A Federated Learning Approach For Cross-Institutional Intangible Cultural Heritage Multimedia Content Retrieval And Sensitive Information Protection In The Digital Cultural And Creative Industry

Qiku Bao

Jilin Animation Institute, 130000, Jilin, China

Corresponding author. E-mail: baoqiku1@163.com

Received July 21, 2026; accepted August 20, 2026

In the digital cultural and creative industry, ICH multimedia resources (images, videos, audio, text, metadata) are scattered across cultural institutions and often involve sensitive information (identity, location, craft knowledge, ritual scenes), so centralized retrieval cannot balance sharing, efficiency, and privacy. This study proposes a federated learning framework combining local multimodal encoding, federated contrastive learning for cross-modal alignment, gradient pruning/DP/secure aggregation against inference risks, and retrieval combining prototype routing, local indexing, sensitivity assessment, and result anonymization for privacy-protected Top-K retrieval. On 31,500 cross-institutional non-IID ICH samples, the method achieves Recall@5 of 0.807, mAP of 0.752, and NDCG@10 of 0.793 (beating FedAvg, FedProx, FedCMR, FedCMR-DP), maintains Recall@5 of 0.738 at non-IID degree 0.9, reduces sensitive-information exposure from 18.6% to 4.1%, lowers member/attribute-inference success to 38.6%/33.7%, and achieves 126 ms latency and 472 queries/s throughput at 20 nodes.

Keywords: federated learning; intangible cultural heritage; multimedia retrieval; sensitive information protection; distributed index; scalable information system; differential privacy

http://dx.doi.org/10.6180/jase.202612_35.020

1. Introduction

Digital preservation of intangible cultural heritage is now central to public cultural service systems, with imaging, video, audio, and annotation technologies shifting ICH resources from text archives to multimodal resources. Research shows multimodal processing is central to cultural resource management [1–3].

In practice, ICH multimedia resources are scattered across institutions: cultural centers hold interview videos, museums hold tool/craft images, university labs hold annotated samples, archives hold application materials. Institutions are generally unwilling or unable to centralize raw data due to ownership, privacy, and regulatory concerns [4].

Traditional centralized retrieval requires uploading all data to a central node, failing cross-institutional privacy requirements. Cross-modal and hash retrieval research shows unified semantic space and indexing are key to performance [5–7], but relies on centralized training without protecting institutional silos.

Federated learning offers a feasible approach for cross-institutional collaboration, achieving collaborative modeling via parameter/gradient aggregation without uploading raw data [8], though remaining susceptible to model inversion, membership inference, and attribute inference, requiring secure aggregation, differential privacy, and robust optimization [9, 10].

This study proposes a cross-institutional ICH retrieval framework based on federated learning: each institution locally completes multimodal encoding and sensitive infor-

mation identification [6, 7], while the central server only performs secure aggregation and routing updates. Contrastive learning achieves cross-modal alignment; retrieval combines prototype routing with local Top-K search filtered by sensitivity and permissions.

ICH multimedia retrieval requires images, videos, audio, texts, and metadata to map into a unified semantic space so users can retrieve craft videos from a tool image or project text from folk-song audio. Multimodal federated learning supports this without centralizing data [11–13], but existing research inadequately addresses sensitive entities and query-stage desensitization for ICH.

ICH multimedia sensitivity combines identity, location, knowledge, and context: identity-sensitive (name, face, voice) [14]; location-sensitive (addresses, sacrificial sites); knowledge-sensitive (process steps, recipe ratios); and context-sensitive (rituals, restricted performances). Table 1 shows risk by data type.

2. Methods

2.1. System Model and Problem Definition

2.1.1. Cross-institutional ICH Multimedia Resource Scenario

This study considers a cross-institutional collaborative retrieval system composed of multiple ICH resource preservation institutions. Let there be K participating institutions in the system, denoted as

$$\mathcal{C} = \{C_1, C_2, \ldots, C_K\} \tag{1}$$

Where C_k represents the k institutional node, which can correspond to a local cultural center, museum, archive, university laboratory, intangible cultural heritage protection center, or industry association. Each institutional node stores a local intangible cultural heritage multimedia dataset.

$$D_k = \{(x_i^v, x_i^a, x_i^t, m_i, y_i, s_i)\}_{i=1}^{n_k} \tag{2}$$

Local data includes visual data (images, video frames, key scenes), audio data (folk songs, operas, interviews), text data (project introductions, lineage records, metadata descriptions), structured metadata, category tags, and sensitivity markers; the total data volume across the system is the sum across institutions.

$$N = \sum_{k=1}^{K} n_k \tag{3}$$

In this scenario, institutional data cannot be centralized; the central server only aggregates updates and manages routing [9]. A query is encoded, candidate institutions located via the global prototype index, and candidates perform local Top-K search, returning anonymized results by permission and sensitivity level.

2.1.2. Intangible Cultural Heritage Multimedia Data and Sensitive Information Types

Sensitivity spans four types: identity-sensitive (inheritor name, face, voice, contact, address) [14]; location-sensitive (precise addresses, sacrificial sites, activity routes, coordinates); knowledge-sensitive (undisclosed process steps, recipe ratios, tool details, apprenticeship rules); and context-sensitive (religious rituals, ethnic-group activities, restricted performances) — detailed in

Resource sensitivity is classified into four levels for differentiated protection: Level 0 (fully public, displayed normally); Level 1 (generally sensitive, precise fields hidden); Level 2 (identity/location-sensitive, requiring de-identification, obfuscation, and access verification); and Level 3 (highly sensitive knowledge, returning only summary/index or controlled-access prompts).

2.1.3. Cross-Institutional Multimedia Retrieval Task

Definition Given a user query q, which can be an image, text, audio, or a multimodal combination, the system needs to return the semantically most relevant Top-K results from the local indexes of all institutions without centralizing the original data. The query encoder maps q to a unified vector:

$$z_q = f_\theta(q) \tag{4}$$

The multimodal representation of the i resource in organization C_k is as follows:

$$z_i = f_\theta\left(x_i^v, x_i^a, x_i^t, m_i\right) \tag{5}$$

The basic similarity is calculated using the cosine similarity as follows:

$$sim\left(z_q, z_i\right) = \frac{z_q^\top z_i}{\| z_q \|_2 \| z_i \|_2} \tag{6}$$

After considering sensitive risks, the final search score was defined as follows:

$$Score\left(q, z_i\right) = sim\left(z_q, z_i\right) - \gamma \rho_i \tag{7}$$

Where ρ_i represents the sensitivity risk level of resource i, and γ represents the sensitivity penalty weight. When user permissions are insufficient, even if a highly sensitive resource has a high semantic similarity, the original content will not be returned directly.

Table 1. Cross-Institutional Intangible Cultural Heritage Multimedia Data Types and Sensitive Information Risks

Data types	Specific content	Main search value	Potential sensitive risks
Image data	Tools, clothing, craft scenes, and photos of inheritors	Visual similarity retrieval and category recognition	Face, location, undisclosed process details
Video data	Performance process, production process, and ceremonial activities	Action process retrieval, scene retrieval	Ritual scenes, identity exposure, and process leaks
Audio data	Folk songs, operas, oral interviews, and explanations of crafts	Voiceprint, melody, and semantic retrieval	Personal voice, interview privacy, local information
Text data	Project introduction, lineage, and metadata description	Semantic retrieval, tag completion	Name, phone number, address, ethnic information
Metadata	Region, Year, Project Category, Organization Code	Cross-organizational routing, index filtering	Precise geographical location, internal organization number

2.1.4. *Optimization Objective*

This study aims to optimize retrieval performance, privacy protection, and system overhead while ensuring that cross-institutional data are not centrally shared. The overall objective function is defined as

$$\min_{\theta} \sum_{k=1}^{K} \frac{n_k}{N} \mathcal{L}_k(\theta) + \lambda_1 \mathcal{R}_p + \lambda_2 \mathcal{R}_c \tag{8}$$

This work's contribution is threefold: a retrieval-oriented federated framework keeping resources inside institutional nodes; federated contrastive learning coupled with sensitive-information identification; and query-stage prototype routing and anonymized results extending privacy protection to retrieval delivery.

The retrieval loss is the primary target, with privacy risk and communication cost as regularization constraints: privacy appears via gradient clipping, DP perturbation, and secure aggregation, while communication appears via parameter compression and candidate-institution routing — preventing the framework from optimizing accuracy alone.

2.2. Cross-Institutional Multimedia Retrieval Framework Based on Federated Learning

2.2.1. *Overall Architecture*

The system has six layers (Figure 1): institutional data (local ICH resources) [9]; local multimodal encoding (unified semantic space); federated optimization (training, pruning, DP, secure aggregation); distributed retrieval index (local indexes plus global prototypes); sensitive information protection (entity ID, sensitivity assessment, permissions); and user query. Institutions never upload originals; the central node only receives protected updates, enabling horizontal scaling.

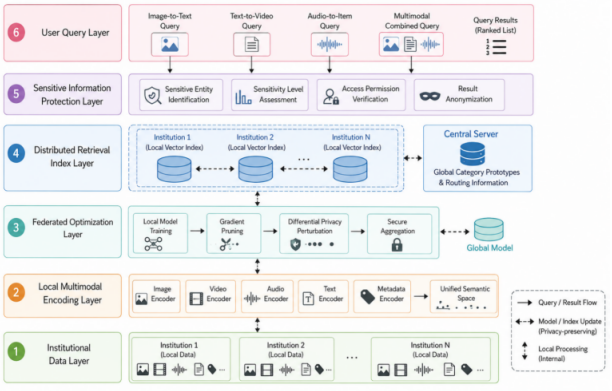

Fig. 1. Overall Architecture of the Cross-Institutional Intangible Cultural Heritage Multimedia Federated Retrieval System

2.2.2. *Local Multimodal Feature Encoding*

ICH multimedia resources show strong modal heterogeneity: images reflect tools and scenes; videos reflect performance and craft processes; audio reflects singing and dialects; text provides labels; metadata provides region and era. This study uses a multi-branch encoder with vision transformers for the visual branch.

$$z_i^v = f_v\left(x_i^v; \theta_v\right) \tag{9}$$

The audio branch first converts the audio to a log-Mel spectrum and then obtains the acoustic representation through an audio encoder:

$$z_i^a = f_a\left(x_i^a; \theta_a\right) \tag{10}$$

Text branches use a pretrained language model to extract semantic representations.

$$z_i^t = f_t\left(x_i^t; \theta_t\right) \tag{11}$$

Metadata branches map category, region, year, and organization number to structured embeddings:

$$z_i^m = f_m\left(m_i; \theta_m\right) \tag{12}$$

Multimodal fusion employs splicing and linear mapping.

$$z_i = W_f\left[z_i^v; z_i^a; z_i^t; z_i^m\right] + b_f \tag{13}$$

To avoid representation failure when a modality is absent, this study uses random modality-dropping during training, letting the model retain retrieval capability even when images, audio, or text are missing.z_i

2.2.3. *Federated Contrastive Learning Mechanism*

Cross-modal retrieval requires semantic correspondence, not simple classification. This study uses federated contrastive learning, building positive pairs from the same ICH project's images/videos/audio/text and negative pairs from different projects/categories.

$$L_{con} = -log\frac{e^{\frac{s_i^p}{\tau}}}{\sum_{j=1}^{B} e^{\frac{s_{ij}}{\tau}}}, \quad s_{ij} = sim\left(z_i, z_j\right) \tag{14}$$

This loss aligns images, videos, audio, and text in the same semantic space. The local training loss has three parts: contrastive, classification, and sensitive-identification losses.

$$\mathcal{L}_k = \mathcal{L}_{con} + \alpha\mathcal{L}_{cls} + \beta\mathcal{L}_{sen} \tag{15}$$

The classification task maintains category discrimination, the sensitive-identification task helps the model recognize potentially sensitive resources at retrieval time, and the contrastive task improves cross-modal retrieval quality.

2.2.4. *Privacy-Preserving Federated Aggregation*

Although ordinary federated learning does not upload the original data, gradient updates may still contain private information. This study implements three steps to protect local gradients before uploading. The first step is gradient clipping.

$$\bar{g}_k = \frac{g_k}{\max\left(1, \| g_k \|_2 / C\right)} \tag{16}$$

Where g_k represents the local gradient of the k institution, and C is the pruning threshold. This operation limits the update magnitude of a single institution, thereby reducing the risk of anomalous updates and privacy leaks. The second step is the differential privacy noise injection.

$$\widetilde{g}_k = \bar{g}_k + \mathcal{N}\left(0, \sigma^2 C^2 I\right) \tag{17}$$

Here, σ represents the noise level. A larger σ can enhance privacy protection but may also reduce model convergence quality; therefore, a suitable balance needs to be found experimentally. The third step is the secure aggregation. The central server does not directly analyze updates from individual institutions but aggregates the protected updates as follows:

$$\theta^{t+1} = \theta^t - \eta\sum_{k=1}^{K}\frac{n_k}{N}\widetilde{g}_k \tag{18}$$

This process reduces the risk of individual institution parameters being back-analyzed and maintains global model training stability under multi-institutional participation.

In implementation, privacy-preserving aggregation runs before each server upload: clients clip updates, inject Gaussian perturbation, and submit only masked/aggregated updates, reducing the chance a malicious server can infer whether a specific record participated in training.

2.2.5. *Distributed Indexing and Retrieval Routing*

To avoid a full search across all institutions, this study uses a two-stage "global prototype routing + local vector retrieval" mechanism, with each institution maintaining a local vector index.

$$I_k = \{z_i, id_i, \rho_i, A_i\} \tag{19}$$

Where id_i is the resource ID, ρ_i is the sensitivity level, and A_i is the access permission requirement. The central server only maintains the category prototype as follows:

$$p_c = \frac{1}{|D_c|}\sum_{i:y_i=c} z_i \tag{20}$$

For a query q, the system computes similarity to the global prototype, selects candidate categories and institutions, then performs local Top-K search among candidates — avoiding brute-force search and reducing latency while improving throughput.

2.3. Sensitive Information Protection Mechanism

2.3.1. *Sensitive Information Classification and Protection Strategy*

Protecting sensitive information related to intangible cultural heritage requires the distinction between resource types and access scenarios [10]. This study classifies resources into four levels and configures differentiated protection strategies for different levels, as shown in

Table 2. Sensitive Information Classification and Protection Strategy

Sensitivity Level	Information type	Example	Protection strategy
Level 0	Public Information	Intangible cultural heritage category, publicly disclosed project name	Normal display
Level 1	Generally Sensitive Information	District/county level areas, activity time	Generalization
Level 2	Identity and Location Sensitive Information	Name of the inheritor, village location, and contact information	Desensitization, obfuscation, and permission verification
Level 3	Highly Sensitive Knowledge Information	Undisclosed formula, ceremony details, and internal files	Returning to the original text is prohibited; only a summary or a controlled access warning is displayed.

2.3.2. *Sensitive Entity Recognition in Text*

Text metadata may leak the inheritor's name, phone, address, village name, ritual name, or undisclosed process steps. A sensitive entity recognition model annotates text content to predict its sensitive category.

$$P\left(s_i|x_i^t\right) = softmax\left(W_s h_i + b_s\right) \tag{21}$$

For Level 1 resources, the system retains category/region information but hides exact addresses; for Level 2, it performs name replacement, location obfuscation, and access verification; for Level 3, it returns only a summary or authorization prompt.

2.3.3. *Visual and Audio Sensitive Content Recognition*

Sensitive visual/video information mainly includes faces, landmarks, ritual scenes, craft details, and unauthorized recordings. This study adds a sensitive-area detection module [12, 13]: faces/identity markers are obfuscated; undisclosed craft details get reduced resolution or text summary only; restricted ritual scenes are withheld from insufficiently authorized users. Sensitive audio (voices, interviews, place/personal names, craft details) is first transcribed and entity-recognized, retaining sensitivity markers to avoid returning unauthorized full audio.

2.3.4. *Privacy Filtering During the Query Phase*

Privacy protection during the training phase cannot replace the protection of results during the query phase. This study defines a query result filtering function as follows:

$$R^{'(q)} = Filter\left(R(q), A_u, \rho_i\right) \tag{22}$$

If the user's permission is lower than the resource's sensitivity level, the system returns a de-identified summary, low-resolution preview, or authorization prompt instead of original content — preventing highly-ranked sensitive resources from direct exposure.

2.4. Algorithm Implementation and Complexity Analysis

2.4.1. *Federated Training Algorithm*

The algorithm has five stages: global initialization; local training (contrastive, classification, and sensitive-recognition losses); privacy-preserving updates (pruning and DP); secure aggregation; and index synchronization (Table 3).

2.4.2. *Complexity Analysis*

With average samples per institution n, model size d, communication rounds T, and local rounds E, federated training computational complexity is approximately O(T·K·E·n·d).

$$O(KTEn) \tag{23}$$

The communication complexity is mainly determined by uploading and downloading the model parameters:

$$O\left(KT|\theta|\right) \tag{24}$$

If Top-k parameter compression or low-rank updates are used, the communication complexity can be reduced to

$$O\left(KT\kappa|\theta|\right) \tag{25}$$

Where κ represents the compression ratio. During the retrieval phase, each institution's local index employs an approximate nearest-neighbor search with a complexity of approximately

$$O\left(\log N_k + r\right) \tag{26}$$

Table 3. Privacy-Preserving Federated Intangible Cultural Heritage Multimedia Retrieval Algorithm

Step	Operation	Description
1	Global Initialization	Initialize the global multimodal retrieval model
2	Client Selection	Institutional nodes were selected to participate in the training.
3	Local Encoding	Extract visual, audio, text, and metadata features.
4	Local Training	Optimize contrastive, classification, and sensitive-identification losses.
5	Gradient Pruning	Limiting the gradient norm reduces the effects of anomalous updates.
6	DP Perturbation	Add Gaussian noise to local updates.
7	Secure Aggregation	Aggregate protected model parameters without exposing individual institutional updates.
8	Model Distribution	Distribute the global model to all institutions
9	Index Update	Update local vector search index
10	Query Filtering	Return the anonymized Top-K search results

Table 4. Cross-institutional ICH Multimedia Simulation Data Partitioning

Institutional Nodes	Data scale	Main modes	Main categories	Non-IID features
Institution 1	8,000	Images, text	Traditional crafts and traditional arts	High proportion of visual data
Institution 2	6,500	Video, text	Traditional dances and folk activities	High proportion of video data
Institution 3	5,000	Audio, text	Traditional music and folk arts	High proportion of audio data
Institution 4	7,200	Image, audio, text	Traditional drama and folk literature	Modal equilibrium
Institution 5	4,800	Video, metadata	Folklore and traditional medicine	There is a lack of metadata

Because global prototype routing filters candidate institutions first, the system avoids a full query across all institutions, so overall query overhead increases slowly with institution count.

2.5. Experimental Design

2.5.1. Dataset Construction and Cross-Institution Partitioning

Because authentic cross-institutional ICH data face ownership restrictions, this study organizes public ICH resources with non-IID partitioning and manual sensitive-field annotation [15], spanning eight categories totaling 31,500 items: 12,400 images, 7,600 video clips, 5,200 audio, and 6,300 text/metadata resources (Table 4).

The simulated split creates category imbalance (different dominant categories per institution) and modality imbalance (varying resource-type proportions), with sensitive labels annotated by risk type — approximating real cultural-resource governance.

2.5.2. Comparison Methods and Evaluation Metrics

Six comparison methods were used: Local Retrieval, Centralized Retrieval (upper-bound, not privacy-preserving), FedAvg, FedProx, FedCMR (standard federated cross-modal retrieval), FedCMR-DP, and the proposed method, evaluated on retrieval performance, privacy protection, system overhead, and scalability (Table 5).

All methods use the same training rounds, embedding dimension, retrieval depth, and evaluation set for fair comparison. Privacy metrics are computed on both returned retrieval results and model-update attack simulations, reflecting service-stage exposure and training-stage leakage risk.

2.5.3. Experimental Parameter Settings

Communication rounds were set to 100, local rounds to 5, batch size 64, learning rate 0.001, embedding dimension 512, temperature 0.07, sensitivity loss weight 0.3, classification loss weight 0.5, gradient clipping threshold 1.0, and default DP noise scale 0.5. Retrieval defaults to Top-10 results with the top three candidate institutions from prototype routing.

Inference attacks use shadow-model simulations: membership inference determines whether a target sample was in a client's training set from prediction confidence and update behavior; attribute inference attempts to recover sensitive identity, location, or knowledge labels from learned

Table 5. Experimental Evaluation Metrics

Indicator Type	Index	Meaning
Search performance	Recall@1, Recall@5, Recall@10, mAP, NDCG@10	Measuring search accuracy and ranking quality
Privacy protection	Sensitive Exposure Rate, Privacy Leakage Rate, Attack Success Rate	Measuring the risk of sensitive information leakage
System overhead	Communication Cost, Training Time, Index Update Time	Measuring training and synchronization costs
Scalability	Query Latency, Throughput, Scalability Ratio	Measuring multi-agency scalability
Robustness	Performance under Non-IID, Missing Modality Ratio	Measuring the ability to adapt to heterogeneous data

representations. Reported success rates are averaged across institutions and categories.

3. Results and discussion

3.1. Comparison of Retrieval Performance of Different Algorithms

Figure 2 compares Recall@1/5/10, mAP, and NDCG@10: Local Retrieval reaches only 0.681 Recall@5; FedAvg improves to 0.724; FedProx reaches 0.746; FedCMR reaches 0.781 via contrastive learning; FedCMR-DP drops slightly from DP noise; the proposed method achieves 0.807 Recall@5, 0.752 mAP, and 0.793 NDCG@10, ahead of all federated baselines.

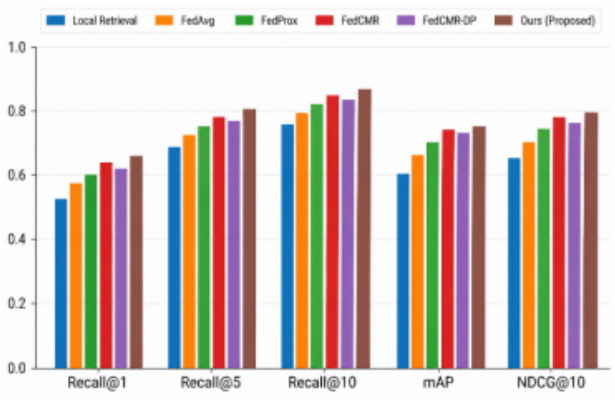

Fig. 2. Performance Comparison of Cross-Institutional Intangible Cultural Heritage Multimedia Retrieval Using Different Methods

This near-centralized performance without centralizing data stems from two factors: federated contrastive learning enhances semantic alignment across images, videos, audio, and text, and the sensitive-identification auxiliary task keeps the model focused on discriminative semantic regions rather than over-relying on sensitive fields.

3.2. Federated Training Convergence Performance Analysis

Figure 3 shows Recall@5 convergence after 100 rounds: FedAvg improves slowly and fluctuates later from non-IID sensitivity; FedProx is more stable but lower than contrastive methods; FedCMR improves rapidly in the first 60 rounds; the proposed method stabilizes after ~70 rounds at Recall@5 0.807 with minimal fluctuation.

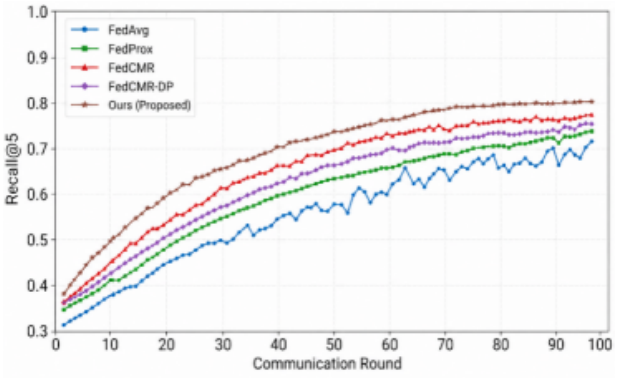

Fig. 3. Convergence Curves of Different Federated Retrieval Methods

The sensitive-identification auxiliary task provides additional semantic constraints, letting institutions form a relatively consistent global feature space despite different modal distributions.

3.3. Impact of Non-IID Degree on Retrieval Performance

Figure 4 shows Recall@5 across non-IID degrees 0.1–0.9: FedAvg drops most (0.758→0.631), FedProx to 0.668, Fed-

CMR to 0.701, while the proposed method still achieves 0.738 even at degree 0.9.

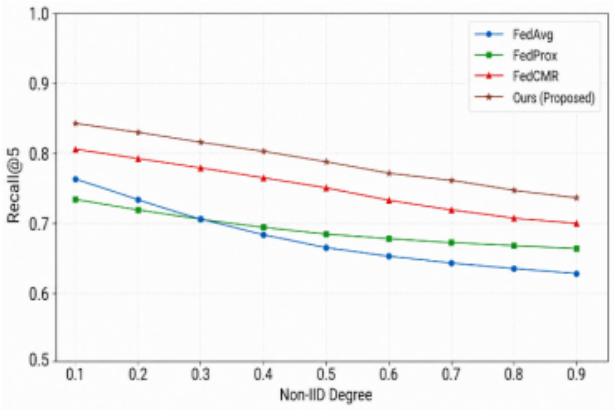

Fig. 4. Retrieval performance variation under different levels of Non-IID

This matches ICH data characteristics, where institutions show resource bias (music audio, craft images, or ritual videos dominating different collections). Ordinary federated learning is vulnerable to this bias; the proposed method mitigates it via cross-modal contrastive learning and global prototype alignment.

3.4. Comparison of Sensitive Information Exposure Rates

Figure 5 compares sensitive-information exposure: Centralized Retrieval is highest at 18.6%; FedAvg/FedProx lack query-stage filtering, exposing 12.4%/11.2%; FedCMR reduces to 9.7%; FedCMR-DP reduces to 6.2%; the proposed method combines sensitivity levels, access control, and anonymization to reach 4.1%.

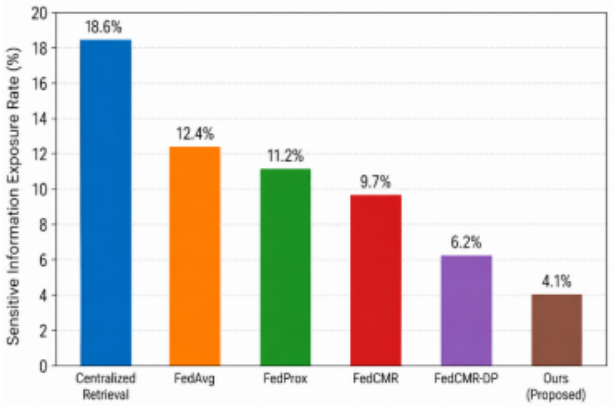

Fig. 5. Sensitive information exposure rates under different retrieval and protection methods

Federated training alone cannot prevent sensitive leakage; effective ICH protection must cover training, indexing, and querying — especially the query-return stage.

3.5. Analysis of Resistance to Inference Attacks

Figure 6 shows attack success rates: FedAvg has 65.2% membership-inference and 58.9% attribute-inference success; FedCMR-DP reduces membership-inference to 45.8%; the proposed method further reduces it to 38.6% and attribute-inference to 33.7% via secure aggregation and sensitive-field constraints.

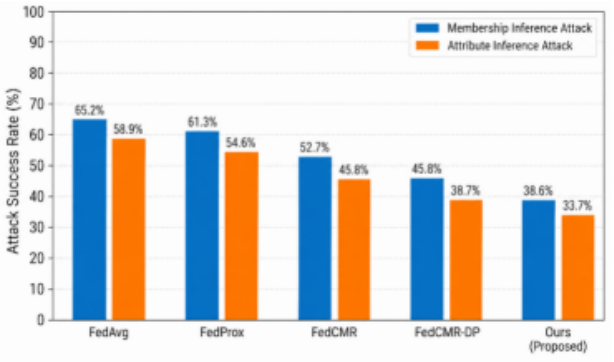

Fig. 6. Membership and attribute inference attack success rates under different methods

Differential privacy perturbation and secure aggregation can effectively reduce the risk of privacy leakage during model updates, whereas the sensitive information identification branch can reduce the model's over-reliance on sensitive attributes such as identity and location.

3.6. Impact of Differential Privacy Noise Intensity

Figure 7 shows DP noise scale σ's effect: at σ=0, Recall@5 is highest but leakage risk is highest; as σ increases, leakage decreases but Recall@5 declines; at σ=0.5, Recall@5 is 0.807 with leakage 0.041 (a good balance); at σ=1.0, leakage decreases further but Recall@5 drops to 0.772.

More noise is not necessarily better protection: excessive noise disrupts the cross-modal embedding structure, preventing similar-resource clustering, so noise intensity must be dynamically tuned to privacy level and retrieval scenario.

3.7. Communication Cost Analysis

Figure 8 shows communication cost as institutions grow from 3 to 20: FedAvg is lowest but with insufficient performance; FedCMR is higher from its contrastive branch; the proposed method stays lower than uncompressed FedCMR via compression, growing roughly linearly from 2.5 to 9.6 GB per round.

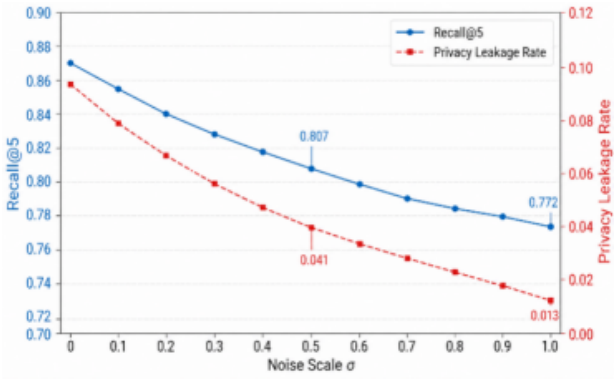

Fig. 7. Impact of Differential Privacy Noise on Retrieval Performance and Privacy Leakage Rate

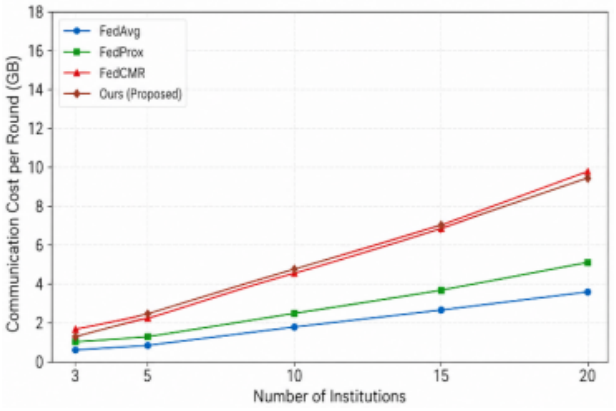

Fig. 8. Comparison of Communication Costs with Different Numbers of Institutions

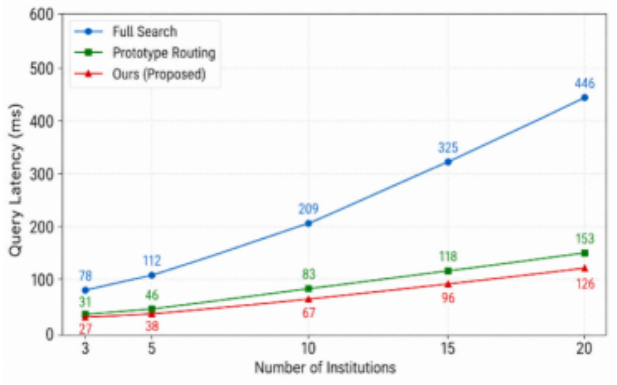

Fig. 9. Query latency comparison under different numbers of institutions

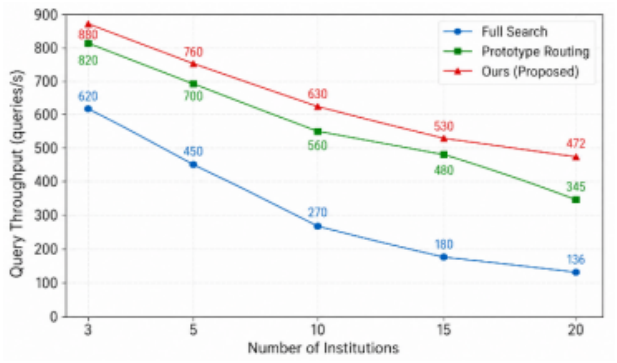

Fig. 10. Query throughput comparison under different numbers of institutions

3.8. Query Latency and Throughput Analysis

Figure 9 shows query latency: full search reaches 446 ms at 20 institutions; prototype routing reduces this to 153 ms; the proposed method, combining sensitivity pre-filtering with local approximate nearest-neighbor indexing, achieves 126 ms at 20 institutions.

3.9. Ablation Experiment

Figure 11 shows ablation results: removing federated contrastive learning drops Recall@5 from 0.807 to 0.761; removing differential privacy slightly improves Recall@5 but significantly raises exposure rate; removing sensitive filtering raises exposure from 4.1% to 15.9%; removing distributed routing raises latency from 78 to 139 ms — confirming each module's distinct necessity.

This shows the framework is not a simple combination of independent modules: contrastive learning contributes semantic alignment, DP/secure aggregation reduce update leakage, sensitive filtering protects retrieval results, and prototype routing controls overhead — removing any module causes distinct degradation, supporting unified training-, indexing-, and query-stage protection.

3.10. Feature Distribution Visualization

Figure 12 shows t-SNE visualization: traditional music, dance, drama, folk art, crafts, customs, medicine, and fine arts form relatively clear clusters, confirming effective cross-modal semantic learning.

3.11. Discussion

This method transforms ICH multimedia sharing into a cross-institutional scalable information-system problem: institutions have data ownership and privacy boundaries making centralized sharing impractical, and federated learning enables collaborative training without uploading raw data.

Results show near-centralized performance, significantly outperforming ordinary federated learning, confirming federated contrastive learning suits ICH multimedia retrieval given natural semantic connections across images, audio, video, and text describing the same cultural phe-

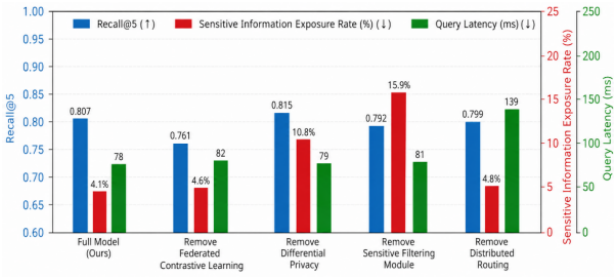

Fig. 11. Impact of key modules on Recall@5, sensitive information exposure rate, and query latency

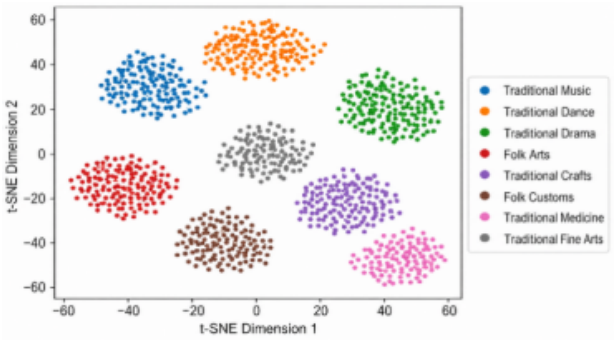

Fig. 12. Visualization of t-SNE in the Multimodal Retrieval Embedding Space

nomenon.

This study emphasizes collaborative protection across training, indexing, and query stages: federated learning alone prevents raw-data upload but not retrieval-result leakage, so ICH systems must incorporate access control and anonymization as core modules.

For scalability, global prototype routing reduces cross-institutional retrieval scope and local vector indexing improves efficiency, keeping latency growth slow as institutions increase — institutions need only update their local index and global routing prototype rather than centralizing all resources.

Limitations: simulated partitioning cannot fully replace real deployment; sensitive-information labeling requires manual intervention; this study focuses on Chinese ICH resources, leaving cross-language retrieval for future work; and audit trails for highly sensitive resources need further development via blockchain or fine-grained permissions.

Deployment should include institutional consent management and expert review for ritual or undisclosed craft knowledge — the framework is a technical layer within broader cultural governance, not a replacement for human authorization.

4. Conclusions

This study addresses centralized-sharing challenges, modal heterogeneity, and scalability limits for cross-institutional ICH retrieval via federated learning combining local multimodal coding, contrastive learning, differential privacy, secure aggregation, distributed routing, and query-stage sensitive filtering, significantly outperforming ordinary federated learning in accuracy, exposure rate, attack resistance, latency, and throughput.

Acknowledgments

This study was supported by Project of Jilin Higher Education Association, Research on the Construction of Talent Cultivation System for "Learning, Research, Production and Innovation" in AI Empowered Design Major - Based on the "Three dimensional Interaction" Model (Project No.: JGJX25D0931).

References

[1] S. Münster, F. Maiwald, I. di Lenardo, J. Henriksson, A. Isaac, M. M. Graf, C. Beck, and J. Oomen, (2024) *"Artificial Intelligence for Digital Heritage Innovation: Setting up a R&D Agenda for Europe"* **Heritage 7**(2): 794–816. DOI: 10.3390/heritage7020038.

[2] F. Ju, (2024) *"Mapping the Knowledge Structure of Image Recognition in Cultural Heritage: A Scientometric Analysis Using CiteSpace, VOSviewer, and Bibliometrix"* **Journal of Imaging *10***(11): 272. DOI: 10 . 3390 / jimaging10110272.

[3] J. Li, X. Zheng, I. Watanabe, and Y. Ochiai, (2024) *"A Systematic Review of Digital Transformation Technologies in Museum Exhibition"* **Computers in Human Behavior *161***: 108407. DOI: 10.1016/j.chb.2024.108407.

[4] L. Tsipi, D. Vouyioukas, G. Loumos, A. Kargas, and D. Varoutas, (2023) *"Digital Repository as a Service (DRaaS): Enhancing Access and Preservation of Cultural Heritage Artifacts"* **Heritage *6***(10): 6881–6900. DOI: 10.3390/heritage6100359.

[5] X. Xia, G. Dong, F. Li, L. Zhu, and X. Ying, (2023) *"When CLIP Meets Cross-Modal Hashing Retrieval: A New Strong Baseline"* **Information Fusion *100***: 101968. DOI: 10.1016/j.inffus.2023.101968.

[6] Y. Sun, Z. Ren, P. Hu, D. Peng, and X. Wang, (2024) *"Hierarchical Consensus Hashing for Cross-Modal Retrieval"* **IEEE Transactions on Multimedia *26***: 824–836. DOI: 10.1109/TMM.2023.3272169.

[7] H. Cai, B. Zhang, J. Li, B. Hu, and J. Chen, (2024) *"Unsupervised Dual Hashing Coding (UDC) on Semantic Tagging and Sample Content for Cross-Modal Retrieval"* **IEEE Transactions on Multimedia *26***: 9109–9120. DOI: 10.1109/TMM.2024.3385986.

[8] J. Liu, J. Huang, Y. Zhou, X. Li, S. Ji, H. Xiong, and D. Dou, (2022) *"From Distributed Machine Learning to Federated Learning: A Survey"* **Knowledge and Information Systems *64***(4): 885–917. DOI: 10.1007/s10115-022-01664-x.

[9] K. Zhang, X. Song, C. Zhang, and S. Yu, (2022) *"Challenges and Future Directions of Secure Federated Learning: A Survey"* **Frontiers of Computer Science *16***(5): 165817. DOI: 10.1007/s11704-021-0598-z.

[10] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava, (2021) *"A Survey on Security and Privacy of Federated Learning"* **Future Generation Computer Systems *115***: 619–640. DOI: 10.1016/j.future.2020.10.007.

[11] Y.-M. Lin, Y. Gao, M.-G. Gong, S.-J. Zhang, Y.-Q. Zhang, and Z.-Y. Li, (2023) *"Federated Learning on Multimodal Data: A Comprehensive Survey"* **Machine Intelligence Research *20***(4): 539–553. DOI: 10.1007/s11633-022-1398-0.

[12] F. Liberti, D. Berardi, and B. Martini, (2024) *"Federated Learning in Dynamic and Heterogeneous Environments: Advantages, Performances, and Privacy Problems"* **Applied Sciences *14***(18): 8490. DOI: 10.3390/app14188490.

[13] P. Qi, D. Chiaro, and F. Piccialli, (2023) *"FL-FD: Federated Learning-Based Fall Detection with Multimodal Data Fusion"* **Information Fusion *99***: 101890. DOI: 10.1016/j.inffus.2023.101890.

[14] G. Feretzakis, K. Papaspyridis, A. Gkoulalas-Divanis, and V. S. Verykios, (2024) *"Privacy-Preserving Techniques in Generative AI and Large Language Models: A Narrative Review"* **Information *15***(11): 697. DOI: 10.3390/info15110697.

[15] Q. Li, Y. Diao, Q. Chen, and B. He. "Federated Learning on Non-IID Data Silos: An Experimental Study". In: *Proceedings of the 2022 IEEE 38th International Conference on Data Engineering (ICDE)*. IEEE, 2022, 965–978. DOI: 10.1109/ICDE53745.2022.00077.