# Inclusive Data Representation in Federated Learning: A Novel Approach Integrating Textual and Visual Prompt


Zihao Zhao
Tsinghua-Berkeley Shenzhen Institute,
Tsinghua University,
Shenzhen, Guangdong, China
zhao-zh21@mails.tsinghua.edu.cn

Zhenpeng Shi
Tsinghua-Berkeley Shenzhen Institute,
Tsinghua University,
Shenzhen, Guangdong, China
shizp22@mails.tsinghua.edu.cn

Yang Liu
Institute for AI Industry Research,
Tsinghua University,
Beijing, China
Shanghai Artificial Intelligence Laboratory,
Shanghai, China
liuy03@air.tsinghua.edu.cn

Wenbo Ding*
Tsinghua-Berkeley Shenzhen Institute,
Tsinghua University,
Shenzhen, Guangdong, China
Shanghai Artificial Intelligence Laboratory,
Shanghai, China
ding.wenbo@sz.tsinghua.edu.cn



## ABSTRACT

Federated Learning (FL) is often impeded by communication overhead issues. Prompt tuning, as a potential solution, has been introduced to only adjust a few trainable parameters rather than the whole model. However, current single-modality prompt tuning approaches fail to comprehensively portray local clients' data. To overcome this limitation, we present Twin Prompt Federated learning (TPFL), a pioneering solution that integrates both visual and textual modalities, ensuring a more holistic representation of local clients' data characteristics. Furthermore, in order to tackle the data heterogeneity issues, we introduce the Augmented TPFL (ATPFL) employing the contrastive learning to TPFL, which not only enhances the global knowledge acquisition of client models but also fosters the development of robust, compact models. The effectiveness of TPFL and ATPFL is substantiated by our extensive evaluations, consistently showing superior performance compared to all baselines.



## CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in ubiquitous and mobile computing**; *Collaborative and social computing*; • **Computing methodologies** → **Distributed algorithms**.

## KEYWORDS

federated learning, prompt tuning, contrastive learning

**ACM Reference Format:**
Zihao Zhao, Zhenpeng Shi, Yang Liu, and Wenbo Ding. 2023. Inclusive Data Representation in Federated Learning: A Novel Approach Integrating Textual and Visual Prompt. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing (UbiComp/ISWC '23 Adjunct), October 08–12, 2023, Cancun, Quintana Roo, Mexico.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3594739.3612914



*Corresponding author





*UbiComp/ISWC '23 Adjunct, October 08–12, 2023, Cancun, Quintana Roo, Mexico*



ACM ISBN 979-8-4007-0200-6/23/10.
https://doi.org/10.1145/3594739.3612914


## 1 INTRODUCTION

The emergence of distributed learning systems has provided considerable advantages across a wide range of domains. Nonetheless, growing privacy concerns about distributed learning have necessitated the advent of *Federated Learning* (FL) [2, 20], a framework expressly developed to protect participants' private information. In FL, instead of uploading their private data, local clients share their local model weights with a central server during each communication round. The server aggregates these models and circulates them back to the local clients, thereby accomplishing the goal of information consolidation.

Recently, FL has confronted a wealth of challenges, including significant communication overheads [19, 26, 27] and data heterogeneity [13]. A variety of recent research initiatives have sought to tackle these obstacles. Specifically, some have proposed innovative efficient encoding and model compression algorithms to reduce the communication cost, such as quantization to a continuous range of values into a finite set and sparsification [24] to clip the full gradient into a sparse one, as well as intelligent scheduling of client participation [21] during the training process. Moreover, some incorporate the original FL framework with an additional step of knowledge distilling [17] to contract larger models into smaller ones, thereby enhancing the robustness of the global model.

Despite these strategies, certain inherent limitations persist. Primarily, they require a substantial volume of labeled training samples, which may be unavailable to many clients in the FL environment, hindering effective training and resulting in model overfitting [12]. In addition, notwithstanding the communication costs reduction achieved by these efficient methods, most IoT devices such as smart home devices or industrial sensors, cannot accommodate large backbone model training due to their limited processing powers [11], infinitesimal memory, and energy constraints. To illustrate,

training a ResNet-50 model [9] involves intensive computation and storage memory. It has approximately 25 million weight parameters and computes 16 million activations in the forward pass. Even after the communication-efficient algorithm to weights and activations, the total storage needed for saving ResNet-50's intermediate gradient results is over 7.5 GB for a mini-batch of 32 on a high-performance GPU. Given the hardware constraints of typical IoT devices, it is clear that they would struggle to support such intensive computations and memory requirements.

To resolve these problems, current research is leaning towards prompt tuning [14]. Unlike conventional fine-tuning methods in FL that tune and aggregate full model parameters, applying prompt learning in FL only adjusts soft prompts for corresponding downstream tasks, while keeping large backbone models static to diminish both the communication and computation costs. Back to the ResNet-50 case, prompt tuning could save gradient results to just a handful of MB, drastically decreasing the communication overhead. However, most existing work only considers a single modality, failing to represent the local clients comprehensively. For instance, Guo et al. [7] exclusively employs textual soft prompts to depict the local clients without taking the visual knowledge into consideration; yet, Feng et al. [5] leverages continuous visual prompts to capture the image data information, disregarding text knowledge. In contrast, our work proposes Twin Prompt Federated learning (TPFL), a method resorting to both visual and textual modalities for a more comprehensive representation of the local clients' data characteristics. First off, we find that merely combining two modalities overlooks the potential for a unified approach. As such, we devise Augmented TPFL (ATPFL) to fuse the contrastive learning approach into the prompt tuning, facilitating the acquisition of global knowledge by client models. To the best of our knowledge, ATPFL is the first to integrate both textual and visual modalities within the context of FL and use contrastive learning to connect them. The contributions of this paper are threefold:

- We present an innovative FL framework named ATPFL, that merges both visual and textual modalities for an improved representation of local clients' data characteristics, surpassing existing work's performance that only considers a single modality.
- The incorporation of contrastive learning to prompt tuning, enabling clients to acquire more global knowledge and improving on the direct combination of modalities that may overlook the potential for a unified approach. This is the first work to integrate two modalities within the context of FL and to utilize contrastive learning for their integration.
- Extensive evaluations have been conducted to ascertain the effectiveness of TPFL and ATPFL. The results demonstrate that ATPFL outperforms all the baselines.

## 2 RELATED WORKS

### 2.1 Communication Efficiency

Communication efficiency has always been a critical challenge in the FL field. Different lines of research have been investigated to tackle this challenge. Firstly, quantization[6] methods are used to represent the full model parameters with lower bits. This technique involves converting the high-precision floating-point values of the model parameters into lower-precision values. For example, stochastic quantization[1] adaptively adjusts the quantization level in a stochastic manner. Secondly, sparsification methods improve communication efficiency by directly reducing the number of model parameters to be sent. More specifically, the sparsification method selects an important subset of model parameters and sets other insignificant parameters to zero before sending them to the global server. Top-k sparsification and rank-k sparsification are common sparsification methods[3]. Han et al.[8]proposed to adaptively change the sparsification level to minimize overall training time. Shi et al.[25] introduced global-k sparsification to compress the down streaming communication from the server to the clients. Thirdly, knowledge distillation is also investigated to alleviate communication overhead[15]. Knowledge distillation methods transfer knowledge from a larger teacher model to a smaller student model. Examples of knowledge-distillation-based federated learning are FedMD[15], FedDF[23], etc. However, all the aforementioned strategies have a high resource requirement and can hardly be implemented in IoT devices due to their limited hardware restrictions.

### 2.2 Prompt Tuning

Houlsby et al. [10] proposed parameter-efficient transfer learning with adapter modules. Liu et al. [18] showed that prompt-tuning can match the performance of fine-tuning with only 0.1% - 3% tuned parameters in the context of Natural Language Understanding. Li and Liang [16] applied prefix-tuning to GPT-2 and BART for downstream tasks and shows that prefix-tuning can outperform fine-tuning in low-data settings. Guo et al. [7] proposed a federated learning framework for prompt-tuning called PromptFL.The PromptFL framework leverages the power of federated learning, which allows training prompts on decentralized data across multiple devices. In this work, only one modality text prompt is used and the result shows that federated prompt tuning achieved better performance compared to fine-tuning FL in many IID and non-IID settings. Nonetheless, the existing research primarily focuses on a single modality, constraining their capability to obtain more information of local clients. In this paper, we present to employ both textual and visual representations to comprehensively characterize the local client.

## 3 METHODOLOGY

This section begins by outlining the basic structure of FL. Subsequently, we introduce the TPFL which considers both visual and textual information. Despite showing improvements, TPFL has certain inherent limitations. Therefore, we propose ATPFL to address these shortcomings and achieve superior performance.

### 3.1 Problem Statement

In the general FL setting, the entail system envelops $M$ clients, while, in every round, $K$ clients will actively participate, each possessing a unique local dataset. Each local dataset on client $k$ consists of $n_k$ samples, with each sample representing a pair, $(x_i^k, y_i^k)$, of a data feature $x$ and its corresponding target label $y$. The primary objective of FL is to construct a global model parameter vector $w$ that minimizes the mean loss across all local datasets, as demonstrated

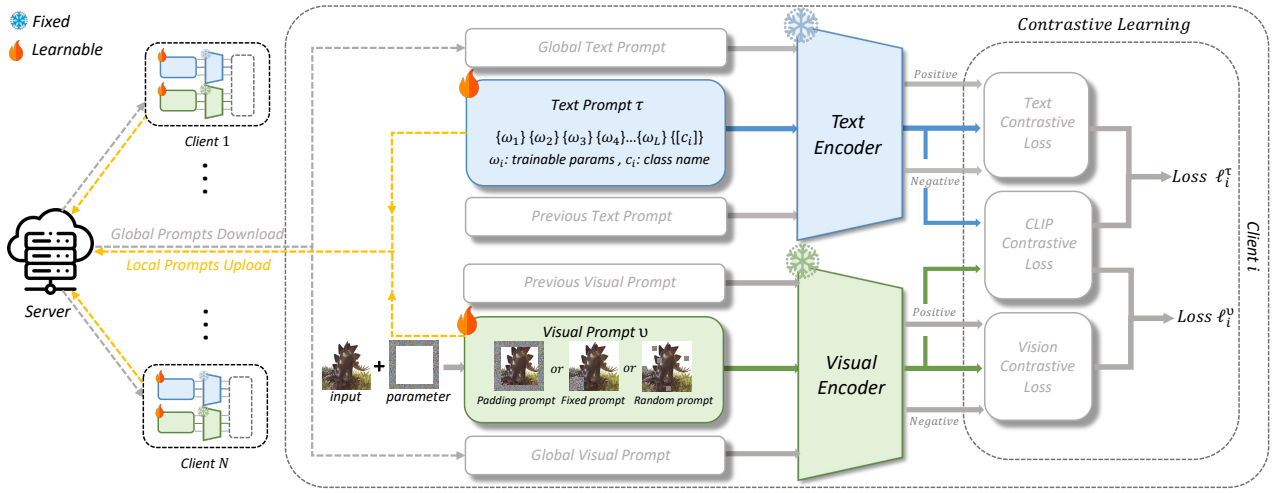


**Figure 1: This figure illustrates the pipeline of ATPFL with contrastive learning. In local training, the current prompt, previous prompt, and received global prompt are passed to each modality encoder. After the encoding, two types of contrastive learning are performed. Text contrastive loss and Visual contrastive loss use the feature extracted from the global prompt as positive contrast and the feature extracted from the previous prompt as negative contrast. CLIP contrastive loss is computed with the test prompt feature and the visual prompt feature.**

in the following optimization problem:

$$w = \arg\min_{w} \frac{1}{K} \sum_{i=1}^{K} \frac{1}{n_i} \sum_{n=1}^{n_i} \mathcal{L}\left(w; x_n^i, y_n^i\right), \tag{1}$$

where $w$ denotes the weights of the prediction model, $\mathcal{L}$ is the loss function.

## 3.2 Twin Prompt Federated Learning (TPFL)

As aforementioned CoOp [28] resorts to a series of continuous learnable parameters as the textual prompts, replacing the manually-designed constant ones. The textual prompt can be denoted as $\boldsymbol{\tau}_i = \{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \ldots, \boldsymbol{c}_i, \ldots, \boldsymbol{\omega}_L\}$, where $\boldsymbol{c}_i$ signifies the word embedding of the $i^{th}$ image class names, $\boldsymbol{\omega}$ is a collection of learnable vectors, denoted as $\{\boldsymbol{\omega}_i|_{i=1}^{L}\}$, and $L$ symbolizes the length of context words. Importantly, the position of $c_i$ can be placed anywhere between $(1, L+1)$. In the training process, the textual prompt will be fed into a text encoder $g(\cdot)$, obtaining the textual feature as $\boldsymbol{g}_i = g(t_i)$. Similarly, the visual feature $\boldsymbol{f} = f(x)$ is calculated by visual encoder $f$. The final prediction probability is computed by the matching score:

$$p(y = c_i \mid \boldsymbol{x}) = \frac{\exp\left(\text{sim}\left(\boldsymbol{f}, \boldsymbol{g}_i\right)/\Gamma\right)}{\sum_j \exp\left(\text{sim}\left(\boldsymbol{f}, \boldsymbol{g}_j\right)/\Gamma\right)}, \tag{2}$$

where $\Gamma \in \mathbb{R}$ is the temperature factor to control the overall distribution of the similarity between the embedding of the visual feature and test feature.

Different from the previous work, which solely obtains a single modal to represent a local client, our study introduces TPFL to resort to two different modalities, vision and text, to enhance the generalization capability and resilience of the global model. More specifically, instead of relying on a constant input visual feature $x$, we incorporate an additional trainable visual prompt $v$ as an extended representation for the local data characterization and conduct $x + v$ to get the final input feature. As illustrated in Figure 1, three templates of the visual prompt are employed: the padding, random patch, and fixed patch patterns, each contributing to varying model performances. After acquiring both the textual and visual prompts, each local client transmits them to the central server. The server then aggregated the received prompts, in light of the number of their training samples:

$$\tau_g \leftarrow \sum_{i=1}^{K} \frac{n_i}{\sum_{j=1}^{K} n_j} \tau_i, \quad v_g \leftarrow \sum_{i=1}^{K} \frac{n_i}{\sum_{j=1}^{K} n_j} v_i. \tag{3}$$

However, the naive aggregation of the uploaded model weights may invite certain problems. To begin with, in practical scenarios, the data distribution across multiple clients may not be independently and identically distributed (IID). In other words, different clients can host data with significantly divergent statistical characteristics. The direct averaging of models struggles to effectively amalgamate local models originating from these devices, owing to this non-IID data distribution, and as a result, the performance of the global model suffers. Moreover, data volume can significantly vary across devices, with certain scenarios providing only a sparse dataset (only a few data points are available). Conventional FL aggregation might lack the robustness required to manage these few-shot learning scenarios, thereby complicating the process of discerning meaningful patterns from such limited data.

## 3.3 Augmented TPFL (ATPFL)

To address these aforementioned challenges, we propose the incorporation of a contrastive learning strategy, thus fortifying the robustness of FL. Specifically, we utilize the InfoNCE loss function [22] to encourage the output distributions of both the local visual and textual prompts to align closely with the output distribution of the global model. This methodology fosters a better comprehension of the global model by the local client, consequently mitigating the

**Algorithm 1** ATPFL

**Input:** The entire $K$ clients are indexed by $i \in \{1, 2, \ldots, K\}$; $T_g$ and $T_{loc}$ is the number of global epochs and local epochs, respectively, and $\alpha$ is the learning rate.
**Server executes:**
Initialize $\tau_g^0, v_g^0$
**for** each round $t = 1, 2, \ldots, T_g$ **do**
**for** each client $i$ **in parallel do**
$\tau_i^{t+1}, v_i^{t+1} \leftarrow ClientUpdate(i, \tau_g^t, v_g^t)$
**end for**
Aggregate the global prompts $\tau_g^{t+1}, v_g^{t+1}$ by (3)
**end for**
**ClientUpdate**($i, \tau_g^t, v_g^t$)**:**
**for** each local epoch from 1 to $T_{loc}$ **do**
Calculate the logits and loss for textual prompt and visual prompt by (6) and (7), respectively
Update the local textual and visual prompts by:
$\tau_i^{t+1} \leftarrow \tau_g^t - \alpha \nabla \ell_i^\tau$, $v_i^{t+1} \leftarrow v_g^t - \alpha \nabla \ell_i^v$
**end for**
Return $\tau_i^{t+1}, v_i^{t+1}$ to the server

adverse effects of non-IID data. The key insight fueling this strategy is that contrastive learning facilitates the distinction between similar and dissimilar data points. It mitigates the discrepancies among local models caused by non-IID data through the learning of invariant features, making local models more amenable to aggregation at the global level. The contrastive (InfoNCE) loss functions for both textual and visual prompts are formulated in (4) and (5), respectively:

$$\ell_{con\_\tau} = -\log \frac{\exp\left(\text{sim}\left(z_\tau, z_{g\_\tau}\right)/\Gamma\right)}{\exp\left(\text{sim}\left(z_\tau, z_{g\_\tau}\right)/\Gamma\right) + \exp\left(\text{sim}\left(z_\tau, z_{p\_\tau}\right)/\Gamma\right)}, \tag{4}$$

$$\ell_{con\_v} = -\log \frac{\exp\left(\text{sim}\left(z_v, z_{g\_v}\right)/\Gamma\right)}{\exp\left(\text{sim}\left(z_v, z_{g\_v}\right)/\Gamma\right) + \exp\left(\text{sim}\left(z_v, z_{p\_v}\right)/\Gamma\right)}, \tag{5}$$

where $sim(\cdot, \cdot)$ function represents the similarity function, $\Gamma$ denotes the temperature function (with a little symbol abuse to (2)), $z_\tau, z_v$ refer to the embedding of local textual and visual prompts, and $z_{g\_\tau}, z_{g\_v}$ represent the global textual and visual prompts. After attaining the contrastive loss, the overall loss of the trainable prompts can be calculated by:

$$\ell_i^\tau = \ell_{con}(w_i^t; x_i, y_i) + \mu \ell_{con\_\tau}(w_i^t; w_i^{t-1}; w_g^t; x_i), \tag{6}$$

$$\ell_i^v = \ell_{con}(w_i^t; x_i, y_i) + \mu \ell_{con\_v}(w_i^t; w_i^{t-1}; w_g^t; x_i), \tag{7}$$

where $\ell_{con}$ denotes the contrastive loss formulated in (2), $w_i^{t-1}$ represents the previous model, $w_g^t$ denotes the global model, and $\mu$ represents a tuning factor to control the influence of $\ell_{con\_\tau}$ and $\ell_{con\_v}$. The overall training process of ATPFL is shown in Algorithm 1.

# 4 EVALUATION

In this section, we perform intensive evaluations to verify the effectiveness of our proposed TPFL and ATPFL.

## 4.1 Evaluation setup.

**Few-shot Dataset and Data partition.** Extended from PromptFL[7] who only evaluates their model on four datasets, we verify ATPFL in seven different datasets: Caltech-101 [4], Oxford-Pets, Stanford Cars, OxfordFlowers-102, EuroSAT, UCF-101, and Describable Textures (DTD). Furthermore, in order to create the few-shot dataset, we set that each client has $n_k$ samples for each class. For the majority of our evaluations, we choose $n_k = 4$ meaning that each client has a four-shot dataset; besides, we investigate the effect of the shot size in the ablation section. For the non-IID setting in FL, we select the label-skewing method to emulate the heterogeneous local clients.

**Models** Following the existing work, we choose the ResNet-50 (RN50) and Visual Transformer model (ViT) as the backbone of the visual encoder, and the Transformer model as the textual encoder.

**Baselines.** In our evaluation, we compare ATPFL with the following baselines: (1) Local training, where all clients train their own models in an offline manner, and no model transmission is conducted; (2) PromptFL, using only the textual modality; (3) TPFL, employing both the textual and visual modalities, but no InfoNCE loss.

**Implementation details.** To prevent the influence of randomness and ensure the fairness of our evaluations, each experiment setting has been performed in three identical random seeds, and then we average the results to get the final result. We use the Adam optimizer with learning rate $\alpha = 1e-3$, and the Cosine scheduler with $max_{epoch} = 20$. Furthermore, For the implementation environment, we conduct our code on Python version 3.11.0 and Pytorch 1.13.0. We also use 4 NTX NVIDIA A6000 GPUs to run our code.

## 4.2 Main results

In this section, the experimental outcomes are assessed. Table 1 and Table 2 present the average test accuracy for ViT and RN50 backbones across seven diverse datasets in a non-IID setting. Both PromptFL and ATPFL consistently surpass local training, with margins extending up to 18.5%. This is intuitive, as local training or full-model fine-tuning may lead to catastrophic forgetting. This issue is exacerbated by client data heterogeneity. These compounded factors significantly impede fine-tuning performance in the federated learning context, necessitating the exploration of PromptFL and ATPFL. For ViT, TPFL excels over PromptFL in six of the seven datasets, with margins spanning 0.1% - 6.2%, except for UCF-101 where TPFL lags by 0.2%. When factoring in the standard error of test accuracy across multiple experiments, TPFL's advancements over previous methods are noticeable. Despite TPFL's success, limitations persist, leading to the proposal of ATPFL to better address these issues. Our ATPFL model outperforms the baseline by 0.4% - 1.1% across all datasets, illustrating ATPFL's potential to mitigate data heterogeneity in prompt federated learning scenarios.

In the ResNet-50 tests, TPFL outperforms local training and PromptFL in six of the seven datasets, except for the EuroAT dataset. Our ATPFL continues to surpass TPFL in four of the seven datasets, except for Oxford-Pets and DTD where ATPFL trails TPFL by 0.2% and 0.5% respectively. This could be due to the model disparities between ViT and ResNet-50.

**Table 1: Test Accuracy (%) Results for ViT model on 7 datasets with 5 different seeds.**

| Algorithm (ViT) | Caltect-101 | Flowers-102 | Oxford-Pets | DTD | EuroSAT | Stanford Car | UCF-101 |
|---|---|---|---|---|---|---|---|
| Local Training | 86.9±0.03 | 58.7±0.04 | 83.6±0.02 | 37.8±0.01 | 25.8±0.32 | 59.5±0.12 | 61.3±0.06 |
| PromptFL[7] | 89.7±0.01 | 67.6±0.01 | 88.5±0.07 | 42.9±0.08 | 48.1±0.22 | 63.0±0.01 | 66.1±0.02 |
| TPFL | 90.6±0.01 | 68.9±0.01 | 89.1±0.00 | 43.0±0.07 | 54.3±0.29 | 63.4±0.01 | 65.9±0.02 |
| **ATPFL** (ours) | **91.3**±0.01 | **69.6**±0.01 | **89.5**±0.01 | **44.1**±0.01 | **54.9**±0.26 | **63.8**±0.00 | **66.5**±0.01 |

**Table 2: Test Accuracy (%) Results for ResNet-50 model on 7 datasets with 5 different seeds.**

| Algorithm (RN50) | Caltect-101 | Flowers-102 | Oxford-Pets | DTD | EuroSAT | Stanford Car | UCF-101 |
|---|---|---|---|---|---|---|---|
| Local Training | 63.1±0.37 | 18.7±2.61 | 30.8±4.83 | 22.5±0.21 | 19.2±0.01 | 20.1±0.93 | 34.3±0.36 |
| PromptFL[7] | 84.8±0.04 | 58.7±0.01 | 85.3±0.04 | 35.7±0.03 | **33.4**±0.03 | 52.9±0.02 | 57.8±0.07 |
| TPFL | 85.2±0.02 | 59.6±0.01 | **85.6**±0.02 | **37.4**±0.03 | 32.2±0.04 | 53.8±0.01 | 58.2±0.03 |
| **ATPFL** (ours) | **85.6**±0.02 | **60.5**±0.00 | 85.4±0.04 | 36.9±0.03 | 32.2±0.01 | **54.1**±0.01 | **59.3**±0.04 |

In conclusion, our proposed ATPFL, leveraging the concept of contrastive learning, offers superior performance in handling data heterogeneity. These results corroborate our prior discussions in the methodology section.

## 4.3 Ablation study

In this section, we examine various factors influencing our model's performance, including the application of InfoNCE loss, number of shot size, and client quantity.

**InfoNCE loss.** First off, we investigates the impact of InfoNCE loss (i.e., the difference between TPFL and ATPFL). As illustrated in Table 1 and Table 2, ATPFL shows a clear advantage compared to TPFL. In 11 out of 14 experiments, ATPFL outperforms TPFL by a margin of up to 1.1%.

**Shot size.** Second, we explore the impact of shot size, and Figure 2 demonstrates a monotonic increase in the F1-score as the number of shots rises, with the F1-score in a 16-shot scenario exceeding that of a 1-shot scenario by 2.3%. Moreover, despite the absence of a consistent increase, accuracy still trends upward with an increasing number of shots. Even at a 1-shot scenario, ATPFL exhibits substantial performance (90.2% accuracy and 87.8% F1-score), but greater shot numbers offer additional potential performance benefits due to the increased feature information provided at each learning round.

**Client volume.** Lastly, the ablation study examines the effect of the number of clients. Figure 3 reveals a decline in both accuracy and the F1-score as the client number rises, with a tenfold increase in clients (from 10 to 100) decreasing accuracy and the F1-score by 2.1% and 3.2%, respectively. However, even with a larger number of clients, ATPFL maintains reasonable performance, achieving 86.1% accuracy in a 100-client scenario.

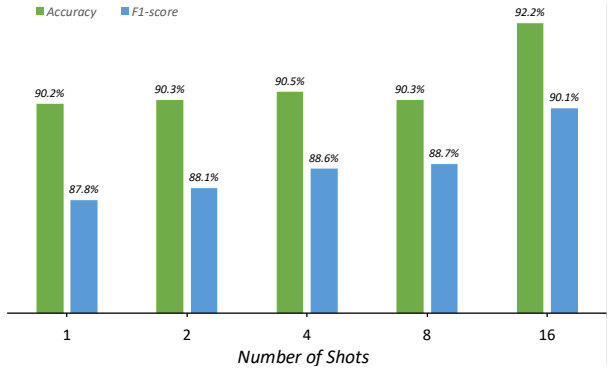


**Figure 2: This figure illustrates how shot number affects the model accuracy and F1-score**


**Figure 3: This figure illustrates how client number affects the model accuracy and F1-score**

## 5 CONCLUSION

In this paper, we propose an FL framework, TPFL, which first considers both visual and textual information in prompt tuning to augment the global model in FL. Notwithstanding, the performance improvement offered by TPFL is limited due to data heterogeneity. To address this issue, we developed ATPFL to facilitate local clients in obtaining more information from the global model, thereby enhancing their representing performance. A series of experiments have been conducted to validate the effectiveness of our methods, demonstrating that ATPFL consistently outperforms all baseline methods across various datasets and scenarios.

## ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China under Grant No.2022ZD0160504, by Tsinghua Shenzhen International Graduate School-Shenzhen Pengrui Young Faculty Program of Shenzhen Pengrui Foundation (No. SZPR2023005), and by Tsinghua-Toyota Joint Research Institute inter-disciplinary Program and Tsinghua University (AIR)-Asiainfo Technologies (China) Inc. Joint Research Center under grant No. 20203910074. We would also like to thank anonymous reviewers for their insightful comments.

## REFERENCES

[1] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. 2017. QSGD: Communication-efficient SGD via gradient quantization and encoding. *Advances in neural information processing systems* 30 (2017).

[2] Kallista Bonawitz, Peter Kairouz, Brendan Mcmahan, and Daniel Ramage. 2022. Federated learning and privacy. *Commun. ACM* 65, 4 (2022), 90–97.

[3] Negar Foroutan Eghlidi and Martin Jaggi. 2020. Sparse communication for training deep networks. *arXiv preprint arXiv:2009.09271* (2020).

[4] Li Fei-Fei, Rob Fergus, and Pietro Perona. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*. IEEE, 178–178.

[5] Chun-Mei Feng, Bangjun Li, Xinxing Xu, Yong Liu, Huazhu Fu, and Wangmeng Zuo. 2023. Learning Federated Visual Prompt in Null Space for MRI Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8064–8073.

[6] Robert M. Gray and David L. Neuhoff. 1998. Quantization. *IEEE transactions on information theory* 44, 6 (1998), 2325–2383.

[7] Tao Guo, Song Guo, Junxiao Wang, and Wenchao Xu. 2022. PromptFL: Let Federated Participants Cooperatively Learn Prompts Instead of Models–Federated Learning in Age of Foundation Model. *arXiv preprint arXiv:2208.11625* (2022).

[8] Pengchao Han, Shiqiang Wang, and Kin K Leung. 2020. Adaptive gradient sparsification for efficient federated learning: An online learning approach. In *2020 IEEE 40th international conference on distributed computing systems (ICDCS)*. IEEE, 300–310.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[10] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*. PMLR, 2790–2799.

[11] Ahmed Imteaj, Urmish Thakker, Shiqiang Wang, Jian Li, and M Hadi Amini. 2021. A survey on federated learning for resource-constrained IoT devices. *IEEE Internet of Things Journal* 9, 1 (2021), 1–24.

[12] Yilun Jin, Xiguang Wei, Yang Liu, and Qiang Yang. 2020. Towards utilizing unlabeled data in federated learning: A survey and prospective. *arXiv preprint arXiv:2002.11545* (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] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 3045–3059.

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

[16] Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 4582–4597.

[17] 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.

[18] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Dublin, Ireland, 61–68. https://doi.org/10.18653/v1/2022.acl-short.8

[19] Yuzhu Mao, Zihao Zhao, Meilin Yang, Le Liang, Yang Liu, Wenbo Ding, Tian Lan, and Xiao-Ping Zhang. 2023. SAFARI: Sparsity-Enabled Federated Learning with Limited and Unreliable Communications. *IEEE Transactions on Mobile Computing* (2023), 1–12. https://doi.org/10.1109/TMC.2023.3296624

[20] 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.

[21] Takayuki Nishio and Ryo Yonetani. 2019. Client selection for federated learning with heterogeneous resources in mobile edge. In *ICC 2019-2019 IEEE international conference on communications (ICC)*. IEEE, 1–7.

[22] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[23] Felix Sattler, Arturo Marban, Roman Rischke, and Wojciech Samek. 2020. Communication-efficient federated distillation. *arXiv preprint arXiv:2012.00632* (2020).

[24] Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. 2019. Robust and communication-efficient federated learning from non-iid data. *IEEE transactions on neural networks and learning systems* 31, 9 (2019), 3400–3413.

[25] Shaohuai Shi, Qiang Wang, Kaiyong Zhao, Zhenheng Tang, Yuxin Wang, Xiang Huang, and Xiaowen Chu. 2019. A distributed synchronous SGD algorithm with global top-k sparsification for low bandwidth networks. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2238–2247.

[26] Zihao Zhao, Yuzhu Mao, Yang Liu, Linqi Song, Ye Ouyang, Xinlei Chen, and Wenbo Ding. 2023. Towards efficient communications in federated learning: A contemporary survey. *Journal of the Franklin Institute* (2023). https://doi.org/10.1016/j.jfranklin.2022.12.053

[27] Zihao Zhao, Yuzhu Mao, Zhenpeng Shi, Yang Liu, Tian Lan, Wenbo Ding, and Xiao-Ping Zhang. 2023. AQUILA: Communication Efficient Federated Learning with Adaptive Quantization of Lazily-Aggregated Gradients. (2023). arXiv:2308.00258 [cs.LG]

[28] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16816–16825.