FedMVP: Federated Multimodal Visual Prompt Tuning for Vision-Language Models

Mainak Singha[1] Subhankar Roy[2] Sarthak Mehrotra[3] Ankit Jha[4]
Moloud Abdar[5] Biplab Banerjee[3] Elisa Ricci[1,6]

[1] University of Trento, Italy [2] University of Bergamo, Italy [3] Indian Institute of Technology Bombay, India
[4] LNMIIT Jaipur, India [5] The University of Queensland, Australia [6] Fondazione Bruno Kessler, Italy

Abstract

*In federated learning, textual prompt tuning adapts Vision-Language Models (e.g., CLIP) by tuning lightweight input tokens (or prompts) on local client data, while keeping network weights frozen. After training, only the prompts are shared by the clients with the central server for aggregation. However, textual prompt tuning suffers from overfitting to known concepts, limiting its generalizability to unseen concepts. To address this limitation, we propose **M**ultimodal **V**isual **P**rompt Tuning (FedMVP) that conditions the prompts on multimodal contextual information – derived from the input image and textual attribute features of a class. At the core of FedMVP is a PromptFormer module that synergistically aligns textual and visual features through a cross-attention mechanism. The dynamically generated multimodal visual prompts are then input to the frozen vision encoder of CLIP, and trained with a combination of CLIP similarity loss and a consistency loss. Extensive evaluation on 20 datasets, spanning three generalization settings, demonstrates that FedMVP not only preserves performance on in-distribution classes and domains, but also displays higher generalizability to unseen classes and domains, surpassing state-of-the-art methods by a notable margin of $+1.57\% - 2.26\%$. Code is available at* https://github.com/mainaksingha01/FedMVP.

1. Introduction

A successful learning paradigm involves training neural networks on large and centralized datasets, enabling models to learn robust and generalizable representations [22, 34]. However, practical constraints often prevent centralized training (*e.g.*, privacy regulations) creating a critical need for decentralized training frameworks that achieve high performance without central data access. As a solution, federated learning (FL) [16] enables multiple clients to collaboratively train a global model without requiring direct data sharing. In FL, clients train models locally and transmit only model updates (*e.g.*, gradients or parameters) to a central server, which aggregates these updates to improve the global model. The research in FL is primarily aimed at: reducing communication rounds between clients and server [31, 32], faster and lightweight training algorithms on clients [19, 23] and improving generalization while exhibiting high data heterogeneity across clients [25, 33].

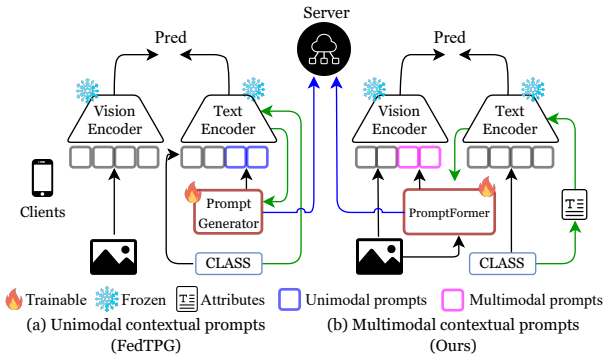


Figure 1. **Comparison of federated prompt tuning methods**: (a) a textual prompt tuning method (*e.g.*, FedTPG [21]) encodes **unimodal** contextual information, *i.e.*, class names, in the learnable prompts, (b) our proposed method FedMVP injects **multimodal** contextual information – image and attributes – into the prompts.

Pretrained Vision-Language Models (VLMs) (*e.g.*, CLIP [22]), have emerged as promising candidates for FL due to their strong generalization capabilities. However, the high communication overhead and computational demands of parameter-heavy VLMs pose challenges for FL, especially in bandwidth- or compute-constrained environments (*e.g.*, robots). To make CLIP viable for the FL setting, prompt tuning (*e.g.*, CoOp [36], VPT [11]) has been adopted as a preferred technique that appends lightweight learnable tokens (or *prompts*) to the input tokens of the CLIP image and/or text encoder. Each client trains the prompts on the local data, while keeping the CLIP backbone frozen, and then dispatches these prompts to the server, where they are updated through aggregation and sent back. This speeds up client training as each client needs access

to only a few samples per class (*e.g.*, 16/class [36]), and reduces communication overhead as only the lightweight prompts (*e.g.*, 0.37% of network parameters) need to be communicated, instead of all the parameters.

Although prompt tuning has proven to be effective in the FL setting, it suffers from reduced generalization to unseen classes and unseen domains – a typical heterogeneous setup native to non independent and identically distributed (non-i.i.d.) FL [21] – due to severe overfitting on the local training data [30]. It is attributed to the fact that prompts learn a *static context* from a set of seen classes, which being fixed once learned, fail to capture the generalizable elements spread across clients. To circumvent the issue of reduced generalization, recent CLIP-based FL methods condition the prompts either on the class information, as in FedTPG [21] (see Fig. 1a), or on visual information, as in FedCoCoOp [35]. This auxiliary conditioning on the input, or *contextual information*, makes the prompts *dynamic* and thus offers a better generalization to unseen distributions.

In this work, we argue that the FL setup being highly heterogeneous, where clients have data from disjoint classes and domains [21], it is not sufficient to solely exploit either of the two modalities for contextual information. Rather, FL demands more comprehensive contextual information that encompasses all the available modalities to condition input prompts. To this end, we propose **Fed**erated **M**ultimodal **V**isual **P**rompt Tuning (FedMVP) that conditions the prompts on *multimodal contextual information*, which comprise: (i) *visual* features of the input image itself, and (ii) *textual* features of attributes of a class (*e.g.*, "two legs", "beak", etc. for the class "hen"). The **key intuition** is that (ii) promotes learning transferable prompts for unseen classes that share attributes similar to seen classes (*e.g.*, unseen class "seagull" will share some attributes with the seen class "hen"), and (i) promotes transferability for unseen classes and domains where attributes cannot be described via text (*e.g.*, texture or abstract concepts). In a nutshell, the dual conditioning offers richer contextual information to the prompts, which helps in generalizing to classes or domains that share similar properties.

In detail, the proposed FedMVP framework contains a **PromptFormer** network (see Fig. 1b) that synergistically merges the information from visual and textual modalities with cross-attention mechanism [26] to generate the prompts. The resulting *multimodal prompts* are then injected at the visual space of the frozen vision encoder of CLIP, unlike the existing methods [7, 21, 30] that inject prompts through text encoder. The key differences between our FedMVP and existing CLIP-based FL methods have been summarized in Tab. 1. The PromptFormer is trained in an end-to-end manner with the CLIP loss and a consistency loss. Once trained, each participating client shares the lightweight PromptFormer network parameters with the server for aggregation through FedAvg [16], and thus enabling knowledge sharing among clients. During inference, the cross-attention learned by the PromptFormer enables the model to dynamically adjust the prompts, enhancing generalizability to unseen classes and domains. We evaluated FedMVP on three generalization settings encompassing 20 datasets that measure generalization to unseen classes, domains, and combinations of them. Experiments demonstrate that FedMVP consistently outperforms the state-of-the-art CLIP-based FL methods by $1.57\% - 2.26\%$.

Table 1. **Summary of federated prompt-tuning methods**. (✗) indicates the learnable prompts are initialized randomly, (✓) signifies learnable prompts are conditioned on contextual information (*e.g.*, textual or visual inputs). *Prompting* column denotes the CLIP encoder wherein learnable prompts are given as input (*e.g.*, "Textual" means the text encoder of CLIP takes prompts as input).

Method	Contextual information		Prompting
	Textual	Visual	
FedKgCoOp [30]	✗	✗	Textual
FedVPT [11]	✗	✗	Visual
FedTPG [21]	✓	✗	Textual
FedCoCoOp [35]	✗	✓	Textual
FedMaPLe [13]	✗	✗	Multimodal
FedMVP (Ours)	✓	✓	Visual

In summary, our main **contributions** are: (**i**) We highlight the importance of leveraging **multimodal contextual information** in prompt tuning based FL setup in order to alleviate the issue of reduced generalizability to unseen data. (**ii**) We propose **FedMVP** that synergistically combines the visual features of the image and textual features of the attributes of classes to generate **multimodal** prompts. This provides richer contextual information to the model and higher transferability to unseen classes and domains.

2. Related Work

Prompt tuning of VLMs. VLMs, such as CLIP [22] and ALIGN [10], leverage large image-text datasets to enable multimodal understanding, performing well in zero-shot classification tasks. However, fully fine-tuning VLMs for downstream tasks without sacrificing their ability to generalize to unseen concepts remains a challenge [27].

Prompt tuning approaches [11, 13, 36] have been proposed for parameter efficient and effective adaptation of CLIP by introducing a set of continuous learnable tokens (or prompts) onto the frozen CLIP backbone. These approaches can be broadly categorized into three categories depending on the manner in which CLIP is prompted: (i) *textual prompting* that inject prompts in the input space of the text encoder [30, 36], (ii) *visual prompting* that injects prompts in the input space of the vision encoder [11], and (iii) *multimodal prompting* that injects prompts in both the text and vision encoder [13, 28].

In particular, MaPLe [13], which injects prompts to both

the text and vision encoders, is not suitable for memory and compute efficient FL since it requires parts of both the encoders to be unlocked during tuning, risking overfitting. Moreover, it does not allow for image-conditioned dynamic prompting, which has shown to be effective for generalization [35]. In contrast, our proposed FedMVP enables multimodal prompting by involving only the vision encoder and supports instance-specific prompting.

Federated prompt tuning of VLMs. Prompt tuning has emerged as an effective technique for FL of VLMs since it allows clients to collaboratively train on a heterogeneously distributed dataset and adapt faster with fewer samples. The fact that the prompts are lightweight lends well to the FL setup as it reduces communication overhead. Due to the benefits offered by prompt tuning, several works [2, 7, 21, 24] have specifically adapted prompt tuning approaches to the FL setup. For example, PromptFL [7] extended CoOp to FL via sharing prompts with the server, enhancing the generalization performance of client-server interactions in vision tasks. FedCLIP [15] learns a lightweight adapter in each client, similar to the CLIP-Adapter [5], and shares parameters with the server. FedOTP [14] introduced global and local prompts, optimizing shared parameters through optimal transport, while DiPrompt [2] studied multi-domain learning by allowing multiple domains per client.

Unlike existing methods, a couple of recent works [21, 28] argue that prompt tuning is prone to overfitting to the seen classes due to learning static context. The closest work to ours is FedTPG [21], where a prompt generation network is learned across multiple clients. Additionally, this network leverages textual inputs, enhancing its ability to generalize to unseen classes. Our work significantly differs from previous methods in two key ways: (i) we exploit both attribute information derived from an LLM and image features to condition the prompts, and (ii) we leverage multimodal prompts from the proposed PromptFormer for visual prompt tuning, instead of textual prompt tuning.

Classification via textual descriptions. Zero-shot accuracy of CLIP on downstream tasks is sensitive to the quality of text prompts [22]. In detail, some works rely on simple hand-crafted prompts (*e.g.*, "a photo of a `[CLS]`") [22] or others augment the hand-crafted prompts with richer attributes obtained from LLMs [17, 20] (*e.g.*, a "hen" has *two legs*, *white feathers*, etc.). There also exists a family of methods, such as LaCLIP [4], LaBo [29] and VFC [18], that have proposed to further refine CLIP weights with enriched captions from LLMs for improved performance. Recognizing the importance of class attributes for improved generalization to unseen concepts and domains – an important desideratum in the FL set up – we, for the first time in FL, introduce LLM generated class attributes for enriching the image-conditioned multimodal prompts through the PromptFormer network.

3. Methods

We introduce FL setup in Sec. 3.1, preliminaries on prompt tuning in Sec. 3.2 and our proposed solution in Sec. 3.3.

3.1. Problem formulation

We consider a FL setup in which multiple remote clients train models on local data, and a central server aggregates models from the clients. Formally, let each remote client i has access to a dataset $\mathcal{D}_i = \{\mathbf{x}_{i,j}, y_{i,j}\}_{j=1}^{n_i}$, where $\mathbf{x}_{i,j}$ is the j^{th} image from the i^{th} client, $y_{i,j}$ is the corresponding class label and n_i is the number of samples in client i. Let the set of text class names in client i be denoted as $\mathcal{C}_i = \{c_{i,k}\}_{k=1}^{K_i}$, where K_i is the total number of classes. As in FedTPG [21], we follow a non-IID FL setup, where the samples in each client are drawn from a disjoint set of classes, *i.e.*, $\mathcal{C}_i \cap \mathcal{C}_l = \emptyset$. As an example, the training data on each mobile device usually depend on the user, which may not be representative of the general population [16].

In particular, we investigate two kinds of non-IID FL setups, where, in addition to the in-distribution performance, we also evaluate the generalization ability of the aggregated server model on (i) unseen *classes*, and (ii) unseen *domains* or *datasets*. The need for generalization to unseen classes and domains makes the non-IID FL setup more challenging than the traditional FL setting [12].

3.2. Preliminaries on Prompt Tuning

Textual prompt tuning (TPT), *i.e.*, CoOp [36], has emerged as an efficient technique in CLIP-based FL that keeps the CLIP backbone frozen and locally learns lightweight prompt vectors on each client. The locally learned prompts are then communicated to the server, as in PromptFL [7]. Specifically, the hand-crafted text prompts (*e.g.*, *"a photo of a* `[CLS]`") input to the text encoder $\mathcal{E}_t$ is augmented with the learnable soft prompts $v_1, v_2, \ldots, v_m$. However, as shown in [35], TPT can often lead to overfitting and reduced generalization due to the static context learned on the seen classes. This can be exacerbated in the non-IID FL setup due to the high data heterogeneity among the clients [21].

Alternatively, visual prompt tuning (VPT) [11] adapts pre-trained vision encoder $\mathcal{E}_v$ by injecting small number of visual learnable parameters into ViT's input space. More formally, let an input image $\mathbf{x}$ be divided in b patches, yielding a collection of patch embeddings (augmented with positional encodings) at the input of the vision encoder as $\mathbf{E} \in \mathbb{R}^{b \times d_v}$, where d_v is the embedding dimension. A `[CLS]` token $\mathbf{z}$ is concatenated with $\mathbf{E}$, encapsulating the initial input representation for the ViT layers of $\mathcal{E}_v$. VPT further augments the input by concatenating a set of randomly initialized learnable visual prompts $\mathbf{P} = \{p_1, p_2, \ldots, p_m\}$, where m is the prompt length. The first layer input of $\mathcal{E}_v$ is then redefined as $\mathbf{I} = [\mathbf{z}; \mathbf{E}; \mathbf{P}] \in \mathbb{R}^{(1+b+m) \times d_v}$, where

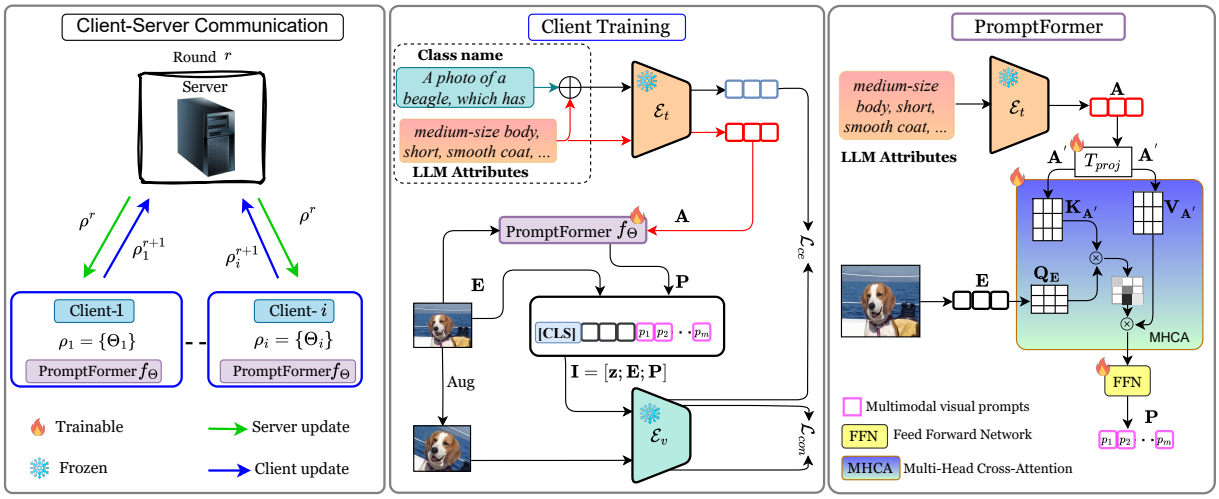


Figure 2. **Overview of Federated Multimodal Visual Prompt Tuning (FedMVP)**. Each client i trains the parameters of the PromptFormer network f_Θ and multi-head self-attention on its local data. At each round, the server aggregates the trainable parameters from the clients and sends the updated parameters back to the clients. The PromptFormer uses cross-attention mechanism to fuse the visual and textual attribute feature to generate multimodal visual prompts $\mathbf{P}$, which are then used for visual prompt tuning.

";" means concatenation. Although VPT enables more nuanced understanding of the visual space, it suffers from the same issues as TPT, *i.e.*, limited generalization (see Sec. 4).

3.3. Federated Multimodal Visual Prompt Tuning

Our proposed method FedMVP deviates from the TPT and VPT paradigms and incorporates **multimodal contextual information** from both the visual and textual modalities to improve generalization in the FL setup. As shown in Fig. 2, at the core of our novel architecture is a prompt generator, called **PromptFormer**, which conditions the visual prompts $\mathbf{P}$ on the task-related information: input image and text attributes corresponding to the class names, rather than randomly initializing them. This makes $\mathbf{P}$ *context-aware*, generalizing better to unseen classes and domains. Once trained on each client, only the parameters of the lightweight PromptFormer network are communicated to the server, thus, keeping the communication overhead low. Next, we describe PromptFormer (Sec. 3.3.1), training objectives (Sec. 3.3.2) and how we conduct client-side training and server-side aggregation (Sec. 3.3.3).

3.3.1. Visual Prompt Generation with PromptFormer

We introduce the PromptFormer network f_Θ, parameterized by Θ, designed to generate contextually rich visual tokens (or prompts) for each client i by leveraging both attribute features and visual patch embeddings (see Fig. 2). Concretely, we use the CLIP text encoder $\mathcal{E}_t$ to extract embeddings $\mathbf{A}_i = \{\mathcal{E}_t(\texttt{LLM}(c_k))\}_{k=1}^K$ of the text attributes[1] $\mathcal{A}_i$ corresponding to the text class names $\mathcal{C}_i$, generated using a large lanuage model (LLM) [1] (see Sec. A.1 of Supp. Mat. for Attribute Generation process). Going forward, we drop the client index i for notational clarity. The PromptFormer takes as input b visual patch embeddings $\mathbf{E}$, corresponding to an image $\boldsymbol{x}$, and the text embeddings $\mathbf{A}$ to generate a set of enriched m multimodal visual prompts $\mathbf{P} \in \mathrm{R}^{m \times d_v}$:

$$\mathbf{P} = \{p_j\}_{j=1}^m = f_\Theta(\mathbf{A}, \mathbf{E}). \tag{1}$$

Consequently, the input to the first layer of the visual encoder $\mathcal{E}_v$ is redefined as $[\mathbf{z}; \mathbf{E}; \mathbf{P}] \in \mathbb{R}^{(1+b+m)\times d_v}$. Intuitively, multimodal visual prompts carry richer contextual information, and helps the network in generalizing to unseen classes and domains that share commonalities.

Architecture of PromptFormer. The PromptFormer architecture comprises: lightweight multi-head cross-attention (MHCA) modules, feed-forward network (FFN) and a linear projection layer; transforming multimodal features into comprehensive visual prompts. In detail, the visual patch embedding $\mathbf{E}$ interacts with the linearly projected attribute features $\mathbf{A}' = T_{\text{proj}}(\mathbf{A})$ to construct multimodal visual prompts as:

$$\begin{aligned} &\mathbf{P}(\mathbf{A}', \mathbf{E}) = \text{FFN}(\text{CrossAttention}(\mathbf{Q}_\mathbf{E}, \mathbf{K}_{\mathbf{A}'}, \mathbf{V}_{\mathbf{A}'})); \\ &\text{with } \mathbf{Q}_\mathbf{E} = \mathbf{E}W_\mathbf{Q}, \mathbf{K}_{\mathbf{A}'} = \mathbf{A}'W_\mathbf{K}, \mathbf{V}_{\mathbf{A}'} = \mathbf{A}'W_\mathbf{V}, \end{aligned} \tag{2}$$

where, $\mathbf{E}$ are transformed into query vectors $\mathbf{Q}_\mathbf{E}$ using query matrices $W_\mathbf{Q}$; and the projected attribute features $\mathbf{A}'$ are transformed into key and value vectors $\mathbf{K}_{\mathbf{A}'}$ and $\mathbf{V}_{\mathbf{A}'}$, using key and value matrices $W_\mathbf{K}$ and $W_\mathbf{V}$, respectively.

Intuitively, through the cross-attention mechanism, visual features learn to attend to the relevant parts of the corresponding attribute features. For example, the patches depicting the *legs* of the class "dog" will learn to attend to

[1]For example, for the class name "giraffe" the LLM outputs "Exceptionally long neck, unique coat pattern with irregular brown patches, ..."

the attributes *four legged*. When an unseen class (*e.g.*, "giraffe") comprises an animal with *four legs*, the final prompts $\mathbf{P}$ will contain this relevant information derived from the attributes *four legged*. Additional details of PromptFormer architecture are reported in Sec. A.3 of the Supp. Mat.

3.3.2. Training of FedMVP

After the visual prompts are obtained with PromptFormer as per Eq. 2, we concatenate the [CLS] token, the patch embeddings, and the visual prompts to get $\mathbf{I} = [\mathbf{z}; \mathbf{E}; \mathbf{P}]$ following Sec. 3.2, and feed it to the visual encoder $\mathcal{E}_v$.

The parameters of PromptFormer $\rho = \{\Theta\}$ are the only trainable parameters of FedMVP, and the rest of the weights are kept frozen. To train the model, we employ the CLIP cross-entropy loss [22], formulated as:

$$\mathcal{L}_{\text{ce}} = - \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{D}} y \log p(y|\mathbf{I}), \tag{3}$$

where y denotes the ground truth label, and the prediction probability of CLIP between the visual features $\mathbf{v} = \mathcal{E}_v(\mathbf{I})$ and text features $\mathbf{t}_k = \mathcal{E}_t([\text{A photo of [CLASS]}; \texttt{LLM}(c_k)])$ for class k is computed as:

$$p(y=k|\mathbf{I}) = \frac{\exp(\cos(\mathbf{v}, \mathbf{t}_k))/\tau)}{\sum_{k'=1}^{K} \exp(\cos(\mathbf{v}, \mathbf{t}_{k'}))/\tau)}, \tag{4}$$

where $\cos(\cdot,\cdot)$ as a similarity function, τ as the temperature parameter, and $[;]$ denotes concatenation. Details on how $\mathbf{t}_k$ is computed are provided in Sec. A.1 of the Supp. Mat.

In addition to the $\mathcal{L}_{\text{ce}}$, we employ a regularization constraint to ensure consistency between two augmented views of an image. We define the consistency loss $\mathcal{L}_{\text{con}}$ as:

$$\mathcal{L}_{\text{con}} = 1 - \cos(\mathcal{E}_v(\mathbf{I}), \mathcal{E}_v(\mathbf{x}')), \tag{5}$$

where $\mathbf{x}'$ is the augmented version of $\mathbf{x}$. Finally, the total loss function is given as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ce}} + \alpha \cdot \mathcal{L}_{\text{con}}, \tag{6}$$

where α is a balancing coefficient for the regularization term, ensuring both alignment and consistency.

3.3.3. Client-side training and server-side aggregation

In our approach, multiple remote clients participate in distinct image classification tasks, collaboratively training the PromptFormer network, parameterized by $\rho = \{\Theta\}$. In each communication round, the client transmits ρ to the server, which is aggregated at the server. Through training on a diverse set of classification tasks distributed across clients, and aggregating the parameters at the server, the resulting model achieves generalization capability to classes and domains unseen during training. Pseudo-code of FedMVP can be found in Sec. A.2 of the Supp. Mat.

Initially, the server randomly initializes the parameters ρ^0 and in each communication round only a random subset of remote clients $\mathcal{S}^r$ will be chosen by the server for sharing the updated parameters ρ. Each communication round r involves the following steps:
(i) **Model initialization:** Each client i in S_r receives the latest model parameters ρ^r to configure its local model.
(ii) **Local optimization:** Following Eq. (6), client i optimizes its local parameters ρ^r using an SGD optimizer over J iterations with learning rate η^r. Updated parameters:

$$\rho_i^{r+1} = \texttt{SGD}_J(\eta^r, \rho^r, \mathcal{C}_i, \mathcal{D}_i). \tag{7}$$

(iii) **Parameter aggregation:** After local training, each client i in S_r transmits its updated parameters ρ_i^{r+1} back to the server, where they are aggregated [16] as:

$$\rho^{r+1} = \frac{1}{|S_r|} \sum_{i \in S_r} \rho_i^{r+1}. \tag{8}$$

The above steps are repeated for R communication rounds, after which the server converges on the final model parameters ρ^R. This iterative process facilitates collaborative and distributed optimization of ρ across client data.
Lightweight fine-tuning in clients. As the FL setup exhibits high data heterogeneity, different clients possess varying numbers of samples. Thus, treating every client as equal and repeatedly fine-tuning the parameters ρ can lead to severe overfitting in some clients, especially those with very limited training data. To prevent this, we propose to inject low-rank decomposition matrices, or LoRA [9], to the PromptFormer module, and fine-tune these LoRA matrices if training loss $\mathcal{L}_{\text{total}}$ in a client starts below a certain loss value σ, which we set to 0.5. In particular, only the LoRA weights corresponding to the query, key and value matrices of the attention layers are trained, while freezing the parameters ρ. In that case, the client shares only the LoRA matrices with the server. This not only helps in boosting generalization, but also brings down the communication overhead of clients by around $\times 267$.

4. Experiments

We evaluate our proposed FedMVP in three distinct settings that are aimed to gauge the (i) generalization to unseen *categories* (or Base-to-New Generalization) in Sec. 4.1, (ii) generalization to unseen *domains* (or Domain Generalization) in Sec. 4.2, and (iii) generalization to unseen *datasets* (or Cross-Dataset Generalization) in Sec. 4.3. Finally, we report the results of our ablation study in Sec. 4.4 that are aimed to assess impact of the design choices made in FedMVP. Due to the lack of space, we report additional experimental results in Sec. B of the Supp. Mat.
Benchmarks and datasets. Following FedTPG [21], we have reported results on the Base-to-New Generalization setting that encompasses nine medium-sized image recognition datasets. For the Domain Generalization setting we

have used the ImageNet [3] and DomainBed [6] benchmarks. In addition, we have also included EuroSAT [8] for the Cross-Dataset Generalization setting, alongside all nine datasets of the Base-to-New Generalization setting. In total, we have used 20 datasets for a comprehensive experimental evaluation. Detailed dataset descriptions and statistics can be found in Sec. B.1 of the Supp. Mat.

Baselines. We have compared our method with the state-of-the-art prompt tuning approaches that use the CLIP backbone [22]: KgCoOp [30], CoCoOp [35], VPT [11], and MaPLe [13], which were originally designed for offline and non-FL setup, but adapted to the federated setting, as in FedTPG [21]. Furthermore, we have compared with CLIP-based methods specifically designed to work in the FL setting: FedTPG [21], FedOTP [14], PromptFL [7], and FedCLIP [15]. We also included zero-shot CLIP (ZS-CLIP).

Implementation details. We have employed the frozen ViT-B/16 CLIP backbone as the vision encoder, and GPT-4o [1] as the LLM in our experiments. Each MHCA module of f_Θ, consists of a 4-head cross-attention layers, complemented by layer normalization and each FFN comprising of a two-layer bottleneck structure (Linear-GeLU-Linear). The query prompt $\mathbf{Q}$ has a length $m = 4$ and matches the patch embedding dimension $d_v = 768$. The projection layer T_{proj} maps the textual dimension of 512 to the patch embedding dimension via a linear transformation. In Eq. (6), we keep $\alpha = 10$. We have set the batch size to 128, number of shots per class to 8, and the learning rate equal to 0.003 with a decay rate of 1e-5 and momentum of 0.9.

4.1. Base-to-New Generalization

Experimental setup. To recap, the base-to-new generalization setting encompasses nine different datasets. The setting is characterized by splitting the classes from each dataset into two groups – base and new. The samples of the base classes are used for training, and the samples of the new classes are held out to test the generalization performance [21]. Furthermore, the samples of bases classes from all nine datasets are distributed to the clients, such that each client does not see samples from more than 20 disjoint classes (referred to as *local* classes). We report the model's classification accuracy on: (i) the client's local classes, (ii) all the base classes (spread across clients), and (iii) the held-out new classes that were never seen by any client. Note that (ii) and (iii) is computed using the aggregated model on the server, and (i) is computed using the model on the client (which is the same server model after convergence).

Main results. In Table 2 we report the performance in the base-to-new generalization setting, summarized across nine datasets. The results show that the textual and multimodal prompting methods, such as PromptFL, FedKgCoOp, and FedMaPLe, excel on the seen local and base class accuracy, but have at-par or worse generalization on the new classes than the ZS-CLIP model. This hints at the potential overfitting caused by learning static context from the seen classes. In contrast, our proposed FedMVP that considers multimodal contextual information and dynamic conditioning demonstrates superior performance in both the seen and unseen classes, achieving the highest HM accuracy. While FedTPG also generates prompts based on task-related input, it does not outperform our FedMVP due to lack of comprehensive visual and attribute information in the prompts. In summary, FedMVP surpasses the state-of-the-art methods by a notable margin of $+1.57\%$ in the base-to-new generalization setting. Detailed breakdown for each of the nine datasets is reported in Tab. A.4 in the Supp. Mat.

Table 2. **Base-to-New Generalization setting.** Accuracies (%) on client's local (seen) classes, base (seen) classes and new (unseen) classes. Harmonic mean (HM) is computed with Base and New class accuracies. Best number in each column is in bold and second-best underlined. Improvement row reports the absolute difference between the best and second-best numbers.

Methods	Local	Base	New	HM
ZS-CLIP [22]	76.72	70.51	75.78	74.24
FedOTP [14]	74.82	65.22	57.04	64.89
FedCoCoOp [35]	81.46	73.76	66.00	73.20
FedVPT [11]	76.29	70.43	74.89	73.79
FedCLIP [15]	76.87	71.04	75.06	74.24
FedMaPLe [13]	81.63	74.44	70.62	75.29
FedKgCoOp [30]	78.38	72.18	75.87	75.39
PromptFL [7]	<u>81.75</u>	<u>74.47</u>	71.70	75.74
FedTPG [21]	80.75	73.68	<u>76.02</u>	<u>76.70</u>
FedMVP (Ours)	**81.89**	**75.37**	**77.82**	**78.27**
Improvement	**+0.14**	**+0.90**	**+1.79**	**+1.57**

4.2. Domain Generalization

Experimental setup. We conduct experiments under two distinct domain generalization (DG) setups: Multi-source Single-target (MSST) and Single-source Multi-target (SSMT). In the MSST DG setting for a given benchmark, we follow a leave-one-domain-out protocol for evaluation, where we hold out one domain for measuring generalization and use the rest of the domains for training the clients. Training classes are split among clients under the constraint that each client can see images from a single domain. This setting mirrors a practical scenario where images taken with a smartphone uniquely reflect the taste of the user or socio-economic biases. In a similar manner, in the SSMT DG setting, we use one domain for training the clients in a distributed manner, while the rest of the domains are used for evaluation. More details on this experimental setup can be found in Sec. B.4 of the Supp. Mat.

Main results. In Table 3, we report the performance of FedMVP and its competitors in the MSST and SSMT DG settings. In particular, SSMT DG is a more challenging setting than the MSST DG setting, as the clients are trained only on a single domain, and expected to generalize to multiple unseen domains. We observe that ZS-CLIP is a strong base-

Table 3. **Domain Generalization setting on the DomainBed [6] benchmark.** Accuracies (%) on the unseen target domain(s) for the Multi-source Single-target (MSST) and Single-source Multi-target (SSMT) settings. Best number in each column is in bold and second-best underlined. Improvement row reports the absolute difference between the best and second-best numbers.

Method	Multi-source Single-target						Single-source Multi-target					
	PACS	OfficeHome	VLCS	Terra Inc.	DomainNet	Average	PACS	OfficeHome	VLCS	Terra Inc.	DomainNet	Average
ZS-CLIP [22]	96.16	81.49	83.29	33.98	57.13	70.41	96.16	81.49	83.29	33.98	57.13	70.41
FedOTP [14]	90.71	76.42	67.41	13.24	49.67	59.49	91.17	74.53	65.16	15.11	39.37	57.07
FedCoCoOp [35]	85.06	81.42	61.73	23.68	57.08	61.79	84.56	50.87	61.25	23.07	58.28	55.61
FedTPG [21]	90.99	82.78	69.77	26.79	56.82	65.43	90.71	82.60	67.63	22.51	58.34	64.36
PromptFL [7]	95.37	81.74	74.87	25.02	56.87	66.77	95.83	81.72	66.97	24.17	58.27	65.39
FedKgCoOp [30]	95.42	81.82	74.90	25.03	56.92	66.82	95.81	81.70	67.49	24.20	58.30	65.50
FedMaPLe [13]	94.51	82.03	71.79	36.30	58.88	68.70	94.25	81.48	72.08	33.59	57.62	67.80
FedVPT [11]	95.36	81.76	83.19	33.62	55.98	69.98	94.79	80.99	82.44	32.23	55.66	69.22
FedCLIP [15]	96.29	81.74	82.70	36.58	57.85	71.03	96.29	81.62	81.81	36.44	57.42	70.72
FedMVP (Ours)	**97.28**	**84.15**	**85.12**	**37.36**	**61.17**	**73.02**	**97.17**	**83.89**	**84.61**	**36.92**	**60.56**	**72.63**
Improvement	**+0.99**	**+1.37**	**+1.83**	**+0.78**	**+2.29**	**+1.99**	**+0.88**	**+1.29**	**+1.32**	**+0.48**	**+2.22**	**+1.91**

Table 4. **Domain Generalization setting on the ImageNet benchmark.** Accuracies (%) on the seen source domain ImageNet (IN) and unseen target domains: ImageNet-V2 (INV2), ImageNet-Sketch (IN-S), ImageNet-A (IN-A) and ImageNet-R (IN-R). Best number in each column is in bold and second-best underlined. Improvement row reports the absolute difference between the best and second-best numbers.

Method	Source	Target				
	IN	INV2	IN-S	IN-A	IN-R	Average
ZS-CLIP [22]	66.75	60.79	46.12	47.79	74.00	57.18
FedOTP [14]	51.68	45.53	34.73	16.64	63.98	40.22
FedMaPLe [13]	66.96	60.65	44.69	46.24	74.62	56.55
FedVPT [11]	66.92	60.34	46.43	48.03	74.56	57.34
PromptFL [7]	67.80	61.59	45.61	48.78	74.49	57.62
FedCLIP [15]	67.26	61.35	46.66	48.24	74.36	57.65
FedKgCoOp [30]	67.53	61.60	46.69	48.37	74.71	57.84
FedCoCoOp [35]	68.51	62.29	46.90	50.33	76.49	59.00
FedTPG [21]	69.51	62.90	47.65	49.97	76.35	59.22
FedMVP (Ours)	**70.87**	**63.72**	**50.93**	**51.76**	**77.23**	**60.91**
Improvement	**+1.36**	**+0.82**	**+3.28**	**+1.43**	**+0.74**	**+1.69**

line that outperforms all textual and multimodal prompting methods, with the exception of FedCLIP and our proposed FedMVP. This suggests that suboptimal prompt tuning can lead to overfitting on the source training domains, thus hurting generalization. Our FedMVP owing to the use of contextual information learns more transferable representations that generalize better to unseen target domains, thus outperforming all competitor methods in most of the benchmarks. More detailed results in Sec. B.4 of the Supp. Mat.

In Table. 4 we report the results in the SSMT DG setting on ImageNet benchmark. We observe that FedTPG and FedCoCoOp come close to our FedMVP in terms of average performance, with FedMVP outperforming all baselines by $+1.69\%$. In particular, we believe FedMVP improves the performance on ImageNet-Sketch (IN-S) by a large margin of $+3.28\%$ due to the use of attribute information, which remains unchanged between real and sketch images.

4.3. Cross-dataset Generalization

Experimental setup. In the Cross-Dataset Generalization setting the classes from ImageNet are split among the clients for training, in a disjoint manner, and the resulting network is evaluated on 10 held out datasets not seen during training. In other words, it is a combination of base-to-new generalization and DG settings, since it requires generalization to both unseen class and data distributions.

Main results. In Table. 5 we report the results of the Cross-Dataset Generalization setting. This setting is considerably challenging, as the model should generalize to different datasets. We find that the results are consistent with the previous findings, where our proposed FedMVP generalizes better to unseen classes and data distributions, outperforming all the text-based and multimodal-based prompting techniques by $+2.26\%$ on average. However, with FedMVP we notice a drop in performance with respect to the ZS-CLIP baseline for some fine-grained datasets, such as OxfordPets and StanfordCars, which is potentially due to overlapping attributes for visually similar categories.

4.4. Ablation Study

Impact of proposed components in FedMVP. In Tab. 6 we ablate each component of FedMVP on the base-to-new and MSST DG settings. We observe that each component contributes positively to the final performance. First, PromptFormer f_Θ plays a major role in combining textual information with visual features, thus improving semantic alignment between the two modalities. Second, adding $\mathcal{L}_{\text{con}}$ (as in Eq. (5)) improves the performance by a small margin. Finally, the impact of LoRA can be observed in the penultimate row, where we fine-tune all the parameters of FedMVP. It shows the dangers of overfitting in the FL setup, as some clients may have limited data.

Participation Rate of the Clients. In Fig. 3 we vary the participation rate from 10% to 100% and show how the participation rate of the clients impacts performance on both seen and unseen classes and domains. We observe that FedMVP always exhibits a gentle, yet positive slope in both settings, indicating that continuous parameter aggregation improves generalization. In contrast, for some competitor methods, such as PromptFL, increased participation leads to saturated performance or negative slope in the plot.

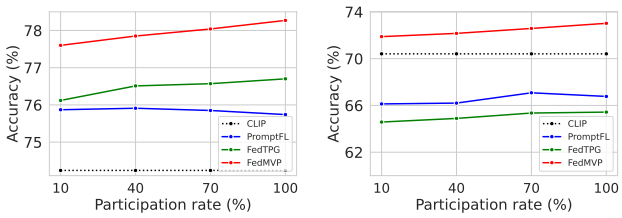
Computational analysis. In Fig. 4 we plot the accuracy vs. communication rounds for FedMVP and competitors in base-to-new and MSST settings. While FedMVP trans-

Table 5. **Cross-Dataset Generalization setting.** Accuracies (%) on the seen classes of the source dataset, *i.e.*, ImageNet, and the unseen classes of the target datasets. Best number in each column is in bold and second-best underlined. Improvement row reports the absolute difference between the best and second-best numbers.

Method	Source	Target										
	ImageNet	Caltech101	Flowers102	FGVCAircraft	UCF101	OxfordPets	Food101	DTD	StanfordCars	SUN397	EuroSAT	Average
ZS-CLIP [22]	66.75	92.90	71.29	24.72	66.75	89.15	86.09	44.33	65.29	62.59	47.68	65.08
FedOTP [14]	51.68	88.60	48.36	10.95	46.10	73.43	61.01	41.55	29.61	48.34	50.26	49.82
FedMaPLe [13]	66.96	92.49	68.25	23.52	60.32	89.67	83.52	44.68	60.16	61.85	45.38	62.98
PromptFL [7]	67.80	91.87	68.13	21.44	64.13	88.70	85.85	42.43	63.59	62.77	43.26	63.22
FedCLIP [15]	67.26	93.39	67.19	24.03	65.64	88.28	85.14	44.44	65.50	63.42	41.67	63.87
FedKgCoOp [30]	67.53	93.63	69.31	23.06	64.46	88.55	85.37	44.74	64.99	63.85	43.29	64.13
FedVPT [11]	68.56	91.42	69.24	21.09	65.70	87.79	85.45	44.57	65.00	64.68	47.82	64.28
FedCoCoOp [35]	68.51	94.11	66.34	20.79	62.75	89.04	85.40	43.20	63.98	64.02	**55.40**	64.50
FedTPG [21]	69.51	93.75	70.04	23.22	64.72	**90.60**	85.91	46.25	63.98	66.78	47.86	65.31
FedMVP (Ours)	**70.87**	**95.37**	**72.80**	**25.94**	**70.58**	89.27	**87.06**	**49.78**	**65.83**	**68.19**	50.84	**67.57**
Improvement	+1.36	+1.62	+1.51	+1.22	+3.83	-1.33	+0.97	+3.53	+0.33	+1.41	-4.56	+2.26

Table 6. **Ablating FedMVP components.** We report numbers for the Base-to-New and MMST DG (on DomainBed) settings.

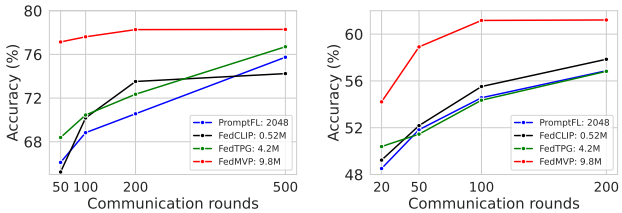
Components	Base-to-New	MSST DG
ZS-CLIP	74.24	70.41
f_Θ *only*	75.94	71.85
f_Θ + $\mathcal{L}_{\text{con}}$	76.27	72.14
w/o LoRA	77.41	72.58
FedMVP (Full)	**78.27**	**73.02**



(a) Base-to-New (b) MSST DG

Figure 3. **Sensitivity to participation rate** of the clients, on (a) Base-to-New and (b) MSST DG (DomainBed) settings.

mits 2× more parameters than FedTPG, it converges nearly 10× faster, requiring fewer rounds. In particular, as training progresses, only the LoRA parameters in PromptFormer are fine-tuned in a client, thus reducing the trainable parameter count by 267×. Overall, FedMVP has lower total computational cost with respect to other methods when normalized by the number of communication rounds.

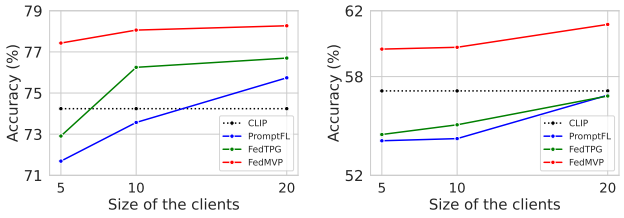


(a) Base-to-New (b) MSST DG

Figure 4. **Computational analysis of the methods** on (a) Base-to-New and (b) MSST DG (DomainNet) setting.

Size of the clients. In Fig. 5, we vary the number of training classes in each client $K = \{5, 10, 20\}$ while keeping the size of the dataset fixed. In other words, decreasing the number of classes in each client results in an increase in the total number of clients participating in the federated setup. We observe that for all the baselines a higher heterogeneity (or fewer classes per client) leads to lower performance and vice versa. Nevertheless, FedMVP still exhibits a much superior performance when compared to the baselines, outperforming them while having 4x more clients.

Additional analyses relating to number of shots, choice of LLMs, hyperparameters, and performance in non-federated setting are reported in Sec. B.5 of the Supp. Mat.



(a) Base-to-New (b) MSST DG

Figure 5. **Sensitivity to number of classes per client** on (a) Base-to-New and (b) MSST DG (DomainNet) setting.

5. Conclusion

In this work, we proposed FedMVP, a novel federated prompt tuning framework for CLIP, to address the challenge of reduced generalization to unseen domains and classes. FedMVP combines visual embeddings with textual attribute features to generate multimodal visual prompts, which are tuned using visual prompt tuning. Our findings revealed that the multimodal contextual information in the prompts enhances the generalizability to unseen classes and domains. FedMVP outperforms all its competitors by a notable margin of $+1.57\% - 2.26\%$ across benchmarks.

Acknowledgements. This work was supported by: JPNP23019, subsidized by the New Energy and Industrial Technology Development Organization (NEDO); SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU; EU project ELIAS (no: 101120237), ANT (no: 101169439), ELLIOT (no: 101214398) and Ministero delle Imprese e del Made in Italy (IPCEI Cloud DM 27 giugno 2022 - IPCEI-CL-0000007).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 4, 6

[2] Sikai Bai, Jie Zhang, Song Guo, Shuaicheng Li, Jingcai Guo, Jun Hou, Tao Han, and Xiaocheng Lu. Diprompt: Disentangled prompt tuning for multiple latent domain generalization in federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27284–27293, 2024. 3

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 6

[4] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[5] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595, 2024. 3

[6] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. 6, 7

[7] Tao Guo, Song Guo, Junxiao Wang, Xueyang Tang, and Wenchao Xu. Promptfl: Let federated participants cooperatively learn prompts instead of models-federated learning in age of foundation model. *IEEE Transactions on Mobile Computing*, 2023. 2, 3, 6, 7, 8

[8] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 6

[9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, page 3, 2022. 5

[10] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 2

[11] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 1, 2, 3, 6, 7, 8

[12] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2): 1–210, 2021. 3

[13] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19113–19122, 2023. 2, 6, 7, 8

[14] Hongxia Li, Wei Huang, Jingya Wang, and Ye Shi. Global and local prompts cooperation via optimal transport for federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12151–12161, 2024. 3, 6, 7, 8

[15] Wang Lu, HU Xixu, Jindong Wang, and Xing Xie. Fedclip: Fast generalization and personalization for clip in federated learning. In *ICLR Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*, 2023. 3, 6, 7, 8

[16] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017. 1, 2, 3, 5

[17] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *International Conference on Learning Representations*, 2023. 3

[18] Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. Verbs in action: Improving verb understanding in video-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15579–15591, 2023. 3

[19] Khalil Muhammad, Qinqin Wang, Diarmuid O'Reilly-Morgan, Elias Tragos, Barry Smyth, Neil Hurley, James Geraci, and Aonghus Lawlor. Fedfast: Going beyond average for faster training of federated recommender systems. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1234–1242, 2020. 1

[20] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15691–15701, 2023. 3

[21] Chen Qiu, Xingyu Li, Chaithanya Kumar Mummadi, Madan Ravi Ganesh, Zhenzhen Li, Lu Peng, and Wan-Yi Lin. Federated text-driven prompt generation for vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2, 3, 5, 6, 7, 8

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 5, 6, 7, 8

[23] Jinhyun So, Chaoyang He, Chien-Sheng Yang, Songze Li, Qian Yu, Ramy E Ali, Basak Guler, and Salman Avestimehr. Lightsecagg: a lightweight and versatile design for secure aggregation in federated learning. *Proceedings of Machine Learning and Systems*, 4:694–720, 2022. 1

[24] Guangyu Sun, Matias Mendieta, Aritra Dutta, Xin Li, and Chen Chen. Towards multi-modal transformers in federated learning. In *European Conference on Computer Vision*, pages 229–246. Springer, 2024. 3

[25] Zhenyu Sun, Xiaochun Niu, and Ermin Wei. Understanding generalization of federated learning via stability: Heterogeneity matters. In *International Conference on Artificial Intelligence and Statistics*, pages 676–684. PMLR, 2024. 1

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 2

[27] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7959–7971, 2022. 2

[28] Yinghui Xing, Qirui Wu, De Cheng, Shizhou Zhang, Guoqiang Liang, Peng Wang, and Yanning Zhang. Dual modality prompt tuning for vision-language pre-trained model. *IEEE Transactions on Multimedia*, 26:2056–2068, 2023. 2, 3

[29] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19187–19197, 2023. 3

[30] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6757–6767, 2023. 2, 6, 7, 8

[31] Xin Yao, Chaofeng Huang, and Lifeng Sun. Two-stream federated learning: Reduce the communication costs. In *2018 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2018. 1

[32] Xin Yao, Tianchi Huang, Chenglei Wu, Rui-Xiao Zhang, and Lifeng Sun. Federated learning with additional mechanisms on clients to reduce communication costs. *arXiv preprint arXiv:1908.05891*, 2019. 1

[33] Liping Yi, Gang Wang, Xiaoguang Liu, Zhuan Shi, and Han Yu. Fedgh: Heterogeneous federated learning with generalized global header. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8686–8696, 2023. 1

[34] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 1

[35] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 2, 3, 6, 7, 8

[36] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1, 2, 3