths as nested prefixes, and aligns their Hamming candidate rankings with the teacher relations. Experiments on multiple benchmarks demonstrate improved retrieval accuracy. *Code is available in the supplementary material.*

1 INTRODUCTION

Cross-modal hashing maps heterogeneous image and text data into a shared Hamming space, enabling efficient retrieval through compact storage and bitwise comparison (Wang et al., 2024). At inference, a query is encoded from its observed modality and ranked against the database codes by Hamming distance. Retrieval quality therefore depends not only on separating relevant from irrelevant samples, but also on whether semantic differences among candidates are reflected in their relative distances.

Recent methods increasingly exploit fine-grained relations encoded in multi-label training structure, from hard-negative objectives that separate relevant from irrelevant samples (Zhang et al., 2026c; Su et al., 2025; Peng et al., 2026), through proxy- and overlap-based objectives that distinguish degrees of observed agreement (Huo et al., 2024b; Yao et al., 2026; Su et al., 2026), to structured-label approaches that capture dataset-level dependencies and already produce continuous relation scores (Cao et al., 2025; Zhang et al., 2026a; Meng et al., 2026). These developments progressively improve label-side resolution, yet a fine-grained relation in label space does not by itself determine how propagation ranges should be encoded in finite bits, nor does it guarantee that the induced order survives in the final Hamming list or across code lengths. The remaining limitation lies less in relation construction than in transferring these relations into finite, multi-length codes.

This transfer gap exposes a relevance resolution bottleneck (RRB): relevance resolution denotes the degree to which semantically grounded differences between a query and its candidates can be distinguished and preserved, rather than the number of bits itself. When label-side relations enter a finite Hamming space without explicit organization and ranking constraints, candidates with different relevance can collide at the same distance, appear in inverted order, or exchange positions across code lengths. The resulting question is how to organize such relations as transferable supervision that is preserved consistently in nested, multi-length Hamming rankings.

As illustrated in Figure 1, RRB manifests as distance collisions and ranking inversions among semantically different candidates; its cross-length form is quantified in Section 4.5. To address it, we propose MULTIBIT, a unified framework that transfers relevance resolution from dataset-level multi-label structure to multi-length Hamming spaces. The framework starts from a scale-decomposable

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.

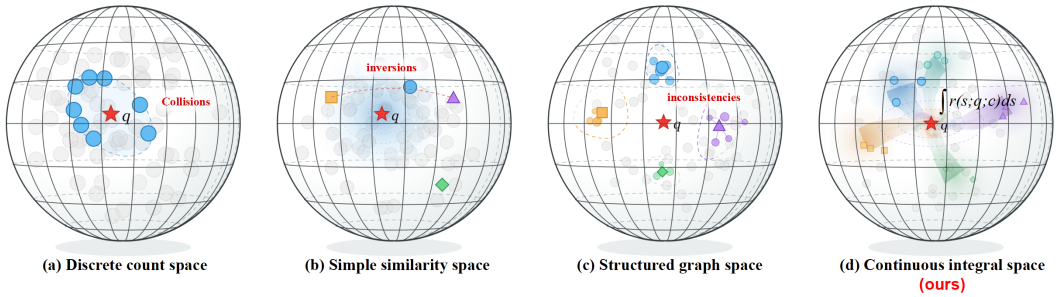

Figure 1: Motivation for multi-scale relation modeling. (a)–(c) illustrate relation spaces built from discrete labels, simple similarity, and graph structure, which capture progressively richer relations but still describe them at a single scale; (d) MultiBit integrates label co-occurrence and label specificity through fractional diffusion, covering dependencies from local neighborhoods to long-range graph paths and exposing the resulting scale structure to finite binary codes.

Fractional Relation Teacher (FRT), which integrates label co-occurrence and label specificity through fractional diffusion and decomposes the result into query-conditioned relevance targets together with the quadrature weight of each diffusion scale. These weights then guide Scale-aware Bit Organization (SBO), which assigns relation scales to dedicated bit subblocks of the maximum-length code, and Multi-length Prefix Alignment (MPA), which organizes the target code lengths as nested prefixes and constrains their Hamming candidate distributions to follow the same teacher-induced ordering. Our contributions are threefold:

- We identify and formalize the relevance resolution bottleneck (RRB), in which label-side relations collide, invert, or exchange ranks across code lengths.
- We propose MULTIBIT, which constructs relevance resolution with FRT and preserves it in nested multi-length Hamming spaces through SBO and MPA.
- We characterize a conditional order-preservation property of the listwise objective in the continuous Hamming surrogate, and evaluate how far that ordering survives sign quantization at all four code lengths.

2 RELATED WORK

Cross-Modal Hashing. Existing cross-modal hashing methods can be grouped into three categories by how they construct semantic supervision. The first employs pairwise, contrastive, triplet, or hard-negative objectives on binary relevance, establishing a basic discriminative boundary (Zhang et al., 2026c; Su et al., 2025; Peng et al., 2026). The second introduces multi-label classification, semantic proxies, shared-label counts, or Jaccard similarity to retain label identities and direct overlap (Huo et al., 2024b; Yao et al., 2026; Cheng et al., 2026b). The third incorporates structured label relations through label graphs, propagation, embeddings, or semantic hierarchies, extending isolated indicators with dataset-level co-occurrence and higher-order dependencies (Zhang et al., 2026b; Cao et al., 2025; Zhang et al., 2026a; Meng et al., 2026).

Fine-Grained Relevance Modeling. Fine-grained relevance modeling describes graded relations that a single relevant/irrelevant indicator cannot express. Direct set-based measures distinguish candidates by observed label overlap, while frequency-aware weighting reduces the dominance of common labels (Yang et al., 2025); structured approaches further exploit label correlations, graph neighborhoods, and indirect semantic paths, and some produce continuous relation scores for pairwise supervision or representation regularization (Su et al., 2026; Cheng et al., 2026b).

Multi-Length and Bit-Scalable Hashing. A separate line of work targets deploying a single model across several code lengths. Central-coding and multi-bit schemes share one representation across lengths (Wu et al., 2024; Zuo et al., 2026), and adaptive-bit methods allocate capacity across bits or lengths (Wang et al., 2025; Zou et al., 2025; Jiang et al., 2026). These approaches reduce deployment cost, but they do not constrain how the induced candidate ranking changes when a code is truncated.

Despite these advances, existing methods focus on constructing stronger semantic relations or improving feature discrimination; the resulting relations are not organized according to propagation

scale and finite bit capacity, and are not constrained to survive as consistent candidate rankings across code lengths (Zuo et al., 2026; Tan et al., 2026; Jiang et al., 2026). Label-side distinctions may therefore still collapse into distance ties, ranking inversions, or cross-length inconsistencies after binary encoding. To address this, MULTIBIT couples FRT, SBO, and MPA to preserve fine-grained semantic relations as consistent candidate rankings in multi-length Hamming spaces.

3 METHOD

3.1 PROBLEM FORMULATION: RELEVANCE RESOLUTION BOTTLENECK

Let $\mathcal{D} = \{(x_i^I, x_i^T, y_i)\}_{i=1}^N$ denote a cross-modal training set, where x_i^I and x_i^T are the image and text views of the i-th instance, respectively, and $y_i \in \{0,1\}^C$ is their shared C-dimensional multi-hot label vector, with N denoting the number of training instances. Given a set of target code lengths $\mathcal{B} = \{B_1, \ldots, B_K\}$ and its maximum length $B_{\max} = \max \mathcal{B}$, we learn modality-specific hash functions F^I and F^T that map heterogeneous inputs into a shared $B_{\max}$-dimensional Hamming space. Since the discrete sign function is non-differentiable, we follow the continuous-relaxation strategy used in deep hashing (Zou et al., 2025; Tan et al., 2026) and employ $\tanh$ as a continuous relaxation during training, so that the maximum-length representation of instance i in modality m is $\widetilde{h}_i^m = \tanh(F^m(x_i^m)) \in (-1,1)^{B_{\max}}$. To support multiple code lengths with a single model, as in multi-bit and unequal-length hashing (Wu et al., 2024; Zuo et al., 2026; Zou et al., 2025; Jiang et al., 2026; Wang et al., 2025), we define a prefix selection matrix $P_B = [I_B \ \mathbf{0}] \in \{0,1\}^{B \times B_{\max}}$ and obtain the binary code at any target length $B \in \mathcal{B}$ from the first B dimensions of the maximum-length representation as $b_i^{m,(B)} = \operatorname{sign}(P_B \widetilde{h}_i^m) \in \{-1,+1\}^B$. Accordingly, the Hamming distance between an image query x_i^I and a text candidate x_j^T at code length B is

$$d_H^{(B)}(i,j) = \frac{1}{2}\left(B - b_i^{I,(B)^\top} b_j^{T,(B)}\right), \tag{1}$$

with a symmetric formulation for text-to-image retrieval. Labels are used only to construct training supervision; inference requires only encoding, prefix truncation, and Hamming ranking. Appendix Section A.3 states RRB formally, in its collision, inversion, and cross-length forms.

3.2 OVERVIEW

As illustrated in Figure 2, MULTIBIT consists of three components; the notation and experimental settings are summarized in Appendix Table A-3. FRT turns training-set label co-occurrence and specificity into continuous local-to-long-range dependencies and decomposes the resulting teacher into finite diffusion scales with their quadrature weights. SBO allocates finite representation capacity by mapping these scales and weights to scale-specific bit subblocks. MPA organizes the subblocks as nested prefixes and aligns their Hamming rankings with one shared teacher ordering.

3.3 SCALE-DECOMPOSABLE FRACTIONAL RELATION TEACHER (FRT)

The first component constructs the relational supervision that the Hamming representations should preserve. Given the training labels defined in Section 3.1, let $n_a = \sum_i y_{ia}$ and $n_{ab} = \sum_i y_{ia} y_{ib}$ denote the frequency of label a and the co-occurrence frequency of labels a and b, respectively. We build a positive pointwise mutual information (PPMI) label graph $W \in \mathbb{R}^{C \times C}$ (Levy & Goldberg, 2014; Hou et al., 2025) with a smoothing constant $\varepsilon > 0$, whose entry W_{ab} is the positive part of the mutual information between labels a and b (Appendix Section B, Eq. (B.1)). With $D_{aa} = \sum_b W_{ab}$, the normalized graph Laplacian is $L = I - D^{-1/2} W D^{-1/2}$. We further assign each label a specificity weight $s_a = \log((N+1)/(n_a+1))$ and obtain the weighted label vector $\widetilde{y}_i = y_i \odot s$; the resulting graph captures dataset-level label dependencies while s modulates the contribution of each label by its frequency.

A single propagation depth describes only one semantic neighborhood. To integrate dependencies from direct co-occurrence to long-range graph paths, we build on heat-kernel diffusion over graphs (Chamberlain et al., 2021) and define the continuous-scale fractional diffusion kernel

$$K_{\alpha,\eta} = (L + \eta I)^{-\alpha} = \frac{1}{\Gamma(\alpha)} \int_0^\infty t^{\alpha-1} e^{-\eta t} e^{-tL} \, \mathrm{d}t, \qquad \alpha > 0, \ \eta > 0. \tag{2}$$

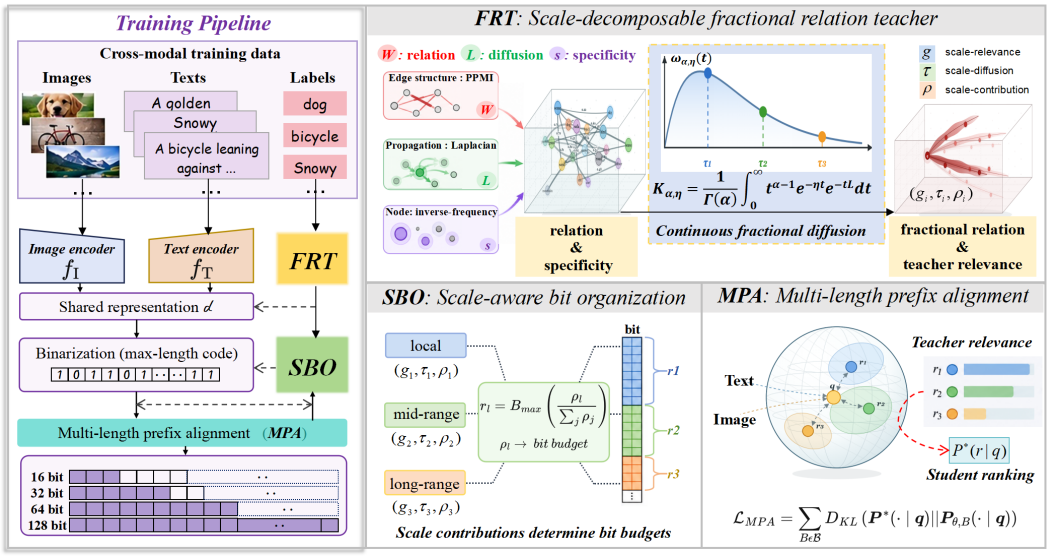

Figure 2: Overview of MULTIBIT. FRT derives an integrated teacher g^* through fractional diffusion and decomposes it into M scale components $\{(g_\ell, \tau_\ell, \rho_\ell)\}$. SBO converts ρ_ℓ into bit budgets m_ℓ within $B_{\max}$, and MPA aligns the Hamming distributions $P_{\theta,B}$ of nested prefixes with the teacher P^* at every target length.

Here t is the diffusion scale, and α and η control the spectral decay and the diffusion range, respectively (Maskey et al., 2023); their interpretation is detailed in Appendix Section B. For a query q and a candidate r, the integrated teacher relevance is defined by the kernel-induced cosine

$$g^*(q,r) = \frac{\widetilde{y}_q^\top K_{\alpha,\eta} \widetilde{y}_r}{\sqrt{(\widetilde{y}_q^\top K_{\alpha,\eta} \widetilde{y}_q)(\widetilde{y}_r^\top K_{\alpha,\eta} \widetilde{y}_r)}}. \tag{3}$$

The continuous kernel must expose a finite scale structure before it can guide finite bits. After the change of variable $x = \eta t$, we apply an M-point generalized Gauss–Laguerre rule (Golub & Welsch, 1969) whose nodes x_ℓ are the roots of $L_M^{(\alpha-1)}(x)$ and whose weights are ω_ℓ. This yields

$$K_{\alpha,\eta} \approx K_{\alpha,\eta}^{(M)} = \sum_{\ell=1}^{M} \rho_\ell K_\ell, \qquad K_\ell = e^{-\tau_\ell L}, \quad \tau_\ell = \frac{x_\ell}{\eta}, \quad \rho_\ell = \frac{\eta^{-\alpha}\omega_\ell}{\Gamma(\alpha)}. \tag{4}$$

The rule supplies only the scale coordinates $\{(\tau_\ell, \rho_\ell)\}$; g^* and every g_ℓ below are computed from the exact matrix functions $(L + \eta I)^{-\alpha}$ and $e^{-\tau_\ell L}$ (Appendix Table B-5). Each component also induces a scale-specific teacher

$$g_\ell(q,r) = \frac{\widetilde{y}_q^\top K_\ell \widetilde{y}_r}{\sqrt{(\widetilde{y}_q^\top K_\ell \widetilde{y}_q)(\widetilde{y}_r^\top K_\ell \widetilde{y}_r)}}, \qquad \ell = 1, \ldots, M. \tag{5}$$

FRT therefore outputs the integrated teacher ranking g^* and the finite set $\{(g_\ell, \tau_\ell, \rho_\ell)\}_{\ell=1}^M$, which exposes the local-to-long-range scale structure to the subsequent bit organization.

3.4 SCALE-AWARE BIT ORGANIZATION (SBO)

The quadrature weights in Section 3.3 describe how the fractional kernel distributes its mass over diffusion scales, and we use them as a prior for allocating finite representation capacity. For the maximum code length $B_{\max}$ defined in Section 3.1, the ideal number of bits assigned to scale ℓ is

$$m_\ell^{\text{ideal}} = B_{\max} \frac{\rho_\ell}{\sum_{j=1}^{M} \rho_j}. \tag{6}$$

Since $\rho_\ell = \eta^{-\alpha}\omega_\ell / \Gamma(\alpha)$, the common factor cancels in the normalized ratio, so the allocation is governed by the Gauss–Laguerre weights $\{\omega_\ell\}$, hence by α rather than η. This ratio is a kernel-mass

prior over scales rather than a measured per-scale contribution (Appendix Section B). Because bit dimensions must be integers, we convert m_ℓ^{ideal} to m_ℓ using a largest-remainder allocation, reserve at least one bit for every retained scale, and adjust the remaining dimensions such that $\sum_{\ell=1}^{M} m_\ell = B_{\max}$. Thus, scales carrying more quadrature mass receive more capacity, while low-weight long-range components still retain a share of it; bit-scalable and unequal-length hashing allocate capacity in the same spirit (Wang et al., 2025; Zou et al., 2025).

Let E^m be the encoder and projection for modality $m \in \{I, T\}$, mapping an input to a shared d-dimensional feature $z_i^m = E^m(x_i^m)$. For each diffusion scale, a shared scale-specific hash head $H_\ell : \mathbb{R}^d \to \mathbb{R}^{m_\ell}$ produces a continuous subblock. We sort the scales from local to long range, denoted by $(\sigma_1, \ldots, \sigma_M) = \operatorname{argsort}(\tau_1, \ldots, \tau_M)$, and concatenate the resulting subblocks in that order to form the maximum-length representation (Appendix Section B, Eq. (B.2)). This refines the hash function F^m introduced in Section 3.1: the maximum-length representation is no longer an unstructured vector, but a concatenation of subblocks with explicit diffusion-scale semantics. During training, each subblock is paired with its corresponding scale teacher g_ℓ; the resulting scale-specific supervision is jointly defined with prefix alignment in the next subsection.

3.5 Multi-Length Prefix Alignment (MPA)

Scale-aware subblocks specify the internal semantics of the maximum-length code, the only deployable length that is itself multi-scale (Table B-1). Consistent rankings across truncations are obtained by aligning the prefix rankings with the shared teacher ordering. All lengths share the same prefix operator, encoder, scale coordinates and bit ordering, so extending a code preserves every bit of its shorter prefix.

For a query q and its cross-modal candidate set $\mathcal{C}_q$, the integrated teacher score from Eq. (3) is converted into a query-conditioned teacher distribution $P^*(r \mid q) \propto \exp(g^*(q, r)/T_t)$ over $r \in \mathcal{C}_q$, where $T_t > 0$ controls the concentration of the teacher ranking without changing its order (Hwang et al., 2025). At length B, we compute a normalized differentiable Hamming surrogate and the corresponding student distribution for retrieval direction $m \to n$ as

$$\begin{aligned} \overline{d}_B^{m\to n}(q, r) &= \frac{1}{2B}\left(B - (\widetilde{h}_q^{m,(B)})^{\mathsf{T}} \widetilde{h}_r^{n,(B)}\right), \\ P_{\theta,B}^{m\to n}(r \mid q) &= \frac{\exp\left(-\overline{d}_B^{m\to n}(q, r)/T_s\right)}{\sum_{u\in\mathcal{C}_q} \exp\left(-\overline{d}_B^{m\to n}(q, u)/T_s\right)}, \end{aligned} \tag{7}$$

where $T_s > 0$ is the student temperature and division by B places different code lengths in the common range $[0, 1]$. For a query set $\mathcal{Q}_m$, the directional listwise loss (Yoon et al., 2025) minimizes the Kullback–Leibler divergence (Kullback & Leibler, 1951) between the two distributions; we denote the resulting loss at length B by $\mathcal{L}_B^{m\to n}$ (Appendix Section B, Eq. (B.3)). The multi-length prefix objective aligns both retrieval directions at every target length:

$$\mathcal{L}_{\text{prefix}} = \sum_{B\in\mathcal{B}} \beta_B \left(\mathcal{L}_B^{I\to T} + \mathcal{L}_B^{T\to I}\right), \qquad \beta_B \geq 0. \tag{8}$$

We additionally apply the same bidirectional distribution matching to every complete scale subblock, replacing g^* with g_ℓ and the length-B prefix with $\widetilde{h}^{m,(\ell)}$. Let the resulting losses be $\mathcal{L}_\ell^{I\to T}$ and $\mathcal{L}_\ell^{T\to I}$. With normalized scale weights $\bar{\rho}_\ell = \rho_\ell / \sum_j \rho_j$, the scale objective and the complete training objective are

$$\begin{aligned} \mathcal{L}_{\text{scale}} &= \sum_{\ell=1}^{M} \bar{\rho}_\ell \left(\mathcal{L}_\ell^{I\to T} + \mathcal{L}_\ell^{T\to I}\right), \\ \mathcal{L} &= \mathcal{L}_{\text{prefix}} + \lambda_s \mathcal{L}_{\text{scale}}. \end{aligned} \tag{9}$$

Scale supervision is designed to make each subblock encode its designated propagation range, whereas prefix supervision aligns the continuous ranking at every deployable length with the integrated teacher order. In particular, if the listwise divergence at a given length vanishes, equality of the two distributions and the monotonicity of Softmax imply

$$g^*(q, r_a) > g^*(q, r_b) \quad \Longrightarrow \quad \overline{d}_B(q, r_a) < \overline{d}_B(q, r_b), \tag{10}$$

which links the teacher's relevance resolution to the student's soft Hamming ranking during training. Appendix Section B.1 and Section 4.4 report how much of this order survives binarization and how well the discrete ranking preserves it. At inference the teacher branch and the listwise objectives are removed, so any requested length B is served by prefix extraction and standard Hamming ranking.

4 EXPERIMENTS

4.1 DATASETS AND EXPERIMENTAL SETUP

We conduct experiments on three public multi-label image–text datasets: MIRFlickr-25K (Huiskes & Lew, 2008), MS-COCO (Lin et al., 2014), and IAPR TC-12 (Grubinger et al., 2006), and report additional results on NUS-WIDE (Chua et al., 2009) in Appendix Table C-1. All methods use the same CLIP ViT-B/32 backbone (Radford et al., 2021). For the data split and the mAP@all metric we report image-to-text (I2T) and text-to-image (T2I) retrieval and follow the protocol of Cheng et al. (2026a); Appendix Section A.1 gives dataset sizes and metrics.

4.2 COMPARISON WITH STATE-OF-THE-ART METHODS

Compared Methods. We compare MultiBit with eight fixed-length cross-modal hashing methods (DSSH, DNPH, DECH, DIMCH, DPBE, DSTH, DPSIH, DGHDGH (Qin et al., 2026a; Huo et al., 2024a; Li et al., 2025; Tu et al., 2025; Cheng et al., 2025; Qin et al., 2026b; Han et al., 2026; Cheng et al., 2026a)) and three multi-length methods serving several lengths from one model (CMCL (Wu et al., 2024), RCMH (Jiang et al., 2026), VH-ABL (Zuo et al., 2026)).

Table 1: I2T and T2I mAP@all (%) on three benchmarks at four code lengths; best in bold, second best underlined.

Task	Method	Reference	MIRFlickr-25K				IAPR TC-12				MS-COCO			
			16	32	64	128	16	32	64	128	16	32	64	128
I2T	DSSH	TKDE'26	75.17	75.59	80.47	81.10	44.09	53.73	57.00	65.21	62.85	67.74	73.24	76.38
	DNPH	TOMM'24	74.12	79.10	81.36	81.70	46.13	49.68	56.07	62.57	64.22	70.30	71.61	72.52
	DECH	AAAI'25	81.77	83.85	85.00	85.44	58.54	64.53	67.65	69.47	64.20	68.70	72.39	74.06
	DIMCH	TIP'25	81.15	82.25	82.45	83.35	58.73	61.93	64.23	65.48	65.18	67.65	69.80	69.33
	DPBE	MM'25	69.31	72.14	74.09	75.80	49.86	55.94	60.98	64.65	55.42	61.78	66.92	72.39
	DSTH	TMM'26	79.67	81.01	82.43	82.85	57.89	62.06	64.73	66.78	62.81	69.00	70.89	72.51
	DPSIH	AAAI'26	77.43	80.76	81.90	82.40	56.55	60.88	64.26	66.38	63.49	68.05	70.90	72.47
	DGHDGH	ICLR'26	79.14	80.47	80.99	81.72	55.44	59.87	62.47	63.86	58.29	61.29	60.44	63.41
I2T (Multi/length)	CMCL	TKDE'24	**83.13**	84.34	85.22	85.44	58.02	62.69	66.91	69.59	**70.23**	**72.95**	75.13	75.67
	RCMH	TPAMI'26	70.80	71.61	72.10	72.45	56.90	59.49	60.56	61.05	67.82	70.79	73.06	75.00
	VH-ABL	TMM'26	80.28	81.73	82.51	82.74	54.71	60.31	63.60	65.37	63.32	70.35	73.61	75.38
	MultiBit	**OURS**	82.86	**84.37**	**85.49**	**85.92**	**60.95**	**65.27**	**68.25**	**70.07**	69.55	72.53	**75.68**	**76.89**
T2I	DSSH	TKDE'26	74.29	73.63	78.65	79.38	43.17	51.10	55.91	68.36	63.26	68.48	74.66	77.68
	DNPH	TOMM'24	71.96	76.43	78.41	79.18	44.28	48.13	56.04	63.62	64.35	70.05	71.05	72.61
	DECH	AAAI'25	78.11	79.86	80.54	81.09	57.41	64.89	68.76	71.05	64.66	69.66	72.55	74.63
	DIMCH	TIP'25	76.89	77.60	79.01	78.76	58.28	61.96	64.49	65.87	64.77	67.51	69.31	68.54
	DPBE	MM'25	71.55	73.57	75.52	76.68	48.30	54.13	59.01	62.73	56.05	61.87	67.02	72.81
	DSTH	TMM'26	75.64	77.60	78.62	79.24	57.93	61.98	65.23	67.14	63.39	68.85	70.83	72.87
	DPSIH	AAAI'26	73.92	77.39	78.09	78.60	56.66	60.62	64.91	66.88	63.48	68.09	70.92	72.71
	DGHDGH	ICLR'26	76.27	77.61	78.39	78.93	54.14	58.38	60.66	62.36	58.61	61.21	62.02	63.53
T2I (Multi/length)	CMCL	TKDE'24	79.09	**80.28**	80.50	80.45	58.72	64.46	**69.52**	**72.80**	**69.95**	**72.87**	75.30	75.73
	RCMH	TPAMI'26	74.11	74.09	74.21	74.25	54.69	57.29	58.71	59.41	67.03	70.52	73.02	75.31
	VH-ABL	TMM'26	74.96	75.94	77.50	78.58	57.22	61.28	65.60	67.01	63.72	70.11	73.55	75.27
	MultiBit	**OURS**	**79.25**	80.02	**80.82**	**81.25**	**59.94**	**65.59**	69.45	71.86	68.81	71.83	**76.15**	**77.84**

Overall Results. Table 1 reports 24 settings, spanning two retrieval directions, three datasets and four code lengths. MultiBit attains the best score in 16 of them, including every 64- and 128-bit cell on MIRFlickr-25K and MS-COCO and all four lengths on IAPR TC-12 in the image-to-text direction. CMCL takes the remaining eight — six at 16 or 32 bits and two in the text-to-image direction of IAPR TC-12 — and RCMH and VH-ABL remain below MultiBit throughout.

Performance–Efficiency. Figure 3 reports the deployment cost on MIRFlickr-25K, where one model serves all four code lengths: the 64-bit query time is 0.0092 ms, about $9\times$ lower than the accuracy-competitive CMCL (0.0861 ms); Table 4 covers all three datasets and Appendix Table C-2 the fixed-length methods.

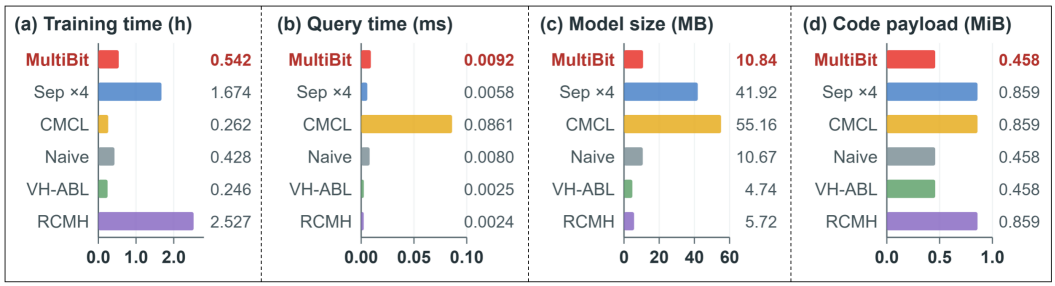

Figure 3: Deployment cost of serving all four code lengths with one model on MIRFlickr-25K: (a) total training time for the four lengths; (b) 64-bit query time excluding the CLIP forward pass; (c) trainable model size excluding the shared frozen backbone; (d) packed code payload.

4.3 Ablation Study

Table 2: Progressive ablation of the core components on three datasets and four code lengths.

Variant	MIRFlickr-25K				IAPR TC-12				MS-COCO			
	16	32	64	128	16	32	64	128	16	32	64	128
Baseline (no FRT/SBO/MPA)	80.500	81.488	81.819	81.861	60.427	65.008	67.206	68.555	66.733	70.423	72.274	72.925
+ FRT	80.803	81.845	82.540	82.749	60.433	65.211	67.939	69.642	68.008	71.311	73.886	74.997
+ FRT + SBO	81.013	82.122	83.005	83.374	60.434	65.391	68.592	70.733	68.635	71.885	75.106	76.834
Full MultiBit	**81.057**	**82.202**	**83.162**	**83.592**	**60.449**	**65.426**	**68.846**	**70.965**	**69.178**	**72.177**	**75.921**	**77.372**

Component Ablation. As shown in Table 2, each component improves Mean mAP in all twelve dataset–length settings. At 128 bits, FRT adds 0.888, 1.087, and 2.072 pp on MIRFlickr-25K, IAPR TC-12, and MS-COCO, SBO a further 0.625, 1.091, and 1.837 pp, and MPA less; the last 64 bits, which span scales 1–4, gain 0.43/2.12/1.45 pp with the full model against 0.04/1.35/0.65 for the baseline. Query-level gains are positive against every design: 43.5% to 70.1% of the 1,500 queries in each group improve, with a mean gain of 0.03 to 1.18 pp (Figure 4; Appendix Table D-2).

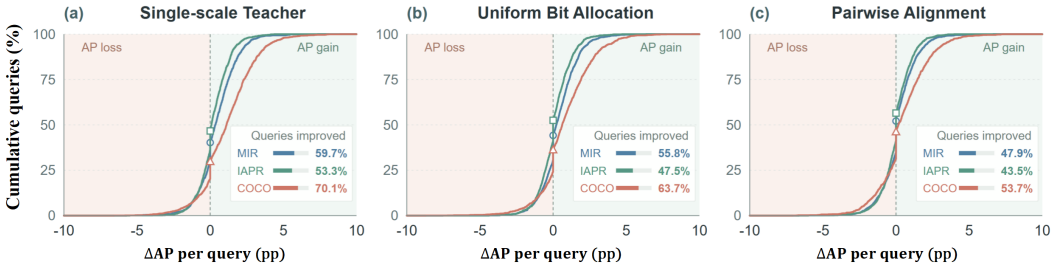

Figure 4: Query-level comparison of Full MultiBit with the three alternative designs on the 64-bit code, over 1,500 queries per dataset: cumulative distributions of the per-query AP difference ΔAP, with the improved share in the legend.

4.4 Relevance Resolution Analysis

Relevance Resolution. Table 3 extends Table 1 with CMCL and VH-ABL; Appendix Table E-3 adds RCMH and Naive Prefix. MultiBit achieves the highest NDCG@100 (Järvelin & Kekäläinen, 2002) and Kendall's τ_b (Kendall, 1938) and the lowest collision and inversion rates, indicating better preservation of the label-derived graded order in finite Hamming spaces. Because this evaluator shares its prior with the relation teacher, we repeat the comparison under a frequency-free label-name reference, where MultiBit again leads on every metric in all nine settings and FRT remains the strongest teacher (Appendix Sections E.1 and E.2; Tables E-1 and E-4).

Parameter Sensitivity. Figure 5 shows a broad high-performance region around intermediate values of α and η on every dataset, whereas extreme settings degrade moderately; FRT therefore does not depend on a narrow configuration (complete grid in Appendix Table E-6).

Table 3: Relevance resolution of MultiBit and recent methods on three datasets at 16, 32, and 64 bits, under the IDF-weighted evaluator.

Dataset	Method	16 bits				32 bits				64 bits			
		N↑	τ_b ↑	C↓	I↓	N↑	τ_b ↑	C↓	I↓	N↑	τ_b ↑	C↓	I↓
MIRFlickr-25K	DECH	0.632	0.503	0.396	0.254	0.650	0.540	0.374	0.237	0.666	0.557	0.356	0.220
	DGHDGH	0.624	0.489	0.405	0.260	0.644	0.527	0.382	0.243	0.659	0.548	0.363	0.228
	CMCL	0.643	0.518	0.387	0.246	0.662	0.554	0.361	0.228	0.677	0.568	0.343	0.212
	VH-ABL	0.651	0.531	0.378	0.241	0.670	0.564	0.352	0.220	0.686	0.579	0.333	0.204
	MultiBit	**0.660**	**0.555**	**0.366**	**0.232**	**0.678**	**0.573**	**0.342**	**0.214**	**0.697**	**0.590**	**0.321**	**0.195**
IAPR TC-12	DECH	0.611	0.424	0.420	0.311	0.634	0.451	0.397	0.291	0.653	0.468	0.375	0.273
	DGHDGH	0.604	0.408	0.429	0.317	0.627	0.447	0.404	0.295	0.646	0.463	0.382	0.279
	CMCL	0.620	0.443	0.408	0.302	0.647	0.460	0.383	0.281	0.663	0.478	0.365	0.263
	VH-ABL	0.629	0.450	0.397	0.295	0.656	0.472	0.373	0.273	0.676	0.489	0.350	0.253
	MultiBit	**0.640**	**0.461**	**0.385**	**0.285**	**0.668**	**0.481**	**0.361**	**0.264**	**0.688**	**0.499**	**0.339**	**0.244**
MS-COCO	DECH	0.672	0.497	0.372	0.263	0.690	0.515	0.349	0.244	0.707	0.531	0.330	0.226
	DGHDGH	0.665	0.492	0.380	0.269	0.683	0.509	0.356	0.249	0.700	0.524	0.336	0.232
	CMCL	0.684	0.508	0.360	0.253	0.704	0.526	0.336	0.233	0.719	0.542	0.316	0.216
	VH-ABL	0.692	0.516	0.351	0.247	0.712	0.537	0.326	0.226	0.729	0.553	0.307	0.209
	MultiBit	**0.700**	**0.528**	**0.340**	**0.237**	**0.721**	**0.547**	**0.316**	**0.216**	**0.739**	**0.564**	**0.297**	**0.199**

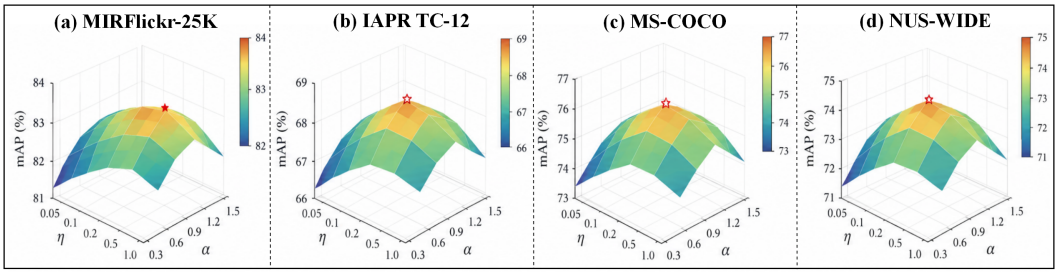

Figure 5: Validation Mean mAP (%) under different combinations of the FRT hyperparameters α and η at 64 bits on all datasets. Red stars mark the best configuration for each dataset, and the response surfaces form a broad high-performance region around intermediate values of both parameters.

Table 4: Retrieval accuracy and deployment cost of serving four code lengths on three datasets (Mean mAP in %). Models gives the number of trained models needed for all four lengths, Model MB the trainable-component size excluding the frozen CLIP backbone, Code MiB the packed code payload, Train (h) the cost of all four lengths, and Query (ms) the 64-bit query time excluding the CLIP forward pass.

Dataset	Method	Mean mAP (%)				Models	Model MB	Code MiB	Train (h)	Query (ms)
		16	32	64	128					
MIRFlickr-25K	Naive Prefix	78.765	80.922	82.311	83.284	1	10.67	**0.458**	0.428	0.0080
	CMCL	**81.107**	**82.312**	82.856	82.946	1	55.16	0.859	0.262	0.0861
	RCMH	72.457	72.851	73.152	73.348	1	5.72	0.859	2.527	0.0024
	VH-ABL	77.621	78.835	80.001	80.657	1	**4.74**	**0.458**	0.246	0.0025
	MultiBit-Sep	80.921	82.104	83.090	83.478	4	41.92	0.859	1.674	0.0058
	MultiBit	81.057	82.202	**83.162**	**83.592**	1	10.84	**0.458**	0.542	0.0092
IAPR TC-12	Naive Prefix	57.884	62.744	66.420	70.422	1	10.67	**0.446**	0.451	0.0069
	CMCL	58.374	63.577	68.216	**71.198**	1	55.16	0.837	0.269	0.0855
	RCMH	55.797	58.386	59.635	60.231	1	5.72	0.837	2.690	0.0022
	VH-ABL	55.964	60.790	64.601	66.192	1	**4.74**	**0.446**	0.249	0.0026
	MultiBit-Sep	60.301	65.291	68.701	70.861	4	41.92	0.837	1.781	0.0058
	MultiBit	**60.449**	**65.426**	**68.846**	70.965	1	10.84	**0.446**	0.579	0.0076
MS-COCO	Naive Prefix	64.201	68.693	73.808	76.955	1	10.67	**3.577**	0.463	0.0088
	CMCL	**70.090**	**72.908**	75.215	75.702	1	55.16	6.707	0.802	0.0901
	RCMH	67.427	70.653	73.041	75.157	1	5.72	6.707	2.602	0.0053
	VH-ABL	63.523	70.228	73.583	75.322	1	**4.74**	**3.577**	0.213	0.0058
	MultiBit-Sep	68.960	71.994	75.764	77.284	4	41.92	6.707	1.834	0.0089
	MultiBit	69.178	72.177	**75.921**	**77.372**	1	10.84	**3.577**	0.601	0.0096

4.5 Multi-Length Consistency and Efficiency

Accuracy and Deployment Cost. Table 4 reports the full comparison. Averaged over the four code lengths, MultiBit attains the highest Mean mAP on all three datasets (82.50, 66.42, and 73.66) and stays above the four separately trained models of MultiBit-Sep at all twelve settings. Across individual lengths it is competitive or better, with CMCL ahead in five of the twelve cells by at most 0.92 pp and MultiBit leading in the other seven. One nested model therefore replaces four at roughly 74% and 47% less model and code storage.

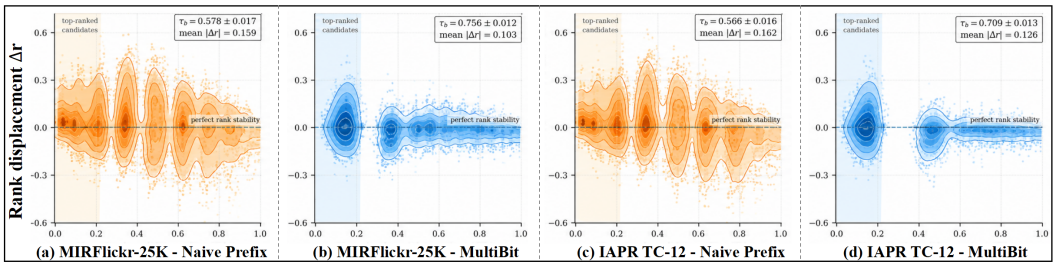

Figure 6: I2T rank shifts of Naive Prefix and MultiBit between 16- and 128-bit codes on two datasets, with five sampled queries per panel and annotated tie-corrected τ_b and mean $|\Delta r|$.

Cross-Length Ranking Consistency. Figure 6 visualizes I2T rank changes from 16 to 128 bits. Against Naive Prefix, the displacements of MultiBit concentrate closer to zero (mean $|\Delta r|$ 0.103 versus 0.159 and 0.126 versus 0.162; Appendix Table F-1), with higher tie-corrected τ_b. The same holds against every published multi-length baseline, including CMCL (0.760 versus 0.633 at 16–128 bits on MIRFlickr-25K); removing MPA lowers the six-pair I2T mean by 0.006 to 0.035. Complete results are in Appendix Table F-2.

4.6 Qualitative Retrieval Analysis

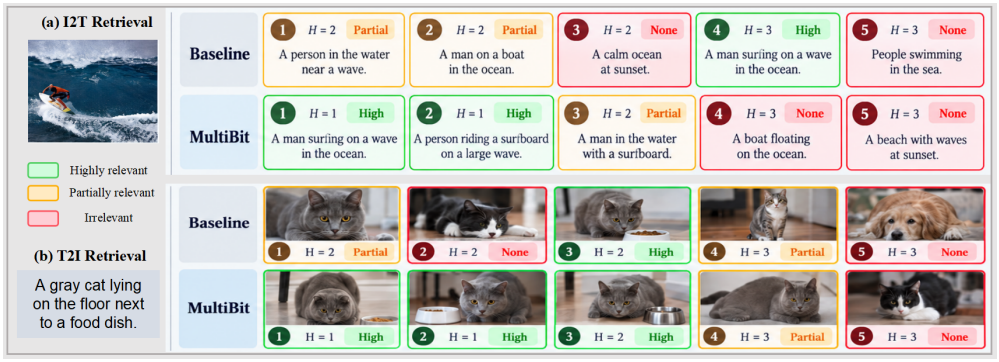

Figure 7: Illustrative 16-bit I2T and T2I rankings on MIRFlickr-25K, annotated with rank, distance, and relevance; MultiBit brings the highly relevant candidates closer than the baseline.

Qualitative Retrieval Analysis. Figure 7 shows 16-bit examples. In the image-to-text case the baseline ranks a highly relevant caption at $H = 3$ behind three weaker candidates at $H = 2$ (an inversion), and in the text-to-image case a highly relevant image shares $H = 2$ with an irrelevant one that precedes it (a collision). Equal-distance candidates are ordered by a fixed rule rather than by relevance, so the gain is distance separation, not re-ranking; the examples do not estimate prevalence.

5 Conclusion

We introduced the relevance resolution bottleneck and proposed MultiBit, which transfers relevance resolution to multi-length Hamming spaces through a scale-decomposable fractional relation teacher, scale-aware bit organization, and multi-length prefix alignment. A single nested model attains the highest Mean mAP averaged over the four code lengths on all three datasets, and attains the highest cross-length ranking agreement among all compared methods. These results indicate that organizing relations by propagation scale is a practical route to consistent multi-length binary retrieval.

AI USE STATEMENT

In this work, generative AI tools were only used to edit the manuscript for improved readability and linguistic quality. Generative AI was not used for data generation, theoretical model development, formulation or proof of mathematical claims, method design and implementation, or result interpretation; none of these disclosure-required tasks apply to this work. All AI-suggested edits were reviewed against the original text by the authors, who take full responsibility for the final manuscript content, including all text, claims, and research artifacts.

REFERENCES

Yuan Cao, Xiangru Chen, Zifan Liu, Wenzhe Jia, Fanlei Meng, and Jie Gui. Deep graph online hashing for multi-label image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1953–1961, 2025. doi: 10.1609/aaai.v39i2.32191.

Ben Chamberlain, James Rowbottom, Maria I. Gorinova, Michael Bronstein, Stefan Webb, and Emanuele Rossi. GRAND: Graph neural diffusion. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1407–1418. PMLR, 2021.

Kun Cheng, Qibing Qin, Wenfeng Zhang, Lei Huang, and Jie Nie. Deep probabilistic binary embedding via learning reliable uncertainty for cross-modal retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 6393–6402, 2025. doi: 10.1145/3746027.3754811.

Kun Cheng, Qibing Qin, Wenfeng Zhang, Lei Huang, and Jie Nie. Deep global-sense hard-negative discriminative generation hashing for cross-modal retrieval. In *International Conference on Learning Representations (ICLR)*, 2026a.

Kun Cheng, Qibing Qin, Wenfeng Zhang, Lei Huang, and Jie Nie. Deep distance weighted sampling hashing for cross-modal retrieval. *IEEE Transactions on Multimedia*, 28:4194–4206, 2026b. doi: 10.1109/TMM.2026.3654452.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. NUS-WIDE: A real-world web image database from national university of singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 1–9, 2009. doi: 10.1145/1646396.1646452.

Gene H. Golub and John H. Welsch. Calculation of Gauss quadrature rules. *Mathematics of Computation*, 23(106):221–230, 1969. doi: 10.1090/S0025-5718-69-99647-1.

Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. The IAPR TC-12 benchmark: A new evaluation resource for visual information systems. In *International Workshop on OntoImage*, 2006.

Shuo Han, Qibing Qin, Kezhen Xie, Wenfeng Zhang, and Lei Huang. Polysemic semantic instance network for cross-modal hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4592–4600, 2026. doi: 10.1609/aaai.v40i6.42459.

Senyu Hou, Yuru Ren, Gaoxia Jiang, and Wenjian Wang. Noisy multi-label learning through co-occurrence-aware diffusion. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, pp. 161117–161146, 2025. doi: 10.52202/085713-4857.

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

Yadong Huo, Qibing Qin, Jiangyan Dai, Wenfeng Zhang, Lei Huang, and Chengduan Wang. Deep neighborhood-aware proxy hashing with uniform distribution constraint for cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 20(6):1–23, 2024a. doi: 10.1145/3643639.

Yadong Huo, Qibing Qin, Wenfeng Zhang, Lei Huang, and Jie Nie. Deep hierarchy-aware proxy hashing with self-paced learning for cross-modal retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):5926–5939, 2024b. doi: 10.1109/TKDE.2024.3401050.

Seong-Hyeon Hwang, Minsu Kim, and Steven Euijong Whang. T-CIL: Temperature scaling using adversarial perturbation for calibration in class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15339–15348, 2025. doi: 10.1109/CVPR52734.2025.01429.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002. doi: 10.1145/582415.582418.

Kaihang Jiang, Wai Keung Wong, Xiaozhao Fang, Weijun Sun, Guoxu Zhou, Shengli Xie, and Xiaochun Cao. An efficient regenerated cross-modal hashing: Improving existing hash codes with the arbitrary length. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(9): 10382–10400, 2026. doi: 10.1109/TPAMI.2026.3688816.

M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1–2):81–93, 1938. doi: 10.1093/biomet/30.1-2.81.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. doi: 10.1214/aoms/1177729694.

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems 27 (NIPS)*, pp. 2177–2185, 2014.

Yuan Li, Liangli Zhen, Yuan Sun, Dezhong Peng, Xi Peng, and Peng Hu. Deep evidential hashing for trustworthy cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18566–18574, 2025. doi: 10.1609/aaai.v39i17.34043.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

Sohir Maskey, Raffaele Paolino, Aras Bacho, and Gitta Kutyniok. A fractional graph Laplacian approach to oversmoothing. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pp. 13022–13063, 2023. doi: 10.52202/075280-0571.

Hui Meng, Huaxiang Zhang, Li Liu, Lei Zhu, Xiaojun Chang, and Xinru Guo. Scene graph knowledge enhanced hashing with contrastive learning for image-text retrieval. *IEEE Transactions on Multimedia*, 28:7719–7733, 2026. doi: 10.1109/TMM.2026.3673545.

Likang Peng, Chao Su, Wenyuan Wu, Yuan Sun, Dezhong Peng, Xi Peng, and Xu Wang. Semantic-consistent bidirectional contrastive hashing for noisy multi-label cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 24811–24819, 2026. doi: 10.1609/aaai.v40i29.39667.

Qibing Qin, Meiling Ge, Wenfeng Zhang, Lei Huang, and Jie Nie. Deep stochastic spherical hashing with Von Mises–Fisher distributions for cross-modal retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 38(5):2655–2669, 2026a. doi: 10.1109/TKDE.2026.3665949.

Qibing Qin, Lei Wu, Wenfeng Zhang, Huihui Zhang, Lei Huang, and Jie Nie. Deep semantic tuplet-based hashing by hypergraph modeling for cross-modal retrieval. *IEEE Transactions on Multimedia*, 28:657–670, 2026b. doi: 10.1109/TMM.2025.3623497.

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, 2021.

Chao Su, Likang Peng, Yuan Sun, Dezhong Peng, Xi Peng, and Xu Wang. Neighbor-aware contrastive disambiguation for cross-modal hashing with redundant annotations. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, pp. 120168–120190, 2025. doi: 10.52202/085713-3620.

Chao Su, Yanan Li, Xu Wang, Yingke Chen, Huiming Zheng, Dezhong Peng, and Yuan Sun. Ambiguity-tolerant cross-modal hashing with partial labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 25636–25644, 2026. doi: 10.1609/aaai.v40i30.39760.

Wentao Tan, Fengling Li, Lei Zhu, Weili Guan, Jingjing Li, Zhiyong Cheng, and Heng Tao Shen. Dynamic bit-wise semantic transformer hashing for multi-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):2954–2969, 2026. doi: 10.1109/TPAMI.2025.3630209.

Junfeng Tu, Xueliang Liu, Zhen Huang, Yanbin Hao, Richang Hong, and Meng Wang. Cross-modal hashing via diverse instances matching. *IEEE Transactions on Image Processing*, 34:2737–2749, 2025. doi: 10.1109/TIP.2025.3561659.

Min Wang, Wengang Zhou, Xin Yao, and Houqiang Li. Adaptive bit selection for scalable deep hashing. *IEEE Transactions on Image Processing*, 34:1048–1059, 2025. doi: 10.1109/TIP.2025.3533215.

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. *Proceedings of the IEEE*, 112(11):1716–1754, 2024. doi: 10.1109/JPROC.2024.3525147.

Qingpeng Wu, Zheng Zhang, Yishu Liu, Jingyi Zhang, and Liqiang Nie. Contrastive multi-bit collaborative learning for deep cross-modal hashing. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):5835–5848, 2024. doi: 10.1109/TKDE.2024.3419577.

Zhan Yang, Binghong Chen, Jiajun Tang, and Yinan Li. Unsupervised similarity-fusion transformer hashing for multimodal retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM)*, pp. 5932–5941, 2025. doi: 10.1145/3746027.3754753.

Tao Yao, Ze Li, Fengling Li, Jing Zhang, Lei Zhu, and Qi Tian. Composite proxy-based hierarchical contrastive hashing for cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(9):13038–13049, 2026. doi: 10.1109/TCSVT.2026.3700328.

Soyoung Yoon, Gyuwan Kim, Gyu-Hwung Cho, and Seung-won Hwang. AcuRank: Uncertainty-aware adaptive computation for listwise reranking. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, pp. 29944–29973, 2025. doi: 10.52202/085713-0890.

Shuang Zhang, Yue Wu, Lei Shi, Huilong Jin, Feifei Kou, Pengfei Zhang, Mingying Xu, and Pengtao Lv. Adaptive graph attention based discrete hashing for incomplete cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28382–28390, 2026a. doi: 10.1609/aaai.v40i33.40067.

Shuang Zhang, Yue Wu, Lei Shi, Feifei Kou, Huilong Jin, Pengfei Zhang, Weiping Ding, Mingying Xu, and Muhammet Deveci. Dual graph network hashing for cross-modal retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 38(9):5781–5797, 2026b. doi: 10.1109/TKDE.2026.3710651.

Zhicheng Zhang, Jiwei Tang, Kuicai Dong, Xiaopeng Li, Jieming Zhu, Jingyu Li, Qianhui Zhu, Fengyuan Lu, Jiaheng Wang, Gang Wang, Hai-Tao Zheng, and Zhaocheng Du. When hard negatives hurt: Bridging the generative-discriminative gap in hard negative synthesis for retrieval. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 6688–6699, 2026c. doi: 10.1145/3770855.3818118.

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 (CVPR)*, pp. 19649–19658, 2025. doi: 10.1109/CVPR52734.2025.01830.

Ruifan Zuo, Chaoqun Zheng, Fengling Li, Tianshi Wang, Tian Gan, Wenpeng Lu, and Xiaoming Wu. Beyond fixed-length constraints: Variable-length hashing via adaptive bit-importance learning for multi-modal retrieval. *IEEE Transactions on Multimedia*, pp. 1–13, 2026. doi: 10.1109/TMM.2026.3729387.

Appendix

Relevance-Resolution Transfer via Scale-Decomposable Fractional Diffusion for Multi-Length Cross-Modal Hash Retrieval

A Method Notation and Experimental Settings

A.1 Data Split, Evaluation Metrics, and Implementation Details

Data Split. All methods share the same split, drawn with the fixed random seed 42, so that the training, validation, query, and database indices are identical across methods and the comparison is fair. Table A-1 summarizes the resulting partitions. The retrieval database is the full dataset minus the query set; it therefore contains the $10{,}000$ training and validation pairs whose labels are used for supervision, and Table A-1 lists the resulting database sizes. The $1{,}000$ validation pairs are disjoint from the $5{,}000$ queries and are used only to select the checkpoint, never as reported results.

Table A-1: Data split shared by all compared methods.

Dataset	Pairs	Classes	Train	Validation	Query	Database
MIRFlickr-25K	20,015	24	9,000	1,000	5,000	15,015
NUS-WIDE	195,834	21	9,000	1,000	5,000	190,834
MS-COCO	122,218	80	9,000	1,000	5,000	117,218
IAPR TC-12	19,627	291	9,000	1,000	5,000	14,627

Evaluation Metrics. We report Image-to-Text (I2T) and Text-to-Image (T2I) retrieval and their arithmetic mean (Mean). The primary metric is mAP@all: for each query, average precision is computed over the complete retrieval database with k set to the database size, and the per-query values are then averaged over the whole query set. The resolution analysis additionally uses graded relevance, NDCG@100, tie-corrected Kendall's τ_b, and the collision and inversion rates defined in Section E.1; the multi-length analysis uses the same τ_b together with the rank-displacement measure defined in Section F.1.

Implementation Details. All experiments are implemented in PyTorch and run on a single NVIDIA GeForce RTX 4090 GPU. All methods adopt CLIP ViT-B/32 as the shared backbone. Code lengths are 16, 32, 64, and 128 bits. Models are trained with Adam for 100 epochs at a batch size of 128, with an initial learning rate of 0.001 and a cosine annealing schedule, and the checkpoint with the best validation Mean mAP is selected. The FRT hyperparameters α and η are selected on the validation set. The final evaluation is performed on the complete retrieval database. Listwise teacher and student distributions are normalized over the current mini-batch, and no memory bank is used. All compared baselines are our own reproductions retrained under this shared protocol. The ablation baseline in Table 2 shares the backbone, encoder, data split, and training schedule of MultiBit, but is trained with the standard cross-modal hashing objective on the CLIP features only: it uses no relation teacher, no scale-aware bit organization, and neither the scale objective nor the prefix objective, and its results at 16, 32, and 64 bits are obtained by training a single 128-bit model and truncating it with the same prefix operator P_B.

Backbone, Heads, and Per-Baseline Settings. The CLIP ViT-B/32 backbone is frozen and only the projection and the scale hash heads are trained. Each of the $M = 4$ scale heads is an independent linear map from the shared 512-dimensional feature to its own m_ℓ dimensions, with no hidden layer and no parameters shared across heads. All four code lengths are read from the same model: one checkpoint is selected on validation and every length is the prefix of its maximum-length code, and the validation criterion is the mean over the four target lengths of the validation Mean mAP. The compared baselines are trained with their own recommended optimizers and learning rates rather than one shared setting (Table A-2); only the data split, the number of epochs, the batch size and the evaluation protocol are shared.

Table A-2: Per-baseline optimizers and learning rates, taken from the original implementations. The data split, the number of epochs, the batch size and the evaluation protocol are shared across methods.

Method	Optimizer	Learning rate
DECH, DGHDGH, DIMCH, DPBE, DPSIH, DSTH, RCMH	AdamW	1×10^{-4}
DNPH	BertAdam	1×10^{-3}
CMCL	BertAdam	0.001 (MIRFlickr-25K) / 0.002 (MS-COCO, NUS-WIDE)
DSSH	Adam	0.1; CLIP learning rate 0.001
VH-ABL	AdamW	1×10^{-4}

A.2 Complete Parameter Register

A.3 Formal Statement of the Relevance Resolution Bottleneck

Let $\mathcal{R}$ be the retrieval database, r_{qi} the graded relevance of candidate i to query q, and π_q^* the ranking of $\mathcal{R}$ by descending r_{qi}. At code length B, let $d_H^{(B)}(q,i)$ be the Hamming distance and $\pi_q^{(B)}$ the ranking by ascending $d_H^{(B)}$. We say that the code preserves relevance resolution at length B if

$$r_{qi} > r_{qj} \quad \Longrightarrow \quad d_H^{(B)}(q,i) < d_H^{(B)}(q,j), \qquad \forall\, i,j \in \mathcal{R}, \tag{A.1}$$

then the code preserves relevance resolution at length B, and an ideal transfer satisfies this simultaneously for every $B \in \mathcal{B}$. The Relevance Resolution Bottleneck (RRB) is the failure of this condition. A B-bit code admits only $B+1$ distinct Hamming distances, so Equation (A.1) is a capacity bound that no encoding can exceed. We therefore report collision and inversion rates as measured quantities and compare them across methods. This failure arises in three forms: (i) a *collision*, where $r_{qi} > r_{qj}$ but $d_H^{(B)}(q,i) = d_H^{(B)}(q,j)$; (ii) an *inversion*, where $r_{qi} > r_{qj}$ but $d_H^{(B)}(q,i) > d_H^{(B)}(q,j)$; and (iii) a *cross-length inconsistency*, where the relative order of i and j differs between two target lengths B and B'.

B Detailed Formulations and Derivations

Label Graph Construction. Let $n_a = \sum_i y_{ia}$ and $n_{ab} = \sum_i y_{ia} y_{ib}$ denote the frequency of label a and the co-occurrence frequency of labels a and b, respectively. The PPMI label graph $W \in \mathbb{R}^{C \times C}$ used in Section 3.3 is constructed as

$$W_{ab} = \mathbb{I}[n_{ab} > 0] \max\left\{0, \log \frac{(n_{ab}+\varepsilon)N}{(n_a+\varepsilon)(n_b+\varepsilon)}\right\}, \quad a \neq b, \qquad W_{aa} = 0, \tag{B.1}$$

where $\varepsilon > 0$ is the smoothing constant.

Scale-Aware Subblocks. Let E^m be the encoder and projection for modality $m \in \{I, T\}$, mapping an input to a shared d-dimensional feature $z_i^m = E^m(x_i^m)$, and let $H_\ell : \mathbb{R}^d \to \mathbb{R}^{m_\ell}$ be the hash head of scale ℓ. Sorting the scales from local to long range as $(\sigma_1, \ldots, \sigma_M) = \mathrm{argsort}(\tau_1, \ldots, \tau_M)$, the maximum-length representation used in Section 3.4 is

$$\widetilde{h}_i^{m,(\ell)} = \tanh(H_\ell(z_i^m)) \in (-1,1)^{m_\ell}, \qquad \widetilde{h}_i^m = \left[\widetilde{h}_i^{m,(\sigma_1)} \,\|\, \cdots \,\|\, \widetilde{h}_i^{m,(\sigma_M)}\right] \in (-1,1)^{B_{\max}}, \tag{B.2}$$

With both configurations the 16-, 32- and 64-bit prefixes lie entirely inside the first subblock; only the 128-bit code combines more than one diffusion scale. SBO therefore acts at two levels: inside the first subblock it reallocates capacity among local scale directions, which benefits every deployable length, and beyond it, it combines the remaining scales, which only the 128-bit code can do. Relative to the +FRT configuration of Table 2, adding SBO raises Mean mAP by $0.210/0.277/0.465$ at 16/32/64 bits and 0.625 at 128 bits on MIRFlickr-25K, by $0.001/0.180/0.653$ and 1.091 on IAPR TC-12, and by $0.627/0.574/1.220$ and 1.837 on MS-COCO. The cross-scale combination is exercised at the maximum length, where the gain is largest. Table D-2 provides the same-capacity control: with the total budget and the subblock structure held fixed and the bits split evenly over the four scales, Full MultiBit improves on 47.5–63.7% of queries with a mean ΔAP of 0.11–0.85 pp.

Resolution improves monotonically with the code length on all three datasets (Table B-2). The last 64 bits, which complete the first subblock and then cover scales 2–4, add 0.032 NDCG@100, 0.028 Kendall's τ_b, -0.036 collision and -0.026 inversion on MIRFlickr-25K, with comparable changes

Table A-3: Notation, model controls, and experimental settings used throughout the manuscript.

Group	Symbol / setting	Definition	Value or setting
Data	N, C	Training-pair count; label count	See Table A-1
Data	$\mathcal{B}, K, B_{\max}$	Target lengths; number of lengths; maximum length	$\{16, 32, 64, 128\}$, 4, 128
Data	d	Shared feature dimension before hash heads	512
Data	P_B	First-B-bit prefix selection	Determined by B and $B_{\max}$
FRT	ε	PPMI smoothing constant	0.1
FRT	W, D, L	PPMI label graph, degree matrix, normalized Laplacian	Derived from training labels
FRT	α	Fractional diffusion order	1.20 (MIRFlickr-25K); 0.90 (others)
FRT	η	Positive kernel shift and diffusion-range control	0.20
FRT	M	Gauss–Laguerre quadrature points / retained scales	4
FRT	τ_ℓ, ρ_ℓ	Diffusion times and quadrature weights, local to long range	α=1.20: $\tau = (2.00, 9.55, 23.89, 48.56)$, $\rho = (3.88, 2.70, 0.32, 0.005)$; α=0.90: $\tau = (1.43, 8.32, 22.08, 46.18)$, $\rho = (2.67, 1.44, 0.15, 0.002)$
FRT	$s_a, \widetilde{y}_i$	Label specificity and weighted label vector	Derived from training-label frequencies
FRT	$K_{\alpha,\eta}, g^*, g_\ell$	Fractional kernel; integrated and scale teacher scores	Derived by Eqs. (2)–(5)
SBO	m_ℓ	Integer bits assigned to scale ℓ, local to long range	$(71, 50, 6, 1)$ for α=1.20 and $(80, 43, 4, 1)$ for α=0.90, with $\sum_\ell m_\ell = B_{\max}$
MPA	T_t, T_s	Teacher and student Softmax temperatures	0.2 / 0.1
MPA	β_B	Prefix-loss weight at length B	$\sqrt{B}/\sum_{B'}\sqrt{B'}$; $(0.138, 0.195, 0.276, 0.391)$ for 16/32/64/128
MPA	λ_s	Scale-loss coefficient	0.7
MPA	$\mathcal{C}_q$	Shared cross-modal candidate set for listwise matching	Current mini-batch
MPA	$P^*, P_{\theta,B}$	Teacher and prefix-student candidate distributions	Defined in Section 3.5
Training	Backbone / hardware	Common encoder and accelerator	CLIP ViT-B/32; one RTX 4090
Training	Optimizer / schedule	Parameter update and learning-rate schedule	Adam; initial LR 0.001; cosine annealing
Training	Batch / epochs	Training batch and maximum epochs	128 / 100
Training	Checkpoint	Validation criterion	Best validation Mean mAP
Evaluation	$\mathcal{R}, \mathcal{P}_q, r_{qi}$	Database, pairwise candidate pool, independent graded relevance	Defined in Section E.1

Table B-1: Scale coverage of the four deployable code lengths. The bit budgets m_ℓ follow from the normalized Gauss–Laguerre weights, and the cumulative column counts the bits up to and including each scale.

α	Scale ℓ	τ_ℓ	ρ_ℓ	m_ℓ	Cumulative
1.20	1	2.00	3.88	71	71
	2	9.55	2.70	50	121
	3	23.89	0.32	6	127
	4	48.56	0.005	1	128
0.90	1	1.43	2.67	80	80
	2	8.32	1.44	43	123
	3	22.08	0.15	4	127
	4	46.18	0.002	1	128

Table B-2: Relevance resolution of MultiBit as a function of the code length, sampled at the scale boundaries of Table B-1, measured under the IDF-weighted evaluator.

Dataset	α	B	**N**↑	τ_b ↑	**C**↓	**I**↓
MIRFlickr-25K	1.20	64	0.697	0.590	0.321	0.195
		71	0.702	0.595	0.315	0.191
		121	0.724	0.613	0.291	0.174
		127	0.728	0.617	0.286	0.170
		128	0.729	0.618	0.285	0.169
IAPR TC-12	0.90	64	0.688	0.499	0.339	0.244
		80	0.697	0.506	0.330	0.237
		123	0.716	0.522	0.307	0.220
		127	0.719	0.525	0.303	0.217
		128	0.720	0.526	0.302	0.216
MS-COCO	0.90	64	0.739	0.564	0.297	0.199
		80	0.746	0.570	0.289	0.193
		123	0.766	0.586	0.268	0.177
		127	0.769	0.589	0.264	0.174
		128	0.770	0.590	0.263	0.173

on the other two datasets. Because this diagnostic lengthens the code and thereby increases scale coverage and bit count together, it isolates the effect of code length rather than the contribution of long-range scales alone. with $b_i^{m,(\ell)} = \operatorname{sign}(\widetilde{h}_i^{m,(\ell)})$.

Directional Listwise Objective. For a query set $\mathcal{Q}_m$, the directional loss of retrieval direction $m \to n$ at code length B used in Section 3.5 is

$$\mathcal{L}_B^{m\to n} = \frac{1}{|\mathcal{Q}_m|} \sum_{q\in\mathcal{Q}_m} D_{\mathrm{KL}}\big(P^*(\cdot \mid q) \,\|\, P_{\theta,B}^{m\to n}(\cdot \mid q)\big)\,. \tag{B.3}$$

Design Rationale. PPMI retains co-occurrences stronger than their independent expectation instead of letting frequent labels dominate the graph, and the specificity weight s_a modulates each label by its frequency so that rare labels remain informative. In Eq. (2), e^{-tL} is the heat diffusion operator at scale t: small t emphasizes local label neighborhoods, whereas large t incorporates more distant dependencies, and fractional orders of this form interpolate between local and anomalous long-range transport. The cosine normalization in Eq. (3) reduces the influence of label cardinality, so that directly overlapping and indirectly connected labels contribute comparably to a query-conditioned relevance score; it also means that g^* is not a linear combination of the scale teachers $\{g_\ell\}$. Accordingly, ρ_ℓ should be read as the coefficient of K_ℓ in the kernel mixture rather than as a per-scale contribution to the final teacher score. Each τ_ℓ in Eq. (4) specifies a diffusion range and $\rho_\ell > 0$ is its coefficient in the integrated kernel; both follow from the fractional kernel rather than manually sampled diffusion times. Because the common factor $\eta^{-\alpha}/\Gamma(\alpha)$ cancels under the normalization in Eq. (6), the bit ratio reflects the Gauss–Laguerre weights, and hence the order α, rather than a dataset-specific measurement of graph response; we therefore treat it as a prior over scales rather than a calibrated quantity, and evaluate it against uniform bit allocation in Table D-2. The normalized weight $\bar{\rho}_\ell$ in Eq. (9) reuses the same prior to weight the per-scale objectives, and likewise does not measure how much each g_ℓ contributes to g^*.

B.1 Quantization Statistics

The property in Eq. (10) is stated for the continuous relaxation $\widetilde{h}$. To measure how much of the continuous ordering survives the sign operation, we evaluate the relaxed representation $\widetilde{h} = \tanh(F(x)) \in (-1,1)^{B_{\max}}$ of the query set and of the retrieval database in each direction and report two statistics for every code length B. The mean absolute deviation $\bar{\epsilon}(B) = \frac{1}{NB}\sum_{i=1}^{N}\sum_{k=1}^{B}\big||\widetilde{h}_{ik}|-1\big|$ measures how far the relaxed values are from the binary values ± 1 they are quantized to, and the tie-corrected Kendall τ_b between the ranking induced by the soft Hamming distance $\frac{1}{2}\big(B - \widetilde{h}_q^{(B)\mathsf{T}}\widetilde{h}_i^{(B)}\big)$ and the ranking induced by the discrete Hamming distance of Eq. (1) measures how much of the soft ordering is preserved after quantization. Here N is the number of query and database instances and $\widetilde{h}^{(B)}$ denotes the first B dimensions of $\widetilde{h}$. Throughout, the distance induced by $\widetilde{h}$ is called the soft Hamming distance and its ranking the soft ranking, in contrast to the discrete Hamming distance and ranking of Eq. (1).

Table B-3: Quantization statistics of the relaxed representation. $\bar{\epsilon}$ is the mean absolute deviation from the binary values, computed over the query set and the retrieval database; τ_b compares the ranking induced by the soft Hamming distance with the ranking induced by the discrete Hamming distance.

Dataset	**Direction**	$\bar{\epsilon}$				τ_b (soft vs. discrete)			
		16	**32**	**64**	**128**	**16**	**32**	**64**	**128**
MIRFlickr-25K	I2T	0.0187	0.0215	0.0258	0.0314	0.907	0.931	0.948	0.960
	T2I	0.0191	0.0220	0.0264	0.0321	0.903	0.928	0.945	0.957
IAPR TC-12	I2T	0.0219	0.0254	0.0308	0.0372	0.899	0.923	0.941	0.953
	T2I	0.0225	0.0261	0.0316	0.0380	0.896	0.920	0.938	0.950
MS-COCO	I2T	0.0168	0.0195	0.0237	0.0289	0.912	0.936	0.952	0.964
	T2I	0.0172	0.0200	0.0242	0.0296	0.909	0.933	0.949	0.961

Table B-4: Tie-only upper bound on the soft-to-discrete ranking agreement. $S(B)$ is the fraction of candidate pairs that share the same discrete Hamming distance; because only the discrete side has ties, even a comparison in which the sign operation perturbs no ordering at all cannot push the tie-corrected τ_b above $\sqrt{1-S(B)}$.

Dataset	B	$S(B)$	$\sqrt{1-S(B)}$	**Observed τ_b (I2T)**	**Gap**
MIRFlickr-25K	16	0.1587	0.917	0.907	0.010
	32	0.1189	0.939	0.931	0.008
	64	0.0856	0.956	0.948	0.008
	128	0.0635	0.968	0.960	0.008
IAPR TC-12	16	0.1718	0.910	0.899	0.011
	32	0.1324	0.931	0.923	0.008
	64	0.0953	0.951	0.941	0.010
	128	0.0736	0.962	0.953	0.009
MS-COCO	16	0.1498	0.922	0.912	0.010
	32	0.1107	0.943	0.936	0.007
	64	0.0781	0.960	0.952	0.008
	128	0.0574	0.971	0.964	0.007

Table B-3 shows that $\bar{\epsilon}$ lies between 0.0168 and 0.0380, so the representation used during training stays close to the codes that are actually stored, and that the soft and discrete rankings agree with τ_b between 0.896 and 0.964; both statistics grow with the code length. If the disagreement between the two rankings were driven by the magnitude of the quantization error, then a larger deviation should go with a smaller τ_b. Instead both statistics grow with the code length, so we find no evidence that the quantization error drives this disagreement; Table B-4 points to the tie structure instead. Because the soft ranking is essentially tie-free whereas the discrete ranking places a fraction $S(B)$ of candidate pairs at equal distance, the tie-corrected denominator carries a factor $\sqrt{1-S(B)}$, so τ_b cannot exceed $\sqrt{1-S(B)}$. The measured τ_b stays within 0.007–0.011 of this bound on every dataset and code length, that is within 0.7%–1.2% of it. The agreement between the two orderings is thus governed mainly by the resolution of the B-bit alphabet, which admits only $B+1$ distinct distances; the sign operation adds only a small perturbation on top of it. Together with Table 3, this quantifies the step from the conditional property of Eq. (10) to the discrete rankings that are actually retrieved.

B.2 Quadrature Fidelity of the Scale Organization

The integrated teacher g^* and every scale teacher g_ℓ are evaluated from the exact matrix functions $(L + \eta I)^{-\alpha}$ and $e^{-\tau_\ell L}$; the Gauss–Laguerre rule of Eq. (4) supplies only the scale coordinates $\{(\tau_\ell, \rho_\ell)\}$, which set the bit budget of Eq. (6) and the relative weight of each subblock loss. Table B-5 therefore measures how faithfully that organization represents the exact kernel, by comparing g^* with the teacher induced by the kernel $\sum_\ell \rho_\ell K_\ell$ rebuilt from the same rule. The rule tracks the exact kernel closely in the low-frequency region that carries the label signal, and the residual deviation is confined to the high-frequency end. Because the comparison weights every frequency by both the label-signal energy and the kernel response, that deviation leaves the teacher nearly unchanged: the signal-weighted error is 6.1%–18.6% at $M = 4$ and at most 1.7% at $M = 8$ on all four datasets, and the induced ordering moves little, with $\tau_b \geq 0.93$ at $M = 4$ and ≥ 0.99 at $M = 8$ (Table B-5). At $M = 4$ the first scale absorbs 56% of the bit budget on MIRFlickr-25K and 63% on the other datasets (Table B-1). We therefore adopt $M = 4$ as the smallest configuration that exposes the local-to-long-range structure used by SBO.

Table B-5: Fidelity of the M-scale Gauss–Laguerre organization to the exact fractional teacher. The signal-weighted error is $\sum_k (f_{\text{ex}} - f_M)(\lambda_k) E_k / \sum_k f_{\text{ex}}(\lambda_k) E_k$, where E_k is the energy of the label signals at eigenvalue λ_k of L. The second block reports the tie-corrected Kendall τ_b between the ranking induced by the quadrature kernel $\sum_\ell \rho_\ell K_\ell$ and that of the exact teacher g^* on the same candidate pool; entries of 1.000 denote $\tau_b \geq 0.9995$.

Dataset	α	Signal-weighted error (%)				τ_b (exact vs. quadrature)			
		2	**4**	**8**	**16**	**2**	**4**	**8**	**16**
MIRFlickr-25K	1.20	14.98	6.08	0.90	0.02	0.885	0.956	0.993	1.000
IAPR TC-12	0.90	54.78	18.57	1.70	0.01	0.857	0.941	0.992	1.000
MS-COCO	0.90	36.83	13.28	1.55	0.02	0.910	0.971	0.994	1.000
NUS-WIDE	0.90	26.18	10.01	1.31	0.03	0.848	0.939	0.991	1.000

C Results for Section 4.2: Main Comparison and Efficiency

C.1 Comparison on NUS-WIDE

Table C-1: I2T and T2I mAP@all results (%) of MultiBit and existing cross-modal hashing methods on NUS-WIDE with four hash-code lengths. MultiBit is shown in bold; the best fixed-length and the best multi-length baseline are underlined.

Task	**Method**	**Reference**	**NUS-WIDE**			
			16	**32**	**64**	**128**
I2T	DSSH	TKDE'26	63.32	67.39	68.98	70.80
	DNPH	TOMM'24	65.19	68.71	70.34	70.60
	DECH	AAAI'25	<u>70.00</u>	<u>72.53</u>	<u>73.42</u>	<u>73.84</u>
	DIMCH	TIP'25	68.51	69.43	70.91	71.12
	DPBE	MM'25	62.17	64.74	67.90	71.01
	DSTH	TMM'26	67.29	68.91	71.04	71.91
	DPSIH	AAAI'26	65.64	67.04	69.40	70.96
	DGHDGH	ICLR'26	63.57	66.20	67.63	67.95
	CMCL	TKDE'24	<u>70.74</u>	<u>72.81</u>	<u>73.57</u>	<u>73.91</u>
	RCMH	TPAMI'26	67.86	69.48	70.61	71.18
	VH-ABL	TMM'26	68.92	71.06	72.45	73.16
	MultiBit	**OURS**	**70.59**	**72.74**	**73.82**	**74.23**
T2I	DSSH	TKDE'26	64.17	68.68	69.73	72.59
	DNPH	TOMM'24	66.84	69.53	71.63	73.00
	DECH	AAAI'25	<u>70.97</u>	<u>72.97</u>	<u>74.02</u>	<u>74.71</u>
	DIMCH	TIP'25	69.02	70.23	71.56	72.04
	DPBE	MM'25	62.41	65.75	68.12	71.53
	DSTH	TMM'26	67.62	69.71	71.55	72.22
	DPSIH	AAAI'26	66.48	67.78	69.64	71.31
	DGHDGH	ICLR'26	65.55	67.90	69.20	69.47
	CMCL	TKDE'24	<u>71.52</u>	<u>73.46</u>	<u>74.31</u>	<u>74.47</u>
	RCMH	TPAMI'26	68.34	70.15	71.26	71.73
	VH-ABL	TMM'26	69.61	71.72	73.08	73.65
	MultiBit	**OURS**	**71.39**	**73.39**	**74.54**	**74.82**

The MultiBit row is shown in bold; within each column the best fixed-length method and the best multi-length method are underlined. I2T denotes image-to-text retrieval, whereas T2I denotes text-to-image retrieval.

Interpretation. NUS-WIDE is reported as an additional benchmark in which the fixed-length baselines of Table 1 and the multi-length baselines CMCL, RCMH and VH-ABL are trained and evaluated under the same protocol. The multi-length methods are listed in the same order as in the main tables (CMCL, then RCMH, then VH-ABL), and the accuracy pattern repeats as well: CMCL is ahead of MultiBit at 16 and 32 bits by 0.07–0.15 pp, whereas MultiBit is ahead at 64 and 128 bits by 0.23–0.35 pp, so the long-code advantage of the nested representation extends to the largest retrieval database.

C.2 Performance–Efficiency Measurements

Table C-2: Raw performance–efficiency measurements of representative methods on MIRFlickr-25K under the 128-bit setting.

Method	Trainable Params (M) ↓	FLOPs (G) ↓	Training Time (h) ↓	Best Epoch ↓	Image Encoding (ms) ↓	Text Encoding (ms) ↓
DECH	10.60	0.0210	0.092	45	0.0022	0.0022
DIMCH	0.86	0.0021	0.167	45	0.0043	0.0043
DSTH	0.66	0.0013	**0.078**	45	0.0020	0.0020
DGHDGH	**0.33**	**0.0007**	0.144	42	0.0017	**0.0015**
MultiBit	2.71	0.0054	0.542	**39**	**0.0013**	**0.0015**

↓ indicates that lower is better; the best result in each column is shown in bold. All measurements follow the common CLIP ViT-B/32 backbone and 100-epoch training protocol described in Section 4.1.

Analysis. Table C-2 separates learnable-component complexity, offline training cost, and per-sample online hash-head and binarization time. Best Epoch is the checkpoint with the highest validation Mean mAP within 100 epochs; its smaller value means an earlier peak, not a shorter total run. At 128 bits, MultiBit uses 2.71M trainable parameters and 0.0054G FLOPs, with 0.542 h of total training, the highest time among these methods. It reaches its best validation checkpoint at epoch 39. Its image and text encoding times are 0.0013 and 0.0015 ms, respectively; the 0.0014 ms arithmetic mean is the smallest among the compared methods. The reported encoding times cover only the hash heads and binarization; the shared CLIP backbone forward pass, database search, and end-to-end retrieval latency are excluded.

D Results for Section 4.3: Component Ablation

D.1 Component Ablation Interpretation

Table 2 reports a monotonic Mean mAP increase as FRT, SBO, and then MPA are added in each of the twelve dataset–length settings; each cell is Mean mAP (%), the arithmetic mean of the I2T and T2I mAP@all values. The increments are small in several MIRFlickr-25K and IAPR TC-12 settings and larger on MS-COCO, indicating that the three components provide complementary gains rather than interchangeable effects. Because the first 16, 32 and 64 bits of the maximum-length code lie inside the first subblock (Table B-1), SBO changes the allocation only at 128 bits; at the shorter lengths its increment reflects the scale-specific supervision of that subblock, and the reallocation of bits between scales takes effect only where several subblocks are combined.

Table D-1 replaces only the relation teacher and holds the backbone, SBO, MPA and the training schedule fixed. FRT attains the highest Mean mAP in all twelve dataset–length settings, and its margin over the strongest remaining variant grows with the code length: 0.02–0.44 pp at 16 and 32 bits against 0.22–1.32 pp at 64 and 128 bits, where it is widest on MS-COCO. At 16 and 32 bits all variants lie close together, consistent with the teacher-level gaps of 0.008–0.011 NDCG@100 reported in Table E-1; on IAPR TC-12 at 16 bits FRT is in fact the only variant that improves on the baseline without a relation teacher. FRT attains its largest margins at the longer code lengths, and it does so with fewer trainable parameters than the learned-weight variant, since its four mixture weights follow from the single order parameter α through the Gauss–Laguerre quadrature instead of being trained independently. With matched nodes, normalization and architecture, UKM reaches 80.850/81.935/82.835/83.215 on MIRFlickr-25K, so replacing the fractional weights by equal weights costs $0.207/0.267/0.327/0.377$ pp at 16/32/64/128 bits; the gap grows with the code length and agrees in direction with the teacher-level comparison of Table E-2.

Table D-1: Mean mAP (%) of the relation-teacher ablation. Rows two to seven replace only the relation teacher, with the backbone, SBO, MPA and the training schedule held fixed. The first and last rows are the two reference points of Table 2, the baseline without a relation teacher and Full MultiBit, reproduced here for comparison; the relation teachers of rows four to eight are the variants compared in Tables E-1 and E-2. The uniform kernel mixture (UKM) mixes the four scale kernels with equal weights and then normalizes, whereas the uniform multi-scale heat (UMH) averages the four per-scale teachers after normalizing each; UKM is evaluated on MIRFlickr-25K, where the scale nodes coincide with those of FRT, and shares the SBO budget, MPA and the training schedule of the full model.

Variant	**MIRFlickr-25K**				**IAPR TC-12**				**MS-COCO**			
	16	**32**	**64**	**128**	**16**	**32**	**64**	**128**	**16**	**32**	**64**	**128**
Baseline	80.500	81.488	81.819	81.861	60.427	65.008	67.206	68.555	66.733	70.423	72.274	72.925
PPMI only	80.691	81.702	82.087	82.142	60.366	65.131	67.561	69.104	67.204	70.802	73.108	72.741
+ specificity	80.634	81.836	82.238	82.401	60.342	65.246	67.748	68.987	67.091	70.711	73.526	73.842
+ Single-scale Graph Relation	80.803	81.781	82.511	82.688	60.397	65.309	67.692	69.768	66.982	71.048	73.401	74.912
+ UMH	80.861	81.991	82.452	83.081	60.381	65.287	67.671	70.021	68.081	70.973	74.603	75.556
+ Learned-weight Multi-scale Heat	80.948	82.113	82.944	82.976	60.424	65.273	68.487	69.944	68.734	71.866	74.511	76.538
+ UKM	80.850	81.935	82.835	83.215	—	—	—	—	—	—	—	—
Full MultiBit (FRT teacher)	**81.057**	**82.202**	**83.162**	**83.592**	**60.449**	**65.426**	**68.846**	**70.965**	**69.178**	**72.177**	**75.921**	**77.372**

Table D-2: Query-level AP comparison of Full MultiBit against alternative designs. Percentages are computed within each 1,500-query group.

Alternative Design	Dataset	Queries	Improved (%)	Tied (%)	Degraded (%)	Mean ΔAP (pp)
Single-scale Teacher	MIRFlickr-25K	1,500	59.73	10.13	30.13	0.48
	IAPR TC-12	1,500	53.33	10.33	36.33	0.23
	MS-COCO	1,500	70.13	9.60	20.27	1.18
Uniform Bit Allocation	MIRFlickr-25K	1,500	55.80	13.47	30.73	0.39
	IAPR TC-12	1,500	47.47	11.47	41.07	0.11
	MS-COCO	1,500	63.67	12.20	24.13	0.85
Pairwise Alignment	MIRFlickr-25K	1,500	47.87	17.00	35.13	0.16
	IAPR TC-12	1,500	43.47	15.27	41.27	0.03
	MS-COCO	1,500	53.73	14.87	31.40	0.46

D.2 Query-Level Gain Statistics

For each alternative design, we compare its AP with that of Full MultiBit on the same query and define $\Delta\mathrm{AP} = \mathrm{AP}_{\mathrm{Full}} - \mathrm{AP}_{\mathrm{Alternative}}$. "Improved", "Tied", and "Degraded" denote $\Delta\mathrm{AP} > 0$, $= 0$, and < 0, respectively. Each dataset–alternative pair contains 1,500 query comparisons. Every comparison uses the 64-bit code in the I2T direction, with 500 candidates per query and a fixed subset of 1,500 queries drawn from the 5,000 queries with seed 42.

Full MultiBit obtains a positive mean $\Delta\mathrm{AP}$ in all nine comparisons, ranging from 0.03 to 1.18 pp. The largest query-level gains occur against the single-scale teacher on MS-COCO, where 70.13% of queries improve. Figure 4 gives the corresponding query-level view: it plots the cumulative distribution of $\Delta\mathrm{AP}$ for each dataset and alternative design, so that the shares of Table D-2 are read off at $\Delta\mathrm{AP} = 0$ and the right tail of a curve reflects the magnitude of the positive differences. These statistics complement the aggregate mAP ablation: they show how the advantage is distributed across individual queries, while the progressive table establishes the contribution of each component across code lengths.

E Results for Section 4.4: Relevance Resolution

E.1 Relevance Resolution Evaluation

IDF-Weighted Shared-Label Evaluator. To score every method with the same external reference rather than with its own objective, all methods are evaluated against one shared IDF-weighted shared-label evaluator. Because this evaluator rewards rare labels through the same frequency prior that the relation teacher uses, the comparison is repeated against a second, frequency-free evaluator in Section E.2. Here $\mathcal{R}$ is the retrieval candidate database, $R = |\mathcal{R}|$, and n_c counts its items labeled

with category c.

$$r_{qi} = \sum_{c=1}^{C} w_c y_{qc} y_{ic}, \qquad w_c = \log \frac{R+1}{n_c+1}, \quad R = |\mathcal{R}|. \tag{E.1}$$

The six variants compared in Table E-1 differ only in how the label-side relation is formed. Binary Relation marks a pair as related when its two label sets intersect and unrelated otherwise, so it carries no graded information. Jaccard Relation replaces this indicator by the Jaccard index $|y_q \cap y_i|/|y_q \cup y_i|$, which distinguishes a single shared label from several. Single-scale Graph Relation propagates the relation through the PPMI label graph of Eq. (B.1) at one diffusion scale, so that a pair without a shared label can still receive a nonzero score. UMH (uniform multi-scale heat) averages the per-scale relation scores g_ℓ over the four retained scales $\tau = (2.00, 9.55, 23.89, 48.56)$ with equal weights and adds no parameter. Learned-weight Multi-scale Heat uses the same four scales but trains the mixture weights $\rho_\ell \geq 0$, $\sum_\ell \rho_\ell = 1$. FRT is the fractional relation teacher of Eqs. (2)–(4), in which all four weights follow from the single order parameter α through the Gauss–Laguerre quadrature rather than being trained independently. The four scales of both multi-scale variants are fixed to the $\alpha = 1.2$ quadrature nodes $\tau = (2.00, 9.55, 23.89, 48.56)$ on every dataset, whereas FRT selects its nodes per dataset on validation ($\alpha = 1.20$ on MIRFlickr-25K and $\alpha = 0.90$ on the other datasets, Table A-3). The learned-weight variant differs from the full model only in how the weights are obtained: it replaces the quadrature weights of Eq. (4) with $\rho = \mathrm{softmax}(\texttt{mixture_logits})$, so that the weights are positive and sum to one, and adds a KL term towards the Gauss–Laguerre prior ρ^* that prevents degeneracy to a single scale. It otherwise uses the same label graph, the same heat kernels at the same nodes, the same objective of Eqs. (8)–(9), and the same shared checkpoint rule as the full model. Each mixture component is the heat kernel $K_\ell = e^{-\tau_\ell L}$ at its own node, so the mixture also induces per-scale teachers g_ℓ of the same form as Eq. (5), and every subblock is supervised by its own g_ℓ exactly as in the full model.

Table E-1: NDCG@100 of different relation teachers under the IDF-weighted shared-label evaluator and the label-name semantic evaluator.

Evaluator	Relation Teacher	MIRFlickr-25K	IAPR TC-12	MS-COCO
IDF-weighted	Binary Relation	0.604	0.436	0.570
	Jaccard Relation	0.631	0.462	0.606
	Single-scale Graph Relation	0.664	0.514	0.648
	UMH	0.682	0.528	0.670
	Learned-weight Multi-scale Heat	0.690	0.535	0.680
	FRT	**0.701**	**0.543**	**0.691**
Label-name semantic	Binary Relation	0.597	0.428	0.562
	Jaccard Relation	0.635	0.468	0.611
	Single-scale Graph Relation	0.656	0.503	0.639
	UMH	0.673	0.520	0.659
	Learned-weight Multi-scale Heat	0.681	0.528	0.670
	FRT	**0.690**	**0.536**	**0.680**

Interpretation. FRT has the highest NDCG@100 under both evaluators on all three datasets, so its advantage does not rest on the frequency prior it shares with the IDF-weighted evaluator. The multi-scale variants are ordered the same way under both evaluators, with FRT ahead of the learned-weight mixture by 0.008–0.011 NDCG@100 on every dataset. These measurements support the teacher construction rather than a particular reference. Table E-2 isolates the two choices that Table E-1 leaves coupled. All three variants use the same label graph, the same specificity weights and the same kernel-cosine normalization at the nodes selected for each dataset, so they differ only in the mixing weights and in the order of mixing and normalization. FRT leads UKM by 0.005–0.014 NDCG@100 under the IDF evaluator and by 0.006–0.012 under the semantic one, and UKM leads UMH by 0.005–0.008; both orderings hold on every dataset and under both evaluators. Matching the normalization therefore leaves the fractional weights worth 0.005–0.014 NDCG@100, while changing only the nodes (UMH at the $\alpha = 1.2$ nodes versus the dataset-selected nodes on IAPR TC-12 and MS-COCO) is worth a further 0.005–0.006.

Table E-3 extends the resolution comparison of Table 3 to every compared multi-length method. MultiBit attains the highest NDCG@100 and Kendall's τ_b and the lowest collision and inversion rates in all 21 dataset–method rows at the three code lengths. The two multi-length baselines that are strongest here behave differently across metrics: VH-ABL ranks above CMCL in NDCG@100 in all nine settings, yet its mAP@all is 2.7–3.6 pp lower once averaged over the four code lengths, and up to 6.9 pp lower at 16 bits on MS-COCO (Table 1), because NDCG@100 evaluates the top of the

Table E-2: Relation-teacher quality under matched normalization and scale nodes. All variants use the same label graph, specificity weights and kernel-cosine normalization at the nodes selected for each dataset ($\alpha = 1.20$ on MIRFlickr-25K, $\alpha = 0.90$ elsewhere); they differ only in the mixing weights and in whether the components are mixed before or after normalization. UKM mixes the four scale kernels with equal weights and then normalizes, whereas UMH normalizes each scale teacher first and then averages. Each cell reports NDCG@100 under the IDF-weighted and the label-name semantic evaluator.

Dataset (nodes)	FRT (fractional)	UKM (uniform kernel mixture)	UMH (uniform multi-scale heat)
MIRFlickr-25K ($\alpha = 1.20$)	**0.701 / 0.690**	0.687 / 0.678	0.682 / 0.673
IAPR TC-12 ($\alpha = 0.90$)	**0.543 / 0.536**	0.538 / 0.530	0.533 / 0.525
MS-COCO ($\alpha = 0.90$)	**0.691 / 0.680**	0.684 / 0.674	0.676 / 0.666

Table E-3: Complete relevance-resolution comparison under the IDF-weighted evaluator, covering every compared multi-length method. NDCG@100, tie-corrected Kendall's τ_b, and the collision and inversion rates at 16, 32, and 64 bits; each value is the arithmetic mean of the I2T and T2I directions. Collision, Inversion and Kendall's τ_b are computed in the pool formed with the full model, whereas NDCG@100 uses the full retrieval database.

Dataset	Method	16 bits				32 bits				64 bits			
		N↑	τ_b ↑	C↓	I↓	N↑	τ_b ↑	C↓	I↓	N↑	τ_b ↑	C↓	I↓
MIRFlickr-25K	DECH	0.632	0.503	0.396	0.254	0.650	0.540	0.374	0.237	0.666	0.557	0.356	0.220
	DGHDGH	0.624	0.489	0.405	0.260	0.644	0.527	0.382	0.243	0.659	0.548	0.363	0.228
	CMCL	0.643	0.518	0.387	0.246	0.662	0.554	0.361	0.228	0.677	0.568	0.343	0.212
	RCMH	0.637	0.509	0.392	0.250	0.656	0.555	0.367	0.232	0.671	0.565	0.350	0.216
	VH-ABL	0.651	0.531	0.378	0.241	0.670	0.564	0.352	0.220	0.686	0.579	0.333	0.204
	Naive Prefix	0.616	0.472	0.414	0.266	0.645	0.521	0.386	0.244	0.674	0.560	0.354	0.219
	MultiBit	**0.660**	**0.555**	**0.366**	**0.232**	**0.678**	**0.573**	**0.342**	**0.214**	**0.697**	**0.590**	**0.321**	**0.195**
IAPR TC-12	DECH	0.611	0.424	0.420	0.311	0.634	0.451	0.397	0.291	0.653	0.468	0.375	0.273
	DGHDGH	0.604	0.408	0.429	0.317	0.627	0.447	0.404	0.295	0.646	0.463	0.382	0.279
	CMCL	0.620	0.443	0.408	0.302	0.647	0.460	0.383	0.281	0.663	0.478	0.365	0.263
	RCMH	0.615	0.437	0.414	0.306	0.641	0.457	0.390	0.286	0.665	0.474	0.369	0.267
	VH-ABL	0.629	0.450	0.397	0.295	0.656	0.472	0.373	0.273	0.676	0.489	0.350	0.253
	Naive Prefix	0.595	0.393	0.437	0.324	0.626	0.440	0.410	0.301	0.653	0.468	0.378	0.275
	MultiBit	**0.640**	**0.461**	**0.385**	**0.285**	**0.668**	**0.481**	**0.361**	**0.264**	**0.688**	**0.499**	**0.339**	**0.244**
MS-COCO	DECH	0.672	0.497	0.372	0.263	0.690	0.515	0.349	0.244	0.707	0.531	0.330	0.226
	DGHDGH	0.665	0.492	0.380	0.269	0.683	0.509	0.356	0.249	0.700	0.524	0.336	0.232
	CMCL	0.684	0.508	0.360	0.253	0.704	0.526	0.336	0.233	0.719	0.542	0.316	0.216
	RCMH	0.679	0.503	0.365	0.258	0.697	0.521	0.342	0.238	0.714	0.537	0.322	0.220
	VH-ABL	0.692	0.516	0.351	0.247	0.712	0.537	0.326	0.226	0.729	0.553	0.307	0.209
	Naive Prefix	0.656	0.482	0.389	0.276	0.684	0.510	0.359	0.250	0.716	0.540	0.320	0.219
	MultiBit	**0.700**	**0.528**	**0.340**	**0.237**	**0.721**	**0.547**	**0.316**	**0.216**	**0.739**	**0.564**	**0.297**	**0.199**

ranking whereas mAP integrates over its full length; MultiBit is the only method that is best under both. Naive Prefix, a 128-bit model trained without explicit prefix objectives and truncated at test time, improves most steeply with the code length.

Metrics. For query q, π_q is the Hamming ranking of the retrieval database $\mathcal{R}$, π_q^* is the ideal ranking by descending r_{qi}, and $\mathcal{P}_q \subseteq \mathcal{R}$ is the common candidate pool for pairwise comparison. NDCG@100 assesses top-ranked utility; Kendall's τ_b, Collision, and Inversion assess agreement, ties, and reversed

orders among candidates with distinct graded relevance.

$$\mathrm{NDCG@100}(q) = \frac{\sum_{k=1}^{100} r_{q,\pi_q(k)} / \log_2(k+1)}{\sum_{k=1}^{100} r_{q,\pi_q^*(k)} / \log_2(k+1)}, \tag{E.2}$$

$$\tau_b(q) = \tau_b\big(\{r_{qi}\}_{i\in\mathcal{P}_q}, \{-d_H(q,i)\}_{i\in\mathcal{P}_q}\big), \tag{E.3}$$

$$\mathrm{Collision}(q) = \frac{\#\{(i,j)\in\mathcal{P}_q^2 : r_{qi} > r_{qj},\ d_H(q,i) = d_H(q,j)\}}{\#\{(i,j)\in\mathcal{P}_q^2 : r_{qi} > r_{qj}\}}, \tag{E.4}$$

$$\mathrm{Inversion}(q) = \frac{\#\{(i,j)\in\mathcal{P}_q^2 : r_{qi} > r_{qj},\ d_H(q,i) > d_H(q,j)\}}{\#\{(i,j)\in\mathcal{P}_q^2 : r_{qi} > r_{qj}\}}. \tag{E.5}$$

Evaluation Protocol. Candidate identities are held fixed within each method comparison. NDCG@100 evaluates the first 100 Hamming-ranked candidates from $\mathcal{R}$; the pairwise measures use a common candidate pool $\mathcal{P}_q \subseteq \mathcal{R}$ for each query. Kendall's τ_b corrects for Hamming ties. Query metrics are averaged separately for I2T and T2I, and their arithmetic mean is reported in the resolution tables, where N denotes NDCG@100, τ_b tie-corrected Kendall's τ_b, C the collision rate and I the inversion rate. The common pool $\mathcal{P}_q$ is the union of the top-100 candidates of the method being reported and of the full model, so $|\mathcal{P}_q| \le 200$; the full model is therefore the common anchor of every row of the resolution tables, and its own row uses its own top-100 candidates. A candidate is relevant for mAP when it shares at least one label with the query, and Hamming ties are ordered by database index (stable sort) for every method. Equal-relevance pairs are excluded from Collision and Inversion, and equal-distance pairs count as collisions rather than inversions. Since every query has a relevant counterpart in the retrieval database, no query is dropped: the ideal NDCG denominator is always positive, and τ_b is 0 only when its tie-corrected denominator vanishes.

E.2 Decoupled Semantic Evaluator

The IDF-weighted evaluator above scores a candidate by frequency-weighted shared labels, which is the same prior the relation teacher uses. To test whether the reported ordering depends on that shared prior, we repeat the comparison against a second evaluator constructed only from label names. Each category name is encoded with the frozen CLIP text tower, the mean label embedding is subtracted to suppress the common prompt component, and the resulting label similarity matrix defines a semantic relevance score that contains no frequency term:

$$\begin{aligned} S_{cd} &= \max\big(0,\ \widehat{e}_c^{\mathsf{T}}\widehat{e}_d\big), \qquad \widehat{e}_c = \mathrm{normalize}(E_{\text{text}}(\text{“a photo of a ”} + \mathrm{name}(c)) - \bar{e}), \\ \bar{e} &= \frac{1}{C}\sum_{c'=1}^{C} E_{\text{text}}(\text{“a photo of a ”} + \mathrm{name}(c')). \end{aligned} \tag{E.6}$$

$$r_{qi}^{\text{sem}} = \frac{y_q^{\mathsf{T}} S\, y_i}{\sqrt{\big(y_q^{\mathsf{T}} S\, y_q\big)\,\big(y_i^{\mathsf{T}} S\, y_i\big)}}, \qquad S_{cc} = 1. \tag{E.7}$$

Unlike the IDF-weighted evaluator, r_{qi}^{sem} is generally nonzero for a candidate pair with no shared label whose labels are semantically related, so this evaluator assigns nonzero relevance to such pairs. Table E-4 reports the same four metrics under this evaluator.

Interpretation. MultiBit attains the highest NDCG@100 and Kendall's τ_b and the lowest collision and inversion rates in all 21 dataset–method rows at the three code lengths under this evaluator as well, so the conclusion drawn under the IDF-weighted evaluator is not an artifact of the shared frequency prior.

To probe the long-range relations that the teacher is designed to model, Table E-5 restricts the comparison to candidate pairs that share no label, and to the subset that the teacher's label graph considers most related, and reports NDCG@100 at 64 bits for both. This stratum can only be measured under the semantic evaluator, because the IDF-weighted reference assigns zero relevance to any pair without a shared label. The label-graph similarity of such a pair is the sum of the PPMI weights between its two label sets in the teacher's label graph, and the second stratum keeps the pairs whose value lies in the top quartile of the no-shared-label pairs of that dataset. MultiBit is best on both strata of all three datasets, so its advantage is not confined to pairs with observed label overlap.

Table E-4: Relevance resolution under the label-name semantic evaluator, measured by the same four metrics as the IDF-weighted evaluator. Collision, Inversion and Kendall's τ_b are computed in the pool formed with the full model, whereas NDCG@100 uses the full retrieval database.

Dataset	Method	16 bits				32 bits				64 bits			
		N↑	τ_b ↑	C↓	I↓	N↑	τ_b ↑	C↓	I↓	N↑	τ_b ↑	C↓	I↓
MIRFlickr-25K	DECH	0.629	0.500	0.392	0.257	0.647	0.536	0.369	0.241	0.664	0.551	0.352	0.224
	DGHDGH	0.622	0.487	0.401	0.263	0.640	0.529	0.377	0.246	0.658	0.543	0.359	0.229
	CMCL	0.640	0.516	0.383	0.249	0.659	0.547	0.356	0.232	0.674	0.561	0.339	0.215
	RCMH	0.634	0.507	0.388	0.253	0.653	0.548	0.363	0.235	0.668	0.558	0.346	0.219
	VH-ABL	0.648	0.538	0.374	0.244	0.667	0.557	0.347	0.223	0.683	0.572	0.329	0.207
	Naive Prefix	0.613	0.468	0.410	0.269	0.642	0.528	0.382	0.247	0.671	0.553	0.350	0.222
	MultiBit	**0.657**	**0.548**	**0.361**	**0.235**	**0.676**	**0.566**	**0.337**	**0.217**	**0.694**	**0.582**	**0.319**	**0.199**
IAPR TC-12	DECH	0.608	0.426	0.414	0.316	0.631	0.444	0.391	0.296	0.650	0.462	0.370	0.278
	DGHDGH	0.601	0.406	0.423	0.322	0.624	0.440	0.398	0.300	0.643	0.456	0.377	0.283
	CMCL	0.617	0.436	0.402	0.306	0.644	0.453	0.377	0.285	0.660	0.471	0.359	0.267
	RCMH	0.612	0.430	0.408	0.310	0.638	0.450	0.384	0.290	0.662	0.467	0.363	0.271
	VH-ABL	0.626	0.443	0.391	0.299	0.653	0.465	0.367	0.277	0.673	0.482	0.344	0.257
	Naive Prefix	0.592	0.391	0.431	0.328	0.623	0.433	0.404	0.305	0.650	0.461	0.372	0.279
	MultiBit	**0.638**	**0.454**	**0.380**	**0.289**	**0.666**	**0.474**	**0.355**	**0.269**	**0.687**	**0.491**	**0.333**	**0.249**
MS-COCO	DECH	0.668	0.490	0.367	0.267	0.686	0.508	0.343	0.249	0.703	0.523	0.325	0.231
	DGHDGH	0.661	0.484	0.375	0.273	0.679	0.501	0.350	0.253	0.697	0.517	0.331	0.236
	CMCL	0.680	0.500	0.355	0.257	0.700	0.518	0.331	0.237	0.715	0.534	0.311	0.220
	RCMH	0.675	0.495	0.360	0.262	0.693	0.513	0.337	0.242	0.710	0.529	0.317	0.224
	VH-ABL	0.688	0.508	0.346	0.251	0.708	0.529	0.321	0.230	0.725	0.545	0.302	0.213
	Naive Prefix	0.652	0.474	0.384	0.280	0.680	0.502	0.354	0.254	0.712	0.532	0.315	0.223
	MultiBit	**0.696**	**0.520**	**0.334**	**0.241**	**0.718**	**0.539**	**0.311**	**0.221**	**0.736**	**0.555**	**0.292**	**0.204**

Table E-5: NDCG@100 at 64 bits under the label-name semantic evaluator, on candidate pairs that share no label, and on the subset of those pairs whose label-graph similarity is in the top quartile.

Dataset	Pair stratum	DECH	DGHDGH	MultiBit	Pairs
MIRFlickr-25K	No shared label	0.332	0.326	**0.355**	182,640
	No shared label, graph top 25%	0.421	0.409	**0.447**	45,660
IAPR TC-12	No shared label	0.266	0.259	**0.282**	392,108
	No shared label, graph top 25%	0.324	0.329	**0.362**	98,027
MS-COCO	No shared label	0.311	0.305	**0.333**	1,486,720
	No shared label, graph top 25%	0.389	0.382	**0.418**	371,680

Table E-4 uses the same layout as the main resolution table, so the two references can be read side by side. To quantify how far they diverge, we compute the query-level tie-corrected Kendall τ_b between the two reference vectors, $\tau_b(\{r_{qi}\}_{i\in\mathcal{P}_q}, \{r^{\text{sem}}_{qi}\}_{i\in\mathcal{P}_q})$, where r_{qi} is the IDF-weighted reference defined above, and average it over queries. This is a consistency check between two references and is distinct from the τ_b that compares a reference with negative Hamming distance in the metrics above. We obtain 0.728, 0.588 and 0.655 on MIRFlickr-25K, IAPR TC-12 and MS-COCO, so the two references are correlated but not interchangeable. Both nevertheless rank MultiBit first on all four metrics in every dataset–length setting (Tables 3 and E-4), so the conclusion does not depend on a single reference.

E.3 FRT Hyperparameter Sensitivity

Interpretation. The grid peaks at $(\eta, \alpha) = (0.20, 1.20)$ on MIRFlickr-25K and $(0.20, 0.90)$ on IAPR TC-12, MS-COCO, and NUS-WIDE. Nearby combinations remain competitive, while edge values are generally lower.

Table E-6: Validation Mean mAP (%) obtained with different combinations of the FRT hyperparameters α and η on all datasets under the 64-bit setting.

Dataset	$\eta \backslash \alpha$	0.3	0.6	0.9	1.2	1.5
MIRFlickr-25K	0.05	82.31	82.55	82.73	82.61	82.22
	0.10	82.56	82.89	83.04	82.98	82.64
	0.20	82.80	83.05	83.16	**83.21**	82.87
	0.50	82.62	82.96	83.08	82.91	82.71
	1.00	82.18	82.45	82.70	82.52	82.30
IAPR TC-12	0.05	67.36	67.70	67.95	67.81	67.29
	0.10	67.62	68.06	68.41	68.18	67.72
	0.20	67.89	68.32	**68.85**	68.67	68.10
	0.50	67.71	68.21	68.58	68.34	67.91
	1.00	67.28	67.69	68.06	67.82	67.43
MS-COCO	0.05	74.62	74.94	75.18	74.89	74.40
	0.10	74.91	75.34	75.60	75.28	74.85
	0.20	75.21	75.56	**75.92**	75.79	75.16
	0.50	74.98	75.48	75.76	75.39	74.91
	1.00	74.44	74.88	75.22	75.01	74.61
NUS-WIDE	0.05	72.91	73.22	73.49	73.31	72.88
	0.10	73.15	73.61	73.91	73.72	73.20
	0.20	73.38	73.88	**74.18**	74.06	73.55
	0.50	73.29	73.73	74.04	73.79	73.39
	1.00	72.86	73.25	73.56	73.42	73.04

F Results for Section 4.5: Multi-Length Consistency and Efficiency

F.1 Multi-Length Evaluation

Configurations. Naive Prefix learns a 128-bit model without explicit prefix objectives and truncates it at test time; it is a separate reference from the ablation baseline of Table 2, which uses only the standard cross-modal hashing objective (Appendix Section A.1), and the two serve the deployment comparison and the component ablation, respectively. MultiBit-Sep trains independent models for 16, 32, 64, and 128 bits. MultiBit uses the shared nested representation. Encoders, data splits, selection rules, and retrieval evaluation are held fixed across the three configurations.

Storage and Time. Model storage counts the parameter files needed for all four lengths, and packed-code storage excludes labels and index overhead. For R database items per modality, reusable 128-bit codes require $2R(128)/8$ payload bytes, whereas four separate databases require $2R(16 + 32 + 64 + 128)/8$ bytes. These formulas describe code payloads, not measured index memory. MultiBit-Sep training time sums its four model runs. Query time covers hash-head encoding and Hamming ranking over the full retrieval database and excludes the shared backbone forward pass. Ranking is common to all methods and depends only on the database size, so the reported differences reflect encoding cost.

Ranking Agreement. For the same query and candidate identities, rankings at lengths b and b' are compared using Kendall's τ_b on negative Hamming distances. Its subscript denotes tie correction, not code length. All six unordered length pairs are computed per query; Table F-2 reports the most distant pair (16–128) and the six-pair mean for every method, separately for I2T and T2I. The four-panel figure visualizes an I2T sample only. Specifically, let $a_i = -d_H^{(b)}(q, i)$ and $a'_i = -d_H^{(b')}(q, i)$ be the ranking scores of candidate i under the two code lengths. For all candidate pairs, let C and D denote the numbers of concordant and discordant pairs, and let T_a and $T_{a'}$ denote the numbers of pairs tied only in the first and second score sequences, respectively. The tie-corrected Kendall coefficient is

$$\tau_b(\boldsymbol{a}, \boldsymbol{a}') = \frac{C - D}{\sqrt{(C + D + T_a)(C + D + T_{a'})}}. \tag{F.1}$$

Thus, $\tau_b = 1$ denotes identical pairwise ordering, while smaller values indicate weaker cross-length agreement; tie correction is essential because finite Hamming spaces contain many equal-distance candidates.

$$A^t_{b,b'} = \frac{1}{|\mathcal{Q}_t^{\text{valid}}|} \sum_{q \in \mathcal{Q}_t^{\text{valid}}} \tau_b\Big(-\boldsymbol{d}_q^{(b)}, -\boldsymbol{d}_q^{(b')}\Big), \quad t \in \{\text{I2T}, \text{T2I}\}. \tag{F.2}$$

F.2 Four-Panel Rank-Displacement Visualization

Visualization Protocol. The plotted data contain five I2T queries with 500 candidates per panel for each dataset–method combination, totaling 10,000 query–candidate points across four panels. For each point, the normalized candidate ranks at 16 and 128 bits are $\widetilde{r}^{16}$ and $\widetilde{r}^{128}$. The horizontal axis of Figure 6 shows $\widetilde{r}^{16}$, and the vertical axis shows

$$\Delta r = \widetilde{r}^{128} - \widetilde{r}^{16}. \tag{F.3}$$

Accordingly, $\Delta r = 0$ denotes perfect rank preservation, whereas a larger absolute displacement indicates a stronger change in candidate ordering. The shaded region marks candidates ranked within the top 20% by the 16-bit representation. Scatter points are individual query–candidate observations, and the overlaid contours summarize their empirical density. Kendall's τ_b is computed between the 16-bit and 128-bit rankings with ties properly handled, while mean $|\Delta r|$ measures the average magnitude of candidate-level rank changes, and the annotated values are those of this sample rather than of Table F-2. A distribution concentrated more tightly around zero, together with larger τ_b and smaller mean $|\Delta r|$, indicates stronger cross-length ranking consistency.

Table F-1: Statistics of the sampled panels behind Figure 6: five I2T queries with 500 candidates each per dataset–method pair. These are sample values and differ from the full-evaluation agreement of Table F-2.

Dataset	Method	τ_b	Mean $\lvert\Delta r\rvert$
MIRFlickr-25K	Naive Prefix	0.578 ± 0.017	0.159
MIRFlickr-25K	MultiBit	0.756 ± 0.012	0.103
IAPR TC-12	Naive Prefix	0.566 ± 0.016	0.162
IAPR TC-12	MultiBit	0.709 ± 0.013	0.126

Table F-2: Cross-length ranking agreement measured by tie-corrected Kendall's τ_b, covering every configuration that serves more than one code length: the published multi-length baselines (CMCL, RCMH, VH-ABL), Naive Prefix, MultiBit-Sep, MultiBit without MPA and the full model. Cross-length agreement compares two code lengths of the same model, so the fixed-length baselines of Table 1 are not applicable.

Dataset	**Method**	**τ_b 16–128 (I2T)**	**τ_b 16–128 (T2I)**	**Mean τ_b (I2T)**	**Mean τ_b (T2I)**
MIRFlickr-25K	Naive Prefix	0.596	0.588	0.716	0.708
	MultiBit-Sep	0.575	0.567	0.693	0.685
	MultiBit w/o MPA (FRT+SBO)	0.706	0.698	0.820	0.812
	CMCL	0.633	0.642	0.687	0.689
	RCMH	0.730	0.672	0.729	0.676
	VH-ABL	0.662	0.629	0.754	0.724
	MultiBit	**0.760**	**0.752**	**0.855**	**0.847**
IAPR TC-12	Naive Prefix	0.527	0.515	0.659	0.647
	MultiBit-Sep	0.509	0.497	0.637	0.625
	MultiBit w/o MPA (FRT+SBO)	0.674	0.667	0.801	0.794
	CMCL	0.359	0.340	0.394	0.379
	RCMH	0.590	0.583	0.645	0.636
	VH-ABL	0.520	0.542	0.650	0.645
	MultiBit	**0.698**	**0.686**	**0.817**	**0.805**
MS-COCO	Naive Prefix	0.580	0.568	0.701	0.689
	MultiBit-Sep	0.560	0.546	0.679	0.665
	MultiBit w/o MPA (FRT+SBO)	0.719	0.711	0.833	0.825
	CMCL	0.440	0.437	0.480	0.476
	RCMH	0.514	0.494	0.609	0.595
	VH-ABL	0.547	0.478	0.642	0.608
	MultiBit	**0.730**	**0.718**	**0.839**	**0.827**

The mean averages the six unordered code-length pairs under a single query and candidate pool; for Naive Prefix and MultiBit-Sep it is the arithmetic mean of their per-pair agreement. The subscript of τ_b denotes tie correction, not code length.

Interpretation. MultiBit obtains the highest Kendall's τ_b for every dataset and retrieval direction on both the most distant 16–128 pair and the six-pair mean. The gains are especially clear for the distant pair, where direct truncation and independently trained length-specific models show weaker agreement. Removing MPA (that is, keeping FRT and SBO without prefix alignment) lowers the six-pair mean from 0.855 to 0.820 on MIRFlickr-25K, from 0.817 to 0.801 on IAPR TC-12 and from 0.839 to 0.833 on MS-COCO, and lowers the 16–128 I2T pair from 0.760 to 0.706 on MIRFlickr-25K,

where the resulting τ_b falls below RCMH (0.730). The effect is largest at the widest length gap, which supports the claim that MPA stabilizes a shared nested ranking across bit budgets.