FedVQA: Personalized Federated Visual Question Answering over Heterogeneous Scenes

Mingrui Lao
m.lao@liacs.leidenuniv.nl
Leiden University
The Netherlands

Nan Pu*
nan.pu@unitn.it
University of Trento
Italy

Zhun Zhong
zhunzhong007@gmail.com
University of Nottingham
United Kingdom

Nicu Sebe
niculae.sebe@unitn.it
University of Trento
Italy

Michael S. Lew
m.s.k.lew@liacs.leidenuniv.nl
Leiden University
The Netherlands

ABSTRACT

This paper presents a new setting for visual question answering (VQA) called personalized federated VQA (FedVQA) that addresses the growing need for decentralization and data privacy protection. FedVQA is both practical and challenging, requiring clients to learn well-personalized models on scene-specific datasets with severe feature/label distribution skews. These models then collaborate to optimize a generic global model on a central server, which is desired to generalize well on both seen and unseen scenes without sharing raw data with the server and other clients. The primary challenge of FedVQA is that, client models tend to forget the global knowledge initialized from central server during the personalized training, which impairs their personalized capacity due to the potential overfitting issue on local data. This further leads to divergence issues when aggregating distinct personalized knowledge at the central server, resulting in an inferior generalization ability on unseen scenes. To address the challenge, we propose a novel federated pairwise preference preserving (FedP^3) framework to improve personalized learning via preserving generic knowledge under FedVQA constraints. Specifically, we first design a differentiable pairwise preference (DPP) to improve knowledge preserving by formulating a flexible yet effective global knowledge. Then, we introduce a forgotten-knowledge filter (FKF) to encourage the client models to selectively consolidate easily-forgotten knowledge. By aggregating the DPP and the FKF, FedP^3 coordinates the generic and the personalized knowledge to enhance the personalized ability of clients and generalizability of the server. Extensive experiments show that FedP^3 consistently surpasses the state-of-the-art in FedVQA task.

CCS CONCEPTS

• **Computing methodologies** → *Computer vision.*

KEYWORDS

Personalized Federated Learning, Visual Question Answering, Knowledge Preserving, Pairwise Preference

ACM Reference Format:
Mingrui Lao, Nan Pu, Zhun Zhong, Nicu Sebe, and Michael S. Lew. 2023. FedVQA: Personalized Federated Visual Question Answering over Heterogeneous Scenes . In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, 12 pages. https://doi.org/10.1145/3581783.3611958

1 INTRODUCTION

In recent years, the field of visual question answering (VQA) has attracted significant attention due to its ability to comprehend textual queries based on images and deduce accurate answers [5, 48]. State-of-the-art VQA models [4, 28, 51, 56] have achieved superior performance across various scenes via large-scale centralized training [62]. However, the utilization of such training paradigms poses a significant challenge to privacy constraints in practical VQA applications [6]. For example, sensitive data obtained from educational settings cannot be shared with other clients or a central server, as shown in Fig. 1. Hence, a decentralized training paradigm is necessary for real-world VQA systems to address this challenge.

Recently, federated learning (FL) [26, 38] has been proposed as a privacy-aware and distributed framework for training models without sharing data with a central server or other clients [46]. To the best of our knowledge, however, there have been limited studies focusing on federated VQA tasks. In addition, compared with the conventional FL on identically distributed (iid) data, the VQA samples collected from different local clients typically involves heterogeneous feature and label distributions, including diverse visual content captured from various realistic scenes (e.g., Fig. 1), as well as inconsistent answer distributions caused by different scene-specific questions. Considering this, we propose a challenging yet practical VQA task, namely personalized federated VQA (FedVQA). The goal of FedVQA task is to train personalized VQA client models for distinct visual scenes, while optimizing a generic model to generalize well on unseen scenes, through client collaboration under the privacy constraint. This target leads to two main challenges. Firstly, local VQA models are prone to forget the generic knowledge aggregated from server during the personalized training, thereby encountering the potential overfitting issue, and performing worse on local data. Secondly, since the training data distributed at local

*Corresponding Author.

MM '23, October 29-November 3, 2023, Ottawa, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3611958

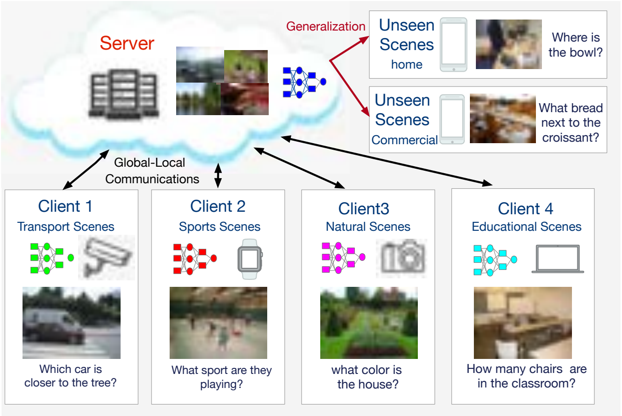

Figure 1: The federated setting for VQA over heterogeneous visual scenes. Given a pre-trained VQA model, we require each participated clients to train personalized model to perform well on their local data (e.g., transports, sports, natural and educational scenes). Meanwhile, the central server is expected to aggregate a generic global model to generalize on the testing data in unseen scenes (e.g., shopping and home).

clients includes scene-specific images and label distributions, the potential conflicts among personalized knowledge are unfavorable for efficient global knowledge aggregation, resulting in the central server with a degraded ability to generalize on unseen visual scenes.

To overcome these challenges, we introduce a novel federated pairwise preference preserving (FedP^3) framework that prevents clients models from forgetting global knowledge when learning from local data, so as to collaboratively optimize both generic and personalized models. Based on the commonly-used FedAvg [46] pipeline (detailed in Sec. 3.2), FedP^3 follows a knowledge preserving (KP) strategy that exploits the soft logits from global model as the generic knowledge, and transfer it to the local model as the regularization during the personalized training. However, we declare that the logits-based constraint achieved by Kullback-Leibler (KL) divergence is overly strict in knowledge preserving, and even disturbs clients' balance between consolidating generic knowledge and acquiring personalized knowledge. To tackle this issue, we propose a novel differentiable pairwise preference (DPP) method that formulates the distilled knowledge as the pairwise binary comparisons among significance of answer prediction, instead of the absolute value of predictive probabilities, which reveals the reasoning behaviour of global model in a relaxed yet effective manner. Besides, we present a forgotten-knowledge filter (FKF) that seeks to generate a forgotten-knowledge driven label distribution to capture the easily-forgotten classes during local training, and then adaptively filters a significant answer subset involved in pairwise preference. Benefited from FKF in DPP, FedP^3 not only further enhances the performance of both local and global models, but also remarkably reduces the computational complexity of knowledge preserving.

After the last round of global-local communication, the aggregated model serves as the generic global model, which iteratively accumulates abundant knowledge over diverse scenes from local clients. Meanwhile, we consider the final-round local model before weighted average as the final personalized VQA model in each client. By integrating the DPP and FKF, our FedP^3 framework coordinates the generic and the personalized knowledge, thereby achieving state-of-the-art performance on our FedVQA setting.

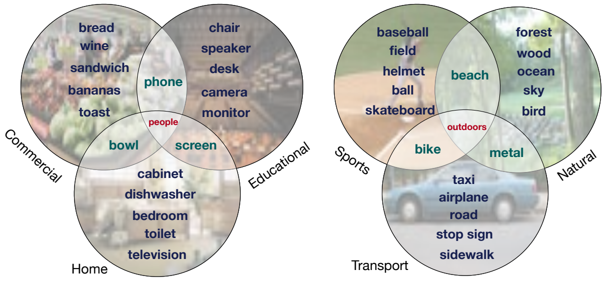

Figure 2: The scene-specific answers (in dark blue) from each local dataset represented in a specific visual scene, and some general answers (in red and green) co-exist in several scenes.

The contributions of this work are summarized as:

- We propose a novel yet practical setting FedVQA for federated VQA over heterogeneous scenes. It not only concerns the performances of local models, but also considers the global model's generalization ability on unseen scenes.
- We propose a novel federated pairwise preference preserving FedP^3 approach to coordinate the generic and personalized knowledge, thereby improving the model's representative ability on both seen and unseen scenes.
- Extensive experiments show that our FedP^3 achieves competitive performance with the state-of-the-art competitors.

2 RELATED WORKS

2.1 Visual Question Answering

Visual Question Answering (VQA) is a prevalent vision-language task, which concentrates on answering natural language question according to the given image, necessitating the comprehensive understanding and reasoning over both visual and textual modalities [5, 48]. Most of earlier VQA works seek to establish efficient model architectures to achieve fine-grained vision-language interactions for answer prediction, such as multimodal fusion [13, 33, 57], attention [4, 11, 21, 51, 56], and large-scale pre-training models [28, 36, 61]. Recently, increasing amount of researches [17, 24, 31, 34, 47] focus on improving reasoning robustness in VQA task, thereby alleviating some undesired model behaviour, such as language bias [2, 7, 16, 32] and multimodal inputs variations [49, 50]. The remarkable performance achieved by these methods is attributed to the centralized training [62] over large-scale and well-collected datasets [15, 22, 25].

However, such a training paradigm is inefficient for real VQA application scenes, due to the growth of the privacy awareness. To investigate this overlooked issue and address additional technical bottleneck, we propose a new FedVQA setting and accordingly introduce a new FedP^3 approach.

2.2 Personalized Federated Learning

Federated Learning (FL) is a learning paradigm that enables the training of a model across multiple client devices while maintaining local data privacy [26, 38]. The most widely adopted FL algorithm is FedAvg [46], which averages the local model parameters across different clients trained on private client datasets to learn a global model. Recent research efforts have focused on improving FedAvg from various aspects, including model convergence [18, 27], robustness [8, 43], communication [29], and non-IID clients [3, 10, 19, 23, 40, 41, 63].

To further handle the heterogeneity of data and models, personalized FL (PFL) has been introduced [30]. In contrast to traditional FL, PFL aims to learn a customized model for each client, tailored to their specific objectives. This method acknowledges the diversity of data among clients by constructing a "personalized" model that fits each client's needs. One group of techniques [41, 42] has leveraged multi-task learning (MTL) methods to incorporate clients' task objectives into the FL framework. The other group contains post-processing techniques [12, 55]. [55] with meta-learning to learn an initial model that can be adapted to each client through local fine-tuning. [55] indicates that fine-tuning can achieve comparable results to other personalized methods. In our framework, we use an MTL-based approach that can optimize generic and personalized VQA models simultaneously. While the benchmarks for conventional FL are well-established, few studies have focused on federated VQA. The most closely related work [44] proposes a vision-language FL framework with shareable networks, but only considers the scenario where clients learn different tasks (e.g., VQA and image captioning) instead of FL VQA across different scenes.

We argue that the proposed FedVQA is a practical and challenging task for two reasons. Firstly, our FedVQA not only aims to improve individual personalized models through collaborative training, but also considers the model's ability to directly deploy on unseen scenes. Secondly, since the heterogeneous data collected from different scenes include scene-specific characteristics (e.g., distinct high-frequency words in Fig. 2), the model trained on FedVQA setting has a high risk of failing to converge. To the best of our knowledge, this work is the first attempt to explore VQA tasks in personalized federated learning.

2.3 Forgetting Issue in Personalized Learning

In the PFL pipeline, models often suffer from a forgetting problem on global knowledge. To cope this issue, FedProx [39] proposes to punish overlarge parameter changes during local training. MOON [37] introduces a model-level contrastive learning to reduce feature discrepancy between the global and local models. Then, FedDyn [1] adopts the averaging of dual variables under partial participation settings to improve convergence. Recently, FedDC [14] proposes drift correction terms as penalized losses on original local objective functions with global gradient estimation. Another typical way to achieve this goal is via knowledge distillation (KD). FedMD [35] aggregates local predictions over a public dataset at the server and transfers the consensus of predictions back to clients for distilling client models. KT-pFL [58] enables each client to maintain a personalized prediction at the server to guide other clients. Recently, FedKD [54] proposes a communication-efficient federated knowledge distillation approach to enhance personalized models by leveraging the assist of global model. However, this may impair the generalizability of global model, inconsistent with the objectives of FedVQA. We experimentally validate this assumption in Tab. 2.

In contrast to these methods that directly adopt entropy-based distillation loss, we propose a novel pairwise preferece preserving approach based on relative comparisons, which flexibly reflects a model's reasoning behavior and coordinates global-local knowledge without requiring a public dataset.

3 METHODOLOGY

In this paper, we present a novel Federated Pairwise Preference Preserving (FedP^3) tailored to the FedVQA setting over heterogeneous scenes. In the following, we first elaborate the benchmark setup, which contains task definition, distribution skews, and training target, respectively. Then, we describe the basic learning pipeline to adapt the typical VQA model into the federated learning scenarios. Finally, we explicitly introduce the FedP^3 strategy.

3.1 Benchmark Formulation

Task Definition: VQA algorithm typically refers to a classification function $\mathcal{F}_{vqa}$ to learn a mapping: $\mathcal{I} \times \mathcal{Q} \rightarrow [0,1]^{|\mathcal{A}|}$ based on a centralized dataset $\mathcal{D} = \{I_i, Q_i, a_i\}_i^N$, where $I_i \in \mathcal{I}$, $Q_i \in \mathcal{Q}$ and $a_i \in \mathcal{A}$ denote image, question and answer respectively. In our FedVQA, there are n clients $C = \{C_1, C_2, \ldots, C_n\}$, each C_i equipped with a local training dataset D_i with personalized image-question training pairs, as well as a target test split $\mathcal{T}_i$. The local clients are to minimize the training loss of the personalized VQA models, i.e., min $\mathcal{L}(\theta_i; \mathcal{T}_i)$, where θ_i refers to the model parameters for the i-th client. As a result, the final learning objective is to acquire the optimal parameters of local models:

$$\left\{\widetilde{\theta}_1, \widetilde{\theta}_2, \ldots, \widetilde{\theta}_n\right\} = \arg\min \sum_{i=1}^{n} \mathcal{L}\left(\theta_i; \mathcal{T}_i\right), \tag{1}$$

where $\widetilde{\theta}_i$ denotes the optimal setting of personalized VQA model from the i-th involved client, $i \in \{1, 2, \ldots, n\}$.

Distribution Skews: As depicted in Fig. 2, FedVQA exists severe feature and label distribution skews among the VQA samples across different clients. To be specific, on the one hand, the training images derived from different local datasets are represented in different visual scenes (e.g., shopping, home, and transports), which leads to the visual domain shifts among the multiple local datasets. On the other hand, for a client responsible to tackle the questions over images in a specific scene (e.g., sports), its label distribution would be inclined to the scene-related answer candidates (e.g., tennis, frisbee, and badminton), which potentially forms the heterogeneous label distribution over participated clients.

Targets:: We summarize two learning targets in FedVQA benchmark, among which one for the personalized models in local clients, and the other for the global model in the central server. 1) The local clients attempt to acquire knowledge from their own private data, and we target on **training an efficient personalized VQA model to perform well on private data represented a specific visual scene**. 2) The central server seeks to aggregate the local models to accumulate knowledge from personalized private datasets, and send the updated global models to each participated client. On the side of server, we focus on **establishing a generic global model with strong generalizability to the VQA samples in unseen scenes.** To our best knowledge, this work is the first attempt to explore the personalized federated setting in VQA task.

3.2 Training Pipeline

To fulfill FedVQA, we use the intuitive and commonly-used FL algorithm **FedAvg** as the baseline strategy for collaborative training

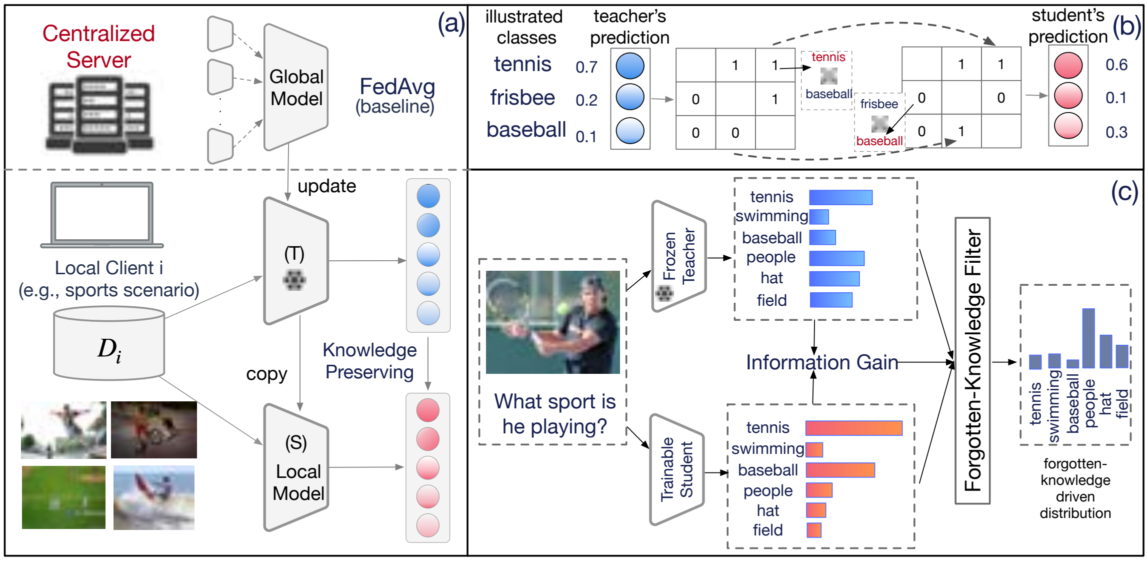

Figure 3: Conceptual illustration of Fed^3 in FedVQA benchmark, which contains three indispensable concepts: (a) knowledge preserving: the global model aggregated by *FedAvg* from central server act as a frozen teacher, so as to transfer generic knowledge to the local model (student) during personalized training. (b) pairwise preference: modelling transferred knowledge via relative comparisons among the answer significance (answer with higher probability in red wins the pairwise matchup). (c) forgotten-knowledge filter: selecting the easily-forgotten answer candidates into pairwise preference for knowledge preserving.

between central server and clients. We define the hyper-parameters C, T and E as the number of clients in the federation, the total communication rounds, and the epochs required for local training, respectively. At the beginning of the global-local communication, the global model is initialized by loading the parameters from the large-scale pre-trained vision-language model. Afterward, according to the pre-defined T and E, the server and participated clients cooperatively accumulate knowledge from distributed data in an iterative learning manner (multiple communication rounds). Specifically, in each round, the server first sends the global model to each client as the initial local model for personalized data training. Then, the client (e.g., the i-th client) locally updates the model using its own private data $\mathcal{D}_i = \left\{I_j, Q_j, a_j\right\}_j^{N_i}$, where N_i implies the total number of training instances. In FedVQA, we adopt the cross-entropy loss function to train the parameters of local model θ_i in the i-th client:

$$\mathcal{L}_{ce} = -\frac{1}{N_i}\sum_{j}^{N_i} \log\left(\mathcal{F}_{vqa}\left(I_j, Q_j; \theta_i\right)\right)\left[a_i\right]. \tag{2}$$

After finishing E-epoch local training, clients are required to return their optimized models back to the central server. Sequentially, the server will integrate a new global model θ_g by conducting a weighted average of uploaded personalized models as follows:

$$\theta_g = \frac{1}{N}\sum_{i}^{C} N_i \cdot \theta_i, \tag{3}$$

where N is the total amount of image-question pairs across all private datasets. Particularly, we exploit the aggregated model in the last communication round as the generic model, which iteratively accumulates abundant knowledge over diverse scenes from clients. Besides, we consider the final-round local model before weighted average as the final personalized VQA model in each client.

Restrictions: Intuitively, the integration of model parameters in FedAvg could effectively accumulate knowledge from decentralized training data. Nevertheless, in FedVQA, or other real-world VQA applications involving federated learning, the statistical heterogeneity inevitably exists among the data across local clients, which significantly impairs the performance of both local and global models. The main reasons are twofold. 1) After obtaining global model, clients attempt to acquire knowledge from private datasets with severe label and feature distribution shifts, which optimizes the model parameters to the local optima and deviates from the global target. 2) The global aggregation process achieved by weighted average often leads to an unwanted drift for the initialization of local clients, which plays a negative role on the model convergence.

3.3 FedP^3: Pairwise Preference Preserving

In this section, built upon the basic FedAvg strategy, we propose a novel federated pairwise preference preserving (FedP^3) for FedVQA benchmark, which contains three indispensable concepts: knowledge preserving (KP), differentiable pairwise preference (DPP), and forgotten-knowledge filter (FKF).

3.3.1 Knowledge Preserving. In FedVQA over heterogeneous scenes, the optimization direction in each local model is typically inconsistent with that in the central server, which potentially leads the clients to forget the aggregated generic knowledge initialized from global model. Particularly, for several classes whose samples do not exist in a specific client, the local training tends to gradually eliminate the predictive probabilities of such classes for local optima, thereby forgetting the general knowledge from global model. To prevent from the overfitting on local data and alleviate the forgetting issue, we introduce an intuitive KP pipeline to preserve the knowledge learned from other participants. Specifically, we store a frozen global model to regularize the local training on each client, and add a distillation term to the local task loss objective (Equ. (2)).

In the beginning of the communication round t ($t \leq T$), the i-th client updates its local model (θ_i^t) from the central server as the trainable student, and meanwhile copies a complete global model ($\widetilde{\theta_g^{t-1}}$) as the frozen teacher to store the aggregated global knowledge in the last communication round. The anti-forgetting process is to exploit the output logits ($p^T = \mathcal{F}_{vqa}\left(I_j, Q_j; \widetilde{\theta_g^{t-1}}\right)$) from teacher model to regularize the student's response ($p^S = \mathcal{F}_{vqa}\left(I_j, Q_j; \theta_i^t\right)$), thereby preventing student from forgetting the previous-learned global knowledge. Specifically, we achieve the aforementioned KP via Kullback-Leibler divergence loss $\mathcal{L}_{\text{KP}}$:

$$\mathcal{L}_{\text{KP}}\left(p^S, p^T\right) = -\sum_{a=1}^{|\mathcal{A}|} p^T(a) \log\left[\frac{p^S(a)}{p^T(a)}\right], \tag{4}$$

where $|\mathcal{A}|$ is the total number of candidates for answer prediction, and $p^S(a)$, $p^T(a)$ refers to the a-th value of p^S and p^T, respectively.

3.3.2 Differentiable Pairwise Preference. Although using KL divergence in KP pipeline can constrain knowledge discrepancy, it might be a "hard" constraint for the probabilities in the label space, especially for the VQA task with severe robustness issues. To be specific, the personalized model would encounter the plasticity issue when acquiring new knowledge from local data, due to the regularization of absolute value for answer prediction. On the contrary, DPP focuses on the relative comparisons among the predictions yielded from different answer candidates (e.g., whether the answer '*baseball*' is more important than '*swimming*' for the training sample labeled by '*tennis*'). It reveals the reasoning behavior of the teacher model in a relaxed yet effective manner. In FedVQA, we seek to fulfill knowledge preserving by leveraging the DPP, which encourages the local models efficiently to learn from local data with less forgetting of global knowledge.

Given the teacher's prediction $p^T = [p^T(0), p^T(1), ..., p^T(|\mathcal{A}|)]$ as $\mathcal{P}^T$, we define DPP by:

$$\mathcal{P}^T = \begin{bmatrix} M\left(p^T(1), p^T(1)\right) & \dots & M\left(p^T(N), p^T(1)\right) \\ \vdots & \ddots & \vdots \\ M\left(p^T(1), p^T(N)\right) & \dots & M\left(p^T(N), p^T(N)\right) \end{bmatrix}, \tag{5}$$

where $M(\cdot)$ is the function of pairwise matchup to compare the significance between two answer candidates. Specifically, given the predictive probabilities of the i-th and j-th answer, the function is:

$$M(p^T(i), p^T(j)) = \begin{cases} 1 & \text{if } p^T(i) \succ p^T(j), \\ 0 & \text{if } p^T(j) \succ p^T(i). \end{cases} \tag{6}$$

Analogously, we can obtain the pairwise preference on the side of student model as $\mathcal{P}^S$. Then, the loss objective of pairwise preference driven knowledge preserving $\mathcal{L}_{pp}$ could be achieved through punishing the inconsistency between $\mathcal{P}^T$ and $\mathcal{P}^S$:

$$\mathcal{L}_{pp} = \sum_i \sum_j \left| M\left(p^T(i), p^T(j)\right) - M\left(p^S(i), p^S(j)\right)\right|. \tag{7}$$

One practical difficulty for pairwise preference is that the matchup function $M(\cdot)$ is discontinuous, which is not compatible with the general deep neural network optimization, such as SGD [9] and AdamW optimizer [45]. To enable the PP to perform the gradients back-propagation in neural networks, we propose to adopt a sigmoid-like function $g(\cdot)$ to approximate the matchup function:

$$g(x) = \frac{1}{1 + e^{-2x}}, \tag{8}$$

Therefore, we reformulate the Equ. (6) as the a differentiable counterpart:

$$M\left(p^T(i), p^T(j)\right) = g\left(p^T(i) - p^T(j)\right) = \frac{1}{1 + e^{-2(p^T(i) - p^T(j))}}, \tag{9}$$

and the derivative of $M(\cdot)$ can be formulated as:

$$\frac{\partial M\left(p^T(i), p^T(j)\right)}{\partial p^T(j)} = \frac{-2e^{-2(p^T(i) - p^T(j))}}{\left[1 + e^{-2(p^T(i) - p^T(j))}\right]^2}, \quad j \neq i. \tag{10}$$

3.3.3 Forgotten-Knowledge Filter. DPP produces a high dimensional binary matrix of quadratic expansion (Equ. (5)), which leads to a non-negligible $O(n^2)$ computational complexity. An intuitive solution to mitigate this issue is to select a subset of answer candidates for DPP, instead of taking all answer pairs into consideration. To this end, we propose a novel forgotten-knolwedge filter (FKF) strategy, which concentrates on creating a rectified label distribution to capture the easily-forgotten knowledge during local training.

In FKF, we assume the selected answers for pairwise preference should be strongly related to the forgotten global knowledge in each local client. Specifically, as illustrated in Fig. 3(c), for the client tailed to sports scenes, its personalized model typically learns from samples labeled by sports-related answers (e.g., *tennis* and *baseball*), while gradually ignoring the learned knowledge involved in some general or label-irrelevant classes (e.g., *people* and *field*). The answer selection for the latter is capable of improving the efficacy of knowledge preserving, and meanwhile reducing the computational complexity caused by pairwise comparisons.

To this end, as shown in Fig. 3(c), we propose to establish a forgotten-knowledge driven label distribution to describe the forgotten knowledge during local training, which is mainly determined by the comparison between predictions from the student and teacher. Specifically, the probability of the i-th class ($r(i)$) in the distribution r can be represented as:

$$r(i) = \text{softmax}\left(\log\left(p^T(i)\right) - \log\left(p^S(i)\right)\right). \tag{11}$$

During the local training, the trainable local model unavoidably forgets the scene-irrelevant knowledge on unrelated classes (e.g., the k-th answer) with lower probability (e.g., $p^S(k)$). According to the Equ. (11), the probability of easily-forgotten class k in the forgotten knowledge driven distribution $r(k)$ would be higher than those of scene-relevant classes. Considering the parameters in local

Algorithm 1: FedP^3

```
Input: Decentralized datasets $\{D_i\}_{i=1}^N$ from $N$ local clients
  $N$ clients' datasets $\{D_i\}_{i=1}^N$, Total communication round
  $T$,Epochs for each communication rounds $E$, learning rate $\eta$,
  batch size $b$
Output: The global model $\theta_g^T$, local models $\theta_1^T$, $\theta_2^T$,..., $\theta_N^T$ in
  the final (T-th) communication round.
ServerExecute:
Initialize the global model $\theta_g^0$ in the server
for $t = 0, \ldots, T-1$ do
    for $i \in N$ in parallel do
        $\theta_i^t \leftarrow$ ClientUpdate $\left(i, \theta_g^t, D_i\right)$
    end
    $\theta_g^{t+1} \leftarrow \frac{1}{|N|} \sum |D_i| \theta_i^t$    ▷ Eq.(3)
end
return $\theta_g^T$
ClientUpdate: $(i, \theta_g^t, D^i)$
$\theta_i^t \leftarrow \theta_g^t$
for epoch $e = 1, \ldots, E$ do
    for batch $b = \{v, q, a\} \in D_i$ do
        $\mathcal{L}_{p^3,i} \leftarrow |\mathcal{P}^T - \mathcal{P}^S|$    ▷ Eq.(14)
        $\mathcal{L}_{ce,i} \leftarrow \log\left(\mathcal{F}_{vqa}\left(v, q; \theta_i^t\right)\right)[a]$    ▷ Eq.(2)
        $\mathcal{L}_i \leftarrow \mathcal{L}_{ce,i} + \mathcal{L}_{p^3,i}$    ▷ Eq.(15)
        $\theta_i^t \leftarrow \theta_i^t - \eta \nabla \mathcal{L}\left(\theta_i^t, b\right)$
    end
end
return $\theta_i^t$ to the server
```

and global models are the same in the beginning of the communication round ($p^T = p^S$), we add an information gain based function into the Equ. (11), and the final distribution r is defined as follows:

$$r(i) = \text{softmax}\left(\log\left(p^T(i)\right) - \log(\frac{H_T}{H_S}) \cdot \log\left(p^S(i)\right)\right), \tag{12}$$

$$H_T = \sum_{i}^{|\mathcal{A}|} P_T(i) \log P_T(i), \tag{13}$$

where H_T and H_S are the information entropies of the teacher's and student's predictions, and H_T/H_S denotes the information gain for local model to accumulate knowledge from decentralized data based on the initialization of global model. For instance, when the client optimizes the model parameters to the local optima, its predictive uncertainty for answer candidates would be gradually decreased, and the influence of student's prediction could be considered more to build the forgotten knowledge based distribution $r(i)$.

Then, we fulfill the FKF via choosing the Top-N most influenced answers in the established distribution $r(i)$, where we formulate the selected answer subset as $\mathcal{S} \subseteq \mathcal{A}$. Finally, the loss function of our propose FedP^3 for knowledge preserving $\mathcal{L}_{p^3}$ is defined as:

$$\mathcal{L}_{p^3} = \sum_{i}^{|\mathcal{S}|} \sum_{j}^{|\mathcal{S}|} \left| M\left(p^T(i), p^T(j)\right) - M\left(p^S(i), p^S(j)\right) \right|. \tag{14}$$

Table 1: The statistics of decentralized datasets over six different visual scenes in FedVQA benchmark.

Scenes	Train	Test	Involved sub-categories of scenes
Commercial	19573	6473	restaurant, market, pharmacy, bakery...
Educational	13472	4225	campus, art gallery , music studio...
Transport	12384	4160	airport, subway , crosswalk, galley...
Natural	14820	4512	forest, mountain, marsh, underwater...
Sports	14784	5120	ballroom, arena, gymnasium, ski slope...
Home	14498	4353	kitchen, bedroom, bathroom, closet...

Algorithmic Pipeline: Based on the aforementioned crucial concepts in our proposed FedP^3, the total loss function in the t-th communication($t \geq 2$ due to the updating process of server) is:

$$\mathcal{L}_{total} = \mathcal{L}_{\text{ce}} + \lambda \mathcal{L}_{p^3}, \tag{15}$$

where the λ is a trade-off factor applied to adjust the contributions of the loss terms between acquiring new knowledge in local data, and preserving previous knowledge from central server. The detailed descriptions about how our method works are summarized in Algorithm 1. The testing phase is performed only once by using aggregated global model and personalized local models obtained in the final communication round.

4 EXPERIMENTS

4.1 Datasets

To build the decentralized datasets for different participated clients under heterogeneous visual scenes, we follow the scene-centric Places365 database [60] and use the pre-trained model to classify the images in GQA [22], which is a large-scale VQA datasets asking about images in realistic scenes. Based on the referenced taxonomy in Place365 [60], we divide the GQA dataset into six personalized datasets, among which each dataset tailored to answer questions about a specific visual scenes (e.g. transportation, sport, natural, home, educational, and commercial scenes). The detailed information including the amount of training and test samples, as well as the involved scene subcategories contained in each decentralized dataset are described in Tab. 1. It is noteworthy that each VQA instance selected in a specific category is computed by a high classification confidence score by pre-trained scene recognition model.

4.2 Implementation Details

For the setting of FL, we define the number of participated clients $N = 4$, and the amount of datasets represented in unseen visual scenes for generalizability testing is $M = 2$. The total communication rounds $T = 5$, and the epochs for local training in each communication round is $E = 2$. To train the personalized model over local dataset, we optimize model parameters via the AdamW optimizer [45] with a learning rate of e^{-4}. The minibatch size is set to 32 distributed on two GPUs. On the side of model architecture, we conduct the federated experiments on the widely-used pretrained ViLT models, where the last 3 layers are trainable. For the structure of task classifier, it contains two layers of non-linear MLP with LayerNorm [45] to predict the probabilities over 1642 answer candidates. Finally, we select the trade-off factor $\lambda = 1$ to adjust contributions between local training and knowledge preserving.

Table 2: Comparisons with state-of-the-art methods for federated learning in FedVQA, where the four datasets (transports, sports, educational, and natural scenes) participate the federated training, and the other two datasets are utilized (home and commercial scenes) for the generalization of unseen scenes. Best and second best numbers are in bold and underlined.

method Scene	DT	FedAvg [46]	FedProx [39]	MOON [37]	FedKD [54]	FedDC [14]	ST [20]	SP [46]	CRD [52]	DKD [59]	FedP^3 (Ours)	CT
Transport	42.97	45.37	45.21	45.83	45.53	45.45	45.24	45.88	45.37	45.57	**46.06**	49.27
Sports	43.19	44.87	45.13	45.97	45.35	45.86	44.66	45.76	45.11	44.91	**46.39**	51.12
Educational	37.56	40.95	41.13	40.85	41.78	41.23	41.41	41.51	41.78	41.83	**42.21**	46.84
Natural	50.29	51.48	51.27	51.41	51.35	51.66	51.52	51.38	51.75	51.54	**52.00**	56.11
Generalization over unseen scenes												
Home	-	35.01	34.89	35.91	34.75	36.18	34.11	35.13	35.49	35.88	**36.76**	41.85
Commercial	-	29.46	30.04	31.13	29.11	31.37	30.60	29.81	30.71	31.17	**32.01**	34.88

4.3 Comparative Approaches

We divide the to-be-compared 9 state-of-the-art methods in two groups. Approaches in first group are specially-designed for FL: 1) *FedAvg* [46]: the baseline strategy to aggregate trained local models by averaging their parameters 2) *FedProx* [39]: restricts the local updates by proposing a regularization of L2-norm distance. 3) *MOON* [37]: utilizes the similarity between model representations to correct the local training of individual clients. 4) *FedKD* [54]: focuses on training efficient personalized models via mutual knowledge distillation without parameter communication between client and server. 5) *FedDC* [14]: exploits a learned local drift variable to bridge the gap between local and global models. The approaches in the other group follow the idea of the knowledge preserving, and form the global knowledge from different perspectives: 6) *ST* [20]: soft targets. 7) *SP* [53]: semantic correlations 8) *CRD* [52]: contrastive representation, and 9) *DKD* [59]: target and non-target logits-based knowledge. We take Decentralized Training (DT) and Centralized Training (CT) as the references for lower and upper bounds.

4.4 State-of-the-art Comparisons

In this section, we aim to compare our propose method with aforementioned state-of-the-art strategies in FedVQA benchmark over six heterogeneous scenes. To simultaneously evaluate the performance for both personalized and generic models, we exploit four datasets to participate the federated training, while the other two datasets only available for generalization over unseen scenes. Besides, to validate the robustness of our method towards scene variations in federated learning, we build two scenarios where involved datasets for generalizability testing are entirely different. From the federated scenarios in Tab. 2 and 3, we have following observations:

1) Even though *FedAvg* improves the performance over the lower bound *DT*, there is still a hugh accuracy gap towards the centralized learning (*CT*) in both scenarios. It verifies that the label and feature distribution skews are severe in FedVQA benchmark. We can also notice that, the clients for sports and natural scenes co-existed in both two federated training perform worse in the second scenario (Tab. 3). It can explained by the fact that, compared with transports and educational scenes, federated learning with clients in home and commercial datasets involves more significant distribution shifts.

2) Among methods specialized for federated learning, *FedProx* yields comparative accuracy with *FedAvg*, and the other three approaches produce better results in terms of local personalization on the first four datasets. For generlizability, *FedKD* slightly impair the performance due to the negligence of global knowledge preserving, while *FedDC* achieves remarkable accuracy boost benefited from the learned local drift variable. Following the idea of knowledge preserving, three advanced knowledge distillation (*SP*, *CRD* and *DKD*) achieve better results than transferring soft logits (*ST*) to local models, mainly because the proposed batch-wise similarity, contrastive learning, and target-based prediction decomposition establish better representations of global knowledge in central server.

3) From results in two scenarios, our proposed FedP^3 is remarkably superior to the baseline *FedAvg* strategy, whose performance occupies all the first places for four participated clients in personalized learning. It powerfully supports that preserving global knowledge in our method facilitates local models to accumulate knowledge from their own private datasets, instead of suppressing their personalization. Besides, the global model trained by FedP^3 shows strong generalizability over unseen visual scenes (last two rows), which reveals that proposed pairwise preference could effectively form the generic knowledge aggregated from central server.

4.5 Ablation Study

We perform extensive ablation studies on the federated scenarios depicted in Tab. 4, where Avg.(Loc) is the average accuracy obtained from four local models in transports, sports, educational, and natural scenes, while Avg.(Glo) denotes the generalization results from global model over unseen home and commercial scenes.

Effectiveness of Different Concepts: We validate the contributions for different concepts in FedP^3 built upon the baseline *FedAvg* approach. From the rows 2-4 in Tab. 4, exploiting soft prediction ($T = 2$) from global model for knowledge preserving would slightly improve the average accuracy, while the other settings ($T = 1, 3$) degrade the performance of *FedAvg*. This is because the predictive value based regularization tends to restrict the local models (student) to obtain personalized knowledge when reviewing global knowledge. In contrast, pairwise preference alleviates this issue via modeling the relative comparisons on the sides of answer significance. The last five rows depict the answer subset selection for pairwise preference according to different label distributions $r(i)$. We can notice that using the distribution from global model performs better than the random ($\mu(i)$) and local distributions ($p^S(i)$), while it fails to reveal the forgotten knowledge during personalized training. Compared with the Equ. (10), leveraging the information

Table 3: Comparisons with the state-of-the-art for federated learning in FedVQA, where sports, home, natural, and commercial scenes participate the federated training, as well as transports and educational scenes for the generalization of unseen scenes.

method / Scene	DT	FedAvg [46]	FedProx [39]	MOON [37]	FedKD [54]	FedDC [14]	ST [20]	SP [46]	CRD [52]	DKD [59]	FedP^3 (Ours)	CT
Sports	43.19	43.62	43.89	44.21	43.71	44.41	44.42	44.01	44.67	44.51	**44.55**	51.10
Home	38.53	39.18	39.01	39.27	39.22	39.07	38.60	39.27	39.28	39.23	**39.43**	46.73
Natural	50.29	50.51	50.24	50.79	50.67	50.97	50.48	51.23	51.03	51.45	**51.65**	56.84
Commercial	37.26	38.37	38.41	38.93	38.95	38.92	38.76	39.15	38.87	39.29	**39.40**	44.74
generalization over unseen scenes												
Transport	-	35.23	35.28	35.88	34.81	36.42	35.51	35.98	35.95	35.63	**37.07**	41.21
Educational	-	34.74	34.84	35.36	34.69	35.48	34.11	34.43	35.19	35.27	**36.03**	39.15

Table 4: Ablation studies of three concepts in FedP^3.

Component	Setting	Avg.(Loc)	Avg.(Glo)
FedAvg	Baseline	45.67	32.24
+Knowledge Preserving	T=1	45.85	31.52
	T=2	45.71	32.36
	T=3	44.42	30.62
+Pairwise Preference	all answers	46.16	33.58
+Forgotten-Knowledge Filter	$r(i) = \mu(i)$	45.97	32.85
	$r(i) = p^S(i)$	46.31	32.05
	$r(i) = p^T(i)$	46.41	33.25
	Equ. (10)	46.50	34.18
	Equ. (11)	**46.67**	**34.57**

gain H_T/H_S in Equ. (11) consistently enhances the performance on both personalized and generic models, with accuracy boosts of 1% and 2.5% over baseline *FedAvg*.

Accuracy vs Complexity: For the personalized answer selection, we explore the trade-off between the computational complexity based on the amount of to-be-selected answer candidates, and the performance of global (Avg.(Glo)) and local (Avg.(Loc)) models. In Fig. 4, we compared the knowledge preserving with soft targets (*KP*), whose the complexity is equal to the total number of classes (1642), with our FedP^3 with different settings. Benefited from proposed forgotten-knowledge based distribution for answer subset selection, our method not only yields better performance than *KP*, but also remarkably reduce the complexity via discarding the non-forgotten answer candidates. Furthermore, when considering 20 most easily-forgotten answers, FedP^3 reaches its highest performance on both generic and personalized learning, with less than one-third the computational complexity of the standard *KP*.

4.6 Case Study

Fig. 5 reveals two VQA training samples in the first federated scenario (Tab. 2), accompanied with different forgotten-knowledge based distributions for answer subset selection. In the first example labeled by high-frequency answer '*airport*' in the transports dataset, the classes with high probabilities are some easily-forgotten general answers (e.g., field and road), or some answers mainly exiting in other scenes (e.g., park and ocean). For the second sample answered by rare label '*computer mouse*' in the educational scene, the selected answers turn to be the visual concepts involved in the image (e.g., keyboard, screen and laptop), which encourages the global model to transfer more informative knowledge for personalized learning.

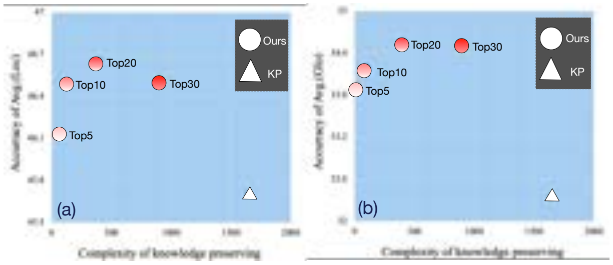

Figure 4: The relationship between computational complexity of distillation, and the local(a)/global(b) accuracy.

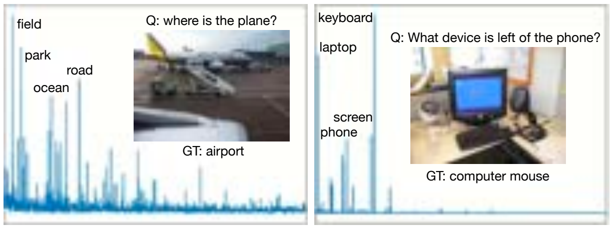

Figure 5: Two VQA examples in transports and educational scenes, respectively. Their forgotten-knowledge based distributions $r(i)$ is marked by answers with Top-4 probabilities.

5 CONCLUSION

In this study, we introduce a relatively unexplored personalized federated visual question answering (FedVQA) task. To tackle this task, we propose a novel federated pairwise preference preserving framework that enables joint optimization of generic and personalized models, leveraging distributed local data in a collaborative manner. Additionally, we construct a multi-scene FedVQA benchmark to facilitate the investigation of FedVQA. The experimental results demonstrate that our proposed method achieves competitive personalized and generalized abilities compared to state-of-the-art approaches. In the future, we attempt to further improve FedVQA task by involving more challenging and practical scenarios.

6 ACKNOWLEDGMENTS

This work was supported mainly by the LIACS Media Lab at Leiden University, and partially by the China Scholarship Council.

REFERENCES

[1] Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N Whatmough, and Venkatesh Saligrama. 2021. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263* (2021).

[2] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4971–4980.

[3] Sabtain Ahmad and Atakan Aral. 2022. FedCD: Personalized federated learning via collaborative distillation. In *2022 IEEE/ACM 15th International Conference on Utility and Cloud Computing (UCC)*. IEEE, 189–194.

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6077–6086.

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.

[6] Cristian-Paul Bara, Qing Ping, Abhinav Mathur, Govind Thattai, Rohith MV, and Gaurav S Sukhatme. 2022. Privacy preserving visual question answering. *arXiv preprint arXiv:2202.07712* (2022).

[7] Abhipsa Basu, Sravanti Addepalli, and R Venkatesh Babu. 2023. RMLVQA: A Margin Loss Approach for Visual Question Answering With Language Biases. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11671–11680.

[8] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečnỳ, Stefano Mazzocchi, Brendan McMahan, et al. 2019. Towards federated learning at scale: System design. *Proceedings of machine learning and systems* 1 (2019), 374–388.

[9] Léon Bottou. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*. Springer, 177–186.

[10] Christopher Briggs, Zhong Fan, and Peter Andras. 2020. Federated learning with hierarchical clustering of local updates to improve training on non-IID data. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–9.

[11] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Learning universal image-text representations. (2019).

[12] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. 2020. Personalized federated learning: A meta-learning approach. *arXiv preprint arXiv:2002.07948* (2020).

[13] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847* (2016).

[14] Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. 2022. Feddc: Federated learning with non-iid data via local drift decoupling and correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10112–10121.

[15] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6904–6913.

[16] Yangyang Guo, Zhiyong Cheng, Liqiang Nie, Yibing Liu, Yinglong Wang, and Mohan Kankanhalli. 2019. Quantifying and alleviating the language prior problem in visual question answering. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*. 75–84.

[17] Yangyang Guo, Liqiang Nie, Zhiyong Cheng, Qi Tian, and Min Zhang. 2021. Loss re-scaling VQA: Revisiting the language prior problem from a class-imbalance view. *IEEE Transactions on Image Processing* 31 (2021), 227–238.

[18] Farzin Haddadpour and Mehrdad Mahdavi. 2019. On the convergence of local descent methods in federated learning. *arXiv preprint arXiv:1910.14425* (2019).

[19] Sungwon Han, Sungwon Park, Fangzhao Wu, Sundong Kim, Chuhan Wu, Xing Xie, and Meeyoung Cha. 2022. FedX: Unsupervised federated learning with cross knowledge distillation. In *European Conference on Computer Vision*. Springer, 691–707.

[20] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[21] Ronghang Hu, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2018. Explainable neural computation via stack neural module networks. In *Proceedings of the European conference on computer vision (ECCV)*. 53–69.

[22] Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6700–6709.

[23] Wonyong Jeong and Sung Ju Hwang. 2022. Factorized-fl: Agnostic personalized federated learning with kernel factorization & similarity matching. *arXiv preprint arXiv:2202.00270* (2022).

[24] Jingjing Jiang, Ziyi Liu, Yifan Liu, Zhixiong Nan, and Nanning Zheng. 2021. X-ggm: Graph generative modeling for out-of-distribution generalization in visual question answering. In *Proceedings of the 29th ACM international conference on multimedia*. 199–208.

[25] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2901–2910.

[26] 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.

[27] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. 2020. Tighter theory for local SGD on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 4519–4529.

[28] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*. PMLR, 5583–5594.

[29] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. 2016. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492* (2016).

[30] Viraj Kulkarni, Milind Kulkarni, and Aniruddha Pant. 2020. Survey of personalization techniques for federated learning. In *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*. IEEE, 794–797.

[31] Mingrui Lao, Yanming Guo, Wei Chen, Nan Pu, and Michael S Lew. 2022. VQA-BC: Robust Visual Question Answering Via Bidirectional Chaining. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4833–4837.

[32] Mingrui Lao, Yanming Guo, Yu Liu, Wei Chen, Nan Pu, and Michael S Lew. 2021. From superficial to deep: Language bias driven curriculum learning for visual question answering. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3370–3379.

[33] Mingrui Lao, Yanming Guo, Nan Pu, Wei Chen, Yu Liu, and Michael S Lew. 2021. Multi-stage hybrid embedding fusion network for visual question answering. *Neurocomputing* 423 (2021), 541–550.

[34] Mingrui Lao, Nan Pu, Yu Liu, Kai He, Erwin M Bakker, and Michael S Lew. 2023. COCA: COllaborative CAusal Regularization for Audio-Visual Question Answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 12995–13003.

[35] Daliang Li and Junpu Wang. 2019. Fedmd: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581* (2019).

[36] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 11336–11344.

[37] Qinbin Li, Bingsheng He, and Dawn Song. 2021. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10713–10722.

[38] Qinbin Li, Zeyi Wen, Zhaomin Wu, Sixu Hu, Naibo Wang, Yuan Li, Xu Liu, and Bingsheng He. 2021. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. *IEEE Transactions on Knowledge and Data Engineering* (2021).

[39] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems* 2 (2020), 429–450.

[40] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. 2019. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189* (2019).

[41] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. 2021. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623* (2021).

[42] Paul Pu Liang, Terrance Liu, Liu Ziyin, Nicholas B Allen, Randy P Auerbach, David Brent, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523* (2020).

[43] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. 2020. Ensemble distillation for robust model fusion in federated learning. *Advances in Neural Information Processing Systems* 33 (2020), 2351–2363.

[44] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2020. Federated learning for vision-and-language grounding problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11572–11579.

[45] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[46] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR, 1273–1282.

[47] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 12700–12710.

[48] Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. *Advances in neural information processing systems* 28 (2015).

[49] Ramprasaath R Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Tulio Ribeiro, Besmira Nushi, and Ece Kamar. 2020. Squinting at vqa models: Introspecting vqa models with sub-questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10003–10011.

[50] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. 2019. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6649–6658.

[51] Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490* (2019).

[52] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2019. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699* (2019).

[53] Frederick Tung and Greg Mori. 2019. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1365–1374.

[54] Chuhan Wu, Fangzhao Wu, Lingjuan Lyu, Yongfeng Huang, and Xing Xie. 2022. Communication-efficient federated learning via knowledge distillation. *Nature communications* 13, 1 (2022), 2032.

[55] Tao Yu, Eugene Bagdasaryan, and Vitaly Shmatikov. 2020. Salvaging federated learning by local adaptation. *arXiv preprint arXiv:2002.04758* (2020).

[56] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6281–6290.

[57] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 1821–1830.

[58] Jie Zhang, Song Guo, Xiaosong Ma, Haozhao Wang, Wenchao Xu, and Feijie Wu. 2021. Parameterized knowledge transfer for personalized federated learning. *Advances in Neural Information Processing Systems* 34 (2021), 10092–10104.

[59] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. 2022. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*. 11953–11962.

[60] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence* 40, 6 (2017), 1452–1464.

[61] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 13041–13049.

[62] Yiyi Zhou, Rongrong Ji, Xiaoshuai Sun, Jinsong Su, Deyu Meng, Yue Gao, and Chunhua Shen. 2019. Plenty is plague: Fine-grained learning for visual question answering. *IEEE transactions on pattern analysis and machine intelligence* 44, 2 (2019), 697–709.

[63] Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. 2021. Data-free knowledge distillation for heterogeneous federated learning. In *International conference on machine learning*. PMLR, 12878–12889.

APPENDIX

A FEDVQA SETTING

Datasets organization: In our proposed FedVQA, we establish six clients’ datasets for personalized training via selectively sampling from the large-scale GQA datasets, whose images are represented in diverse real-world scenes. Specifically, to introduce the scene-based distribution skews among different local datasets, we leverage the taxonomy of Place 365 database, as well as the pretrained ResNet152-places365 model to select the samples in commercial, educational, transport, natural sports and home scenes. The detailed sub-categories in each local datasets are in Tab. 5. It is noteworthy that, we only sample the instances whose Top-1 predictive probability are higher than 0.6 into the corresponding local datasets. Besides, to avoid the similarity of local label distributions caused by excessive general binary answers *‘yes’* and *‘no’*, we only consider 20% of the selected samples labeled by *‘yes’* and *‘no’* into the final local datasets.

Table 5: The representative sub-categories of scenes in the six clients datasets specialized for answering the questions to commercial, educational, transport, natural, sports, home scenes.

Dataset	Representative sub-categories of scenes
Commercial	restaurant, market, pharmacy, bakery, ticket booth, discotheque, beauty salon, restaurant kitchen, repair shop, bank vault, bookstore
Educational	campus, art gallery, music studio, church, museum, temple, lecture room, science museum, biology laboratory, computer room, library
Transport	airport, subway , crosswalk, galley, bus, train station, airfield, boat deck, bridge, highway, gas station, boathouse, bus station, garage
Natural	forest, mountain, marsh, underwater, fishpond, waterfall, ocean, lake, iceberg, desert, rainforest, swamp, marsh, snowfield, river, vineyard
Sports	ballroom, arena, gymnasium, ski slope, basketball court, bowlling alley, locker room, athletic field, football field, swimming pool, sandbox
Home	kitchen, bedroom, bathroom, closet, utility room, shower, living room, child’s room, dining room, alcove, bedchamber, wet bar

Label distributions in local datasets: The label distributions over training samples in six local datasets are depicted in Fig. 6, and the scene-specific answer candidates in each dataset are in the Fig. 2. We can see that the mainstream correct answers across different scenes are inconsistent in our MS-FedVQA benchmark, which poses more challenges for federated solutions to mitigate label distribution skews in FedVQA task.

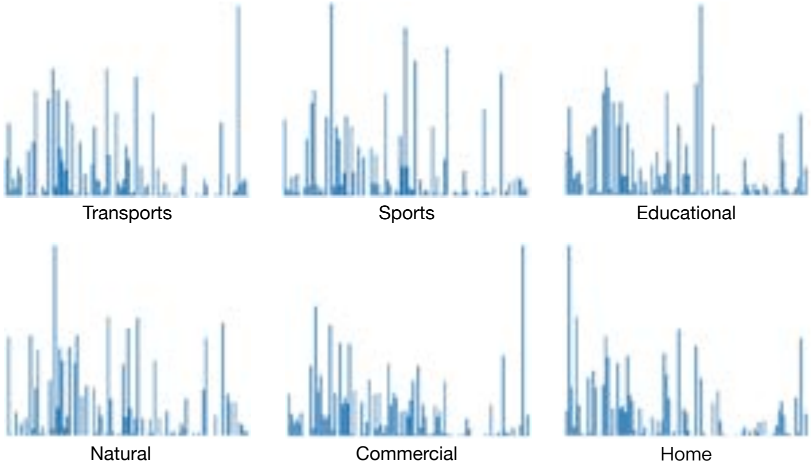

Figure 6: Label distributions of six scene-specific datasets over the first 100 answer candidates (overally high-frequency labels).

B MORE EXPERIMENTAL RESULTS

Ablation study on the second federated scenario: to further demonstrate the effectiveness of our FedP^3 approach, we also conduct ablation study on the second federated scenarios depicted in Tab. 3. In Tab. 6, based on FedAvg, the improvement caused by knowledge preserving is still limited, which can be explained by the hard constraint of logits-based distillation. On the contrary, through the combination of DPP and FKF, our method significantly boosts both the performance of local and global models. These results are consistent to those in the first federated scenarios validated in the Tab. 4.

The settings of factor λ: we conduct experiments for knowledge preserving KP ($T = 2$) and our FedP^3 with different setting of trade-off factor λ in Equ. (15). For KP, using lower value of $\lambda = 0.1, 0.5$ fails to alleviate the strict regularization when reviewing global knowledge, without achieving any accuracy boost. In contrast, our method consistently perform better from both side of local (Avg. (Loc)) and global (Avg. (Glo)) performance. Finally, we select the $\lambda = 1$ in our work, which is the optimal setting for both KP and our method.

Table 6: Ablation studies of three concepts in our proposed FedP^3 according to different settings.

Component	Setting	Avg.(Loc)	Avg.(Glo)
FedAvg	Baseline	42.92	34.99
+Knowledge Preserving	T=1	42.81	34.33
	T=2	43.07	34.81
	T=3	41.98	32.72
+Pairwise Preference	all answers	43.41	35.71
+Forgotten-Knowledge Filter	$r(i) = \mu(i)$	43.35	35.46
	$r(i) = p^S(i)$	43.28	35.23
	$r(i) = p^T(i)$	43.66	35.98
	Equ. (10)	43.73	36.21
	Equ. (11)	**43.76**	**36.55**

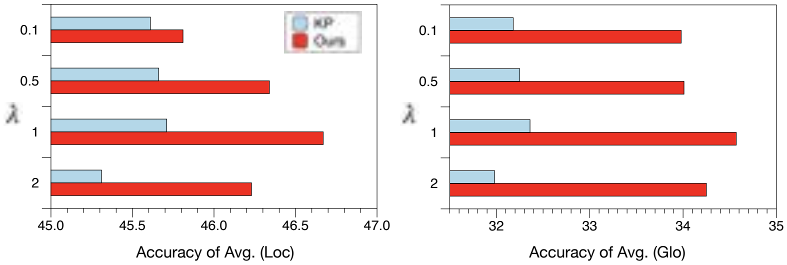

Figure 7: The performance of standard knowledge preserving (KP) and FedP^3 (Ours) under different setting of trade-off factor λ.

More Qualitative Results: In Fig. 8, we introduce more training examples with their forgotten-knowledge driven distributions according to our proposed FKF strategy, which are derived from the local datasets in transports, sports, natural and educational scenes. Overall, compared with ground-truth (GT), the selected easily-forgotten labels are more likely to be some more general or semantic-related answer candidates. The generated distribution encourages local models to adaptively reviewing more useful global knowledge during the personalized learning.

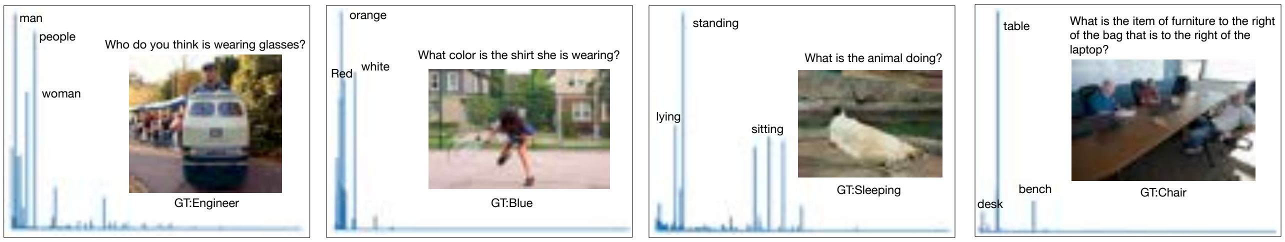

Figure 8: Four VQA training examples of case study from transports, sports, natural and educational scenes, which are trained under the federated scenario in Tab. 2. Their corresponding forgotten-knowledge based distributions $r(i)$ is marked by answer candidates with Top-3 probabilities.