TOFA: Training-Free One-Shot Federated Adaptation for Vision-Language Models

Li Zhang[1], Zhongxuan Han[1], XiaoHua Feng[1], Jiaming Zhang[1], Yuyuan Li[2*], Linbo Jiang[3], Jianan Lin[3], Chaochao Chen[1]

[1]College of Computer Science and Technology, Zhejiang University
[2]School of Communication Engineering, Hangzhou Dianzi University
[3]Ant Group
zhanglizl80@gmail.com, {zxhan, fengxiaohua, 22321350}@zju.edu.cn, y2li@hdu.edu.cn
xiaobo.jlb@antgroup.com, jianan.linjn@myxiaojin.cn, zjuccc@zju.edu.cn

Abstract

Efficient and lightweight adaptation of pre-trained Vision-Language Models (VLMs) to downstream tasks through collaborative interactions between local clients and a central server is a rapidly emerging research topic in federated learning. Existing adaptation algorithms are typically trained iteratively, which incur significant communication costs and increase the susceptibility to potential attacks. Motivated by the one-shot federated training techniques that reduce client-server exchanges to a single round, developing a lightweight one-shot federated VLM adaptation method to alleviate these issues is particularly attractive. However, current one-shot approaches face certain challenges in adapting VLMs within federated settings: (1) *insufficient exploitation of the rich multimodal information inherent in VLMs*; (2) *lack of specialized adaptation strategies to systematically handle the severe data heterogeneity*; and (3) *requiring additional training resource of clients or server*. To bridge these gaps, we propose a novel **T**raining-free **O**ne-shot **F**ederated **A**daptation framework for VLMs, named TOFA. To fully leverage the generalizable multimodal features in pre-trained VLMs, TOFA employs both visual and textual pipelines to extract task-relevant representations. In the visual pipeline, a hierarchical Bayesian model learns **personalized**, class-specific prototype distributions. For the textual pipeline, TOFA evaluates and globally aligns the generated local text prompts for **robustness**. An adaptive weight calibration mechanism is also introduced to combine predictions from both modalities, balancing personalization and robustness to handle data heterogeneity. Our method is training-free, not relying on additional training resources on either the client or server side. Extensive experiments across 9 datasets in various federated settings demonstrate the effectiveness of the proposed TOFA method.

1 Introduction

Federated learning (FL) (McMahan et al. 2017), a distributed machine learning paradigm, enables multiple clients to collaboratively refine a shared model while preserving their data privacy. Recent advancements in large-scale pre-trained models, especially Vision-Language Models (VLMs) such as CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021), have gained widespread attention within FL, driven by their impressive capability to learn transferable representations. An increasing body of research focuses on adapting VLMs to improve performance in downstream tasks, particularly through fine-tuning and prompt learning (Zhang et al. 2025a; Lu, Tong, and Ye 2025; Lu and Yin 2025; Pan, Huang, and Shi 2024; Li et al. 2025a,b; Wang et al. 2025a,b; Cui et al. 2024; Xu et al. 2025; Zeng et al. 2025; Zhang et al. 2025b; Feng et al. 2025; Qiu et al. 2024).

However, the large parameter sizes inherent to VLMs result in substantial communication and computation overhead, significantly hindering their practical applicability in FL scenarios. Most federated VLM adaptation methods, whether fine-tuning or prompt learning, heavily rely on multi-round interactions between the server and clients, increasing communication burdens and requiring sustained system robustness and reliability (Kairouz et al. 2021; Liu et al. 2024). Moreover, a large number of clients and servers (e.g., most mobile devices and capability-limited servers) lack the computational resources necessary for model training (Abreha, Hayajneh, and Serhani 2022; Bonawitz et al. 2019; Mammen 2021), thereby impeding the deployment of VLMs in distributed settings. Recently, the one-shot FL technique emerges as an effective method to minimize communication overhead by consolidating client-server interactions into a single round (Liu et al. 2024; Allouah et al. 2024a; Tang et al. 2024; Zhang, Liu, and Wang 2024), significantly reducing communication overhead while preserving privacy. Motivated by this, it is compelling to formulate a **lightweight one-shot federated adaptation framework for VLMs** to address the aforementioned issues.

Despite the promising advancements of one-shot techniques, developing both training-free and one-shot federated VLMs adaptation methods still faces certain challenges: (1) *Insufficient exploitation of the rich multimodal information inherent in VLMs.* Most one-shot training frameworks are designed for traditional federated model training (Liu et al. 2024; Allouah et al. 2024a; Tang et al. 2024; Zhang, Liu, and Wang 2024), primarily underlining the visual modality and lacking the capacity to leverage the rich multimodal in-

*Yuyuan Li is the corresponding author.

formation from VLMs. These methods are ill-suited for efficiently adapting pre-trained VLMs to the FL setting, as they fail to capture the critical interactions between visual and textual modalities. (2) *Lack of tailored adaptation strategies to effectively handle severe data heterogeneity.* The data heterogeneity within decentralized training frameworks further impedes the development of lightweight adaptation methods for VLMs in FL, where the data distributions among clients are non-identically and non-independently distributed (non-IID) (Li et al. 2020, 2021b; Liu et al. 2025a,b). Such non-IID nature causes distributional shifts from client data to the global data, leading to discrepancies in local and global optimization objectives, which prevent existing VLM adaptation methods from meeting clients' personalized demands, underscoring the necessity for adaptation strategies tailored to federated settings. (3) *Additional training resource requirements for the clients or server.* Existing VLMs adaptation methods within FL typically depend on additional model training resources on the client or server side (Yang et al. 2024; Luo, Chen, and Wu 2025; Tran et al. 2025). These methods are incompatible in the ubiquitous resource-constrained distributed environments (Mammen 2021).

In this paper, we introduce a novel training-free one-shot federated adaptation approach for VLMs to bridge these gaps, named TOFA. **To fully leverage the generalizable multimodal features within VLMs**, our approach employs both visual and textual pipelines to extract task-relevant representations for downstream classification. For the visual pipeline, we utilize a hierarchical Bayesian model to model the heterogeneous feature representations of class-specific prototypes derived from the visual encoder, with the global information serving as the prior for the inference of local feature distributions. The classification probability for downstream tasks is derived using Gaussian Discriminant Analysis (GDA) on the posterior prompt distributions of class prototypes. For the textual pipeline, TOFA aims to extract robust and generalizable augmented textual inputs from the textual modality. The raw text inputs in the original CLIP-based classification are first augmented using Large Language Models. These inputs are then evaluated for quality on each client and globally aligned to select robust text prompts that consistently show high importance scores across diverse local environments. **To address the impact of data heterogeneity**, our approach fuses personalized visual representations with robust text augmentations to integrate both global and local information, learning locally adaptive features while preventing the model from overfitting. Specifically, TOFA introduces an adaptive weight calibration technique that adjusts the sample-wise contributions of the textual and visual modalities based on their prediction confidence to strike a balance between personalization and generalization. Furthermore, throughout the training process, our approach operates **without relying on training resources on either the server or client side**, thereby enhancing its practicality and flexibility.

We demonstrate the effectiveness of TOFA by conducting experiments across nine datasets containing vision datasets and domain datasets within various data heterogeneous environments. TOFA has a consistent and significant improvement over existing one-shot baselines, and even surpasses several training-based federated VLM adaptation methods. Our main contributions can be summarized as follows:

- To our best of knowledge, we are the first to propose an effective training-free one-shot adaptation method for VLMs in the FL setting.
- We propose an one-shot visual pipeline learn the personalized class-specific prompt distribution over visual representation and a textual pipeline to extract robust text augmentations though global text alignment.
- We propose a sample-adaptive modality weight calibration method to integrate personalized visual representations and robust text representations, allowing the model to handle data heterogeneity within FL.
- We conducted extensive experiments on widely adopted datasets in various data heterogeneity, and significant result improvement verifies the superiority of TOFA.

2 Preliminary

In this section, we focus on the background of VLMs and FL. Detailed related work is presented in **Appendix A**.

Contrastive Language-Image Pretraining (CLIP). CLIP (Radford et al. 2021) consists of a visual encoder $\Phi_{\mathbf{V}}(\mathbf{x})$ and text encoder $\Phi_{\mathbf{T}}(\mathbf{t})$, each producing a normalized d-dimensional embedding from an arbitrary image $\mathbf{x}$, and word embeddings $\mathbf{t}$. Once trained, CLIP enables zero-shot C-class image classification by generating each of the c classifier weights $\mathbf{w}_c$ as the d-dimensional text encoding $\Phi_{\mathbf{T}}(\mathbf{t}_c)$. Here $\mathbf{t}_c$ results from adding the class-specific word embedding $\mathbf{e}_c$ to a pre-defined prompt p, i.e., $\mathbf{w}_c = \Phi_{\mathbf{T}}(\mathbf{t}_c)$ with $\mathbf{t}_c = \{\mathbf{p}, \mathbf{e}_c\}$. The prompt $\mathbf{p}$ is manually crafted to capture the semantic meaning of the downstream task, e.g., $\mathbf{t}_c$ = "A photo of a {class}". Given the image embedding $\mathbf{z} = \Phi_{\mathbf{V}}(\mathbf{x})$, the probability of the image $\mathbf{x}$ being classified as $y \in \{1, \ldots, C\} := [C]$ is thus defined as

$$p(y \mid \mathbf{x}) = \frac{\exp\left(\mathbf{z}^\top \mathbf{w}_y / \tau\right)}{\sum_{i=1}^{C} \exp\left(\mathbf{z}^\top \mathbf{w}_i / \tau\right)} \tag{1}$$

where τ is a temperature parameter.

Federated Learning (FL). Consider a federated learning scenario involving K clients and a central server, and each client k holds a local dataset $D^k = \left\{(\mathbf{x}_i^k, y_i^k)\right\}_{i=1}^{N^k}, k = 1, \cdots, K$ containing N^k samples. Let $D = \bigcup_{k=1}^{K} D^k$ represent the total datasets where each dataset is derived from a distinct data distribution $\mathcal{D}_k$. Generally, federated learning is defined as an optimization problem (McMahan et al. 2017; Kairouz et al. 2021) for maximizing a global objective function $\mathbb{F}(\theta)$, which is a mixture of local objective functions $\mathbb{F}^k(\theta, D^k)$, namely $\mathbb{F}(\boldsymbol{\theta}) = \sum_{k=1}^{K} \mathbb{F}_k(\boldsymbol{\theta}, D_k)$, where θ is the parameter vector of the global model.

Notations. In this paper, vectors are denoted by bold lowercase letters, and matrices by bold uppercase letters. In the context of federated dataset, a superscript k denotes data belonging to client k, a subscript c denotes class-c samples. For example, the data belonging to c-th class in total dataset is denoted as D_c, with its cardinality N_c. Similarly, the data

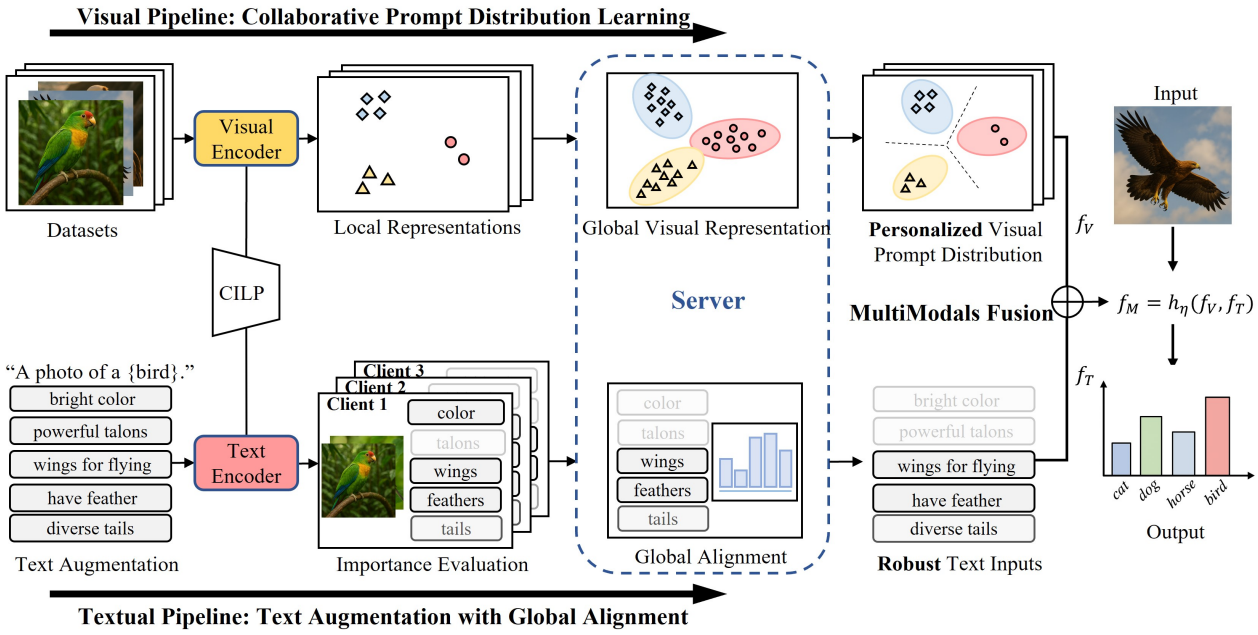

Figure 1: Overall Framework of TOFA

of the k-th client belonging to class c is denoted as D_c^k, and its cardinality as N_c^k. The visual representation of i-th data $\mathbf{x}_{c,i}^k$ from D_c^k can be computed as $\mathbf{z}_{c,i}^k = \Phi_{\mathbf{V}}(\mathbf{x}_{c,i}^k)$.

3 Methodology

3.1 Overview

In this section, we present the design of our TOFA framework, as illustrated in Figure 1, consisting of a visual pipeline, a textual pipeline, and an adaptive multimodal fusion module. To **fully exploit the rich modal information in pre-trained VLMs** and **strike a balance between global generalization and local personalization**, TOFA utilizes prompt distribution learning to extract personalized visual representations, enhances the robustness of LLM-based text augmentation through global alignment of importance scores, and integrates these components with an adaptive multimodal fusion mechanism. To reduce communication costs and address data heterogeneity, without loss of generality, each client adopts a pre-trained CLIP model as the backbone model and has access to an LLM with consistent versions and synchronized parameters. TOFA is a training-free, one-shot federated framework that completes VLM adaptation in a single round without relying on gradient-based model optimization, thereby bolstering both flexibility and implementation simplicity.

3.2 Collaborative Prompt Distribution Learning

To capture the diverse visual variations, our method seeks to model the distribution of class prototypes in feature space. Prior research (Wang et al. 2024b; Lu et al. 2022; Zhu et al. 2024a) have shown that Gaussian distributions effectively model the distribution of CLIP features, resulting in significant performance improvements. Motivated by these observations, we assume $\mathcal{N}(\mathbf{w}_{1:C}, \mathbf{\Sigma})$ with identical covariance is the underlying class-specific prompt distribution over global and each client's local visual representations, where $\mathbf{w}_c \in \mathbb{R}^d$ presents the mean of the embedding distribution for the c-th class, $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ denotes the shared covariance, and the density $\mathcal{N}(\mathbf{z}; \mathbf{w}_c, \mathbf{\Sigma})$ is

$$\frac{1}{\sqrt{(2\pi)^d|\mathbf{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{w}_c)^\top \mathbf{\Sigma}^{-1}(\mathbf{z}-\mathbf{w}_c)\right\}.$$

For global dataset, the prompt distribution over visual representation $p(\mathbf{z} \mid y=c)$ approximated by $\mathcal{N}(\mathbf{w}_c, \mathbf{\Sigma})$. To obtain personalized representations for each client, the visual feature embeddings $p(\mathbf{z} \mid y=c, D^k)$ for $c \in [C]$ are modeled as $\mathcal{N}(\mathbf{w}_{1:C}^k, \mathbf{\Sigma}^k)$, adopting the same distribution form as the global one.

Our goal is to extract personalized visual representations from the global features that are better adapted to the client's local objectives. We first derive the parameters of the global prompt distribution $\theta = (\mathbf{w}_{1:C}, \mathbf{\Sigma})$ using the Bayes' formula. Given a prior distribution $\pi(\theta)$ on the parameters of the Gaussian distribution, the prompt prototypes of the global visual representation are derived from the posterior probability $q(\theta) = \pi(\theta \mid D)$. By Bayes' formula,

$$\pi(\theta \mid D) \propto L(D \mid \theta)\pi(\theta), \tag{2}$$

where $L(D \mid \theta) := \prod_{c\in[C]} \prod_{i=1}^{N_c} \mathcal{N}(\mathbf{z}_{c,i}; \mathbf{w}_c, \mathbf{\Sigma})$ presents the global likelihood function given the parameter θ. Then, hierarchically, the personalized presentation extraction problem aims to deduce the local posterior distribution, with the global prompt distribution serving as the informative prior. Denoting the local parameter as $\theta^k = (\mathbf{w}_{1:C}^k, \mathbf{\Sigma}^k)$, Bayesian inference suggests that the posterior

$$q(\theta^k) \propto L(D^k \mid \theta^k)\pi(\theta^k) \propto L(D^k \mid \theta^k)L(D \mid \theta)\pi(\theta),$$

where we plug in prior $\pi(\theta^k) = q(\theta)$, and $L(D^k \mid \theta^k)$ presents the local likelihood.

Hence, we adopt the power prior to ensure that the posterior is not overly impacted by global information, which introduces a scalar prior parameter $\alpha \in [0, 1]$ that weights the prior distribution relative to the global likelihood. The posterior distribution of θ^k can be written as

$$q(\theta^k) \propto p(D^k \mid \theta^k)[L(D \mid \theta)]^\alpha \pi(\theta). \tag{3}$$

The above personalized prompt distribution learning problem is equivalent to solving the local visual representation $q(\theta^k)$. To ensure computational efficiency and facilitate one-shot adaptation, we design a conjugate prior to construct the hierarchical Bayesian framework with an explicit posterior formulation, as presented in the following lemma.

Lemma 1 *Assume that the mean of each prompt prototype $\mathbf{w}_c$ is independent given shared covariance $\mathbf{\Sigma}$, the hierarchical Bayesian model characterized in* (2) *and* (3) *exists a conjugate prior $\pi(\theta)$ over parameter $(\mathbf{w}_{1:C}, \mathbf{\Sigma})$:*

$$\mathbf{\Sigma} \sim \mathcal{IW}(\mathbf{\Sigma}; \mathbf{S}_0, \nu_0), \quad \mathbf{w}_c|\mathbf{\Sigma} \sim \mathcal{N}(\mathbf{z}; \mathbf{m}_{0,c}, \frac{1}{\kappa_{0,c}}\mathbf{\Sigma}),$$

where $c \in [C]$ and $\mathcal{IW}(\cdot)$ denote the Inverse-Wishart distribution. Specifically, denoting the sample number count for class c as N_c and the embedding for the i-th sample in class c as $\mathbf{z}_{c,i}$, the posterior distribution can be formulated as $\mathcal{N}(\mathbf{z}; \mathbf{w}^*_{1:C}, \mathbf{\Sigma}^*)$, and for $c \in [C]$,

$$\mathbf{\Sigma}^* \sim \mathcal{IW}(\mathbf{\Sigma}; \mathbf{S}_q, \nu_q), \quad \mathbf{w}^*_c|\mathbf{\Sigma}^* \sim \mathcal{N}(\mathbf{z}; \mathbf{m}_{q,c}, \frac{1}{\kappa_{q,c}}\mathbf{\Sigma}^*).$$

The parameters are specified as

$$\begin{aligned}
&\kappa_{q,c} = \kappa_{0,c} + N_c, \qquad \mathbf{m}_{q,c} = \frac{\kappa_{0,c}\mathbf{m}_{0,c} + N_c\bar{\mathbf{z}}_c}{N_c + \kappa_{0,c}} \\
&\nu_q = \nu_0 + \sum_{c\in[C]} N_c \\
&\mathbf{S}_q = \mathbf{S}_0 + \sum_{c\in[C]} \mathbf{S}_c + \sum_{c\in[C]} \left(\kappa_{0.c}\mathbf{m}_{0,c}\mathbf{m}^\top_{0,c} - \kappa_{q,c}\mathbf{m}_{q,c}\mathbf{m}_{q,c}\right),
\end{aligned} \tag{4}$$

where $\mathbf{S}_c := \sum_{i=1}^{N_c} \mathbf{z}_{c,i}\mathbf{z}^\top_{c,i}$ and $\bar{\mathbf{z}}_c := \frac{1}{N_c}\sum_{i=1}^{N_c} \bar{\mathbf{z}}_{c,i}$.

For the above hierarchical Bayesian model, an uninformative prior has the form $\mathbf{S}_0 = s_0\mathbf{I}, \mathbf{m}_{c,0} = \mathbf{0}, \nu_0 = 0$ and $s_0 \approx \kappa_{c,0} \approx 0$ to some small positive number. We now provide a detailed description of each step in the visual pipeline:

Step 1: Clients transmit local statistics to the server, which computes the posterior under uninformative prior via Lemma 1, as global class-specific prototypes characterized by $\mathbf{S}_g, \nu_g, \mathbf{m}_{g,c}, \kappa_{g,c}$ for $c \in [C]$.

Step 2: The server sends global prototypes to each client.

Step 3: Each client k derives personalized, class-specific prototypes, parameterized by $\mathbf{S}^k_l, \nu^k_l, \mathbf{m}^k_{l,c}, \kappa^k_{l,c}$, for $c \in [C]$, using the global prototypes distribution as the prior via Lemma 1.

The proof of Lemma 1 and the detailed expressions for the global and personalized prompt prototype parameters are provided in **Appendix B.1** with the **computation and privacy analysis**.

Furthermore, we utilize the maximum a posteriori (MAP) estimation of the local prompt distribution for the hierarchical model presented above,

$$\arg\max_{\mathbf{m},\mathbf{\Sigma}} f_{p(\mathbf{w}^k_{l,c},\mathbf{\Sigma}^k)}(\mathbf{m}, \mathbf{\Sigma}|D, D^k) = (\mathbf{m}^k_{l,c}, \frac{\mathbf{S}^k_l}{\nu^k_l + d + 2}).$$

If $\widehat{\mathbf{\Sigma}}^k_l = \mathbf{S}^k_l/(\nu^k_l + d + 2))$ is positive definite, (3.2) indicates that $\mathcal{N}(\mathbf{z}; \mathbf{m}^k_{l,c}, \widehat{\mathbf{\Sigma}}^k_l)$ is the most probable distribution within the Gaussian family. By applying Gaussian Discriminant Analysis (GDA), the personalized classification probability $f^k_V(\mathbf{z})$ can be formulated as

$$\begin{aligned}
p(y = c \mid \mathbf{z}, D^k) &= \frac{p(\mathbf{z} \mid y = c, D^k)p(y = c \mid D^k)}{\sum_{i=1}^C p(\mathbf{z} \mid y = i, D^k)p(y = i \mid D^k)} \\
&= \frac{\exp\left((\mathbf{m}^k_{l,c})^\top\mathbf{G}^k\mathbf{z} - \frac{1}{2}(\mathbf{m}^k_{l,c})^\top\mathbf{G}^k\mathbf{m}^k_{l,c} + \log p^k_c\right)}{\sum_{i=1}^C \exp\left((\mathbf{m}^k_{l,i})^\top\mathbf{G}^k\mathbf{z} - \frac{1}{2}(\mathbf{m}^k_{l,i})^\top\mathbf{G}^k\mathbf{m}^k_{l,i} + \log p^k_i\right)},
\end{aligned}$$

where $\mathbf{G}^k := \left(\widehat{\mathbf{\Sigma}}^k_l\right)^{-1}$, and $p^k_c = p(y = c \mid D^k) := 1/C$ is the class-wise prior probability for $c \in [C]$, which is set to be uniform during the prediction phase.

3.3 Text Augmentation with Global Alignment

The textual pipeline produces globally robust text prompts to complement personalized prompt distributions and mitigate data heterogeneity. Each client uses a local LLM to generate augmented descriptions, evaluates their reliability, and assigns weights. A global alignment procedure then aggregates these weighted descriptions into a robust, dataset-wide text augmentation.

To enhance the textual modality, we adopt the two-step prompt augmentation from (Zhu et al. 2024b), using LLMs to generate dataset-aware class descriptions. LLMs produce probing questions from dataset-level summaries. These are then combined with class names to form tailored descriptions, ensuring both diversity and visual relevance. For each class c, the augmented text set is $\{\mathbf{t}^m_c\}^M_{m=1}$, supplemented by the manual prompt "A photo of a class" denoted as $\mathbf{t}^0_c$.

Subsequently, we propose a global text prompt alignment method within FL environments, designed to extract text augmentations that exhibit both generalization and robustness at the global level. To determine the importance of text description for the j-th class, each client calculates the classification probability with the text input $\mathbf{t}$:

$$p^k_c(\mathbf{t}) = \frac{\exp\left(\frac{1}{N^k_c}\sum_{i=1}^{N^c_j} \Phi_\mathbf{T}(\mathbf{t})^\top\mathbf{z}^k_{c,i}\right)}{\sum_{c'=1}^C \exp\left(\frac{1}{N^k_{c'}}\sum_{i=1}^{N^k_{c'}} \Phi_\mathbf{T}(\mathbf{t})^\top\mathbf{z}^k_{c',i}\right)}. \tag{5}$$

With augmented texts $\mathbf{t}^m_c$, $p^k_j(\mathbf{t}^m_c)$ represents the significance of the text prompt $\mathbf{t}^m_c$ for class j, as evaluated from the visual feature embeddings. For the robust text augmentation of c-th class, the most important property is its capability

to distinguish the c-th class's image features from those of other classes. Therefore, on the server side, we introduce a significance scoring criterion based on the manually defined prompts $\mathbf{t}_c^0$,

$$r(\mathbf{t}_c^m) = \frac{1}{K}\sum_{k=1}^{K} u^k(\mathbf{t}_c^0)\log\left(\frac{u^k(\mathbf{t}_c^m)}{u^k(\mathbf{t}_c^0)}\right), \tag{6}$$

where $u^k(\mathbf{t}_c^m) := p_c^k(\mathbf{t}_c^m) - \max_{j\neq c} p_j^k(\mathbf{t}_c^m)$ indicates the confidence of the text input $\mathbf{t}_c^m$ in local classification tasks. Since manually designed inputs $\mathbf{t}_c^0,\ c \in [C]$ are considered the most robust text inputs across various environments, the score function (7), which is analogous to the KL divergence, assigns a higher importance score to text prompts with stronger robustness, approaching or exceeding that of $\mathbf{t}_c^0$, in federated downstream tasks. We then weight the descriptions within the c-th class based on the importance scores as follows:

$$\mathbf{b}(\mathbf{t}_c^m) = \frac{\exp\left(r(\mathbf{t}_c^m)/\tau_t\right)}{\sum_{m=0}^{M}\exp\left(r(\mathbf{t}_c^m)/\tau_t\right)}, \quad m = 0,\ldots,M, \tag{7}$$

where τ_t is the temperature parameter. Here we set $\tau_t = 0.5$ to assign a higher weight to robust text prompts. During the classification phase, denoting $\mathbf{z} = \Phi_{\mathbf{V}}(\mathbf{x})$, the prediction probability of image $\mathbf{x}$ being classified as class c can be computed as

$$f_T(\mathbf{z}) = \frac{\exp\left(\sum_{m=0}^{M}\mathbf{b}(\mathbf{t}_c^m)\mathbf{z}^\top\Phi_{\mathbf{T}}(\mathbf{t}_c^m)\right)}{\sum_{j=1}^{C}\exp\left(\sum_{m=0}^{M}\mathbf{b}(\mathbf{t}_c^m)\mathbf{z}^\top\Phi_{\mathbf{T}}(\mathbf{t}_j^m)\right)}. \tag{8}$$

3.4 Adaptive Multimodals Fusion

As a general rule within FL, combining the global robust model with locally personalized models can further improve performance under data heterogeneity (Li et al. 2021a, 2020; Guo, Guo, and Wang 2023; Wang et al. 2024a; Zhu et al. 2024a). To further enhance the effectiveness of our method in data heterogeneous environments, we propose a sample-wise ensembling technique that adaptively calibrates inter-modal weights to fuse the personalized visual representations and robust language modalities.

The key in our prediction fusion lies in introducing a sample-wise mixing coefficient $\eta(\mathbf{z})$ to balance the contributions of both modalities, namely

$$f_M^k(\mathbf{z}) = \eta(\mathbf{z})f_V^k(\mathbf{z}) + (1-\eta(\mathbf{z}))f_T(\mathbf{z}).$$

The following theorem provides the theoretical motivation for this module's design.

Theorem 1 *Let $f(\mathbf{z}) := \eta(\mathbf{z})f_1(\mathbf{z}) + (1-\eta(\mathbf{z}))f_2(\mathbf{z})$, where $f_1, f_2 \in \mathcal{H}$ are two prediction functions, and $\mathcal{H} : \mathcal{X} \to \{-1,+1\}$ is a hypothesis set. Denoting the empirical predictive errors on $\mathcal{D}_{train} = \{\mathbf{z}_i, y_i\}_{i=1}^N$ as $\widehat{\mathcal{R}}(f_i), i = 1,2$, and the VC dimension of $\mathcal{H}$ as $d_{VC}(\mathcal{H})$, then with probability at least $1-\delta$ over the samples,*

$$\mathcal{R}(f) \le B\sqrt{\frac{d_{VC}(\mathcal{H}) + \log 1/\delta}{N}} + \sum_{i=1,2}\widehat{\mathcal{R}}(f_i) + \mathrm{Cov}\left(\eta(\mathbf{z}), \ell_1(\mathbf{z}) - \ell_2(\mathbf{z})\right),$$

where $\mathcal{R}(f) = \mathbb{E}_{(\mathbf{z},y)\sim\mathcal{D}}[\ell(f(\mathbf{z}),y)]$, ℓ is the cross-entropy loss, Cov *denotes the covariance and B is a constant.*

The proof is detailed in Appendix B.2. According to the above expression, minimizing the generalization error requires η to be proportional to $\ell_2 - \ell_1$, while $\ell := -\log p_{true}$.

Kumar et al. (2022) have shown that a well-calibrated classifier's confidence serves as a reasonable surrogate for its true accuracy p_{true}. Formally, the average confidence over the dataset $\{x_i\}_{i=1}^N$ scaled by a temperature $\tau > 0$ is given by the average of the model's probability for its prediction:

$$\mathrm{conf}(f,\tau) = \frac{1}{N}\sum_{i=1}^{N}\max_j\left[\mathrm{softmax}\left(f\left(x_i\right)/\tau\right)\right]_j.$$

To approximate a well-calibrated classifier, the average confidence of model f should reflect the predicted accuracy, i.e., $\mathrm{conf}(f,\tau) \approx \mathrm{Acc}(f)$. This can be implemented by binary search of τ, which works since the confidence increases when τ decreases. Aiming to ensure that the fused classifier minimizes the generalization error, the weight $\eta(\mathbf{z})$ can be designed as $\eta(\mathbf{z}) = \frac{1}{1+e^{-L(\mathbf{z})}}$, where

$$L(\mathbf{z}) := \log\left(\frac{\max_j[\mathrm{softmax}(f_V^k(\mathbf{z}))]_j}{\max_j[\mathrm{softmax}(f_T(\mathbf{z}))]_j}\right).$$

Thus, the fused classifier naturally favors the modality with higher predictive accuracy for given regions, thereby enhancing overall performance.

4 Experiment

4.1 Setup

Due to space limitations, the detailed information in this section is provided in **Appendix C**.

Datasets. We assess the effectiveness of the proposed TOFA across nine publicly available benchmark datasets under various federated configurations to simulate different types of data heterogeneity: (1) Five representative visual classification datasets commonly employed to test few-shot performance in the CLIP benchmark (Radford et al. 2021): **OxfordPets**, **Flowers102**, **DTD**, **Caltech101**, and **Food101**, hereafter referred to collectively as the CLIP datasets. (2) Two standard image benchmarks, **CIFAR-10** and **CIFAR-100** (Gong et al. 2012). (3) Two multi-domain datasets with feature shift (Li et al. 2021b): **DomainNet** (Peng et al. 2019) (six domains), and **Office-Caltech10** (Gong et al. 2012) (four domains).

Baselines. Since *no previous work was found that investigates training-free and one-shot distributed adaptation method for VLMs* within a FL framework, we compare the performance of TOFA with four categories of baselines: (1) Four existing prompt learning federated learning methods: **PromptFolio** (Pan, Huang, and Shi 2024), **DP-PFL** (Tran et al. 2025), **PromptFL** (Guo et al. 2023); **pFedPrompt** (Guo, Guo, and Wang 2023). (2) Three local adapting methods: **Zero-shot CLIP** (Radford et al. 2021) with hand-crafted text prompt templates; **CLIP-GDA** (Wang et al. 2024b); **CoOp** (Zhou et al. 2022). (3) Three adapted methods derived from advanced one-shot techniques that operate solely with client-side training resources, combined

Method	Training-free	One-shot	OxfordPets	Flowers102	Food101	Caltech101	DTD
CoOp	✗	✗	89.18	69.03	82.54	90.62	63.97
PromptFolio	✗	✗	92.08	74.61	86.50	93.59	65.04
DP-PFL	✗	✗	96.91	85.75	86.08	96.76	86.23
PromptFL	✗	✗	90.79	92.26	88.17	87.90	50.46
pFedPromp	✗	✗	91.84	96.46	92.26	96.54	77.14
Zero-Shot CLIP	✓	✓	85.77	66.14	77.31	86.29	42.32
CLIP-GDA	✓	✓	88.81	91.23	79.05	92.55	60.64
FedLPA+PromptFL	✗	✓	83.42	78.60	74.74	88.69	52.75
FENS+PromptFL	✗	✓	90.51	81.19	80.80	84.37	68.43
FedBEns+PromptFL	✗	✓	79.84	86.27	78.45	86.58	65.82
TOFA (Ours)	✓	✓	**91.23**	**95.78**	**85.49**	**94.58**	**71.68**

* Blue denotes the highest results of multi-round methods. **Bold** denotes the highest results of one-shot methods.

Table 1: Few-shot Performance on CLIP Datasets over 10 Clients.

with the backbone prompt learning method (Guo et al. 2023) in FL: **FedLPA** (Liu et al. 2024), **FENS** (Allouah et al. 2024b), **FedBEns** (Talpini, Savi, and Neglia 2025). (4) **FedAvg** (McMahan et al. 2017) is included as a traditional baseline in experiments on image datasets.

Implementation details. (1) CLIP datasets. Each dataset in CLIP datasets is partitioned into $N = 10$ clients, defaulting if not explicitly specified, each with a disjoint set of classes evenly and randomly assigned to the clients. (2) CIFAR10 and CIFAR100. We split $N = 100$ clients resulting from $Dir(\beta = 0.3)$ partition. (3) DomainNet and Office-Caltech10. Each client in the federated system is assigned data from a single unique domain, Consistent with prior work (Li et al. 2021b; Cui et al. 2024). We present the results using ViT-B16 (Dosovitskiy et al. 2020) backbones. For other hyperparameters, such as learning rate and local epochs in the aforementioned baselines, we adhere to the original configurations from these studies. Given that the results of our model are deterministic due to its training-free nature, we do not present results with statistical variations, which is typical in zero-shot or training-free studies (Zhu et al. 2024a,b; Wang et al. 2024b).

4.2 Overall Comparison

Model evaluation on label shifts. We began by assessing performance of TOFA against baselines on datasets with label shift. We conducted few-shot experiments on the CLIP datasets and standard training experiments on the CIFAR datasets. In Table 1, we present the numerical results of the training-required/training-free and multi-round/single-round baselines under the 16-shot setting. Results show that TOFA consistently exceeds the performance of one-shot baselines on all five datasets, even exceeds many muti-round prompt learning methods, owing to full exploration of modal interaction information from pre-trained VLMs. For example, our method consistently outperforms CoOp, PromptFolio, and PromptFL across five vision datasets, except for Food101. Notably, on the DTD dataset, where training-free methods struggle to perform well, TOFA still outperforms all one-shot baselines and shows only a slight performance gap compared to multi-round methods like PromptFolio and PromptFL. Demonstrating TOFA's remarkable ability to adapt to resource-constrained scenarios in few-shot settings. Table 2 presents the comparison results on CIFAR datasets, which are partitioned under Dirichlet setting $\beta = 0.3$ over 100 clients. This result further corroborates the efficacy of our method in handling extreme data heterogeneity. It also empirically demonstrates the scalability of TOFA with respect to the number of clients.

Method	Cifair10	Cifair100
FedAvg	75.10	42.52
Zero-Shot CLIP	87.71	64.92
CoOp	93.11	74.83
PromptFL	92.30	73.67
TOFA (Ours)	93.18	76.63

Table 2: Results on CIFAR10 & CIFAR100

Model evaluation on feature shifts. To assess the performance of our method in scenarios more closely resembling real-world FL applications, we examine TOFA on real-world data with feature shift (Li et al. 2021b) using DomainNet and Office-Caltech10, where each client is assigned with single domain dataset, resulting in 6 clients for DomainNet and 4 clients for Office-Caltech10. In Table 3, we report the performance of our method against other baselines on these two domain datasets. We present the maximum accuracy of these methods combining one-shot techniques with prompt-based FL (Oneshot+PromptFL). The experimental results show that training-free methods, such as Zero-Shot CLIP and CLIP-GDA, struggle to benefit the clients, particularly on the DomainNet dataset. However, our method achieved the highest average accuracies 93.05% and 98.69% on Office-Caltech10 and DomainNet, respectively. This result surpasses most methods requiring multiple rounds of training and achieves a performance within 2% of the optimal prompt-based FL baseline. This validates the effectiveness and robustness of TOFA in scenarios closer to real-world federated settings.

Datasets	DomainNet							Office-Caltech10				
Domains	C	I	P	Q	R	S	Avg.	A	C	D	W	Avg.
CoOp	98.32	83.01	98.18	82.37	98.21	97.70	92.97	96.38	97.24	100	98.31	97.98
PromptFolio	98.38	83.07	98.24	82.43	98.27	97.84	93.04	95.64	96.50	99.26	97.57	97.24
DP-PFL	98.93	84.52	98.89	87.87	98.64	98.02	94.48	97.92	97.68	100	100	98.90
PromptFL	98.23	79.91	97.89	66.52	96.83	97.31	89.45	96.41	96.39	96.90	100	97.43
pFedPromp	98.14	82.43	98.26	86.52	96.98	98.42	93.46	97.12	98.18	96.85	100	98.04
Zero-Shot CLIP	72.32	47.15	53.63	31.30	48.40	50.18	50.50	19.30	18.20	21.90	18.60	19.50
CLIP-GDA	71.72	60.75	64.39	67.70	66.48	69.19	66.71	96.08	**98.20**	98.30	100	98.15
Oneshot+PromptFL	85.42	75.68	86.97	73.96	84.74	89.61	82.73	94.93	96.21	96.89	99.20	96.81
TOFA (Ours)	**98.86**	**82.69**	**97.45**	**83.37**	**98.12**	**97.83**	**93.05**	**96.94**	97.81	**100**	**100**	**98.69**

* Blue denotes the highest results of multi-round training methods. **Bold** denotes the highest results of one-shot methods.

Table 3: Experimental Results on Office-Caltech10 and DomainNet Datasets with Feature Shift.

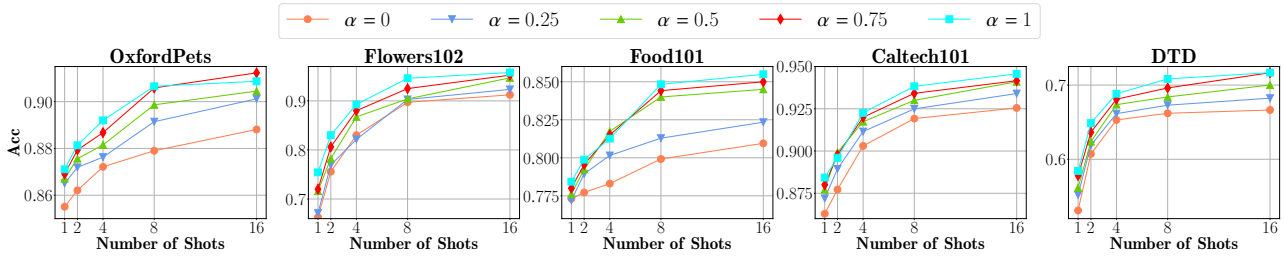

Figure 2: Comparisons on CLIP datasets across varying shot numbers and parameter α in TOFA over 10 clients.

4.3 Ablation Experiments

Impact of number of α. In TOFA, $\alpha \in [0, 1]$ is the coefficient for adjusting the contribution of global information in local distribution posterior inference. Figure 2 displays the results for α values ranging from 0 to 1 across the CLIP datasets. The trends reveal that although the optimal α value varies across datasets, assigning a higher weight to global information (e.g. $\alpha \geq 0.75$) can achieve near-optimal performance. Based on this observation, we adopt $\alpha = 1$ for experiments in this section.

Impact of number of shot. We also investigate the impact of shots in the few-shot learning for TOFA. Figure 2 also presents the impact of shots in the few-shot learning for TOFAacross the CLIP datasets. The number of shots varies from $[1, 2, 4, 8, 16]$. The result shows an explicit improvement in test accuracy with an increase in the number of shots. Based on the accuracy shown in the figure, our method achieves stable results starting from the 8-shot classification.

Inter-modality ablation experiments. We conducted an inter-modality ablation study to analyze the impact of different modalities on the overall performance of TOFA, as presented in Figure 3. The results indicate that the accuracy of both the visual and textual modalities before fusion is lower than that of the fused TOFA model. This demonstrates that combining personalized visual information with robust textual prompts effectively prevents overfitting in the fused model, thereby improving accuracy. It further validates the importance of multimodal information fusion within downstream tasks of VLMs, which, compared to single-modal approaches, enables models to capture more precise and generalizable representations.

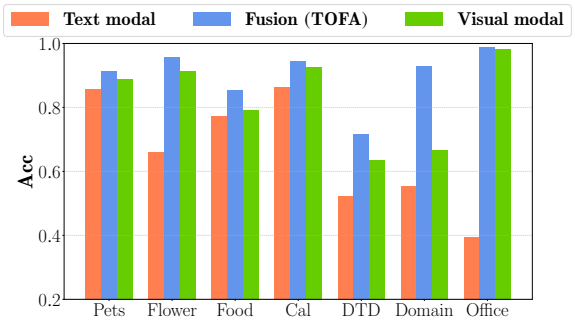

Figure 3: Results on Inter-Modality Ablation Experiments.

5 Conclusion

We propose TOFA, a novel Training-free One-shot Federated Adaptation framework for VLMs in federated learning. To our knowledge, it is the first approach to realize training-free VLM adaptation with a single communication round. TOFA leverages both visual and textual pipelines to extract task-specific, generalizable multimodal representations. In the visual pipeline, a hierarchical Bayesian model learns personalized, class-specific prototype distributions over visual features. In the textual pipeline, TOFA evaluates and globally aligns generated descriptions to ensure robustness. An adaptive weight calibration mechanism then fuses predictions from both modalities, trading off personalization and robustness to mitigate data heterogeneity. The method requires no additional model training on either clients or the server. Extensive experiments on nine datasets under diverse federated settings demonstrate the effectiveness of TOFA.

Acknowledgments

This work was supported by the National Key R&D Program of China (2024YFB3908400, 2024YFB3908403), the National Natural Science Foundation of China (No.62172362), and the Ant Group Research Fund.

References

Abreha, H. G.; Hayajneh, M.; and Serhani, M. A. 2022. Federated learning in edge computing: a systematic survey. *Sensors*, 22(2): 450.

Allouah, Y.; Dhasade, A.; Guerraoui, R.; Gupta, N.; Kermarrec, A.-M.; Pinot, R.; Pires, R.; and Sharma, R. 2024a. Revisiting Ensembling in One-Shot Federated Learning. *arXiv preprint arXiv:2411.07182*.

Allouah, Y.; Dhasade, A.; Guerraoui, R.; Gupta, N.; Kermarrec, A.-M.; Pinot, R.; Pires, R.; and Sharma, R. 2024b. Revisiting Ensembling in One-Shot Federated Learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Bonawitz, K.; Eichner, H.; Grieskamp, W.; Huba, D.; Ingerman, A.; Ivanov, V.; Kiddon, C.; Konečnỳ, J.; Mazzocchi, S.; McMahan, B.; et al. 2019. Towards federated learning at scale: System design. *Proceedings of machine learning and systems*, 1: 374–388.

Cui, T.; Li, H.; Wang, J.; and Shi, Y. 2024. Harmonizing Generalization and Personalization in Federated Prompt Learning. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 9646–9661. PMLR.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Feng, X.; Li, Y.; Chen, C.; Zhang, L.; Li, L.; ZHOU, J.; and Zheng, X. 2025. Controllable Unlearning for Image-to-Image Generative Models via ϵ-Constrained Optimization. In *The Thirteenth International Conference on Learning Representations*.

Gong, B.; Shi, Y.; Sha, F.; and Grauman, K. 2012. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE conference on computer vision and pattern recognition*, 2066–2073. IEEE.

Guo, T.; Guo, S.; and Wang, J. 2023. Pfedprompt: Learning personalized prompt for vision-language models in federated learning. In *Proceedings of the ACM Web Conference 2023*, 1364–1374.

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

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 4904–4916. PMLR.

Kairouz, P.; McMahan, H. B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A. N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. 2021. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2): 1–210.

Kumar, A.; Ma, T.; Liang, P.; and Raghunathan, A. 2022. Calibrated ensembles can mitigate accuracy tradeoffs under distribution shift. In *The 38th Conference on Uncertainty in Artificial Intelligence*.

Li, T.; Hu, S.; Beirami, A.; and Smith, V. 2021a. Ditto: Fair and robust federated learning through personalization. In *International conference on machine learning*, 6357–6368. PMLR.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2: 429–450.

Li, X.; Jiang, M.; Zhang, X.; Kamp, M.; and Dou, Q. 2021b. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623*.

Li, Z.; Chen, Z.; Wen, H.; Fu, Z.; Hu, Y.; and Guan, W. 2025a. Encoder: Entity mining and modification relation binding for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5101–5109.

Li, Z.; Fu, Z.; Hu, Y.; Chen, Z.; Wen, H.; and Nie, L. 2025b. FineCIR: Explicit Parsing of Fine-Grained Modification Semantics for Composed Image Retrieval. *https://arxiv.org/abs/2503.21309*.

Liu, J.; Liu, Y.; Shang, F.; Liu, H.; Liu, J.; and Feng, W. 2025a. Improving Generalization in Federated Learning with Highly Heterogeneous Data via Momentum-Based Stochastic Controlled Weight Averaging. In *Forty-second International Conference on Machine Learning*.

Liu, J.; Shang, F.; Tian, Y.; Liu, H.; and Liu, Y. 2025b. Consistency of Local and Global Flatness for Federated Learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, 3875–3883. New York, NY, USA: Association for Computing Machinery. ISBN 9798400720352.

Liu, X.; Liu, L.; Ye, F.; Shen, Y.; Li, X.; Jiang, L.; and Li, J. 2024. Fedlpa: One-shot federated learning with layer-wise posterior aggregation. *Advances in Neural Information Processing Systems*, 37: 81510–81548.

Lu, W.; Tong, Y.; and Ye, Z. 2025. DAMMFND: Domain-Aware Multimodal Multi-view Fake News Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 559–567.

Lu, W.; and Yin, L. 2025. DMMD4SR: Diffusion Model-based Multi-level Multimodal Denoising for Sequential

Recommendation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 6363–6372.

Lu, Y.; Liu, J.; Zhang, Y.; Liu, Y.; and Tian, X. 2022. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5206–5215.

Luo, J.; Chen, C.; and Wu, S. 2025. Mixture of Experts Made Personalized: Federated Prompt Learning for Vision-Language Models. In *The Thirteenth International Conference on Learning Representations*.

Mammen, P. M. 2021. Federated learning: Opportunities and challenges. *arXiv preprint arXiv:2101.05428*.

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

Pan, B.; Huang, W.; and Shi, Y. 2024. Federated Learning from Vision-Language Foundation Models: Theoretical Analysis and Method. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Peng, X.; Bai, Q.; Xia, X.; Huang, Z.; Saenko, K.; and Wang, B. 2019. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1406–1415.

Qiu, C.; Li, X.; Mummadi, C. K.; Ganesh, M. R.; Li, Z.; Peng, L.; and Lin, W.-Y. 2024. Federated Text-driven Prompt Generation for Vision-Language Models. In *The Twelfth International Conference on Learning Representations*.

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

Talpini, J.; Savi, M.; and Neglia, G. 2025. FedBEns: One-Shot Federated Learning based on Bayesian Ensemble. arXiv:2503.15367.

Tang, Z.; Zhang, Y.; Dong, P.; Cheung, Y.-m.; Zhou, A.; Han, B.; and Chu, X. 2024. Fusefl: One-shot federated learning through the lens of causality with progressive model fusion. *Advances in Neural Information Processing Systems*, 37: 28393–28429.

Tran, L.; Sun, W.; Patterson, S.; and Milanova, A. 2025. Privacy-Preserving Personalized Federated Prompt Learning for Multimodal Large Language Models. In *The Thirteenth International Conference on Learning Representations*.

Wang, F.; Chen, C.; Liu, W.; Fan, T.; Liao, X.; Tan, Y.; Qi, L.; and Zheng, X. 2024a. CE-RCFR: Robust counterfactual regression for consensus-enabled treatment effect estimation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3013–3023.

Wang, F.; Chen, C.; Liu, W.; Lei, M.; Chen, J.; Liu, Y.; Zheng, X.; and Yin, J. 2025a. DR-VAE: Debiased and Representation-enhanced Variational Autoencoder for Collaborative Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wang, F.; Chen, C.; Liu, W.; Qi, L.; Zhang, X.; Tan, Y.; Zhu, M.; and Zheng, X. 2025b. Cluster-Enhanced Dual Discrete Collaborative Filtering for Efficient Recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 1–13.

Wang, Z.; Liang, J.; Sheng, L.; He, R.; Wang, Z.; and Tan, T. 2024b. A Hard-to-Beat Baseline for Training-free CLIP-based Adaptation. In *The Twelfth International Conference on Learning Representations*.

Xu, W.; Xiang, D.; Wang, R.; Hu, Y.; Zhang, L.; Chen, J.; and Lu, Z. 2025. Learning explainable stock predictions with tweets using mixture of experts. *arXiv preprint arXiv:2507.20535*.

Yang, M.; Su, S.; Li, B.; and Xue, X. 2024. Exploring One-Shot Semi-supervised Federated Learning with Pre-trained Diffusion Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(15): 16325–16333.

Zeng, Z.; Wu, J.; Luo, M.; Kong, X.; Ma, Z.; Dai, G.; and Zheng, Q. 2025. Understand, Refine and Summarize: Multi-View Knowledge Progressive Enhancement Learning for Fake News Video Detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 9216–9225.

Zhang, J.; Duan, Y.; Niu, S.; CAO, Y.; and Lim, W. Y. B. 2025a. Enhancing Federated Domain Adaptation with Multi-Domain Prototype-Based Federated Fine-Tuning. In *The Thirteenth International Conference on Learning Representations*.

Zhang, J.; Liu, S.; and Wang, X. 2024. One-shot federated learning via synthetic distiller-distillate communication. *arXiv preprint arXiv:2412.05186*.

Zhang, Y.; Li, Z.; Wang, Y.; Chen, F.; Fan, X.; and Zhou, F. 2025b. Navigating Towards Fairness with Data Selection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(21): 22632–22640.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhu, X.; Zhu, B.; Tan, Y.; Wang, S.; Hao, Y.; and Zhang, H. 2024a. Enhancing zero-shot vision models by label-free prompt distribution learning and bias correcting. *Advances in Neural Information Processing Systems*, 37: 2001–2025.

Zhu, Y.; Ji, Y.; Zhao, Z.; Wu, G.; and Wang, L. 2024b. Awt: Transferring vision-language models via augmentation, weighting, and transportation. *arXiv preprint arXiv:2407.04603*.