# RELIABILITY-DISENTANGLED DISTILLATION HASHING: SELECTIVE BIT AND SEMANTIC PURIFICATION FOR MULTIMODAL RETRIEVAL

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Unsupervised multimodal hashing is critical for retrieval efficiency, yet suffers from excessive parameters and redundant code bits. Standard distillation methods treat the teacher model as a perfect supervisor, forcing the student model to mimic representations that inherently contain stochastic noise and errors. To break this, we propose a **R**eliability-**D**isentangled **D**istillation **H**ashing (**RDDH**) framework for multimodal retrieval, which energizes the student to selectively inherit reliable knowledge via a dual-granularity purification mechanism. At the micro level (bit-wise), we present an information-theoretic reliability module based on secondary clustering that explicitly disentangles discriminative bits from redundant noise via entropy and polarity analysis, focusing the student on semantically stable directions. At the macro level (semantic-wise), we rectify the geometry of teacher through a filtered bidirectional contrastive objective that acts as a gatekeeper, masking the false-positive alignment and false-negative repulsion to preserve the intrinsic topological structure. Specifically, to balance the supervision sparsity from strict filtering, an intra-student contrastive task is integrated to ensure global data utilization. Extensive experiments on three benchmark datasets demonstrate that **RDDH** effectively purifies the distilled knowledge, realizing a lightweight student to significantly surpass the performance of teacher.

## 1 INTRODUCTION

With the advent of the big data era, information is growing exponentially and becoming diverse in terms of modalities. Efficient multimodal information retrieval methodologies have become a challenging task Ji et al. (2023); Yang et al. (2025b), and the storage mode of data also leads to higher retrieval costs. Fortunately, the presence of hash learning greatly alleviated the above phenomenon Qin et al. (2024); Sun et al. (2024); Huo et al. (2024b); Yang et al. (2025a). Specifically, the external knowledge is encoded through pre-trained models and mapped into common semantic space, then binarization in hash codes via hash networks Huo et al. (2024a); Hao et al. (2024); Pu et al. (2025); Kuai et al. (2025), which enables information to be indexed by a sequence of 0 and 1, significantly improving the retrieval efficiency.

For the past few years, the network structure has become increasingly complex, it is time-consuming and labor-intensive to train a high-performance cross-modal retrieval model. In consequence, knowledge distillation has been introduced to the field of hash learning Hu et al. (2020); Li et al. (2024); Yu et al. (2024), which enables the student model to fit the feature distribution from teacher model, realizing the goal of utilizing a lightweight model to learn the knowledge and approach the performance of complex teacher model. Particularly, clustering is widely used in hash learning Wang et al. (2021; 2025), by which we can categorize and determine the valid hash codes from teacher model. However, a central challenge lies in determining which parts of the teacher representation should be preserved. Besides, hash bits often exhibit redundancy and varying semantic stability, motivating the necessity for a more principled assessment of bit reliability.

To address this, we analyze the hash-bit behavior through semantic consistency rather than simple clusters. By mapping the hash bits of each cluster into probability space and employing information entropy, we identify the bits concentrated at $\pm\mathbf{1}$ as informative, and employ a polarization metric



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.

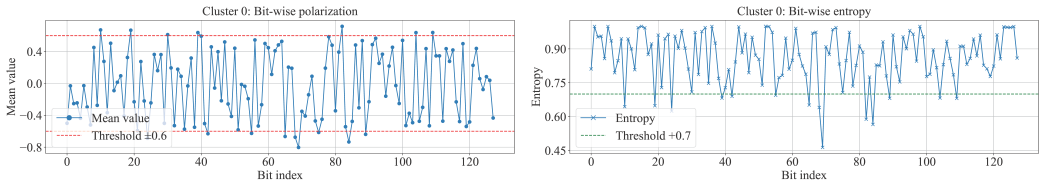


Figure 1: Bit-wise polarization and entropy of hash codes generated by the teacher on MIRFlickr within cluster zeroth.

to discard the bits that hover near **0** and carry no valid class information. As depicted in Fig. 1, DHRL Shu et al. (2025) is employed as the teacher model on MIRFlickr dataset Huiskes & Lew (2008) within code length of 128 bits, where the left is bit-wise mean values, indicating the degree of polarization, *i.e.*, closer to **+1** or **-1** implies higher polarization, and the red dashed lines denote the polarization threshold. Moreover, the right are bit-wise entropy values, where the green dashed line denotes the entropy threshold. The observations motivate the semantic-filtered reliability modeling used in this work.

However, the student model inherently possesses a learning capability limit with the teacher model as the upper bound. The question arises whether to improve the retrieval precision or pursue the retrieval speed of the lightweight student model. Our answer is: **both**. Based on the above goal, the following issues are encountered.

**Question 1**: How to exploit redundant hash bits without discarding the available semantics?

Previous hashing methods Tan et al. (2022); Wu et al. (2024); He et al. (2024) applied a single global clustering in hash codes to identify redundant bits. However, this tends to overlook bits that are discriminative within finer semantic regions. Instead of relying purely on cluster boundaries, we evaluate the bit efficiency through **semantic reliability**, which preserves both coarse discriminative signals and fine-grained distinctions.

**Question 2**: How to prevent the student from inheriting inaccuracy supervision (False Positives)?

Common distillation approaches assume the teacher is infallible, aiming to match distributions via soft labels Yang et al. (2023); Xiao & He (2024) or feature imitation Huang et al. (2022); Hu et al. (2023). However, modality gaps cause the teacher to generate high-confidence pseudo-labels for mismatched pairs (False Positives). To reduce the contamination, we incorporate a **consistency-based filtering mechanism**. By detecting conflicting cluster assignments between modalities, we identify and suppress the erroneous supervision signal, preventing the student from fitting to semantic noise.

**Question 3**: How to overcome the False Negative problem in standard contrastive learning?

Standard contrastive learning van den Oord et al. (2018) suffers from an inherent flaw: sampling bias. It typically treats non-paired samples as negatives. This assumes that any sample other than the input itself is semantically distinct, which is often false in large datasets (False Negatives). This mechanism indiscriminately repels latent semantic neighbors. To address this, we employ **targeted negative masking**, *i.e.*, explicitly identify false repelled neighbors via teacher clustering and exclude from the negative set, ensuring the embedding space preserves the correct semantic topology.

In summary, this paper proposes a semantic-filtered bit-mask distillation method for multimodal hashing, which rectifies both false positive supervision and false negative repulsion, and enables the selective transfer of reliable knowledge, realizing the lightweight student to surpass its teacher.

**Contributions.** The main contributions are summarized as follows:

- **Reliability-Disentangled Bit Purification:** A fine-grained bit purification strategy based on reliability disentanglement is presented. By identifying and weighting the hash dimensions, the student focuses on stable and discriminative bits while purifying the unreliable dimensions that carry high entropy or noise.
- **Semantic-Filtered Bidirectional Distillation:** A dual filtering mechanism is proposed to rectify the teacher supervision. We employ the False Positive Suppression to prevent the student from overfitting to erroneous teacher predictions, while implementing the false

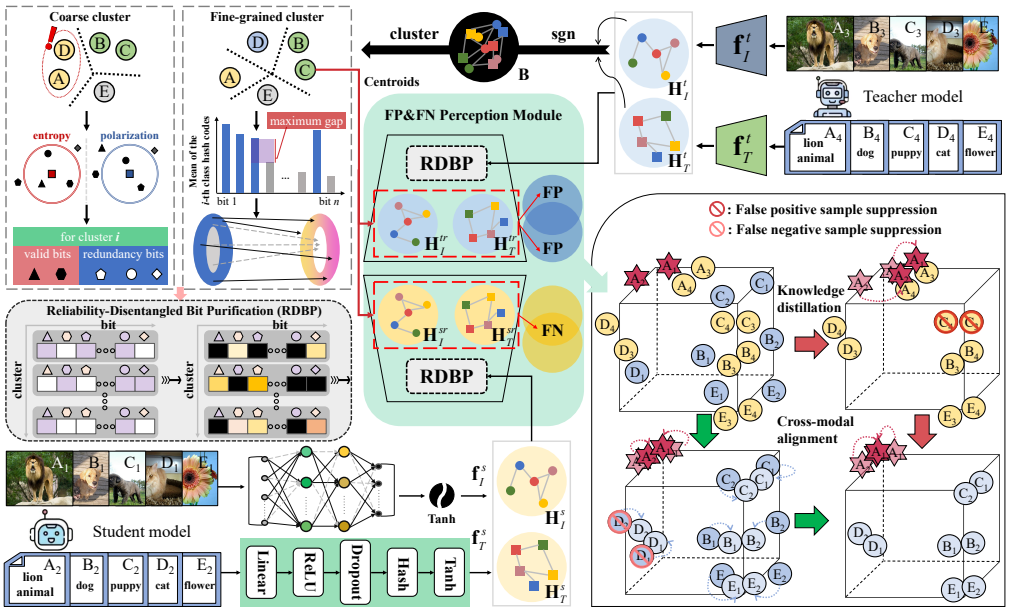


Figure 2: The framework of **R**eliability-**D**isentangled **D**istillation **H**ashing: selective bit and semantic purification for multimodal retrieval (**RDDH**).

negative exclusion to shield latent semantic neighbors from incorrect repulsion, ensuring the student learns a clean and topologically correct semantic structure.

- **Holistic Complementary Optimization:** To balance the sparse supervision caused by strict filtering, we integrate the semantic-filtered distillation with a holistic intra-student contrastive objective. This complementary task maximizes data utilization and enforces manifold consistency, enabling the lightweight student to learn robust representations and surpass the teacher.

## 2 METHOD

### 2.1 NOTATIONS

Consider the dataset $\mathbf{X} = \{\mathbf{X}_1, ..., \mathbf{X}_M\}$ that involves $M$ modalities, where $\mathbf{X}_m = \{\mathbf{x}_1, ..., \mathbf{x}_n\}^\top \in \mathbb{R}^{n \times d_m}$, $n$ is the number of instances, $d_m$ is the dimension of modality $\mathbf{X}_m$. For simplicity, matrices and vectors are denoted by bold uppercase and lowercase letters, respectively, and the transpose of matrix $\mathbf{N}$ is denoted by $\mathbf{N}^\top$. Particularly, this paper mainly deals with image and text modalities, the respective real-valued hash representations are denoted by $\mathbf{H}_I \in \mathbb{R}^{n \times k}$ and $\mathbf{H}_T \in \mathbb{R}^{n \times k}$, where $k$ is the length of hash code, further generate compact binary codes $\mathbf{B} \in \{-1, 1\}^{n \times k}$. To distinguish between the teacher and student models, $\mathbf{H}_I^t$ and $\mathbf{H}_I^s$ are utilized to represent the image hash codes generated via teacher and student models, respectively. Similarly, $\mathbf{H}_T^t$ and $\mathbf{H}_T^s$ can be derived.

### 2.2 **RDDH** FRAMEWORK

**RDDH** is a reliability-disentangled distillation framework that transfers cross-modal knowledge from the teacher hashing model to a lightweight student model, where the student does not blindly mimic the output of teacher but reconstructs a pure embedding geometry by correcting defects. As demonstrated in Fig. 2, the proposed **RDDH** mainly contains three modules, *i.e.*, knowledge distillation from multi-perspective, reliability-disentangled bit purification, and semantic-filtered bidirectional distillation optimization. The fine-grained explanations are as follows.

### 2.3 KNOWLEDGE DISTILLATION FROM MULTI-PERSPECTIVE

The basic module primarily investigates the lightweight distillation model from three perspectives: probabilistic hashing, real-valued hashing, and discrete hashing. Specifically, the student network

consists of an image network and a text network, where the respective original features $\mathbf{X}_I \in \mathbb{R}^{n\times d_I}$ and $\mathbf{X}_T \in \mathbb{R}^{n\times d_T}$ are passed through the networks to generate the student image real-valued hash codes $\mathbf{H}_I^s$ and student text real-valued hash codes $\mathbf{H}_T^s$. Meanwhile, the straight-through estimator method is adopted to binarize the hash codes in $\mathbf{B}_I^s$ and $\mathbf{B}_T^s$ preserving the original gradients, enabling the subsequent loss function based on the discrete hash codes to propagate the gradients. Similarly, the teacher model generates corresponding $\mathbf{H}_I^t, \mathbf{H}_T^t, \mathbf{B}_I^t, \mathbf{B}_T^t$. Besides, $\mathbf{B}$ can be obtained via adding $\mathbf{H}_I^t$ and $\mathbf{H}_T^t$, followed by the $\mathrm{sgn}$ operation. To make the data distributions between modalities of the student approach to each other, the KL divergence loss Kullback & Leibler (1951) is introduced. Thus, the sigmoid function is utilized to convert $\mathbf{H}_I^s$ and $\mathbf{H}_I^t$ into probability values $\mathbf{P}_I \in \mathbb{R}^{n\times k}$ and $\mathbf{P}_T \in \mathbb{R}^{n\times k}$.

If only one modality is allowed to converge towards the other, then it results in a significant change in the probability distribution of the former. However, if we make the image and text modalities actively converge to each other, then the change in respective probability distributions becomes moderate. Therefore, the cross-modal bidirectional divergence loss of the student is defined as,

$$\begin{aligned}&\mathcal{L}_{\text{bkl}} = \tfrac{1}{2}[\mathrm{KL}(\mathbf{p}_I\|\mathbf{p}_T) + \mathrm{KL}(\mathbf{p}_T\|\mathbf{p}_I)],\\ &s.t.\ \mathrm{KL}(\mathbf{p}_I\|\mathbf{p}_T) = \frac{1}{k}\sum_{j=1}^{k}\{\mathbf{p}_{Ij}\,(\log\mathbf{p}_{Ij} - \log\mathbf{p}_{Tj}) + (1-\mathbf{p}_{Ij})[\log(1-\mathbf{p}_{Ij}) - \log(1-\mathbf{p}_{Tj})]\},\end{aligned} \tag{1}$$

where $\mathbf{p}_{Ij}$ represents the value of the $j$-th bit in the hash code of $\mathbf{p}_I$. Meanwhile, contrastive learning on cross-modal hash codes generated by the student is performed, obtaining the student sample image hash codes set $\mathbf{H}_I^s = \{\mathbf{h}_I^s(1), \mathbf{h}_I^s(2), \ldots, \mathbf{h}_I^s(n)\}^\top$ and student sample text hash codes set $\mathbf{H}_T^s = \{\mathbf{h}_T^s(1), \mathbf{h}_T^s(2), \ldots, \mathbf{h}_T^s(n)\}^\top$. We treat diverse modalities of the same sample as positive pairs, while the different samples as negative pairs for contrastive learning, aiming to reduce the semantic gap between modalities and simultaneously push the distance between negative pairs further apart, which can be formulated as,

$$\mu_i = \begin{cases}\mathbf{h}_I^s(i), \\ \mathbf{h}_T^s(i-n),\end{cases} \qquad i^+ = \begin{cases} i+n,\ 1 \le i \le n \\ i-n,\ n < i \le 2n\end{cases}, \tag{2}$$

where $\mu_i$ represents the ordered hash codes of image-text pairs placed in a set, and $i^+$ indicates the index position of corresponding positive sample in the set for cross-modal samples. Furthermore, to shield the interference from the same sample, we define the loss as,

$$\mathcal{L}_{\text{contra}} = \frac{1}{2n}\sum_{i=1}^{2n} -\log\frac{\exp(\phi(\mu_i, \mu_{i^+}))}{\sum_{m=1,m\neq i}^{2n}\exp(\phi(\mu_i,\mu_m))}, \tag{3}$$

where $\phi(\cdot,\cdot)$ denotes the cosine similarity between vectors. To guarantee the student learns available knowledge of the teacher, the mean squared error loss for each modality is calculated, enabling the student to capture the data distribution patterns of the teacher as,

$$\begin{aligned}\mathcal{L}_{\text{con}} &= \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{h}_I^s(i) - \mathbf{h}_I^t(i)\|_2^2 + \|\mathbf{h}_T^s(i) - \mathbf{h}_T^t(i)\|_2^2,\\ \mathcal{L}_{\text{dis}} &= \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{B}_I^s(i) - \mathbf{B}_I^t(i)\|_2^2 + \|\mathbf{B}_T^s(i) - \mathbf{B}_T^t(i)\|_2^2.\end{aligned} \tag{4}$$

The student learns hash codes of the teacher from both real-valued and binary perspectives. We preserve the original gradients, ensuring that the loss among student discrete hash codes is available.

### 2.4 Reliability-Disentangled Bit Purification

In the problem of multi-label clustering, relying solely on the elbow method to categorize samples limits the granularity of semantic representation, typically grouping distinguishable samples (*e.g.*, cats and dogs) into a single fuzzy category. Conversely, division based strictly on label counts becomes loose due to the one-to-many mapping of multi-label data. Therefore, an innovative reliability-disentangled purifying strategy is proposed: constructing a reliability mask matrix **MaskE** via a

coarse-grained clustering number based on the elbow rule, and a semantic disentanglement mask matrix **MaskC** based on the number of labels, respectively. The former is utilized for purifying the global key information bits with high reliability, while the latter is employed for disentangling the detailed differences and local variations among labels.

The cluster centers determined by elbow method correspond to hash bits possess significant reliability and discriminative capacity, and the bits are assigned at the value of **1**, *i.e.*, maximum weight. On the other hand, the clustering based on label count disentangles the semantics of multi-label samples that are similar, and assigns dynamic weights to each hash bit based on importance. Specifically, the elbow method determines the optimal number of clusters by observing the squared errors sum of clustering with the cluster number, and the core idea is to find out the point with the largest rate of decrease in the error and regard it as the optimal number of clusters. The sum of squared errors for the $k$-Means clustering is defined as,

$$\mathrm{SSE}(u) = \sum_{i=1}^{n} \sum_{j=1}^{u} \mathbb{I}(y_i = j) \cdot ||\mathbf{x}_i - \delta_j||^2, \tag{5}$$

where $u$ is the number of clusters, $\delta_j$ is the centroid of the $j$-th cluster, $\mathbb{I}(y_i = j)$ is the indicator function, which is **1** if sample $\mathbf{x}_i$ belongs to cluster $j$, and **0** otherwise. Besides, the dimension-less second-order difference formula for $\mathrm{SSE}(u)$ is defined as,

$$\begin{aligned} \Delta^2 \mathrm{SSE}(u) &= \mathrm{SSE}(u) - 2 \cdot \mathrm{SSE}(u-1) + \mathrm{SSE}(u-2), \\ J^* &= \arg\max_j [\mathrm{SSE}(j) - 2 \cdot \mathrm{SSE}(j-1) + \mathrm{SSE}(j-2)], \end{aligned} \tag{6}$$

where $J^*$ represents the optimal number of clusters.

Thus, the cluster numbers $c_1$ and $c_2$ based on elbow method and the number of labels are obtained, respectively, further the $n$ samples are classified into mapping sets: $L^{(1)}$ and $L^{(2)}$, where the hash codes $\mathbf{B}$ are stored in a couple of sets correspond to the classification methods: $\mathbf{S}_{c_1}^{(1)} = \{\mathbf{B}_i \mid L^{(1)}(\mathbf{B}_i) = c_1\}$ and $\mathbf{S}_{c_2}^{(2)} = \{\mathbf{B}_i \mid L^{(2)}(\mathbf{B}_i) = c_2\}$.

Particularly, for the elbow clustering, $c_1$ clusters are obtained and represented as set $\mathbf{S}_c^{(1)}$. To determine whether each bit of the hash code in the cluster provides reliable information for training, we calculate the centroid of each cluster with absolute value, because for unreliable hash bits, there exist lots of positive and negative values, such that the centroid averages close to **0**. By comparing the value of centroid for each bit, we can judge whether the hash bit at that position is reliable for the cluster. Therefore, the cluster center can be formulated as,

$$\mathbf{A}_{c,j}^{(1)} = \frac{1}{|\mathbf{S}_c^{(1)}|} \sum_{\mathbf{B}_i \in \mathbf{S}_c^{(1)}} \mathbf{B}_{i,j}, \tag{7}$$

where $c \in [0, c_1)$ represents the $c$-th cluster under elbow method. If a certain bit of the hash code belonging to the same cluster tends to **1** or **-1**, then it indicates the bit contains purified reliable information of the class. Further, we map the hash codes into the probability space to capture the valid hash bits via the means of information entropy. However, if a certain bit tends to **0**, then the corresponding information entropy should also decrease. Thus, the polarization value is introduced to eliminate the part of stable unreliable information as,

$$P_{i,j}^{+} = \frac{1 + \mu_{i,j}}{2}, \; P_{i,j}^{-} = 1 - P_{i,j}^{+}, \tag{8}$$

$$\mathbf{H}_{i,j} = -P_{i,j}^{+} \log_2(P_{i,j}^{+}) - P_{i,j}^{-} \log_2(P_{i,j}^{-}). \tag{9}$$

Moreover, the coarse-grained mask can be obtained as,

$$\mathrm{MaskE}_{c,j} = \begin{cases} 1, \; \mathbf{H}_{c,j} < \theta_H \text{ and } \left|\mathbf{A}_{c,j}^{(1)}\right| > \theta_P \\ 0, \; \text{otherwise} \end{cases}, \tag{10}$$

where $\theta_H$ and $\theta_P$ denote the entropy and polarization thresholds, respectively. As an example, let $\theta_H$ = 0.7 and $\theta_P$ = 0.6 in Fig. 1. Besides, the **MaskE** matrix records the reliability of each bit for the class under elbow method.

For the semantic disentanglement based on the number of labels, $\mathbf{S}_c^{(2)}$ and $\mathbf{A}_{c,j}^{(2)}$ can be obtained via the same steps, where $c \in [0, c_2)$. However, $\mathbf{A}_{c,j}^{(2)}$ is operated in absolute value. We assign a relatively low weight to this part to disentangle noise, it is worth noting that we set a high threshold $\alpha$=0.8. When $\mathbf{A}_{c,j}^{(2)} > \alpha$, we consider the bit in the label clustering to be equally important. Thus, we aim to increase the importance of this part as much as possible, and the optimization for $\mathbf{S}_c^{(2)}$ and $\mathbf{A}_{c,j}^{(2)}$ can be formulated as,

$$\mathbf{A}_{c,j}^{(2)} = \mathbf{A}_{c,j}^{(2)} \cdot \mathbf{S}_{c,j}^{(2)}, \quad \mathbf{S}_{c,j}^{(2)} = \begin{cases} 1, & \mathbf{A}_{c,j}^{(2)} > \alpha \\ \mathbf{S}_{c,j}^{(2)}, & \mathbf{A}_{c,j}^{(2)} \leq \alpha \end{cases}. \tag{11}$$

After optimizing $\mathbf{A}_{c,j}^{(2)}$, the hash codes of the $c$-th cluster are sorted with length $k$ based on the size of each bit in the absolute value cluster center. Then, we calculate the difference between adjacent sorted absolute values and find the position with maximum, which is denoted as $z$. Particularly, we consider the hash bits after the $z$-th position as unreliable hash bits,

$$\begin{aligned} &z = \arg\max_{1 \leq j < k} d_{c,j} + 1, \\ &s.t.\ d_{c,j} = \mathbf{A}_{c,\epsilon(j)}^{(2)} - \mathbf{A}_{c,\epsilon(j+1)}^{(2)}, \mathbf{A}_{c,\epsilon(j)}^{(2)} \geq \mathbf{A}_{c,\epsilon(j+1)}^{(2)} \geq \cdots \geq \mathbf{A}_{c,\epsilon(k)}^{(2)}, 1 \leq j < k. \end{aligned} \tag{12}$$

We set hyper-parameters $v_{\max}, v_{\min}$, and map the weights to obtain the clustering mask matrix as,

$$\mathrm{MaskC}_{c,j} = \begin{cases} v_{\min} + v_{\mathrm{dif}} \mathbf{A}_{c,j}^{(2)}, & j \in \{\epsilon(1), \ldots, \epsilon(z)\} \\ 0, & \text{otherwise} \end{cases}, \tag{13}$$

where $v_{\mathrm{dif}} = v_{\max} - v_{\min}$, $v_{\min}$=0.3, $v_{\max}$=0.8. Finally, the clustering mask matrix is employed to supplement the elbow mask matrix, obtaining the ultimate bit weight mask matrix, which records the weight of each bit for the whole sample set as,

$$\mathrm{MaskB}_{i,j} = \begin{cases} 1, & \mathbf{B}_i \in L_c^{(1)} \& \mathrm{MaskE}_{c,j} = 1 \\ \mathrm{MaskC}_{d,j}, & \mathbf{B}_i \in L_c^{(1)} \cap L_d^{(2)} \& \mathrm{MaskE}_{c,j} = 0 \end{cases}. \tag{14}$$

### 2.5 Semantic-Filtered Bidirectional Distillation Optimization

The optimization module, incorporating reliability analysis, is designed for cross-modal and cross-teacher-student distillation. The core idea is to purify the interference from binary codes $\mathbf{B}$, real-valued hash codes $\mathbf{h}_I^t$ and $\mathbf{h}_T^t$ obtained from the teacher. After the mentioned codes are clustered by elbow method, samples categorized as the same class in $\mathbf{B}$ are regarded as potential false-negative samples. If $\mathbf{h}_I^t$ and $\mathbf{h}_T^t$ come from the same sample but belong to different classes, they are regarded as false-positive samples. With the purification mechanism described above, we can both suppress unreliable hash codes in the teacher model and avoid misclassifying the true positive sample pairs as negative samples. Through elbow clustering, the pseudo-labels $\mathbf{y}_I^s, \mathbf{y}_T^s, \mathbf{y}_I^t, \mathbf{y}_T^t$ for $\mathbf{h}_I^s, \mathbf{h}_T^s, \mathbf{h}_I^t, \mathbf{h}_T^t$ can be obtained, respectively. In turn, the purified representations $\mathbf{h}_I^{sr}, \mathbf{h}_T^{sr}, \mathbf{h}_I^{tr}, \mathbf{h}_T^{tr}$ after reliability-disentangled purification can be derived.

Additionally, for image hash codes $\mathbf{h}_I^s$ of the student model, we utilize teacher $\mathbf{h}_I^t(i)$ and $\mathbf{h}_T^{tr}(i)$ as anchor sample and positive sample, respectively, which enables the model to simultaneously learn the data distribution of the teacher and reduce the cross-modal semantic gap. Then, the negative set $N_I(i)$ comprises teacher hash codes $\mathbf{h}_I^{tr}(j)$ and $\mathbf{h}_T^{tr}(j)$, whose image cluster labels $\mathbf{y}_I^t(j)$ are different from $\mathbf{y}_I^t(i)$. Crucially, this constraint explicitly excludes latent semantic neighbors (*i.e.*, potential false negatives) within the same cluster, preventing the optimization from pushing away semantically similar samples. Further, for any sample $j$ in this set, the cluster labels of both $\mathbf{h}_I^{tr}(j)$ and $\mathbf{h}_T^{tr}(j)$ are inconsistent, *i.e.*, false positives. Thus, we construct an indicator $\mathbf{fp}$ to estimate the positive teacher-student pairs, ensuring that the student only aligns with the teacher when cross-modal consensus is achieved, which can be formulated as,

$$\mathrm{fp}_i = \begin{cases} 1, & \mathbf{y}_I^t(i) = \mathbf{y}_T^t(i) \\ 0, & \mathbf{y}_I^t(i) \neq \mathbf{y}_T^t(i) \end{cases}. \tag{15}$$

Moreover, contrastive candidate set $\mathbf{R}(i)$ includes both the positive teacher targets and negative samples from $N_I(i)$ as,

$$\mathbf{R}(i) = \{\mathbf{h}_I^t(i), \mathbf{h}_T^t(i)\} \cup N_I(i). \tag{16}$$

In consequence, the ultimate bidirectional distillation optimization consists of two symmetric branches, and the loss function for image-to-text retrieval branch $\mathcal{L}_{\text{bI}}$ is defined as,

$$\mathcal{L}_{\text{bI}} = \frac{1}{n}\sum_{i=1}^{n} -\log \frac{e^{\text{fp}_i\left(\gamma\cdot\phi\left(\mathbf{h}_I^{srt}(i)\right)+(1-\gamma)\cdot\phi\left(\mathbf{h}_I^{sr}(i),\mathbf{h}_T^{tr}(i)\right)\right)/\tau}}{\sum_{\mathbf{h}\in\mathbf{R}(i)} e^{\phi\left(\mathbf{h}_I^{sr}(i),\mathbf{h}\right)/\tau}}, \tag{17}$$

where $\phi\left(\mathbf{h}_I^{srt}(i)\right) = \phi\left(\mathbf{h}_I^{sr}(i), \mathbf{h}_I^{t}(i)\right)$, $\gamma \in (0,1)$ represents the weighting coefficient, $\tau$ = 0.3 represents the temperature coefficient, and the loss function for text-to-image retrieval branch $\mathcal{L}_{\text{bT}}$ is symmetrical to $\mathcal{L}_{\text{bI}}$, thus the ultimate semantic-filtered bidirectional distillation cross-modal loss $\mathcal{L}_{\text{bmbdc}}$ can be formulated as,

$$\mathcal{L}_{\text{bmbdc}} = \mathcal{L}_{\text{bI}} + \mathcal{L}_{\text{bT}}. \tag{18}$$

Note that the ultimate student not only learns the distribution patterns, but also effectively avoids the unreliable knowledge provided by the teacher. Through learning cross-modal knowledge from both the student and teacher simultaneously, and the proposed reliability-disentangled purifying strategy based on secondary clustering, the lightweight student network realizes available cross-modal learning, surpassing the existing teacher model.

### 2.6 Overall Loss Function

With the aid of distillation losses ($\mathcal{L}_{\text{bkl}}, \mathcal{L}_{\text{con}}, \mathcal{L}_{\text{dis}}$), the student approximates the distribution characteristics of teacher. Meanwhile, the optimization integrates two complementary contrastive objectives: semantic-filtered $\mathcal{L}_{\text{bmbdc}}$ and intra-student $\mathcal{L}_{\text{contra}}$, where $\mathcal{L}_{\text{bmbdc}}$ purifies the transfer by circumventing erroneous teacher knowledge, $\mathcal{L}_{\text{contra}}$ serves as a holistic supplement, compensating for the data sparsity caused by filtering, and ensuring global semantic consistency. The synergy not only moderates the inter-modal difference but also enhances the discriminative ability of the generated hash codes. Integrating Eqs. (1), (3), (4), and (18) into a unified learning framework, the overall loss function can be derived as,

$$\mathcal{L} = \lambda_1\mathcal{L}_{\text{contra}} + \lambda_2\mathcal{L}_{\text{bmbdc}} + \mathcal{L}_{\text{bkl}} + \mathcal{L}_{\text{con}} + \mathcal{L}_{\text{dis}}. \tag{19}$$

Particularly, we focus on optimization through contrast learning, meanwhile, as a universal method, there exist minimal number of hyper-parameters $\lambda_1$ and $\lambda_2$, which remain uniform with the settings of 1. Distillation loss, on the other hand, is necessary but not the priority, thus there is no extra setting for these parts, *i.e.*, default to 1.

## 3 Experiment

In this paper, 5,000/2,000 image-text pairs are randomly selected from MIRFlickr, NUS-WIDE Chua et al. (2009), and IAPR-TC12 Escalante et al. (2010) datasets as the training/query sets for evaluation, respectively. For MIRFlickr dataset, the residual 18,015 image-text pairs are utilized as the retrieval set within 24 categories, where the text modality uses 1386-D BoW feature. For NUS-WIDE dataset, 10 typical tags with the residual 184,577 image-text pairs are employed as the retrieval set, where the text modality uses 1000-D BoW feature. For IAPR-TC12 dataset, the residual 18,000 image-text pairs are utilized as the retrieval set within 255 labels, where the text modality employs 2912-D BoW feature. Particularly, we employ the 4096-D image feature extracted from a 19-layer VGG network Simonyan & Zisserman (2015).

**Baselines.** In the experiment, DJSRH Su et al. (2019), DGCPN Yu et al. (2021), PT-FUCH Li et al. (2023), CIRH Zhu et al. (2023), DHRL, SMSH Kuai et al. (2025), USFTH Yang et al. (2025c) are employed as teacher models. We uniformly utilize a lightweight student model and compare the retrieval performance under training of each teacher model with **RDDH** in retrieving textual data by visual query (I2T) and retrieving visual data by text query (T2I).

**Evaluation Metrics.** To assess the retrieval performance of **RDDH**, mean average precision (**mAP**), precision-recall (**PR**), and top-$k$ precision (**P**@k) curves act as evaluation metrics. The number of retrieved instances is set to 50.

**Implementation Details.** As stated in the loss objective, hyper-parameters are defaulted to 1. Besides, we set the learning rate to $1e^{-4}$, and the total iteration number to 100. For the baselines, we realized their works via open-source codes. Experiments were conducted on a server equipped with an Intel Xeon Silver 4210 Processor @2.20 GHz, NVIDIA GeForce RTX 4070 GPU, 128G RAM.

Table 1: Performance comparison of **RDDH** and baselines on the datasets with various code lengths.

| Task | Method | MIRFlickr | | | | NUS-WIDE | | | | IAPR-TC12 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 16 bits | 32 bits | 64 bits | 128 bits | 16 bits | 32 bits | 64 bits | 128 bits | 16 bits | 32 bits | 64 bits | 128 bits |
| I → T | DJSRH | 0.716 | 0.767 | 0.790 | 0.804 | 0.594 | 0.658 | 0.687 | 0.706 | 0.484 | 0.551 | 0.570 | 0.591 |
| | DJSRH-**RDDH** | **0.827** | **0.865** | **0.891** | **0.906** | **0.705** | **0.773** | **0.802** | **0.831** | **0.550** | **0.597** | **0.648** | **0.666** |
| | DGCPN | 0.877 | 0.915 | 0.925 | 0.937 | 0.791 | 0.818 | 0.839 | 0.854 | 0.607 | 0.675 | 0.691 | 0.705 |
| | DGCPN-**RDDH** | **0.887** | **0.920** | **0.929** | **0.943** | **0.793** | **0.830** | **0.847** | **0.862** | **0.617** | **0.693** | **0.712** | **0.725** |
| | PTFUCH | 0.814 | 0.890 | 0.908 | 0.922 | 0.756 | 0.778 | 0.813 | 0.837 | 0.596 | 0.630 | 0.639 | 0.660 |
| | PTFUCH-**RDDH** | **0.848** | **0.907** | **0.916** | **0.936** | **0.768** | **0.797** | **0.829** | **0.845** | **0.626** | **0.656** | **0.669** | **0.696** |
| | CIRH | 0.884 | 0.902 | 0.919 | 0.933 | 0.781 | 0.816 | 0.838 | 0.848 | 0.623 | 0.669 | 0.688 | 0.702 |
| | CIRH-**RDDH** | **0.888** | **0.912** | **0.931** | **0.938** | **0.790** | **0.822** | **0.846** | **0.861** | **0.647** | **0.685** | **0.707** | **0.726** |
| | DHRL | 0.841 | 0.889 | 0.904 | 0.918 | 0.756 | 0.812 | 0.822 | 0.833 | 0.376 | 0.581 | 0.687 | 0.701 |
| | DHRL-**RDDH** | **0.885** | **0.920** | **0.936** | **0.948** | **0.809** | **0.850** | **0.867** | **0.882** | **0.446** | **0.652** | **0.745** | **0.766** |
| | SMSH | 0.904 | 0.910 | 0.928 | 0.933 | 0.785 | 0.816 | 0.836 | 0.848 | 0.630 | 0.671 | 0.692 | 0.699 |
| | SMSH-**RDDH** | **0.905** | **0.912** | **0.932** | **0.939** | **0.789** | **0.823** | **0.852** | **0.862** | **0.634** | **0.675** | **0.699** | **0.706** |
| | USFTH | 0.898 | 0.905 | 0.925 | 0.935 | 0.807 | 0.827 | 0.840 | 0.856 | 0.629 | 0.663 | 0.689 | 0.709 |
| | USFTH-**RDDH** | **0.908** | **0.919** | **0.945** | **0.954** | **0.833** | **0.853** | **0.866** | **0.877** | **0.665** | **0.679** | **0.723** | **0.730** |
| T → I | DJSRH | 0.700 | 0.763 | 0.768 | 0.769 | 0.565 | 0.619 | 0.652 | 0.664 | 0.481 | 0.556 | 0.587 | 0.603 |
| | DJSRH-**RDDH** | **0.777** | **0.846** | **0.853** | **0.875** | **0.643** | **0.727** | **0.759** | **0.781** | **0.548** | **0.606** | **0.653** | **0.678** |
| | DGCPN | 0.860 | 0.881 | 0.888 | 0.894 | 0.752 | 0.773 | 0.779 | 0.797 | 0.601 | 0.686 | 0.701 | 0.714 |
| | DGCPN-**RDDH** | **0.862** | **0.885** | **0.897** | **0.901** | **0.756** | **0.792** | **0.797** | **0.816** | **0.619** | **0.701** | **0.726** | **0.740** |
| | PTFUCH | 0.758 | 0.822 | 0.821 | 0.849 | 0.710 | 0.748 | 0.743 | 0.771 | 0.589 | 0.621 | 0.625 | 0.651 |
| | PTFUCH-**RDDH** | **0.772** | **0.837** | **0.842** | **0.867** | **0.724** | **0.760** | **0.765** | **0.793** | **0.614** | **0.653** | **0.668** | **0.694** |
| | CIRH | 0.861 | 0.881 | 0.895 | 0.897 | 0.775 | 0.792 | 0.806 | 0.817 | 0.648 | 0.699 | 0.714 | 0.727 |
| | CIRH-**RDDH** | **0.865** | **0.882** | **0.897** | **0.902** | **0.776** | **0.795** | **0.811** | **0.821** | **0.653** | **0.707** | **0.730** | **0.746** |
| | DHRL | 0.824 | 0.858 | 0.864 | 0.865 | 0.763 | 0.764 | 0.767 | 0.768 | 0.413 | 0.572 | 0.697 | 0.726 |
| | DHRL-**RDDH** | **0.858** | **0.876** | **0.881** | **0.904** | **0.778** | **0.790** | **0.802** | **0.813** | **0.467** | **0.661** | **0.755** | **0.795** |
| | SMSH | 0.864 | 0.875 | 0.885 | 0.889 | 0.763 | 0.788 | 0.802 | 0.810 | 0.637 | 0.681 | 0.705 | 0.705 |
| | SMSH-**RDDH** | **0.866** | **0.876** | **0.890** | **0.895** | **0.768** | **0.789** | **0.804** | **0.812** | **0.640** | **0.689** | **0.707** | **0.715** |
| | USFTH | 0.855 | 0.871 | 0.884 | 0.891 | 0.767 | 0.781 | 0.796 | 0.801 | 0.629 | 0.674 | 0.713 | 0.723 |
| | USFTH-**RDDH** | **0.868** | **0.883** | **0.890** | **0.903** | **0.784** | **0.800** | **0.809** | **0.815** | **0.650** | **0.702** | **0.723** | **0.737** |

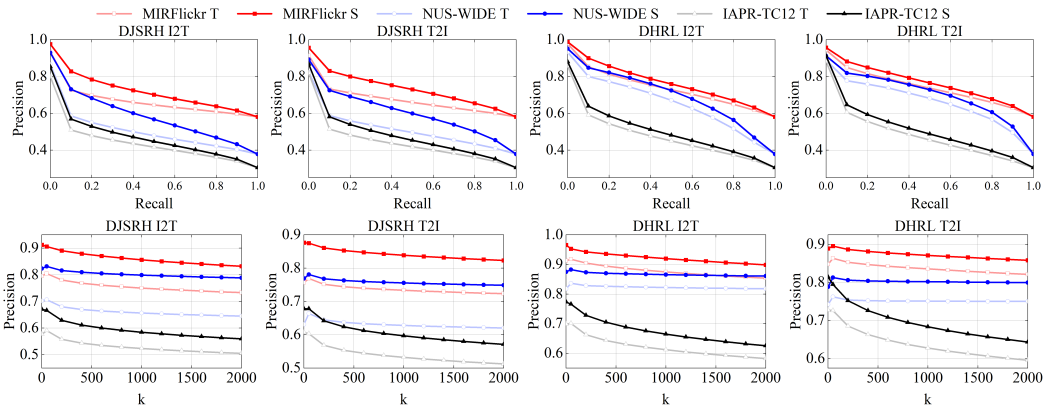


Figure 3: PR and top-$k$ curves of DJSRH and DHRL with teacher-**RDDH** on three datasets within code length of 128 bits.

### 3.1 Results

Table 1 presents the mAP scores for each method on the three datasets, utilizing hash code lengths from 16 to 128 bits. The PR and top-$k$ precision curves are illustrated in Fig. 3. The analysis of results yields the following key insights:

- **RDDH** surpassed all the baselines. Specifically, on the I2T task, the average mAP scores of **RDDH** exceed the classic baseline DJSRH by 13.4%, 17.7%, and 12.1% on MIRFlickr, NUS-WIDE, and IAPR-TC12 datasets, respectively. Similarly, on the T2I task, **RDDH** exhibits improvements of at least 11.7%, 16.3%, and 11.7% on the three datasets, respectively.

Table 2: The mAP results of teacher-**RDDH** and variants on MIRFlickr and IAPR-TC12 datasets.

| Method | Task | MIRFlickr | | IAPR-TC12 | | Task | MIRFlickr | | IAPR-TC12 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 64 bits | 128 bits | 64 bits | 128 bits | | 64 bits | 128 bits | 64 bits | 128 bits |
| DJSRH-**RDDH**-$\mathcal{R}$ | I → T | 0.881 | 0.903 | 0.643 | 0.660 | T → I | 0.848 | 0.854 | 0.594 | 0.667 |
| DJSRH-**RDDH**-$\mathcal{C}$ | | 0.821 | 0.836 | 0.572 | 0.581 | | 0.768 | 0.788 | 0.587 | 0.591 |
| DJSRH-**RDDH**-$\mathcal{B}$ | | 0.883 | 0.897 | 0.641 | 0.657 | | 0.842 | 0.851 | 0.596 | 0.665 |
| DJSRH-**RDDH** | | **0.891** | **0.906** | **0.648** | **0.666** | | **0.853** | **0.875** | **0.653** | **0.678** |
| DHRL-**RDDH**-$\mathcal{R}$ | | 0.929 | 0.933 | 0.724 | 0.730 | | 0.871 | 0.892 | 0.743 | 0.789 |
| DHRL-**RDDH**-$\mathcal{C}$ | | 0.933 | 0.942 | 0.656 | 0.680 | | 0.877 | 0.894 | 0.672 | 0.720 |
| DHRL-**RDDH**-$\mathcal{B}$ | | 0.928 | 0.932 | 0.733 | 0.752 | | 0.857 | 0.868 | 0.740 | 0.784 |
| DHRL-**RDDH** | | **0.936** | **0.948** | **0.745** | **0.766** | | **0.881** | **0.904** | **0.755** | **0.795** |

- Although CIRH and DHRL are efficient hashing models, **RDDH** achieves higher mAP scores than corresponding teachers, indicating that the student benefits from correcting the unreliable semantic relations of teacher. Particularly, the proposed semantic-filtered reliability modeling and bidirectional contrastive optimization jointly enhance the ability of student to capture stable cross-modal structure while discarding erroneous teacher associations.
- The hierarchical distribution of hash code quality exhibits a significant positive correlation with the improvement in student performance. Fig. 1 and Fig. 5 further confirm that the determinism (low entropy) and discriminative ability (high polarization) of the teacher network are core factors determining the efficacy of knowledge distillation.

### 3.2 Ablation Experiments

In this section, we conduct ablation experiments to verify the efficacy of semantic-filtered bit masking and bidirectional contrastive optimization in **RDDH**. Specifically, **RDDH**-$\mathcal{R}$ denotes the variant without semantic reliability masking, where both the teacher and student directly use the original hash codes without purifying via Eq. (14). **RDDH**-$\mathcal{C}$ refers to the variant without intra-student contrastive learning, *i.e.*, removes Eq. (3), thereby missing false positives without positive supervision and significantly reducing effective training signals. **RDDH**-$\mathcal{B}$ eliminates the bidirectional contrastive refinement, *i.e.*, removes Eq. (18).

The ablation results under teachers DJSRH and DHRL are depicted in Table 2, it can be found that both the semantic-filtered bit weighting and bidirectional distillation strategy are essential for strengthening the representation learning of student. Without masking, the student inherits unstable teacher bits. Similarly, without bidirectional refinement, the cross-modal relation cannot be corrected. Together, the key components enable the student to form a structured semantic space that is more compact and discriminative.

### 3.3 Parameters Sensitivity Analysis

To investigate the impact of key hyper-parameters on three datasets, we take DHRL as the teacher and conduct the grid search for $\lambda_1$ and $\lambda_2$ preserving the remaining parameters constant, with the hash code length of 128 bits. Obviously, the parameters exhibit insensitivity to cross-modal tasks across a wide value spectrum, highlighting the generalization of **RDDH**.

## 4 Conclusion

In this paper, we propose a reliability-disentangled distillation framework that moves beyond blind mimicry. Through the bit disentanglement and intra-student contrastive learning in micro level, **RDDH** achieves superior results where every student model surpasses the performance of corresponding teacher. Particularly, the reliability-disentangled bit purification module utilizes polarization and entropy to estimate reliability, offering a promising metric for quantifying hash code quality. This work redefines distillation as a potent paradigm for semantic rectification in multimodal retrieval. There are several limitations that the reliance on the Elbow method may increase computational costs on massive datasets, and the purification thresholds are heuristically set. Future efforts will focus on automated parameter optimization strategies and extending the framework to complex modalities.

### AI USE STATEMENT

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

## REFERENCES

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. NUS-WIDE: a real-world web image database from national university of singapore. In *Proceedings of the 8th ACM International Conference on Image and Video Retrieval, CIVR 2009*. ACM, 2009.

Hugo Jair Escalante, Carlos A. Hernández, Jesús A. González, Aurelio López-López, Manuel Montes-y-Gómez, Eduardo F. Morales, Luis Enrique Sucar, Luis Villaseñor Pineda, and Michael Grubinger. The segmented and annotated IAPR TC-12 benchmark. *Comput. Vis. Image Underst.*, 114(4): 419–428, 2010.

Zhifeng Hao, Yaochu Jin, Xueming Yan, Chuyue Wang, Shangshang Yang, and Hong Ge. Cross-modal hashing retrieval with compatible triplet representation. *Neurocomputing*, 602:128293, 2024.

Liyang He, Zhenya Huang, Jiayu Liu, Enhong Chen, Fei Wang, Jing Sha, and Shijin Wang. Bit-mask robust contrastive knowledge distillation for unsupervised semantic hashing. In *Proceedings of the ACM on Web Conference 2024, WWW 2024*, pp. 1395–1406. ACM, 2024.

Hengtong Hu, Lingxi Xie, Richang Hong, and Qi Tian. Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, pp. 3120–3129. Computer Vision Foundation / IEEE, 2020.

Tie Hu, Mingbao Lin, Lizhou You, Fei Chao, and Rongrong Ji. Discriminator-cooperated feature map distillation for GAN compression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023*, pp. 20351–20360. IEEE, 2023.

Zhenhua Huang, Shunzhi Yang, MengChu Zhou, Zhetao Li, Zheng Gong, and Yunwen Chen. Feature map distillation of thin nets for low-resolution object recognition. *IEEE Trans. Image Process.*, 31: 1364–1379, 2022.

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

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 Trans. Multim. Comput. Commun. Appl.*, 20(6):169:1–169:23, 2024a.

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 Trans. Knowl. Data Eng.*, 36(11): 5926–5939, 2024b.

Wei Ji, Yinwei Wei, Zhedong Zheng, Hao Fei, and Tat-Seng Chua. Deep multimodal learning for information retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023*, pp. 9739–9741. ACM, 2023.

Mingjin Kuai, Jun Long, and Zhan Yang. Statistical model-driven similarity hashing: Bridging modalities for efficient unsupervised retrieval. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, AAAI 2025*, pp. 11977–11986. AAAI Press, 2025.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86, 1951.

Jiaxing Li, Wai Keung Wong, Lin Jiang, Xiaozhao Fang, Shengli Xie, and Yong Xu. CKDH: clip-based knowledge distillation hashing for cross-modal retrieval. *IEEE Trans. Circuits Syst. Video Technol.*, 34(7):6530–6541, 2024.

Jingzhi Li, Fengling Li, Lei Zhu, Hui Cui, and Jingjing Li. Prototype-guided knowledge transfer for federated unsupervised cross-modal hashing. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023*, pp. 1013–1022. ACM, 2023.

Ruitao Pu, Yuan Sun, Yang Qin, Zhenwen Ren, Xiaomin Song, Huiming Zheng, and Dezhong Peng. Robust self-paced hashing for cross-modal retrieval with noisy labels. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, AAAI 2025*, pp. 19969–19977. AAAI Press, 2025.

Qibing Qin, Yadong Huo, Lei Huang, Jiangyan Dai, Huihui Zhang, and Wenfeng Zhang. Deep neighborhood-preserving hashing with quadratic spherical mutual information for cross-modal retrieval. *IEEE Trans. Multim.*, 26:6361–6374, 2024.

Zhenqiu Shu, Yibing Bai, Kailing Yong, and Zhengtao Yu. Deep cross-modal hashing with ranking learning for noisy labels. *IEEE Trans. Big Data*, 11(2):553–565, 2025.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

Shupeng Su, Zhisheng Zhong, and Chao Zhang. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019*, pp. 3027–3035. IEEE, 2019.

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

Wentao Tan, Lei Zhu, Weili Guan, Jingjing Li, and Zhiyong Cheng. Bit-aware semantic transformer hashing for multi-modal retrieval. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 982–991. ACM, 2022.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

Lu Wang, Jie Yang, Masoumeh Zareapoor, and Zhonglong Zheng. Cluster-wise unsupervised hashing for cross-modal similarity search. *Pattern Recognit.*, 111:107732, 2021.

Zhongwen Wang, Xingfeng Li, Yinghui Sun, Quansen Sun, Yuan Sun, Han Ling, Jian Dai, and Zhenwen Ren. TPCH: tensor-interacted projection and cooperative hashing for multi-view clustering. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, AAAI 2025*, pp. 21420–21428. AAAI Press, 2025.

Qingpeng Wu, Zheng Zhang, Yishu Liu, Jingyi Zhang, and Liqiang Nie. Contrastive multi-bit collaborative learning for deep cross-modal hashing. *IEEE Trans. Knowl. Data Eng.*, 36(11): 5835–5848, 2024.

Lingao Xiao and Yang He. Are large-scale soft labels necessary for large-scale dataset distillation? In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*, 2024.

Fan Yang, Xinqi Liu, Fumin Ma, Xiaojian Ding, and Kaixiang Wang. Online asymmetric supervised discrete cross-modal hashing for streaming multimedia data. *Pattern Recognit.*, 165:111604, 2025a.

Wei Yang, Jingjing Fu, Rui Wang, Jinyu Wang, Lei Song, and Jiang Bian. OMGM: orchestrate multiple granularities and modalities for efficient multimodal retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, ACL 2025*, pp. 24545–24563. Association for Computational Linguistics, 2025b.

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, MM 2025*, pp. 5932–5941. ACM, 2025c.

Zhendong Yang, Ailing Zeng, Zhe Li, Tianke Zhang, Chun Yuan, and Yu Li. From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023*, pp. 17139–17148. IEEE, 2023.

Jun Yu, Hao Zhou, Yibing Zhan, and Dacheng Tao. Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, pp. 4626–4634. AAAI Press, 2021.

Yang Yu, Meiyu Liang, Mengran Yin, Kangkang Lu, Junping Du, and Zhe Xue. Unsupervised multimodal graph contrastive semantic anchor space dynamic knowledge distillation network for cross-media hash retrieval. In *40th IEEE International Conference on Data Engineering, ICDE 2024*, pp. 4699–4708. IEEE, 2024.

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

# A Appendix: Theoretical Analysis

We provide the theoretical justification for the core component of **RDDH**: reliability-disentangled bit purification.

The core premise is that high-quality hash bits should exhibit high determinism (assessed by polarization) and high consistency within the semantic clusters discovered by the elbow method (assessed by entropy). We formally prove that the selection and down-weighting strategy maximizes the **mutual information** between the generated hash codes $\mathbf{B}$ and the underlying cluster structure $\mathbf{Y}_{elbow}$.

**Definition A.1.** ***(Elbow-Induced Semantic Clusters)*** *The elbow method partitions the training data into $c_1$ disjoint clusters $L^{(1)} = \{1, 2, \ldots, c_1\}$, which serve as the pseudo-semantic labels $\mathbf{Y}_{elbow}$ for the unsupervised data. Ideally, samples within the same cluster share the same latent semantics.*

**Definition A.2.** ***(Hash Bit Quality Assessment)*** *We evaluate the quality of a hash bit $b$ using two metrics:*

- ***Determinism Assessment:*** *We use the polarization value to assess the capacity of each bit in a certain cluster. A high polarization value indicates that the bit tends towards the binary boundaries $\{-1, +1\}$ rather than hovering near **0**, suggesting that it has the potential to carry information.*
- ***Consistency Assessment:*** *We use the intra-cluster entropy to assess stability. A low intra-cluster entropy indicates that the bit remains invariant for samples within the same cluster.*

**Proposition A.3.** *Selecting bits with high polarization and low intra-cluster entropy, while down-weighting bits with low polarization or high intra-cluster entropy, is equivalent to maximizing the lower bound of the mutual information $I(\mathbf{B}; \mathbf{Y}_{elbow})$ between the hash codes and the cluster assignments.*

*Proof.* The objective of hash learning is to maximize the relevance between the code and the semantic structure, defined as the mutual information,

$$I(\mathbf{B}; \mathbf{Y}_{elbow}) = H(\mathbf{B}) - H(\mathbf{B} \mid \mathbf{Y}_{elbow}), \tag{20}$$

where $H(\mathbf{B})$ is the marginal entropy of the bit distribution (representing channel capacity), and $H(\mathbf{B} \mid \mathbf{Y}_{elbow})$ is the conditional entropy within the given cluster assignment (representing uncertainty).

**Step 1: Polarization assesses the potential for maximizing marginal entropy $H(\mathbf{B})$.**
The polarization metric measures whether a bit effectively utilizes the binary space. In information theory, bits concentrated near **0** (low polarization) have minimal variance, and the corresponding marginal entropy $H(\mathbf{B})$ approaches **0**, representing a lack of discriminative capacity. By selecting bits with high polarization, we ensure that each bit possesses sufficient variance to support a large marginal entropy $H(\mathbf{B})$. This operation maximizes the total channel capacity of the hash codes.

**Step 2: Entropy-based down-weighting minimizes $H(\mathbf{B} \mid \mathbf{Y}_{elbow})$.**
The reliability-disentangled bit purification module explicitly calculates the bit-wise entropy within each cluster determined by the elbow method, which mathematically corresponds to the conditional entropy of an individual bit $b$ as,

$$\mathcal{L}^{(k)}_{entropy} = \mathbb{E}_{c \sim \mathbf{Y}_{elbow}}[H(b_k \mid c)] \equiv H(b_k \mid \mathbf{Y}_{elbow}). \tag{21}$$

Bits with high intra-cluster entropy exhibit disorder within the same semantic category, implying a large $H(b \mid \mathbf{Y}_{elbow})$. By down-weighting the unreliable bits via the masking mechanism, we suppress their influence on the final representation. Assuming the bits are conditionally independent, the code-wise conditional entropy is the summation of individual bit-wise entropies: $H(\mathbf{B} \mid \mathbf{Y}_{elbow}) = \sum_1^k H(b_k \mid \mathbf{Y}_{elbow})$. Thus, the proposed bit purification operation effectively minimizes the collective conditional entropy $H(\mathbf{B} \mid \mathbf{Y}_{elbow})$ in the global optimization objective.

**Combining** Step 1 and Step 2, the optimization strategy (filtering and processing bits under the constraints of polarization and entropy) is equivalent to,

$$\max\left[H(\mathbf{B}) - H(\mathbf{B} \mid \mathbf{Y}_{elbow})\right] = \max I(\mathbf{B}; \mathbf{Y}_{elbow}), \tag{22}$$

which proves that retaining bits within the reliable intersection guarantees the preservation of maximum semantic information inherent in the data structure. □

## B APPENDIX: ADDITIONAL EXPERIMENTS

The t-distributed stochastic neighbor embedding (t-SNE) is utilized to visualize the hashing embeddings of teachers DJSRH, DGCPN, PT-FUCH, CIRH, DHRL, SMSH, USFTH with teacher-**RDDH** on MIRFlickr dataset in Fig. 4. Compared with the teacher, obvious cluster boundaries and compact intra-class groupings are achieved after performing the proposed semantic-filtered bit-mask distillation strategy, which indicates that **RDDH** yields a more discriminative and semantically coherent embedding space than the teacher.

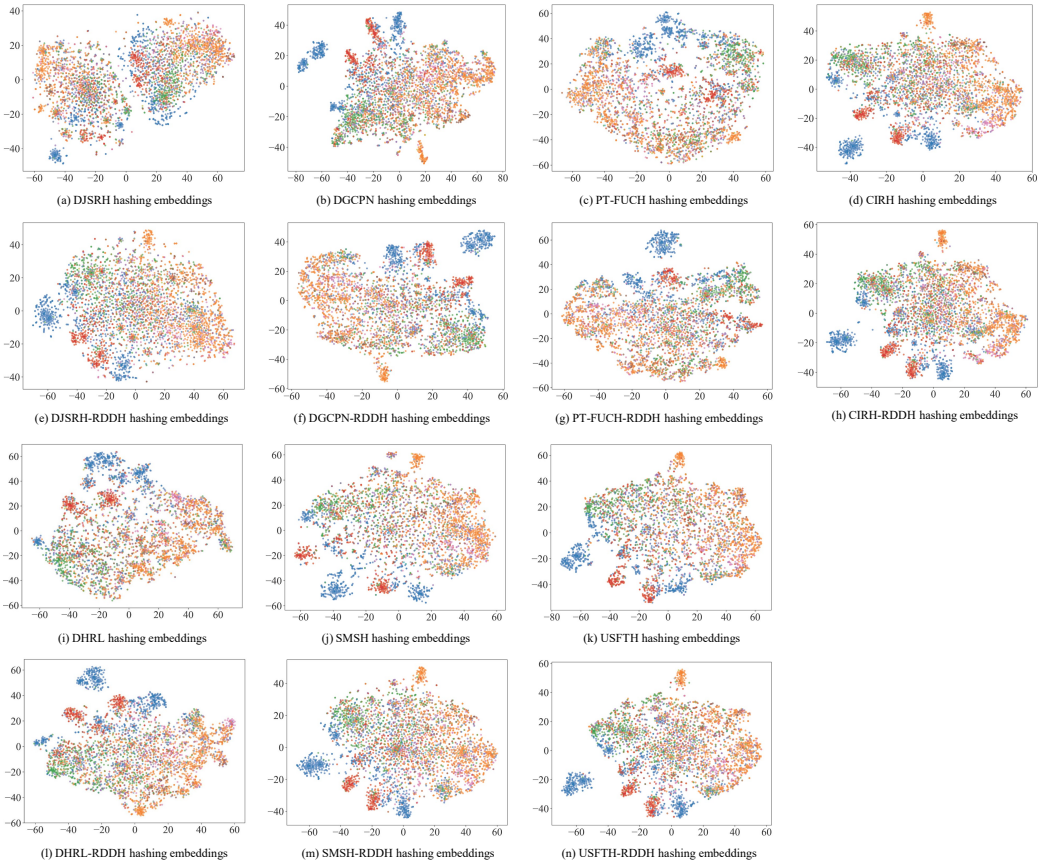


Figure 4: t-SNE visualizations of DJSRH, DGCPN, PT-FUCH, CIRH, DHRL, SMSH, USFTH with teacher-**RDDH** hashing embeddings on MIRFlickr dataset within code length of 128 bits.

We provide the additional bit-wise polarization and entropy of hash codes, PR curves, top-$k$ precision curves, parameters sensitivity analysis, ablation results. The details are as follows:

- Fig. 5: The bit-wise polarization and entropy of hash codes include teachers DJSRH, DGCPN, PT-FUCH, CIRH, SMSH, USFTH on MIRFlickr dataset presented in Section 1.
- Fig. 6: The PR curves include teachers DGCPN, PT-FUCH, CIRH, SMSH, USFTH with teacher-**RDDH** on MIRFlickr, NUS-WIDE, IAPR-TC12 datasets presented in Section 3.1.
- Fig. 7: The top-$k$ precision curves include teachers DGCPN, PT-FUCH, CIRH, SMSH, USFTH with teacher-**RDDH** on MIRFlickr, NUS-WIDE, IAPR-TC12 datasets presented in Section 3.1.
- Fig. 8: The parameters sensitivity analysis of $\lambda_1$ and $\lambda_2$ within DHRL-**RDDH** on MIRFlickr, NUS-WIDE, IAPR-TC12 datasets presented in Section 3.3.
- Table. 3: The ablation results include teachers DGCPN, PT-FUCH, CIRH, SMSH, USFTH with teacher-**RDDH** on MIRFlickr and IAPR-TC12 datasets presented in Section 3.2.

Through visual analysis of polarization and entropy in the reliability-disentangled bit purification module, *i.e.*, Fig. 1 and Fig. 5, significant performance variations can be observed: DHRL and DJSRH exhibit optimal characteristics, featuring numerous effective bits with polarization values strongly tending toward **1** or **-1** and extremely low entropy values; SMSH and USFTH excel in polarization while maintaining relatively good entropy performance. In contrast, CIRH and PT-FUCH exhibit chaotic entropy distributions; DGCPN possesses the worst performance, with polarization values heavily clustered near **0** accompanied by extremely high entropy, which corresponds to Section 3.1. Besides, the PR and top-$k$ precision curves exhibit a parallel trend to the mAP scores.

To sum up, similar trends are observed, the student models consistently exhibit superior structured representations with enhanced semantic compaction, which further demonstrates the generalizability of **RDDH** in terms of diverse teacher architectures.

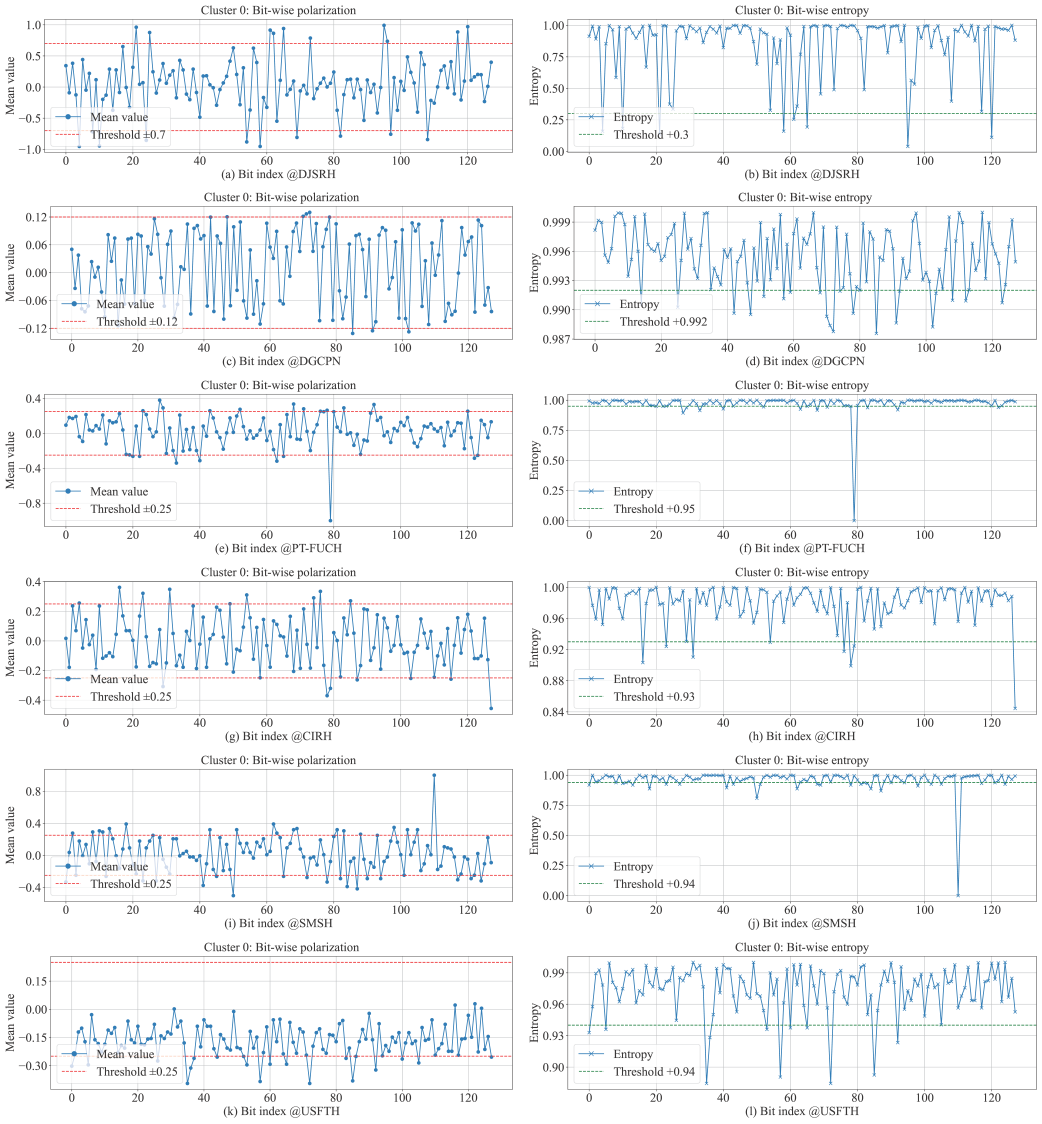


Figure 5: Bit-wise polarization and entropy of hash codes generated through teachers DJSRH, DGCPN, PT-FUCH, CIRH, SMSH, USFTH on MIRFlickr dataset within cluster zeroth at code length of 128 bits.

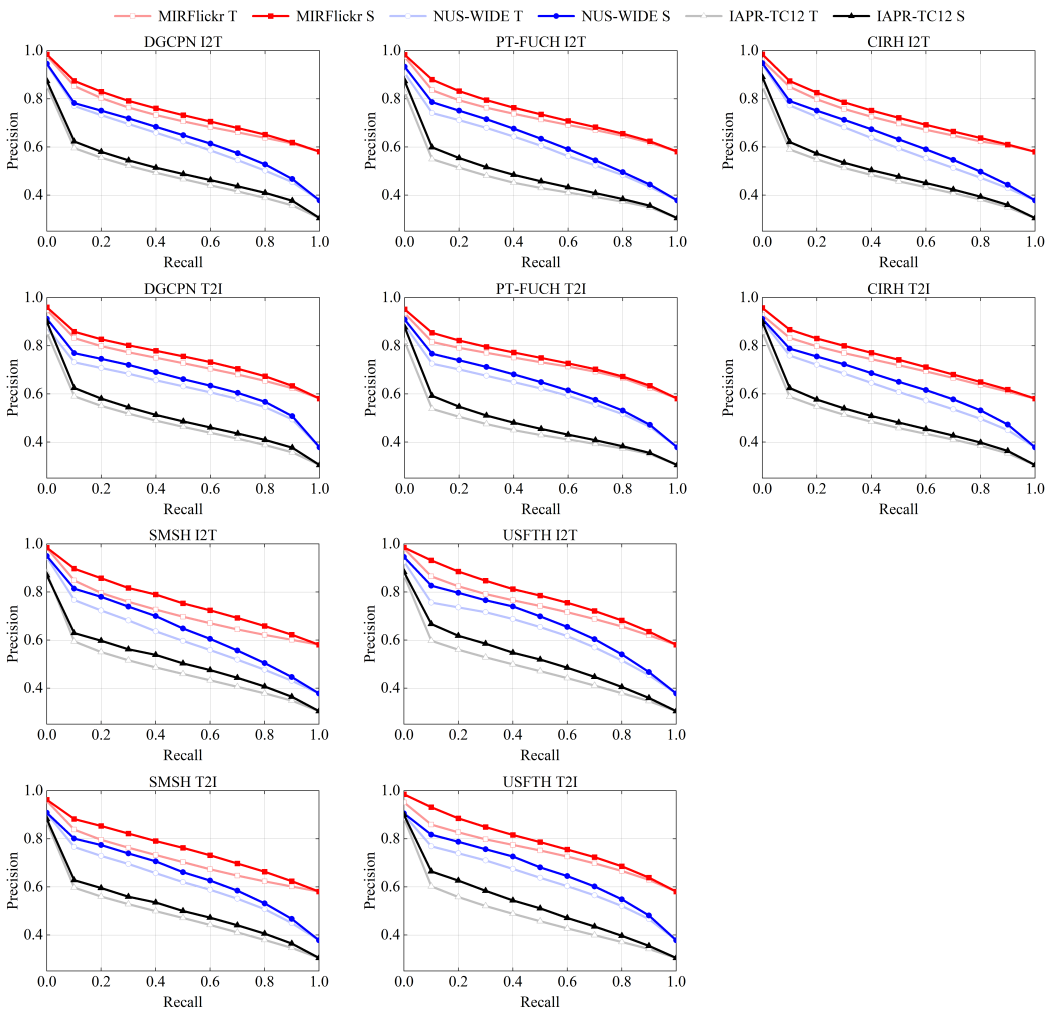


Figure 6: PR curves of DGCPN, PT-FUCH, CIRH, SMSH, USFTH with teacher-**RDDH** on MIRFlickr, NUS-WIDE, IAPR-TC12 datasets within code length of 128 bits.

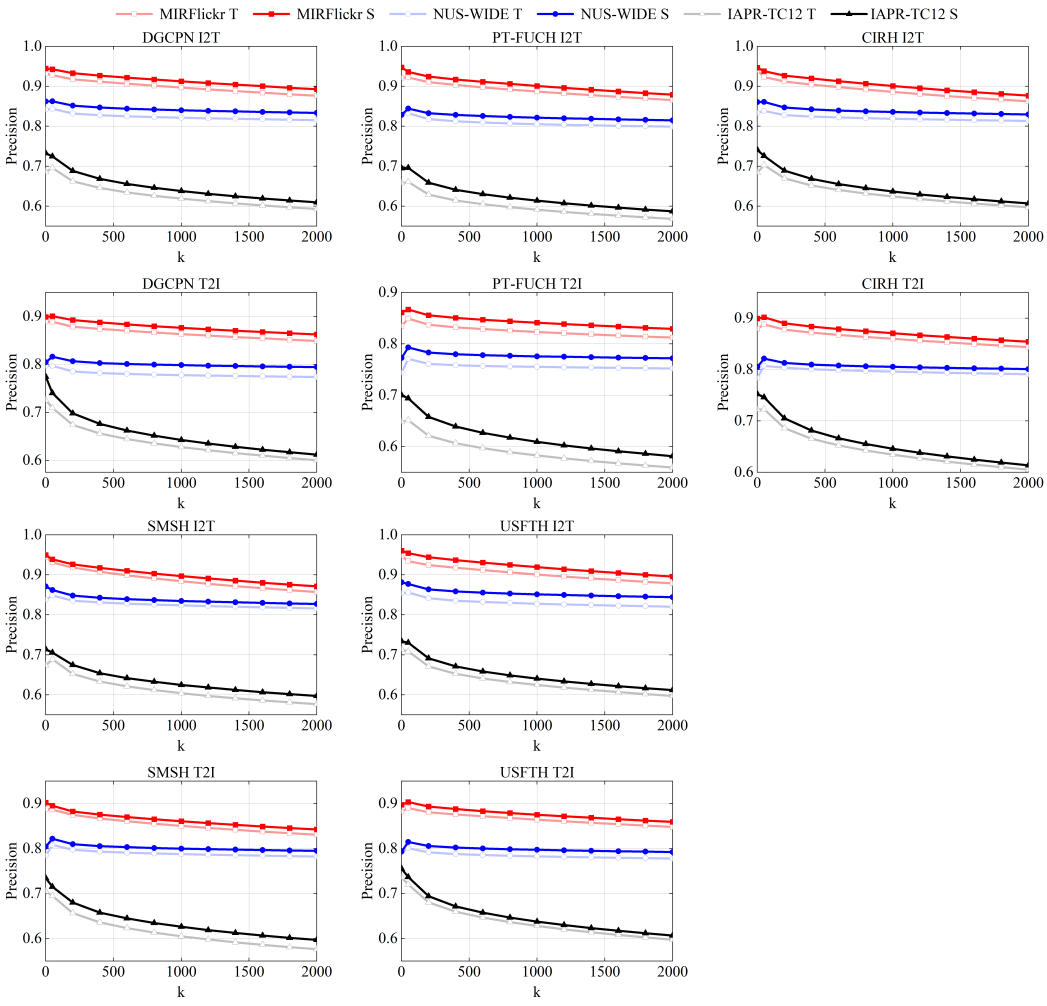


Figure 7: Top-$k$ curves of DGCPN, PT-FUCH, CIRH, SMSH, USFTH with teacher-**RDDH** on MIRFlickr, NUS-WIDE, IAPR-TC12 datasets within code length of 128 bits.

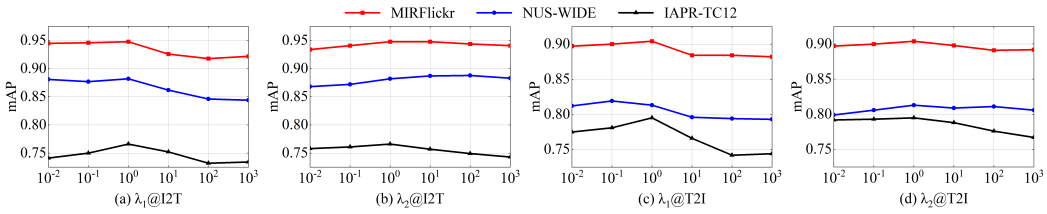


Figure 8: Parameters sensitivity analysis of $\lambda_1$ and $\lambda_2$ within DHRL-**RDDH** on MIRFlickr, NUS-WIDE, IAPR-TC12 datasets within code length of 128 bits.

Table 3: The mAP results of teacher-**RDDH** and variants on MIRFlickr and IAPR-TC12 datasets.

| Method | Task | MIRFlickr | | IAPR-TC12 | | Task | MIRFlickr | | IAPR-TC12 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 64 bits | 128 bits | 64 bits | 128 bits | | 64 bits | 128 bits | 64 bits | 128 bits |
| DGCPN-**RDDH**-$\mathcal{R}$ | I → T | 0.914 | 0.921 | 0.688 | 0.699 | T → I | 0.882 | 0.884 | 0.708 | 0.718 |
| DGCPN-**RDDH**-$\mathcal{C}$ | | 0.925 | 0.932 | 0.679 | 0.693 | | 0.877 | 0.893 | 0.678 | 0.693 |
| DGCPN-**RDDH**-$\mathcal{B}$ | | 0.898 | 0.906 | 0.692 | 0.695 | | 0.856 | 0.869 | 0.705 | 0.714 |
| DGCPN-**RDDH** | | **0.929** | **0.943** | **0.712** | **0.725** | | **0.897** | **0.901** | **0.726** | **0.740** |
| PT-FUCH-**RDDH**-$\mathcal{R}$ | | 0.898 | 0.915 | 0.651 | 0.684 | | 0.827 | 0.829 | 0.649 | 0.676 |
| PT-FUCH-**RDDH**-$\mathcal{C}$ | | 0.908 | 0.928 | 0.625 | 0.636 | | 0.813 | 0.838 | 0.615 | 0.645 |
| PT-FUCH-**RDDH**-$\mathcal{B}$ | | 0.901 | 0.918 | 0.648 | 0.682 | | 0.830 | 0.852 | 0.651 | 0.672 |
| PT-FUCH-**RDDH** | | **0.916** | **0.936** | **0.669** | **0.696** | | **0.842** | **0.867** | **0.668** | **0.694** |
| CIRH-**RDDH**-$\mathcal{R}$ | | 0.915 | 0.923 | 0.682 | 0.699 | | 0.869 | 0.883 | 0.714 | 0.731 |
| CIRH-**RDDH**-$\mathcal{C}$ | | 0.927 | 0.936 | 0.678 | 0.695 | | 0.891 | 0.892 | 0.704 | 0.717 |
| CIRH-**RDDH**-$\mathcal{B}$ | | 0.914 | 0.921 | 0.686 | 0.703 | | 0.874 | 0.881 | 0.712 | 0.728 |
| CIRH-**RDDH** | | **0.931** | **0.938** | **0.707** | **0.726** | | **0.897** | **0.902** | **0.730** | **0.746** |
| SMSH-**RDDH**-$\mathcal{R}$ | | 0.930 | 0.935 | 0.691 | 0.702 | | 0.883 | 0.885 | 0.698 | 0.712 |
| SMSH-**RDDH**-$\mathcal{C}$ | | 0.926 | 0.932 | 0.679 | 0.682 | | 0.874 | 0.872 | 0.686 | 0.691 |
| SMSH-**RDDH**-$\mathcal{B}$ | | 0.927 | 0.937 | 0.692 | 0.704 | | 0.887 | 0.889 | 0.695 | 0.710 |
| SMSH-**RDDH** | | **0.932** | **0.939** | **0.699** | **0.706** | | **0.890** | **0.895** | **0.707** | **0.715** |
| USFTH-**RDDH**-$\mathcal{R}$ | | 0.938 | 0.945 | 0.711 | 0.725 | | 0.886 | 0.895 | 0.717 | 0.729 |
| USFTH-**RDDH**-$\mathcal{C}$ | | 0.934 | 0.939 | 0.702 | 0.720 | | 0.887 | 0.897 | 0.696 | 0.723 |
| USFTH-**RDDH**-$\mathcal{B}$ | | 0.934 | 0.943 | 0.711 | 0.724 | | 0.884 | 0.893 | 0.707 | 0.727 |
| USFTH-**RDDH** | | **0.945** | **0.954** | **0.723** | **0.730** | | **0.890** | **0.903** | **0.723** | **0.737** |