Retrieval-Augmented Vertical Federated Learning for Recommender System

Wenjie Li[1,2], Zhongren Wang[1,2], Jianghui Zhang[1], Chenghui Song[1], Shu-Tao Xia[1]
Jile Zhu[2], Mingjian Chen[2], Jiangke Fan[2], Jia Cheng[2], Jun Lei[2]
Tsinghua University[1], Meituan[2]
liwj20,wangzr23,jh-zhang21,sch21@mails.tsinghua.edu.cn,xiast@sz.tsinghua.edu.cn
zhujile,chenmingjian,jiangke.fan,jia.cheng.sh,leijun@meituan.com

ABSTRACT

As an emerging privacy-preserved fashion capable of leveraging cross-platform user interactions to enhance user modeling, vertical federated learning (VFL) has been increasingly applied in recommender systems. However, vanilla VFL is only applicable to overlapped users among participants, which may incur the loss of important universal interest patterns hidden in non-overlapped users. Such *limited user application scope* reduces the efficacy of VFL in real-world recommendation scenarios.

In this paper, we propose the **R**etrieval-**A**ugmented **VFL** to tackle this issue, yet a groundbreaking initiative of exploring retrieval-enhanced machine learning fashion in VFL. Specifically, We set up a general "retrieve-and-fusion" algorithm to enhance the quality of representations in all parties. We design a flexible user-grained retrieval mechanism for VFL and conduct *(i) Cross-RA for information missing in passive parties* and *(ii) Local-RA for intra-domain representations enhancement*. Extensive experiments conducted on a real-world industry dataset from Meituan demonstrate that our method achieved ***percentile-level AUC lifts*** in appropriate settings and validates the effectiveness of both components.

KEYWORDS

Recommender System, Retrieval Augmentation, Vertical Federated Learning, Split Neural Network

ACM Reference Format:
Wenjie Li[1,2], Zhongren Wang[1,2], Jianghui Zhang[1], Chenghui Song[1], Shu-Tao Xia[1], Jile Zhu[2], Mingjian Chen[2], Jiangke Fan[2], Jia Cheng[2], Jun Lei[2]. 2018. Retrieval-Augmented Vertical Federated Learning for Recommender System. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Recommender systems[24], as one of the essential tools for information retrieval (IR), have become ubiquitous in our daily lives, from reading news and watching movies to reviewing restaurants, online shopping, and identifying points of interest (POIs). In this context, user behaviors are recorded by many kinds of service platforms, which is diverse and complementary to depict user interests in different domains. Collectively leveraging these multi-platform user behaviors is helpful to achieve more fine-grained and comprehensive user interest modeling, which is beneficial to many tasks. However, it is nearly impossible to directly share raw data across platforms due to data privacy regulations [33] and commercial confidentiality among service agencies.

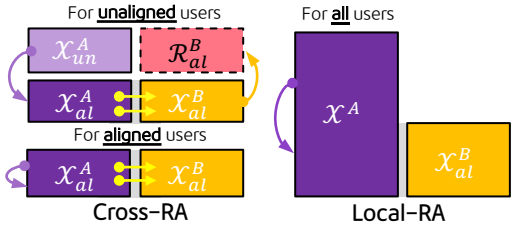

Figure 1: Two kinds of augmentation in RA-VFL. (a) Cross-RA aims to fill the information missing gap for unaligned users with retrieved set $\mathcal{R}^B_{al}$, and can be further extended for aligned users to enhance B-side representation. (b) Local-RA aims to enhance A-side representation for all users

To address these issues, vertical federated learning [3, 23, 32, 34] has been adopted to utilize cross-platform attributes without compromising user privacy. It has been explored in various recommendation tasks[36], such as click-through prediction, conversion rate prediction, and item recommendation [5, 6, 10, 11]. In a typical VFL process, participants first execute the PSI (Private Set Intersection[26]) to obtain the aligned (i.e., overlapped) dataset and then perform distributed training. This essentially restrains the models to only train and infer on aligned users. This constraint of ***narrowed data scope*** makes VFL impractical in two ways:

- **Insufficient interest representation**: Aligned users for dissimilar businesses are often limited and constitute only a small portion of the user population. This reduced training set size can increase the risk of overfitting and result in low-quality embeddings and hidden representations, especially in sparse high-dimensional recommendation datasets. In particular, item interactions of unaligned users are completely ignored, despite their potential to enhance item representations.
- **helpless for unaligned users**: The lack of fields in other parties makes it impossible for a federated model to make predictions for unaligned users, which further undermines the practicability of VFL. If a participant holds more unaligned users than aligned users, or weights unaligned users more heavily in their business,

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

there may be insufficient motivation to join the federation. The expected performance gains on the aligned user set are not useful in this case. Although default filling with default values can alleviate this problem, it is superficial and does not provide a meaningful information supplement.

In this paper, we propose a retrieval-enhanced approach named **RA-VFL** (Retrieval Augmented VFL) to tackle these issues, motivated by the success of retrieval-enhanced machine learning (REML)[37] in the fields of open-domain question answering [4, 15], large language models[2, 7], and recommendation[1, 27]. Specifically, we design two kinds of retrieval augmentation strategies:

- **Cross-RA for Inter-Domain Field Missing** For a target user in the unaligned user set, we retrieve relevant users in the aligned user set (only depending on their characters in the active party's domain) and then use the corresponding features in passive party's domain as supplement features to fill the information missing gap.
- **Local-RA for Intra-Domain Enhancement** To enhance mutual correlation learning among aligned and unaligned users, we further retrieve relevant users in the entire user set for each user to enhance representations in the active party's feature domain.

To achieve these strategies, we design an interaction-based user retriever to conduct similarity-based Top-k user search and record selection. And propose attention-based fusion modules to learn enhanced representations for target users based on retrieved records. With RA-VFL, training, and inference can be carried out for the full user set, while representation in all domains can be enhanced. In summary, our contributions are:

(1) We propose the first retrieval-based framework for vertical federated learning, which is ***the groundbreaking initiative attempt to show how can retrieval systems help VFL***.
(2) We propose the Cross-side RA and Local-side RA mechanism under a "retrieve-and-fusion" framework to effectively achieve enhanced modeling for all parties, which also ***tackles the problem of full-set user modeling in VFL*** and systematically enhances all parties' representation.
(3) We conducted extensive experiments on a ***real-world industry dataset*** of Meituan and our method ***achieves significant performance lift*** against baseline models.

Our paper focuses on the typical two-party vertical federated learning setting and experiments on CTR task as a demonstration. However, our method is compatible to other recommendation tasks and can be directly used in multi-party VFL scenarios.

2 PRELIMINARY

2.1 Two-Party VFL

Without loss of generality, we focus on a typical two-party VFL setting [3, 32] with an **active party A** which posses the label and a **passive party B** which doesn't. The two parties first perform private set intersection (PSI) to obtain an aligned sample set for training:

$$\mathcal{D}_{al} = \{(\mathbf{x}_A, \mathbf{x}_B, \mathbf{y})_{u,i} | u \in \mathcal{U}_{al}, i \in \mathcal{I}_{al}\} \tag{1}$$

where $\mathbf{x}_A, \mathbf{x}_B$ denote the sample's feature provided by two parties respectively, the label $\mathbf{y}$ is a one-hot vector, u and i are indices from the aligned user set $\mathcal{U}_{al}$ and aligned item set $\mathcal{I}_{al}$. For brevity, *we may also use the term "local" as a special pronoun for the active party* in the following sections. In the forward pass, Each party holds a bottom model (f_A, f_B) for extracting hidden representations and the active party additionally holds a top model g_A to fuse two sides of representations to make predictions $\hat{\mathbf{y}}$, and further computes the cross-entropy loss $\mathcal{L}$.

$$h_A = f_A(\mathbf{x}_A), h_B = f_B(\mathbf{x}_B) \tag{2}$$

$$\mathbf{l} = g_A(\mathbf{h}_A, \mathbf{h}_B), \hat{y} = softmax(\mathbf{l}) \tag{3}$$

$$\mathcal{L} = -\mathbf{y}^\top log\hat{\mathbf{y}} \tag{4}$$

Note that $\mathbf{h}_A$ and $\mathbf{h}_B$ are distributedly computed in each party's server. Once $\mathbf{h}_B$ is ready, it will be transferred via the network from party B to party A to finish the subsequent computation. In the backward pass, the active party A will send the gradient of $\mathbf{h}_B$ to party B to conduct subsequent backward pass and parameter updates.

$$\mathbf{g} := \nabla_{\mathbf{h}_B}\mathcal{L} = \frac{\partial \mathcal{L}}{\partial \mathbf{h}_B} = \frac{\partial \mathbf{l}}{\partial \mathbf{h}_B} \cdot \frac{\partial \mathcal{L}}{\partial \mathbf{l}} \tag{5}$$

During the whole training process, only $\mathbf{h}_B$ and $\mathbf{g}$ are shared between parties, the original data is never exposed. The security of intermediate data transmission can be satisfied by combining methods proposed by related works which are complementary to us, so we focus on the aspect of model training.

2.2 Problem Formulation

First, we define the **fullset** of labeled data as the combination of the aligned dataset from both parties and the unaligned sample set from the active party:

$$\mathcal{D}_{full} = \{\mathcal{D}_{al}, \mathcal{D}_{un}\} \tag{6}$$

$$\mathcal{D}_{un} = \{(\mathbf{x}_A, \mathbf{y})_{u,i} | u \in \mathcal{U}_{un}, i \in \mathcal{I}_{un} \cup \mathcal{I}_{al}\} \tag{7}$$

The unaligned part of party B is naturally excluded since no corresponding label exists in party A. One of the key drawbacks of two-party VFL is the degraded sample size from the fullset $\mathcal{D}_{full}$ to the aligned set $\mathcal{D}_{al}$. Thus in this paper, we focus on a more general scenario that the fullset of labeled data is utilized to maximize the data utilization in VFL.

However, the challenge of additionally exploiting the active party's unaligned samples is that they do not have passive party fields, they can not be directly used. To tackle this issue, we propose a brand new fashion namely Retrieval-Augmented Vertical Federated Learning (RA-VFL) to release the potential of fullset modeling. The problem is defined as follows:

Given the fullset of labeled dataset $\mathcal{D}_{full}$, RA-VFL aims to leverage some retrieval mechanism $\mathcal{R}$ to acquire Top-k relevant samples to augment the modeling, either to fill the missing part of $\mathcal{D}_{un}$ to enable full user-set modeling or enhancing representations with complemented intra-party information.

2.3 Overview

RA-VFL, shown in Fig. 2, introduces two new components relative to the Basic Two-Tower VFL model: (**i**) **The Retrieval Module** that produces top-k similar users and their corresponding augmentation records (referred to as the "neighbor set") for both parties and (**ii**) **The Fusion module**, which aggregates the representation of the neighborhood set.

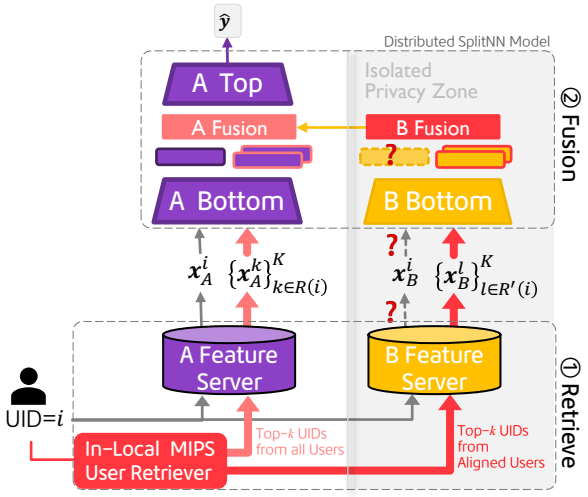

Figure 2: Overview of our RA-VFL Framework.

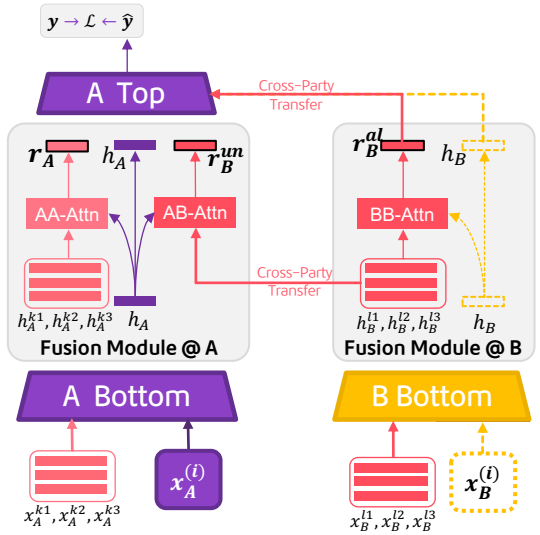

Figure 4: Architecture of the Enhanced VFL Model.

2.4 Retrieval Module

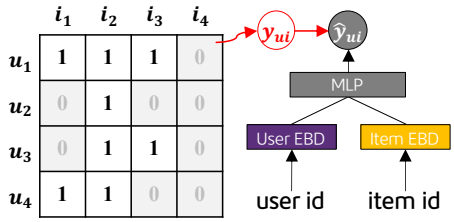

Figure 3: The Lightweight Interaction-Based User Retriever

We introduce the retrieval module in two aspects: (**i**) The in-local user retrieval: Given a random user id, how can we get meaningful top-k related user ids without breaking privacy? (**ii**) The federated sample retrieval strategy: how can we get helpful samples on each party based on user retrieval results?

2.4.1 ***In-Local User Retrieval.*** Given the user-item interaction nature of recommendation datasets and the entity-level PSI in VFL, we assume that user-level retrieval is more reasonable and feasible. Therefore, to retrieve similar records for augmentation, we first design a user-centric dense retriever to obtain users similar to the target user, and then select records corresponding to these users. Specifically, we encode the interaction interests of each user using an NCF model [9], which allows us to derive user embeddings. To select relevant users, we calculate a relevance score between the target user and the rest of the users based on their embeddings. Following REALM [8], we define the relevance score with an inner product:

$$s(u_i, u_j) = \mathbf{v}_i \cdot \mathbf{v}_j$$

where $\mathbf{v}_i$ and $\mathbf{v}_j$ represent the embeddings of the target user u_i and a candidate user u_j, respectively. We use this function to compute the relevance score of each candidate user, and then select the top-k relevant users as the augmentation users. After selecting the users, we choose one sample from all records by that user with a selector $\mathcal{S}$. We utilize the latest selector (select the most recent record from that user) in this paper, but leave the design of $\mathcal{S}$ for future work.

We adopt the Maximum Inner Product Search (MIPS) algorithm for accelerating the embedding retrieval, which has a sub-linear time complexity over the number of candidates. After retrieval, we can obtain a set of k relevant augmentation records for $\mathbf{x}_i$, denoted by $\mathcal{NS}_i = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$.

2.4.2 ***Federated Sample Retrieval.*** Given a retriever $\mathcal{R}$ from the active party, we aim to conduct data augmentation for both-sides. Specifically, we design two mechanisms as depicted in Figure 1

- **Cross-RA**: For both unaligned and aligned users, we use the aligned user set as jumper to retrieve the fields in the passive party's domain. Such indirect retriever is very practical since it only relies on a Local retriever in the active party and does not have any requirements for the passive party. Cross-RA can fill the information gap for unaligned users thus enabling full user set modeling. It can also enhance B-side representation for aligned users. Overall, we retrieve top-k augmentation from the aligned user set for all users, denoted as $\mathcal{U}^t_{full} \leftarrow \mathcal{U}^{t-N}_{al}$, here, t denotes the day of the query and $t - N$ denotes the day of the retrieval pool. The push back of N days is set to prevent data leakage.
- **Local-RA**: we retrieve top-k augmentation from the full user set for each user in fullset, denoted as $\mathcal{U}^t_{full} \leftarrow \mathcal{U}^{t-N}_{full}$. This is designed to enhance A-side representation for both the aligned and unaligned users.

2.5 Fusion Module

Once the retrieval is finished, We feed all the retrieved samples into bottom models to get their initial representations. Specifically, for a sample k from the target user t's neighbour set $\mathcal{NS}(t) = \{1, \ldots, K\}$, its representations in both-side are: $\mathbf{h}^k_A = f_A(\mathbf{x}^k_A), \mathbf{h}^k_B = f_B(\mathbf{x}^k_B)$.

2.5.1 ***Fusion for Cross-RA.*** For a target user t in unaligned set $\mathcal{U}_A$, there is no corresponding $\mathbf{h}_B^t$, so we use cross attention based on $\mathbf{h}_A^t$ to fuse the retrieved set:

$$\mathbf{r}_B^{un} = \sum_{k=1}^{K} \alpha_k \mathbf{h}_B^k, \quad \alpha_k = \frac{exp((\mathbf{h}_B^j)^\top \mathbf{\Phi} \mathbf{h}_A^t)}{\sum_{j=1}^{K} exp((\mathbf{h}_B^j)^\top \mathbf{\Phi} \mathbf{h}_A^t)} \tag{8}$$

where $\mathbf{\Phi} \in \Re^{d_B \times d_A}$ is the weight matrix used to project vectors from domain B to domain A. While for user t in the aligned set $\mathcal{U}_{al}$, we just use its B-side representation $\mathbf{h}_B^t$ as attention key:

$$\mathbf{r}_B^{al} = \sum_{k=1}^{K} \beta_k \mathbf{h}_B^k, \quad \beta_k = \frac{exp((\mathbf{h}_B^j)^\top \mathbf{\Theta} \mathbf{h}_B^t)}{\sum_{j=1}^{K} exp((\mathbf{h}_B^j)^\top \mathbf{\Theta} \mathbf{h}_B^t)} \tag{9}$$

where $\mathbf{\Theta} \in \Re^{d_B \times d_B}$ is the weight matrix to project vectors into metric space.

2.5.2 ***Fusion for Local-RA.*** For Local-RA mode, all users are treated equally with $\mathbf{h}_A^t$-based attention since all users have A-side features:

$$\mathbf{r}_A = \sum_{k=1}^{K} \gamma_k \mathbf{h}_A^k, \quad \gamma_k = \frac{exp((\mathbf{h}_A^j)^\top \mathbf{W} \mathbf{h}_A^t)}{\sum_{j=1}^{K} exp((\mathbf{h}_A^j)^\top \mathbf{W} \mathbf{h}_A^t)} \tag{10}$$

Once the fusion is finished, we concatenate all collected hidden representations and feed them into the top model to finish subsequent forward passes:

$$\hat{y}_t = g_A(concat([\mathbf{h}_A^t, \mathbf{r}_A^t, \mathbf{h}_B^t, \mathbf{r}_B^t]))$$

Here $\mathbf{h}_B^t = f_B(x_B^*)$ is a hidden vector computed for default values x_B^* for unaligned samples, $\mathbf{r}_B^t$ is computed via Eq.8 and Eq.9 for unaligned and aligned samples, respectively.

3 EXPERIMENTS

Our experiments aim to answer the following research questions.

- **RQ1**: How much does RA-VFL outperform traditional VFL?
- **RQ2**: How does the retrieval size k affect performance?
- **RQ3**: How much do different RA modules contribute?

3.1 Experimental Settings

3.1.1 ***Dataset.*** We conducted experiments on a click-through rate (CTR) dataset collected from a real-world advertising platform. The dataset consists of transaction records from 9 days in some cities on the Meituan platform[1]. It includes user profiles, item profiles, and two interaction domains: search actions and browse actions.

3.1.2 ***Party Segmentation.*** We organize attributes of "user profile + item profile + search" into the **Search** domain, and attributes of browse actions into the **Browse** domain. Two federated scenarios were simulated by alternatively choosing one domain as the active party and the other domain as the passive party. The data statistics of the two scenarios after preprocessing are summarized in Table 2.

[1]https://www.meituan.com

3.1.3 ***User Segmentation.*** In each scenario, we segmented the unaligned and aligned user sets according to the absence ratio of fields in the passive party. For example, when we selected the Browse domain as the active party, users with more than $threshold\%$ of missing fields in the Search domain were classified into the unaligned set, and all corresponding values in the Search domain were eliminated for these users. Since the user set, item set, and the alignment status of a specific user among days are all varying, the statistics in Table 2 are calibrated by mean ratio among days to reflect the overall characteristics of the dataset for conciseness.

3.1.4 ***Train & Test & Pool Splitting.*** We use days 1 to 5 as the training set, day 7 as the validation set, and day 8 as the test set. To avoid future data leakage, we use day $t-1$ (days 0 to 4) as the retrieval pool for the training set, and day 6 for both validation and test sets.

3.1.5 ***Evaluation Metrics.*** We choose the commonly used AUC and log loss to measure the performance for CTR prediction task, which reflects pairwise ranking performance and point-wise likelihood, respectively. In Meituan's industry practice, AUC lift in thousands is considered effective, and percentiles are seen as highly significant.

3.1.6 ***Baselines.*** We setup the following approaches to evaluate the effectiveness of RA-VFL:

- **`Local`** is a model trained on the active party's features only, without using any fields from the passive part.
- **`Fed`** This is a VFL model trained on aligned records. It only covers the aligned user set.
- **`Fed-Fill`** This further exploits the active party's unaligned records, in which the missing fields of the passive party are filled with default values. That means the model will learn a default embedding for each missing field. It covers the full user set and fills the missing fields with heuristic method.
- **`RA-both`** This is the full version of our model depicted in Figure 2. "Both" denotes conducting augmentation on both parties.
- **`RA-a & RA-b`** They are two degraded versions of our model, with only one-sided augmentation conducted.
- **`Fed-Raw`** It is trained on full attribute set that the passive party's fields of unaligned users are known in advance. Therefore its performance is the upper bound, and just a reference.

3.1.7 ***Implementation Details.*** User embeddings for retrieval can be trained based on interactions with items across any domain related to the target users, and are easy to acquire in industry. For our experiments, we just leveraged user embeddings that were pre-trained on item domains commonly used in Meituan advertising system. To create a more realistic federated simulation, we used user embeddings from two different domains for two scenarios. To simplify the experimentation process, we pre-retrieved results for all records using FAISS. Our model architecture consists of bottom models with a 2-layer MLP using $64 \rightarrow 32$ units and a top model with a 2-layer MLP with $d_{cat} \rightarrow 16 \rightarrow 1$ units. Here d_{cat} is the width of concatenated vectors from the bottom model and fusion modules. Embedding dimensions for all fields are set to 10. We train our models using Adam optimizer with L_2 regularization and set the learning rate $\eta = 0.001$ and $\lambda = 0.001$ for all methods. We use a

Table 1: Main results of RA-VFL show significant improvement over baselines in both scenarios and across different user sets. We conducted separate experiments for unaligned, aligned, and full-set users, comparing RA-VFL to the most competitive baseline in each setting. *Note: AUC was evaluated on the corresponding user set and shown in percentile for readability.

Scenario	S1:Search-Browse					S2:Browse-Search				
User Set	Method	Logloss↓	Diff	*AUC↑	Diff	Method	Logloss↓	Diff	*AUC↑	Diff
$\mathcal{U}_{un}$	Fed-Fill	0.1386	–	64.12%	–	Fed-Fill	0.1402	–	60.61%	–
	RA-top2	0.1385	-0.0001	64.35%	0.23%	RA-top2	0.1399	-0.0003	61.16%	0.55%
	RA-top4	0.1381	-0.0005	64.94%	0.82%	RA-top4	0.1399	-0.0003	61.32%	0.71%
	RA-top6	0.1380	-0.0006	65.00%	0.88%	RA-top6	0.1399	-0.0003	61.42%	0.81%
	RA-top8	**0.1379**	**-0.0007**	**65.13%**	**1.01%**	RA-top8	**0.1398**	**-0.0004**	**61.66%**	**1.05%**
$\mathcal{U}_{al}$	Fed	0.1762	–	62.96%	–	Fed	0.1760	–	63.03%	–
	RA-top2	0.1768	0.0007	62.95%	-0.01%	RA-top2	0.1761	0.0001	63.07%	0.04%
	RA-top4	0.1762	0.0000	63.18%	0.22%	RA-top4	0.1761	0.0000	63.17%	0.15%
	RA-top6	0.1760	-0.0002	63.27%	0.31%	RA-top6	0.1767	0.0007	63.20%	0.17%
	RA-top8	**0.1760**	**-0.0002**	**63.38%**	**0.42%**	RA-top8	**0.1759**	**-0.0001**	**63.32%**	**0.29%**
$\mathcal{U}_{full}$	Fed-Fill	0.1608	–	64.09%	–	Fed-Fill	0.1612	–	63.43%	–
	RA-top2	0.1606	-0.0002	64.44%	0.34%	RA-top2	0.1611	-0.0001	63.60%	0.17%
	RA-top4	0.1604	-0.0004	64.67%	0.57%	RA-top4	0.1611	-0.0001	63.70%	0.27%
	RA-top6	0.1604	-0.0004	64.80%	0.71%	RA-top6	0.1610	-0.0002	63.78%	0.34%
	RA-top8	**0.1603**	**-0.0005**	**64.81%**	**0.71%**	RA-top8	**0.1610**	**-0.0002**	**63.85%**	**0.42%**

Table 2: Dataset statistics for two scenarios. The underline represents the active party.

Scenario	S1:Search–Browse		S2:Browse–Search	
User Group	unaligned	aligned	unaligned	aligned
Users	337,733	654,671	340,755	651,649
Items*	4618|8.2%	4772|11.1%	4633|8.4%	4763|10.8%
Impressions	1,625,741	2,382,273	1,672,070	2,335,944
Clicks%	1.31%	2.49%	1.33%	2.47%

*: item amount is shown in the "total | unique ratio" format.

batch size of 10000 and adopt early stopping with a patience of 3 epochs. All experiments are conducted on a Linux GPU workstation.

3.2 Results & Analysis

3.2.1 ***Performance of RA (RQ1).*** As shown in 1, RA-VFL significantly outperforms baselines more than 0.15% on AUC, across all user sets and both scenarios, except only for the case of aligned users when $k = 2$. On the other hand, RA-VFL achieves a maximum percentile AUC lift for unaligned users (1.01% in S1 and 1.05% in S2), showing the significant effectiveness in filling the missed information. Although a lower log loss is not always achieved, the gap is subtle. We believe that further improvements can be made by fine-tuning hyperparameters, as we use the same learning rate and regularization ratio for all methods which maybe sub-optimal. Besides, We also observed that:**(i)** Unaligned users achieved greater improvements than aligned users. This is reasonable, as the retrieved top-k records for aligned users are harder to improve beyond the ground truth one, whereas those for unaligned users significantly fill the information gaps and result in a notable difference. **(ii)** In most cases, the performance improvement in S1 is higher than S2 (except for $k = 2$ in $\mathcal{U}_{un}$, $\mathcal{U}_{al}$). This again reflects the relative importance of domains against the label.

3.2.2 ***Ablation Study on K (RQ2).*** The impact of k can be observed in both Table 1 and Figure 5. The results show that the AUC monotonically increases as k increases, demonstrating the scalability of ranking performance with respect to the size of the top-k user set, and indicating the effectiveness of both our retrieval and fusion modules. We also find that: **(i)** In scenario 1 for $\mathcal{U}_{full}$, the increase in AUC significantly slows down when k changes from 6 to 8, suggesting that further improvement is bottlenecked. This is reasonable, as overly large values of k may introduce noise. **(ii)** The reduction of log loss is not significant when k further increases and sometimes it keeps remaining and even increases, indicating that log loss is more difficult to optimize, or RA is less efficient in reducing log loss. We leave the performance improvement of RA-VFL on log loss in future work.

3.2.3 ***Ablation Study on Fusion Modules (RQ3).*** Figure 5 summarizes the contribution of fusion components, revealing that **(i)** each component outperforms the baseline method of Fed-Fill independently in most cases (gap less than 0.01% for bad cases), and **(ii)** combining both fusion methods achieves higher performance in almost all cases (only 1 bad case with subtle gap of 0.04%). Although the superiority does not always hold, the differences are subtle. This can be resolved by implementing fine-grained hyper-parameter settings, while more sophisticated architecture designs may also aid in better combining the two fusion modules.

3.2.4 ***Supplementary.*** Due to differences in data properties, such as variations in user group distribution and the importance of column sets, the performance levels of models can differ in different

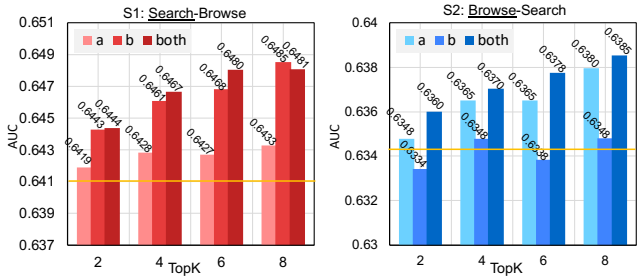

Figure 5: Results of different fusion components and amount of top-k users. The model is trained on $\mathcal{U}_{full}$. *The orange line denotes the value of baselines* (64.09% for S1 and 63.43% for S2).

Table 3: Supplementary results showing basic characters. Models are trained on various user sets and evaluated for both unaligned and aligned users.

	Fields	Search-only		Browse-only	
Method	Train\Test	un	al	un	al
Local	un	64.10%	59.38%	60.57%	59.82%
	al	63.35%	61.08%	60.51%	60.64%
	full	64.41%	61.30%	61.52%	61.06%
		Search-default		Browse-default	
Fed-Fill	un	64.12%	59.32%	60.61%	59.86%
	full	64.29%	63.08%	61.33%	63.32%
		Search-Browse		Browse-Search	
Fed-Raw*	un	65.41%	62.07%	65.22%	61.39%
	al	64.58%	62.96%	65.02%	63.03%
	full	65.95%	63.32%	66.14%	63.29%

data scenarios. Although these variations do not affect the fair comparison and validation of the superiority of our methods, they are important to consider in understanding the results and avoiding ambiguity. Therefore, we conducted supplementary experiments to help understand some basic characteristics in our VFL settings. The results are shown in Table 3, where "Fed-Raw*" is the hypothetical perfect condition referred as upper bound, "Local" is the basic condition off non-federation referred as lower bound, "Fed-Fill" is the realistic setting we consider. We summarize some key observations: **1) Search > Browse**: As shown in "Local", all experiments using search domain achieves higher AUC. columns in domain search are more useful to make prediction. **2) User Volume Benefits**: Results under fullset always achieve better results, no matter which method is chosen. This shows the necessity of including unaligned set in VFL. **3) User Set Bias**: As shown in "Local", Model trained on $\mathcal{U}_{al}$ generalizes well to $\mathcal{U}_{un}$, but the reverse is not true. This implies bias among user sets, as like the non-i.i.d problem in horizontal federated learning[13]. **4) Segmentation matters**: As shown in "Fed-Raw*", results in S1 and S2 are all different even the whole column set is identity. This is reasonable since the change of bottom models and cut layers would lead to some difference.

4 RELATED WORK

Retrieval Enhanced Machine Learning: The idea of retrieval-enhancement machine learning were firstly introduced in open-domain question answering [4, 12, 15, 18, 31] and have since been continuously adopted in large language modeling [2, 7, 16–18, 28, 30]. In these context of natural language processing (NLP), the primary purpose of adopting retrieval is to disentangle the parametric model from memorization. Typically, retrieval-based models work in a "retrieve-and-read" manner, in which a retriever searches from a huge text database to find relevant sentences or passages for a query, and the reader represent and fuse the knowledge in retrieved results to get enhanced representation for that query. Inspired by the success of retrieval-enhancement in NLP, [1, 27] have adopted it in recommendation tasks. They designed recommendation-oriented retrievers to search for relevant samples or users for data augmentation, thus utilizing cross-sample or cross-user correlation to enhance user interest representation. To summarize, the key distinguishing factor among these works lies in the design of the retriever and the reader, which are tailored to specific problem settings. For a more comprehensive understanding, interested readers can refer to a recent survey paper [37]. Our paper focuses on utilizing REML to address the challenges associated with VFL, which has not been considered in the works mentioned above.

Federated Recommendation Our paper focus on improving the utility of recommendation tasks in two-party VFL settings, considering the narrow data scope problem. This motivation clearly distinguishes our work from many related works in horizontal FL[21, 22, 25, 36]. Thus, we focus on discussing most relevant VFL works with similar purposes to ours. FedMVT[14] considers to use all parties data by complement both the missed representations and labels for unaligned samples in both party. While this solution comprehensively uses all available data, it is computationally inefficient and not tailored for recommendation tasks. It is also differs to us in the consideration of the unaligned unlabeled data of the passive paties. Three distillation based works [19, 20, 29] propose to decouple the dependence of online serving with federation and meanwhile enables full user set inference. However, their goal of achieving local serving bans the use of B-side inputs completely for prediction, creating a significantly different and more challenging ill-pose setting from ours. [35] proposes an novel diffusion-based alternative training algorithm to utilize the unaligned data from active parties. However, its focus is only on improving performance for aligned samples, and it does not include unaligned samples in federated serving. All of these works, including ours, have a common interest in utilizing unaligned data for training. However, they differ in the serving stage settings: 1) local serving for all samples[19, 20, 29], 2) federated serving for aligned samples only[35], and 3) federated serving for all samples[14]. Although developed for different industry scenarios, our retrieval augmentation methods can be incorporated into these works to further enhance their performance. Overall, Our main contribution is proposing the first retrieval-based algorithm to enhance VFL's performance by explicitly incorporating cross-user correlations. This approach shows notable promise and warrants further exploration in other VFL settings and recommendation tasks.

5 CONCLUSION AND FUTURE WORK

In this paper, we introduce RA-VFL, ***the first retrieval-based VFL framework*** designed to address the narrowed data scope problem of vertical federated learning. We propose a general "*retrieve-and-fusion*" framework to acquire enhanced representations for all parties, which results in improved performance. Our experimental results demonstrate that RA-VFL can ***achieve percentile-level AUC lifts on an industry CTR dataset***, revealing its enormous potential in real applications and further research. In the future, we plan to investigate finer-grained retrieval strategies and tailored fusion modules for different tasks to enhance the effectiveness of RA-VFL. Additionally, we will explore the feasibility of using RA-VFL as a plug-in for related works to further boost its potential.

6 ACKNOWLEDGEMENTS

We thank Jinpeng Wang and Guanghao Meng for the valuable sharing experience of retrieval-based learning in natural language, recommendation, and multi-modality. We also thank Tao Dai and Bin Chen for their support of this project.

REFERENCES

[1] Shuqing Bian, Wayne Xin Zhao, Jinpeng Wang, and Ji-Rong Wen. 2022. A Relevant and Diverse Retrieval-enhanced Data Augmentation Framework for Sequential Recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2923–2932.

[2] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*. PMLR, 2206–2240.

[3] Iker Ceballos, Vivek Sharma, Eduardo Mugica, Abhishek Singh, Alberto Roman, Praneeth Vepakomma, and Ramesh Raskar. 2020. SplitNN-driven Vertical Partitioning. *CoRR* abs/2008.04137 (2020). arXiv:2008.04137 https://arxiv.org/abs/2008.04137

[4] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051* (2017).

[5] Tianyi Chen, Xiao Jin, Yuejiao Sun, and Wotao Yin. 2020. Vafl: a method of vertical asynchronous federated learning. *arXiv preprint arXiv:2007.06081* (2020).

[6] Fangcheng Fu, Huanran Xue, Yong Cheng, Yangyu Tao, and Bin Cui. 2022. BlindFL: Vertical Federated Machine Learning without Peeking into Your Data. In *Proceedings of the 2022 International Conference on Management of Data*. 1316–1330.

[7] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*. PMLR, 3929–3938.

[8] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*. PMLR, 3929–3938.

[9] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. 173–182.

[10] Yaochen Hu, Di Niu, Jianming Yang, and Shengping Zhou. 2019. FDML: A collaborative machine learning framework for distributed features. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2232–2240.

[11] Mingkai Huang, Hao Li, Bing Bai, Chang Wang, Kun Bai, and Fei Wang. 2020. A federated multi-view deep learning framework for privacy-preserving recommendations. *arXiv preprint arXiv:2008.10808* (2020).

[12] Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282* (2020).

[13] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. 2021. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning* 14, 1–2 (2021), 1–210.

[14] Yan Kang, Yang Liu, and Tianjian Chen. 2020. FedMVT: Semi-supervised Vertical Federated Learning with MultiView Training. *ArXiv* abs/2008.10838 (2020).

[15] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906* (2020).

[16] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172* (2019).

[17] Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2021. Internet-augmented dialogue generation. *arXiv preprint arXiv:2107.07566* (2021).

[18] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[19] Wenjie Li, Qiaolin Xia, Hao Cheng, Kouyin Xue, and Shu-Tao Xia. 2022. Vertical semi-federated learning for efficient online advertising. *arXiv preprint arXiv:2209.15635* (2022).

[20] Wenjie Li, Qiaolin Xia, Junfeng Deng, Hao Cheng, Jiangming Liu, Kouying Xue, Yong Cheng, and Shu-Tao Xia. 2022. Semi-Supervised Cross-Silo Advertising with Partial Knowledge Transfer. *arXiv preprint arXiv:2205.15987* (2022).

[21] Feng Liang, Weike Pan, and Zhong Ming. 2021. Fedrec++: Lossless federated recommendation with explicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 4224–4231.

[22] Yujie Lin, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Dongxiao Yu, Jun Ma, Maarten de Rijke, and Xiuzhen Cheng. 2020. Meta matrix factorization for federated rating predictions. In *SIGIR*. 981–990.

[23] Yang Liu, Yan Kang, Tianyuan Zou, Yanhong Pu, Yuanqin He, Xiaozhou Ye, Ye Ouyang, Ya-Qin Zhang, and Qiang Yang. 2022. Vertical Federated Learning. *arXiv preprint arXiv:2211.12814* (2022).

[24] Linyuan Lü, Matúš Medo, Chi Ho Yeung, Yi-Cheng Zhang, Zi-Ke Zhang, and Tao Zhou. 2012. Recommender systems. *Physics reports* 519, 1 (2012), 1–49.

[25] Khalil Muhammad, Qinqin Wang, Diarmuid O'Reilly-Morgan, Elias Tragos, Barry Smyth, Neil Hurley, James Geraci, and Aonghus Lawlor. 2020. Fedfast: Going beyond average for faster training of federated recommender systems. In *SIGKDD*. 1234–1242.

[26] Daniel Peterson, Pallika Kanani, and Virendra J Marathe. 2019. Private federated learning with domain adaptation. *arXiv preprint arXiv:1912.06733* (2019).

[27] Jiarui Qin, Weinan Zhang, Rong Su, Zhirong Liu, Weiwen Liu, Ruiming Tang, Xiuqiang He, and Yong Yu. 2021. Retrieval & interaction machine for tabular data prediction. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1379–1389.

[28] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083* (2023).

[29] Zhenghang Ren, Liu Yang, and Kai Chen. 2022. Improving Availability of Vertical Federated Learning: Relaxing Inference on Non-overlapping Data. *ACM Transactions on Intelligent Systems and Technology (TIST)* (2022).

[30] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567* (2021).

[31] Devendra Singh, Siva Reddy, Will Hamilton, Chris Dyer, and Dani Yogatama. 2021. End-to-end training of multi-document reader and retriever for open-domain question answering. *Advances in Neural Information Processing Systems* 34 (2021), 25968–25981.

[32] Praneeth Vepakomma, Otkrist Gupta, Tristan Swedish, and Ramesh Raskar. 2018. Split learning for health: Distributed deep learning without sharing raw patient data. *arXiv preprint arXiv:1812.00564* (2018).

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

[34] Kang Wei, Jun Li, Chuan Ma, Ming Ding, Sha Wei, Fan Wu, Guihai Chen, and Thilina Ranbaduge. 2022. Vertical federated learning: Challenges, methodologies and experiments. *arXiv preprint arXiv:2202.04309* (2022).

[35] Penghui Wei, Hongjian Dou, Shaoguo Liu, Rongjun Tang, Li Liu, Liang Wang, and Bo Zheng. 2023. FedAds: A Benchmark for Privacy-Preserving CVR Estimation with Vertical Federated Learning. *arXiv preprint arXiv:2305.08328* (2023).

[36] Liu Yang, Ben Tan, Vincent W. Zheng, Kai Chen, and Qiang Yang. 2020. *Federated Recommendation Systems*. Springer International Publishing, Cham, 225–239. https://doi.org/10.1007/978-3-030-63076-8_16

[37] Hamed Zamani, Fernando Diaz, Mostafa Dehghani, Donald Metzler, and Michael Bendersky. 2022. Retrieval-enhanced machine learning. *arXiv preprint arXiv:2205.01230* (2022).