FedMVP: Federated Multimodal Visual Prompt Tuning for Vision-Language Models
(Supplementary Material)

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

The supplementary material is organized as follows: in Sec. A we lay down additional implementation details of our proposed FedMVP. In Sec. B we provide a detailed breakdown of the experimental results that have been reported in the main paper.

A. Additional implementation details

A.1. Attribute generation and usage

Attribute generation. As discussed in Sec. 3.3.1 of the main paper, one of the core proposals in our proposed FedMVP is to integrate the class attribute information into the multimodal prompt generation process orchestrated by the PromptFormer network.

The attributes for a given class are generated using a large language model (LLM), such as GPT-4o [2] in our case. For the k^{th} class name in the i^{th} client $c_{i,k}$, we query GPT-4o using a structured instructional prompt, following [24] as:

> **LLM Prompt**
>
> "What are the most useful detailed generic visual features for distinguishing a [class name] in an image? Please act as an expert with comprehensive knowledge of all aspects of generic objects."

where the "class name" is replaced with the value of $c_{i,k}$. For instance, when we prompt GPT-4o with the class name "giraffe" we get a comma separated list of attributes:

> **Attributes generated by LLM**
>
> "Distinctive coat pattern with large, irregular brown patches", "unique coat pattern with large, irregular brown patches", "exceptionally long neck, a primary distinguishing feature", "small, rounded ossicones or horns on the head", "slender, elongated legs, emphasizing their height", "tall, narrow body frame with prominent shoulders."

Composing text prompts. In addition to the attributes, we utilize generic hand-crafted prefixes (*e.g.* "a photo of a [class name]"), as used in [29], or domain-specific prefixes (*e.g.* "a sketch of a [class name]") tailored to each dataset. Details of the prefix templates for each benchmark are reported in Table A2. We then combine these prefixes with GPT-4o generated attributes using connector phrases "which is a/an" or "which has", to form composite text prompts for the CLIP text-encoder feature extractor. A complete example of text prompt for the class "giraffe" that is used for CLIP text feature attribute extraction is given as follows:

> **Composite prompt for CLIP text encoder**
>
> "A photo of a giraffe, which has a distinctive coat pattern with large, irregular brown patches."

We provide more examples of the LLM-generated attributes and the complete text prompts in Fig. 1.

Using attributes during training. Note from Fig. 2 of the main paper that there is a distinction between how attributes are used by the PromptFormer network and for training the FedMVP. To recap, the PromptFormer network takes as input only the LLM-generated attributes for

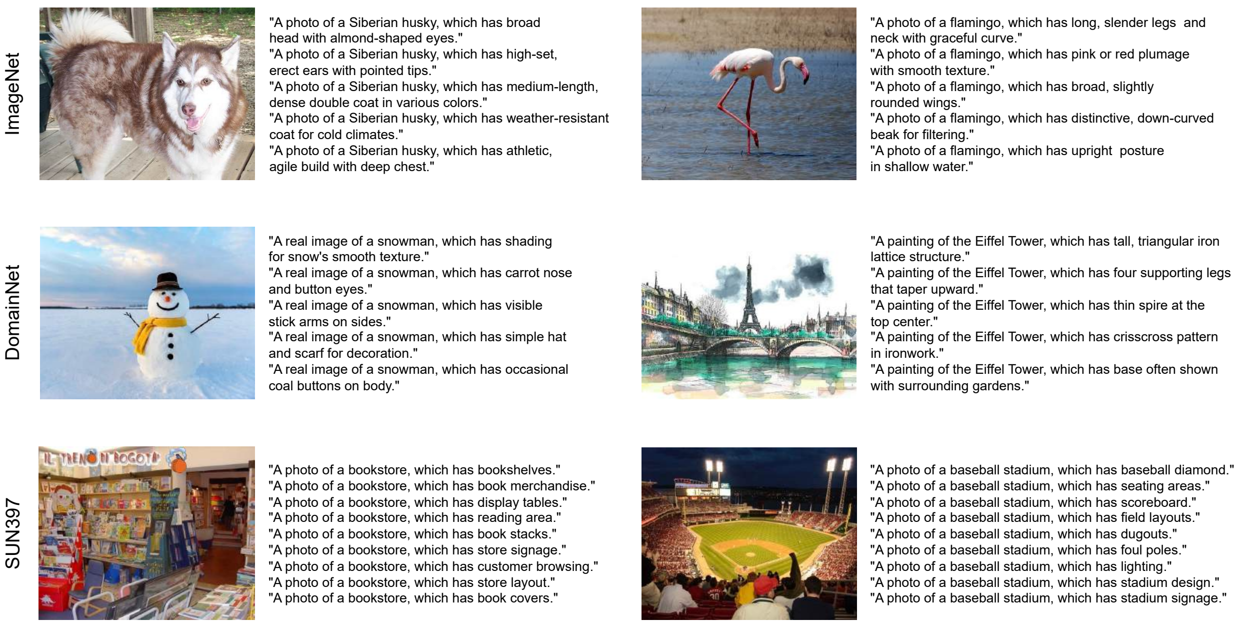


Figure 1. **Examples of LLM-generated attributes and complete text prompts.** We report some example text prompts used in FedMVP for the datasets ImageNet, DomainNet and SUN397.

constructing the multimodal visual prompts. Whereas, for training the FedMVP we use the composite text prompts, which has just been described above. Given a class name c_k[1] we obtain the text feature $\mathbf{t}_k$ (used in Eq. 5 of the main paper) from the CLIP text encoder $\mathcal{E}_t$ by using a composite text prompt. Then we repeat this step for all the attributes generated by the LLM for a given class k, which gives us a matrix of CLIP text embeddings $\mathbf{T}_k$, where the entry $\mathbf{T}_{k,j}$ is the text embedding corresponding the j^{th} attribute for class k. Given all the text embeddings for a class, we are now ready to compute the CLIP similarity in Eq. 5 of the main paper.

We follow the scoring proposed in DESC [24], where for a given image, belonging to class label y and class name c_k, the final prediction probability $p(y = k \mid \mathbf{I})$ are calculated by averaging the score for each attribute j as $p(y = k|\mathbf{I}) =:$

$$\frac{1}{\dim_2(\mathbf{T}_{k,J})} \sum_{j \in J} \frac{\exp(\cos(\mathbf{v}, \mathbf{T}_{k,j}))/\tau)}{\sum_{k'=1}^{K} \exp(\cos(\mathbf{v}, \mathbf{T}_{k',j}))/\tau)}, \quad (1)$$

where $\dim_2(\cdot)$ denotes the dimension or total number of attributes J for a given class k.

Using attributes during inference. For inference, we follow the same scoring as per Eq. 1. The predicted class is given by taking an argmax of the probability distribution over all the classes. Since we utilize the scoring method of DESC in our method, we also compare the performance of DESC with FedMVP on two generalization settings – base-to-new generalization and multi-source single-target generalization – in Tab. A1. From DESC numbers we can observe that using LLM-generated attributes alone is not sufficient to improve much over the ZS-CLIP baseline. Interestingly, DESC leads to a drop in performance over the ZS-CLIP baseline that does not use any attributes on the DomainBed benchmark. Contrarily, our FedMVP can better utilize the LLM attributes through the cross-attention mechanism of the proposed PromptFormer network, as described in Sec. 3.3.1 of the main paper.

[1] Omitting client index i for brevity.

Table A1. **Comparison of our FedMVP with zero-shot CLIP and DESC.** Average harmonic mean is reported for Base-to-New generalization and average accuracy of DomainBed benchmarks of multi-source single-target generalization.

Methods	Prompts	Base-to-New	MSST
ZS-CLIP [29]	hand-crafted	74.24	70.41
DESC [24]	hand-crafted	75.18	69.89
FedMVP (Ours)	textual+visual	**77.52**	**72.24**

Since all the evaluations are done on the server, except for local accuracy, as described in Sec. 4.1 of the main paper, each client sends the LLM-generated attributes to the server, which constitutes a relatively tiny communication overhead. Alternatively, the server can also generate the attributes for all the classes, since the clients will have no knowledge about the disjoint classes in other clients.

Algorithm 1 Federated Multimodal Visual Prompt Tuning (FedMVP) algorithm

procedure SERVER EXECUTION
    $\rho^0 = \{\Theta^0\}$ ▷ Initialize ρ^0 parameters of PromptFormer module
    for $r \leftarrow 0$ **to** R **do** ▷ For total communication rounds R
        Choose a random subset of remote clients as S_r
        for $i \in S_r$ in parallel **do** ▷ For a client i, belongs to S_r
            Send the current global model ρ^r to client i
            Receive locally updated ρ_i^{r+1} from **Client Training**
        end for
        Aggregate the updated model parameters, $\rho^{r+1} = \frac{1}{|S_r|} \sum_{i \in S_r} \rho_i^{r+1}$
    end for
    Get the final model parameter $\rho^R = \{\Theta^R, \theta^R\}$
end procedure
procedure CLIENT TRAINING
    Generate the attributes of set of classes, $\mathcal{C}_i = \{c_{i,k}\}_{k=1}^{K_i}$, by a LLM ▷ K_i is the total number of classes of i
    Extract the attribute embeddings, $\mathbf{A}_i = \{\mathcal{E}_t(\texttt{LLM}(\mathcal{C}_i))\}$
    for $l \leftarrow 0$ **to** L **do** ▷ For local epochs L
        if loss >threshold **then**
            Generate the visual prompts $\mathbf{P}$ using eq. 2 and eq. 3 ▷ Follow eq. 2 and eq. 3 from the main paper
            Concatenate the `[CLS]` token, the patch embeddings, and the visual prompts, as $\mathbf{I} = [\mathbf{z}; \mathbf{E}; \mathbf{P}]$
            Extract the visual features from $\mathcal{E}_v$
        else
            Start LoRA fine-tuning of the PromptFormer module
        end if
        Estimate the prediction scores using Eq. 1
        Calculate and update the losses using eq. 4 to eq. 7 ▷ Follow eq. 4 to eq. 7 from the main paper
        Update the parameters ρ^r to ρ_i^{r+1} locally using eq. 8 on $(x, y) \sim \mathcal{D}_i$ ▷ Follow eq. 8 from the main paper
    end for
end procedure

A.2. Pseudo-code of FedMVP

In Algorithm 1 we provide a pseudo-code of the full FedMVP algorithm. We split the algorithm into two parts: one for server execution and another for client training.

A.3. Architecture details of FedMVP

The only trainable parameters of FedMVP are composed of PromptFormer network parameters. Below we describe additional architecture details of each trainable network. Note that we do not tune the vision and text encoder backbones of CLIP, and hence refer the reader to the original paper [29] for the architecture details of CLIP.

PromptFormer. The PromptFormer network f_Θ consists of two multi-head cross-attention (MHCA) modules, two feed-forward networks (FFN), a projection layer T_{proj} and a learnable query prompt $\mathbf{Q}$. Each MHCA module consists of a 4-head cross-attention mechanism, followed by LayerNorm. Each FFN comprises of a two-layer bottleneck structure (Linear-GeLU-Linear). T_{proj} performs a linear transformation to convert the textual feature space of dimension 512 into the patch embedding space with a dimension of 768. $\mathbf{Q}$ is comprised of prompt length 4, initialized with a Gaussian distribution of $\sigma = 0.02$.

B. Additional experimental results

B.1. Dataset details.

In Tabs. A2(a) and (b) we provide detailed information of all the datasets used in the experiments of the main paper and the associated statistics, such as the number of classes, number of samples, and the prefix templates. In detail, the Tab. A2(a) includes the datasets used in the experiments corresponding to Tab. 2 of the main paper. The Tab. A2(a) includes the datasets used in Tabs. 1, 2, and 4 of the main paper. We refer the reader to the corresponding papers that have proposed the original datasets for further details and example images.

B.2. Experimental setup

Metrics. We assess the performance of all methods using classification accuracy. In the base-to-novel generalization setting, we additionally report the harmonic mean

Table A2. **Dataset Details**

(a) Domain Generation dataset statistical details on class, training and test splits, prefix template.

Dataset	Domain	Classes	Train	Test	Prefix template
PACS [19]	Art Painting	7	1,024	614	An art painting of a [CLASS]
	Cartoon		1,171	704	A cartoon of a [CLASS]
	Photo		835	502	A photo of a [CLASS]
	Sketch		1,964	1,179	A sketch of a [CLASS]
OfficeHome [33]	Art	65	1,214	728	An art of a [CLASS]
	Clipart		2,191	1,298	A clipart of a [CLASS]
	Product		2,226	1,324	A product image of a [CLASS]
	RealWorld		2,180	1,304	A realworld image of a [CLASS]
VLCS [8]	CALTECH	5	891	424	A high quality photo of a [CLASS], as a standalone object
	LABELME		1,672	797	A realworld photo of the [CLASS]
	PASCAL-VOC		2,127	1,013	A realworld photo of a [CLASS]
	SUN		2,067	985	A photo of a [CLASS], in diverse scenic environments
Terra Incognita [3]	Location-38	10	4,883	2,930	A photo of a [CLASS]
	Location-43		2,009	1,207	A photo of a [CLASS]
	Location-46		3,061	1,836	A photo of a [CLASS]
	Location-100		2,439	1,466	A photo of a [CLASS]
DomainNet [27]	Clipart	345	24,417	14,647	A clipart of a [CLASS]
	Infograph		26,609	15,948	An infograph of a [CLASS]
	Painting		37,873	22,744	A painting of a [CLASS]
	Quickdraw		86,250	51,750	A quickdraw image of a [CLASS]
	Real		87,663	52,604	A real image of a [CLASS]
	Sketch		35,195	21,109	A sketch of a [CLASS]

(b) Dataset statistical details on class, training and test splits, prefix template.

Dataset	Classes	Train	Test	Prefix template
Caltech101 [9]	101	4,128	2,465	A photo of a [CLASS]
Flowers102 [25]	102	4,093	2,463	A photo of a [CLASS], a type of flower
FGVCAircraft [23]	100	3,334	3,333	A photo of a [CLASS], a type of aircraft
UCF101 [32]	101	7,639	3,783	A photo of a person doing [CLASS]
OxfordPets [26]	37	2,944	3,369	A photo of a [CLASS], a type of pet
Food101 [5]	101	50,500	30,300	A photo of a [CLASS], a type of food
DTD [6]	47	2,820	1,692	A photo of a [CLASS], a type of texture
StanfordCars [18]	196	6,509	8,041	A photo of a [CLASS]
SUN397 [35]	397	15,880	19,850	A photo of a [CLASS]
EuroSAT [12]	10	13500	8,100	A centered satellite photo of [CLASS]
ImageNet [7]	1000	1.28M	50,000	A photo of a [CLASS]
ImageNetV2 [30]	1000	N/A	10,000	A photo of a [CLASS]
ImageNet-Sketch [34]	1000	N/A	50,889	A photo of a [CLASS]
ImageNet-A [14]	200	N/A	7500	A photo of a [CLASS]
ImageNet-R [13]	200	N/A	30,000	A photo of a [CLASS]

(HM) of the accuracies on base and new classes. All the performances are reported on the test split of each dataset, unless stated otherwise.

Base-to-New Generalization. In this setup (which corresponds to Sec. 4.1 of the main paper), we keep the participation of clients to 100% and the number of classes per client, $K = 20$, similar to [28] that produces 30 remote clients over 9 datasets. We set the batch size to 128, the training sample per class to 8, and the number of communication rounds to 200.

Domain Generalization. For both the Multi-source Single-target (MSST) and Single-source Multi-target (SSMT) domain generalization settings (corresponding to Sec. 4.2 of the main paper) on DomainBed benchmark, we keep the participation of clients to 100%, shots to 8 and batch size to 128. However, we set the number of classes per client, $K = 2$, and global communication round to 20 for PACS, VLCS, and Terra Incognita datasets. In contrast, for OfficeHome and DomainNet datasets, we fix the number of classes per client, $K = 20$, and global communication rounds to 100.

For the DG setting on ImageNet benchmark, we follow the setup of [28], keeping participation of clients to 10%, shots to 8, and number of classes per client, $K = 5$, for ImageNet training. In this case, we fix the batch size to 128 and the number of communication rounds to 200.

Cross-Dataset Generalization. Similar to [28], we keep the participation of clients to 10%, shots to 8, and the number of classes per client, $K = 5$ for ImageNet training, and perform the evaluation on 10 datasets. The batch size and number of communication rounds are fixed to 128 and 200,

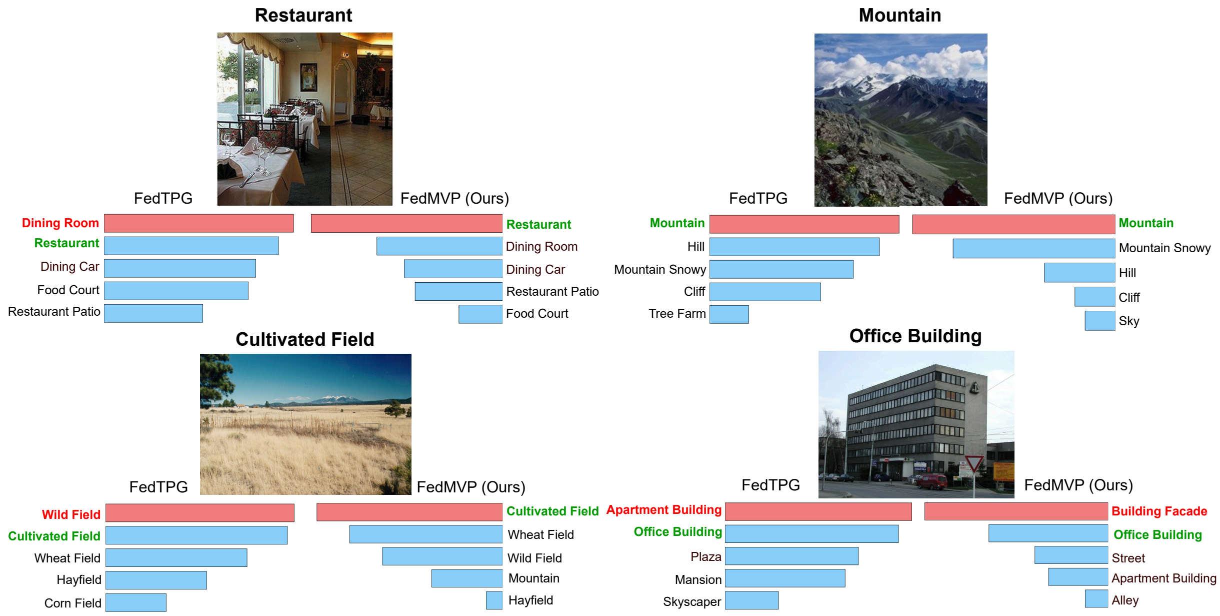


Figure 2. **Qualitative comparison** of top-5 predictions on SUN397 dataset in Base-to-New Generalization setting. The correct predictions (and annotations) are highlighted with green and the incorrect predictions are highlighted with red.

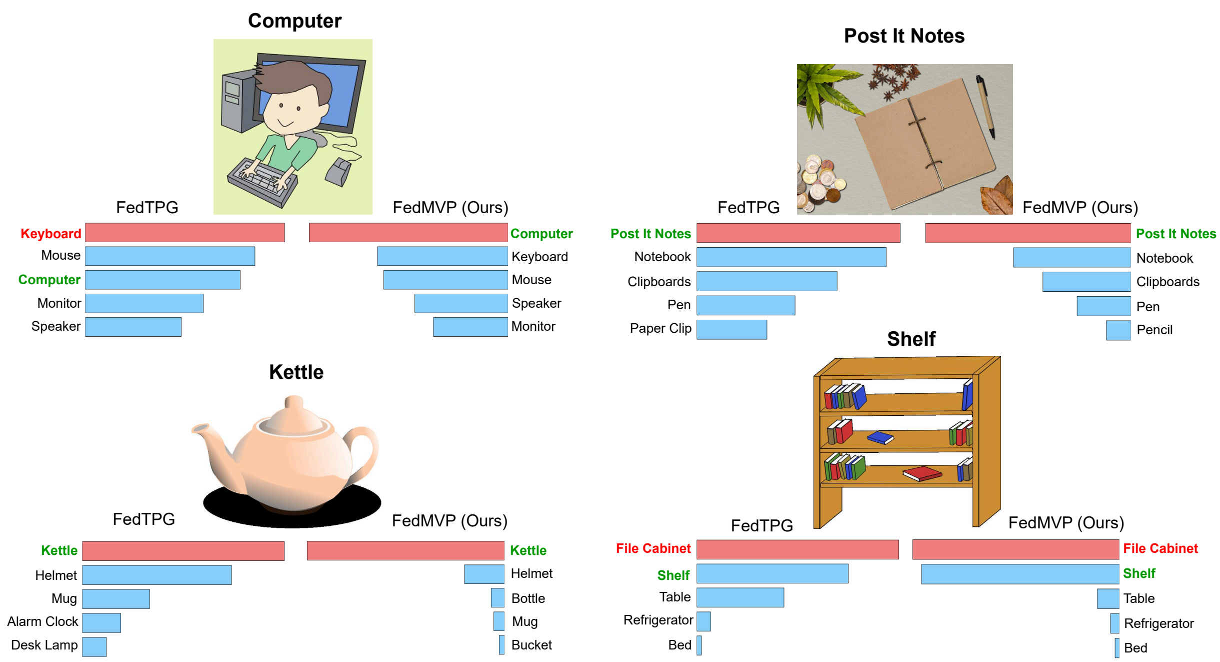


Figure 3. **Qualitative comparison** of top-5 predictions on Clipart domain of OfficeHome dataset in MSST DG setting. The correct predictions (and annotations) are highlighted with green and the incorrect predictions are highlighted with red.

respectively.

B.3. Detailed Results

Qualitative results. In Figs. 2 and 3 we randomly pick few test samples and visualize the top-5 prediction probabilities of our FedMVP and compare it with that of FedTPG [28].

In Fig. 2 we report the top-5 predictions of our FedMVP and FedTPG for a few samples corresponding to the base-to-new generalization generalization setting. On the top-left example, we notice that FedTPG confuses a "Restaurant" with a "Dining room", two classes that share a lot of visual similarities, whereas our FedMVP correctly classifies it as the "Restaurant". This behaviour can be attributed to the usage of attributes of our method that imparts a more fine-grained knowledge into the visual prompts. Interestingly, for the example of "Mountain" on top-right, while both the methods can predict the correct class, we see that for our FedMVP, the second most confident class is "Mountain Snowy", which is more accurate for the given example. This can be attributed due to the use of attributes during visual prompt tuning, as a LLM would generate "*snow on mountains*" as a characteristic attribute of a mountain. Thus, given enough training samples of mountains with snow, the model will even start recognizing snowy mountains, which is not the case for FedTPG, where the second most confident prediction is "Hill" – a more generic form of a mountainous landscape.

In Fig. 3 bottom-left example of the class "Kettle", both FedTPG and FedMVP predicts the correct class. However, the top-5 classes predicted by FedTPG include completely unrelated classes such as "Helmet", "Alarm Clock" and "Desk Lamp". Whereas, in our FedMVP we notice classes semantically similar to the class "Kettle" such as "Bottle", "Mug" and "Bucket". This indicates that the presence of the attributes helps the model to generalize across clients or domains unseen during training, and make more reasonable predictions, as long as objects share similar visual parts. However, in the bottom right example of Fig. 3 we observe that both the models get the prediction incorrect, but it is a more reasonable mistake as a "File Cabinet" shares a lot of visual similarity with the class "Shelf" which is the ground truth annotation.

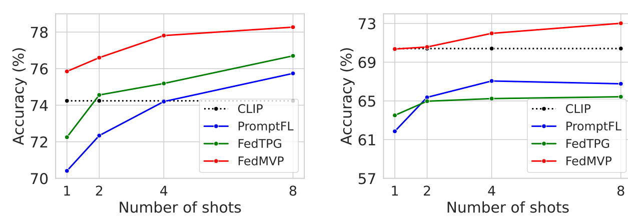
This underscores the importance of incorporating attributes during the visual prompt tuning step, enhancing the model's accuracy and ensuring that its errors remain reasonable even when the top-1 predictions are incorrect.

B.4. Detailed quantitative results

In this section we report the detailed experimental results corresponding to the Tabs. 1, and 2 of the main paper, which are essentially the summarized versions of the tables in the supplementary material.

Base-to-New Generalization. In Tab. A4, we present the expanded version of Tab. 1 of the main paper. Here, we showcase the base-to-new generalization performances of 9 datasets. In detail, we report the seen class accuracies, *i.e.*, the local and base accuracy, the unseen class accuracy or the new accuracy, and the harmonic mean (HM) of the base and new class accuracies, separately for each dataset. Note that all the 9 datasets participates in the federated training set up. The Tab. A4(a) is the same as Tab. 1 of the main paper, and the Tabs. A4(b)-(h) reports the performance on each dataset separately.

From these tables on individual datasets we observe that our proposed FedMVP outperforms the baselines in majority of the datasets, with a few exceptions. This demonstrates that the improvement brought by FedMVP is consistent across datasets, and the average performance in Tab. A4(a) or Tab. 1 of the main paper is not dominated by some particular dataset. In summary, we have demonstrated that FedMVP can successfully generalize on the base, *i.e.*, combined classes from multiple clients, and completely new classes. Notably, FedMVP has achieved better performances with significant margin on unseen classes, where other FL methods fail to do so.



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

Figure 4. **Sensitivity to number of shots of images** on (a) Base-to-New and (b) MSST DG (DomainNet) setting.

Domain Generalization. In Tab. A5, A6 & A7 we present the detailed results of the summarized results of Tab. 2 of the main paper on the DomainBed benchmark in MSST DG setting. Here, the accuracy of a particular domain refer that the model is tested on that domain, while trained on rest of the domains. FedMVP has shown superior performances over other FL methods in all of the datasets, except Terra Incognita, where FedCLIP and FedMaPLe are able to classify fine-grained animal classes better than others.

In addition, we have provided the detailed results on the SSMT DG setting (reported in Tab. 2 of the main paper) for the datasets PACS (in Tab. A8), OfficeHome in (Tab. A9), VLCS (in Tab. A10), Terra Incognita (in Tab. A11) and DomainNet (in Tab. A12). We also notice similar trend, with our FedMVP outperforming the baselines consistently on several datasets, with a few exceptions.

Table A3. **Comparison of effects of different prompting methods used in FedMVP on the Base-to-New, Multi-source Single-target (MSST) and Single-source Multi-target (SSMT) Domain Generalization settings**.

Method	B2N	MSST	SSMT
Textual Prompting	73.56	66.26	65.89
Multi-modal Prompting	75.95	69.14	68.76
Visual Prompting (ours)	**78.27**	**73.02**	**72.63**

B.5. Detailed ablation studies

Number of shots of images. In Fig. 4, we present the performance of FedMVP compared FL baselines across varying numbers of shots (or images) in base-to-new and MSST DG settings. Both of the results clearly demonstrate that FedMVP consistently outperforms others at all shot levels. Interestingly, even with as few as 2 samples per class, FedMVP can outperform the baselines with four times the data, indicating better data efficiency.

Sensitivity to α hyperparamer in our FedMVP. In Fig. 5, we demonstrate that maintaining a constant value of $\alpha = 10$ yields consistent results across both the base-to-new generalization and MSST DG tasks. It is evident from the plot that as the value of α increases, the influence of the cross-entropy loss term, $\mathcal{L}_{ce}$, diminishes. This reduction in influence ultimately leads to less accurate backpropagation of loss functions for classification task, highlighting the delicate balance between α and the model's performance.

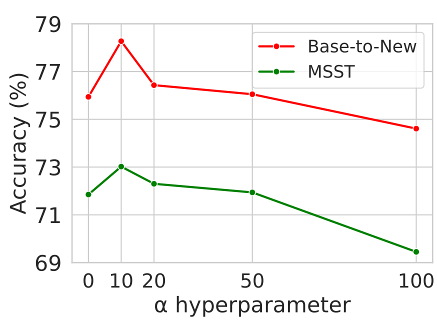


Figure 5. **Sensitivity to α hyperparamer in our FedMVP** on Base-to-New and MSST DG (DomainBed) setting.

Effect of LLMs in our FedMVP. We evaluate the performance of our FedMVP using four different large language models (LLMs), namely Llama-3.2-3B [10], Qwen2.5-14B [36], Phi-4 [1], and GPT-4o [2], and present their results in Table A13 for the MSST DG settings on both the PACS and OfficeHome datasets. The results demonstrate that GPT-4o outperforms the other models, achieving the highest performance overall. This superior performance can be attributed to its ability to capture more accurate and nuanced feature descriptions. Interestingly, all of the evaluated LLMs significantly outperform the second-best competitor, FedCLIP, on the MSST DG task. This highlights the importance of detailed feature representations in the success of our FedMVP, further emphasizing the value of precise feature description for enhancing the performance.

Superiority of FedMVP in non-federated offline setting. In Table A14, we present the performance of FedMVP, demonstrating its potential to significantly improve vision-language alignment, even in non-federated offline settings. For comparison, we include a range of state-of-the-art methods, such as zero-shot CLIP (ZS-CLIP) [29] and several recent non-federated prompt learning techniques, including CoOp [41], CoCoOp [40], VPT [15], KgCoOp [37], MaPLe [16], PromptSRC [17], StyLIP [4], CoPrompt [31], TCP [38], DePT [39], and DeKgTCP [21]. These methods are evaluated within the base-to-new generalization setting. Our findings reveal that while DePT achieves the highest performance on the base classes, our FedMVP outperforms all the competitors when it comes to unseen new classes, as well as the harmonic mean across both base and new classes. This highlights the robustness of FedMVP, which not only demonstrates a reduced tendency to overfit but also exhibits superior adaptability to unseen samples during inference. These results underscore the effectiveness of FedMVP in generalizing across both known and novel data distributions, making it a promising approach for real-world vision-language tasks.

Effect of different prompting methods. While using the multi-modal prompts for multi-modal prompting (*i.e.*, through both vision and text encoder) is an interesting proposal, we find that it hurts performance (Table A3). We postulate that tuning prompts through both encoders is redundant as long as the prompts are multi-modal. While we expected the textual prompting with multi-modal prompts to work equally well as FedMVP, we find it further degrades performance. This can be attributed to overfitting, as shown in [37]. In addition, we can cache the text representations once and backpropagate through the vision encoder alone in FedMVP, which is computation friendlier.

Table A4. **Comparison of methods on the Base-to-new generalization setting**. Tables (b)-(k) report the performance on each dataset.

(a) **Average over 9 datasets**

Method	Local	Base	New	HM
ZS-CLIP [29]	76.72	70.51	75.78	74.24
FedOTP [20]	74.82	65.22	57.04	64.89
FedCoCoOp [40]	81.46	73.76	66.00	73.20
FedVPT [15]	76.29	70.43	74.89	73.79
FedCLIP [22]	76.87	71.04	75.06	74.24
FedMaPLe [16]	81.63	74.44	70.62	75.29
FedKgCoOp [37]	78.38	72.18	75.87	75.39
PromptFL [11]	81.75	74.47	71.70	75.74
FedTPG [28]	80.75	73.68	76.02	76.70
FedMVP (Ours)	**81.89**	**75.37**	**77.82**	**78.27**

(b) **Caltech101**

Method	Local	Base	New	HM
ZS-CLIP [29]	97.57	96.97	93.89	96.12
FedOTP [20]	95.72	94.83	86.46	92.14
FedCoCoOp [40]	96.71	94.41	91.59	94.19
FedVPT [15]	96.23	95.31	94.53	95.35
FedCLIP [22]	97.71	97.29	94.21	96.38
FedMaPLe [16]	97.00	95.41	90.06	94.06
FedKgCoOp [37]	97.65	97.24	94.79	96.54
PromptFL [11]	96.97	96.69	92.79	95.44
FedTPG [28]	97.59	97.08	95.24	96.63
FedMVP (Ours)	**97.85**	**97.73**	**95.48**	**97.01**

(c) **Flowers102**

Method	Local	Base	New	HM
ZS-CLIP [29]	82.58	72.18	77.94	77.33
FedOTP [20]	86.95	65.90	62.06	70.11
FedCoCoOp [40]	94.00	**77.49**	65.63	77.36
FedVPT [15]	81.67	73.09	76.10	76.79
FedCLIP [22]	79.72	71.51	75.96	75.58
FedMaPLe [16]	94.28	76.44	68.51	78.36
FedKgCoOp [37]	84.59	72.11	77.06	77.59
PromptFL [11]	**94.44**	76.40	70.12	79.07
FedTPG [28]	90.76	71.80	77.76	79.35
FedMVP (Ours)	94.05	76.34	**78.24**	**82.16**

(d) **FGVCAircraft**

Method	Local	Base	New	HM
ZS-CLIP [29]	30.59	27.55	35.81	30.96
FedOTP [20]	28.35	24.01	15.53	21.23
FedCoCoOp [40]	35.21	31.93	22.67	28.89
FedVPT [15]	31.36	27.92	32.67	30.51
FedCLIP [22]	31.41	28.45	34.07	31.14
FedMaPLe [16]	35.83	31.39	32.34	33.08
FedKgCoOp [37]	33.68	29.79	34.01	32.38
PromptFL [11]	**36.29**	32.41	30.95	33.07
FedTPG [28]	34.68	30.82	35.18	33.44
FedMVP (Ours)	35.50	**32.54**	**37.32**	**35.01**

(e) **UCF101**

Method	Local	Base	New	HM
ZS-CLIP [29]	80.75	70.58	77.50	76.04
FedOTP [20]	70.99	59.61	51.54	59.68
FedCoCoOp [40]	84.92	75.23	64.25	73.83
FedVPT [15]	82.43	71.32	77.05	76.66
FedCLIP [22]	79.67	70.22	75.66	74.98
FedMaPLe [16]	84.17	75.12	68.68	75.47
FedKgCoOp [37]	82.66	73.14	76.36	77.19
PromptFL [11]	**86.13**	75.65	70.60	76.94
FedTPG [28]	85.64	74.89	76.64	78.79
FedMVP (Ours)	85.41	**75.92**	**80.25**	**80.34**

(f) **OxfordPets**

Method	Local	Base	New	HM
ZS-CLIP [29]	91.33	91.33	97.04	93.16
FedOTP [20]	88.62	88.62	71.92	82.26
FedCoCoOp [40]	92.34	92.34	87.36	90.62
FedVPT [15]	89.70	89.70	96.73	91.93
FedCLIP [22]	91.17	91.18	93.12	91.81
FedMaPLe [16]	**95.00**	**95.00**	**97.09**	**95.69**
FedKgCoOp [37]	91.58	91.58	96.53	93.17
PromptFL [11]	93.31	93.32	95.39	94.00
FedTPG [28]	94.70	94.69	95.79	95.06
FedMVP (Ours)	94.05	94.05	96.12	94.73

(g) **Food101**

Method	Local	Base	New	HM
ZS-CLIP [29]	94.39	90.16	91.25	91.90
FedOTP [20]	87.06	77.12	69.09	77.07
FedCoCoOp [40]	93.24	87.57	84.95	88.45
FedVPT [15]	90.26	88.39	89.45	89.36
FedCLIP [22]	94.23	89.57	90.67	91.45
FedMaPLe [16]	93.95	89.43	89.60	90.95
FedKgCoOp [37]	94.19	89.94	91.81	91.95
PromptFL [11]	93.52	88.63	88.47	90.15
FedTPG [28]	94.09	89.87	91.64	91.83
FedMVP (Ours)	**95.06**	**91.89**	**92.57**	**93.15**

(h) **DTD**

Method	Local	Base	New	HM
ZS-CLIP [29]	53.13	53.01	58.21	54.68
FedOTP [20]	**69.21**	**69.21**	55.31	63.86
FedCoCoOp [40]	68.63	68.63	45.77	58.83
FedVPT [15]	52.06	52.06	60.13	54.50
FedCLIP [22]	56.48	56.48	60.39	57.72
FedMaPLe [16]	68.28	68.28	46.61	59.12
FedKgCoOp [37]	58.76	58.75	59.61	59.04
PromptFL [11]	68.67	68.67	52.74	62.39
FedTPG [28]	63.62	63.62	60.51	62.55
FedMVP (Ours)	67.32	67.32	**64.96**	**66.51**

(i) **StanfordCars**

Method	Local	Base	New	HM
ZS-CLIP [29]	71.51	63.44	74.90	69.61
FedOTP [20]	58.89	45.25	44.00	48.54
FedCoCoOp [40]	**76.62**	**66.51**	66.40	69.53
FedVPT [15]	73.47	65.98	71.47	70.16
FedCLIP [22]	72.32	64.42	75.04	70.30
FedMaPLe [16]	74.76	66.26	71.33	70.61
FedKgCoOp [37]	71.89	64.33	75.71	70.32
PromptFL [11]	74.53	66.16	72.32	70.82
FedTPG [28]	74.54	66.34	74.26	71.50
FedMVP (Ours)	75.30	66.45	**75.94**	**72.29**

(j) **SUN397**

Method	Local	Base	New	HM
ZS-CLIP [29]	88.66	69.41	75.46	77.05
FedOTP [20]	87.54	62.39	57.42	66.87
FedCoCoOp [40]	91.44	69.76	65.36	73.94
FedVPT [15]	89.43	70.13	75.92	77.69
FedCLIP [22]	89.10	70.22	76.42	77.82
FedMaPLe [16]	91.40	72.66	71.33	77.47
FedKgCoOp [37]	90.38	72.72	76.94	79.34
PromptFL [11]	91.93	72.34	71.89	77.70
FedTPG [28]	91.11	74.01	77.13	80.10
FedMVP (Ours)	**92.44**	**76.09**	**79.51**	**82.11**

Table A5. **Comparison of methods on the Multi-source Single-target (MSST) Domain Generalization setting**. The results are reported for the PACS and OfficeHome datasets.

Method	**PACS**					**OfficeHome**				
	A. Painting	Cartoon	Photo	Sketch	Average	Art	Clipart	Product	RealWorld	Average
ZS-CLIP [29]	97.55	98.72	100.00	88.38	96.16	80.63	67.25	87.99	90.10	81.49
FedOTP [20]	93.49	93.75	98.00	77.61	90.71	71.98	65.33	83.08	85.28	76.42
FedCoCoOp [40]	81.43	92.05	81.67	85.07	85.06	78.54	66.73	90.18	90.24	81.42
FedTPG [28]	90.55	95.17	89.84	88.38	90.99	80.63	68.56	90.78	91.14	82.78
PromptFL [11]	96.42	97.72	99.80	87.53	95.37	79.61	66.70	90.42	90.22	81.74
FedKgCoOp [37]	96.48	97.77	99.83	87.58	95.42	79.67	66.87	90.48	90.26	81.82
FedMaPLe [16]	90.72	97.44	99.80	90.08	94.51	79.67	68.34	90.33	89.80	82.03
FedVPT [15]	95.73	96.37	100.00	89.32	95.36	80.94	68.21	88.56	89.34	81.76
FedCLIP [22]	**97.56**	98.72	100.00	88.89	96.29	80.50	67.26	89.05	90.15	81.74
FedMVP (Ours)	96.92	**99.35**	**100.00**	**92.86**	**97.28**	**82.20**	**70.05**	**91.78**	**92.56**	**84.15**

Table A6. **Comparison of methods on the Multi-source Single-target (MSST) Domain Generalization setting**. The results are reported for the VLCS and Terra Incognita datasets.

Method	**VLCS**					**Terra Incognita**				
	Caltech	Labelme	Pascal-VOC	Sun	Average	L38	L43	L46	L100	Average
ZS-CLIP [29]	100.00	**68.88**	89.24	75.02	83.29	20.14	33.64	29.19	52.93	33.98
FedOTP [20]	86.79	57.21	62.98	62.64	67.41	4.30	27.92	16.94	3.82	13.24
FedCoCoOp [40]	54.95	58.59	65.35	68.02	61.73	10.58	8.87	19.55	55.73	23.68
FedTPG [28]	88.92	61.48	62.98	65.69	69.77	**46.11**	19.14	21.78	20.12	26.79
PromptFL [11]	98.82	55.71	75.42	69.54	74.87	14.35	14.62	21.60	49.51	25.02
FedKgCoOp [37]	98.87	55.80	75.36	69.57	74.90	14.37	14.66	21.57	49.52	25.03
FedMaPLe [16]	99.53	52.70	71.08	63.86	71.79	30.10	33.64	28.81	52.66	36.30
FedVPT [15]	99.45	69.47	89.87	73.98	83.19	20.46	34.76	26.08	53.17	33.62
FedCLIP [22]	100.00	67.88	87.48	75.43	82.70	25.60	**35.21**	29.25	56.28	36.58
FedMVP (Ours)	**100.00**	70.15	**90.93**	**79.40**	**85.12**	23.95	34.95	**33.19**	**57.34**	**37.36**

Table A7. **Comparison of methods on the Multi-source Single-target (MSST) Domain Generalization setting**. The results are reported for the DomainNet dataset.

Method	**DomainNet**						
	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Average
ZS-CLIP [29]	70.88	45.94	66.27	14.19	83.22	62.25	57.13
FedOTP [20]	64.33	38.42	54.60	11.63	73.45	55.61	49.67
FedCoCoOp [40]	70.84	46.26	65.90	14.35	83.13	62.00	57.08
FedTPG [28]	71.35	46.03	66.10	13.96	81.50	61.97	56.82
PromptFL [11]	70.95	45.98	65.57	13.80	82.76	62.15	56.87
FedKgCoOp [37]	70.96	46.00	66.02	13.83	82.83	61.90	56.92
FedMaPLe [16]	72.91	49.93	67.13	16.20	82.73	64.36	58.88
FedVPT [15]	71.74	42.57	61.47	13.65	83.66	62.80	55.98
FedCLIP [22]	71.88	46.46	67.09	15.13	83.56	62.96	57.85
FedMVP (Ours)	**73.93**	**52.06**	**69.16**	**18.50**	**86.64**	**66.72**	**61.17**

Table A8. **Comparison of methods on the Single-source Multi-target (SSMT) Domain Generalization setting**. The results are reported for the PACS dataset. Here Ap, Cr, Ph, Sk denote to Art painting, Cartoon, Photo and Sketch domains respectively.

Method	**Ap**				**Cr**				**Ph**				**Sk**			
	Cr	Ph	Sk	Avg.	Ap	Ph	Sk	Avg.	Ap	Cr	Sk	Avg.	Ap	Cr	Ph	Avg.
ZS-CLIP [29]	98.72	100.00	88.38	95.70	97.55	100.00	88.38	95.31	97.55	98.72	88.38	94.88	97.55	98.72	100.00	98.76
FedOTP [20]	93.75	98.21	82.36	91.44	91.69	98.61	76.42	88.91	90.39	94.74	78.63	87.92	94.46	95.99	98.81	96.42
FedCoCoOp [40]	92.90	80.88	85.58	86.45	79.97	79.88	84.73	81.53	79.97	90.77	84.48	85.07	81.76	91.90	81.87	85.18
FedTPG [28]	93.89	84.46	88.63	88.99	89.90	91.03	88.04	89.66	91.04	95.31	87.87	91.41	90.88	95.60	91.83	92.77
PromptFL [11]	98.72	100.00	92.11	96.94	96.09	98.60	88.97	94.55	94.13	97.02	92.36	94.50	96.41	96.31	99.20	97.31
FedKgCoOp [37]	98.65	99.89	92.15	96.90	96.15	98.63	88.91	94.56	94.24	96.87	92.40	94.50	96.35	96.25	99.27	97.29
FedMaPLe [16]	97.59	99.80	90.25	95.88	90.39	99.80	90.16	93.45	90.88	97.73	90.08	92.89	90.23	97.30	96.81	94.78
FedVPT [15]	97.33	99.53	84.24	93.70	95.44	98.25	85.89	93.19	97.85	98.67	86.13	94.22	96.56	98.06	99.47	98.03
FedCLIP [22]	98.72	100.00	88.89	95.87	**97.56**	100.00	88.89	95.48	**97.56**	98.72	88.89	95.06	**97.56**	98.72	100.00	98.76
FedMVP (Ours)	**99.02**	**100.00**	**93.15**	**97.39**	96.45	**100.00**	**92.19**	**96.21**	96.87	**98.91**	**92.56**	**96.11**	97.49	**99.40**	**100.00**	**98.96**

Table A9. **Comparison of methods on the Single-source Multi-target (SSMT) Domain Generalization setting**. The results are reported for the OfficeHome dataset. Here Ar, Cl, Pr, Rw denote to Art, Clipart, Product and RealWorld domains respectively.

Method	Ar				Cl				Pr				Rw			
	Cl	Pr	Rw	Avg.	Ar	Pr	Rw	Avg.	Ar	Cl	Rw	Avg.	Ar	Cl	Pr	Avg.
ZS-CLIP [29]	67.25	87.99	90.10	81.78	80.63	87.99	90.10	86.24	80.63	67.25	90.10	79.33	80.63	67.25	87.99	78.62
FedOTP [20]	63.33	80.97	83.59	75.96	69.78	80.36	80.75	76.97	68.00	62.33	84.75	71.69	71.29	64.87	84.37	73.51
FedCoCoOp [40]	36.98	55.21	54.37	48.85	52.20	53.70	53.99	53.30	57.00	39.98	60.74	52.57	52.20	37.44	56.65	48.76
FedTPG [28]	68.41	90.33	91.18	83.31	81.04	90.78	91.26	87.69	79.94	**69.10**	90.33	79.79	80.35	68.49	90.03	79.62
PromptFL [11]	66.79	89.95	90.80	82.51	80.49	90.26	90.64	87.13	79.26	66.25	90.34	78.62	80.77	65.95	89.12	78.61
FedKgCoOp [37]	66.83	90.02	90.85	82.57	80.41	90.29	90.56	87.09	79.45	66.12	90.56	78.71	80.61	65.72	89.02	78.45
FedMaPLe [16]	66.03	89.35	88.73	81.37	75.14	86.18	85.20	82.17	**90.08**	67.26	89.19	**82.18**	**82.69**	68.03	89.88	80.20
FedVPT [15]	68.05	88.25	89.50	81.93	78.46	84.67	90.89	84.67	79.32	68.94	90.35	79.54	80.87	66.30	86.29	77.82
FedCLIP [22]	65.95	88.07	90.49	81.50	79.95	88.90	90.57	86.47	80.91	68.26	91.41	80.19	80.22	66.72	87.99	78.31
FedMVP (Ours)	**70.08**	**91.76**	**92.14**	**84.66**	**81.88**	**92.00**	**92.21**	**88.70**	80.79	69.03	**92.35**	80.72	82.16	**70.25**	**92.07**	**81.49**

Table A10. **Comparison of methods on the Single-source Multi-target (SSMT) Domain Generalization setting**. The results are reported for the VLCS dataset. Here C, L, V, S denote to Caltech, LabelMe, Pascal-VOC and SUN domains respectively.

Method	C				L				V				S			
	L	V	S	Avg.	C	V	S	Avg.	C	L	S	Avg.	C	L	V	Avg.
ZS-CLIP [29]	68.88	89.24	75.02	77.71	100.00	89.24	75.02	88.09	**100.00**	68.88	75.02	81.30	**100.00**	68.88	**89.24**	**86.04**
FedOTP [20]	58.09	69.89	63.49	63.83	69.10	54.00	52.69	58.60	98.82	62.54	71.70	77.69	75.00	49.69	56.86	60.52
FedCoCoOp [40]	57.34	67.42	68.63	64.46	50.94	62.39	67.21	60.18	56.84	55.83	67.21	59.96	56.84	58.59	65.75	60.39
FedTPG [28]	70.89	81.93	62.84	71.89	46.22	56.56	59.08	53.95	98.35	65.12	72.49	78.65	79.95	51.69	66.43	66.02
PromptFL [11]	61.94	80.12	67.05	69.70	45.20	60.97	59.13	55.10	98.78	36.60	56.09	63.82	99.23	59.76	78.82	79.27
FedKgCoOp [37]	61.98	80.45	67.11	69.85	45.28	61.10	64.16	56.85	98.82	36.76	56.04	63.87	99.29	59.85	79.07	79.40
FedMaPLe [16]	66.00	76.11	65.99	69.37	87.50	57.65	59.59	68.25	99.29	53.20	66.90	73.13	99.53	56.34	76.90	77.59
FedVPT [15]	69.36	89.68	75.93	78.32	98.45	88.60	75.34	87.46	98.78	68.60	73.98	80.45	96.59	66.27	87.67	83.51
FedCLIP [22]	68.63	87.07	76.04	77.25	**100.00**	85.89	75.74	87.21	**100.00**	60.48	75.03	78.50	**100.00**	66.63	86.18	84.27
FedMVP (Ours)	**70.93**	**90.40**	**78.13**	**79.82**	**100.00**	**90.54**	**78.37**	**89.64**	**100.00**	**70.06**	**78.93**	**83.00**	**100.00**	**69.44**	88.51	85.98

Table A11. **Comparison of methods on the Single-source Multi-target (SSMT) Domain Generalization setting**. The results are reported for the Terra Incognita dataset. Here L38, L43, L46, L100 denote to Location-38, Location-43, Location-46 and Location-100 domains respectively.

Method	L38				L43				L46				L100			
	L43	L46	L100	Avg.	L38	L46	L100	Avg.	L38	L43	L100	Avg.	L38	L43	L46	Avg.
CLIP [29]	33.64	29.19	52.93	38.59	20.14	29.19	52.93	34.09	20.14	33.64	52.93	35.57	20.14	33.64	29.19	27.66
FedOTP [20]	27.42	17.05	1.84	15.44	28.12	19.34	9.62	19.03	0.38	28.50	3.82	10.90	0.99	27.76	16.45	15.06
FedCoCoOp [40]	14.00	14.71	51.50	26.74	6.59	29.41	54.98	30.33	9.86	7.87	46.93	21.55	10.27	14.17	16.50	13.65
FedTPG [28]	20.21	13.29	10.98	14.83	**38.46**	15.68	11.32	21.82	**45.42**	12.34	15.41	24.39	**46.10**	19.47	21.40	28.99
PromptFL [11]	19.97	13.40	45.70	26.36	21.81	11.38	24.21	19.13	19.73	14.08	57.50	30.44	29.73	15.08	17.43	20.75
FedKgCoOp [37]	19.94	13.44	45.72	26.37	21.83	11.42	24.22	19.16	19.73	14.09	57.55	30.46	29.78	15.14	17.48	20.80
FedMaPLe [16]	**34.38**	28.11	44.61	35.70	24.64	20.75	46.59	30.66	30.38	30.99	48.09	36.48	34.16	33.55	26.85	**31.52**
FedVPT [15]	32.75	29.75	48.64	37.05	18.45	30.56	50.58	33.20	18.93	30.67	48.18	32.59	16.59	33.08	28.57	26.08
FedCLIP [22]	**34.38**	27.56	52.87	38.27	25.67	26.31	**58.73**	**36.90**	24.81	**36.70**	**57.78**	**39.76**	27.13	**38.61**	26.74	30.83
FedMVP (Ours)	33.90	**33.08**	**55.34**	**40.77**	26.78	**33.10**	56.16	38.68	23.64	34.39	55.67	37.90	25.43	34.28	**31.32**	30.34

Table A12. **Comparison of methods on the Single-source Multi-target (SSMT) Domain Generalization setting**. The results are reported for the DomainNet dataset. Here Cl, Ig, Pt, Qd, Re, Sk denote to Clipart, Infograph, Panting, Quickdraw, Real and Sketch domains respectively.

Method	Cl						Ig						Pt					
	Ig	Pt	Qd	Re	Sk	Avg.	Cl	Pt	Qd	Re	Sk	Avg.	Cl	Ig	Qd	Re	Sk	Avg.
ZS-CLIP [29]	45.94	66.27	14.19	83.22	62.25	54.37	70.88	66.27	14.19	83.22	62.25	59.36	70.88	45.94	14.19	83.22	62.25	55.30
FedOTP [20]	30.68	44.32	8.68	58.81	46.08	37.71	51.16	42.96	7.56	57.81	45.28	40.95	52.28	30.93	8.20	59.04	46.35	39.36
FedCoCoOp [40]	50.51	67.71	15.62	83.81	64.72	56.47	71.87	66.02	14.56	84.15	63.81	60.08	71.36	48.02	13.04	82.76	63.78	55.79
FedTPG [28]	50.02	67.06	12.70	84.24	63.31	55.47	71.92	67.36	13.12	84.26	63.79	60.09	72.58	50.12	12.55	84.24	63.44	56.59
PromptFL [11]	50.27	67.48	15.60	83.55	64.52	56.28	71.59	65.95	14.25	83.89	63.62	59.86	71.52	47.91	14.67	82.84	63.71	56.13
FedKgCoOp [37]	50.35	67.52	15.68	83.70	64.61	56.37	71.72	65.98	14.27	83.94	63.67	59.92	71.58	47.78	14.52	82.88	63.62	56.08
FedMaPLe [16]	50.35	66.69	15.74	82.99	63.59	55.87	71.37	66.51	15.74	83.15	63.67	60.09	71.70	48.82	14.30	82.79	62.41	56.00
FedVPT [15]	47.65	66.50	14.94	83.65	63.01	55.15	71.41	66.36	14.63	80.45	60.70	58.71	67.46	42.84	12.57	80.82	60.40	52.82
FedCLIP [22]	45.72	66.62	14.83	83.17	64.42	54.95	71.34	66.63	14.81	83.17	62.42	59.67	71.37	45.72	14.82	83.17	62.41	55.50
FedMVP (Ours)	**51.26**	**69.72**	**18.96**	**86.53**	**65.92**	**58.48**	**73.47**	**69.74**	**18.25**	**85.94**	**65.48**	**62.58**	**74.00**	**51.38**	**18.95**	**85.73**	**64.78**	**58.97**

Method	Qd						Re						Sk					
	Cl	Ig	Pt	Re	Sk	Avg.	Cl	Ig	Pt	Qd	Sk	Avg.	Cl	Ig	Pt	Qd	Re	Avg.
ZS-CLIP [29]	70.88	45.94	66.27	83.22	62.25	65.71	70.88	45.94	66.27	14.19	62.25	51.91	70.88	45.94	66.27	14.19	83.22	56.10
FedOTP [20]	48.55	29.38	40.06	55.21	43.48	43.34	51.93	29.53	43.81	8.52	45.23	35.80	52.54	30.88	44.26	8.45	59.03	39.03
FedCoCoOp [40]	71.45	49.26	67.34	81.87	63.67	66.72	71.66	48.70	67.29	13.04	63.78	52.89	72.34	49.98	67.50	14.89	83.96	57.73
FedTPG [28]	71.52	50.00	67.87	83.47	64.64	67.50	72.48	49.75	66.97	12.26	62.88	52.87	72.81	50.14	67.41	13.00	84.22	57.52
PromptFL [11]	71.29	49.01	67.24	82.09	63.76	66.68	71.47	48.96	67.14	13.25	63.54	52.87	72.16	50.41	67.35	15.30	83.87	57.82
FedKgCoOp [37]	71.33	49.15	67.13	82.14	63.85	66.72	71.52	48.87	67.19	13.34	63.65	52.91	72.08	50.49	67.42	15.37	83.74	57.82
FedMaPLe [16]	68.63	45.32	61.38	78.03	59.69	62.61	72.07	49.42	67.60	12.63	63.24	52.99	**72.96**	49.64	68.09	16.21	83.75	58.13
FedVPT [15]	71.70	44.28	60.73	79.44	62.83	63.80	66.58	44.78	61.22	15.29	60.57	49.69	67.27	46.37	61.00	12.78	81.68	53.82
FedCLIP [22]	71.45	45.78	66.64	83.20	62.48	65.91	71.32	45.73	66.61	14.83	62.43	52.18	71.37	45.72	66.59	14.81	83.18	56.34
FedMVP (Ours)	**73.69**	**51.90**	**68.72**	**84.50**	**65.14**	**68.79**	**72.57**	**51.72**	**68.91**	**18.59**	**65.28**	**55.41**	72.70	**50.63**	**68.37**	**18.29**	**85.59**	**59.12**

Table A13. **Comparison of effects of different LLMs used in FedMVP on the Multi-source Single-target (MSST) Domain Generalization setting**. The results are reported for the PACS and OfficeHome datasets.

Method	PACS					OfficeHome				
	A. Painting	Cartoon	Photo	Sketch	Average	Art	Clipart	Product	RealWorld	Average
Llama-3.2-3B [10]	96.34	98.92	99.67	92.21	96.79	81.84	69.94	**91.84**	92.30	83.98
Qwen2.5-14B [36]	96.20	99.04	100.00	**93.08**	97.08	82.05	69.80	91.63	92.18	83.92
Phi-4 [1]	96.67	98.78	99.46	92.17	96.77	82.04	70.02	91.54	92.45	84.01
GPT-4o [2]	96.92	**99.35**	**100.00**	92.86	**97.28**	**82.20**	**70.05**	91.78	**92.56**	**84.15**

Table A14. **Comparison of methods on the Base-to-new generalization task in non-federated offline setting.**

Method	Sets	CLIP ICML21	CoOp IJCV22	CoCoOp CVPR22	VPT ECCV22	KgCoOp CVPR23	MaPLe CVPR23	PromptSRC ICCV23	StyLIP WACV24	CoPrompt ICLR24	TCP CVPR24	DePT CVPR24	DeKgTCP ICLR25	FedMVP -
Average	Base	69.34	82.69	80.47	82.11	80.73	82.28	84.12	83.22	84.00	84.13	**85.18**	84.96	85.00
	New	74.22	63.22	71.69	71.73	73.61	75.14	75.02	75.94	77.23	75.36	76.17	76.38	**77.85**
	H	71.70	71.66	75.83	76.57	77.00	78.55	79.31	79.41	80.48	79.51	80.42	80.44	**81.27**
ImageNet	Base	72.43	76.47	75.98	75.90	75.83	76.66	77.75	77.15	77.67	77.27	**78.20**	77.40	78.11
	New	68.14	67.88	70.43	68.10	69.96	70.54	70.70	71.34	71.27	69.87	70.27	69.20	**72.26**
	H	70.22	71.92	73.10	71.79	72.78	73.47	74.06	74.13	74.33	73.38	74.02	73.07	**75.07**
Caltech101	Base	96.84	98.00	97.96	98.03	97.72	97.74	98.13	98.23	98.27	98.23	98.57	**98.64**	98.48
	New	94.00	89.81	93.81	94.30	94.39	94.36	93.90	94.91	94.90	94.67	94.10	**95.20**	94.43
	H	95.40	93.73	95.84	96.13	96.03	96.02	95.97	96.54	96.55	96.42	96.28	**96.89**	96.41
OxfordPets	Base	91.17	93.67	95.20	95.13	94.65	95.43	95.50	**95.96**	95.67	94.67	95.43	94.47	95.70
	New	97.26	95.29	97.69	96.47	97.76	97.76	97.40	98.14	98.10	97.20	97.33	97.76	**98.45**
	H	94.12	94.47	96.43	95.80	96.18	96.58	96.44	97.04	96.87	95.92	96.37	96.09	**97.06**
StanfordCars	Base	63.37	78.12	70.49	71.63	71.76	72.94	78.40	75.19	76.97	80.80	80.80	**81.18**	80.95
	New	74.89	60.40	73.59	72.20	75.04	74.00	74.73	74.46	74.40	74.13	**75.00**	74.75	74.67
	H	68.65	68.13	72.01	71.92	73.36	73.47	75.52	74.82	75.66	77.32	77.79	**77.83**	77.68
Flowers102	Base	72.08	97.60	94.87	95.93	95.00	95.92	97.90	96.54	97.27	97.73	98.40	**98.58**	98.51
	New	77.80	59.67	71.75	70.37	74.73	72.46	76.77	73.08	76.60	75.57	77.10	75.18	**78.76**
	H	74.83	74.06	81.71	81.18	83.65	82.56	86.06	83.19	85.71	85.23	86.46	85.30	**87.53**
Food101	Base	90.10	88.33	90.70	89.80	90.50	90.71	90.63	91.20	90.73	90.57	90.87	90.73	**91.35**
	New	91.22	82.26	91.29	90.37	91.70	92.05	91.50	92.48	92.07	91.37	91.57	91.55	**93.04**
	H	90.66	85.19	90.99	90.08	91.09	91.38	91.06	91.84	91.40	90.97	91.22	91.14	**92.19**
FGVCAircraft	Base	27.19	40.44	33.41	35.90	36.21	37.44	42.30	37.65	40.20	41.97	**45.70**	45.20	42.38
	New	36.29	22.30	23.71	30.37	33.55	35.61	36.97	35.93	39.33	34.43	36.73	35.09	**39.82**
	H	31.09	28.75	27.74	32.90	34.83	36.50	39.46	36.77	39.76	37.83	40.73	39.51	**41.06**
SUN397	Base	69.36	80.60	79.74	79.50	80.29	80.82	82.83	82.12	82.63	82.63	83.27	82.52	**83.41**
	New	75.35	65.89	76.86	76.17	76.53	78.70	79.00	79.95	80.03	78.20	78.97	78.30	**79.50**
	H	72.23	72.51	78.27	77.80	78.36	79.75	80.87	81.02	81.31	80.35	81.06	80.35	**81.41**
DTD	Base	53.24	79.44	77.01	80.90	77.55	80.36	82.60	81.57	83.13	82.77	**84.80**	83.80	83.28
	New	59.90	41.18	56.00	52.73	54.99	59.18	57.50	61.72	**64.73**	58.07	61.20	59.66	61.94
	H	56.37	54.24	64.85	63.85	64.35	68.16	67.80	70.27	**72.79**	68.25	71.09	69.70	71.04
EuroSAT	Base	56.48	92.19	87.49	95.83	85.64	94.07	92.40	**94.61**	94.60	91.63	93.23	94.02	94.33
	New	64.05	54.74	60.04	65.03	64.34	73.23	68.43	74.06	78.57	74.73	77.90	**81.69**	81.59
	H	60.03	68.69	71.21	77.48	73.48	82.30	78.63	83.08	85.84	82.32	84.88	87.42	**87.50**
UCF101	Base	70.53	84.69	82.33	84.63	82.89	83.00	86.93	85.19	86.90	87.13	87.73	88.06	**88.46**
	New	77.50	56.05	73.45	72.90	76.67	78.66	78.33	79.22	79.57	80.77	77.70	81.77	**81.92**
	H	73.85	67.46	77.64	78.33	79.65	80.77	82.41	82.10	83.07	83.83	82.46	84.80	**85.06**

References

[1] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024. 7, 11

[2] 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. 1, 7, 11

[3] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018. 4

[4] Shirsha Bose, Ankit Jha, Enrico Fini, Mainak Singha, Elisa Ricci, and Biplab Banerjee. Stylip: Multi-scale style-conditioned prompt learning for clip-based domain generalization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5542–5552, 2024. 7

[5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 - mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 4

[6] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014. 4

[7] 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. 4

[8] Chen Fang, Ye Xu, and Daniel N. Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013. 4

[9] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Conference on Computer Vision and Pattern Recognition Workshop*, pages 178–178. IEEE, 2004. 4

[10] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 7, 11

[11] 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. 8, 9, 10, 11

[12] 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. 4

[13] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 4

[14] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15262–15271, 2021. 4

[15] 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. 7, 8, 9, 10, 11

[16] 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. 7, 8, 9, 10, 11

[17] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15190–15200, 2023. 7

[18] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. 4

[19] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017. 4

[20] 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. 8, 9, 10, 11

[21] Yilun Li, Miaomiao Cheng, Xu Han, and Wei Song. Divergence-enhanced knowledge-guided context optimization for visual-language prompt tuning. In *The Thirteenth International Conference on Learning Representations*, 2025. 7

[22] 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. 8, 9, 10, 11

[23] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *CoRR*, abs/1306.5151, 2013. 4

[24] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *International Conference on Learning Representations*, 2023. 1, 2

[25] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008*

Sixth Indian Conference on Computer Vision, Graphics & Image Processing, pages 722–729, 2008. 4

[26] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 4

[27] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 4

[28] 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. 4, 6, 8, 9, 10, 11

[29] 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, 7, 8, 9, 10, 11

[30] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 4

[31] Shuvendu Roy and Ali Etemad. Consistency-guided prompt learning for vision-language models. In *International Conference on Learning Representations*, 2024. 7

[32] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 4

[33] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017. 4

[34] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, 2019. 4

[35] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 4

[36] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 7, 11

[37] 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. 7, 8, 9, 10, 11

[38] Hantao Yao, Rui Zhang, and Changsheng Xu. Tcp: Textual-based class-aware prompt tuning for visual-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23438–23448, 2024. 7

[39] Ji Zhang, Shihan Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Dept: Decoupled prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12924–12933, 2024. 7

[40] 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. 7, 8, 9, 10, 11

[41] 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. 7