Federated Vision-Language-Recommendation with Personalized Fusion

Zhiwei Li[1], **Guodong Long**[1], **Jing Jiang**[1], **Chengqi Zhang**[2], **Qiang Yang**[2]

[1] Australian AI Institute, Faculty of Engineering and IT, University of Technology Sydney
[2] Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University
zhw.li@outlook.com, {guodong.long, jing.jiang}@uts.edu.au, {chengqi.zhang, profqiang.yang}@polyu.edu.hk

Abstract

Applying large pre-trained Vision-Language Models to recommendation is a burgeoning field, a direction we term Vision-Language-Recommendation (VLR). Bringing VLR to user-oriented on-device intelligence within a federated learning framework is a crucial step for enhancing user privacy and delivering personalized experiences. This paper introduces FedVLR, a federated VLR framework specially designed for user-specific personalized fusion of vision-language representations. At its core is a novel bi-level fusion mechanism: The server-side multi-view fusion module first generates a diverse set of pre-fused multimodal views. Subsequently, each client employs a user-specific mixture-of-expert mechanism to adaptively integrate these views based on individual user interaction history. This designed lightweight personalized fusion module provides an efficient solution to implement a federated VLR system. The effectiveness of our proposed FedVLR has been validated on seven benchmark datasets.

Code — https://github.com/mtics/FedVLR
Extended Version — https://arxiv.org/abs/2410.08478

Introduction

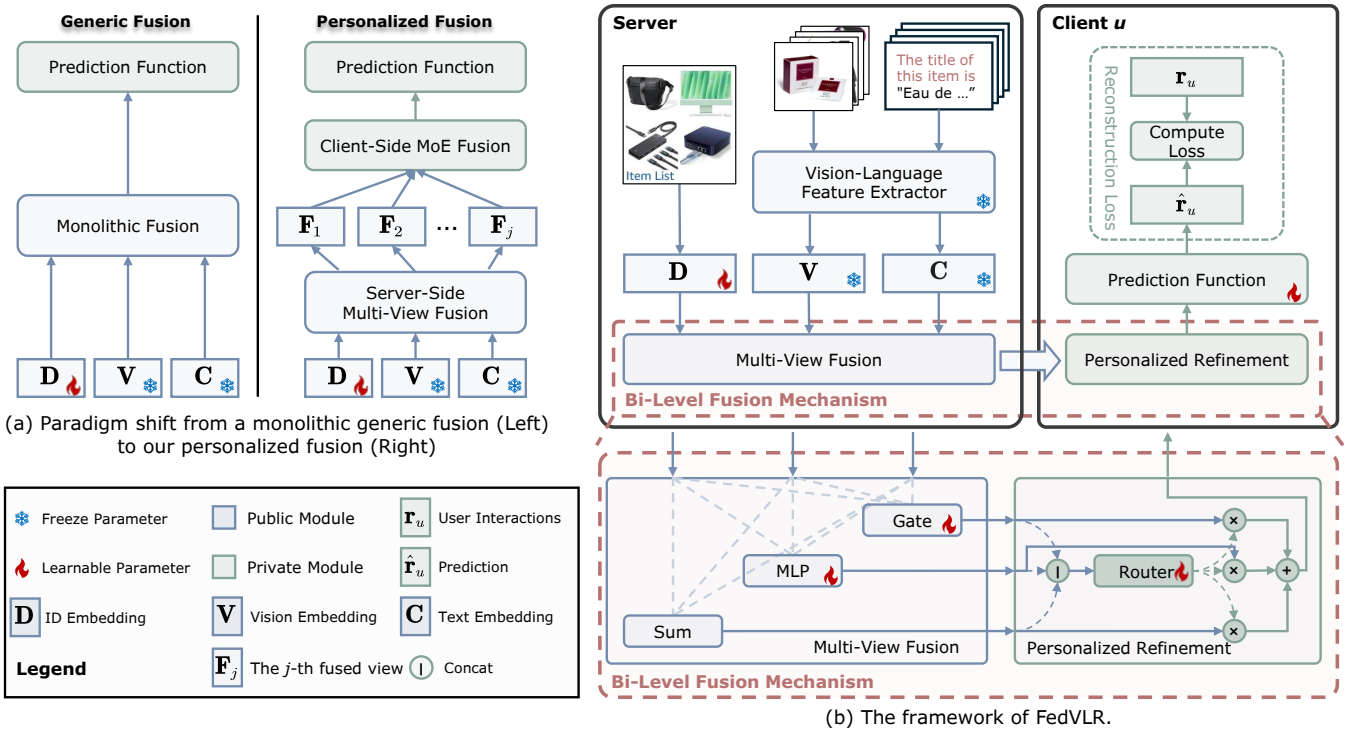
Vision-Language Models (VLMs) are pushing the boundaries of personalized recommendation by interpreting the rich content of items (Wei et al. 2024), a direction we conceptualize as Vision-Language-Recommendation (VLR). By understanding visual aesthetics and textual semantics, VLR models can move beyond simple item identifiers (IDs) to capture a deeper, more nuanced understanding of user preferences (Zhou et al. 2023a). Deploying these powerful VLR models directly on user devices is a significant step forward, as this on-device approach enhances user privacy, reduces network latency, and grants users direct ownership of their data (Yin et al. 2024), aligning with current privacy-centric principles (Voigt and Von dem Bussche 2017).

This imperative has catalyzed the burgeoning field of Federated Vision-Language Models (FedVLMs), which aims to train powerful VLMs on decentralized data without compromising privacy (Ren et al. 2024). Initial research in FedVLMs has primarily focused on foundational challenges, such as adapting large model architectures to the federated setting (Liu et al. 2020), and mitigating prohibitive communication costs (Guo et al. 2023). While these efforts are vital, they often overlook a more subtle challenge specific to the on-device recommendation tasks: how to fuse vision-language signals in an user-oriented personalized way.

Consider the process of choosing a movie. What factors contribute to the decision? One user might be attracted by a beautiful movie poster, another might be impacted by the story described in the text summary, while a third might rely on collaborative signals from friends with similar tastes (Li et al. 2023; Liu et al. 2024a,b; Ma et al. 2025; Li et al. 2025a; Zhang et al. 2025b). These factors, spanning visual, textual, and collaborative signals, contribute to each user's final decision in a highly individualized manner (Wei et al. 2024), and points to a challenge deeper than just the statistical heterogeneity of data, i.e., *preference heterogeneity*. We define this as the phenomenon where users exhibit diverse and personal criteria when evaluating and weighing information from different modalities. Inspired by this, we believe that a personalized multi-modal fusion module is the critical component needed to enhance federated VLR by capturing these fine-grained user preferences across all modalities.

This diversity means that a single and one-size-fits-all module for fusing multimodal signals is inherently suboptimal. Yet, existing federated recommendation systems often fail to address this. They are either content-agnostic, relying only on interaction IDs (Lin et al. 2020; Zhang et al. 2023b), or they impose a globally uniform fusion logic on all users (Li et al. 2024; Feng et al. 2024). They neglect the critical need for the fusion module itself to be personalized.

To address this gap, we propose a novel **Fed**erated **V**ision-**L**anguage-**R**ecommendation framework with Personalized Fusion (**FedVLR**). Our framework learns under a standard federated recommendation setting, where all item features are stored on the server, while user interaction histories remain privately on each client's device (Zhang et al. 2024c,b), respecting data ownership and mitigates privacy risks. The core innovation lies in a personalized and dynamic fusion strategy, realized through our proposed Bi-Level Fusion Mechanism (BLFM). As shown in Fig. 1a, instead of imposing a generic fusion logic on all users through a shared single module (Left), our BLFM enables personalized fusion by decoupling the fusion into two levels (Right). The server firstly generates multiple feature views using diverse



(a) Paradigm shift from a monolithic generic fusion (Left) to our personalized fusion (Right)

(b) The framework of FedVLR.

Figure 1: The overall framework of FedVLR, detailing: (a) The paradigm shift enables fine-grained on-device personalization by decoupling server-side view generation from client-side refinement; (b) FedVLR comprises two components: (1) **Server-Side Multi-View Fusion**, which generates diverse pre-fused feature views from visual-language content, and (2) **Client-Side Personalized Refinement**, which dynamically combines these views based on the user's private interaction history.

operators, handling the major computational load. Subsequently, each client learns a lightweight, personalized refinement over these views using a Mixture-of-Experts (MoE) module based on user's private history. This architecture maximizes the utility of server-side operators to deliver personalization in a lightweight manner, effectively addressing preference heterogeneity in the federated VLR settings.

Our main contributions are summarized as follows:

- We are the first to formalize and tackle the personalized multimodal fusion problem in federated VLR settings.
- Our proposed Bi-Level Fusion Mechanism contributes a new paradigm for modality fusion that can be generalized to the broader domain of multimodal learning, enabling fine-grained multimodal personalization.
- We design FedVLR as a versatile framework that seamlessly enhances a wide range of existing ID-based federated models with content-aware personalization.
- Extensive empirical validation shows that FedVLR not only substantially improves existing baselines but can also outperform centralized models in certain low-data regimes. Source code is provided for reproducibility.

Related Work

Vision-Language-Recommendation Models

A prominent direction in representation learning is translating multimodal perception into personalized decisions (Zhang et al. 2020; Liu et al. 2024a). Among these modalities, vision and language are particularly powerful for creating rich item representations (Wei et al. 2019; Yu et al. 2023; Zhou et al. 2023a; Ren et al. 2024; Malitesta et al. 2025). We conceptualize the specific task of leveraging them for recommendation as Vision-Language-Recommendation (VLR). VLR models aim to move beyond simple interaction data by interpreting the actual content of items, which can alleviate data sparsity and capture a more nuanced understanding of user preferences (Liu et al. 2019; Ren et al. 2024; Wei et al. 2024; Yu et al. 2025; Zhou et al. 2025).

The typical approach in a centralized setting involves a two-stage process. First, powerful pre-trained foundation models are used to extract high-level semantic features from item images and textual descriptions (Zhang et al. 2021; Hou et al. 2022; Bian et al. 2023; Geng et al. 2023; Sun et al. 2023; Wang et al. 2023; Lu et al. 2023; Fu et al. 2024; Pan, Huang, and Shi 2024; Zhou et al. 2025). Second, a dedicated fusion module combines these unimodal representations into a single comprehensive embedding for each item (Radford et al. 2021b; Zhang et al. 2021; Liu et al. 2024b; Wei et al. 2024; Zhang et al. 2024b). This fused representation is then used to predict user preferences. While effective, these centralized models require unrestricted access to all user and item data, which motivates our work to adapt these sophisticated VLR techniques to a privacy-preserving environment where data remains decentralized on devices.

Federated Learning for On-Device Personalization

On-device recommendation involves training models on a user's local device to preserve privacy (Li, Long, and Zhou 2024; Yin et al. 2024). Federated Learning (FL) (McMahan et al. 2017) provides the foundational framework, enabling collaborative training without centralizing sensitive data. In the context of recommendation, this creates a one-user-per-client setting for personalization (Zhang et al. 2023b,a,c). Incorporating rich item content into this setting introduces a key personalization challenge. Users weigh modalities like vision and text differently when making choices (Liu et al. 2019; Niu, Zhong, and Yu 2021; Lei et al. 2023). This diversity makes a single global modal fusion module across all users suboptimal, as a fixed rule for all users fails to capture their nuanced individual tastes for different modality (Zhang et al. 2024a; Yuan et al. 2024; Kong et al. 2025).

Most on-device recommendation models are content-agnostic (Ammad-Ud-Din et al. 2019; Lin et al. 2020; Liang, Pan, and Ming 2021; Luo, Xiao, and Song 2022; Perifanis and Efraimidis 2022; Zhang et al. 2023b,a, 2025a), relying solely on user-item interaction IDs. While these methods address statistical heterogeneity from non-IID data (Allouah et al. 2023; Li, Long, and Zhou 2024; Yuan et al. 2024), their dependence on IDs prevents the integration of rich item content. A few recent studies have incorporated multiple modalities of content (Li et al. 2024; Feng et al. 2024), but they typically impose a uniform fusion logic on all users. Therefore, developing a user-oriented fusion module that can be personalized on each user's device remains a critical open problem in content-aware on-device recommendation.

Problem Formulation

We consider the user-oriented federated setting for on-device VLR tasks, where the system consists of a set of users $\mathcal{U}$ and a set of items $\mathcal{I}$. Each user $u \in \mathcal{U}$ acts as a distinct client, holding their private interaction data. We represent the historical interactions as a binary matrix $\mathbf{R} \in \{0,1\}^{|\mathcal{U}|\times|\mathcal{I}|}$, where $r_{u,i} = 1$ signifies that user u has interacted with item i. The set of observed interactions for user u is $\mathcal{O}_u = \{i \in \mathcal{I} \mid r_{u,i} = 1\}$. Each item $i \in \mathcal{I}$ is described by features from a set of modalities $\mathcal{M}$. For a given modality $m \in \mathcal{M}$, the features for all items are represented by a matrix $\mathbf{E}_m \in \mathbb{R}^{|\mathcal{I}|\times d_m}$, where $\mathbf{e}_{i,m} \in \mathbb{R}^{d_m}$ is the feature vector for item i. Visual (v) and textual (c) modalities all are stored on the server to ensure client-side efficiency.

Methodology

This section presents our proposed FedVLR framework, with its overall architecture depicted in Fig. 1b. The core of FedVLR lies in a novel Bi-Level Fusion Mechanism that learns across the server and clients to achieve on-device VLR within the user-oriented federated framework.

Preliminary

As depicted on the server-side of Fig. 1b, FedVLR's initial step is to prepare a comprehensive set of item representations. To avoid burdening user devices, this process is handled entirely by the server. A frozen pre-trained VLM $h(\cdot)$ is employed in a one-time process to transform raw visual features $\mathbf{E}_v$ and textual features $\mathbf{E}_c$ into high-level semantic embeddings, yielding the vision and text embedding $\mathbf{V}$ and $\mathbf{C}$:

$$\mathbf{V} = h(\mathbf{E}_v), \quad \mathbf{C} = h(\mathbf{E}_c)\,. \tag{1}$$

Alongside these static content embeddings, FedVLR also maintains a globally learnable ID embedding $\mathbf{D} \in \mathbb{R}^{|\mathcal{I}|\times d}$, designed to capture collaborative signals from user-item interactions. Collectively, the set $\{\mathbf{D}, \mathbf{V}, \mathbf{C}\}$ provides a multi-view representation for all items available to the system.

Bi-Level Fusion Mechanism

To address both statistical and preference heterogeneity, FedVLR introduces the Bi-Level Fusion Mechanism (BLFM), which learns at two coordinated levels: server-side multi-view fusion and client-side personalized refinement.

Server-Side Multi-View Fusion. Instead of modeling a single unified item representation, the server generates a diverse set of fused feature views by employing a collection of distinct learnable fusion operators $\mathcal{G} = \{g_j\}_{j=1}^{|\mathcal{G}|}$. Each operator g_j takes the full set of item representations $\{\mathbf{V}, \mathbf{C}, \mathbf{D}\}$ as input to produce a unique fused view $\mathbf{F}_j$:

$$\mathbf{F}_j = g_j\big(\mathbf{V}, \mathbf{C}, \mathbf{D};\, \gamma_j\big) \tag{2}$$

where γ_j are the learnable parameters of the j-th fusion operator. As shown in the lower panel of Fig. 1b, these operators can include simple strategies like element-wise *Sum*, as well as more complex parameterized functions like a multi-layer perceptron *MLP* or a *Gate* mechanism. The fused views $\{\mathbf{F}_j\}$ are then broadcast to the client devices.

Client-Side Personalized Refinement. Upon receiving the set of pre-fused views, each client u performs personalized refinement via a MoE module, allowing FedVLR to adapt to the individual user's preferences. As shown in Fig. 1b, the client employs a local lightweight *Router* module ϕ_u, parameterized by the local parameters φ_u. It takes the server-provided item views $\{\mathbf{F}_j\}$ and the user's history $\mathbf{r}_u$ as input to dynamically compute a set of importance weights $\mathbf{w}_u$:

$$\mathbf{w}_u = \text{softmax}\left(\phi_u\left(\{\mathbf{F}_j\}_{j=1}^{|\mathcal{G}|},\, \mathbf{r}_u;\, \varphi_u\right)\right). \tag{3}$$

Each weight $w_{u,j}$ in the vector $\mathbf{w}_u \in \mathbb{R}^{|\mathcal{G}|}$ quantifies how much user u values the j-th feature view. The client then models its final personalized item representation $\bar{\mathbf{F}}^{(u)}$ by computing a weighted sum of the global views:

$$\bar{\mathbf{F}}^{(u)} = \sum_{j=1}^{|\mathcal{G}|} w_{u,j}\, \mathbf{F}_j. \tag{4}$$

Objective Function

With the personalized item representation $\bar{\mathbf{F}}^{(u)}$ for user u, the preference score $\hat{r}_{ui}$ for an item i is computed by a local prediction function f, parameterized by local parameters θ_u:

$$\hat{r}_{ui} = f(\bar{\mathbf{F}}_i^{(u)};\, \theta_u). \tag{5}$$

The training process aims to find the optimal set of parameters $\mathbf{\Theta} = (\{\mathbf{D}, \{\gamma_j\}\}, \{\theta_u, \varphi_u\})$ for minimizing the reconstruction loss between the predicted scores $\hat{r}_{ui}$ and the user

interactions r_{ui}, aggregated across all users. The joint optimization problem of FedVLR is then formulated as follows:

$$\begin{aligned} \min_{\Theta} \quad & \mathcal{J}(\Theta) = \sum_{u \in \mathcal{U}} \alpha_u \mathcal{J}_u(\Theta), \\ \text{s.t.} \quad & \mathbb{E}_{\xi}\left[\|g_u(\Theta) - \nabla \mathcal{J}_u(\Theta)\|^2\right] \le \sigma^2, \\ & \mathbb{E}_{u}\left[\|\nabla \mathcal{J}_u(\Theta) - \nabla \mathcal{J}(\Theta)\|^2\right] \le \zeta^2. \end{aligned} \tag{6}$$

In Eq. (6), we assume that each client's objective, denoted by $\mathcal{J}_u(\mathbf{\Theta}) = \mathcal{L}_u(\hat{r}_{ui}, r_{ui})$, belongs to the class of L-smooth functions $\mathcal{F}_L$. Furthermore, the optimization learns under the standard conditions of bounded variance σ^2 for stochastic gradients on each client and ζ^2 for the gradient dissimilarity across the population of clients, which are standard prerequisites for establishing convergence in federated training. The procedure for optimizing this objective is detailed in Alg. 1.

Theoretical Analysis

Convergence Analysis

In each communication round t, clients perform A local updates, which inevitably introduces a drift between the locally trained parameters and the global model state. Despite this drift, by building on the standard assumptions for the objective function in Eq. (6), we can guarantee that our proposed FedVLR converges to a stationary point of $\mathcal{J}$ in a manner consistent with rates of typical non-convex FL problems:

Theorem 1 (Convergence of FedVLR)**.** *Let the number of participating clients per round be n_s. After T communication rounds with learning rate η, the algorithm satisfies:*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla \mathcal{J}(\mathbf{\Theta}^t)\|^2\right] \le O\left(\frac{1}{\sqrt{T}}\right) + O\left(\frac{A}{T}\right) + O\left(\frac{\zeta^2}{n_s}\right). \tag{7}$$

Theorem 1 demonstrates that FedVLR is theoretically sound, achieving a standard federated convergence rate, validating that our personalized fusion architecture can effectively address preference heterogeneity without impeding convergence. The detailed proof is provided in the appendix.

Complexity Analysis

Here, we analyze the computational and storage costs of FedVLR. The server stores the global item embeddings, including the ID embeddings $\mathbf{D}$ and the pre-computed multimodal embeddings $\mathbf{V}$ and $\mathbf{C}$, requiring $O(|\mathcal{I}|d)$ space, where d is the embedding dimension. In each communication round, the server's main computational cost comes from generating the $|\mathcal{G}|$ fused views, resulting in a time complexity of $O(|\mathcal{G}||\mathcal{I}|d)$. Each client stores its local model parameters K. The primary storage overhead is from receiving the $|\mathcal{G}|$ feature views, requiring $O(|\mathcal{G}||\mathcal{I}|d + K)$ space temporarily during training. The client's per-iteration time complexity for local updates, involving the prediction function f and the BLFM router ϕ_u, is denoted as $O(P)$. Therefore, the total system space complexity is roughly $O(|\mathcal{U}|(|\mathcal{G}||\mathcal{I}|d + K) + |\mathcal{I}|d)$, and the total time complexity per round involves server computation plus aggregated client computation, scaling approximately as $O(|\mathcal{G}||\mathcal{I}|d + n_s AP)$ after performing A local update steps. Therefore, FedVLR adds a manageable overhead that is well-justified by its ability to model complex, personalized multimodal preferences.

Algorithm 1: The training algorithm for FedVLR

Input: User Interaction History $\mathbf{R}$, Learning Rate η, Number of Communication Round T, Number of Local Training Epoch A
Initialize: User Embeddings $\{\theta_u\}$, User-specific Parameters $\{\varphi_u\}$, Strategy Parameters $\{\gamma_j\}$, ID Embeddings $\mathbf{D}$
MultiStrategyFusion:

$\mathbf{V} \leftarrow h(\mathbf{E}_v)$, $\mathbf{C} \leftarrow h(\mathbf{E}_c)$ according to Eq. (1);
for $t = 1, 2, \ldots, T$ **do**
  $S_t \leftarrow$ randomly select n_s from n clients;
  Compute $\mathbf{F}_j$ according to Eq. (2), $j = 1, 2, \cdots |\mathcal{G}|$;
  for all client index $u \in S_t$ **do**;
    $\hat{\mathbf{r}}_u, \nabla_{\mathbf{D}}^{(u)}, \{\nabla_j^{(u)}\gamma\} \leftarrow \text{ClientUpdate}(\{\mathbf{F}_j\}_{j=1}^{|\mathcal{G}|})$;
  end for
  $\mathbf{D} \leftarrow \mathbf{D} - \eta \sum_{u \in S_t} \alpha_u \nabla_{\mathbf{D}}^{(u)}$;
  $\gamma_j \leftarrow \gamma_j - \eta \sum_{u \in S_t} \alpha_u \nabla_j^{(u)}\gamma$, $j = 1, 2, \cdots, |\mathcal{G}|$;
end for
return: $\hat{\mathbf{R}} = [\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_n]^T$

PersonalizedRefinement:

$\nabla_{\mathbf{D}}^{(u)} \leftarrow 0$;
$\nabla_j^{(u)}\gamma \leftarrow 0$, $j = 1, 2, \cdots |\mathcal{G}|$;
for $a = 1, 2, \ldots, A$ **do**
  Compute ∇_{θ_u}, ∇_{φ_u}, $\nabla_{\mathbf{D}}$ and $\{\nabla_{\gamma_j}\}$ according to Eq. (6);
  Update local prediction parameters $\theta_u \leftarrow \theta_u - \eta \nabla_{\theta_u}$;
  Update local router parameters $\varphi_u \leftarrow \varphi_u - \eta \nabla_{\varphi_u}$;
  Accumulate gradients $\nabla_{\mathbf{D}}^{(u)} \leftarrow \nabla_{\mathbf{D}}^{(u)} + \nabla_{\mathbf{D}}$;
  Accumulate $\nabla_j^{(u)}\gamma \leftarrow \nabla_j^{(u)}\gamma + \nabla_{\gamma_j}$, $j = 1, 2, \cdots |\mathcal{G}|$;
  Compute $\hat{\mathbf{r}}_u$ according to Eq. (5);
end for
return: $\hat{\mathbf{r}}_u, \nabla_{\mathbf{D}}^{(u)}, \{\nabla_j^{(u)}\gamma\}$

Privacy Preservation

FedVLR ensures privacy based on the foundational FL principle of data localization, where raw user interactions $\mathbf{r}_u$ remain on-device. Crucially, all personalization components, including the user-specific router ϕ_u and its parameters φ_u, are kept entirely local. This design ensures individual modality preferences are never exposed. The gradients transmitted to the server are structurally analogous to those in standard ID-based frameworks (Li, Long, and Zhou 2024; Zhang et al. 2023b; Li et al. 2025b; Feng et al. 2024; Wu et al. 2024), thus introducing no new attack surfaces. Furthermore, FedVLR is compatible with advanced privacy-enhancing technologies, which our experiments show can be incorporated with an acceptable performance trade-off.

Experiments

Datasets

We conduct a comprehensive evaluation on seven public datasets spanning three distinct domains shown in Table 1:

- **Amazon Review Datasets**[1] (Hou et al. 2024): Two review datasets are selected for the e-commerce domain, i.e., All_Beauty (Beauty) and Gift_Cards (Cards).

[1] https://amazon-reviews-2023.github.io/

Dataset	#Users	#Items	#Ratings	Sparsity
Beauty	253	356	2,535	97.19%
Cards	377	129	2,429	95.01%
ML	610	3,650	90,274	95.95%
KU	204	560	3,488	96.95%
Dance	10,231	1,676	80,086	99.53%
Food	5,990	1,125	36,482	99.46%
Movie	15,908	2,528	111,091	99.72%

Table 1: The statistical information of the datasets used.

- **MovieLens-latest-small (ML)**[2] (Harper and Konstan 2015): A classic dataset for the movie recommendation.
- **NineRec Datasets**[3] (Zhang et al. 2024d): Four datasets from the short-video domain are selected: KU, Bili_Food (Food), Bili_Dance (Dance), and Bili_Movie (Movie).

This diverse selection across different domains, platforms, and scales provides a robust testbed for assessing FedVLR's adaptability and generalizability under realistic and extreme sparse data conditions (over 95%) in the user-oriented federated settings. Following standard practice, we filter out users with fewer than 5 interactions to mitigate cold-start issues. For data preprocessing, missing images are imputed using the average visual features of existing items, and missing titles are replaced with the placeholder "The title is missing."

Experimental Setup

We evaluate FedVLR against two categories: (1) centralized multimodal models and (2) ID-based federated frameworks.

Centralized VLR Baselines serve as performance references with full data access in a non-federated environment:

- **VBPR** (He and McAuley 2016): A Bayesian personalized ranking method incorporating visual features;
- **BM3** (Zhou et al. 2023b): A self-supervised recommendation framework for multimodal data integration;
- **MGCN** (Yu et al. 2023): A model that employs multi-view graph convolutions to fuse item's modal features;
- **PGL** (Yu et al. 2025): Mines local graph structures of the user-item interaction to enhance performance.

ID-based Federated Frameworks represent diverse architectures for personalization in federated settings and are used as the backbone to test our approach[4]. For each baseline, we compare its original performance with its performance when enhanced by our framework ("+ours"):

- **FedAvg** (McMahan et al. 2017): The foundational federated algorithm based on weighted model averaging;
- **FCF** (Ammad-Ud-Din et al. 2019): Federated collaborative filtering using equal-weighted averaging;
- **FedNCF** (Perifanis and Efraimidis 2022): A framework with local user embeddings and global item embeddings;
- **PFedRec** (Zhang et al. 2023b): Enables two-way personalization through server-side aggregation;
- **FedRAP** (Li, Long, and Zhou 2024): Uses additive modeling for fine-grained user and item personalization.

To further validate FedVLR's design, we also include two additional variants: a naive federated adaptation of VBPR using FedAvg ("+VBPR"), and an architectural ablation that replaces our BLFM module with standard MLPs ("+MLP").

Evaluation Protocol. To ensure a fair comparison, all methods are evaluated under the same protocol. We treat observed interactions as positive samples and all other items as negative samples. We then evaluate performance by ranking the predicted scores for all candidate items in descending order, after masking items seen during training. We use two standard metrics for implicit feedback (He et al. 2020): Hit Rate (HR@K) and Normalized Discounted Cumulative Gain (NDCG@K), setting K=50 for all experiments. The rank-sensitive nature of NDCG is particularly crucial for our evaluation, as it directly measures the quality of fine-grained personalization that our multimodal fusion aims to achieve.

Experimental Setting

Following prior work (Zhang et al. 2023b; Li, Long, and Zhou 2024), we employ negative sampling during training and use the leave-one-out strategy for validation and testing. Hyperparameters for all baselines are tuned via grid search on the validation set, including the learning rate η within $\{10^i \mid i = -4, \ldots, -1\}$ and method-specific parameters. We use CLIP (Radford et al. 2021b) as the foundation model for extracting visual and textual features, implemented via OpenCLIP (Ilharco et al. 2021) with pre-trained ViT-B-32 weights from OpenAI (Radford et al. 2021a; Schuhmann et al. 2022). A linear mapping layer projects multimodal and ID embeddings to a shared latent dimension of 64 for efficiency. The training batch size is 2048. All models incorporating hidden layers use a three-layer MLP structure. For federated methods, clients perform 5 local training epochs per communication round using a consistent aggregation strategy for both vanilla and FedVLR-integrated versions.

Performance Analysis

Tables 2 and 3 evaluate our proposed FedVLR's performance and efficiency in the original federated frameworks and when enhanced with three different multimodal module variants: our personalized fusion framework ("+ours"), a generic MLP-based module for ablation ("+MLP"), and a federated adaptation of VBPR ("+VBPR").

Effectiveness of the Federated Architecture. The performance gains lies in FedVLR's ability to learn a personalized fusion model for each user. The generic MLP-based fusion provides a crucial comparison. As shown in Table 3, despite having a comparable parameter count, FedVLR consistently demonstrates superior performance, indicating the gain is a direct result of our BLFM architecture, not simply more parameters. A notable exception is the HR performance on the small-scale Beauty dataset, where the simpler MLP-based fusion excels. This is likely because on datasets with very limited user data, a less adaptive model can be

[2]https://grouplens.org/datasets/movielens/latest/

[3]https://github.com/westlake-repl/NineRec

[4]We do not include direct comparisons with AMMFRS (Feng et al. 2024) and FedMMR (Li et al. 2024) as their official code implementations were not publicly available at the time of our experiments, preventing a fair and reproducible comparison.

Method	Beauty		Cards		ML		KU		Dance		Food		Movie	
	HR	NDCG	HR	NDCG	HR	NDCG	HR	NDCG	HR	NDCG	HR	NDCG	HR	NDCG
VBPR	24.11	6.65	73.74	29.34	23.11	6.98	44.12	15.47	23.06	7.47	24.72	7.90	14.29	4.86
BM3	18.58	4.15	83.55	34.46	18.36	4.91	41.67	15.95	26.54	8.56	25.63	8.20	19.17	6.66
MGCN	29.25	6.94	80.11	33.27	25.74	7.10	33.82	15.08	23.63	7.67	24.09	7.77	16.97	5.74
PGL	22.53	6.60	81.17	32.18	26.07	8.37	46.57	20.13	24.83	8.25	36.16	13.56	17.37	5.94
FedAvg	17.00	3.82	64.19	23.20	11.80	3.79	11.27	5.62	7.82	2.01	6.96	1.91	5.52	1.58
+ VBPR	18.18	4.42	64.99	19.03	7.87	3.36	51.47	14.91	\	\	9.82	2.91	\	\
+ MLP	18.58	5.39	68.17	23.20	7.87	3.28	39.22	12.40	8.94	2.18	9.63	2.70	5.63	1.60
+ ours	19.37	5.51	70.29	27.60	12.30	3.81	54.90	**16.04**	**10.07**	**2.61**	9.48	2.68	5.64	1.60
FCF	28.46	8.87	69.50	25.16	11.64	3.72	22.55	6.43	6.76	1.77	8.20	2.19	5.81	1.63
+ MLP	**41.11**	11.77	71.35	20.27	11.64	3.72	36.76	12.40	7.26	1.74	8.99	2.27	6.28	1.73
+ ours	37.15	**12.04**	72.68	28.51	**13.28**	3.91	**55.39**	15.97	8.06	2.17	9.18	2.47	**6.67**	1.80
FedNCF	17.79	4.32	42.71	11.13	3.93	1.14	11.76	2.49	2.21	0.55	5.48	1.63	1.85	0.54
+ MLP	20.16	5.07	70.82	22.73	10.82	3.54	28.43	6.91	6.01	1.47	9.18	2.40	3.70	1.04
+ ours	21.74	5.08	72.68	28.38	11.31	3.01	43.63	12.18	7.08	1.87	8.43	2.43	4.06	1.25
PFedRec	17.79	4.56	32.63	8.67	11.31	3.67	6.37	2.30	2.52	0.68	3.27	0.93	1.70	0.42
+ MLP	21.34	4.67	69.76	25.31	11.97	3.38	39.22	8.52	8.52	2.10	8.68	2.50	5.41	1.41
+ ours	24.90	5.72	70.29	27.76	12.30	3.81	45.59	15.05	7.46	1.92	10.10	2.55	5.72	1.48
FedRAP	23.72	9.18	62.60	22.29	11.64	3.76	47.06	10.86	6.82	1.86	6.96	1.91	4.53	1.36
+ MLP	32.02	10.80	64.19	27.57	12.30	3.58	48.53	12.12	7.36	1.76	8.68	2.21	4.56	1.42
+ ours	38.34	11.39	**73.74**	**28.76**	12.79	**4.15**	48.24	13.05	8.95	2.47	**11.42**	**2.94**	5.86	**1.99**

Table 2: Performance comparison of our FedVLR ("+ours") against centralized and federated baselines, reporting HR@50 (HR) and NDCG@50 (NDCG) in percent. The highest federated result per column is in **bold**, and \ denotes failed runs.

Model	Beauty	Cards	ML	KU	Dance	Food	Movie
FedAvg	22	8	233	287	107	72	161
+ VBPR	441	452	2,493	495	11,334	838	17,584
+ MLP	68	24	700	107	321	216	485
+ ours	80	37	713	120	334	228	497
FCF	22	8	233	35	858	576	161
+ MLP	68	24	700	107	321	216	485
+ ours	80	37	713	120	334	228	497
FedNCF	88	75	556	108	1,535	921	1,182
+ MLP	403	273	2,556	527	2,525	1,629	3,797
+ ours	440	311	2,594	564	2,561	1,666	3,833
PFedRec	22	8	233	287	107	72	161
+ MLP	68	24	700	107	321	216	485
+ ours	80	37	713	120	334	228	497
FedRAP	45	16	467	71	214	144	323
+ MLP	91	33	934	143	429	288	647
+ ours	103	45	946	155	441	300	659

Table 3: Comparison of client-side parameters in thousands, detailing the parameter count for each ID-based federated baseline and its variants with different fusion modules.

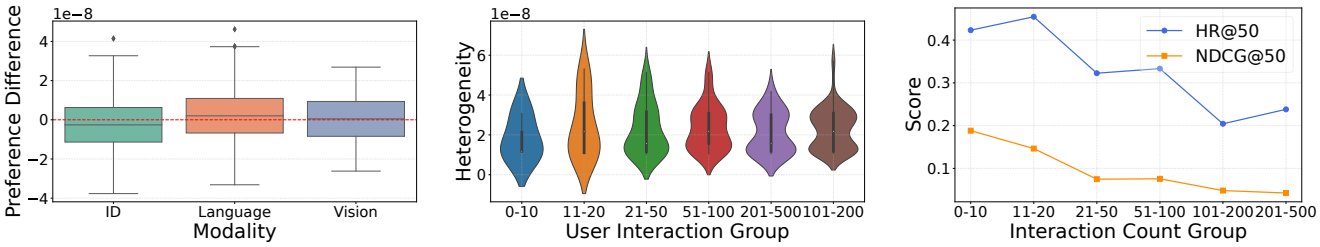
less prone to overfitting. Nevertheless, the overall trend confirms that our client-side refinement successfully captures user-specific nuances that a one-size-fits-all mechanism cannot. The direct federated adaptation of VBPR provides further evidence for our design. This naive adaptation proves impractical, as its excessive parameter count leads to out-of-memory failures on larger datasets and inferior performance where it runs. This underscores that an effective solution requires a purpose-built architecture, like our decoupled BLFM, designed fundamentally for the federated setting.

Bridging the Gap to Centralized Performance. Our framework also demonstrates strong performance when compared to centralized models. Centralized methods learn from global interaction patterns, while our approach learns from isolated user data in a private decentralized manner. The performance dynamic between these two paradigms depends on the dataset's scale. On smaller datasets such as Beauty and KU, the global collaborative signal is sparse. In this regime, FedVLR's ability to learn a rich and personalized content model directly from user interactions becomes the decisive factor, often leading to superior performance. As the dataset scale and user population grow, the centralized model's access to a massive interaction matrix provides a strong advantage in discovering subtle and high-order collaborative patterns that are invisible to any single client. Even in these scenarios, FedVLR consistently narrows the performance gap. This key finding suggests that deep personalization from local data is a powerful alternative to learning from global signals, validating our privacy-preserving framework as a practical and effective approach.

Ablation Studies

We conduct a series of ablation studies to dissect the key components and properties of FedVLR. We analyze the impact of client heterogeneity, contribution of each modality, and the framework's robustness while privacy-enhancing.

Analysis of Client Heterogeneity. We empirically investigate client heterogeneity learned by FedRAP enhanced with our FedVLR on the KU dataset. Fig. 2 reveals several key findings. Fig. 2a shows that users not only exhibit diverse activity levels, but they also show significant variance in their preferences for different modalities, especially for visual and textual content. Furthermore, this preference heterogeneity tends to increase for more active user groups, as indicated in Fig. 2b. This rising heterogeneity directly



(a) Modality Preferences (b) Heterogeneity Distribution (c) Performance Trend

Figure 2: Analysis of user characteristics learned by FedRAP enhanced with our proposed FedVLR on the KU dataset: (a) User preference for different modalities; (b) User activity heterogeneity distribution; (c) Performance trend across user groups.

Method	Metric	w/ ours	w/ noise	Degrade
FedAvg	HR	51.47	49.69	3.46% ↓
	NDCG	14.91	14.16	5.03% ↓
FCF	HR	55.39	53.78	2.91% ↓
	NDCG	15.97	14.02	12.21% ↓
FedNCF	HR	19.61	16.80	14.33% ↓
	NDCG	6.09	4.20	31.03% ↓
PFedRec	HR	21.57	19.65	8.90% ↓
	NDCG	6.16	5.86	4.87% ↓
FedRAP	HR	38.24	37.25	2.59% ↓
	NDCG	13.05	12.36	5.29% ↓

Table 4: Analysis of the privacy–utility trade-off on KU.

impacts model performance. As seen in Fig. 2c, accuracy tends to decrease for the most active and thus most heterogeneous user groups. Capturing the diverse and nuanced preferences of these users is a significant challenge for models with non-adaptive fusion logic, providing strong empirical validation for the central problem our paper addresses. FedVLR's client-side refinement module is designed specifically to tackle this preference heterogeneity, enabling the model to tailor its representations to each user.

Contribution of Different Modalities. We analyze the contribution of the visual (**V**), textual (**C**), and collaborative ID (**D**) signals by systematically removing each of them from the model, as shown in Fig. 3. Performance consistently degrades when any modality is removed, confirmed that all signals are integral to the accuracy. The more critical finding is that the relative importance of these modalities is dynamic, depending on the underlying federated architecture. For instance, some frameworks are severely impacted by the removal of the collaborative ID signal, while others are more sensitive to the loss of visual or textual content. This strong architectural dependency demonstrates that no universal hierarchy of modality importance exists, indicating that the optimal way to combine these diverse signals is highly context-dependent. It validates the need for an adaptive fusion mechanism capable of learning this balance.

Robustness to Privacy Enhancement. We assess FedVLR's compatibility with privacy-enhancing technologies by adding Gaussian noise to the gradients before server aggregation, a common method for achieving differential privacy. Table 4 shows that it introduces a predictable and modest trade-off between privacy and utility, with a slight decrease in performance across all models. Notably, the performance degradation is less severe for simpler or more adaptive methods like FedAvg and FedRAP, suggesting that complex, rigid models may be more sensitive to the perturbations introduced by noise. Overall, the results confirm that FedVLR integrates effectively with standard privacy-enhancing mechanisms, demonstrating its practical viability for applications in privacy-conscious environments.

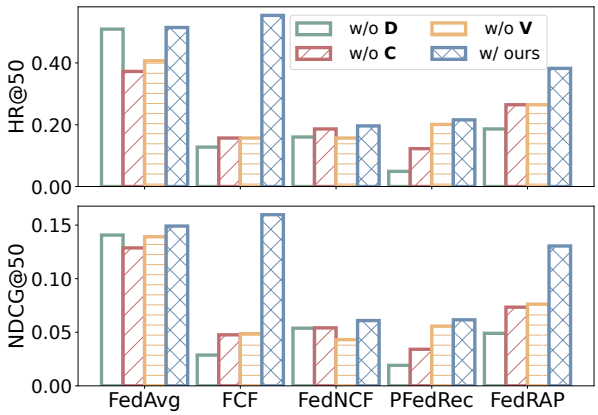


Figure 3: Impact on the performance after removing visual (**V**), textual (**C**), or collaborative ID (**D**) features on KU.

Conclusion

This paper tackles the challenge of personalized modality fusion in on-device VLR, where uniform fusion logic fails to capture users' diverse preferences for visual and language content. We propose FedVLR, a framework with a bi-level fusion mechanism that decouples server-side view generation from lightweight on-device refinement. This design empowers each client to learn a fine-grained personalized multimodal fusion module. Extensive experiments and theoretical analysis confirm FedVLR substantially improves existing federated baselines, offering a principled pathway toward more personal and privacy-preserving VLR systems.

References

Allouah, Y.; Guerraoui, R.; Gupta, N.; Pinot, R.; and Rizk, G. 2023. Robust distributed learning: Tight error bounds and breakdown point under data heterogeneity. *NeuRIPS*, 36: 45744–45776.

Ammad-Ud-Din, M.; Ivannikova, E.; Khan, S. A.; Oyomno, W.; Fu, Q.; Tan, K. E.; and Flanagan, A. 2019. Federated collaborative filtering for privacy-preserving personalized recommendation system. *arXiv preprint arXiv:1901.09888*.

Bian, S.; Pan, X.; Zhao, W. X.; Wang, J.; Wang, C.; and Wen, J.-R. 2023. Multi-modal mixture of experts represetation learning for sequential recommendation. In *CIKM*, 110–119.

Feng, C.; Feng, D.; Huang, G.; Liu, Z.; Wang, Z.; and Xia, X.-G. 2024. Robust Privacy-Preserving Recommendation Systems Driven by Multimodal Federated Learning. *TNNLS*.

Fu, J.; Yuan, F.; Song, Y.; Yuan, Z.; Cheng, M.; Cheng, S.; Zhang, J.; Wang, J.; and Pan, Y. 2024. Exploring adapter-based transfer learning for recommender systems: Empirical studies and practical insights. In *WSDM*, 208–217.

Geng, S.; Tan, J.; Liu, S.; Fu, Z.; and Zhang, Y. 2023. Vip5: Towards multimodal foundation models for recommendation. *arXiv preprint arXiv:2305.14302*.

Guo, T.; Guo, S.; Wang, J.; Tang, X.; and Xu, W. 2023. Promptfl: Let federated participants cooperatively learn prompts instead of models–federated learning in age of foundation model. *IEEE TMC*, 23(5): 5179–5194.

Harper, F. M.; and Konstan, J. A. 2015. The movielens datasets: History and context. *TIIS*, 5(4): 1–19.

He, R.; and McAuley, J. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *AAAI*, volume 30.

He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; and Wang, M. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *SIGIR*, 639–648.

Hou, Y.; Li, J.; He, Z.; Yan, A.; Chen, X.; and McAuley, J. 2024. Bridging Language and Items for Retrieval and Recommendation. *arXiv preprint arXiv:2403.03952*.

Hou, Y.; Mu, S.; Zhao, W. X.; Li, Y.; Ding, B.; and Wen, J.-R. 2022. Towards universal sequence representation learning for recommender systems. In *SIGKDD*, 585–593.

Ilharco, G.; Wortsman, M.; Wightman, R.; Gordon, C.; Carlini, N.; Taori, R.; Dave, A.; Shankar, V.; Namkoong, H.; Miller, J.; Hajishirzi, H.; Farhadi, A.; and Schmidt, L. 2021. OpenCLIP. If you use this software, please cite it as below.

Kong, X.; He, X.; Ma, X.; Yan, X.; Wang, L.; Shen, G.; and Liu, Z. 2025. Oh-FedRec: one-shot and heterogeneous vertical federated recommendation system. *IEEE Transactions on Consumer Electronics*, 1–1.

Lei, F.; Cao, Z.; Yang, Y.; Ding, Y.; and Zhang, C. 2023. Learning the user's deeper preferences for multi-modal recommendation systems. *TOMM*, 19(3s): 1–18.

Li, G.; Ding, X.; Yuan, L.; Zhang, L.; and Rong, Q. 2024. Towards Resource-Efficient and Secure Federated Multimedia Recommendation. In *ICASSP*, 5515–5519. IEEE.

Li, M.; Zhang, Z.; Zhao, X.; Wang, W.; Zhao, M.; Wu, R.; and Guo, R. 2023. Automlp: Automated mlp for sequential recommendations. In *WWW*, 1190–1198.

Li, Y.; Shan, Y.; Liu, Y.; Wang, H.; Wang, W.; Wang, Y.; and Li, R. 2025a. Personalized Federated Recommendation for Cold-Start Users via Adaptive Knowledge Fusion. In *WWW*, 2700–2709.

Li, Z.; Long, G.; and Zhou, T. 2024. Federated Recommendation with Additive Personalization. In *The Twelfth International Conference on Learning Representations*.

Li, Z.; Long, G.; Zhou, T.; Jiang, J.; and Zhang, C. 2025b. Personalized federated collaborative filtering: A variational autoencoder approach. In *AAAI*, volume 39, 18602–18610.

Liang, F.; Pan, W.; and Ming, Z. 2021. Fedrec++: Lossless federated recommendation with explicit feedback. In *AAAI*, volume 35, 4224–4231.

Lin, G.; Liang, F.; Pan, W.; and Ming, Z. 2020. Fedrec: Federated recommendation with explicit feedback. *IEEE Intell. Syst.*, 36(5): 21–30.

Liu, F.; Cheng, Z.; Sun, C.; Wang, Y.; Nie, L.; and Kankanhalli, M. 2019. User diverse preference modeling by multimodal attentive metric learning. In *ACM MM*, 1526–1534.

Liu, F.; Wu, X.; Ge, S.; Fan, W.; and Zou, Y. 2020. Federated learning for vision-and-language grounding problems. In *AAAI*, volume 34, 11572–11579.

Liu, Q.; Hu, J.; Xiao, Y.; Zhao, X.; Gao, J.; Wang, W.; Li, Q.; and Tang, J. 2024a. Multimodal recommender systems: A survey. *ACM Computing Surveys*, 57(2): 1–17.

Liu, Q.; Zhu, J.; Yang, Y.; Dai, Q.; Du, Z.; Wu, X.-M.; Zhao, Z.; Zhang, R.; and Dong, Z. 2024b. Multimodal Pretraining, Adaptation, and Generation for Recommendation: A Survey. In *SIGKDD*, 6566–6576.

Lu, S.; Liu, Z.; Liu, T.; and Zhou, W. 2023. Scaling-up medical vision-and-language representation learning with federated learning. *Engineering Applications of Artificial Intelligence*, 126: 107037.

Luo, S.; Xiao, Y.; and Song, L. 2022. Personalized federated recommendation via joint representation learning, user clustering, and model adaptation. In *CIKM*, 4289–4293.

Ma, S.; Zeng, Y.; Wu, S.; and Xu, G. 2025. Refining Contrastive Learning and Homography Relations for Multi-Modal Recommendation. In *ACM MM*, 6316–6324.

Malitesta, D.; Cornacchia, G.; Pomo, C.; Merra, F. A.; Di Noia, T.; and Di Sciascio, E. 2025. Formalizing multimedia recommendation through multimodal deep learning. *TORS*, 3(3): 1–33.

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

Niu, Z.; Zhong, G.; and Yu, H. 2021. A review on the attention mechanism of deep learning. *Neurocomputing*, 452: 48–62.

Pan, B.; Huang, W.; and Shi, Y. 2024. Federated learning from vision-language foundation models: Theoretical analysis and method. *NeuRIPS*, 37: 30590–30623.

Perifanis, V.; and Efraimidis, P. S. 2022. Federated neural collaborative filtering. *KBS*, 242: 108441.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021a. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021b. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

Ren, C.; Yu, H.; Peng, H.; Tang, X.; Zhao, B.; Yi, L.; Tan, A. Z.; Gao, Y.; Li, A.; Li, X.; et al. 2024. Advances and Open Challenges in Federated Foundation Models. *arXiv preprint arXiv:2404.15381*.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C. W.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; Schramowski, P.; Kundurthy, S. R.; Crowson, K.; Schmidt, L.; Kaczmarczyk, R.; and Jitsev, J. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Sun, W.; Xie, R.; Bian, S.; Zhao, W. X.; and Zhou, J. 2023. Universal Multi-modal Multi-domain Pre-trained Recommendation. *arXiv preprint arXiv:2311.01831*.

Voigt, P.; and Von dem Bussche, A. 2017. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676): 10–5555.

Wang, J.; Zeng, Z.; Wang, Y.; Wang, Y.; Lu, X.; Li, T.; Yuan, J.; Zhang, R.; Zheng, H.-T.; and Xia, S.-T. 2023. MISSRec: Pre-training and transferring multi-modal interest-aware sequence representation for recommendation. In *ACM MM*, 6548–6557.

Wei, T.; Jin, B.; Li, R.; Zeng, H.; Wang, Z.; Sun, J.; Yin, Q.; Lu, H.; Wang, S.; He, J.; et al. 2024. Towards unified multi-modal personalization: Large vision-language models for generative recommendation and beyond. *arXiv preprint arXiv:2403.10667*.

Wei, Y.; Wang, X.; Nie, L.; He, X.; Hong, R.; and Chua, T.-S. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *ACM MM*, 1437–1445.

Wu, X.; Huang, A.; Yang, H.; He, H.; Tai, Y.; and Zhang, W. 2024. Towards Bridging the Cross-modal Semantic Gap for Multi-modal Recommendation. *arXiv preprint arXiv:2407.05420*.

Yin, H.; Qu, L.; Chen, T.; Yuan, W.; Zheng, R.; Long, J.; Xia, X.; Shi, Y.; and Zhang, C. 2024. On-device recommender systems: A comprehensive survey. *arXiv preprint arXiv:2401.11441*.

Yu, P.; Tan, Z.; Lu, G.; and Bao, B.-K. 2023. Multi-view graph convolutional network for multimedia recommendation. In *ACM MM*, 6576–6585.

Yu, P.; Tan, Z.; Lu, G.; and Bao, B.-K. 2025. Mind Individual Information! Principal Graph Learning for Multimedia Recommendation. In *AAAI*, volume 39, 13096–13105.

Yuan, W.; Qu, L.; Cui, L.; Tong, Y.; Zhou, X.; and Yin, H. 2024. Hetefedrec: Federated recommender systems with model heterogeneity. In *ICDE*, 1324–1337. IEEE.

Zhang, C.; Long, G.; Guo, H.; Fang, X.; Song, Y.; Liu, Z.; Zhou, G.; Zhang, Z.; Liu, Y.; and Yang, B. 2024a. Federated Adaptation for Foundation Model-based Recommendations. *arXiv preprint arXiv:2405.04840*.

Zhang, C.; Long, G.; Guo, H.; Liu, Z.; Zhou, G.; Zhang, Z.; Liu, Y.; and Yang, B. 2025a. Multifaceted user modeling in recommendation: A federated foundation models approach. In *AAAI*, volume 39, 13197–13205.

Zhang, C.; Long, G.; Zhou, T.; Yan, P.; Zhang, Z.; and Yang, B. 2023a. Graph-guided Personalization for Federated Recommendation. *arXiv preprint arXiv:2305.07866*.

Zhang, C.; Long, G.; Zhou, T.; Yan, P.; Zhang, Z.; Zhang, C.; and Yang, B. 2023b. Dual personalization on federated recommendation. *arXiv preprint arXiv:2301.08143*.

Zhang, C.; Yang, Z.; He, X.; and Deng, L. 2020. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3): 478–493.

Zhang, H.; Li, H.; Chen, J.; Cui, S.; Yan, K.; Wuerkaixi, A.; Zhou, X.; Shen, Z.; and Li, Y. 2024b. Beyond similarity: Personalized federated recommendation with composite aggregation. *arXiv preprint arXiv:2406.03933*.

Zhang, H.; Liu, H.; Li, H.; and Li, Y. 2024c. Transfr: Transferable federated recommendation with pre-trained language models. *arXiv preprint arXiv:2402.01124*.

Zhang, H.; Luo, F.; Wu, J.; He, X.; and Li, Y. 2023c. LightFR: Lightweight federated recommendation with privacy-preserving matrix factorization. *TOIS*, 41(4): 1–28.

Zhang, J.; Cheng, Y.; Ni, Y.; Pan, Y.; Yuan, Z.; Fu, J.; Li, Y.; Wang, J.; and Yuan, F. 2024d. Ninerec: A benchmark dataset suite for evaluating transferable recommendation. *TPAMI*.

Zhang, Q.; Li, J.; Jia, Q.; Wang, C.; Zhu, J.; Wang, Z.; and He, X. 2021. UNBERT: User-News Matching BERT for News Recommendation. In *IJCAI*, volume 21, 3356–3362.

Zhang, Z.; Liu, S.; Liu, Z.; Zhong, R.; Cai, Q.; Zhao, X.; Zhang, C.; Liu, Q.; and Jiang, P. 2025b. Llm-powered user simulator for recommender system. In *AAAI*, volume 39, 13339–13347.

Zhou, H.; Zhou, X.; Zeng, Z.; Zhang, L.; and Shen, Z. 2023a. A comprehensive survey on multimodal recommender systems: Taxonomy, evaluation, and future directions. *arXiv preprint arXiv:2302.04473*.

Zhou, P.; Liu, C.; Ren, J.; Zhou, X.; Xie, Y.; Cao, M.; Rao, Z.; Huang, Y.-L.; Chong, D.; Liu, J.; et al. 2025. When Large Vision Language Models Meet Multimodal Sequential Recommendation: An Empirical Study. In *WWW*, 275–292.

Zhou, X.; Zhou, H.; Liu, Y.; Zeng, Z.; Miao, C.; Wang, P.; You, Y.; and Jiang, F. 2023b. Bootstrap latent representations for multi-modal recommendation. In *WWW*, 845–854.