Retrieval-Guided Completion Hashing with Token–Patch Alignment for Incomplete Cross-Modal Retrieval

Zhixi Luo , Zhenqiu Shu*

Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China

luozx@stu.kust.edu.cn, shuzhenqiu@163.com

Abstract

Cross-modal hashing has been extensively adopted in cross-modal retrieval tasks owing to its high storage efficiency and fast retrieval capability. However, in practical applications, multimodal data often suffer from modality missing issues, which cause semantic incompleteness and thus severely impair both cross-modal alignment and hashing-based retrieval performance. To address these issues, this paper proposes a novel incomplete cross-modal retrieval framework, called retrieval-guided completion hashing with token–patch alignment (RGCH-TPA). Specifically, it first constructs a dynamic memory bank based on the features of complete samples and uses available modalities to retrieve and locate similar regions of incomplete samples. Thus, it achieves intra-modal reconstruction of missing data under the condition of available modalities using retrieved expert signals. Moreover, a token-patch level cross-modal alignment mechanism is introduced to model the interaction between text tokens and image patches, while global representations are jointly integrated to further enhance the discriminability and robustness of the learned cross-modal representations. Extensive evaluations on three benchmark datasets demonstrate that the proposed RGCH-TPA method consistently outperforms other state-of-the-art incomplete cross-modal hashing methods across missing-modality ratios. The source code is available at https://github.com/szq0816/RGCH-TPA.

1 Introduction

With the explosive growth of intelligent devices and social networks, efficient information retrieval [Jiang and Li, 2017] has become a critical topic in both academia and industry. As a promising similarity retrieval technique, cross-modal retrieval (CMR) [Zhen *et al.*, 2019; Shen *et al.*, 2020; Hu *et al.*, 2023; Wang *et al.*, 2015; He *et al.*, 2022] aims to retrieve semantically relevant instances from one modality when given a query from another modality [Zhang *et al.*, 2022b]. Among them, hashing (CMH) [Tu *et al.*, 2022; Zhang *et al.*, 2022a; Bai *et al.*, 2023; Li *et al.*, 2024] is considered an efficient mechanism for CMR, which learns a set of hash functions to map raw data into compact binary codes, thereby enabling efficient retrieval. The recent advances in deep CMH [Jiang and Li, 2017; Tu *et al.*, 2022; Li *et al.*, 2018; Shu *et al.*, 2024] have attracted widespread attention. Compared with shallow CMH [Wang *et al.*, 2020], deep CMH achieves superior retrieval performance, owing to its powerful representation capability derived from end-to-end training and nonlinear hash functions.

However, existing CMH methods generally assume that the multimodal data are complete without missing modalities. However, in practical applications, multimodal data often suffer from missing modalities due to device limitations, privacy concerns, and high annotation costs [Lang *et al.*, 2025]. The information gap caused by missing modalities disrupts the correlations between cross-modal features, thereby affecting the discriminability of hash codes. Most existing incomplete CMH methods rely on the available modality to recover features of the missing modality. [Peng *et al.*, 2025]. These methods rely on static data distribution assumptions and often ignore the structure of the retrieval sample space during completion. They neglect prior knowledge from globally similar samples, resulting in significant distribution deviations and inconsistencies between the completed features and retrieval objectives. In addition, images contain rich and continuous information, whereas texts are inherently sparse and abstract in expression [Zareapoor *et al.*, 2025]. This asymmetry in information density between modalities makes it difficult for strategies relying solely on global feature alignment to establish sufficient and stable cross-modal semantic correlations.

To address these challenges, this paper proposes a novel completion hashing method, called retrieval-guided completion hashing with token–patch alignment (RGCH-TPA), for cross-modal retrieval. In the proposed model, we employ cross-instance retrieval as dynamic structural guidance for modality completion. It effectively exploits retrieval results as reliable expert signals to guide the recovery of missing modality information. By dynamically maintaining a feature memory bank and jointly optimizing completion, the proposed framework establishes a closed-loop mechanism in which retrieval feedback continuously refines the completion process. Meanwhile, to mitigate cross-modal misalignment

*Corresponding author.

caused by asymmetric information density, we introduce a fine-grained token-to-patch alignment strategy that enables bidirectional local semantic interaction between text and image modalities. By jointly modeling retrieval-guided completion and local cross-modal alignment, the two components form a mutually reinforcing process: the completion module supplies reliable feature foundations, while the alignment module provides precise semantic correspondences for robust hash learning. This collaborative design significantly enhances retrieval performance under incomplete-modality scenarios.

The main contributions of this work are summarized as follows:

- We propose a novel retrieval-guided modality completion strategy that leverages cross-instance retrieval results as expert priors to recover missing modalities. Instead of treating completion as an isolated feature generation, it formulates modality recovery as a collaborative optimization process driven by retrieval objectives.
- We design a token-to-patch alignment and enhancement module that explicitly models the semantic correspondence between text tokens and image patches. Therefore, it effectively alleviates insufficient cross-modal alignment caused by information asymmetry.
- Extensive experiments on three benchmark datasets show that our RGCH-TPA method consistently outperforms other state-of-the-art approaches, demonstrating its superiority in incomplete CMH tasks.

2 Related Work

2.1 Complete Cross-modal Retrieval

Cross-modal hashing (CMH) seeks to learn a set of hash functions that project heterogeneous multimodal data into a common Hamming space to generate fixed-length hash codes. Existing CMH methods are typically grouped into unsupervised and supervised paradigms. Unsupervised CMH methods preserve semantic proximity by exploiting inter-modal correlations without manual annotations. For instance, DJSRH [Su *et al.*, 2019] and DGCPN [Yu *et al.*, 2021] model inter-instance semantic neighborhoods, thereby providing structural priors for discriminative binary code learning. In contrast, UCMFH [Xia *et al.*, 2023] uses frozen CLIP encoders and contrastive learning to jointly supervise multimodal fusion and binary code generation. Unlike methods that rely solely on contrastive learning, DDSS [Huang *et al.*, 2025] further incorporates dual-branch representation learning and dual-path instance hashing to explicitly capture intra-modal topology and inter-modal semantic consistency. To enhance retrieval performance, several supervised CMH methods leverage manually annotated labels as explicit semantic priors to guide hash function learning. A representative CNN-based method, DCMH [Jiang and Li, 2017], utilizes the strong nonlinear representation power of CNNs along with label supervision to learn highly separable binary hash codes. With the emergence of large-scale vision–language models (i.e., CLIP), Transformer-based hashing methods have shown superior multimodal contextual modeling capabilities. For example, DCMHT [Tu *et al.*, 2022] adopts ViTs to obtain more expressive visual embeddings and enables end-to-end differentiable hashing, thereby improving fine-grained cross-modal representation learning. In addition, DNPH [Huo *et al.*, 2024] incorporates learnable semantic class proxies and alignment constraints to improve class-wise compactness and cross-modal semantic separation in the binary space. However, hard label-induced decision margins may conflict with cross-modal semantic similarity distributions due to modality heterogeneity. Therefore, DScPH [Qin *et al.*, 2025] models asymmetric intra- and inter-class correlations and introduces a semantic-consistent penalization loss, improving its retrieval results.

2.2 Incomplete Cross-modal Retrieval

Cross-modal retrieval remains highly challenging when multimodal data is incomplete. In early studies, DAVAE [Jing *et al.*, 2020] learns aligned latent spaces via modality-specific VAEs and generates missing-modality latent embeddings through VAE-based synthesis, thereby bypassing explicit retrieval-space modeling. Additionally, MCCN [Zeng *et al.*, 2021a] and PAN [Zeng *et al.*, 2021b] facilitate missing-modality learning by constructing category prototypes. In hashing-based retrieval, early methods, such as CYC-DGH [Wu *et al.*, 2018], employ adversarial training to generate missing-modality features in a cycle-consistent manner. Subsequently, SICH [Peng *et al.*, 2025] introduces an autoencoder as a data recovery network to recover missing-modality features using information entropy theory. This is an important attempt for unsupervised incomplete hashing, but the lack of label supervision limits its retrieval performance. CICH [Luo *et al.*, 2024] reconstructs global similarity through label supervision and contrastive learning, and further aligns contextual correspondences to handle missing modalities and mismatched pairs. Beyond cross-modal hashing, several works also explore robust multimodal hashing in incomplete data settings. For example, GCIMH [Shen *et al.*, 2023] employs a teacher-student framework, in which a GCN-based autoencoder encodes incomplete multimodal structural semantics as reconstruction priors. These priors are then distilled into a student hashing network to learn semantically separable binary hash codes.

3 Methodology

3.1 Notations and Problem Definition

We aim to project high-dimensional image and text features into a shared K-bit discrete Hamming space, and learn unified hash codes in supervised incomplete-modality scenarios, where K denotes the hash code length. We assume that the incomplete multimodal dataset $\boldsymbol{O} = \{\boldsymbol{O}^{vt}, \boldsymbol{O}^{v}, \boldsymbol{O}^{t}\}$ containing n instances. $\mathbf{O}^{vt} = \{(V, T, L)\}^{n_f}$ denotes the complete image–text pairs with semantic labels L, where n_f is the number of complete image–text pairs. $\boldsymbol{O}^{v} = \{(V, L)\}^{n_v}$ denotes image-only instances, where n_v is the number of image-only samples. Similarly, $\boldsymbol{O}^{t} = \{(T, L)\}^{n_t}$ denotes text-only samples, where n_t is the number of text-only samples. Note that the total instance number satisfies $n = n_f + n_v + n_t$. Figure 1

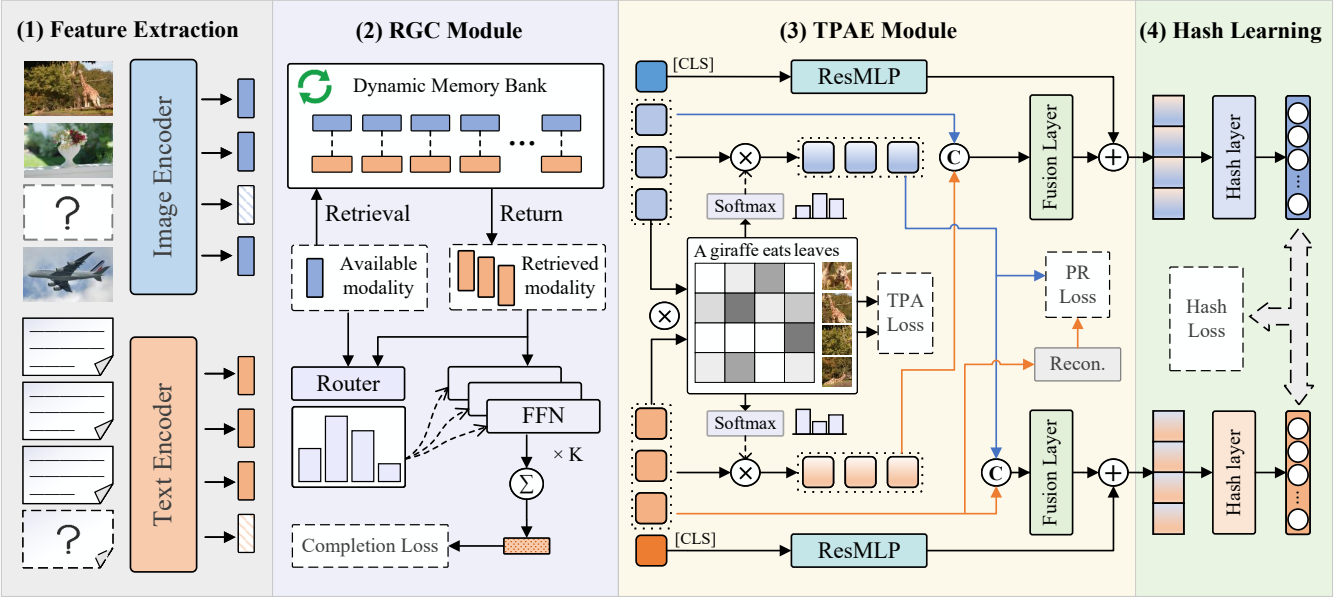

Figure 1: Overall Framework of the proposed RGCH-TPA method. Our framework consists of four components: a) Multimodal feature extraction using the pre-trained CLIP network; b) The RGC module for accurate completion of missing modalities; c) The TPAE module for precisely aligned cross-modal global representations; d) Hash learning with similarity preservation loss.

illustrates the overall workflow of our proposed RGCH-TPA framework.

3.2 Multimodal Feature Extraction

To fully utilize the Transformer's global contextual modeling capacity, the proposed RGCH-TPA method employs pretrained ViTs [Dosovitskiy, 2020] and GPT-2 [Radford *et al.*, 2019] as the image and text encoders, respectively. Even when only a single modality is available, self-attention models long-range dependencies to extract robust unimodal features $F_v = \{g_v, z_v\}$ and $F_t = \{g_t, z_t\}$. Here, $F_v \in \mathbb{R}^{B\times(L_v+1)\times d}$ and $F_t \in \mathbb{R}^{B\times(L_t+1)\times d}$, where g represents the global class embedding, z refers to local patch/token embeddings, B denotes the batch size, and $d = 512$ is the shared embedding dimension. To ensure compatibility with pretrained Transformers, the proposed RGCH-TPA method applies explicit zero-padding to all missing-modality inputs before encoding, thus preserving input alignment without altering encoder interfaces.

3.3 Retrieval-Guided Completion

In incomplete cross-modal hashing retrieval, existing methods complete missing modalities without modeling the intrinsic structure of the retrieval space. To address this issue, we design a retrieval-guided completion (RGC) module in our proposed method.

The RGC module maintains a dynamic memory bank that stores the image and text sequence features in a paired manner. During model training, only complete image–text pairs are stored in the memory bank to update features from the previous epoch. When a modality is missing, our RGC module uses the remaining available modality features as queries to retrieve semantically similar cross-modal expert features from the memory bank. Specifically, for query feature $F_q \in \mathbb{R}^{S_q\times d}$ and memory bank features $F_m \in \mathbb{R}^{N\times S_m\times d}$, we first average along the sequence dimension to obtain embeddings f_q and f_m, and then compute cosine similarity as follows:

$$\mathrm{sim}(f_q, f_m) = \frac{f_q \cdot f_m}{\|f_q\|\|f_m\|}. \tag{1}$$

We then select the Top-K memory samples with the highest similarity, and preserve their full sequence features as expert candidates of F_q: $\mathcal{E} = \{F_m^{(1)}, \ldots, F_m^{(K)}\} \in \mathbb{R}^{K\times S_m\times d}$.

To adaptively allocate expert importance, we introduce a modality-aware context router that estimates routing scores by semantic relevance between available modality features and expert features. To enhance representations, we first linearly project the remaining available modality feature F_q and expert features $\mathcal{E} \in \mathbb{R}^{K\times S\times d}$ as follows:

$$\hat{F}_q = \mathcal{W}_q F_q, \quad \hat{\mathcal{E}} = \mathcal{W}_m \mathcal{E}, \tag{2}$$

where $\mathcal{W}_q$ and $\mathcal{W}_m$ are learnable projection matrices. We flatten the projected features along the sequence dimension to obtain global representations $\bar{F}_q$ and $\bar{\mathcal{E}}$. Next, we calculate the correlations between the remaining modality and each expert via matrix multiplication, and obtain routing scores using softmax normalization:

$$\alpha = \mathrm{Softmax}(\bar{F}_q \bar{\mathcal{E}}^\top) \in \mathbb{R}^{1\times K}, \tag{3}$$

where α_k denotes the routing weight of the k-th expert for the query feature F_q, satisfying $\sum_{k=1}^{K} \alpha_k = 1$. This design ensures routing weights are conditioned on available modality semantics, enabling retrieval-consistent expert selection.

After obtaining routing weights, the proposed RGC module generates missing-modality features using a conditional mixture of experts (CMoE) generator. The CMoE generator comprises K parallel expert networks, each implemented as a feed-forward network (FFN). After forwarding all experts, we compute a routing-weighted aggregation of expert outputs to generate the missing-modality features as follows:

$$F^{gen} = \sum_{k=1}^{K} \alpha_k \cdot \mathrm{FFN}_k(\mathcal{E}_k). \tag{4}$$

Overall, the RGC module is symmetrically designed to handle both image-missing and text-missing cases. It synthesizes missing features conditioned on available modalities and reinserts them into missing positions to form completed features $\tilde{F}_v = \{\tilde{g}_v, \tilde{z}_v\}$ and $\tilde{F}_t = \{\tilde{g}_t, \tilde{z}_t\}$. Meanwhile, to ensure semantic fidelity between generated and original features, the proposed RGC method applies a completion-consistency regularization on complete image–text pairs during training:

$$\mathcal{L}_{\mathrm{com}} = \frac{1}{M} \sum_{*\in\{v,t\}} \sum_{i=1}^{M} \|F^{gen}_{*,i} - F^{ori}_{*,i}\|_2^2, \tag{5}$$

where M denotes the number of complete sample pairs within a minibatch.

3.4 Token-to-Patch Alignment and Enhancement

After obtaining the completed image and text features, we introduce a token-to-patch alignment and enhancement (TPAE) module to achieve fine-grained semantic alignment between text tokens and image patches, and then construct enhanced global representations for hash learning.

Token to Patch Interaction. To measure semantic alignment between text tokens and image patches, we explicitly model their token–patch similarity matrix as follows: $S = \tilde{z}_t \cdot \tilde{z}_v^{\top}$. Then we apply softmax along the patch dimension to derive text-to-image soft alignment weights as follows:

$$A = \mathrm{Softmax}(S, \mathrm{dim} = -1) \in \mathbb{R}^{B\times L_t \times L_v}, \tag{6}$$

where $A_{b,t,v}$ denotes the attention allocation from the t-th text token to the v-th image patch for the b-th instance, satisfying $\sum_{v=1}^{L_v} A_{b,t,v} = 1$. We aggregate image patch embeddings to produce token-aligned visual representations as follows:

$$z_v^{align} = A \cdot \tilde{z}_v \in \mathbb{R}^{B\times L_t \times d}. \tag{7}$$

Symmetrically, we perform image-guided token aggregation by normalizing the similarity matrix along the textual dimension to obtain z_t^{align}. To ensure text tokens preserve reconstructive semantics for aligned visual patches, we design a two-layer MLP reconstructor as follows: $z_v^{recon} = \mathrm{Reconstructor}(\tilde{z}_t)$. The patch reconstruction is supervised using aligned patch embeddings with a feature-level L_1 loss:

$$\mathcal{L}_{\mathrm{PR}} = \frac{1}{L_t} \sum_{i=1}^{L_t} \|z_{v,i}^{recon} - z_{v,i}^{align}\|_1. \tag{8}$$

This constraint forces text tokens to preserve reconstructive semantics for corresponding visual patches, improving local cross-modal feature consistency. Furthermore, to reinforce discriminative token–patch correspondence, we design an alignment loss with hard positive and soft negative patch sampling. For each text token, we select the highest-scoring image patch as the hard positive: $s_j^+ = \max_i S_{j,i}$. We also sample the lowest-scoring patch among the remaining patches as the soft negative: $s_j^- = \min_i S_{j,i}$. The token-to-patch alignment loss is given as follows:

$$\mathcal{L}_{\mathrm{TPA}} = -\mathbb{E}\left[\log \sigma(s^+)\right] - \mathbb{E}\left[\log(1 - \sigma(s^-))\right], \tag{9}$$

where $\sigma(\cdot)$ denotes the Sigmoid function. It aims to promote high similarity between each text token and its most semantically relevant image patch, while suppressing interference from irrelevant patches.

Then we enhance local features through bidirectional cross-modal fusion as follows:

$$\begin{aligned} z_v^{fused} &= \mathrm{Linear}_v[\tilde{z}_v; z_t^{align}] \in \mathbb{R}^{B\times L_v \times d}, \\ z_t^{fused} &= \mathrm{Linear}_t[\tilde{z}_t; z_v^{align}] \in \mathbb{R}^{B\times L_t \times d}, \end{aligned} \tag{10}$$

where $[\cdot;\cdot]$ denotes concatenation, and $\mathrm{Linear} \in \mathbb{R}^{2d\to d}$ is a learnable linear fusion projection layer.

Global Representation Enhancement. To improve the discriminability of global features, a residual MLP is introduced to nonlinearly enhance the global embeddings: $\hat{g}_v = \mathrm{ResMLP}(\tilde{g}_v)$, $\hat{g}_t = \mathrm{ResMLP}(\tilde{g}_t)$. Then, we use a symmetric contrastive learning loss to align the enhanced high-order features $\hat{g}_v$ and $\hat{g}_t$. Specifically, in a minibatch M, we learn discriminative cross-modal representations by contrasting the similarity relationships between positive and negative pairs:

$$\begin{aligned} \mathcal{L}_{\mathrm{global}} = &-\frac{1}{2M} \sum_{i=1}^{M} \log \frac{\exp\left((\hat{g}_i^v)^{\top} \hat{g}_i^t / \tau\right)}{\sum_{j=1}^{M} \exp\left((\hat{g}_i^v)^{\top} \hat{g}_j^t / \tau\right)} \\ &-\frac{1}{2M} \sum_{i=1}^{M} \log \frac{\exp\left((\hat{g}_i^t)^{\top} \hat{g}_i^v / \tau\right)}{\sum_{j=1}^{M} \exp\left((\hat{g}_i^t)^{\top} \hat{g}_j^v / \tau\right)}, \end{aligned} \tag{11}$$

where τ is the temperature hyperparameter.

Finally, we apply average pooling on the fused local features and integrate them with the enhanced global class embeddings to form the final global representations:

$$f_v = \hat{g}_v + \frac{1}{L_v} \sum_{i=1}^{L_v} z_{v,i}^{fused}, \quad f_t = \hat{g}_t + \frac{1}{L_t} \sum_{j=1}^{L_t} z_{t,j}^{fused}. \tag{12}$$

This design preserves global semantic consistency while incorporating fine-grained token–patch correspondence, thus yielding more discriminative global representations for supervised hash code learning.

3.5 Hash Learning

The hash learning stage aims to encode the aligned cross-modal global representations into binary hash codes, while preserving retrieval discriminability and cross-modal consistency. We design a hashing layer that projects global representations into K-dimensional real-valued features to obtain

H_v and H_t, and then apply the Sign function to obtain compact hash codes B_v and B_t, respectively. To improve the discriminability of hash codes and preserve the semantic structure of multimodalities, the hash loss consists of the following loss functions:

1) Intra-modal similarity preserving loss:

$$\mathcal{L}_{\text{intra}} = \frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(-S_{ij}\Omega_{ij}^{*} + \log(1+e^{\Omega_{ij}^{*}})\right), \quad (13)$$

where $\Omega_{ij}^{(*)} = \frac{1}{2}(h_i^{(*)})^{\top}h_j^{(*)}$ measures the intra-modal similarity via inner product, S_{ij} is the label-derived semantic similarity between the i-th and j-th instances, and $* \in (v, t)$.

2) Cross-modal similarity preserving loss:

$$\begin{aligned}\mathcal{L}_{\text{inter}} = -\frac{1}{MN}\Bigg(&\sum_{i=1}^{N}\sum_{j=1}^{M}\left(S_{ij}\Theta_{ij} - \log\left(1+e^{\Theta_{ij}}\right)\right)\\ &+\sum_{i=1}^{N}\sum_{j=1}^{M}\left(S_{ij}\Phi_{ij} - \log\left(1+e^{\Phi_{ij}}\right)\right)\Bigg),\end{aligned} \quad (14)$$

where $\Theta_{ij} = \frac{1}{2}(h_i^t)^{\top}h_j^v$ and $\Phi_{ij} = \frac{1}{2}(h_i^v)^{\top}h_j^t$ denote the inter-modal similarity via inner product.

3) Quantization loss: To learn unified binary hash codes, we fuse the real-valued outputs from both modalities into a shared binary code $b_i = \text{sign}\left(h_i^v + h_i^t\right)$, and minimize the quantization errors between b_i and the real-valued features of each modality:

$$\mathcal{L}_{\text{quan}} = \frac{1}{KM}\sum_{*\in\{x,y\}}\sum_{i=1}^{M}\left(\left\|b_i - h_i^{*}\right\|_2^2\right). \quad (15)$$

Finally, the overall loss function of the proposed RGCH-TPA method is given as follows:

$$\begin{aligned}\mathcal{L} = \alpha\mathcal{L}_{\text{com}} + \beta\mathcal{L}_{\text{global}} + \gamma_1\mathcal{L}_{\text{TPA}} + \gamma_2\mathcal{L}_{\text{PR}}\\ +\lambda_1\mathcal{L}_{\text{intra}} + \lambda_2\mathcal{L}_{\text{inter}} + \mathcal{L}_{\text{quan}},\end{aligned} \quad (16)$$

where α, β, γ_1, γ_2, λ_1, λ_2 are weight hyperparameters.

4 Experiments

4.1 Experimental Settings

Datasets. To validate the effectiveness of the proposed method, we conduct experiments on three widely-used datasets: MIRFLICKR-25K [Huiskes and Lew, 2008], NUS-WIDE [Chua *et al.*, 2009], and MS-COCO [Lin *et al.*, 2014]. Among them, the MIRFLICKR-25K dataset consists of 24,581 image-text pairs, covering 24 categories. The NUS-WIDE dataset comprises 195,834 image–text pairs from the 21 most frequent semantic classes. The MS-COCO dataset contains 122,218 image–text pairs spanning 80 categories.

Implementation. To evaluate the performance of our approach in incomplete cross-modal hashing retrieval tasks, we follow the experimental settings of CICH [Luo *et al.*, 2024]. We divide the training set and retrieval set into three predefined modality incompleteness levels. Case 1 includes 10% complete pairs, 45% image-only, and 45% text-only instances, simulating severe modality incompleteness. We further adopt two additional incomplete proportions: Case 2 (0.3, 0.35, 0.35) and Case 3 (0.5, 0.25, 0.25). We train with a batch size of 128 for 30 epochs, using $K = 4$ experts in the CMoE generator. We optimize the network using AdamW [Loshchilov and Hutter, 2017] with a weight decay of 0.01 for hash learning. We set the learning rate of both encoders to 1e-6, while the other modules use dataset-specific learning rates: 1e-3 for MIRFLICKR-25K, 2e-3 for NUS-WIDE and MS COCO.

Metric. Following standard cross-modal hashing evaluation, we conduct two retrieval tasks: Image-to-Text (I2T) and Text-to-Image (T2I). We adopt mean Average Precision (mAP) as the primary metric for retrieval performance. In addition, Precision-Recall (PR) curves and TopK-Precision curves are employed for comprehensive evaluation.

Baselines. We compare the proposed method against nine state-of-the-art baselines, including seven complete CMH methods: UCMFH [Xia *et al.*, 2023], DDSS [Huang *et al.*, 2025], DSPH [Huo *et al.*, 2023], DNPH [Huo *et al.*, 2024], CMCL [Wu *et al.*, 2024], DNcH [Zhu *et al.*, 2025], and DScPH [Qin *et al.*, 2025]. Due to the limited availability of incomplete CMH methods and the absence of public implementations, we consider only two incomplete methods as baselines: CICH [Luo *et al.*, 2024] and GCIMH [Shen *et al.*, 2023]. Among them, GCIMH has been modified from a multimodal hashing retrieval method to a cross-modal one.

4.2 Comparisons with State-of-The-Art Methods

To comprehensively evaluate the effectiveness of our approach under incomplete multimodal settings, we conduct comparisons with several state-of-the-art cross-modal hashing methods across three levels of incompleteness on three datasets. Due to limited space, only the 32-bit retrieval results are reported in Table 1 and Figure 2.

The results indicate that the proposed RGCH-TPA method consistently achieves the best performance across all datasets and missing-modality ratios, demonstrating its robustness under incomplete multimodal scenarios. In the most challenging Case 1, the proposed method still attains mean mAPs of 0.8776, 0.7620, and 0.7547 on three datasets, respectively, clearly outperforming other state-of-the-art hashing methods. It can be observed that the performance of all methods improves as the proportion of complete samples increases (Cases 2 and 3). Nevertheless, our method shows a more pronounced improvement and consistently outperforms others, demonstrating its effectiveness in exploiting complete data via retrieval-guided completion. Figure 2 further reports the PR and TopK-Precision curves under Case 2. The proposed RGCH-TPA method consistently yields higher precision over a wide recall range, especially in the high-recall region, showing improved ranking stability. Moreover, it achieves superior TopK-Precision for large K, confirming that the learned hash codes preserve reliable semantic neighborhoods in the Hamming space. These results verify that retrieval-guided expert cues and token–patch level alignment allow the proposed method to learn more discriminative and robust hash representations under incomplete multimodal settings.

Proportions	Methods	MIRFLICKR-25K			NUS-WIDE			MS COCO		
		I2T	T2I	Mean	I2T	T2I	Mean	I2T	T2I	Mean
Case 1 (0.1, 0.45, 0.45)	UCMFH (2023)	0.5746	0.5633	0.5690	0.3665	0.3390	0.3527	0.4616	0.3863	0.4239
	GCIMH (2023)	0.7668	0.7707	0.7688	0.6359	0.6448	0.6403	0.6496	0.6367	0.6431
	DSPH (2023)	0.7889	0.7998	0.7944	0.6842	0.7038	0.6940	0.7230	0.7118	0.7174
	CICH (2024)	0.7504	0.7449	0.7477	0.6162	0.6379	0.6271	0.5534	0.5311	0.5423
	DNPH (2024)	0.7838	0.8001	0.7920	0.6893	0.7233	0.7063	0.6886	0.6827	0.6856
	CMCL (2024)	0.8321	0.8367	0.8344	0.7141	0.7372	0.7257	0.7124	0.7086	0.7105
	DDSS (2025)	0.6208	0.5704	0.5956	0.4479	0.3568	0.4024	0.5563	0.4421	0.4992
	DNcH (2025)	0.8023	0.8125	0.8074	0.7170	0.7328	0.7249	0.7351	0.7369	0.7360
	DScPH (2025)	0.8181	0.8300	0.8240	0.7009	0.7229	0.7119	0.7203	0.7190	0.7197
	RGCH-TPA	**0.8510**	**0.9042**	**0.8776**	**0.7446**	**0.7795**	**0.7620**	**0.7597**	**0.7497**	**0.7547**
Case 2 (0.3, 0.35, 0.35)	UCMFH (2023)	0.5801	0.5750	0.5776	0.3900	0.3632	0.3766	0.4912	0.4318	0.4615
	GCIMH (2023)	0.7774	0.7756	0.7765	0.6434	0.6497	0.6466	0.6498	0.6359	0.6429
	DSPH (2023)	0.8027	0.8116	0.8072	0.6948	0.7118	0.7033	0.7292	0.7348	0.7320
	CICH (2024)	0.7621	0.7457	0.7538	0.6162	0.6187	0.6175	0.5606	0.5321	0.5463
	DNPH (2024)	0.7911	0.8071	0.7991	0.6942	0.7183	0.7062	0.7014	0.7043	0.7029
	CMCL (2024)	0.8414	0.8358	0.8388	0.7219	0.7417	0.7318	0.7529	0.7489	0.7509
	DDSS (2025)	0.6222	0.5980	0.6101	0.4400	0.3884	0.4142	0.5983	0.5794	0.5888
	DNcH (2025)	0.8134	0.8186	0.8160	0.7120	0.7317	0.7219	0.7361	0.7340	0.7351
	DScPH (2025)	0.8241	0.8321	0.8281	0.7144	0.7373	0.7258	0.7262	0.7268	0.7265
	RGCH-TPA	**0.8767**	**0.9095**	**0.8931**	**0.7651**	**0.7879**	**0.7765**	**0.7776**	**0.7714**	**0.7745**
Case 3 (0.5, 0.25, 0.25)	UCMFH (2023)	0.5915	0.5839	0.5877	0.4082	0.3845	0.3963	0.5221	0.4755	0.4988
	GCIMH (2023)	0.7905	0.7802	0.7853	0.6422	0.6544	0.6483	0.6523	0.6491	0.6507
	DSPH (2023)	0.8052	0.7973	0.8012	0.7031	0.7155	0.7093	0.7408	0.7284	0.7346
	CICH (2024)	0.7731	0.7535	0.7633	0.6230	0.6284	0.6257	0.5626	0.5400	0.5513
	DNPH (2024)	0.7787	0.7983	0.7885	0.6873	0.7210	0.7041	0.6980	0.6777	0.6878
	CMCL (2024)	0.8464	0.8408	0.8436	0.7330	0.7481	0.7406	0.7722	0.7646	0.7684
	DDSS (2025)	0.6297	0.6112	0.6205	0.4503	0.4187	0.4345	0.6015	0.5971	0.5993
	DNcH (2025)	0.8200	0.8197	0.8199	0.7185	0.7357	0.7271	0.7381	0.7343	0.7362
	DScPH (2025)	0.8297	0.8387	0.8342	0.7145	0.7300	0.7223	0.7216	0.7202	0.7209
	RGCH-TPA	**0.8959**	**0.9138**	**0.9049**	**0.7761**	**0.7927**	**0.7844**	**0.7864**	**0.7799**	**0.7831**

Table 1: mAP results of 32 bit with three incomplete proportions on three datasets

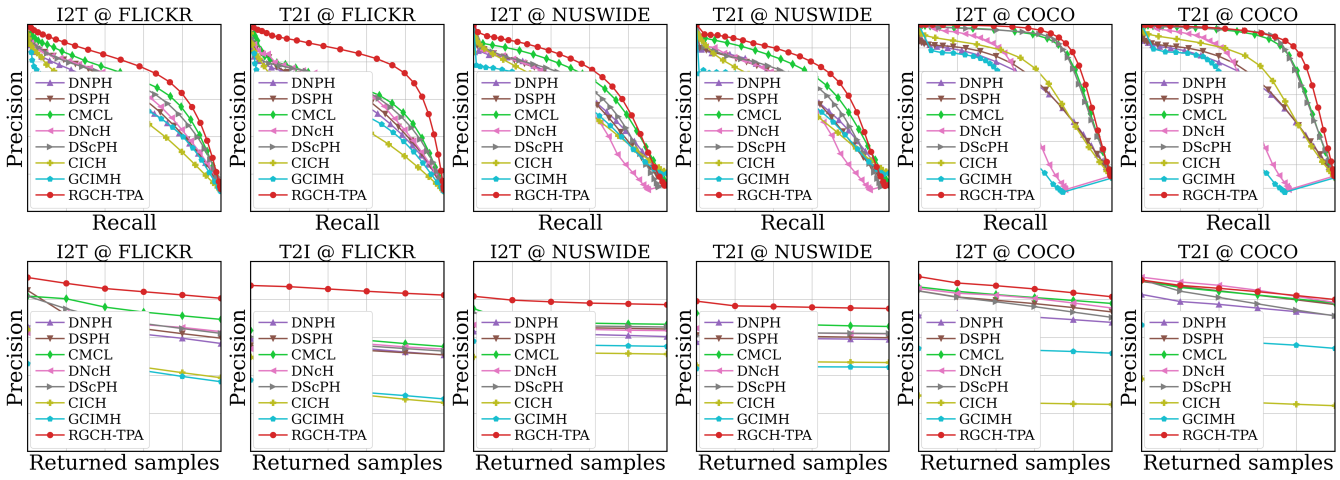

Figure 2: PR and TopK-Precision curves on three datasets w.r.t. Case 2, 32 bit.

4.3 Ablation Study

In this part, we conduct a systematic ablation study across three incompleteness ratio settings on the MIRFLICKR-25K dataset. Specifically, we take Case 1 to construct the variants as follows: a) BASE: It acts as the baseline; b) BASE+RGC: This variant adds the retrieval-guided completion (RGC) module to the baseline; c) BASE+RGC+TPI: It further adds token-patch interaction (TPI) to the model in row

Proportions	BASE	RGC	TPI	GRE	16 bits			32 bits			64 bits		
					I2T	T2I	Mean	I2T	T2I	Mean	I2T	T2I	Mean
Case 1 (0.1, 0.45, 0.45)	✓				0.7991	0.8138	0.8064	0.8104	0.8233	0.8168	0.8161	0.8252	0.8207
	✓	✓			0.8095	0.8201	0.8148	0.8222	0.8318	0.8270	0.8327	0.8387	0.8357
	✓	✓	✓		0.8226	0.8852	0.8539	0.8406	0.9013	0.8709	0.8472	0.9097	0.8784
		✓	✓	✓	0.7381	0.8181	0.7781	0.7577	0.8440	0.8008	0.7773	0.8529	0.8151
	✓	✓	✓	✓	**0.8399**	**0.8909**	**0.8654**	**0.8510**	**0.9042**	**0.8776**	**0.8618**	**0.9122**	**0.8870**
Case 2 (0.3, 0.35, 0.35)	✓				0.8121	0.8201	0.8161	0.8279	0.8300	0.8290	0.8328	0.8354	0.8341
	✓	✓			0.8191	0.8214	0.8203	0.8328	0.8342	0.8335	0.8443	0.8421	0.8432
	✓	✓	✓		0.8471	0.8897	0.8684	0.8636	0.9040	0.8838	0.8745	0.9144	0.8944
		✓	✓	✓	0.7884	0.8337	0.8111	0.8117	0.8622	0.8369	0.8228	0.8710	0.8469
	✓	✓	✓	✓	**0.8636**	**0.8988**	**0.8812**	**0.8767**	**0.9095**	**0.8931**	**0.8880**	**0.9195**	**0.9037**
Case 3 (0.5, 0.25, 0.25)	✓				0.8185	0.8252	0.8219	0.8371	0.8354	0.8363	0.8417	0.8406	0.8411
	✓	✓			0.8214	0.8263	0.8238	0.8355	0.8400	0.8377	0.8458	0.8467	0.8463
	✓	✓	✓		0.8682	0.8926	0.8804	0.8816	0.9068	0.8942	0.8953	0.9176	0.9065
		✓	✓	✓	0.8093	0.8380	0.8236	0.8334	0.8661	0.8498	0.8493	0.8767	0.8630
	✓	✓	✓	✓	**0.8811**	**0.9017**	**0.8914**	**0.8959**	**0.9138**	**0.9049**	**0.9075**	**0.9239**	**0.9157**

Table 2: mAP results of different components with three incomplete proportions on the MIRFLICKR-25K dataset.

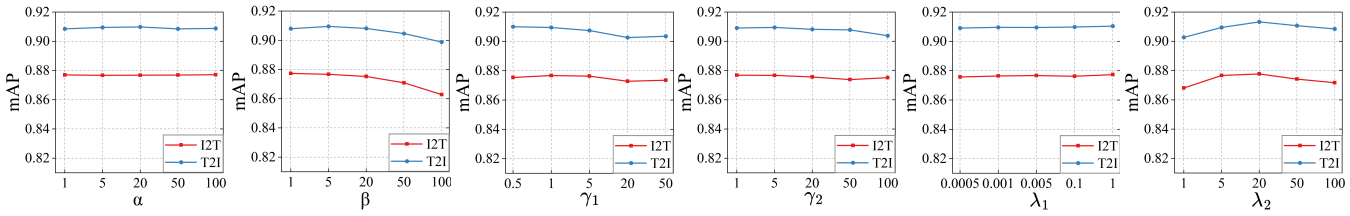

Figure 3: Performances with different parameters on the MIRFLICKR-25K dataset w.r.t. Case 2, 32-bit.

2; d) RGC+TPI+GRE: It replaces the original modal similarity preservation loss in BASE; e) BASE+RGC+TPI+GRE: This is the full model, obtained by adding the global representation enhancement (GRE) component to the model in row 3.

As observed in table 2, the BASE model achieves the lowest performance. By adding the RGC module, this variant shows consistent improvements across all incomplete configurations in both I2T and T2I tasks, indicating that retrieval-guided completion effectively mitigates semantic discontinuity caused by missing modalities. We further incorporate the TPI block, thereby leading to additional performance gains. It confirms the critical role of fine-grained cross-modal interaction in boosting cross-modal semantic consistency. Finally, by strengthening global representations through the GRE module, the full model achieves optimal performance across all incomplete scenarios. This demonstrates that collaborative modeling of global and local features is essential for learning discriminative hash representations, thereby validating the contribution of each component.

4.4 Parameter Sensitivity

This subsection analyzes the sensitivity of the proposed RGCH-TPA method to key hyperparameters on the MIRFLICKR-25K dataset under Case 2 with 32-bit hash codes. Specifically, we evaluate the influence of α, β, γ_1, γ_2, λ_1, and λ_2, with the corresponding results illustrated in Figure 3. It can be observed that when these hyperparameters vary over a wide range, the mAP curves for both I2T and T2I retrieval remain stable with only minor fluctuations. This demonstrates that the proposed RGCH-TPA method is insensitive to hyperparameter settings and consistently maintains reliable retrieval performance, confirming its robustness and practical applicability in incomplete-modality scenarios.

5 Conclusion

In this paper, we propose a novel retrieval-guided completion hashing with token–patch alignment (RGCH-TPA) framework for incomplete cross-modal retrieval. Specifically, the RGC module reconstructs high-quality missing modality features by leveraging retrieved complete-modal samples as expert guidance. The TPAE module enhances local cross-modal consistency through fine-grained token–patch semantic alignment and interaction, and performs non-linear enhancement on global features. Finally, we integrate hierarchical semantic information to generate highly discriminative hash representations. Extensive experiments on three benchmark datasets validate the superiority of the proposed framework in incomplete-modality multimodal retrieval scenarios.

Acknowledgements

This work was supported by the National Natural Science Foundation of China [Grant 62162033, 62462038], Yunnan

Xingdian Talent Support Plan Project.

References

[Bai *et al.*, 2023] Yibing Bai, Zhenqiu Shu, Jun Yu, Zhengtao Yu, and Xiao-Jun Wu. Proxy-based graph convolutional hashing for cross-modal retrieval. *IEEE Transactions on Big Data*, 10(4):371–385, 2023.

[Chua *et al.*, 2009] 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*, pages 1–9, 2009.

[Dosovitskiy, 2020] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[He *et al.*, 2022] Shiyuan He, Weiyang Wang, Zheng Wang, Xing Xu, Yang Yang, Xiaoming Wang, and Heng Tao Shen. Category alignment adversarial learning for cross-modal retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):4527–4538, 2022.

[Hu *et al.*, 2023] Peng Hu, Zhenyu Huang, Dezhong Peng, Xu Wang, and Xi Peng. Cross-modal retrieval with partially mismatched pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9595–9610, 2023.

[Huang *et al.*, 2025] Cheng Huang, Wenzhe Liu, Jinghua Wang, Jinrong Cui, and Jie Wen. Dual-driven cross-modal contrastive hashing retrieval network via structural feature and semantic information. *Information Fusion*, page 103252, 2025.

[Huiskes and Lew, 2008] Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 39–43, 2008.

[Huo *et al.*, 2023] Yadong Huo, Qibing Qin, Jiangyan Dai, Lei Wang, Wenfeng Zhang, Lei Huang, and Chengduan Wang. Deep semantic-aware proxy hashing for multi-label cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(1):576–589, 2023.

[Huo *et al.*, 2024] Yadong Huo, Qin Qibing, 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, 2024.

[Jiang and Li, 2017] Qing-Yuan Jiang and Wu-Jun Li. Deep cross-modal hashing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3232–3240, 2017.

[Jing *et al.*, 2020] Mengmeng Jing, Jingjing Li, Lei Zhu, Ke Lu, Yang Yang, and Zi Huang. Incomplete cross-modal retrieval with dual-aligned variational autoencoders. In *Proceedings of the 28th ACM international conference on multimedia*, pages 3283–3291, 2020.

[Lang *et al.*, 2025] Jian Lang, Rongpei Hong, Zhangtao Cheng, Ting Zhong, Yong Wang, and Fan Zhou. Redeeming modality information loss: Retrieval-guided conditional generation for severely modality missing learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 1241–1252, 2025.

[Li *et al.*, 2018] Chao Li, Cheng Deng, Ning Li, Wei Liu, Xinbo Gao, and Dacheng Tao. Self-supervised adversarial hashing networks for cross-modal retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4242–4251, 2018.

[Li *et al.*, 2024] Li Li, Zhenqiu Shu, Zhengtao Yu, and Xiao-Jun Wu. Robust online hashing with label semantic enhancement for cross-modal retrieval. *Pattern Recognition*, 145:109972, 2024.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[Luo *et al.*, 2024] Haoyang Luo, Zheng Zhang, and Liqiang Nie. Contrastive incomplete cross-modal hashing. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):5823–5834, 2024.

[Peng *et al.*, 2025] Shouyong Peng, Tao Yao, Ying Li, Gang Wang, Lili Wang, and Zhiming Yan. Self-supervised incomplete cross-modal hashing retrieval. *Expert Systems with Applications*, 262:125592, 2025.

[Qin *et al.*, 2025] Qibing Qin, Lei Wu, Wenfeng Zhang, Lei Huang, and Jie Nie. Deep semantic-consistent penalizing hashing for cross-modal retrieval. *IEEE Transactions on Multimedia*, 2025.

[Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[Shen *et al.*, 2020] Heng Tao Shen, Luchen Liu, Yang Yang, Xing Xu, Zi Huang, Fumin Shen, and Richang Hong. Exploiting subspace relation in semantic labels for cross-modal hashing. *IEEE Transactions on Knowledge and Data Engineering*, 33(10):3351–3365, 2020.

[Shen *et al.*, 2023] Xiaobo Shen, Yinfan Chen, Shirui Pan, Weiwei Liu, and Yuhui Zheng. Graph convolutional incomplete multi-modal hashing. In *Proceedings of the 31st ACM international conference on multimedia*, pages 7029–7037, 2023.

[Shu *et al.*, 2024] Zhenqiu Shu, Yibing Bai, Kailing Yong, and Zhengtao Yu. Deep cross-modal hashing with ranking learning for noisy labels. *IEEE Transactions on Big Data*, 11(2):553–565, 2024.

[Su *et al.*, 2019] Shupeng Su, Zhisheng Zhong, and Chao Zhang. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3027–3035, 2019.

[Tu *et al.*, 2022] Junfeng Tu, Xueliang Liu, Zongxiang Lin, Richang Hong, and Meng Wang. Differentiable cross-modal hashing via multimodal transformers. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 453–461, 2022.

[Wang *et al.*, 2015] Kaiye Wang, Ran He, Liang Wang, Wei Wang, and Tieniu Tan. Joint feature selection and subspace learning for cross-modal retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):2010–2023, 2015.

[Wang *et al.*, 2020] Yongxin Wang, Xin Luo, Liqiang Nie, Jingkuan Song, Wei Zhang, and Xin-Shun Xu. Batch: A scalable asymmetric discrete cross-modal hashing. *IEEE Transactions on Knowledge and Data Engineering*, 33(11):3507–3519, 2020.

[Wu *et al.*, 2018] Lin Wu, Yang Wang, and Ling Shao. Cycle-consistent deep generative hashing for cross-modal retrieval. *IEEE Transactions on Image Processing*, 28(4):1602–1612, 2018.

[Wu *et al.*, 2024] 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.

[Xia *et al.*, 2023] Xinyu Xia, Guohua Dong, Fengling Li, Lei Zhu, and Xiaomin Ying. When clip meets cross-modal hashing retrieval: A new strong baseline. *Information Fusion*, 100:101968, 2023.

[Yu *et al.*, 2021] Jun Yu, Hao Zhou, Yibing Zhan, and Dacheng Tao. Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 4626–4634, 2021.

[Zareapoor *et al.*, 2025] Masoumeh Zareapoor, Pourya Shamsolmoali, and Yue Lu. Bimac: Bidirectional multimodal alignment in contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 22290–22298, 2025.

[Zeng *et al.*, 2021a] Zhixiong Zeng, Ying Sun, and Wenji Mao. Mccn: Multimodal coordinated clustering network for large-scale cross-modal retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5427–5435, 2021.

[Zeng *et al.*, 2021b] Zhixiong Zeng, Shuai Wang, Nan Xu, and Wenji Mao. Pan: Prototype-based adaptive network for robust cross-modal retrieval. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 1125–1134, 2021.

[Zhang *et al.*, 2022a] Chao Zhang, Huaxiong Li, Yang Gao, and Chunlin Chen. Weakly-supervised enhanced semantic-aware hashing for cross-modal retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 35(6):6475–6488, 2022.

[Zhang *et al.*, 2022b] Zheng Zhang, Haoyang Luo, Lei Zhu, Guangming Lu, and Heng Tao Shen. Modality-invariant asymmetric networks for cross-modal hashing. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):5091–5104, 2022.

[Zhen *et al.*, 2019] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10394–10403, 2019.

[Zhu *et al.*, 2025] Congcong Zhu, Qibing Qin, Wenfeng Zhang, and Lei Huang. Deep neighbor-coherence hashing with discriminative sample mining for supervised cross-modal retrieval. *Expert Systems with Applications*, 279:127365, 2025.