Learning Together Securely: Prototype-Based Federated Multi-Modal Hashing for Safe and Efficient Multi-Modal Retrieval

Ruifan Zuo[1,2], Chaoqun Zheng[1,2*], Lei Zhu[3], Wenpeng Lu[1,2], Yuanyuan Xiang[4], Zhao Li[1,5], Xiaofeng Qu[6]

[1]Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences)
[2]Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing, Shandong Fundamental Research Center for Computer Science
[3]School of Electronic and Information Engineering, Tongji University
[4]Shandong Branch of National Computer Network Emergency Response Technical Team/Coordination Center (CNCERT/SD)
[5]Evay Info
[6]School of Information Science and Engineering, University of Jinan
{zrfan9928, leizhu0608}@gmail.com, {cqzhengwork, luwpeng, qquan_2014, ise_quxf}@163.com, liz@sdas.org

Abstract

With the proliferation of multi-modal data, safe and efficient multi-modal hashing retrieval has become a pressing research challenge, particularly due to concerns over data privacy during centralized processing. To address this, we propose *Prototype-based Federated Multi-modal Hashing* (PFMH), an innovative framework that seamlessly integrates federated learning with multi-modal hashing techniques. PFMH achieves fine-grained fusion of heterogeneous multi-modal data, enhancing retrieval accuracy while ensuring data privacy through prototype-based communication, thereby reducing communication costs and mitigating risks of data leakage. Furthermore, using a prototype completion strategy, PFMH tackles class imbalance and statistical heterogeneity in multi-modal data, improving model generalization and performance across diverse data distributions. Extensive experiments demonstrate the efficiency and effectiveness of PFMH within the federated learning framework, enabling distributed training for secure and precise multi-modal retrieval in real-world scenarios.

Code — https://github.com/vindahi/PMFH

Introduction

The geometric increase in multi-modal data has drawn considerable attention to multi-modal hashing in the field of information retrieval, owing to its advantages of efficient querying and low-cost storage (Wang et al. 2023b; He et al. 2023; Wang et al. 2023a; Shi et al. 2023). As shown in Figure 1, previous methods have primarily focused on centralized training scenarios, where data from various sources is aggregated for model training and optimization. While these methods have shown encouraging performance, they also raise prominent concerns regarding data security and privacy protection. Consequently, achieving efficient and accurate multi-modal retrieval while ensuring data privacy has become critical in this field (Wang et al. 2017; Li et al. 2023; Zhu et al. 2024; Cohen et al. 2023; Cui et al. 2024).

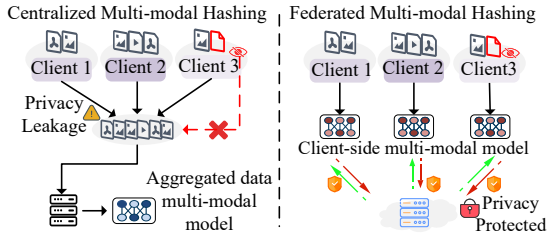


Figure 1: Centralized vs Federated multi-modal hashing.

Fortunately, federated learning (McMahan et al. 2017; Li et al. 2020; Li, He, and Song 2021; Tan et al. 2022c; Li et al. 2023) presents a compelling solution to tackle this pressing challenge. Its core advantage lies in enabling effective model collaboration across multiple participants, without the need for directly transmitting users' private data. By training models on local devices, federated learning can significantly reduce the risk of data leakage, safeguard the security of local client data, and thus provide a theoretical and practical bridge for secure multi-modal retrieval.

However, it still faces a series of urgent issues and challenges in real-world applications. 1) Although federated learning aims to ensure data security through parameter sharing rather than directly exchanging raw data, the protection of parameter transmission remains a significant challenge. Attackers still can intercept multi-modal information or even infer original data from model parameters through reverse engineering, posing a serious threat to the security of the federated learning process. Ensuring that information during the parameter transmission process is not illegally accessed has become a critical issue that needs urgent resolution. 2) The characteristics of class imbalance and statistical heterogeneity in multi-modal data pose a major challenge for federated learning. In real-world scenarios, data often exhibits an imbalanced distribution where certain classes have far more samples than others. This can lead to model training that is biased towards classes with larger sample sizes,

*Corresponding author.

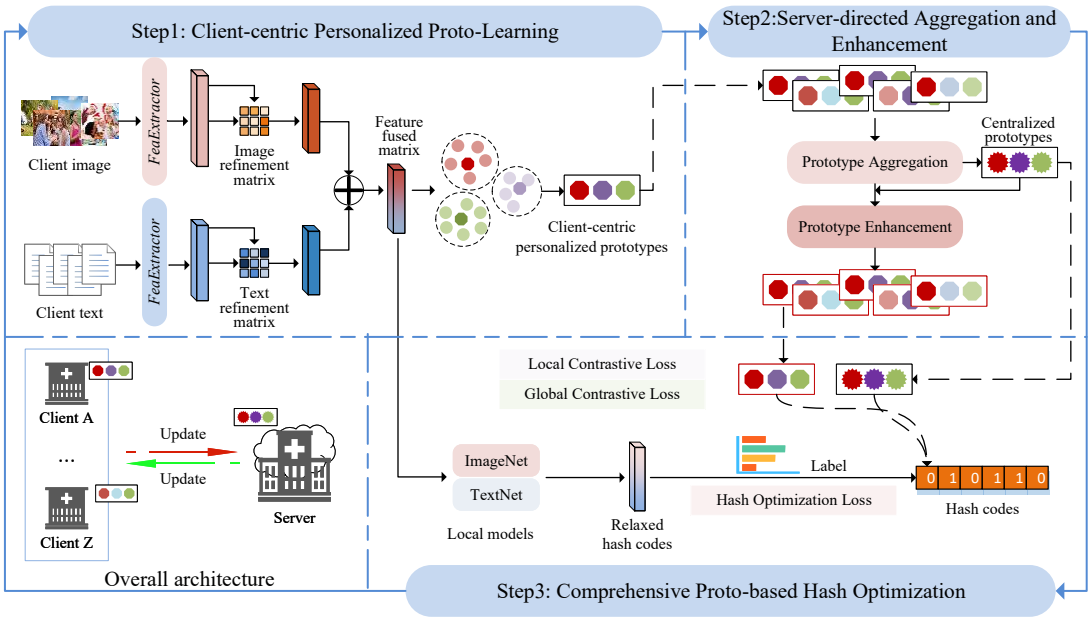


Figure 2: Overview of our proposed federated multi-modal hashing framework. Specifically, it operates efficiently within a distributed data environment, accommodating data distributed across multiple clients.

compromising the model's ability to handle diverse classes and resulting in an overall performance imbalance. Furthermore, varying data distributions across clients indicate that models trained on a limited number of clients may struggle to adapt to the data characteristics of all clients, significantly reducing model generalization and retrieval effectiveness. 3) Traditional multi-modal hashing methods often employ insufficiently robust data fusion strategies, such as cascade fusion or direct summation. While these methods have achieved some level of information integration, they may lead to crude and inadequate information fusion. This is primarily due to their failure to adequately consider the unique characteristics and inherent relationships of multi-modal data. Therefore, there is an urgent need to delve deeper into the potential value of multi-modal data fusion, as it holds the key to unlocking significant advancements and breakthroughs specifically in multi-modal hashing retrieval.

Motivated by the above analyses, we propose a *Prototype-based Federated Multi-modal Hashing* (PFMH) model for multi-modal retrieval by seamlessly integrating federated learning with multi-modal hashing techniques. Specifically, our method offers significant advantages and innovations in achieving the dual goals of privacy security and efficient retrieval, with specific contributions as follows:

- We propose a federated multi-modal hashing model that adeptly addresses privacy concerns inherent in existing centralized multi-modal data processing. Notably, it is the first work to specifically consider and address security and privacy issues in multi-modal retrieval.
- Technically, PFMH innovatively integrates fine-grained multi-modal data and harnesses representative prototypes as the medium for information exchange, adopting a dual global and local perspective. It not only reduces communication costs and mitigates data leakage risks but also addresses the common and challenging issue of data distribution imbalance in federated environments.
- Comprehensive experimental results, conducted in both distributed and centralized environments, demonstrate our model's robustness and effectiveness, while ensuring data privacy, highlighting its practical value in real-world applications requiring both efficiency and security.

Related Work

Federated Learning. Federated learning (McMahan et al. 2017; Arevalo et al. 2024) allows collaborative training across multiple clients without sharing raw data, prioritizing user privacy and data security. It can be broadly classified into two categories based on encryption mechanisms: direct and indirect encryption. Direct encryption methods (Xu et al. 2021; Zuo et al. 2024) rely on cryptographic keys, but they become vulnerable and costly as key lengths increase. Indirect encryption methods, typically including cryptography-based and prototype-based methods, facilitate data processing without direct access to raw data, albeit with high communication requirements. Specifically, prototype-based methods address the challenges in federated learning by aggregating and transmitting the essential knowledge encapsulated within client models. It is attractive due to the compact representation of prototypes, which significantly reduces communication overhead. Representative works include FedProto (Tan et al. 2022b), FedProc (Mu et al. 2023), PT-FUCH (Li et al. 2023), and FedTGP (Zhang et al. 2024).

Multi-modal Hashing. Multi-modal hashing involves mapping high-dimensional multi-modal data into compact hash codes, enabling efficient retrieval and reducing storage costs. Traditional methods (Lu et al. 2019c,a; Liu, Zhang, and Huang 2020; Lu et al. 2019b), which are trained in a

centralized environment, can be categorized as unsupervised (Liu, Yu, and Shao 2015; Shen et al. 2015, 2018; Lu et al. 2024) or supervised (Zheng et al. 2019; Zhu et al. 2020) based on whether they utilize label information. Unsupervised methods discover meaningful representations and similarities across modalities, suitable for scenarios with limited labeled data. Supervised methods rely on accurate labels to guide learning, enabling the model to learn more discriminative hash codes for better class distinction. With the continuous advancement and reinforcement of deep network models, deep multi-modal hashing models (Tan et al. 2022a, 2023; Shen et al. 2023; Tu et al. 2024), leveraging deep learning for feature extraction, effective fusion, and mapping into hash codes, have gradually gained more attention. Deep models are highly sensitive to the quality of training data and may encounter privacy and security challenges when dealing with sensitive information.

Some multimedia retrieval works have incorporated federated learning to address data privacy concerns, but these methods primarily focus on cross-modal retrieval (Liu et al. 2023; Li et al. 2023; Zuo et al. 2024), *i.e.*, using query data from one modality to retrieve data from another modality. In contrast, multi-modal retrieval is more flexible and comprehensive, as it can leverage the complementarity among multiple modalities to learn useful information from limited data. This advantage is particularly pronounced in scenarios involving privacy-sensitive data.

Methodology

Our primary objective is to perform multi-modal retrieval in a distributed setting, eliminating reliance on traditional centralized data storage systems. To achieve this, we consider a federated learning scenario that consists of one central server and M local clients. In this setup, the central server acts as a coordinator, responsible for aggregating and optimizing the global model by collecting and integrating local model knowledge from each client. Meanwhile, each client is tasked with executing local multi-modal hashing retrieval tasks and independently possesses a unique multi-modal private dataset $\mathcal{O}_i = \{(x_{ij}, y_{ij}, l_{ij})_{j=1}^{m_i}\}$, where x_{ij} and y_{ij} are the image and text of the j-th sample, and l_{ij} is its true label.

Proposed Framework

As shown in Figure 2, the training procedure of our proposed model encompasses an iterative training framework executed across the server and clients, consisting of the following three key components:

⋆ *Step 1: Client-centric Personalized Proto-Learning*: Each client independently integrates multi-modal data, learns the client-specific prototype captures fine-grained semantic nuances, and leverages the prototype as a mediator for semantic transmission with the server.

⋆ *Step 2: Server-directed Aggregation and Enhancement*: The server aggregates all client-centric personalized prototypes and further enhances these prototypes with rich global knowledge, fostering a holistic understanding and enriching the semantic representations.

⋆ *Step 3: Comprehensive Proto-based Hash Optimization*: Each client receives both the centralized prototype and the enhanced client prototype from the server, deepening the understanding of the prototype semantics and enabling robust local hash learning.

Client-centric Personalized Proto-Learning

Since each modality exhibits unique characteristics, we initially capture distinct features from each modality using a tailored feature extraction network. Subsequently, these extracted features are mapped to a unified k-dimensional latent space, which can be mathematically expressed as:

$$\boldsymbol{f}_* = \textit{FeaExtractor}_*(*; \theta_{1*}) \in \mathbb{R}^{k \times m_i}, * \in \{x, y\}, \quad (1)$$

where m_i represents the number of samples in the multi-modal dataset $\mathcal{O}_i$ of the i-th client, and θ_{1*} represents the training parameters specific to the feature extraction network for either the image (x) or text (y) modality.

Following this, to further enhance the representation of the extracted multi-modal features, we learn a feature refinement matrix $\mathbf{C}_*$ for each modality, designed to adaptively remove redundant and noisy information while emphasizing more critical features. Formally, the refinement matrix is:

$$\mathbf{C}_* = \textit{FeaRefiner}(\boldsymbol{f}_*; \theta_{2*}) \in \mathbb{R}^{k \times m_i}, * \in \{x, y\}, \quad (2)$$

where θ_{2*} denotes the training parameters specific to each modality. The significance of $\mathbf{C}_*$ lies in its ability to dynamically enable the model to focus on the most discriminative features for each modality. Taking advantage of this fact, we then obtain more fine-grained features $\boldsymbol{f}_*^{'}$ by applying $\mathbf{C}_*$ through dot product operation:

$$\boldsymbol{f}_*^{'} = \boldsymbol{f}_* \odot \mathbf{C}_* \in \mathbb{R}^{k \times m_i}, * \in \{x, y\}. \quad (3)$$

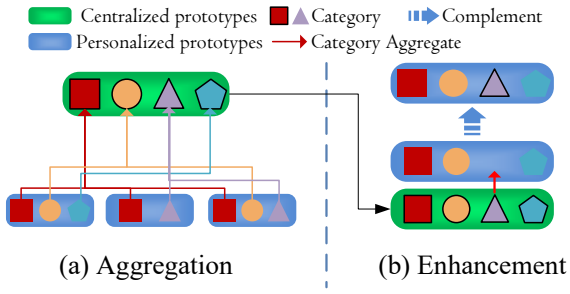
Based on that, to intelligently integrate complementary information from different modalities, we integrate the refined multi-modal features to obtain the feature fusion matrix $\mathbf{H}$, which is formulated as follows:

$$\mathbf{H} = \textit{Integrate}(\boldsymbol{f}_x^{'} \oplus \boldsymbol{f}_y^{'}; \theta_{3*}), \quad (4)$$

where $\oplus$ is the concatenation operation, and θ_{3*} is the training parameters. This fusion process is crucial for synthesizing disparate yet complementary data from various sources.

To securely and effectively facilitate communication and knowledge exchange among clients in federated learning, we calculate the client-centric personalized prototypes $\mathbf{P}_i = \{\mathbf{P}_i^1, \mathbf{P}_i^2, ..., \mathbf{P}_i^j\}$ for each client i, which serve as vital information carriers for interaction between the server and clients. Specifically, each prototype $\mathbf{P}_i^j$ represents the class prototype, that is, the mean of the features of samples belonging to the j-th class. After calculating each prototype, we normalize it to have a unit norm so that the learned prototypes maintain a consistent scale, allowing for easy comparison and combination across client and server. Formally, this process uses the following formula:

$$\mathbf{P}_i^j = \frac{\sum_{i=1}^{m_i} \mathbf{H}_i \mathbb{1}\left(l_{ij} = 1\right)}{\sum_{i=1}^{m_i} \mathbb{1}\left(l_{ij} = 1\right)}, \quad \text{then} \quad \mathbf{P}_i^j \leftarrow \frac{\mathbf{P}_i^j}{||\mathbf{P}_i^j||_2}, \quad (5)$$



(a) Aggregation (b) Enhancement

Figure 3: Server-directed aggregation and enhancement.

where m_i is the number of samples on the i-th client, and $\mathbb{1}$ is an indicator function used to determine whether a sample belongs to the j-th class based on its label l_{ij}. In this process, it is worth emphasizing that by utilizing class prototypes as representative embeddings, we can more accurately capture the unique data attributes from different clients, thereby facilitating the realization of personalized learning.

Server-directed Aggregation and Enhancement

In the preceding learning process, each client computes the personalized prototype that effectively captures the essential features of its local dataset. Subsequently, as illustrated in Figure 3, the central server orchestrates the aggregation of these prototype sets from all clients, enabling the consolidation of knowledge and facilitating the sharing of sample feature information across the entire network.

To achieve the aforementioned objectives, for each category j, we compute the centralized prototype $\bar{\mathbf{P}}^j$ by:

$$\bar{\mathbf{P}}^j = \frac{\sum_i^{\mathcal{M}_j} \mathbf{P}_i^j \mathbb{1}\left(l_{ij}=1\right)}{\sum_i^{\mathcal{M}_j} \mathbb{1}\left(l_{ij}=1\right)}, \quad \text{then} \quad \bar{\mathbf{P}}^j \leftarrow \frac{\bar{\mathbf{P}}^j}{||\bar{\mathbf{P}}^j||_2}, \tag{6}$$

where $\bar{\mathbf{P}}^j$ represents the centralized prototype for the j-th category that incorporates knowledge from all clients. The set $\mathcal{M}_j$ denotes the clients that contain samples from the j-th category. The indicator function $\mathbb{1}\left(l_{ij}=1\right)$ ensures that only clients with samples from the j-th class contribute to the computation of the centralized prototype for that class. The final set of centralized prototypes, $\bar{\mathbf{P}} = \{\bar{\mathbf{P}}^1, \bar{\mathbf{P}}^2, ..., \bar{\mathbf{P}}^j\}$, encapsulates consolidated knowledge and serves as a valuable resource for sharing sample feature information across the entire network.

Importantly, an often overlooked fact is that in reality, the categories in each client are not balanced, that is, some clients may be missing certain categories. This imbalance poses a significant disadvantage as it can lead to biased models that perform poorly on categories that are underrepresented or completely missing in the local data of some clients. Such models may struggle to generalize well to new, unseen data, especially if it belongs to categories that were scarce or absent during training. Additionally, this imbalance can hinder the effectiveness of collaborative learning scenarios, where clients are expected to contribute knowledge about all categories to improve the global model.

This observation inspires us to propose a simple yet effective prototype enhancement strategy. As depicted in Figure 3, the server undertakes a systematic comparison between the learned centralized prototypes and the client-specific personalized prototypes. Through this comparison, the server identifies any missing category prototypes and fills them in using the following formula:

$$\mathbf{P}_i^j = \begin{cases} \mathbf{P}_i^j, & i \in \mathcal{M}_j \\ \bar{\mathbf{P}}^j, & i \notin \mathcal{M}_j \end{cases} \tag{7}$$

This process ensures that each client's prototype set is expanded to include prototype information for all categories. It not only balances the distribution of categories but also enhances the model's learning and generalization capabilities across all global categories. More specifically, the advantage of using the centralized prototype is that it provides a higher level of semantic expression and knowledge-sharing within the federated learning system. Furthermore, leveraging the centralized prototypes to augment client-specific prototypes enables the model to comprehend the global data diversity and complexity, ultimately enhancing its generalization ability and performance across various data distributions.

Comprehensive Proto-based Hash Optimization

Upon receiving the globally aggregated prototypes and the refined client-specific prototypes from the server, our immediate task is to empower clients with the ability to leverage both the shared global knowledge across clients and the unique characteristics inherent in each client's data. This, in turn, will ultimately enable effective multi-modal hash learning on the client side. With this idea, we design a HashEncoder that incorporates three key loss functions: the global contrastive loss ($\mathcal{L}_{global}$), the local contrastive loss ($\mathcal{L}_{local}$) and the hash optimization loss ($\mathcal{L}_{hash}$). These loss functions work together to generate hash codes $\mathbf{B}_i$ from the fused multi-modal features $\mathbf{H}$. The process is:

$$\mathbf{B}_i = \textit{HashEncoder}(\mathbf{H}; \theta_{4*}), \tag{8}$$

where θ_{4*} is the parameter.

Specifically, the global contrastive loss is designed to promote the alignment of features with the global centers, ensuring that the learned representations are consistent with the overall data distribution. Formally, it is expressed as:

$$\mathcal{L}_{global} = -\sum_{z=1}^{|\bar{\mathbf{P}}|} \log \frac{\exp\left(\mathbf{H} \cdot \bar{\mathbf{P}}^c / \tau\right)}{\sum_{j \neq c} \exp\left(\mathbf{H} \cdot \bar{\mathbf{P}}^j / \tau\right)}, \tag{9}$$

where τ is the temperature parameter controlling the sharpness of the softmax function, and the dot product measures the similarity between the features and the global centers.

The local contrastive loss facilitates the alignment of features with each client's local prototypes, ensuring that the unique characteristics of each client's data are preserved:

$$\mathcal{L}_{local} = -\sum_{(x,y)\in\mathcal{O}_i} \frac{1}{m_i} \sum_{z=1}^{m_i} \log \frac{\exp\left(\mathbf{H} \cdot \mathbf{P}_z^i / \tau\right)}{\sum_{i \neq j} \exp\left(\mathbf{H} \cdot \mathbf{P}_z^j / \tau\right)}, \tag{10}$$

where $\mathcal{O}_i$ is the multi-modal dataset of the i-th client.

To enhance the hash codes' quality and discriminative power, we design the hash optimization loss $\mathcal{L}_{hash}$ to integrate similarity matching and information preservation.

Algorithm 1: Prototype-based Federated Multi-modal Hashing

Input: Training set $\mathcal{O}_i = \{(x_{ij}, y_{ij}, l_{ij})_{j=1}^{m_i}\}$ for each client i, communication rounds T, client numbers M, local epoch E, hyper-parameters α, β.
Output: Multi-modal retrieval modal for each client.

```
Initialize Local Prototypes P_i for each client i
for Round t = 0, 1, 2, ..., T − 1 do
    Client-centric Personalized Proto-Learning:
    for each client i = 1, 2, ..., M do
        P_i ← UpdateClient(i, P_i)
    end for
    Server-directed Aggregation and Enhancement:
    P̄ ← Aggregate(P_1, P_2, ..., P_M)
    for each client i = 1, 2, ..., M do
        P_i ← Enhance(P_i, P̄)
    end for
    Comprehensive Proto-based Hash Optimization:
    Client receive P_i and P̄ from the server
    for each epoch k = 1, 2, ..., E do
        for each batch in batches do
            Compute overall loss L using P_i and P̄
            Update model parameters to minimize L
        end for
        P_i ← AggregateLocal(current batch)
    end for
end for
```

Similarity matching encourages the hash codes to preserve the semantic similarity between input samples, while information preservation ensures that the hash codes capture the essential information from the input features.

$$\mathcal{L}_{hash} = \underbrace{\alpha\| \cos(\boldsymbol{b}_i, \boldsymbol{b}_j) - \mathbf{s}_{ij}\|_2^2}_{\textit{Similarity Matching}} + \underbrace{\beta\|\mathbf{B}_i - \boldsymbol{b}_i\|_2^2}_{\textit{Information Preservation}}, \quad (11)$$

where α and β are hyper-parameters. $\mathbf{s}_{ij}$ is the similarity affinity distances between different samples, which guides the similarity matching process. Specifically, $\boldsymbol{b}_i$ denotes the relaxed hash representation, obtained from the semantic features $\mathbf{H}$ via a hashing layer. This relaxed hash representation allows for a smooth and gradual optimization process. Notably, $\mathbf{B}_i$, the strictly binary hash codes, are obtained from $\boldsymbol{b}_i$ through the sign function: $\mathbf{B}_i = sign(\boldsymbol{b}_i) \in \{-1, 1\}^{1\times k}$.

Finally, by jointly considering the losses presented in Eqs. (9), (10), and (11), we can subsequently obtain the final loss function as follows:

$$\min \mathcal{L} = \mathcal{L}_{global} + \mathcal{L}_{local} + \mathcal{L}_{hash}. \quad (12)$$

Main algorithm flow is outlined in Algorithm 1 for clarity.

Out of Sample Extension

After several rounds of iterative optimization across client and server, each multi-modal retrieval model on individual clients has been meticulously trained, yielding optimal network parameters. Given a target query sample ($\mathbf{x}_q$, $\mathbf{y}_q$), the trained model on each client can generate the corresponding binary hash codes ($\mathbf{B}_q$). This process can be succinctly summarized as:

$$\begin{aligned}\mathbf{x}_q, \mathbf{y}_q \xrightarrow{\text{input}} &\textit{FeaExtractor}(\mathbf{x}_q, \mathbf{y}_q) \rightarrow \textit{FeaRefiner}(\boldsymbol{f}_x, \boldsymbol{f}_y) \\ &\rightarrow \textit{Integrate}(\boldsymbol{f}_x^{'}, \boldsymbol{f}_y^{'}) \rightarrow \textit{HashEncoder}(\mathbf{H}) \xrightarrow{\text{output}} \mathbf{B}_q\end{aligned} \quad (13)$$

Experiments

Experimental Setups

Dataset. We perform experiments on three datasets. Wikipedia (Rasiwasia et al. 2010) comprises 2,866 image-text pairs from the top 10 categories, split into 2,173 for training and 693 for querying. Given its limited size, the retrieval set is the same as the training set. MIR-Flickr (Huiskes and Lew 2008) includes 20,015 image-text pairs from 24 categories, with 2,243 unique samples for querying, 17,772 for retrieval, and 5,000 randomly chosen for training. NUS-WIDE (Chua et al. 2009) has 195,834 instances from the top 21 categories, with 2,085 for querying, 193,749 for retrieval, and 21,000 randomly selected for training.

Evaluation Metric. We use Mean Average Precision (MAP) and Top-K precision curve to evaluate retrieval performance. MAP serves as a comprehensive metric, assessing overall retrieval effectiveness through the weighted averaging of precision and recall rates. Top-K precision curve measures the accuracy of the top search results, emphasizing the system's practical ability to present relevant items in the retrieval process.

Baselines. There is no existing work on federated multi-modal hashing. To validate our proposed federated multi-modal hashing model, we compare it with three advanced federated learning methods: FedAvg (McMahan et al. 2017), FedProto (Tan et al. 2022b), and PT-FUCH (Li et al. 2023). We utilize these three federated learning frameworks as the core to independently train our proposed multi-modal hashing model on each client with private data, while ensuring that the multi-modal hashing model architecture on each client remains unchanged. This allows us to compare the performance of our proposed model when trained with different federated learning frameworks. Furthermore, to provide a comprehensive evaluation, we select two advanced multi-modal hashing methods, BSTH (Tan et al. 2022a) and GCIMH (Shen et al. 2023), for comparison, to demonstrate the effectiveness and competitiveness of our model, particularly highlighting its advantages in retrieval performance.

Implementation Details. In our model, the *FeaExtractor* network extracts and maps features from various modalities into a common 512-dimensional space, consisting of an input layer, hidden layers, and an output layer for feature transformation and integration. Specifically, we use Dropout layers to enhance generalization. Additionally, the modality-specific *FeaRefiner* network employs a sequence of linear layers with nonlinear activations to learn optimal feature fusion weights. To simulate a federated learning environment, we use 10 clients and create a non-IID data scenario using the Dirichlet distribution. We conduct 15 communication rounds for three datasets, optimizing with Adam (learning rate $1e^{-5}$, weight decay $1e^{-6}$). The hyper-parameters are

Methods	Wikipedia				MIR-Flickr				NUS-WIDE			
	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
BSTH	0.6302	0.6719	0.6813	0.6847	0.8145	0.8340	0.8482	0.8571	0.6321	0.6578	0.6696	0.6783
GCIMH	0.6280	0.6317	0.6552	0.6693	0.8332	0.8358	0.8508	0.8529	0.6768	0.6842	0.6955	0.7012
CentralizedOur	0.6750	0.6862	0.6947	0.7083	0.8421	0.8553	0.8662	0.8711	0.7222	0.7478	0.7649	0.7777
FedAvg	0.7059	0.7555	0.7750	0.7971	0.8450	0.8552	0.8649	0.8789	0.6517	0.6528	0.6775	0.7009
FedProto	0.7150	0.7651	0.7952	0.8105	0.8467	0.8538	0.8658	0.8776	0.6513	0.6574	0.6831	0.7008
PT-FUCH	0.7232	0.7666	0.7956	0.8112	0.8486	0.8591	0.8690	0.8795	0.6515	0.6622	0.6867	0.7058
Ours	**0.7435**	**0.7687**	**0.7998**	**0.8126**	**0.8640**	**0.8669**	**0.8778**	**0.8819**	**0.7426**	**0.7663**	**0.7903**	**0.8091**

Table 1: MAP results on Wikipedia, MIR-Flickr and NUS-WIDE. The best result in each column is marked in bold.

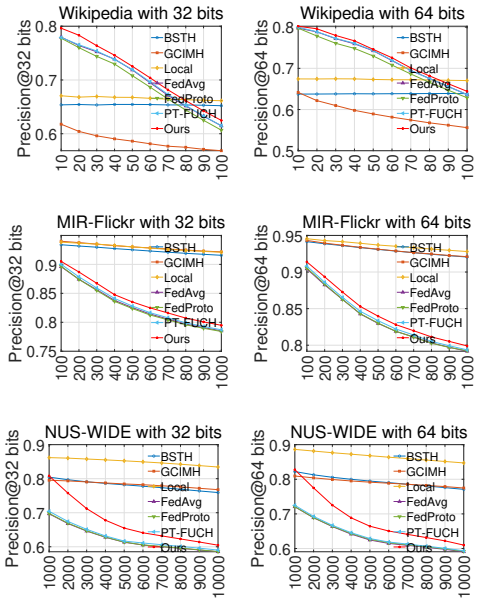


Figure 4: Top-K precision curves.

set to $\alpha = 1e^{-2}$ and $\beta = 1e^{-3}$. We use a batch size of 64 and set the temperature parameter τ to 0.07.

Federated Performance Comparison

We first compare PFMH against advanced federated learning methods in the federated environment. The experimental results, as shown in Table 1, indicate that PFMH outperforms these advanced federated learning methods, achieving commendable retrieval accuracy improvements on three datasets. Meanwhile, Figure 4 visually represents the Top-K retrieval performance of PFMH and baseline methods. Notably, Top-K precision curves of PFMH consistently outperform the baselines, indicating high retrieval accuracy even for larger K. Behind this excellent performance lie our utilization of prototypes for information exchange and two key innovative factors: 1) Our prototype learning strategy deeply mines and fuses rich semantic information to create a highly recognizable and comprehensive prototype, enhancing the model's delicate and thorough analysis in multi-modal information retrieval. 2) Our prototype enhancement strategy effectively strengthens the model's semantic information during training, reinforces the prototype-category connection, and fills semantic gaps, ensuring accuracy and generalization in complex multi-modal data processing.

Centralized Performance Comparison

In this section, we introduce an additional experimental setup for centralized training as a supplementary measure. This setup simulates the traditional multi-modal hashing method environment where all data is centralized and pre-aggregated for model training. Specifically, we formulate a *CentralizedOur* variant for comparison with existing multi-modal hashing methods. Notably, this variant involves aggregating multi-modal data and inputting it into a model with the same architecture as the client models in this paper, followed by training to ensure a fair evaluation. The results in Table 1 demonstrate that *CentralizedOur* outperforms state-of-the-art methods, underscoring the strength of our model architecture in effectively leveraging multi-modal data in a non-federated manner. However, it is important to highlight that while *CentralizedOur* exhibits certain advantages, our federated multi-modal hashing model, PFMH, even surpasses it in overall performance. This may be attributed to the fact that data from different clients often exhibit unique characteristics and biases, and a model trained in a federated manner is inherently more likely to generalize well to unseen query data due to its exposure to this diversity. Moreover, by training on distributed data without aggregating it centrally, PFMH may be less prone to overfitting the specific data distribution of any single aggregated set.

Ablation Experiments

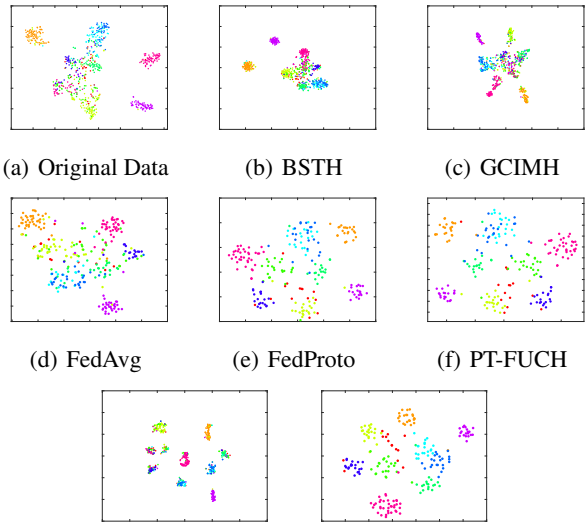
To comprehensively assess the impact of individual modules, we design four variants: Ours-w/o-*P* (removing local/global prototype contrast), Ours-w/o-*PE* (removing prototype enhancement), Ours-w/o-$\mathcal{L}_s$ (excluding similarity matching), and Ours-w/o-$\mathcal{L}_l$ (excluding information preservation). Results are shown in Table 2. The experimental results of variants Ours-w/o-*P* and Ours-w/o-*PE* indicate that both local and global prototypes significantly impact our model. The performance drop observed when the global enhancement module is removed highlights its crucial role in the model. Additionally, the results of variants Ours-w/o-$\mathcal{L}_s$ and Ours-w/o-$\mathcal{L}_l$ demonstrate the importance of similarity matching and information preservation in achieving optimal results in the local multi-modal hash learning process.

Visualization

To visualize and analyze binary hash codes, we apply t-SNE to the single-label Wikipedia dataset. This mapping helps show data structure, keeping similar points close. We compare our model's t-SNE results with BSTH and GCIMH

Methods	Wikipedia				MIR-Flickr				NUS-WIDE			
	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
Ours-w/o-P	0.7332	0.7571	0.7783	0.7906	0.8571	0.8611	0.8734	0.8796	0.7324	0.7517	0.7821	0.7949
Ours-w/o-PE	0.7406	0.7655	0.7984	0.8103	0.8621	0.8625	0.8754	0.8791	0.7411	0.7649	0.7887	0.8052
Ours-w/o-$\mathcal{L}_s$	0.7312	0.7434	0.7749	0.7866	0.8479	0.8538	0.8616	0.8789	0.7266	0.7465	0.7794	0.7850
Ours-w/o-$\mathcal{L}_l$	0.7413	0.7526	0.7800	0.8017	0.8612	0.8635	0.8726	0.8799	0.7401	0.7579	0.7897	0.7963
Ours	**0.7435**	**0.7687**	**0.7998**	**0.8126**	**0.8640**	**0.8669**	**0.8778**	**0.8819**	**0.7426**	**0.7663**	**0.7903**	**0.8091**

Table 2: Ablation study results on Wikipedia, MIR-Flickr and NUS-WIDE. The best result in each column is marked in bold.



(a) Original Data (b) BSTH (c) GCIMH

(d) FedAvg (e) FedProto (f) PT-FUCH

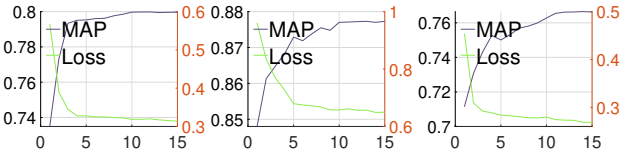
(g) Ours (centralized) (h) Ours (federated)

Figure 5: t-SNE visualization with 64-bit hash code length.

in a centralized environment. The results shown in Figures 5 (a-c,g) demonstrate that, in a centralized environment, our model achieves precise aggregation of data from different categories, forming more compact and distinct category clusters compared to the original data and existing methods. Furthermore, we also compare the t-SNE results of our model with those generated by three federated learning baselines in a federated environment, as illustrated in Figures 5 (d-f,h). To ensure fairness, we select a fixed client from all clients under the IID data scenario as the experimental subject. The results indicate that our method demonstrates significant technical advantages in achieving precision and clarity of category clusters.

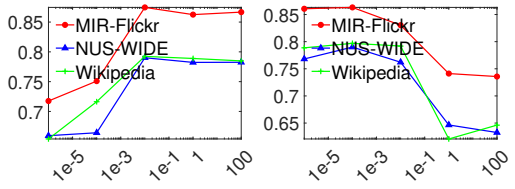
Communication Rounds and Convergence Analysis

To investigate our model's convergence, particularly the relationship between performance and communication efficiency, we conduct comprehensive communication round and convergence analysis experiments on three datasets. The experimental results in Figure 6 demonstrate a significant improvement in model performance as the number of communication rounds increases, ultimately achieving convergence. From the perspective of client-server interaction within the federated learning framework, our model can effectively converge and achieve optimal performance even with a limited number of communication rounds.



(a) Wikipedia (b) MIR-Flickr (c) NUS-WIDE

Figure 6: Communication rounds results.



(a) α (b) β

Figure 7: Parameter sensitivity results.

Parameter Sensitivity Analysis

To investigate the impact of hyper-parameters on our model performance, we conduct a comprehensive parameter sensitivity analysis across three datasets, with a focus on the 64-bit hash code length. As shown in Figure 7, our findings unequivocally reveal that even slight adjustments to the hyper-parameters α and β can result in substantial variations in experimental outcomes within a specific range. This observation not only underscores the crucial role of parameter adjustment in optimizing model performance but also demonstrates the notable sensitivity and adaptability of PFMH to variations in hyper-parameters.

Conclusion

Existing multi-modal hashing methods exhibit notable deficiencies in addressing both the effectiveness and security of multi-modal retrieval, motivating our proposal of an innovative *Prototype-based Federated Multi-modal Hashing* (PFMH) model. This multi-modal hashing model seamlessly combines federated learning, aiming to facilitate secure and efficient multi-modal data retrieval in distributed environments. By leveraging prototypes as the primary conduit for information exchange, PFMH effectively addresses data privacy concerns, as well as the intricacies of class imbalance and statistical heterogeneity inherent in multi-modal datasets. Our extensive experiments have convincingly validated the practicality and robustness of PFMH, particularly its superior performance in retrieval accuracy, communication efficiency, and privacy preservation.

Acknowledgments

This work is supported by the Natural Science Foundation of Shandong (ZR2024QF054), Qilu University of Technology (Shandong Academy of Sciences) Science, Talent Research Project (2023RCKY142), Program of Innovation Improvement for Small and Medium-sized Enterprises of Shandong (2024TSGC0039), National Natural Science Foundation of China (62376130), and Program of New Twenty Policies for Universities of Jinan (202333008), the Pilot Project for Integrated Innovation of Science, Education, and Industry of Qilu University of Technology (Shandong Academy of Sciences) (2024ZDZX08).

References

Arevalo, C. A.; Noorbakhsh, S. L.; Dong, Y.; Hong, Y.; and Wang, B. 2024. Task-Agnostic Privacy-Preserving Representation Learning for Federated Learning against Attribute Inference Attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, 10909–10917.

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)*, 1–9.

Cohen, E.; Nelson, J.; Sarlós, T.; and Stemmer, U. 2023. Tricking the hashing trick: A tight lower bound on the robustness of countsketch to adaptive inputs. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 37, 7235–7243.

Cui, H.; Zhao, L.; Li, F.; Zhu, L.; Han, X.; and Li, J. 2024. Effective Comparative Prototype Hashing for Unsupervised Domain Adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, 8329–8337.

He, L.; Huang, Z.; Chen, E.; Liu, Q.; Tong, S.; Wang, H.; Lian, D.; and Wang, S. 2023. An efficient and robust semantic hashing framework for similar text search. *ACM Transactions on Information Systems*, 41(4): 1–31.

Huiskes, M. J.; and Lew, M. S. 2008. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval (ACM MM)*, 39–43.

Li, J.; Li, F.; Zhu, L.; Cui, H.; and Li, J. 2023. Prototype-guided knowledge transfer for federated unsupervised cross-modal hashing. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, 1013–1022.

Li, Q.; He, B.; and Song, D. 2021. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 10713–10722.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems (MLSys)*, 2: 429–450.

Liu, J.; Zhan, Y.-W.; Luo, X.; Chen, Z.-D.; Wang, Y.; and Xu, X.-S. 2023. Prototype-Based Layered Federated Cross-Modal Hashing. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–2. IEEE.

Liu, L.; Yu, M.; and Shao, L. 2015. Multiview alignment hashing for efficient image search. *IEEE Transactions on image processing*, 24(3): 956–966.

Liu, L.; Zhang, Z.; and Huang, Z. 2020. Flexible discrete multi-view hashing with collective latent feature learning. *Neural Processing Letters*, 52: 1765–1791.

Lu, X.; Liu, L.; Ning, L.; Zhang, L.; Mu, S.; and Zhang, H. 2024. Multi-Facet Weighted Asymmetric Multi-Modal Hashing Based on Latent Semantic Distribution. *IEEE Transactions on Multimedia*.

Lu, X.; Zhu, L.; Cheng, Z.; Li, J.; Nie, X.; and Zhang, H. 2019a. Flexible online multi-modal hashing for large-scale multimedia retrieval. In *Proceedings of the 27th ACM international conference on multimedia (ACM MM)*, 1129–1137.

Lu, X.; Zhu, L.; Cheng, Z.; Nie, L.; and Zhang, H. 2019b. Online multi-modal hashing with dynamic query-adaption. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval (SIGIR)*, 715–724.

Lu, X.; Zhu, L.; Li, J.; Zhang, H.; and Shen, H. T. 2019c. Efficient supervised discrete multi-view hashing for large-scale multimedia search. *IEEE Transactions on Multimedia*, 22(8): 2048–2060.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics (AISTATS)*, 1273–1282. PMLR.

Mu, X.; Shen, Y.; Cheng, K.; Geng, X.; Fu, J.; Zhang, T.; and Zhang, Z. 2023. Fedproc: Prototypical contrastive federated learning on non-iid data. *Future Generation Computer Systems*, 143: 93–104.

Rasiwasia, N.; Costa Pereira, J.; Coviello, E.; Doyle, G.; Lanckriet, G. R.; Levy, R.; and Vasconcelos, N. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia (ACM MM)*, 251–260.

Shen, X.; Chen, Y.; Pan, S.; Liu, W.; and Zheng, Y. 2023. Graph Convolutional Incomplete Multi-modal Hashing. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, 7029–7037.

Shen, X.; Shen, F.; Liu, L.; Yuan, Y.-H.; Liu, W.; and Sun, Q.-S. 2018. Multiview discrete hashing for scalable multimedia search. *ACM Transactions on Intelligent Systems and Technology.*, 9(5): 1–21.

Shen, X.; Shen, F.; Sun, Q.-S.; and Yuan, Y.-H. 2015. Multi-view latent hashing for efficient multimedia search. In *Proceedings of the 23rd ACM international conference on Multimedia. (ACM MM)*, 831–834.

Shi, D.; Zhu, L.; Li, J.; Zhang, Z.; and Chang, X. 2023. Unsupervised adaptive feature selection with binary hashing. *IEEE Transactions on Image Processing*, 32: 838–853.

Tan, W.; Zhu, L.; Guan, W.; Li, J.; and Cheng, Z. 2022a. Bit-aware semantic transformer hashing for multi-modal retrieval. In *Proceedings of the 45th International ACM SIGIR*

Conference on Research and Development in Information Retrieval (SIGIR), 982–991.

Tan, W.; Zhu, L.; Li, J.; Zhang, Z.; and Zhang, H. 2023. Partial multi-modal hashing via neighbor-aware completion learning. *IEEE Transactions on Multimedia*, 25.

Tan, Y.; Long, G.; Liu, L.; Zhou, T.; Lu, Q.; Jiang, J.; and Zhang, C. 2022b. Fedproto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, 8432–8440.

Tan, Y.; Long, G.; Ma, J.; Liu, L.; Zhou, T.; and Jiang, J. 2022c. Federated learning from pre-trained models: A contrastive learning approach. *Advances in neural information processing systems (NeurIPS)*, 35: 19332–19344.

Tu, R.-C.; Mao, X.-L.; Liu, J.; Wei, W.; Huang, H.; et al. 2024. Similarity Transitivity Broken-Aware Multi-Modal Hashing. *IEEE Transactions on Knowledge and Data Engineering*.

Wang, H.; Sun, J.; Wei, X.; Zhang, S.; Chen, C.; Hua, X.-S.; and Luo, X. 2023a. Dance: Learning a domain adaptive framework for deep hashing. In *Proceedings of the ACM Web Conference 2023 (WWW)*, 3319–3330.

Wang, J.; Zhang, T.; Sebe, N.; Shen, H. T.; et al. 2017. A survey on learning to hash. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4): 769–790.

Wang, L.; Pan, Y.; Liu, C.; Lai, H.; Yin, J.; and Liu, Y. 2023b. Deep hashing with minimal-distance-separated hash centers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 23455–23464.

Xu, R.; Baracaldo, N.; Zhou, Y.; Anwar, A.; Joshi, J.; and Ludwig, H. 2021. Fedv: Privacy-preserving federated learning over vertically partitioned data. In *Proceedings of the 14th ACM workshop on artificial intelligence and security (AISec)*, 181–192.

Zhang, J.; Liu, Y.; Hua, Y.; and Cao, J. 2024. Fedtgp: Trainable global prototypes with adaptive-margin-enhanced contrastive learning for data and model heterogeneity in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, 16768–16776.

Zheng, C.; Zhu, L.; Lu, X.; Li, J.; Cheng, Z.; and Zhang, H. 2019. Fast discrete collaborative multi-modal hashing for large-scale multimedia retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 32(11): 2171–2184.

Zhu, L.; Lu, X.; Cheng, Z.; Li, J.; and Zhang, H. 2020. Deep collaborative multi-view hashing for large-scale image search. *IEEE Transactions on Image Processing*, 29: 4643–4655.

Zhu, L.; Zheng, C.; Guan, W.; Li, J.; Yang, Y.; and Shen, H. T. 2024. Multi-Modal Hashing for Efficient Multimedia Retrieval: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(1): 239–260.

Zuo, R.; Zheng, C.; Li, F.; Zhu, L.; and Zhang, Z. 2024. Privacy-Enhanced Prototype-based Federated Cross-modal Hashing for Cross-modal Retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*.