World Wide Web (2024) 27:41
https://doi.org/10.1007/s11280-024-01249-4

Federated learning for supervised cross-modal retrieval

Ang Li[1] · Yawen Li[2] · Yingxia Shao[1]

Received: 31 December 2023 / Revised: 17 January 2024 / Accepted: 24 January 2024

Abstract

In the last decade, the explosive surge in multi-modal data has propelled cross-modal retrieval into the forefront of information retrieval research. Exceptional cross-modal retrieval algorithms are crucial for meeting user requirements effectively and offering invaluable support for subsequent tasks, including cross-modal recommendations, multi-modal content generation, and so forth. Previous methods for cross-modal retrieval typically search for a single common subspace, neglecting the possibility of multiple common subspaces that may mutually reinforce each other in reality, thereby resulting in the poor performance of cross-modal retrieval. To address this issue, we propose a **Fed**erated **S**upervised **C**ross-**M**odal **R**etrieval approach (FedSCMR), which leverages competition to learn the optimal common subspace, and adaptively aggregates the common subspaces of multiple clients for dynamic global aggregation. To reduce the differences between modalities, FedSCMR minimizes the semantic discrimination and consistency in the common subspace, in addition to modeling semantic discrimination in the label space. Additionally, it minimizes modal discrimination and semantic invariance across common subspaces to strengthen cross-subspace constraints and promote learning of the optimal common subspace. In the aggregation stage for federated learning, we design an adaptive model aggregation scheme that can dynamically and collaboratively evaluate the model contribution based on data volume, data category, model loss, and mean average precision, to adaptively aggregate multi-party common subspaces. Experimental results on two publicly available datasets demonstrate that our proposed FedSCMR surpasses state-of-the-art cross-modal retrieval methods.

✉ Yawen Li
warmly0716@126.com

Ang Li
david.lee@bupt.edu.cn

Yingxia Shao
shaoyx@bupt.edu.cn

[1] School of Computer Science, Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876, People's Republic of China

[2] School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876, People's Republic of China

Published online: 26 June 2024

Keywords Federated learning · Cross-modal retrieval · Supervised learning · Multi-modal learning

1 Introduction

Cross-modal retrieval has garnered substantial attention in recent years, driven by the increasing availability of multimodal data [1–3]. Its primary objective is to retrieve semantically relevant results from different modalities when given a query from a specific modality [4–6]. As the demand for privacy preservation grows, coupled with the prevalence of data silos, cross-modal retrieval involving multi-party data that remains within local domains has become increasingly practical and valuable [7]. The pursuit of high-quality cross-modal retrieval holds the potential to enable a wide array of applications, such as cross-modal style transformation [8], the generation of images and text across modalities [9], personalized cross-modal retrieval [10], and recommendations for cross-modal methodologies [11].

Cross-modal retrieval confronts two fundamental challenges: the heterogeneous gap problem [12] and the privacy-preserving problem [13]. The former revolves around the quest for a unified subspace to gauge semantic similarity across diverse modalities, while the latter pertains to the secure exchange of information without divulging it to external parties. In tackling these challenges, it becomes imperative to explore *how to discern a common subspace enabling the direct assessment of semantic similarity between distinct modal data, all while safeguarding the privacy of local data*.

Prior to the most recent cross-modal retrieval methods, linear combination transformations were commonly used to measure the similarity between different modalities [14–19], but they ignored the possibility of nonlinear correlations between multimodal data. In recent years, deep learning-based methods have become popular for cross-modal retrieval due to their ability to capture complex nonlinear relationships between different modalities [20–26]. For bidirectional cross-modal retrieval, He et al. [22] proposed a deep bidirectional representation learning method using two convolutional networks to map images and texts to a common subspace and measure similarity using cosine distance. To find a common subspace, Zheng et al. [23] developed a deep supervised cross-modal retrieval method (DSCMR), which learned modal invariant and discriminant features from a supervised model. However, these methods assume that the data is public and shared, which does not guarantee privacy. To address the public privacy concerns, Google proposed federated learning, which allows large-scale data learning without exposing multi-party privacy [27–32]. However, this method is not directly applicable to cross-modal retrieval due to the large number of parameters or gradients that need to be aggregated, which could lead to significant privacy risks and high maintenance costs. To solve this problem, Zong et al. [32] proposed a federated cross-modal retrieval method (FedCMR), which uses common subspaces to replace model parameters or gradients for global federated training. However, current cross-modal retrieval methods only find a single common subspace, ignoring the possibility of multiple common subspaces that can promote each other and potentially lead to better performance. Furthermore, these methods measure model contributions based on the proportion of number or categories of each client, which does not account for dynamic changes in client contributions during iterative training and may result in reduced global model performance.

To solve the aforementioned problems, we propose FedSCMR - a Federated Supervised Cross-Modal Retrieval method that leverages the competition among multiple common subspaces to promote optimal common subspace learning. Specifically, FedSCMR exploits

inter-modal label prediction to preserve semantic discrimination in the label subspace. To reduce modal discrepancy, FedSCMR simultaneously minimizes the distance between all modal samples within a class and between sample pairs before and after cross-common-subspace mapping. Additionally, to preserve modal discrimination and semantic invariance across common subspaces, FedSCMR minimizes the distance between different modalities before and after the sample pair's cross-common-subspace mapping while maximizing the distance between the same modalities. In the federated aggregation stage, FedSCMR dynamically aggregates the common subspace based on the contribution of each client in each round, utilizing an adaptive federated model aggregation scheme. Experimental results demonstrate that our proposed FedSCMR outperforms state-of-the-art cross-modal retrieval methods. The main contributions of this paper are summarized as follows:

- We propose FedSCMR, a Federated Supervised Cross-Modal Retrieval method that learns the optimal common subspace by minimizing discrimination in label subspaces, semantic discrimination and consistency in common subspaces, and modal discrimination and semantic invariance across common subspaces.
- We design an adaptive federated model aggregation scheme that dynamically evaluates model contribution based on data volume, data category, model loss, and mean average precision to collaboratively aggregate multi-party models.
- Experimental results on two publicly available datasets demonstrate that FedSCMR outperforms the state-of-the-art cross-modal retrieval methods, including methods based on deep learning and federated learning.

The remaining structure of the paper is organized as follows: Section 2 provides a comprehensive summary of related work. Section 3 introduces the problem definition. Section 4 delves into the details of our proposed Federated Supervised Cross-Modal Retrieval (FedSCMR). Section 5 furnishes readers with the specifics of the experimental setup. Moving on to Section 6, we present the experimental results and conduct a thorough analysis. Finally, Section 7 concludes the paper.

2 Related work

2.1 Cross-modal retrieval

Cross-modal retrieval, a task aimed at learning a shared subspace for different modalities, facilitates the projection of data into this space for direct similarity measurement. The canonical correlation analysis (CCA)-based approaches [14–19] and deep learning-based methods [20–26] have been prominent in addressing this task.

Canonical correlation analysis, a set of multivariate statistical techniques, gauges the correlation between two variable sets. For example, Hardoon et al. [14] proposed a method employing kernel canonical correlation analysis for learning semantic representations of network images and corresponding texts. Hwang et al. [15] introduced an unsupervised learning process based on kernel canonical correlation analysis, revealing the correlation between human-labeled images and the significance of objects and layouts in the scene. While effective, these methods suffer from limited representation power due to fixed kernels and variable representation calculation times for new data points, depending on the training set size.

The emergence of deep learning has led to increased exploration of deep models for cross-modal retrieval. Andrew et al. [20] advocated for deep canonical correlation analysis,

leveraging nonlinear transformations for learning highly linearly correlated representations from two data views. Wang et al. [21] enhanced this approach by introducing a deep canonical correlation autoencoder. He et al. [22] proposed a deep bidirectional representation learning model for image-text cross-modal retrieval. Zheng et al. [23] presented a deep supervised cross-modal retrieval method (DSCMR), integrating label space and common representation space discrimination losses for identifying discriminative features. Despite their advancements, existing methods typically search for a single common subspace, overlooking the potential benefits of multiple common subspaces that could complement each other in practice, thereby limiting cross-modal retrieval performance.

2.2 Federated learning

Federated learning is a promising paradigm designed to facilitate collaborative model training across multiple data holders without the necessity of sharing local data [27–31, 33, 34]. The iterative model averaging approach proposed by McMahan et al. [27], known as FedAvg, forms a practical foundation for deep network federated learning. Konečn'y et al. [28] further enhanced communication efficiency in federated learning by introducing methods like structured update and thumbnail update to reduce uplink communication costs.

However, direct parameter averaging in model parameters can lead to a significant decline in performance due to the invariance of neural network parameter permutations [35, 36]. Addressing this challenge, Wang et al. [29] proposed the Federated Matching Average algorithm, tailored for modern neural network architectures like convolutional neural networks and short-term memory networks. This algorithm constructs a shared global model in a hierarchical manner by matching and averaging hidden elements with similar features.

Li et al. [30] introduced the FedProx framework, aiming to handle heterogeneity in federated networks by improving and re-parameterizing FedAvg. In the domain of cross-modal retrieval tasks, Zong et al. [32] proposed a federated cross-modal retrieval method (FedCMR) to mitigate high privacy risks and maintenance costs associated with aggregating large amounts of parameters or gradients. FedCMR leverages distributed multimodal data to train a cross-modal retrieval model, utilizing local data from each client to learn common subspaces across multiple modalities.

However, existing federated learning methods often aggregate models based on the proportion of the number or category of each client, neglecting the dynamic changes in the contribution of different clients to global model training during iterative training. This oversight can lead to a decrease in global model performance.

3 Problem formulation

To avoid losing generality, we describe the federated cross-modal retrieval task using data in both image and text modalities. Given multi-party clients $C = \{C_1, C_2, \cdots, C_N\}$, each client k holds its own data $D_k = \left\{m_1^k, m_2^k, \cdots, m_{n^k}^k\right\}$, where $m_i^k = (v_i^k, t_i^k)$ represents the ith sample (image-text pair) with semantic relevance under client k, and n^k represents the total number of samples under client k. In the client k, each sample m_i^k is assigned a semantic label vector $l_i^k = [y_{i1}^k, y_{i2}^k, \cdots, y_{ic^k}^k] \in \mathbb{R}^c$, where c^k denotes total number of categories. If the category corresponding to the ith sample is j, then $y_{ij}^k = 1$, otherwise $y_{ij}^k = 0$. The image feature matrix is $V^k = [v_1^k, v_2^k, \cdots, v_{n^k}^k]$, where v_i^k represents the image

feature representation learned by the ith sample in the client k; The text feature matrix is $T^k = [t_1^k, t_2^k, \cdots, t_{n^k}^k]$, where t_i^k represents the text feature representation learned from the ith sample in client k.

Data in the federated cross-modal retrieval task is partitioned among multiple parties, resulting in data within each party being inaccessible to other parties in the system. In order to overcome this data isolation, multiple parties must collaborate to train a cross-modal retrieval model. However, image and text data are inherently different in terms of their statistical characteristics and representation spaces, making direct comparison difficult. To address this challenge, the goal of federated cross-modal retrieval is to learn a common subspace S that allows image features V and text features T to be projected into $S_V = f_V(V; \phi_V) \in \mathbb{R}^d$ and $S_T = f_T(T; \phi_T) \in \mathbb{R}^d$, respectively, where ϕ_V and ϕ_T are parameters that can be learned during training of the two mapping functions. By mapping both modalities into the same subspace, the cross-modal retrieval model can directly compare image and text features to facilitate accurate retrieval.

4 Federated supervised cross-modal retrieval method

We propose a federated supervised cross-modal retrieval method called FedSCMR, which is illustrated in Figure 1. In each local client, image and text features are extracted from images and text, respectively. The extracted features are then mapped to a common subspace through a basic network, enabling direct comparison between the modalities. Additionally, the image and text features mapped from the basic network are mapped to a common subspace through an enhancement network, further improving retrieval performance. The cross-modal retrieval model under each client is collaboratively trained by learning in the label space, common subspace, and cross common subspace. In the global model, the common subspaces uploaded

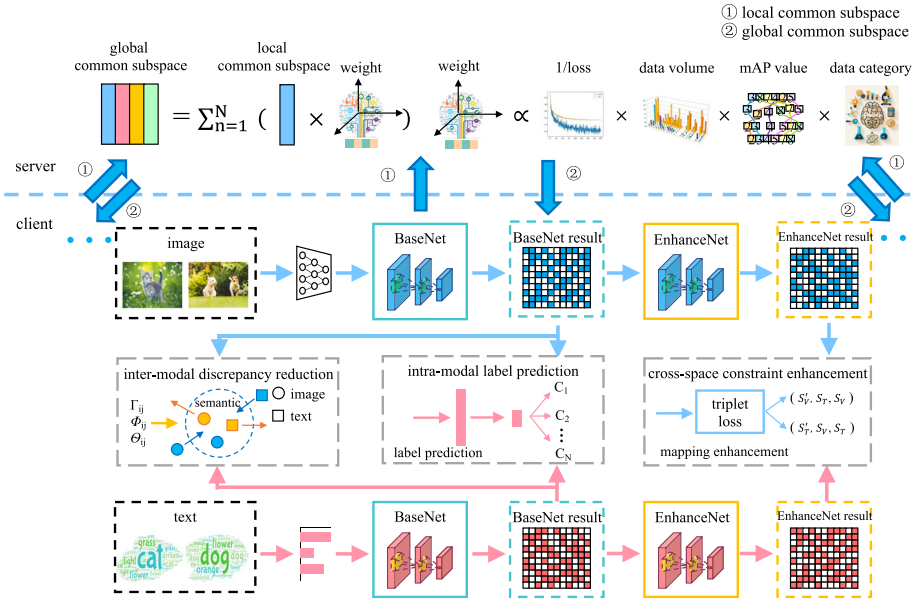


Figure 1 The overall framework of FedSCMR

by each local client are aggregated, and the aggregated common subspaces are distributed to each client for local updates.

4.1 Local cross-modal retrieval model

Our local cross-modal retrieval model aims to map data from different modalities into a unified common subspace that enables direct comparison. The model comprises three parts: intra-modal label prediction, inter-modal discrepancy reduction, and cross-space constraint enhancement. The intra-modal label prediction ensures that the mapped features retain their original semantics. The inter-modal discrepancy reduction reduces interference from modal differences in semantic discrimination and consistency. Finally, the cross-space constraint enhancement preserves modal discrimination and semantic invariance across common subspaces.

4.1.1 Intra-modal label prediction

To preserve the semantic information of the mapped features, we employ a linear classifier to perform label prediction on the mapped sample (image, text pair) $m_i = (v_i, t_i)$ in the common subspace S. For this, we add a linear layer on top of the BaseNet of the image and text modes, respectively, and obtain a prediction label vector of dimension c^k for each sample. To calculate the discrimination loss L_{imd} in the label space, we use the loss function introduced in [23]:

$$L_{imd}(\phi_{imd}) = \frac{1}{n^k} \left\| \hat{p}_i(v_i) - l_i \right\|_F + \frac{1}{n^k} \left\| \hat{p}_i(t_i) - l_i \right\|_F, \tag{1}$$

where $\|\cdot\|_F$ represents the Frobenius norm, ϕ_{imd} represents the parameters of the classifier, l_i is the semantic label for each sample m_i, $\hat{p}_i$ is the probability distribution predicted for the sample m_i.

4.1.2 Inter-modal discrepancy reduction

The inter-modal discrepancy reduction module aims to ensure that the semantic discrimination and consistency of all samples are not affected by different modalities. Specifically, it ensures that samples with the same semantics but different modalities are close to each other, while those with different semantics and modalities are far apart. For samples with all modalities in the common subspace S, we employ the loss function proposed in [32] to calculate the semantic discrimination loss:

$$l_{csd} = \frac{1}{(n^k)^2} \sum_{x,y=1}^{2} \sum_{i,j=1}^{n^k} (log(1 + e^{\Gamma_{ij}^{xy}}) - \delta_{ij}^{xy} \Gamma_{ij}^{xy}), \tag{2}$$

where $\Gamma_{ij}^{xy} = \frac{1}{2} cos(x_i, y_j)$ represents the cosine similarity between x modality and y modality of sample m_i, δ_{ij}^{xy} is an indicator function. When the samples m_i and m_j belong to the same category, the value is 1, otherwise, the value is 0. If $x = y$, the two samples are samples within the same modality. To ensure semantic consistency before and after the sample features in the common subspace S are mapped to the common subspace $S^{'}$, the constructed semantic consistency loss is:

$$l_{csc} = \left\| S_{V^k} - S_{T^k} \right\|_F + \left\| S'_{V^k} - S'_{T^k} \right\|_F, \tag{3}$$

where S_{V^k} and S_{T^k} represents the characteristics of images and text mapped in the common subspace S through the BaseNet, $S^{'}_{V^k}$ and $S^{'}_{T^k}$ represents the features of the image and text features in the subspace S projected into the common subspace $S^{'}$ again through an EnhanceNet.

Semantic discrimination and consistency loss L_{cs} in common subspaces is expressed as:

$$L_{cs}(\phi_V, \phi_T) = l_{csd} + l_{csc}, \tag{4}$$

where l_{csd} represents the semantic discriminant loss of all modalities in the common subspace S, l_{csc} represents the loss of semantic consistency before and after the mapping of sample features from the common subspace S to $S^{'}$.

4.1.3 Cross-space constraint enhancement

The cross-space constraint enhancement module is designed to enhance the image features $S_{V^k} = f_{V^k}(V^k; \phi_{V^k})$ and text features $S_{T^k} = f_{T^k}(T^k; \phi_{T^k})$ in the common subspace S using an enhancement network, which helps to further learn the common subspace $S^{'}$. The enhanced features are then mapped to $S^{'}_{V^k} = g_{V^k}(V^k; \phi_{V^k})$ and $S^{'}_{T^k} = g_{T^k}(T^k; \phi_{T^k})$. The motivation behind this module is to encourage competition between the new common subspace $S^{'}$ and the original common subspace S in order to promote the learning of the optimal common subspace. The goal is to ensure that modal discriminancy and semantic invariance are still preserved across the common subspace. In other words, data with the same semantics but different modes should still have similar distances before and after mapping to the common subspace. The modal discrimination and semantic invariance loss across the common subspace is calculated using the following equations:

$$L_{ccs,V}(\phi_V, \phi_T, \phi_{S_V}) = max(0, \left\| S^{'}_V - S_T \right\|_F - \left\| S^{'}_V - S_V \right\|_F), \tag{5}$$

$$L_{ccs,T}(\phi_T, \phi_V, \phi_{S_T}) = max(0, \left\| S^{'}_T - S_V \right\|_F - \left\| S^{'}_T - S_T \right\|_F), \tag{6}$$

where $L_{ccs,V}(\phi_V, \phi_T, \phi_{S_V})$ and $L_{ccs,T}(\phi_T, \phi_V, \phi_{S_T})$ are the loss of modal discrimination and semantic invariance for images and text across common spaces, respectively. Therefore, the loss L_{ccs} of modal discrimination and semantic invariance across common subspaces is:

$$L_{ccs}(\phi_V, \phi_T, \phi_{S_V}, \phi_{S_T}) = L_{ccs,V}(\phi_V, \phi_T, \phi_{S_V}) + L_{ccs,T}(\phi_T, \phi_V, \phi_{S_T}). \tag{7}$$

4.1.4 Loss of local cross-modal retrieval model

The local cross-modal retrieval model loss L_{total} is composed of three terms: discrimination loss in the label space L_{imd}, semantic discrimination and consistency loss in common subspaces L_{cs}, and modal discrimination and semantic invariance loss across common subspaces L_{ccs}. These terms are jointly composed to obtain the final loss function L_{total}.

$$L_{total}(\phi_V, \phi_T, \phi_{S_V}, \phi_{S_T}, \phi_{imd}) = \alpha \cdot L_{cs} + \beta \cdot L_{ccs} + L_{imd}, \tag{8}$$

where α and β are coefficiency of L_{cs} and L_{ccs} respectively, which are utilized to adjust the proportion of L_{cs} and L_{ccs} in the total loss of model.

4.2 Adaptive model aggregation scheme

4.2.1 Model aggregation

Follow previous work [32], we leverage common subspace to share the knowledge learned by various local clients. Since the parameters of the linear layer play an important role in learning the common subspace, the central server aggregates the linear layer parameters of each client through (9), thereby aggregating the common subspace of the global model.

$$W_t = \sum_{k=1}^{K} \lambda_t^k W_t^k, \tag{9}$$

where K is the total number of clients, W_t^k represents the linear layer parameter uploaded by the client k to represent the common subspace during the tth round of communication, λ_t^k represents the weight of W_t^k during the tth round of communication, i.e. the contribution of the client k during the tth round of communication.

4.2.2 Adaptive contribution measurement

Assume the contribution λ_t^k of common subspace W_t^k of client k is proportional to the number of samples n^k and the number of sample label categories c^k, and inversely proportional to the loss of the model l^k:

$$\lambda_t^k \propto \Omega^k = \frac{n^k}{\sum_{k=1}^{K} n^k} \cdot \frac{c^k}{\sum_{k=1}^{K} c^k}, \tag{10}$$

$$\lambda_t^k \propto \Psi_t^k = e^{-e^{l_t^k}/(\frac{\sum_{k=1}^{K} l_t^k}{K})}, \tag{11}$$

Considering the contribution of common subspace W_t^k of client k to global common subspace W_t varies dynamically with training results, W_t^k's contribution is proportional to the average mean precision (mAP) v_t^k on the cross-modal retrieval task during the tth round training:

$$\lambda_t^k \propto \Phi_t^k = \frac{v_t^k}{\sum_{k=1}^{K} v_t^k}, \tag{12}$$

The adaptive contribution λ_t^k of the common subspace W_t^k in the client k is calculated as follows:

$$\lambda_t^k = \frac{e^{\Omega^k + \Psi_t^k + \gamma \Phi_t^k}}{\sum_{k=-1}^{K} e^{\Omega^k + \Psi_t^k + \gamma \Phi_t^k}}. \tag{13}$$

where γ is an adjustable parameter to control Φ_t^k's participation.

4.3 Training process of FedSCMR

The training process of FedSCMR is presented in Algorithm 1, which includes two roles: client and server. In each client k, FedSCMR undergoes R rounds of iterative training. During each round of training, the client receives the current global common subspace W_t from the server, updates its local common subspace W_t^k based on its own data, and uploads the updated W_t^k to the server. On the server side, it receives the uploaded W_t^k from all clients, aggregates

Algorithm 1 The training process of FedSCMR.

Input: client set $C = \{C_1, C_2, \cdots, C_N\}$, data set $D = \{D_1, D_2, \cdots, D_N\}$, communication epoch R, number of clients k.
Output: W_R: global common subspace in epoch R.

Initialize W_0.
Role: client k
for *epoch* $t \leftarrow 0, 1, 2, \ldots, R-1$ **do**
  Receive global common subspace W_t sent from server.
  Update local common subspace W_{t+1}^k by using received W_t on client k's local data:
  $W_{t+1}^k \leftarrow$ Update(k, W_t).
  Upload W_{t+1}^k to server.
end
Role: server
for client $k \in C_t$ **do**
  Receive common subspace W_{t+1}^k from client k.
  Calculate client k's contribution λ_{t+1}^k by using (13).
end
Calculate global common subspace W_{t+1} by using (9).
Send W_{t+1} to client k.

them based on the contribution λ_t^k of each client, and obtains the updated global common subspace W_t which is then sent back to all clients for the next round of training.

5 Experimental setup

5.1 Research questions

The research questions that lead to the rest of this paper are as follows:

RQ1: Is the performance of the supervised federated cross-modal retrieval method FedSCMR superior to the state-of-the-art federated cross-modal retrieval methods?

RQ2: Does the main components of our FedSCMR contribute to the federated cross-modal retrieval task?

RQ3: How model parameters affect FedSCMR in federated cross-modal retrieval tasks?

5.2 Dataset

We conducted experiments using the Pascal Sentence [37] dataset and Wikipedia data [38]. The Pascal dataset is a subset of the Pascal VOC dataset, comprising 1,000 pairs of images and text descriptions. It consists of 20 categories, each containing 50 pairs of images and corresponding text. The Wikipedia dataset is the most widely used dataset, containing 10 categories and a total of 2,866 pairs of images and text descriptions.

5.3 Baselines

We compare our FedSCMR with the state-of-the-art cross-media retrieval methods, and the baselines are as follows

Deep Supervised Cross-Modal Retrieval (DSCMR): This method [23] minimizes discrimination loss in the label space and common representation space while supervising model learning to identify features. It also minimizes the loss of modal invariance. We run DSCMR on each client and do not aggregate each model.

Federated Averaging (FedAvg): This method [27] combines local random gradient descent on each client with a server that performs model averaging. It is robust to unbalanced and non-independent identically distributed data. In our experiments, DSCMR are trained on each client, and FedAvg is used for model aggregation.

FedProx: This method [30] enables multiple users to obtain convergence guarantees when learning data from non-independent identically distributed distributions (statistical heterogeneity). In our experiments, DSCMR are trained on each client, and FedProx is used for model aggregation.

Federated Cross-Modal Retrieval (FedCMR): This method [32] trains a cross-modal retrieval model and uses local data from each client to learn common spaces across multiple modalities.

To further analyze the contribution of the main component of our FedSCMR to cross-modal retrieval, we use three versions of FedSCMR for comparison:

Federated Supervised Cross-Modal Retrieval (FedSCMR): This method promotes learning of the optimal common subspace by constructing multiple common subspaces to compete together while performing adaptive model aggregation based on the contribution of multiple models during the federated aggregation phase.

FedSCMR$_{\neg Lcs}$: This method removes semantic discrimination and consistency loss L_{cs} from the common subspace based on FedSCMR.

FedSCMR$_{\neg Lccs}$: This method removes modal discrimination and semantic invariance losses L_{ccs} across common subspaces based on FedSCMR.

FedSCMR$_{\neg \Phi}$: This method removes adaptive federated model aggregation scheme according to dynamic mAP performance during each training round.

5.4 Parameter setting

We use the fc7 layer of VGG-19 [39] to extract 4,096 dimensional image feature vectors from image samples, and use sentence BERT [40] to extract 1,024 dimensional text feature vectors from text samples. Our proposed FedSCMR sets the α and β to [0.1, 1, 10, 100], respectively. PySyft [41] is used to simulate federated cross-modal retrieval.

6 Results and analysis

6.1 Effectiveness of FedSCMR

To answer *RQ1*, we compared FedSCMR with all state-of-the-art methods on Pascal dataset and Wikipedia dataset, and evaluated their mean Average Precision (mAP@N) and Normalized Discounted Cumulative Gain (NDCG@N) in three clients for image-to-text retrieval (img2txt), text-to-image retrieval (txt2img), and average retrieval, where $N = 10, 20, 30,$ and 50. Tables 1 and 2 present the performance of all methods. From Tables 1 and 2, we have the following findings: i) FedSCMR outperformed all state-of-the-art federated cross-modal retrieval methods in all retrieval tasks on all clients, demonstrating the effectiveness of

Table 1 The cross-modal retrieval performance, mAP@N and NDCG@N, of FedSCMR and the state-of-the-art methods on Pascal dataset, where $N = \{10, 20, 30, 50\}$

Metric	Task	Method	N=10			N=20			N=30			N=50		
			A	B	C	A	B	C	A	B	C	A	B	C
mAP	Img2Txt	DSCMR	0.6970	0.7157	0.7086	0.6891	0.7036	0.6989	0.6859	0.6973	0.6865	0.6760	0.6882	0.6764
		FedAvg	0.6883	0.6913	0.6829	0.6747	0.6835	0.6697	0.6702	0.6826	0.6636	0.6634	0.6784	0.6457
		FedProx	0.6562	0.6691	0.6571	0.6479	0.6668	0.6415	0.6454	0.6637	0.6332	0.6369	0.6583	0.6137
		FedCMR	0.7094	0.7295	0.7121	0.6985	0.7182	0.7008	0.6907	0.7083	0.6880	0.6770	0.7019	0.6823
		FedSCMR	0.7163	0.7398	0.7172	0.7067	0.7359	0.7037	0.7029	0.7326	0.6916	0.6989	0.7246	0.6895
	Txt2Img	DSCMR	0.7015	0.7120	0.6850	0.7023	0.7097	0.6728	0.6947	0.7076	0.6581	0.6807	0.6873	0.6385
		FedAvg	0.6699	0.6795	0.6647	0.6610	0.6794	0.6595	0.6537	0.6730	0.6478	0.6495	0.6532	0.6369
		FedProx	0.5870	0.6217	0.5821	0.5858	0.6205	0.5784	0.5822	0.6142	0.5686	0.5721	0.5970	0.5574
		FedCMR	0.7301	0.7331	0.7225	0.7182	0.7301	0.7133	0.7099	0.7158	0.7013	0.6918	0.7031	0.6830
		FedSCMR	0.7541	0.7727	0.7245	0.7480	0.7608	0.7140	0.7338	0.7437	0.7066	0.7281	0.7297	0.6939
	Average	DSCMR	0.6992	0.7138	0.6968	0.6957	0.7067	0.6858	0.6903	0.7025	0.6723	0.6784	0.6878	0.6574
		FedAvg	0.6791	0.6854	0.6738	0.6679	0.6815	0.6646	0.6620	0.6778	0.6557	0.6564	0.6658	0.6413
		FedProx	0.6216	0.6454	0.6196	0.6169	0.6437	0.6099	0.6138	0.6389	0.6009	0.6045	0.6277	0.5856
		FedCMR	0.7198	0.7313	0.7173	0.7084	0.7241	0.7070	0.7003	0.7121	0.6946	0.6844	0.7025	0.6826
		FedSCMR	0.7352	0.7563	0.7208	0.7273	0.7484	0.7088	0.7184	0.7382	0.6991	0.7135	0.7271	0.6917
NDCG	Img2Txt	DSCMR	0.7702	0.7645	0.7404	0.7775	0.7695	0.7479	0.7801	0.7665	0.7528	0.7830	0.7751	0.7533
		FedAvg	0.7652	0.7480	0.7314	0.7655	0.7578	0.7326	0.7672	0.7655	0.7437	0.7677	0.7687	0.7424
		FedProx	0.7431	0.7449	0.7303	0.7571	0.7533	0.7310	0.7604	0.7620	0.7342	0.7594	0.7655	0.7372
		FedCMR	0.7801	0.7735	0.7736	0.7924	0.7846	0.7889	0.7948	0.7929	0.7924	0.7968	0.7955	0.7912
		FedSCMR	0.7981	0.8084	0.8140	0.8114	0.8139	0.8140	0.8139	0.8165	0.8159	0.8147	0.8170	0.8114

Table 1 continued

Metric	Task	Method	N=10			N=20			N=30			N=50		
			A	B	C	A	B	C	A	B	C	A	B	C
	Txt2Img	DSCMR	0.7774	0.7966	0.7642	0.7920	0.8078	0.7705	0.7898	0.8101	0.7783	0.7906	0.8070	0.7759
		FedAvg	0.7747	0.7868	0.7540	0.7816	0.8046	0.7683	0.7809	0.8055	0.7716	0.7837	0.8019	0.7722
		FedProx	0.7552	0.7618	0.7182	0.7623	0.7714	0.7305	0.7635	0.7759	0.7337	0.7616	0.7716	0.7316
		FedCMR	0.8015	0.8020	0.7648	0.8029	0.8131	0.7788	0.8054	0.8171	0.7838	0.8071	0.8159	0.7828
		FedSCMR	0.8085	0.8147	0.8070	0.8174	0.8172	0.8095	0.8185	0.8256	0.8106	0.8217	0.8241	0.8133
	Average	DSCMR	0.7738	0.7805	0.7523	0.7848	0.7886	0.7592	0.7850	0.7883	0.7655	0.7868	0.7911	0.7646
		FedAvg	0.7700	0.7674	0.7427	0.7735	0.7812	0.7505	0.7741	0.7855	0.7576	0.7757	0.7853	0.7573
		FedProx	0.7491	0.7533	0.7242	0.7597	0.7623	0.7308	0.7619	0.7690	0.7339	0.7605	0.7685	0.7344
		FedCMR	0.7908	0.7877	0.7692	0.7976	0.7989	0.7839	0.8001	0.8050	0.7881	0.8019	0.8057	0.7870
		FedSCMR	0.8033	0.8115	0.8105	0.8144	0.8155	0.8117	0.8162	0.8211	0.8133	0.8182	0.8206	0.8124

Table 2 The cross-modal retrieval performance, mAP@N and NDCG@N, of FedSCMR and the state-of-the-art methods on Wikipedia dataset, where $N = \{10, 20, 30, 50\}$

Metric	Task	Method	N=10			N=20			N=30			N=50		
			A	B	C	A	B	C	A	B	C	A	B	C
mAP	Img2Txt	DSCMR	0.4726	0.4623	0.4497	0.4720	0.4512	0.4524	0.4644	0.4566	0.4490	0.4625	0.4514	0.4383
		FedAvg	0.4507	0.4417	0.4300	0.4419	0.4332	0.4244	0.4342	0.4273	0.4211	0.4225	0.4218	0.4108
		FedProx	0.4066	0.4204	0.3897	0.4041	0.4190	0.3874	0.4014	0.4110	0.3844	0.3912	0.4021	0.3753
		FedCMR	0.4857	0.4679	0.5060	0.4732	0.4619	0.5029	0.4686	0.4587	0.4958	0.4661	0.4571	0.4822
		FedSCMR	0.4998	0.4997	0.5148	0.4942	0.4974	0.5107	0.4904	0.4958	0.5071	0.4889	0.4931	0.4867
	Txt2Img	DSCMR	0.4330	0.4345	0.3934	0.4316	0.4318	0.3946	0.4257	0.4251	0.3909	0.4199	0.4173	0.3805
		FedAvg	0.4025	0.4226	0.3897	0.3727	0.4133	0.3782	0.3564	0.3885	0.3543	0.3426	0.3583	0.3275
		FedProx	0.3530	0.3542	0.3208	0.3518	0.3520	0.3217	0.3470	0.3466	0.3187	0.3323	0.3402	0.3102
		FedCMR	0.4412	0.4448	0.4612	0.4335	0.4412	0.4292	0.4281	0.4358	0.4058	0.4242	0.4264	0.3989
		FedSCMR	0.4516	0.5163	0.4811	0.4479	0.4678	0.4407	0.4364	0.4376	0.4129	0.4359	0.4371	0.4031
	Average	DSCMR	0.4528	0.4484	0.4215	0.4518	0.4415	0.4235	0.4451	0.4408	0.4200	0.4412	0.4344	0.4094
		FedAvg	0.4266	0.4321	0.4098	0.4073	0.4233	0.4013	0.3953	0.4079	0.3877	0.3826	0.3901	0.3691
		FedProx	0.3798	0.3873	0.3552	0.3780	0.3855	0.3546	0.3742	0.3788	0.3516	0.3618	0.3711	0.3427
		FedCMR	0.4635	0.4563	0.4836	0.4534	0.4516	0.4661	0.4484	0.4473	0.4508	0.4452	0.4417	0.4405
		FedSCMR	0.4757	0.5080	0.4980	0.4710	0.4826	0.4757	0.4634	0.4667	0.4600	0.4624	0.4651	0.4449
NDCG	Img2Txt	DSCMR	0.5231	0.5275	0.5211	0.5461	0.5335	0.5204	0.5353	0.5364	0.5235	0.5346	0.5389	0.5225
		FedAvg	0.5133	0.5243	0.5135	0.5329	0.5299	0.5158	0.5239	0.5297	0.5133	0.5292	0.5024	0.4914
		FedProx	0.4892	0.4906	0.4481	0.5070	0.5097	0.4676	0.5292	0.5224	0.4926	0.5181	0.4978	0.4760
		FedCMR	0.5326	0.5342	0.5352	0.5500	0.5349	0.5577	0.5727	0.5500	0.5826	0.6205	0.6024	0.6336
		FedSCMR	0.5381	0.5419	0.5692	0.5604	0.5712	0.5897	0.5871	0.5877	0.6091	0.6444	0.6416	0.6507

Table 2 continued

Metric	Task	Method	N=10			N=20			N=30			N=50		
			A	B	C	A	B	C	A	B	C	A	B	C
	Txt2Img	DSCMR	0.5152	0.5197	0.4900	0.5201	0.5213	0.4983	0.5208	0.5203	0.5006	0.5195	0.5214	0.4991
		FedAvg	0.4980	0.5176	0.4860	0.4998	0.5174	0.4900	0.4991	0.5145	0.4895	0.4993	0.5122	0.4944
		FedProx	0.4755	0.5132	0.4823	0.4847	0.5077	0.4888	0.4879	0.5065	0.4837	0.4903	0.5035	0.4905
		FedCMR	0.5266	0.5204	0.5432	0.5260	0.5241	0.5436	0.5258	0.5242	0.5404	0.5218	0.5259	0.5391
		FedSCMR	0.5305	0.5477	0.5585	0.5448	0.5485	0.5589	0.5531	0.5437	0.5599	0.5588	0.5401	0.5606
	Average	DSCMR	0.5192	0.5236	0.5055	0.5331	0.5274	0.5093	0.5281	0.5283	0.5121	0.5271	0.5302	0.5108
		FedAvg	0.5056	0.5210	0.4997	0.5164	0.5236	0.5029	0.5115	0.5221	0.5014	0.5142	0.5073	0.4929
		FedProx	0.4824	0.5019	0.4652	0.4959	0.5087	0.4782	0.5086	0.5144	0.4881	0.5042	0.5007	0.4832
		FedCMR	0.5296	0.5273	0.5392	0.5380	0.5295	0.5506	0.5493	0.5371	0.5615	0.5712	0.5642	0.5864
		FedSCMR	0.5343	0.5448	0.5638	0.5526	0.5598	0.5743	0.5701	0.5657	0.5845	0.6016	0.5908	0.6057

our proposed method; ii) FedSCMR considers both discrimination in label space, semantic discrimination and consistency in common subspaces, and modal discrimination and semantic invariance across common subspaces. Its performance is generally superior to FedCMR, which only considers discrimination in label space, semantic discrimination in common subspaces, and consistency between modalities. This finding indicates that minimizing modal discrimination and semantic invariance across common subspaces can improve the performance of cross-modal retrieval; iii) FedSCMR and FedCMR, which use adaptive federated aggregation methods, perform better than DSCMR, FedAvg, and FedProx, which perform federated aggregation based on a fixed ratio. This finding demonstrates that dynamically assessing the client's contribution to the server can enhance the performance of cross-modal retrieval; iv) The results demonstrate superior performance of federated learning methods, namely FedSCMR, FedCMR, FedProx, and FedAvg, compared to DSCMR, which is trained independently on each client without information aggregation. This suggests that federated learning effectively enhances model performance; v) The performance of our FedSCMR on the Wikipedia dataset is consistent with the conclusions drawn from the Pascal dataset, indicating that FedSCMR's performance is not limited to a specific dataset but rather exhibits generalization.

To further validate the effectiveness of FedSCMR, we compare the average mAP@N performance of FedSCMR and the state-of-the-art methods for Pascal and Wikipedia datasets with different top N searched results, where $N = 10$, 20, 30, and 50. The performance is reported in Figures 2 and 3. In the scenarios with N=10, N=20, N=30, and N=50, it is evident that FedSCMR consistently outperforms state-of-the-art cross-modal retrieval methods. This superior performance is observed across tasks, including text retrieval given image (img2txt),

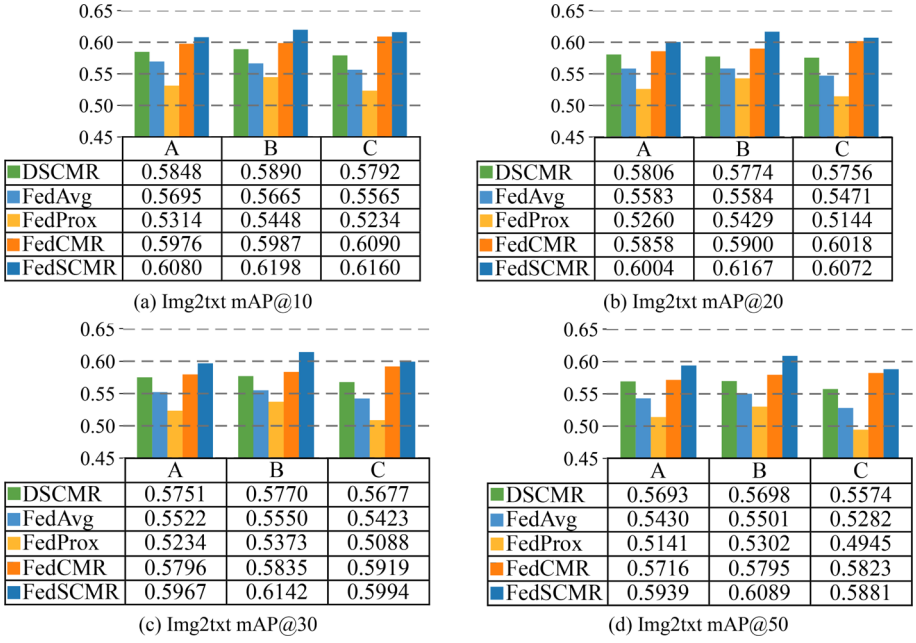
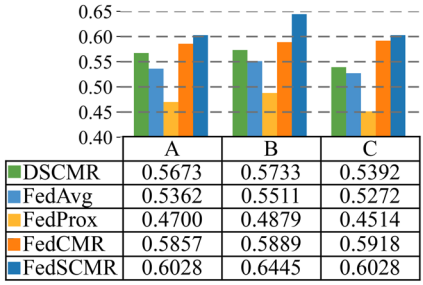
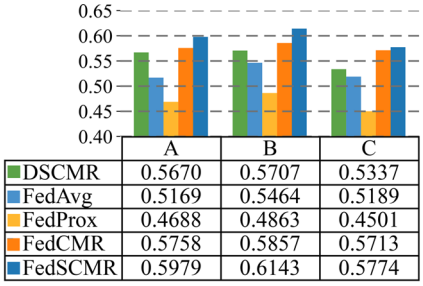


Figure 2 The average mAP performance of FedSCMR and the state-of-the-art methods for Pascal and Wikipedia datasets with different top N searched results for image-to-text retrieval (img2txt)



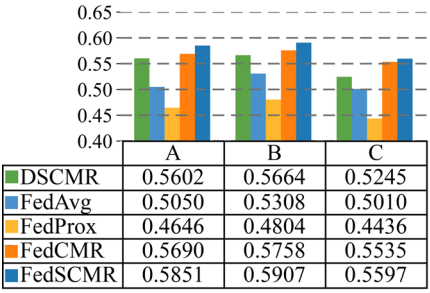
	A	B	C
DSCMR	0.5673	0.5733	0.5392
FedAvg	0.5362	0.5511	0.5272
FedProx	0.4700	0.4879	0.4514
FedCMR	0.5857	0.5889	0.5918
FedSCMR	0.6028	0.6445	0.6028

(a) Txt2img mAP@10



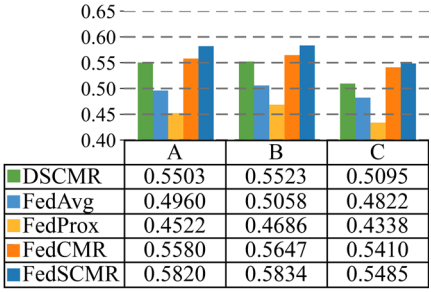
	A	B	C
DSCMR	0.5670	0.5707	0.5337
FedAvg	0.5169	0.5464	0.5189
FedProx	0.4688	0.4863	0.4501
FedCMR	0.5758	0.5857	0.5713
FedSCMR	0.5979	0.6143	0.5774

(b) Txt2img mAP@20



	A	B	C
DSCMR	0.5602	0.5664	0.5245
FedAvg	0.5050	0.5308	0.5010
FedProx	0.4646	0.4804	0.4436
FedCMR	0.5690	0.5758	0.5535
FedSCMR	0.5851	0.5907	0.5597

(c) Txt2img mAP@30



	A	B	C
DSCMR	0.5503	0.5523	0.5095
FedAvg	0.4960	0.5058	0.4822
FedProx	0.4522	0.4686	0.4338
FedCMR	0.5580	0.5647	0.5410
FedSCMR	0.5820	0.5834	0.5485

(d) Txt2img mAP@50

Figure 3 The average mAP performance of FedSCMR and the state-of-the-art methods for Pascal and Wikipedia datasets with different top N searched results for text-to-image retrieval (txt2img)

and image retrieval given text (txt2img), as measured by mAP@N. This, again, demonstrates the effectiveness of FedSCMR on cross-modal retrieval tasks.

6.2 Effectiveness of main components of FedSCMR

We now turn to answer *RQ2*. We compare FedSCMR with its variants, which are FedSCMR$_{\neg Lcs}$, FedSCMR$_{\neg Lccs}$, and FedSCMR$_{\neg \Phi}$. FedSCMR$_{\neg Lcs}$ represents FedSCMR that removes semantic discrimination and consistency loss L_{cs} in the common subspace. FedSCMR$_{\neg Lccs}$ represents FedSCMR that removes modal discrimination across the common subspace and semantic invariance loss L_{ccs}. FedSCMR$_{\neg \Phi}$ represents FedSCMR that removes adaptive federated model aggregation scheme according to dynamic mAP performance during each training round. Since the discrimination loss L_{imd} in the label space is not an innovation in this paper, no ablation experiment is conducted on it.

The results of the ablation experiment are shown in Tables 3 and 4. From the results, we have the following findings: i) The performance of FedSCMR on the tasks of image-to-text retrieval, text-to-image retrieval, and average retrieval surpasses all variants, as evidenced by mAP@N and NDCG@N. This, once again, demonstrates the superiority of FedSCMR in cross-modal retrieval tasks; ii) The FedSCMR, which takes into account semantic discrimination and consistency loss L_{cs}, outperforms the FedSCMR$_{\neg Lcs}$, which ignores L_{cs}. This suggests that incorporating semantic discrimination and consistency to achieve inter-modal discrepancy reduction contributes to improving the performance of cross-modal retrieval; iii) The FedSCMR, which considers modal discrimination and semantic invariance loss L_{ccs}, outperforms the FedSCMR$_{\neg Lcs}$, which removes L_{ccs} from FedSCMR. This indicates that

Table 3 The ablation study for cross-modal retrieval performance on Pascal dataset

Metric	Task	Method	N=10			N=20			N=30			N=50		
			A	B	C	A	B	C	A	B	C	A	B	C
mAP	Img2Txt	FedSCMR$_{\neg Lcs}$	0.7002	0.6240	0.6914	0.6932	0.6241	0.6819	0.6810	0.6236	0.6704	0.6683	0.6219	0.6523
		FedSCMR$_{\neg Lccs}$	0.7078	0.7177	0.7079	0.6975	0.7159	0.7009	0.6895	0.7100	0.6902	0.6790	0.7007	0.6853
		FedSCMR$_{\neg \Phi}$	0.6913	0.7180	0.6764	0.6738	0.7042	0.6823	0.6717	0.7015	0.6781	0.6719	0.6914	0.6658
		FedSCMR	0.7163	0.7398	0.7172	0.7067	0.7359	0.7037	0.7029	0.7326	0.6916	0.6989	0.7246	0.6895
	Txt2Img	FedSCMR$_{\neg Lcs}$	0.7518	0.7536	0.7021	0.7474	0.7352	0.6902	0.7201	0.7140	0.6793	0.6851	0.6885	0.6598
		FedSCMR$_{\neg Lccs}$	0.7510	0.7707	0.7216	0.7436	0.7566	0.7026	0.7308	0.7411	0.6925	0.7210	0.7284	0.6860
		FedSCMR$_{\neg \Phi}$	0.7476	0.7674	0.7142	0.7414	0.7582	0.7112	0.7308	0.7418	0.7036	0.7207	0.7235	0.6861
		FedSCMR	0.7541	0.7727	0.7245	0.7480	0.7608	0.7140	0.7338	0.7437	0.7066	0.7281	0.7297	0.6939
	Average	FedSCMR$_{\neg Lcs}$	0.7260	0.6888	0.6967	0.7203	0.6796	0.6861	0.7006	0.6688	0.6748	0.6767	0.6552	0.6561
		FedSCMR$_{\neg Lccs}$	0.7294	0.7442	0.7147	0.7206	0.7363	0.7017	0.7102	0.7255	0.6914	0.7000	0.7146	0.6857
		FedSCMR$_{\neg \Phi}$	0.7195	0.7427	0.6953	0.7076	0.7312	0.6967	0.7012	0.7217	0.6908	0.6963	0.7075	0.6760
		FedSCMR	0.7352	0.7563	0.7208	0.7273	0.7484	0.7088	0.7184	0.7382	0.6991	0.7135	0.7271	0.6917
NDCG	Img2Txt	FedSCMR$_{\neg Lcs}$	0.7801	0.7467	0.7439	0.7893	0.7620	0.7587	0.7963	0.7686	0.7676	0.7955	0.7744	0.7750
		FedSCMR$_{\neg Lccs}$	0.7944	0.7904	0.7584	0.8085	0.8106	0.7738	0.8127	0.8130	0.7820	0.8113	0.8144	0.7868
		FedSCMR$_{\neg \Phi}$	0.7969	0.7910	0.7761	0.8105	0.8079	0.8012	0.8107	0.8134	0.8035	0.8129	0.8156	0.8033
		FedSCMR	0.7981	0.8084	0.8140	0.8114	0.8139	0.8140	0.8139	0.8165	0.8159	0.8147	0.8170	0.8114
	Txt2Img	FedSCMR$_{\neg Lcs}$	0.7687	0.8129	0.7392	0.7714	0.8147	0.7547	0.7622	0.8084	0.7581	0.7644	0.8034	0.7600
		FedSCMR$_{\neg Lccs}$	0.8037	0.8012	0.7544	0.8146	0.8130	0.7729	0.8159	0.8140	0.7762	0.8198	0.8123	0.7791
		FedSCMR$_{\neg \Phi}$	0.7921	0.8050	0.7712	0.8029	0.8105	0.7836	0.8048	0.8131	0.7864	0.8079	0.8163	0.7856
		FedSCMR	0.8085	0.8147	0.8070	0.8174	0.8172	0.8095	0.8185	0.8256	0.8106	0.8217	0.8241	0.8133
	Average	FedSCMR$_{\neg Lcs}$	0.7744	0.7798	0.7415	0.7803	0.7883	0.7567	0.7792	0.7885	0.7629	0.7799	0.7889	0.7675
		FedSCMR$_{\neg Lccs}$	0.7991	0.7958	0.7564	0.8115	0.8118	0.7733	0.8143	0.8135	0.7791	0.8155	0.8134	0.7830
		FedSCMR$_{\neg \Phi}$	0.7945	0.7980	0.7736	0.8067	0.8092	0.7924	0.8077	0.8133	0.7950	0.8104	0.8159	0.7944
		FedSCMR	0.8033	0.8115	0.8105	0.8144	0.8155	0.8117	0.8162	0.8211	0.8133	0.8182	0.8206	0.8124

Table 4 The ablation study for cross-modal retrieval performance on Wikipedia dataset

Metric	Task	Method	N=10			N=20			N=30			N=50		
			A	B	C	A	B	C	A	B	C	A	B	C
mAP	Img2Txt	FedSCMR$_{\neg Lcs}$	0.4795	0.4813	0.4635	0.4748	0.4782	0.4628	0.4705	0.4733	0.4593	0.4657	0.4666	0.4542
		FedSCMR$_{\neg Lccs}$	0.4900	0.4882	0.4705	0.4853	0.4799	0.4744	0.4820	0.4758	0.4727	0.4761	0.4708	0.4693
		FedSCMR$_{\neg \Phi}$	0.4880	0.4767	0.4757	0.4814	0.4733	0.4755	0.4798	0.4684	0.4712	0.4763	0.4733	0.4592
		FedSCMR	0.4998	0.4997	0.5148	0.4942	0.4974	0.5107	0.4904	0.4958	0.5071	0.4889	0.4931	0.4867
	Txt2Img	FedSCMR$_{\neg Lcs}$	0.4300	0.4342	0.4025	0.3962	0.4028	0.3796	0.3784	0.3837	0.3625	0.3542	0.3520	0.3334
		FedSCMR$_{\neg Lccs}$	0.4503	0.5138	0.4519	0.4147	0.4594	0.4181	0.3932	0.4270	0.3915	0.3622	0.3880	0.3615
		FedSCMR$_{\neg \Phi}$	0.4490	0.5006	0.4559	0.4247	0.4591	0.4216	0.4022	0.4329	0.3978	0.3779	0.4038	0.3653
		FedSCMR	0.4516	0.5163	0.4811	0.4479	0.4678	0.4407	0.4364	0.4376	0.4129	0.4359	0.4371	0.4031
	Average	FedSCMR$_{\neg Lcs}$	0.4547	0.4577	0.4330	0.4355	0.4405	0.4212	0.4244	0.4285	0.4109	0.4100	0.4093	0.3938
		FedSCMR$_{\neg Lccs}$	0.4701	0.5010	0.4612	0.4500	0.4697	0.4463	0.4376	0.4514	0.4321	0.4192	0.4294	0.4154
		FedSCMR$_{\neg \Phi}$	0.4685	0.4887	0.4658	0.4531	0.4662	0.4486	0.4410	0.4506	0.4345	0.4271	0.4385	0.4123
		FedSCMR	0.4757	0.5080	0.4980	0.4710	0.4826	0.4757	0.4634	0.4667	0.4600	0.4624	0.4651	0.4449
NDCG	Img2Txt	FedSCMR$_{\neg Lcs}$	0.5087	0.5405	0.5145	0.5434	0.5705	0.5389	0.5740	0.5848	0.5702	0.6225	0.6403	0.6217
		FedSCMR$_{\neg Lccs}$	0.5347	0.5403	0.5220	0.5538	0.5669	0.5394	0.5794	0.5856	0.5674	0.6324	0.6405	0.6173
		FedSCMR$_{\neg \Phi}$	0.5266	0.4893	0.5205	0.5446	0.5148	0.5439	0.5741	0.5460	0.5737	0.6250	0.6124	0.6152
		FedSCMR	0.5381	0.5419	0.5692	0.5604	0.5712	0.5897	0.5871	0.5877	0.6091	0.6444	0.6416	0.6507
	Txt2Img	FedSCMR$_{\neg Lcs}$	0.5106	0.5242	0.5455	0.5188	0.5244	0.5459	0.5244	0.5241	0.5457	0.5288	0.5191	0.5433
		FedSCMR$_{\neg Lccs}$	0.5290	0.5368	0.5258	0.5353	0.5378	0.5349	0.5378	0.5362	0.5370	0.5381	0.5340	0.5377
		FedSCMR$_{\neg \Phi}$	0.5230	0.5301	0.5229	0.5324	0.5265	0.5295	0.5337	0.5245	0.5339	0.5318	0.5206	0.5328
		FedSCMR	0.5305	0.5477	0.5585	0.5448	0.5485	0.5589	0.5531	0.5437	0.5599	0.5588	0.5401	0.5606
	Average	FedSCMR$_{\neg Lcs}$	0.5097	0.5323	0.5300	0.5311	0.5474	0.5424	0.5492	0.5544	0.5579	0.5756	0.5797	0.5825
		FedSCMR$_{\neg Lccs}$	0.5318	0.5386	0.5239	0.5445	0.5523	0.5372	0.5586	0.5609	0.5522	0.5853	0.5873	0.5775
		FedSCMR$_{\neg \Phi}$	0.5248	0.5097	0.5217	0.5385	0.5207	0.5367	0.5539	0.5353	0.5538	0.5784	0.5665	0.5740
		FedSCMR	0.5343	0.5448	0.5638	0.5526	0.5598	0.5743	0.5701	0.5657	0.5845	0.6016	0.5908	0.6057

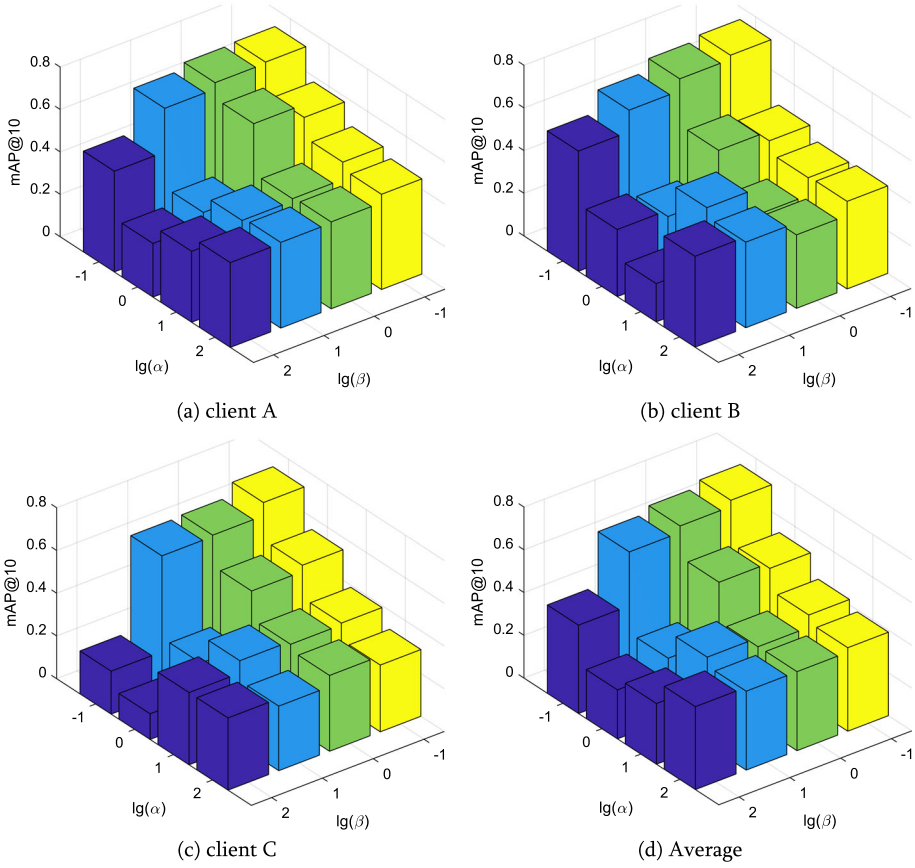


Figure 4 The mAP@10 performance of FedSCMR in the image-to-text retrieval task was evaluated on different clients with various parameters α and β using the Pascal dataset

incorporating modal discrimination and semantic invariance loss to achieve cross-space constraint enhancement is beneficial for cross-modal retrieval. The FedSCMR, equipped with adaptive federated aggregation capability, outperforms the FedSCMR$_{\neg\Phi}$, which aggregates federated models solely based on proportions. This suggests that dynamically adapting model aggregation according to the contribution of each model helps improve the performance of cross-modal retrieval models.

6.3 Impact of model parameters

Finally, to answer the third research question (*RQ3*), we investigated the impact of the parameters in (8) on the performance of the FedSCMR algorithm. Specifically, we tested the values of α and β, setting them to [0.1, 1, 10, 100], respectively. We then evaluated the mAP@10 of the FedSCMR algorithm on image-to-text retrieval and text-to-image retrieval tasks across three clients, while fixing the value of one parameter and varying the value of the other. The experimental results are presented in Figures 4 and 5. We find that FedSCMR performs best

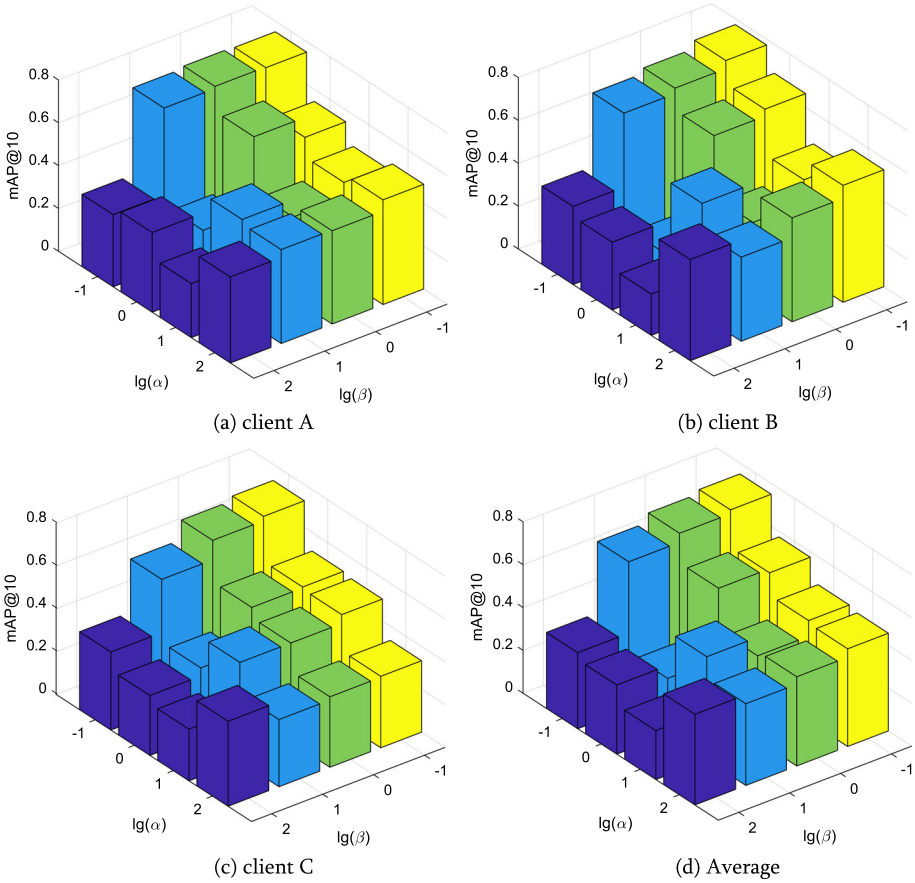


Figure 5 The mAP@10 performance of FedSCMR in the text-to-image retrieval task was evaluated on different clients with various parameters α and β using the Pascal dataset

on image-to-text retrieval and text-to-image retrieval tasks across all clients when α and β are both set to 0.1.

To investigate the impact of the parameter γ in (13) on FedSCMR, we conducted experiments by setting γ to different values, specifically [1, 5, 10, 15, 20, 30, 50, 100]. The objective was to observe how varying γ influenced the performance of FedSCMR across image-to-text retrieval, text-to-image retrieval, and average retrieval tasks on three different clients. The results are presented in Figure 6. As seen from Figure 6, with the change in γ, the overall performance of FedSCMR on all three clients remains relatively stable. When γ is set to 30, FedSCMR achieves the highest mAP@10 value, indicating that this value is the most suitable one for achieving optimal performance.

Figure 7 demonstrates the variation of the overall loss of FedSCMR on three clients as the federated training epoch changes. From Figure 7, we find that the loss values decrease rapidly, demonstrating that the proposed FedSCMR algorithm is easy to train. The use of the Frobenius norm as the loss function ensures that the gradient is continuous, enabling faster training.

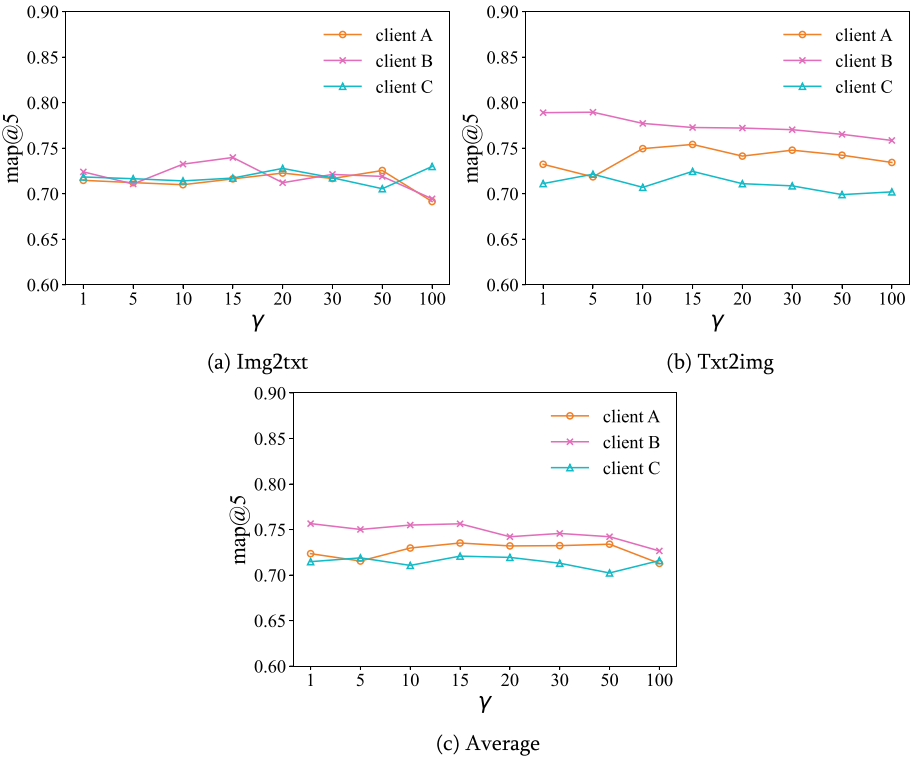


Figure 6 The mAP@10 Performance of FedSCMR across image-to-text retrieval, text-to-image retrieval, and average retrieval tasks was evaluated on different clients with various parameter γ using the Pascal dataset

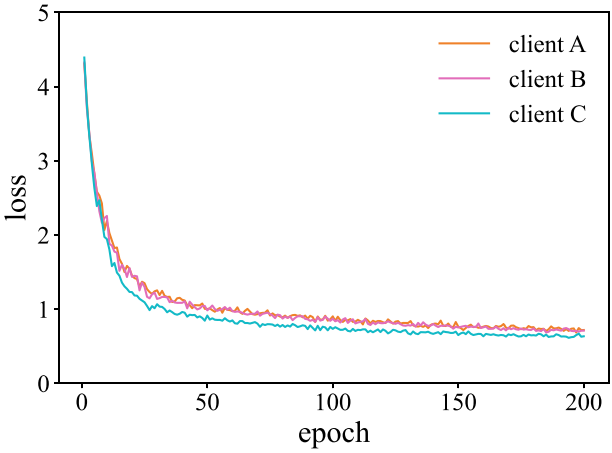


Figure 7 Loss of FedSCMR in three clients with training epoch varies

7 Conclusion

We proposed Federated Supervised Cross-Modal Retrieval (FedSCMR) method. FedSCMR is designed as a comprehensive approach to supervised cross-modal retrieval, with the overarching goal of learning an optimal common subspace by constructing multiple competing subspaces. One of the distinguishing features of FedSCMR is its adaptive model aggregation during the federated aggregation stage. Rather than a one-size-fits-all approach, FedSCMR intelligently aggregates models based on their individual contributions. This adaptability ensures that the federated learning process is finely tuned to the characteristics and strengths of each participating model. In constructing a local cross-modal retrieval model, FedSCMR takes into account various critical factors. These include discrimination in label space, ensuring that the model is capable of distinguishing between different labels effectively. Semantic discrimination and consistency in common subspaces are also key considerations, allowing the model to capture meaningful semantic relationships across modalities. Additionally, FedSCMR addresses modal discrimination and semantic invariance across common subspaces, contributing to a more robust and versatile retrieval model. The federated aggregation phase in FedSCMR is a dynamic and collaborative process. The model contributions are evaluated based on multiple factors, including data volume, data category, loss, and mean Average Precision (mAP). This comprehensive evaluation process aims to determine the optimal aggregation strategy for achieving superior cross-modal retrieval performance. Experimental results have substantiated the effectiveness and superiority of FedSCMR over existing state-of-the-art cross-modal retrieval methods. This confirmation through empirical evaluation underscores the potential of FedSCMR to significantly advance the capabilities of cross-modal retrieval model. The empirical improvement achieved by the proposed FedSCMR signifies the effectiveness of leveraging competition to learn the optimal common subspace, and adaptively aggregating the common subspaces of multiple clients for dynamic global aggregation.

Author Contributions Ang Li developed and designed the FedSCMR method, and drafted the methodology. Yawen Li led experimental design, and conducted an extensive literature review. Yingxia Shao provided expertise in privacy preservation and federated learning, and contributed to experiment design.

Funding This work was supported by the Program of the National Natural Science Foundation of China (62192784, U22B2038, 62172056), by the 8th Young Elite Scientists Sponsorship Program by CAST (2022QNRC001), and by the Postgraduate Education and Teaching Reform Project of Beijing University of Posts and Telecommunications (2023Y028).

Data Availability The data that support the results of this study are openly available at https://github.com/rupy/PascalSentenceDataset and http://www.svcl.ucsd.edu/projects/crossmodal.

Declarations

Ethics approval and consent to participate This article does not contain any studies involving human participants and/or animals by any of the authors.

Consent for publication All authors have read and agreed to the published version of the manuscript.

Human and animal Ethics Not applicable.

Competing interests The authors declare no competing interests.

References

1. Hu, P., Huang, Z., Peng, D., Wang, X., Peng, X.: Cross-modal retrieval with partially mismatched pairs. IEEE Trans. Pattern Anal. Mach. Intell., 1–15 (2023)
2. Li, Y., Yuan, Y., Wang, Y., Lian, X., Ma, Y., Wang, G.: Distributed multimodal path queries. IEEE Trans. Knowl. Data Eng. **34**(7), 3196–3210 (2022)
3. Yuehua, Y., Junping, D., Yuan, P.: Ontology-based intelligent information retrieval system. J Softw. **26**(7), 1675–1687 (2015)
4. Li, A., Du, J., Kou, F., Xue, Z., Xu, X., Xu, M., Jiang, Y.: Scientific and technological information oriented semantics-adversarial and media-adversarial cross-media retrieval. arXiv:2203.08615 (2022)
5. Liang, M., Du, J., Yang, C., Xue, Z., Li, H., Kou, F., Geng, Y.: Cross-media semantic correlation learning based on deep hash network and semantic expansion for social network cross-media search. IEEE Trans. Neural Netw. Learn. Syst. **31**(9), 3634–3648 (2020)
6. Liang, M., Du, J., Liu, W., Xue, Z., Geng, Y., Yang, C.: Fine-grained cross-media representation learning with deep quantization attention network. In: Proceedings of the 27th ACM International Conference on Multimedia. MM '19, pp. 1313–1321 (2019)
7. Rodríguez-Barroso, N., Jiménez-López, D., Luzón, M.V., Herrera, F., Martínez-Cámara, E.: Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges. Inf. Fusion **90**, 148–173 (2023)
8. Yin, W., Yin, H., Baraka, K., Kragic, D., Björkman, M.: Dance style transfer with cross-modal transformer. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 5058–5067 (2023)
9. Żelaszczyk, M., Mańdziuk, J.: Cross-modal text and visual generation: A systematic review. part 1: Image to text. Inf. Fusion **93**, 302–329 (2023)
10. Shi, L., Luo, J., Zhu, C., Kou, F., Cheng, G., Liu, X.: A survey on cross-media search based on user intention understanding in social networks. Inform. Fusion **91**, 566–581 (2023)
11. Cao, X., Shi, Y., Wang, J., Yu, H., Wang, X., Yan, Z.: Cross-modal knowledge graph contrastive learning for machine learning method recommendation. In: Proceedings of the 30th ACM International Conference on Multimedia. MM '22, pp. 3694–3702 (2022)
12. Liang, M., Du, J., Cao, X., Yu, Y., Lu, K., Xue, Z., Zhang, M.: Semantic structure enhanced contrastive adversarial hash network for cross-media representation learning. In: Proceedings of the 30th ACM International Conference on Multimedia. MM '22, pp. 277–285 (2022)
13. Zhang, P., Bai, G., Yin, H., Huang, Z.: Proactive privacy-preserving learning for cross-modal retrieval. ACM Trans. Inf. Syst. **41**(2), 1–23 (2023)
14. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: An overview with application to learning methods. Neural Comput. **16**(12), 2639–2664 (2004)
15. Hwang, S.J., Grauman, K.: Learning the relative importance of objects from tagged images for retrieval and cross-modal search. Int. J. Comput. Vision **100**, 134–153 (2012)
16. Xiao, S., Shao, Y., Li, Y., Yin, H., Shen, Y., Cui, B.: LECF: Recommendation via learnable edge collaborative filtering. Sci. China Inf. Sci. **65**(1), 1–15 (2022)
17. Cao, T., Xu, C., Du, J., Li, Y., Xiao, H., Gong, C., Zhong, L., Niyato, D.: Reliable and efficient multimedia service optimization for edge computing-based 5G networks: Game theoretic approaches. IEEE Trans. Netw. Serv. Manage. **17**(3), 1610–1625 (2020)
18. Li, A., Li, Y., Shao, Y., Liu, B.: Multi-view scholar clustering with dynamic interest tracking. IEEE Trans. Knowl. Data Eng. **35**(9), 9671–9684 (2023)
19. Shao, Y., Huang, S., Li, Y., Miao, X., Cui, B., Chen, L.: Memory-aware framework for fast and scalable second-order random walk over billion-edge natural graphs. VLDB J. **30**(5), 769–797 (2021)
20. Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: Proceedings of the 30th International Conference on Machine Learning, pp. 1247–1255 (2013)
21. Wang, W., Arora, R., Livescu, K., Bilmes, J.: On deep multi-view representation learning. In: Proceedings of the 32nd International Conference on Machine Learning, pp. 1083–1092 (2015)
22. He, Y., Xiang, S., Kang, C., Wang, J., Pan, C.: Cross-modal retrieval via deep and bidirectional representation learning. IEEE Trans. Multimedia **18**(7), 1363–1377 (2016)
23. Zhen, L., Hu, P., Wang, X., Peng, D.: Deep supervised cross-modal retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10394–10403 (2019)
24. Li, Y., Zeng, I.Y., Niu, Z., Shi, J., Wang, Z., Guan, Z.: Predicting vehicle fuel consumption based on multi-view deep neural network. Neurocomputing **502**, 140–147 (2022)
25. Li, Y., Jiang, D., Lian, R., Wu, X., Tan, C., Xu, Y., Su, Z.: Heterogeneous latent topic discovery for semantic text mining. IEEE Trans. Knowl. Data Eng. **35**(1), 533–544 (2023)

26. Huang, J., Wang, H., Sun, Y., Fan, M., Huang, Z., Yuan, C., Li, Y.: HGAMN: Heterogeneous graph attention matching network for multilingual POI retrieval at Baidu maps. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 3032–3040 (2021)
27. McMahan, B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, pp. 1273–1282 (2017)
28. Konečný, J., McMahan, H.B., Yu, F.X., Richtárik, P., Suresh, A.T., Bacon, D.: Federated learning: Strategies for improving communication efficiency. arXiv:1610.05492 (2016)
29. Wang, H., Yurochkin, M., Sun, Y., Papailiopoulos, D.S., Khazaeni, Y.: Federated learning with matched averaging. arXiv:2002.06440 (2020)
30. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. In: Proceedings of Machine Learning and Systems, pp. 429–450 (2020)
31. Guan, Z., Li, Y., Xue, Z., Liu, Y., Gao, H., Shao, Y.: Federated graph neural network for cross-graph node classification. In: 2021 IEEE 7th International Conference on Cloud Computing and Intelligent Systems (CCIS), pp. 418–422 (2021)
32. Zong, L., Xie, Q., Zhou, J., Wu, P., Zhang, X., Xu, B.: FedCMR: Federated cross-modal retrieval. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1672–1676 (2021)
33. Li, Y., Li, W., Xue, Z.: Federated learning with stochastic quantization. International Journal of Intelligent Systems (2022)
34. Guan, Z., Li, Y., Pan, Z., Liu, Y., Xue, Z.: RFDG: Reinforcement federated domain generalization. IEEE Trans. Knowl. Data Eng., 1–14 (2023)
35. Zang, Y., Xue, Z., Ou, S., Long, Y., Zhou, H., Du, J.: FedPcF: An integrated federated learning framework with multi-level prospective correction factor. In: Proceedings of the 2023 ACM International Conference on Multimedia Retrieval, pp. 490–498 (2023)
36. Long, Y., Xue, Z., Chu, L., Zhang, T., Wu, J., Zang, Y., Du, J.: FedCD: A classifier debiased federated learning framework for non-IID data. In: Proceedings of the 31st ACM International Conference on Multimedia (2023)
37. Rashtchian, C., Young, P., Hodosh, M., Hockenmaier, J.: Collecting image annotations using Amazon's mechanical Turk. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pp. 139–147 (2010)
38. Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G.R., Levy, R., Vasconcelos, N.: A new approach to cross-modal multimedia retrieval. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 251–260 (2010)
39. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
40. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 (2018)
41. Ryffel, T., Trask, A., Dahl, M., Wagner, B., Mancuso, J., Rueckert, D., Passerat-Palmbach, J.: A generic framework for privacy preserving deep learning. arXiv:1811.04017 (2018)