Adaptive Graph Attention Based Discrete Hashing for Incomplete Cross-modal Retrieval

Shuang Zhang[1], Yue Wu[1], Lei Shi[2*], Huilong Jin[1], Feifei Kou[3], Pengfei Zhang[4]
Mingying Xu[5], Pengtao Lv[6]

[1]College of Engineering, Hebei Normal University, China
[2]State Key Laboratory of Media Convergence and Communication, Communication University of China, China
[3]School of Computer Science (National Pilot School of Software Engineering), BUPT, China
[4]School of Computer Science and Engineering, Anhui University of Science of Technology, China
[5]School of Artificial Intelligence and Computer Science, North China University of Technology, China
[6] College of Information Science and Engineering, Henan University of Technology, China
zshuang@hebtu.edu.cn; wuyue@stu.hebtu.edu.cn; leiky_shi@cuc.edu.cn; JHL981@hebtu.edu.cn; koufeifei000@bupt.edu.cn; zpf.bupt@bupt.cn; xumingying@ncut.edu.cn; pengtaolv@haut.edu.cn

Abstract

Cross-modal hashing has emerged as a pivotal solution for efficient retrieval across diverse modalities, such as images and texts, by mapping them into compact binary hash spaces. However, in real-world scenarios, the modalities data is often missing or misaligned. Existing methods are most rely on fully paired training data and ignore missing or misaligned modalities data, resulting in the semantic inconsistencies. To address these challenges, we propose an Adaptive Graph Attention-Based Discrete Hashing (AGADH) method, which consists of three parts. First, to solve the problem of missing modalities, AGADH employs a masked completion strategy to reconstruct missing modalities. Second, to mitigate semantic misalignment, AGADH leverages a Graph Attention Network (GAT) encoder-decoder architecture with alignment module to construct features from different modalities. Additionally, to enhance the fusion performance, an adaptive fusion module dynamically adjusting the contributions of image and text modalities with learnable weighting coefficients is proposed. Extensive experiments on three benchmark datasets, MS-COCO, NUS-WIDE, and MIRFlickr-25K, demonstrating that AGADH outperforms state-of-the-art methods in both fully paired and incompletely paired scenarios, showing its robustness and effectiveness in cross-modal retrieval tasks.

Introduction

The exponential proliferation of multimodal data including images, texts, and videos across platforms such as social media and search engines has spurred significant interest in cross-modal retrieval methodologies (Wang et al. 2025; Zhen et al. 2020; Li et al. 2025). These methodologies enable the querying of information across different modalities, providing users with versatile means to access data. However, as the volume of data continues to escalate, the computational cost of conducting precise searches in high-dimensional feature spaces becomes increasingly prohibitive. In response to this challenge, cross-modal hashing has emerged as a robust solution (Hu et al. 2024; Zhu et al. 2023), effectively enhancing retrieval efficiency and storage performance by encoding diverse modalities into compact binary hash spaces. Consequently, cross-modal hashing has found widespread application in the realm of multimodal retrieval.

Despite its advantages, most existing cross-modal hashing techniques (Liu et al. 2023a; Jiang et al. 2023; Liang et al. 2024) operate under the assumption that fully paired modal data is available during the training phase. This assumption implies a one-to-one correspondence between image and text samples, thereby simplifying the alignment of feature spaces across modalities and establishing semantic consistency. However, real-world applications frequently present scenarios where modalities originate from disparate data sources, leading to issues such as data loss during acquisition and transmission. In these incompletely paired situations (Shi et al. 2024; Hu et al. 2023a), traditional hashing methods struggle to accurately capture the deep semantic relationships between images and texts, resulting in significant declines in model performance. Thus, developing a hash function that maintains both semantic consistency and retrieval efficiency in the face of missing modalities has emerged as a critical challenge in contemporary cross-modal hashing research.

To address the problem of missing modalities, most of the existing studies have focused on reconstructing the missing modality in incomplete multimodal data. Unsupervised Deep Imputed Hashing (UDIH) framework (Chen et al. 2020) was the first to propose an unsupervised approach that leverages a two-stage learning strategy, mining the relative semantic similarities among multimodal data via correlation graphs, which shows promise in handling missing modalities. However, by relying only on the available modal data for training hash networks, significant semantic inconsistencies can arise when substantial portions of data are missing,

*Corresponding author.

ultimately diminishing retrieval performance. To reduce the computational overhead associated with extensive use of the original modal data, the Flexible Dual Multimodal Hashing (FDMH) framework (Wei and An 2024) also introduces a two-stage approach that maps original features from each modality into a low-dimensional anchor graph. This framework utilizes existing incomplete anchor graph to reconstruct the anchor graph of the absent modality, which also relies heavily on data augmentation and complex training pipelines. To avoid the above-mentioned problems, a Graph Convolutional Incomplete Multi-modal Hashing (GCIMH) (Peng et al. 2025) is proposed to learn hash code on incomplete multi-modal data. GCIMH develops Graph Convolutional Autoencoder to reconstruct incomplete multi-modal data with effective exploit of its semantic structure. Mutual information between different modalities, at the same time, maximizing mutual information between different modalities is also adopt to reconstruct the missing modality to avoid large-scale data pre-training. Summarizing the above, a **significant computational cost** due to the high reliance on data augmentation and complex training procedures is first need to be considered. In addition, more attention should also be paid to the **inconsistency between the reconstructed modal data and the complete modal data**. Reducing the impact of modal deficiency on feature fusion is also one of the problems that need to be solved.

To this end, we propose an Adaptive Graph Attention-Based Discrete Hashing (AGADH) framework, which is designed to handle both fully paired and incompletely paired modal data. Our method employs a masked completion strategy within the modality missing data completion module to generate semantically coherent samples for the missing modality. We utilize a Graph Attention Network (GAT) encoder to map features from various modalities into a shared low-dimensional space, while a GAT decoder reconstructs the features of the missing modality during training, thereby reducing semantic loss and enhancing the integrity of the generated hash codes. Furthermore, an adaptive weight fusion module is introduced to dynamically adjust the contributions of image and text modalities through learnable coefficients, optimizing feature representation for missing modalities. In summary, our primary contributions are as follows:

- We propose a novel framework capable of unsupervised handling of both fully paired and incompletely paired modal data, employing a masked completion strategy to reconstruct missing modalities, which simplify the training process by focusing on reconstruction tasks based on the design of masking ratio and decoder.
- We introduce a GAT encoding-GAT decoding structure with an alignment module to facilitate collaborative feature representation across diverse sources within a shared semantic space.
- We propose an adaptive weight fusion module that enriches feature representation among different modalities, thereby enhancing semantic content and improving the quality of the generated hash codes.
- We conduct extensive experiments on three benchmark datasets, MS-COCO, NUS-WIDE, and MIRFlickr-25K, demonstrating that our method significantly outperforms existing cross-modal hashing techniques.

The Proposed Method

Notation and Problem Definition

In this section, we introduce the notation employed throughout this paper. We construct a dataset $\mathbf{O}$ comprising three distinct components to simulate scenarios involving missing modality data. Specifically, $\mathbf{O}_w = \{(x_i^w, y_i^w) \,|i \in [1, N_w]\}$ denotes a fully paired dataset, while $\mathbf{O}_i = \left\{\left(x_i^I, *\right) |i \in [1, N_i]\right\}$ represents a dataset with missing text modalities. Conversely, $\mathbf{O}_t = \left\{\left(*, y_i^T\right) |i \in [1, N_t]\right\}$ indicates a dataset where image modalities are absent. Here, $*$ represents the missing data of the corresponding modality. x_i symbolizing the data from the i-th image modality, and y_i denoting the data from the i-th text modality. The image and text features obtained by processing modal data through the corresponding feature extraction network are $\mathbf{x}_i$ and $\mathbf{y}_i$. $N_* \, (* = w, i, t)$ indicates the number of samples in each respective dataset.

The objective of our study is to derive a hash code $\mathbf{B} \in \{-1, 1\}^{N_* \times c}$ from the feature representations $\mathbf{F}_i^I$ and $\mathbf{F}_i^T$ associated with images and text. In this context, c denotes the length of the hash code. This framework serves as the foundation for our proposed approach to cross-modal hashing retrieval, particularly in scenarios characterized by incomplete modality data.

Network Architecture of AGADH

The architecture of the Adaptive Graph Attention-Based Discrete Hashing (AGADH) framework comprises two primary modules, as illustrated in Figure 1. The Modality Missing Data Completion module uses a mask completion strategy and a GAT encoder-decoder to complete missing modality data and reconstruct the modality's feature representation, enriching the modality's semantic information and enhancing semantic alignment. The Hash Learning module utilizes an adaptive weight fusion mechanism to dynamically assign weights to the modality's latent representations, thereby generating optimal hash codes and enhancing modality retrieval efficiency.

Modality Missing Data Completion Module

In the context of incomplete multimodal data, many existing methods resort to discarding samples with missing modalities, thereby resulting in a significant loss of valuable information. To address this limitation, our approach begins with a mask completion strategy to generate semantically consistent samples for the missing modalities, followed by a reconstruction process utilizing a graph attention encoder-decoder framework to enhance retrieval accuracy.

Mask Completion Strategy Inspired by the construction of missing modal data in CPCMR (Xinyu et al. 2025), we employ a masking strategy to generate placeholder features for the absent modalities. For image samples lacking corresponding text, we create a placeholder text feature, while

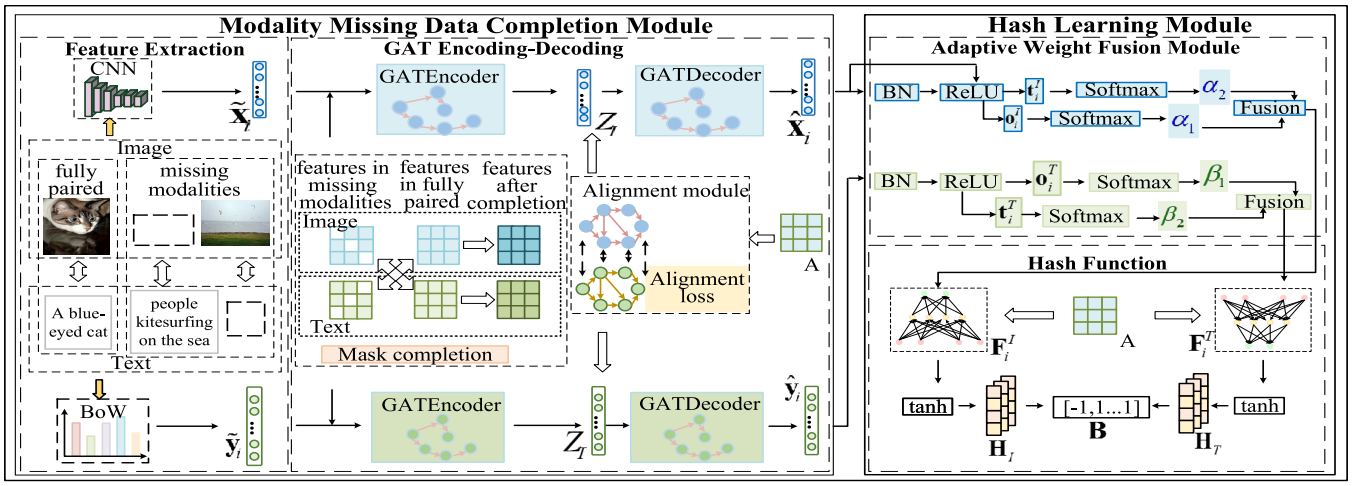

Figure 1: Overall framework of AGADH. AGADH mainly consists of two modules: 1) modality missing data completion module: utilizes GAT encoding-decoding and alignment modules to reconstruct features of missing modalities; 2) hash learning module: strengthens feature representation and generates hash codes through adaptive weight fusion module.

for text samples with missing images, a placeholder image feature is generated. The process can be mathematically represented as follows:

$$\begin{cases} \tilde{f}_i^T = \frac{\lambda}{w+m^T}\left(\sum_{j=1}^{n} f_j^T + \sum_{s=1}^{m^T} \bar{f}_s^T\right) \\ \tilde{f}_i^I = \frac{\lambda}{w+m^I}\left(\sum_{j=1}^{n} f_j^I + \sum_{s=1}^{m^I} \bar{f}_s^I\right) \end{cases}, \tag{1}$$

where $\tilde{f}_i^T$ and $\tilde{f}_i^I$ represent the placeholder text features and image features for datasets $\mathbf{O}_i$ and $\mathbf{O}_t$, respectively. λ denotes the scaling factor, n indicates the number of completely paired samples, m^T and m^I refer to the number of samples in the scenarios where text and image modalities are missing, respectively. The terms f_j^T and f_j^I represent fully paired text and image features, while $\bar{f}_s^T$ and $\bar{f}_s^I$ denote the text features corresponding to missing images and image features corresponding to missing texts, respectively. By generating these placeholder features, we ensure that the semantic information of the modalities is preserved, allowing the model to maintain semantic consistency in the presence of missing data.

Consequently, $\tilde{\mathbf{x}}_i = \left\{\mathbf{x}_i, \tilde{f}_i^I\right\}$ represents the image features input into the GAT encoder, and $\tilde{\mathbf{y}}_i = \left\{\mathbf{y}_i, \tilde{f}_i^T\right\}$ denotes the text features sent to the GAT encoder.

GAT Encoding-Decoding To tackle the challenges posed by missing modal data, we utilize a Graph Attention Network (GAT) as both the encoder and decoder. The GAT dynamically adjusts the reliance on the missing modality through its attention mechanism, thereby mitigating the impact of missing data. Initially, the image modality features, $\tilde{\mathbf{x}}_i$, are encoded using a two-layer GAT (GATConv), yielding the encoded feature h_I:

$$h_I = \mathrm{GAT}_2(\mathrm{ELU}(\mathrm{GAT}_1(\tilde{\mathbf{x}}_i, \varepsilon_x))\varepsilon_x). \tag{2}$$

where ELU represents the Exponential Linear Unit, which is employed to introduce nonlinearity, while ε_x signifies the edge index of the adjacency graph (with a complete graph utilized in the absence of external graphs). Subsequently, the parameters of the Gaussian distribution are derived through linear mapping:

$$\mu = \mathbf{W}_\mu h_I, \quad \log \sigma^2 = \mathbf{W}_{\log \sigma} h_I. \tag{3}$$

The latent representation of the image modality, Z_I is then computed as follows:

$$Z_I = \mu + \delta \cdot \sigma, \delta \in \mathcal{N}(0, I). \tag{4}$$

Given that the distribution in the latent space is constrained to approximate the prior standard normal distribution, $\mathcal{N}(0, 1)$, we incorporate KL divergence regularization to prevent overfitting and promote a smooth latent space:

$$\mathcal{L}_{KL} = -\frac{1}{2}\sum_{i=1}^{d_Z}\left(1 + \log \sigma_i^2 - \mu_i^2 - \sigma_i^2\right). \tag{5}$$

Next, Z_I is passed to the GAT decoder, where the decoded features are transformed into reconstructed data that matches the original input dimensions, yielding the reconstructed image data $\hat{\mathbf{x}}_i$:

$$\hat{\mathbf{x}}_i = \mathrm{FC}_{\mathrm{out}}(\mathrm{GAT}_2(\mathrm{ELU}(\mathrm{GAT}_1(Z_I, \varepsilon_x)), \varepsilon_x)). \tag{6}$$

The reconstruction of the text data $\hat{\mathbf{y}}_i$ follows an analogous process to that of the image data. Throughout the encoding and decoding phases, we incorporate a structural alignment module to ensure the structural similarity between different modalities. To further enhance this structural alignment, we introduce a structural alignment loss, ensuring that the heterogeneous modalities maintain consistent semantic topology within the latent space. The adjacency matrix $\mathbf{A}$ serves as an "ideal" similarity supervision, aligning the modal similarity of the embedding matrices for each modality with it. The alignment loss can be defined as follows:

$$\begin{aligned} \mathcal{L}_{\mathrm{align}} = &\left\|A - \cos(Z_I, Z_I)\right\|_F^2 + \left\|A - \cos(Z_I, Z_T)\right\|_F^2 \\ &+ \left\|A - \cos(Z_T, Z_T)\right\|_F^2 \end{aligned}. \tag{7}$$

Here, $\|\cdot\|_F^2$ denotes the Frobenius norm.

To ensure high-quality restoration of each modality post-encoding and decoding, we introduce a single-modal reconstruction loss:

$$\mathcal{L}_{recon} = \|\hat{\mathbf{x}}_i - \tilde{\mathbf{x}}_i\|_2^2 + \|\hat{\mathbf{y}}_i - \tilde{\mathbf{y}}_i\|_2^2. \tag{8}$$

Moreover, we account for reconstruction error when utilizing the existing modality to complete the missing modality:

$$\mathcal{L}_{cross-recon} = \|\hat{\mathbf{x}}_i - \mathbf{x}_i\|_2^2 + \|\hat{\mathbf{y}}_i - \mathbf{y}_i\|_2^2. \tag{9}$$

Ultimately, the comprehensive objective of the modality missing data completion module is to minimize the following expression:

$$\mathcal{L}_{total} = \alpha_1\mathcal{L}_{recon} + \alpha_2\mathcal{L}_{align} + \alpha_3\mathcal{L}_{cross-recon} - \lambda\mathcal{L}_{KL}, \tag{10}$$

where α_1, α_2, α_3, and λ are hyperparameters employed to balance the various loss components effectively.

Hash Learning Module

Adaptive Weight Fusion Module The adaptive weight fusion module is designed to integrate image and text modality features adaptively, learning to assign importance weights to both the original and transformed features of each modality.

For the image reconstruction feature representation $\hat{\mathbf{x}}_i$, we commence by normalizing it through a batch normalization (BN) layer, followed by activation using the ReLU function. This process yields an initial dimensionality reduction output, facilitating the extraction of a low-dimensional yet stable intermediate feature representation:

$$\mathbf{u}_i^I = \mathrm{ReLU}(\mathrm{BN}(\mathbf{W}_1^I\hat{\mathbf{x}}_i^I + \mathbf{b}_1^I)). \tag{11}$$

To capture higher-level semantic features within the image data, we apply an additional linear transformation to the intermediate feature representation, coupled with ReLU activation:

$$\mathbf{t}_i^I = \mathrm{ReLU}(\mathbf{W}_2^I\mathbf{u}_i^I + \mathbf{b}_2^I). \tag{12}$$

To ensure that the low-dimensional information conveyed by the original features complements the deep compression features, we directly perform a linear mapping of the original reconstructed features, followed by activation:

$$\mathbf{o}_i^I = \mathrm{ReLU}(\mathbf{W}_1\hat{\mathbf{x}}_i + \mathbf{b}_1). \tag{13}$$

Next, we utilize the softmax function to generate attention weights, allowing the model to adaptively balance the significance of the original reconstructed image feature representation against the deeper image semantic features:

$$[\alpha_{i,1}^I, \alpha_{i,2}^I]^{\mathrm{T}} = \mathrm{Softmax}([\mathbf{W}_w\mathbf{o}_i^I + \mathbf{b}_w, \mathbf{W}_w\mathbf{t}_i^I + \mathbf{b}_w]^{\mathrm{T}}). \tag{14}$$

The adaptive fused feature representation for the image modality is then computed as follows:

$$\mathbf{F}_i^I = \alpha_{i,1}^I\mathbf{o}_i^I + \alpha_{i,2}^I\mathbf{t}_i^I. \tag{15}$$

For the text reconstruction feature representation, the process mirrors that of the images, ultimately yielding the adaptive fusion feature representation $\mathbf{F}_i^T$ for the text modality.

Adjacency Matrix Module Initially, the fused feature representations for both image and text modalities are normalized and processed through a dot product to produce the corresponding similarity matrices:

$$\mathbf{S}_I(i,j) = \frac{\mathbf{F}_i^I\cdot\mathbf{F}_j^I}{\|\mathbf{F}_i^I\|\|\mathbf{F}_j^I\|} \quad \mathbf{S}_T(i,j) = \frac{\mathbf{F}_i^T\cdot\mathbf{F}_j^T}{\|\mathbf{F}_i^T\|\|\mathbf{F}_j^T\|}. \tag{16}$$

These matrices are subsequently integrated using the weight coefficient γ_1:

$$\mathbf{S}_1 = \gamma_1\mathbf{S}_I + (1-\gamma_1)\mathbf{S}_T. \tag{17}$$

To encode the potential correspondence between images and texts into a probability distribution, we utilize the similarity matrix $\mathbf{S}_2$:

$$\mathbf{S}_2(i,j) = exp(\frac{\mathbf{F}_i^I\cdot\mathbf{F}_j^T - m}{\xi})/\textstyle\sum_k exp(\frac{\mathbf{F}_i^I\cdot\mathbf{F}_k^T - m}{\xi}). \tag{18}$$

In this expression, $exp(\cdot)$ denotes the exponential function, m serves as a numerical stability term, and ξ is the temperature coefficient, ensuring that the similarity is transformed into a distribution with a row sum of 1. The final adjacency matrix $\mathbf{A}$ is obtained through the weight coefficient γ_2:

$$\mathbf{A} = \gamma_2\mathbf{S}_1 + (1-\gamma_2)\mathbf{S}_2. \tag{19}$$

Hash Function To derive the final continuous hash vector, we apply independent fully connected layers and nonlinear mapping. We first concatenate the adaptive fusion features of the image and text with the adjacency matrix $\mathbf{A}$ to yield the hash vector $\mathbf{H}$ corresponding to the mapped image and text:

$$\begin{aligned}\mathbf{H}_I &= \tanh(\mathbf{W}_I[\mathbf{A}\mathbf{F}_i^I||\mathbf{F}_i^I + \mathbf{b}_I)\\ \mathbf{H}_T &= \tanh(\mathbf{W}_T[\mathbf{F}_i^T||\mathbf{A}\mathbf{F}_i^T + \mathbf{b}_T)\end{aligned}. \tag{20}$$

where $||$ represents concatenation, $\mathbf{W}_I$ and $\mathbf{W}_T$ are weights for mapping image and text features to hash space, $\mathbf{b}_I$ and $\mathbf{b}_T$ are corresponding biases. In order to obtain a discrete binary hash code $\mathbf{B}$, we use the $sign(\cdot)$ function to obtain the final required hash code:

$$\mathbf{B} = sign(\mathbf{H}) = \begin{cases}-1, & \mathbf{H} < 0\\ 1, & \mathbf{H} \geq 0\end{cases} \tag{21}$$

Through the adaptive weight fusion module, our model not only fuses cross-modal complementary information but also enhances the global and local consistency among samples. Ultimately, this process yields a compact discrete hash code that encapsulates multimodal semantics while preserving the neighborhood structure.

Experiments

Implementation Details

In this study, we utilize three widely recognized datasets for cross-modal retrieval: MIRFlickr-25K(Huiskes and Lew 2008), NUS-WIDE(Chua et al. 2009), and MS-COCO(Lin et al. 2014). We adopt the methodology outlined in SRGMH (Li et al. 2024b), which involves removing an equal number of samples from both modalities. Specifically, we vary the partial data ratio (PDR) from 0 to 0.8 in increments of 0.2, where PDR=0 indicates that there are no missing samples

Task	Method	MIRFlickr-25K			NUS-WIDE			MS-COCO		
		16bits	32bits	64bits	16bits	32bits	64bits	16bits	32bits	64bits
I→T	DJSRH	0.6652	0.6873	0.6987	0.5271	0.5582	0.6015	0.5257	0.5454	0.5646
	JDSH	0.7276	0.7426	0.7468	0.6536	0.6601	0.6900	0.5928	0.6348	0.6517
	DCHMT	0.8177	0.8221	0.8261	0.6711	0.6812	0.6932	0.6450	0.6331	0.6647
	CDTH	0.7317	0.7461	0.7477	0.6596	0.6613	0.6700	0.5853	0.6411	0.6573
	UCCH	0.7606	0.7620	0.7674	0.6718	0.6738	0.6891	0.6039	0.6249	0.6398
	HNH	0.7742	0.8252	0.8402	0.6626	0.7650	0.7994	0.5603	0.6509	0.7108
	DNPH	0.8101	0.8114	0.8100	0.6633	0.6790	0.6891	0.6094	0.6866	0.7016
	MITH	0.8192	0.8300	0.8329	0.6992	0.7135	0.7291	0.6959	0.7188	0.7582
	AGADH	**0.8844**	**0.9101**	**0.9227**	**0.8066**	**0.8297**	**0.8466**	**0.7146**	**0.7612**	**0.8034**
T→I	DJSRH	0.6710	0.6958	0.7043	0.5575	0.5680	0.5952	0.5590	0.5591	0.5519
	JDSH	0.7304	0.7326	0.7481	0.6439	0.6640	0.6921	0.5888	0.6510	0.6635
	DCHMT	0.8007	0.8021	0.8065	0.6852	0.6963	0.7009	0.6298	0.6176	0.6616
	CDTH	0.7315	0.7464	0.7503	0.6788	0.6815	0.6910	0.5846	0.6427	0.6573
	UCCH	0.7343	0.7342	0.7410	0.6740	0.6812	0.6945	0.6023	0.6258	0.6371
	HNH	0.8004	0.8142	0.8338	0.6903	0.7130	0.7525	0.5594	0.6096	0.7291
	DNPH	0.7968	0.8171	0.8158	0.6811	0.6936	0.7098	0.6164	0.7012	0.7200
	MITH	0.8004	0.8163	0.8221	0.7135	0.7270	0.7306	0.6858	0.7089	0.7560
	AGADH	**0.8479**	**0.8724**	**0.8846**	**0.7650**	**0.7818**	**0.8051**	**0.7094**	**0.7881**	**0.8340**

Table 1: The mAP comparison results on three datasets in PDR=0

Task	Method	MIRFlickr-25K			NUS-WIDE			MS-COCO		
		16bits	32bits	64bits	16bits	32bits	64bits	16bits	32bits	64bits
I→T	UCCH	0.6925	0.6963	0.7138	0.6029	0.6544	0.6706	0.5450	0.5763	0.6075
	HNH	0.7107	0.7162	0.7402	0.6330	0.6778	0.7250	0.5448	0.6479	0.7042
	AGADH	**0.8711**	**0.8973**	**0.9176**	**0.7991**	**0.8251**	**0.8392**	**0.6990**	**0.7544**	**0.7922**
T→I	UCCH	0.6971	0.6936	0.7092	0.5958	0.6629	0.6699	0.5464	0.5767	0.6091
	HNH	0.6241	0.6766	0.6750	0.5939	0.6742	0.6710	0.4780	0.5937	0.7124
	AGADH	**0.8433**	**0.8558**	**0.8705**	**0.7542**	**0.7682**	**0.7906**	**0.6956**	**0.7656**	**0.8139**

Table 2: The mAP comparison results on three datasets in PDR=0.2

in either modality. As the PDR increases, missing samples are evenly distributed between the image and text modalities. For instance, at a PDR of 0.8, 80% of the data will have missing modalities, with 40% of the data in both the image and text modalities being affected.

The AGADH framework is executed on an NVIDIA GeForce RTX 3080 GPU, equipped with 20GB of RAM, and is implemented using the PyTorch library (Paszke et al. 2019). To optimize the overall network, we employ the Adam optimizer (Kingma and Ba 2017), setting the learning rate at 10^{-4}. In addition, we configure several hyperparameters to facilitate the learning process, including the parameter ξ. The sensitivity of these hyperparameters is assessed through a series of experimental studies, ensuring that our model is finely tuned for optimal performance in cross-modal retrieval tasks.

Evaluation Metrics

We employ widely adopted evaluation metrics, specifically mean Average Precision (mAP) and the precision-recall (PR) curve. The mAP is calculated by first determining the average precision (AP) for each query, followed by averaging these values across all queries to derive an overall performance measure. The PR curve illustrates the variation in precision across different recall rates, providing insight into the model's performance under various conditions.

Comparison with Baseline

We conducted experiments involving two retrieval tasks: "I→T" (image-to-text retrieval) and "T→I" (text-to-image retrieval), comparing the results across three datasets—MIRFlickr-25K, NUS-WIDE, and MS-COCO—to assess the efficacy of the proposed AGADH model. Our baseline model comprises eight advanced cross-modal hashing methods, including DJSRH(Su, Zhong, and Zhang 2019), JDSH(Liu et al. 2020), CDTH(Li et al. 2024a), UCCH(Hu et al. 2023b), HNH(Zhang et al. 2021), DCHMT(Tu et al. 2022), DNPH(Huo et al. 2024), and MITH(Liu et al. 2023b). Among them, DJSRH, JDSH, UCCH, HNH, and CDTH are unsupervised methods, and the rest are supervised methods. To ensure a fair comparison, we re-executed the code for the comparative experiments. In instances where the original papers did not provide code, we utilized the published results to maintain consistent dataset partitioning. Performance was evaluated using mAP and PR curve as evaluation criteria.

For fully paired modal data (PDR=0), the specific results are shown in Table 1. As the hash code length increases,

Task	Method	MIRFlickr-25K			NUS-WIDE			MS-COCO		
		16bits	32bits	64bits	16bits	32bits	64bits	16bits	32bits	64bits
I→T	UCCH	0.6820	0.6898	0.7074	0.5678	0.6508	0.6504	0.5489	0.5716	0.5978
	HNH	0.6224	0.6386	0.7392	0.5379	0.5487	0.5625	0.5980	0.6118	0.6206
	AGADH	**0.8615**	**0.8909**	**0.9046**	**0.7804**	**0.8241**	**0.8281**	**0.6616**	**0.7320**	**0.7689**
T→I	UCCH	0.6864	0.6870	0.7046	0.5648	0.6557	0.6452	0.5441	0.5705	0.6024
	HNH	0.5731	0.5426	0.5432	0.5301	0.4470	0.4079	0.4913	0.5045	0.6715
	AGADH	**0.8332**	**0.8497**	**0.8502**	**0.7433**	**0.7641**	**0.7690**	**0.6770**	**0.7426**	**0.7884**

Table 3: The mAP comparison results on three datasets in PDR=0.4

Task	Method	MIRFlickr-25K			NUS-WIDE			MS-COCO		
		16bits	32bits	64bits	16bits	32bits	64bits	16bits	32bits	64bits
I→T	AGADH-AF	0.8507	0.8882	0.8990	0.7779	0.8176	0.8194	0.6332	0.7319	0.7686
	AGADH-AM	0.8509	0.8872	0.9027	0.7680	0.7991	0.8238	0.6349	0.6916	0.7446
	AGADH	**0.8615**	**0.8909**	**0.9046**	**0.7804**	**0.8241**	**0.8281**	**0.6616**	**0.7320**	**0.7689**
T→I	AGADH-AF	0.8305	0.8391	0.8461	0.7225	0.7568	0.7282	0.6460	0.7269	0.7862
	AGADH-AM	0.7644	0.8101	0.8124	0.6881	0.7216	0.7541	0.6132	0.6961	0.7533
	AGADH	**0.8332**	**0.8497**	**0.8502**	**0.7433**	**0.7641**	**0.7690**	**0.6770**	**0.7426**	**0.7884**

Table 4: Ablation study on the three datasets in PDR=0.4

our method consistently outperforms other baseline models on all three datasets. In particular, on MIRFlickr-25K, our method significantly outperforms the best baseline method across all three hash code lengths. Specifically, for "I→T", performance improves by 6.5%, 8%, and 8.2%, respectively; and for "I→T", performance improves by 4.7%, 5.5%, and 5%, respectively. This shows that our method effectively enhances semantic information, thereby improving hash code retrieval efficiency. Figure 2 illustrates the PR curves for the three datasets with hash code lengths of 32 and 64 bits when PDR=0. In most cases, our proposed method outperforms the other methods. Notably, for datasets characterized by lower complexity, our method demonstrates a clear advantage over competing approaches. This demonstrates that our approach can more fully capture the fine-grained semantic correlations between modalities, thereby improving the accuracy of cross-modal retrieval.

Given the high experimental cost of testing all baseline methods under multiple modal loss combinations, and to avoid experimental redundancy, we selected two unsupervised hashing methods—UCCH and HNH—that performed best in modal integrity experiments on the three datasets as representatives for comparative analysis under modal loss. Although UCCH and HNH were originally designed for modal integrity input, to assess their robustness under modal loss scenarios, we artificially simulated different modal loss scenarios during the testing phase and evaluated and compared their performance accordingly. Table 2 and Table 3 display the mAP results for the three datasets with PDR set to 0.2 and 0.4, respectively. To ensure experimental consistency, the distribution of missing modalities for UCCH and HNH was aligned with our experimental design. The results indicate that regardless of the PDR level, our method consistently achieves superior performance, thereby affirming its robustness in handling incomplete modality data. This demonstrates that the modality missing data completion module can effectively reconstruct missing modal data and enhance inter-modal feature representation. The baseline method's performance degrades significantly with increasing missing modal ratios. Especially at high missing data ratios (PDR=0.4), our method still significantly outperforms the baseline. This illustrates that our method is capable of preserving information integrity under conditions of minimal data loss, and can reconstruct discriminative representations even when faced with severely missing data.

Ablation Study

To evaluate the contributions of the key components of the AGADH framework, we conducted ablation experiments on the MIRFlickr-25K, NUS-WIDE, and MS-COCO datasets. Table 4 shows the mAP results of different model variants under the PDR=0.4.

AGADH-AF: In this experiment, we removed the adaptive fusion module, while maintaining the remaining architecture consistent with the full AGADH. Experimental results show that, while the performance drop is small, some degradation is still observed, demonstrating that the module plays a positive role in maintaining semantic consistency and improving the quality of fused features.

AGADH-AM: In this experiment, we removed the modality alignment module, while maintaining the remaining architecture identical to AGADH. Experimental results show that removing this module leads to a significant performance drop, particularly on the semantically complex MS-COCO dataset. This illustrates the importance of preserving modality alignment and the superiority of the suggested alignment module.

The adaptive fusion module helps maintain semantic consistency and enhances the discriminative power of fused features, while the modality alignment module effectively strengthens the structural alignment of different modalities in a shared semantic space, thereby improving the over-

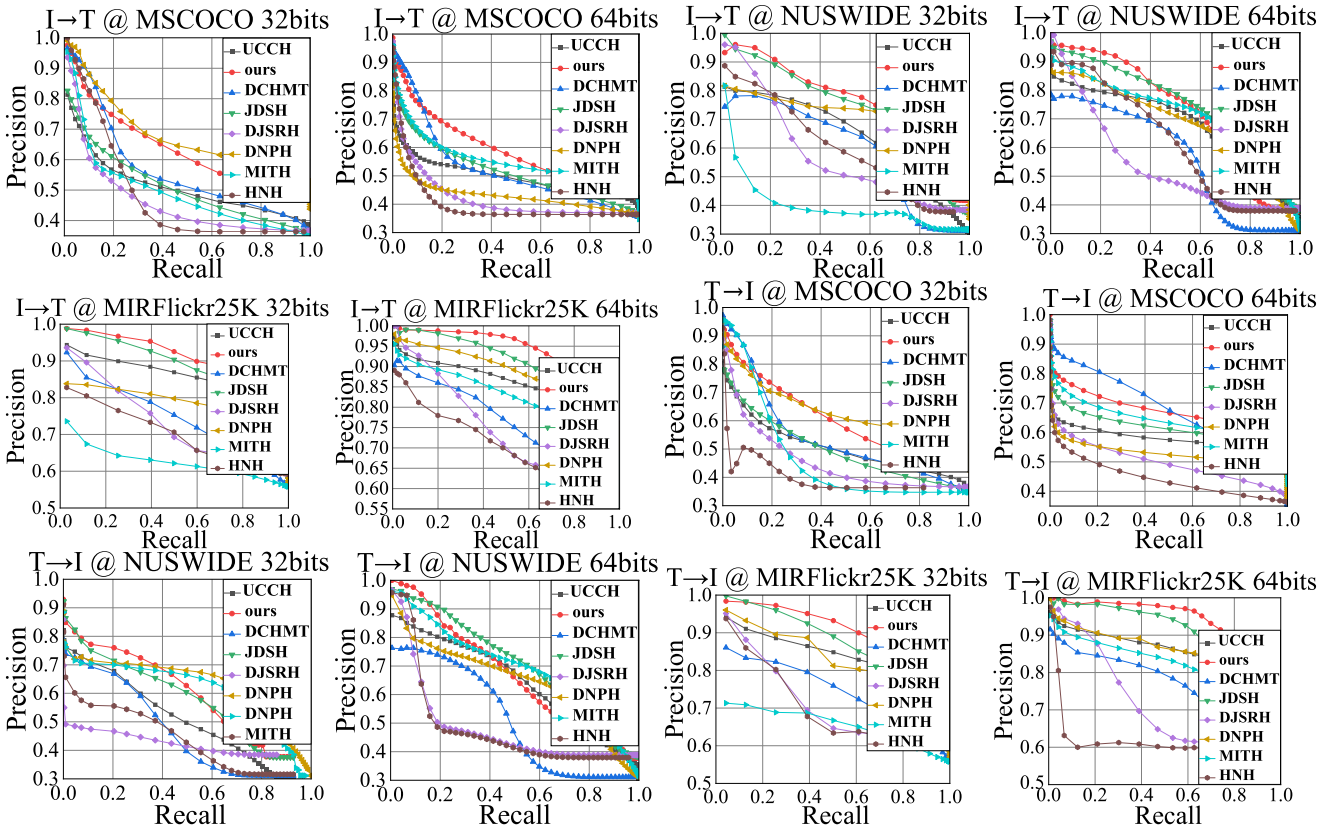

Figure 2: The PR curves for different hash code lengths on MIRFlickr-25k, NUS-WIDE, and MS-COCO datasets.

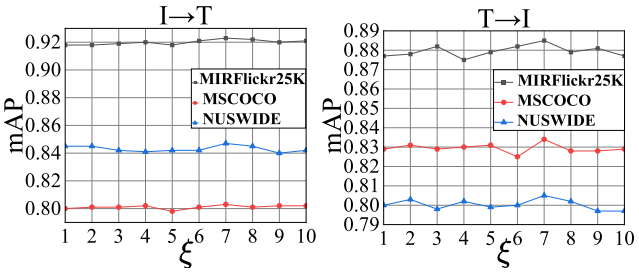

Figure 3: The influence of hyper-parameters.

all performance of cross-modal retrieval. Overall, the ablation experiments clearly demonstrate the key role of these two modules in the AGADH framework. Their synergy not only significantly enhances the semantic association between modalities but also significantly improves the robustness and accuracy of cross-modal retrieval in the presence of modality loss.

Parameter Sensitivity

We establish a series of parameters for experimental optimization, focusing on one specific parameter set for verification. We utilized a consistent 64-bit hash code to conduct the experiments on three datasets. We analyze the impact of the hyperparameter ξ on model performance, this parameter is used to adjust the structural similarity in the hash learning process. Specifically, we set the range of values of ξ from 1 to 10 with a step size of 1 and conduct independent experiments for each value. Figure 3 shows the mAP change trend of the model on three datasets under different ξ values. It can be seen that when ξ is equal to 7, the mAP achieves the best value on three datasets.

Conclusion

We propose Adaptive Graph Attention-Based Discrete Hashing (AGADH), a framework designed to address the challenges of incomplete cross-modal retrieval by reconstructing missing modalities and enhancing semantic alignment. Through the integration of a mask completion strategy, a GAT encoder-decoder, and an adaptive weight fusion module, the method effectively generates robust hash codes that maintain semantic consistency across modalities. Experimental results on benchmark datasets demonstrate our method in both fully paired and incomplete scenarios. These findings underscore the potential of adaptive graph attention mechanisms in advancing cross-modal retrieval tasks, while also opening avenues for future exploration into broader applications and dataset variations.

Acknowledgments

This work was supported by the Agricultural Scientific and Technological Achievements 496 transformation Funds of Hebei Province (2025JNZ-S24), the Joint Fund Key Program of the National Natural Science Foundation of China(No.U23B2029), and the Funda-mental Research Funds for the Central Universities (No.CUC25SG013).

References

Chen, D.; Cheng, M.; Min, C.; and Jing, L. 2020. Unsupervised Deep Imputed Hashing for Partial Cross-modal Retrieval. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. 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*, CIVR '09, 1–9. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-60558-480-5.

Hu, P.; Huang, Z.; Peng, D.; Wang, X.; and Peng, X. 2023a. Cross-modal retrieval with partially mismatched pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 9595–9610.

Hu, P.; Zhu, H.; Lin, J.; Peng, D.; Zhao, Y.-P.; and Peng, X. 2023b. Unsupervised Contrastive Cross-Modal Hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3877–3889.

Hu, Z.; Cheung, Y.-M.; Li, M.; and Lan, W. 2024. Cross-modal hashing method with properties of hamming space: A new perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 7636–7650.

Huiskes, M. J.; and Lew, M. S. 2008. The MIR flickr retrieval evaluation. MIR '08, 39–43. New York, NY, USA: Association for Computing Machinery. ISBN 9781605583129.

Huo, Y.; Qibing, Q.; Dai, J.; Zhang, W.; Huang, L.; and Wang, C. 2024. Deep Neighborhood-aware Proxy Hashing with Uniform Distribution Constraint for Cross-modal Retrieval. *ACM Trans. Multimedia Comput. Commun. Appl.*, 20(6).

Jiang, K.; Wong, W. K.; Fang, X.; Li, J.; Qin, J.; and Xie, S. 2023. Random online hashing for cross-modal retrieval. *IEEE Transactions on Neural Networks and Learning Systems*.

Kingma, D. P.; and Ba, J. 2017. Adam: A Method for Stochastic Optimization. .

Li, F.; Wang, B.; Zhu, L.; Li, J.; Zhang, Z.; and Chang, X. 2024a. Cross-Domain Transfer Hashing for Efficient Cross-Modal Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10): 9664–9677.

Li, Q.; Li, X.; Chang, Z.; Zhang, Y.; Ji, C.; and Wang, S. 2025. Multimodal knowledge retrieval-augmented iterative alignment for satellite commonsense conversation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 8168–8176.

Li, Y.; Zheng, C.; Zuo, R.; and Lu, W. 2024b. Semantic Reconstruction Guided Missing Cross-modal Hashing. In *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Liang, X.; Yang, E.; Yang, Y.; and Deng, C. 2024. Multi-relational deep hashing for cross-modal search. *IEEE Transactions on Image Processing*, 33: 3009–3020.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In Fleet, D.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision – ECCV 2014*, 740–755. Cham: Springer International Publishing. ISBN 978-3-319-10602-1.

Liu, S.; Qian, S.; Guan, Y.; Zhan, J.; and Ying, L. 2020. Joint-Modal Distribution-based Similarity Hashing for Large-scale Unsupervised Deep Cross-modal Retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, 1379–1388. New York, USA: Association for Computing Machinery. ISBN 978-1-4503-8016-4.

Liu, X.; Zeng, H.; Shi, Y.; Zhu, J.; Hsia, C.-H.; and Ma, K.-K. 2023a. Deep cross-modal hashing based on semantic consistent ranking. *IEEE Transactions on Multimedia*, 25: 9530–9542.

Liu, Y.; Wu, Q.; Zhang, Z.; Zhang, J.; and Lu, G. 2023b. Multi-Granularity Interactive Transformer Hashing for Cross-modal Retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, 893–902. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701085.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Peng, S.; Yao, T.; Li, Y.; Wang, G.; Wang, L.; and Yan, Z. 2025. Self-supervised incomplete cross-modal hashing retrieval. *Expert Systems with Applications*, 262: 125592–125603.

Shi, D.; Zhu, L.; Li, J.; Dong, G.; and Zhang, H. 2024. Incomplete cross-modal retrieval with deep correlation transfer. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(5): 1–21.

Su, S.; Zhong, Z.; and Zhang, C. 2019. Deep Joint-Semantics Reconstructing Hashing for Large-Scale Unsupervised Cross-Modal Retrieval. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3027–3035.

Tu, J.; Liu, X.; Lin, Z.; Hong, R.; and Wang, M. 2022. Differentiable Cross-modal Hashing via Multimodal Transformers. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, 453–461. New York,

NY, USA: Association for Computing Machinery. ISBN 9781450392037.

Wang, T.; Li, F.; Zhu, L.; Li, J.; Zhang, Z.; and Shen, H. T. 2025. Cross-modal retrieval: a systematic review of methods and future directions. *Proceedings of the IEEE*.

Wei, Y.; and An, J. 2024. Flexible Dual Multi-Modal Hashing for Incomplete Multi-Modal Retrieval. *International Journal of Image and Graphics*, 1–24.

Xinyu, X.; Lei, Z.; Xiushan, N.; Guohua, D.; and Huaxiang, Z. 2025. Typical Concept-Driven Modality-Missing Deep Cross-Modal Retrieval. *Journal of Computer-Aided Design & Computer Graphics*, 37(3): 519–532.

Zhang, P.-F.; Luo, Y.; Huang, Z.; Xu, X.-S.; and Song, J. 2021. High-order nonlocal hashing for unsupervised cross-modal retrieval. *World Wide Web*, 24(2): 563–583.

Zhen, L.; Hu, P.; Peng, X.; Goh, R. S. M.; and Zhou, J. T. 2020. Deep multimodal transfer learning for cross-modal retrieval. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2): 798–810.

Zhu, L.; Zheng, C.; Guan, W.; Li, J.; Yang, Y.; and Shen, H. T. 2023. Multi-modal hashing for efficient multimedia retrieval: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(1): 239–260.