A Experiment Setup Details

A.1 Dataset and Hyper-parameter Details

We follow the settings in Zhou et al. (2022b) to conduct the experiments in this paper with the nine classification datasets for generalization on seen to unseen classes, and four variants of ImageNet datasets for domain shifting, where the statistical details are presented in Table 5.

For each compared FL approach and each classification task, via grid search, the learning rate of the SGD optimizer was set to $\eta = 0.003$ with a decay rate $1\mathrm{e}{-5}$ and a momentum of 0.9. The local SGD training step is set to $K = 1$. By default, all the experimental results in the paper are obtained by averaging from three runs with different random seeds.

A.2 Federated Learning Setup Details

Experimental Setup for Seen and Unseen Classes in Table 1 To evaluate the generalization ability for the proposed FedTPG and compared FL approaches from in the paper, we monitor the model performance on the following three benchmark accuracies: (1) The local classification accuracy, representing the performance of local clients' classification tasks on local available classes; (2) The base classification accuracy, representing the performance against all seen classes (combining classes from multiple clients) in a dataset in the FL network; (3) The new classification accuracy, which indicates the performance on unseen classes but within the domain of seen classes. We report the harmonic mean (HM) of these three accuracies on each classification task, as shown in Table 1.

In the FL data partition process for Table 1, we first split the classes of the considered 9 classification datasets equally into two groups $\mathcal{D}^s$ and $\mathcal{D}^u$, denotes seen and unseen groups respectively. Then we split the classes within $\mathcal{D}^s$ to the 30 remote clients, where each remote client has $n = 20$ classes in each local dataset $\mathcal{D}_i$. For each class, the number of image-text paired data shots is set to 8. During the FL training process, the participation rate of remote clients is set to 100% and the communication round is set to 500.

Experimental Setup for Unseen Datasets in Table 2 and Table 3 To evaluate the generalization ability of FedTPG on unseen datasets during training, we consider the following two settings: (1) Domain Shifting, where we monitor the performance of model by training with ImageNet and testing on four variants of ImageNet, including ImageNetV2, ImageNet-Sketch, ImageNet-A, and ImageNet-R; (2) Unseen Datasets, where we evaluate the performance of trained model in (1) on nine unseen datasets, including Caltech101, OxfordPets, StanfordCars, Flowers102, Food101, FGVCAircraft, SUN397, UCF101, and DTD. During the training process, we set the FL network with 200 remote clients where each client has $n = 5$ classes of 8-shots training data disjointly. The participation rate of remote clients is set to 10% that $|\mathcal{S}^r| = 20$ and the global communication round is set to $R = 500$ to obtain θ^R.

Experimental Setup for Ablation Study in Table 4 and Figure 4 We study the impact of the number of classes owned by each client at Table 4 from the introduced local, base and new classification accuracies with the same setup in Table 1 where a full client participation is performed with $R = 500$ and number of shots is 8. Specifically, we perform the data partition with the disjoint rule during class splitting: when $n = 5$, we set the number of clients to 119; when $n = 10$, we set the number of clients to 59; and when $n = 20$, we set the number of clients to 20, respectively.

The study of the number of shots is shown in Figure 4(b), where we set the number of clients to 30 with $n = 20$ and the client participation rate is 100% in each round where $R = 500$. The study of the participation rate is shown in Figure 4(b), where we set the number of clients to 30 with $n = 20$ and the number of shots is 8.

Then, we monitor the impact of the FL client participation rate in each communication round as shown in Figure 4(a). We formulate the FL network with 30 clients where $n = 20$ and the number of shots is 8. Four client participation rates in $\{10\%, 40\%, 70\%, 100\}\%$ are considered during the model training process with $R = 500$.

Table 5: Dataset statistical details on class, training and test splits, prompt template.

Dataset	Classes	Train	Test	Hand-crafted prompt template
ImageNet	1000	1.28M	50,000	A photo of a **[class]**
Caltech101	101	4,128	2,465	A photo of a **[class]**
Flowers102	102	4,093	2,463	A photo of a **[class]**, a type of flower
FGVCAircraft	100	3,334	3,333	A photo of a **[class]**, a type of aircraft
UCF101	101	7,639	3,783	A photo of a person doing **[class]**
OxfordPets	37	2,944	3,369	A photo of a **[class]**, a type of pet
Food101	101	50,500	30,300	A photo of a **[class]**, a type of food
DTD	47	2,820	1,692	A photo of a **[class]**, a type of texture
StanfordCars	196	6,509	8,041	A photo of a **[class]**
SUN397	397	15,880	19,850	A photo of a **[class]**
ImageNetV2	1000	N/A	10,000	A photo of a **[class]**
ImageNet-Sketch	1000	N/A	50,889	A photo of a **[class]**
ImageNet-A	200	N/A	7500	A photo of a **[class]**
ImageNet-R	200	N/A	30,000	A photo of a **[class]**

B ADDITIONAL RESULTS

Table 6 and Table 7 show the detailed results of FedTPG and the compared FL baselines on the benchmark of seen and unseen classes with $n = 5$ and $n = 10$, respectively. The results of Table 6 and Table 7 are the detailed results of Table 4 in the main paper, where we would like to claim that the HM results in the main paper are the harmonic mean of the base accuracy and the new accuracy, while the results in Table 6 and Table 7 are the harmonic mean of the local accuracy, the base accuracy and the new accuracy that leads to the difference in some columns.

The results show that similar to the results of $n = 20$ in Table 1, the proposed FedTPG achieves the best average accuracy on unseen classes, and achieves the best new performance for 3 tasks while the second best new performance for most of the other tasks. We can also observe that as n increases, the advantage of FedTPG against other approaches becomes more significant. This supports our theoretical claim that the unified prompt generator in FedTPG generalizes better on unobserved classification tasks, especially for challenging scenarios.

Table 6: Accuracies (%) on clients' local tasks (seen), base (seen) classes, and new (unseen) classes. Each client has labeled images from five disjoint classes. The number of shot is 8 and $n = 5$.

(a) Average over 9 datasets.

	Local	Base	New	HM
CLIP	86.25	70.52	75.78	76.98
FedCoOp	89.38	69.53	70.05	74.74
FedKgCoOp	86.63	70.83	75.55	77.12
FedTPG	87.78	71.08	75.51	**77.51**

(b) Caltech101.

	Local	Base	New	HM
CLIP	97.40	96.97	93.89	96.06
FedCoOp	97.19	93.67	92.14	94.28
FedKgCoOp	97.95	96.57	94.21	**96.22**
FedTPG	97.31	94.00	94.43	95.22

(c) Flowers102.

	Local	Base	New	HM
CLIP	91.12	72.18	77.94	**79.66**
FedCoOp	97.89	70.65	74.47	79.37
FedKgCoOp	89.96	70.27	76.51	78.09
FedTPG	94.20	70.23	76.77	79.20

(d) FGVCAircraft.

	Local	Base	New	HM
CLIP	49.04	27.55	35.81	35.45
FedCoOp	55.82	25.45	26.57	31.63
FedKgCoOp	51.98	28.89	33.75	**35.93**
FedTPG	53.62	26.38	33.92	34.87

(e) UCF101.

	Local	Base	New	HM
CLIP	88.78	70.58	77.50	**78.25**
FedCoOp	90.71	69.75	65.33	73.77
FedKgCoOp	87.68	70.06	76.14	77.29
FedTPG	88.53	71.20	75.96	77.91

(f) OxfordPets.

	Local	Base	New	HM
CLIP	96.75	91.33	97.04	94.96
FedCoOp	98.08	91.92	94.57	94.79
FedKgCoOp	96.65	91.34	96.16	94.66
FedTPG	97.96	91.39	96.03	**95.04**

(g) Foods102.

	Local	Base	New	HM
CLIP	97.57	90.16	91.25	**92.88**
FedCoOp	97.17	88.27	86.67	90.48
FedKgCoOp	97.42	89.59	91.52	92.72
FedTPG	97.34	89.24	91.31	92.51

(h) DTD.

	Local	Base	New	HM
CLIP	79.55	53.01	58.21	61.71
FedCoOp	86.94	54.40	51.45	60.83
FedKgCoOp	80.50	55.47	60.26	63.77
FedTPG	82.72	60.19	61.53	**66.73**

(i) StanfordCars.

	Local	Base	New	HM
CLIP	83.06	63.44	74.90	72.90
FedCoOp	86.06	64.84	71.77	73.22
FedKgCoOp	83.42	63.84	75.85	**73.46**
FedTPG	83.75	63.92	72.35	72.45

(j) SUN397.

	Local	Base	New	HM
CLIP	93.02	69.41	75.46	78.10
FedCoOp	94.55	66.83	67.44	74.32
FedKgCoOp	94.12	71.45	75.52	79.23
FedTPG	94.56	73.17	77.24	**80.67**

Table 7: Accuracies ($\%$) on clients' local tasks (seen), base (seen) classes, and new (unseen) classes. Each client has labeled images from ten disjoint classes. The number of shot is 8 and $n = 10$.

(a) Average over 9 datasets.

	Local	Base	New	HM
CLIP	80.57	70.52	75.78	75.40
FedCoOp	85.64	72.15	70.61	75.57
FedKgCoOp	81.39	71.18	75.81	75.90
FedTPG	83.49	72.17	75.84	**76.89**

(b) Caltech101.

	Local	Base	New	HM
CLIP	97.83	96.97	93.89	**96.20**
FedCoOp	97.45	94.56	93.46	95.13
FedKgCoOp	97.64	96.80	93.99	96.12
FedTPG	98.03	95.83	94.58	96.13

(c) Flowers102.

	Local	Base	New	HM
CLIP	84.58	72.18	77.94	77.91
FedCoOp	97.17	73.33	71.10	**78.96**
FedKgCoOp	84.77	71.93	76.80	77.48
FedTPG	90.03	71.58	77.08	78.85

(d) FGVCAircraft.

	Local	Base	New	HM
CLIP	37.88	27.55	35.81	33.10
FedCoOp	44.00	27.23	25.76	30.53
FedKgCoOp	38.53	26.86	35.06	32.71
FedTPG	41.74	28.44	35.05	**34.21**

(e) UCF101.

	Local	Base	New	HM
CLIP	83.65	70.58	77.5	76.87
FedCoOp	87.56	73.53	71.76	77.01
FedKgCoOp	84.00	71.25	76.11	76.77
FedTPG	85.78	72.15	76.05	**77.59**

(f) OxfordPets.

	Local	Base	New	HM
CLIP	93.26	91.33	97.04	93.82
FedCoOp	95.95	92.36	91.60	93.27
FedKgCoOp	92.55	90.32	96.36	93.01
FedTPG	95.86	93.92	96.73	**95.48**

(g) Foods102.

	Local	Base	New	HM
CLIP	95.94	90.16	91.25	**92.38**
FedCoOp	95.18	88.21	89.91	90.72
FedKgCoOp	95.81	89.88	91.66	**92.38**
FedTPG	95.73	89.93	91.63	92.36

(h) DTD.

	Local	Base	New	HM
CLIP	62.74	53.01	58.21	57.71
FedCoOp	78.15	63.11	49.65	61.50
FedKgCoOp	68.10	57.12	60.26	61.49
FedTPG	71.41	59.52	60.18	**63.26**

(i) StanfordCars.

	Local	Base	New	HM
CLIP	78.29	63.44	74.9	71.62
FedCoOp	81.23	65.76	70.93	72.09
FedKgCoOp	78.82	64.13	75.52	72.25
FedTPG	80.15	65.33	74.62	**72.84**

(j) SUN397.

	Local	Base	New	HM
CLIP	90.96	69.41	75.46	77.61
FedCoOp	94.07	71.32	72.10	77.88
FedKgCoOp	92.28	72.36	76.47	79.51
FedTPG	92.71	72.90	76.62	**79.88**