pFEDVSA: PERSONALIZED FEDERATED VISUAL-TO-SEMANTIC ALIGNMENT FOR CLIP ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated few-shot CLIP learning aims to exploit client-specific visual evidence for personalization while preserving transferable recognition over global base classes and unseen novel classes. A key challenge is how to use client-specific class-level visual evidence for collaborative adaptation without directly sharing such evidence, especially when local visual priors can be sparse or unreliable. We propose Personalized Federated Visual-to-Semantic Alignment (pFedVSA), which separates private visual personalization from shared semantic adaptation. Each client maintains a personalized image adapter and a Local Visual Prior Memory (LVPM) that remain local throughout federated training. A shared Prototype-to-Semantic Transfer (PST) module learns to map locally available visual priors into CLIP-compatible semantic pseudo-tokens, enabling collaborative visual-to-semantic adaptation without aggregating the priors themselves. Semantic Alignment in CLIP Space (SACS) further incorporates prototype-conditioned evidence through reliability-controlled residual calibration based on prior support and image–prior consistency. For classes without local visual support, prediction falls back to the standard adapted semantic branch. Extensive experiments on eleven datasets under label and feature heterogeneity show that pFedVSA consistently improves the balance among client-specific recognition, global base-class transfer, and novel-class generalization.

1 INTRODUCTION

Vision-Language Models (VLMs), such as CLIP Radford et al. (2021), enable open-vocabulary recognition by aligning visual and textual representations in a shared semantic space Jia et al. (2021); Li et al. (2022); Yao et al. (2021); Zhang et al. (2022); Zhu et al. (2023). Their strong zero-shot and few-shot generalization makes them attractive for data-scarce applications, where lightweight prompt tuning, adapters, and LoRA can adapt pretrained models efficiently Jia et al. (2022); Chen et al. (2022); Hu et al. (2021). In many practical scenarios, however, downstream data are distributed across multiple clients and cannot be centralized, making Federated Learning (FL) a natural paradigm for collaborative CLIP adaptation McMahan et al. (2017). Federated few-shot CLIP learning is particularly challenging because clients may simultaneously exhibit label and feature heterogeneity: each client may observe only a sparse subset of classes, while its images can differ in domain, style, and acquisition conditions Li et al. (2021b;a). A successful method must therefore exploit client-specific observations for personalization while preserving transferable recognition over global base classes and unseen novel classes.

Existing federated VLM adaptation methods address this tension through global or personalized prompts, prompt generation, client-specific experts, and local–global adapter decomposition Guo et al. (2023); Qiu et al. (2023); Cui et al. (2024); Li et al. (2024); Luo et al. (2025); Ghiasvand et al. (2025). Recent studies have further explored richer cross-modal interactions and prototype-level knowledge; for example, FedSPA investigates semantic–visual prototype alignment for federated adaptation Yang & Liu (2026). Despite this progress, effectively exploiting client-specific class-level visual evidence remains challenging. Such evidence is valuable for personalization but is also tied to local appearance and domain characteristics, making direct sharing susceptible to client-specific bias, while keeping it entirely local limits cross-client knowledge transfer. Moreover, visual priors estimated from few observations can be unreliable, and their relevance may vary across query samples. This raises a more specific question: *how can private class-level visual evidence contribute*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

to collaborative semantic adaptation without directly exchanging the evidence itself, and how should its influence be controlled when the evidence is sparse or inconsistent?

We address this problem by separating private visual personalization from shared semantic adaptation. Instead of globally sharing client-specific visual priors, each client retains them locally and uses them to optimize a shared transformation into CLIP's semantic space. Cross-client transfer is therefore realized through a collaboratively learned visual-to-semantic mapping rather than direct aggregation of class prototypes or personalized visual states. At the same time, the contribution of local evidence should depend on both the amount of supporting data and its consistency with the current image, preventing unreliable priors from dominating semantic prediction.

Based on this principle, we propose **pFedVSA**, a personalized federated visual-to-semantic alignment framework for CLIP adaptation. Each client maintains a personalized image adapter and a Local Visual Prior Memory (LVPM), both of which remain client-private. A shared Prototype-to-Semantic Transfer (PST) module maps available class-level visual priors into CLIP-compatible semantic pseudo-tokens, allowing heterogeneous local evidence to supervise a collaboratively learned visual-to-semantic transformation without uploading the priors themselves. A shared text adapter further coordinates semantic adaptation across clients, while Semantic Alignment in CLIP Space (SACS) incorporates prototype-conditioned semantic evidence as a bounded residual according to prior support and image–prototype consistency. During federated training, only PST and the text adapter are uploaded and aggregated, whereas the personalized image adapter and LVPM remain local. For locally supported classes, SACS uses prototype-conditioned semantic evidence to calibrate the standard adapted prediction; for locally unsupported base classes and all novel classes, no local visual prior is used and the model falls back to the standard adapted semantic branch. Therefore, prototype-conditioned evidence directly affects supported-class prediction, while its influence on unsupported and novel classes is mediated through the semantic modules optimized during federated training. Semantic anchoring and knowledge distillation further constrain this adaptation to preserve CLIP's transferable semantic structure. Extensive experiments on eleven benchmarks covering base-to-new generalization, label heterogeneity, and feature heterogeneity show that pFedVSA achieves a strong balance among client-specific recognition, global base-class transfer, and novel-class generalization, with additional analyses examining visual-prior construction, semantic transfer, reliability calibration, and module sharing.

Our contributions are summarized as follows:

- We propose **pFedVSA**, a personalized federated CLIP adaptation framework that separates private visual personalization from collaboratively learned semantic adaptation. Client-specific image adapters and class-level visual priors remain local, while lightweight semantic modules are shared across clients.
- We introduce **Prototype-to-Semantic Transfer (PST)**, which learns a shared mapping from private class-level visual priors to CLIP-compatible semantic signals, and **Semantic Alignment in CLIP Space (SACS)**, which incorporates locally available semantic evidence through support- and consistency-aware bounded residual calibration.
- We conduct extensive experiments across eleven benchmarks covering base-to-new generalization, label heterogeneity, and feature heterogeneity. The results demonstrate consistent improvements in balancing client-specific recognition, global base-class transfer, and novel-class generalization, together with detailed analyses of the proposed visual-prior and semantic-alignment designs.

2 Related Work

CLIP Adaptation and Base-to-Novel Generalization. Large-scale vision-language models such as CLIP Radford et al. (2021) learn transferable visual–textual representations that support strong zero-shot recognition. Parameter-efficient adaptation methods seek to specialize these representations to downstream tasks while preserving their pretrained generalization ability. CoOp Zhou et al. (2022b) replaces hand-crafted prompts with learnable context vectors, but the resulting prompts may overfit the observed classes. CoCoOp Zhou et al. (2022a) introduces instance-conditioned prompts to improve generalization to unseen categories, while MaPLe Khattak et al. (2023) jointly adapts the visual and textual branches for stronger cross-modal alignment. Feature- and adapter-based ap-

proaches provide complementary alternatives for lightweight CLIP adaptation Chen et al. (2022); Zhang et al. (2022). These studies establish an important trade-off between task-specific adaptation and preservation of CLIP's transferable semantic structure. However, they assume centralized training and do not address the additional client heterogeneity introduced by federated learning.

Personalized Federated VLM Adaptation. Federated VLM adaptation extends parameter-efficient CLIP adaptation to decentralized and heterogeneous client data. PromptFL Guo et al. (2023) collaboratively learns prompts over a frozen CLIP backbone, while FedTPG Qiu et al. (2023) introduces a federated prompt generator for transfer across seen and unseen classes. FedPGP Cui et al. (2024) combines global and personalized prompts to balance cross-client generalization and client-specific adaptation, and FedOTP Li et al. (2024) coordinates global and local prompts through cross-modal alignment. Adapter-based approaches further provide explicit mechanisms for personalization. FedCLIP Lu et al. (2023) introduces lightweight visual adaptation, pFedMoAP Luo et al. (2025) combines local prompt experts through a gating mechanism, and pFedMMA Ghiasvand et al. (2025) decomposes multimodal adapters into client-specific and globally shared components. Together, these approaches demonstrate that federated VLM adaptation benefits from carefully coordinating globally transferable and client-specific model components. Nevertheless, how client-specific visual evidence itself should participate in collaborative semantic learning remains a distinct design problem.

Visual Prototype and Cross-Modal Knowledge Transfer. A growing line of work moves beyond purely parameter-level personalization by explicitly incorporating visual information into cross-modal adaptation. FedMVP Singha et al. (2025) conditions federated prompt learning on image-derived visual context and class-level textual information through cross-modal interaction. FedSPA Yang & Liu (2026) further investigates semantic–visual prototype alignment for federated adaptation. These methods highlight the value of visual context and prototype-level knowledge in federated vision-language learning, but they also raise two questions that are central to our setting. First, class-level visual evidence can encode client-specific appearance and domain characteristics, making it desirable to exploit such information without directly treating the visual priors themselves as globally shared states. Second, locally estimated visual priors may be unreliable when class support is limited or when the current image is inconsistent with the stored prototype, requiring explicit control over their influence on prediction. pFedVSA addresses these issues by keeping client-specific image adapters and class-level visual priors local, learning a shared Prototype-to-Semantic Transfer module that maps local visual evidence into CLIP-compatible semantic signals, and using SACS to incorporate such signals through support- and consistency-aware bounded residual calibration.

3 METHODOLOGY

3.1 OVERVIEW

We propose pFedVSA, a personalized federated visual-to-semantic alignment framework for CLIP adaptation. The key challenge is to exploit client-specific visual evidence for personalization while preserving CLIP's transferable semantic space. To this end, pFedVSA adopts an asymmetric private-visual and shared-semantic design built upon Prototype-to-Semantic Transfer (PST) and Semantic Alignment in CLIP Space (SACS). Each client retains a personalized image adapter and a Local Visual Prior Memory (LVPM) that capture client-specific visual characteristics. PST learns a shared mapping from locally available class-level visual priors to CLIP-compatible semantic pseudo-tokens, without exchanging the visual priors themselves. SACS further incorporates prototype-conditioned evidence as a bounded, reliability-controlled residual to the adapted semantic prediction. For locally unsupported base classes and all novel classes, no local visual prior is used and prediction falls back to the standard adapted semantic branch. During federated training, only PST and the text adapter are uploaded and aggregated, whereas the image adapter and LVPM remain client-specific. The overall framework is illustrated in Figure 1.

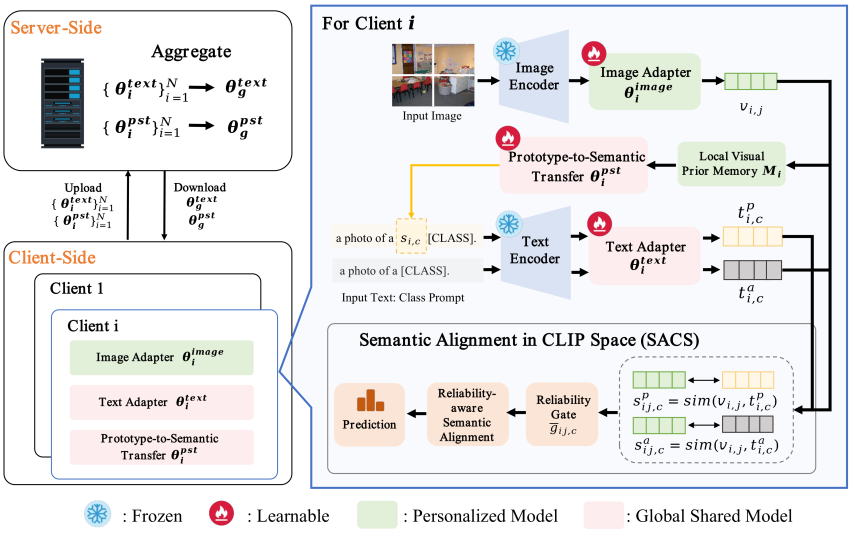

Figure 1: Overview of pFedVSA. Each client maintains a personalized image adapter and a LVPM to capture client-specific visual evidence. PST maps locally available visual priors into CLIP-compatible semantic pseudo-tokens, while the shared text adapter coordinates semantic adaptation across clients. SACS integrates the prototype-conditioned evidence through reliability-controlled residual calibration. During federated training, only PST and text-adapter parameters are uploaded and aggregated, whereas the image adapter and LVPM remain local to each client.

3.2 Private Visual Prior and Prototype-to-Semantic Transfer

Let f_{image} denote the frozen CLIP image encoder equipped with client i's personalized image adapter $\theta_i^{\text{image},r}$. For a local image $x_{i,j}$, its normalized client-aware representation is $v_{i,j}^r = \text{Norm}\big(f_{\text{image}}(x_{i,j}; \theta_i^{\text{image},r})\big)$. The image adapter remains private throughout federated training.

Local visual prior memory. At the beginning of communication round r, client i reconstructs its LVPM from unique local training samples. Let $\mathcal{Y}_i^{\text{obs}}$ denote the set of base classes observed by client i. For each $c \in \mathcal{Y}_i^{\text{obs}}$, let $\mathcal{D}_{i,c}^{\text{tr}} = \{(x_{i,j}, y_{i,j}) \in \mathcal{D}_i^{\text{tr}} \mid y_{i,j} = c\}$ and $N_{i,c} = |\mathcal{D}_{i,c}^{\text{tr}}|$. The corresponding class-level visual prior is

$$p_{i,c}^r = \text{Norm}\left(\frac{1}{N_{i,c}} \sum_{(x_{i,j}, y_{i,j}) \in \mathcal{D}_{i,c}^{\text{tr}}} \text{sg}\left(v_{i,j}^r\right)\right), \tag{1}$$

where $\text{sg}(\cdot)$ denotes stop-gradient and $n_{i,c} = N_{i,c}$ records the number of unique support samples. The resulting memory is $M_i^r = \{(p_{i,c}^r, n_{i,c}) \mid c \in \mathcal{Y}_i^{\text{obs}}\}$. It is reconstructed before local optimization, kept fixed within the current round, and never uploaded. This round-wise lifecycle avoids repeated-count inflation and stale representations from earlier image-adapter states.

Prototype-to-semantic transfer. Directly using local visual priors as classifiers may preserve client-specific visual biases and does not explicitly align them with CLIP's language space. PST therefore maps each available local visual prior into a text-compatible semantic pseudo-token:

$$s_{i,c}^r = g_{\text{pst}}\left(p_{i,c}^r; \theta_i^{\text{pst},r}\right), \qquad c \in \mathcal{Y}_i^{\text{obs}}, \tag{2}$$

where $\theta_i^{\text{pst},r}$ is the client-side copy of the globally shared PST parameters. The resulting token is inserted into the prototype-conditioned prompt $\pi_{i,c}^p =$ "a photo of a $[s_{i,c}^r]$ [CLASS].", while the standard branch uses $\pi_c^a =$ "a photo of a [CLASS].". Both prompts are processed by the frozen CLIP text encoder equipped with the shared text adapter θ_i^{text}, producing $t_{i,c}^p$ and $t_{i,c}^a$, respectively.

The former carries semantic evidence derived from the local visual prior, whereas the latter provides the standard adapted class semantics.

Only PST and the text-adapter parameters are aggregated across clients. Thus, federated learning shares the visual-to-semantic mapping rather than the class-level visual priors themselves.

3.3 Reliability-Controlled Semantic Alignment

For a locally supported class $c \in \mathcal{Y}_i^{\text{obs}}$, SACS computes the prototype-conditioned and standard adapted similarities as

$$s_{ij,c}^{p} = \text{sim}\left(v_{i,j}, t_{i,c}^{p}\right), \qquad s_{ij,c}^{a} = \text{sim}\left(v_{i,j}, t_{i,c}^{a}\right). \tag{3}$$

Since local visual priors may be sparse or inconsistent with the current image, SACS evaluates their reliability using both support sufficiency and image–prior consistency:

$$g_{ij,c} = \frac{n_{i,c}}{n_{i,c}+k} \cdot \sigma\left(\gamma\left[\text{sim}(v_{i,j}, p_{i,c}) - \tau_g\right]\right), \qquad c \in \mathcal{Y}_i^{\text{obs}}, \tag{4}$$

and bounds it as $\bar{g}_{ij,c} = \text{clip}(g_{ij,c}, 0, g_{\max})$. Here, k controls the support scale, γ and τ_g control the similarity response, and $g_{\max}$ limits the maximum contribution of local evidence. The stored visual prior remains a no-gradient buffer.

The final prediction is defined according to whether a local visual prior is available:

$$\ell_{ij,c} = \begin{cases} \alpha\left[s_{ij,c}^{a} + \bar{g}_{ij,c}\left(s_{ij,c}^{p} - s_{ij,c}^{a}\right)\right], & c \in \mathcal{Y}_i^{\text{obs}}, \\ \alpha s_{ij,c}^{a}, & c \notin \mathcal{Y}_i^{\text{obs}}, \end{cases} \tag{5}$$

where α is the CLIP logit scale. Thus, prototype-conditioned evidence acts as a bounded correction for locally supported classes, while locally unsupported base classes use only the standard adapted branch. Novel classes are excluded from federated training and have no local visual prior, so they also use the standard adapted branch at inference. Accordingly, PST and SACS directly calibrate supported classes, whereas their effect on novel classes can only arise indirectly through the semantic modules optimized during training.

Training objective. Let $\mathcal{Y}^{\text{tr}} = \mathcal{Y}^{\text{base}}$ denote the global training vocabulary. The local objective is

$$\mathcal{L}_i = \mathcal{L}_{\text{ce}} + \lambda_{\text{anc}}\mathcal{L}_{\text{anc}} + \lambda_{\text{kd}}\mathcal{L}_{\text{kd}}. \tag{6}$$

The classification loss is

$$\mathcal{L}_{\text{ce}} = -\frac{1}{|\mathcal{B}_i|} \sum_{(x_{i,j}, y_{i,j}) \in \mathcal{B}_i} \log \frac{\exp(\ell_{ij,y_{i,j}})}{\sum_{c \in \mathcal{Y}^{\text{tr}}} \exp(\ell_{ij,c})}. \tag{7}$$

For locally supported classes, the semantic anchoring loss is

$$\mathcal{L}_{\text{anc}} = \frac{1}{|\mathcal{Y}_i^{\text{obs}}|} \sum_{c \in \mathcal{Y}_i^{\text{obs}}} \left[1 - \text{sim}\left(t_{i,c}^{p}, \text{sg}(t_{i,c}^{a})\right)\right]. \tag{8}$$

This term prevents the prototype-conditioned representation from drifting excessively from the adapted class semantics. We further use the original frozen CLIP prediction as the distillation teacher. Let $s_{ij,c}^{0} = \text{sim}(v_{i,j}^{0}, t_c^{0})$ denote the original CLIP similarity. The distillation loss is

$$\mathcal{L}_{\text{kd}} = T^2 \text{KL}\left(\text{sg}\left[\text{softmax}\left(\frac{\alpha s^0}{T}\right)\right] \,\middle\|\, \text{softmax}\left(\frac{\ell}{T}\right)\right), \tag{9}$$

Method	SUN397				Flowers102				DTD				OxfordPets			
	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM
CLIP	69.41	69.38	75.52	71.32	67.89	69.23	76.88	71.12	54.26	54.86	59.18	56.02	89.45	89.42	96.81	91.77
PromptFL	77.73	77.71	72.96	76.07	97.37	97.06	63.62	82.66	80.23	80.21	45.29	63.81	96.35	96.28	97.26	96.63
FedCLIP-Adapter	72.82	72.78	77.22	74.22	71.46	73.22	77.38	73.94	56.99	55.90	59.78	57.51	93.01	92.93	97.09	94.30
FedCLIP-LoRA	74.67	80.46	72.28	75.65	76.77	**97.72**	68.16	79.09	68.16	81.94	59.42	68.64	90.71	95.64	97.71	94.59
FedPGP	94.29	54.88	57.76	65.02	99.67	72.44	58.65	73.37	89.03	71.03	50.94	66.75	96.62	95.17	97.15	96.31
pFedMoAP	**95.93**	31.18	35.40	42.41	99.81	43.70	48.37	55.99	96.43	53.60	48.21	60.28	99.92	77.61	92.05	88.87
pFedMMA	94.06	70.99	76.37	79.34	95.58	71.54	76.00	79.79	97.45	55.44	61.55	67.35	**100.00**	88.50	96.60	94.78
pFedVSA (Ours)	95.40	78.82	**80.28**	**84.21**	**100.00**	95.68	**82.45**	**92.08**	**99.49**	**75.33**	**65.19**	**77.59**	**100.00**	**95.45**	**98.83**	**98.05**

Method	Caltech101				Food101				UCF101				Average on 7 datasets			
	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM
CLIP	96.14	97.22	94.21	95.84	89.40	89.42	90.70	89.84	68.00	68.15	75.18	70.29	76.36	76.81	81.21	78.07
PromptFL	97.77	98.19	92.58	96.11	90.48	90.50	91.37	90.78	82.57	82.73	64.47	75.55	88.93	88.95	75.36	83.90
FedCLIP-Adapter	96.28	97.35	94.10	95.89	90.07	90.10	91.19	90.45	72.13	72.23	77.88	73.98	78.97	79.22	82.09	80.07
FedCLIP-LoRA	96.76	97.93	94.43	96.35	89.18	90.32	91.49	90.32	75.58	**87.13**	74.80	78.79	81.69	90.16	79.76	83.63
FedPGP	99.42	94.94	90.88	94.95	93.70	86.38	87.14	88.96	94.94	60.62	59.26	68.33	95.38	76.49	71.68	79.98
pFedMoAP	99.92	94.07	92.43	95.37	97.49	69.86	83.51	82.09	95.79	62.74	66.23	72.33	97.89	61.82	66.60	72.45
pFedMMA	**100.00**	96.53	94.29	96.88	97.45	89.15	90.77	92.32	95.63	69.61	74.88	78.58	97.17	77.40	81.49	84.55
pFedVSA (Ours)	**100.00**	**98.33**	**96.19**	**98.15**	**98.16**	**91.21**	**94.76**	**94.62**	**97.48**	76.63	**80.52**	**83.97**	**98.65**	**87.35**	**85.46**	**90.13**

Table 1: Top-1 accuracy (%) across seven datasets in the 16-shot setting. HM denotes the harmonic mean of Local, Base, and Novel accuracy. For the seven-dataset average, Local, Base, and Novel are first averaged across datasets, after which HM is computed.



(a) CIFAR-10 (b) CIFAR-100

Figure 2: Accuracy comparison (%) on the Dirichlet Non-IID setting across different β values. Subfigure (a) shows the results on CIFAR-10, and subfigure (b) shows the results on CIFAR-100. The horizontal axis denotes the Dirichlet concentration parameter β, and the vertical axis denotes classification accuracy (%).

where T is the distillation temperature and both distributions are normalized over $\mathcal{Y}^{\mathrm{tr}}$. Together, semantic anchoring and knowledge distillation constrain semantic drift and preserve CLIP's transferable structure during local adaptation.

Federated optimization. At round r, the server broadcasts $\Theta_g^r = \{\theta_g^{\mathrm{text},r}, \theta_g^{\mathrm{pst},r}\}$. Client i initializes its shared modules from Θ_g^r, combines them with the private image adapter $\theta_i^{\mathrm{image},r}$, and reconstructs M_i^r. The image adapter, PST, and text adapter are optimized locally, while LVPM remains fixed throughout the round. After local training, client i retains its image adapter and LVPM and uploads only $\Theta_i^{r+1} = \{\theta_i^{\mathrm{text},r+1}, \theta_i^{\mathrm{pst},r+1}\}$. For $m \in \{\mathrm{text}, \mathrm{pst}\}$, the server performs

$$\theta_g^{m,r+1} = \sum_{i \in \mathcal{S}_r} \frac{N_i}{\sum_{k \in \mathcal{S}_r} N_k} \theta_i^{m,r+1}, \tag{10}$$

where $N_i = |\mathcal{D}_i^{\mathrm{tr}}|$ and $\mathcal{S}_r$ denotes the selected-client set. Therefore, visual personalization remains client-specific, while cross-client semantic learning is realized through the aggregated PST and text adapter.

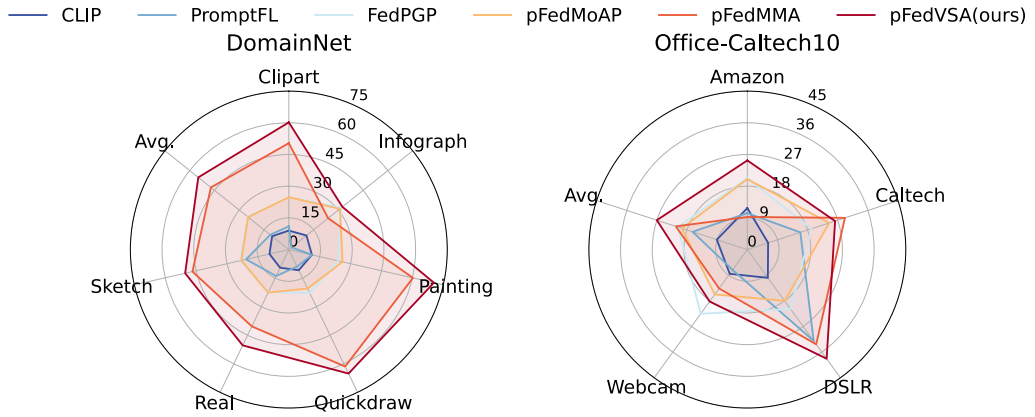

Figure 3: Radar plot of test accuracy (%) for different methods on DomainNet and Office-Caltech10 with Dirichlet partitioning ($\beta = 0.5$). The left radar chart corresponds to DomainNet with 7 domains (including average), and the right radar chart corresponds to Office-Caltech10 with 5 domains (including average). Each line represents one method.

4 Experiments

4.1 Experimental Setup

Datasets and federated splits. We evaluate pFedVSA on eleven public benchmarks covering base-to-new generalization, cross-domain federated adaptation, and personalized federated classification. We follow the dataset partitions and federated evaluation protocols commonly adopted in prior personalized federated CLIP studies Ghiasvand et al. (2025). For base-to-new generalization, we use seven visual recognition datasets: SUN397 Xiao et al. (2010), OxfordPets Parkhi et al. (2012), Flowers102 Nilsback & Zisserman (2008), DTD Cimpoi et al. (2014), Caltech101 Fei-Fei et al. (2004), UCF101 Soomro et al. (2012), and Food101 Bossard et al. (2014). Training is performed on base classes, while evaluation considers client-observed Local classes, global Base classes, and unseen Novel classes. For cross-domain federated adaptation, we use DomainNet Peng et al. (2019) and Office-Caltech10 Gong et al. (2012), where different domains exhibit distinct visual distributions. For personalized federated classification, we use CIFAR-10 Krizhevsky et al. (2010) and CIFAR-100 Krizhevsky et al. (2009) with Dirichlet label-skew partitions.

Baseline methods. We compare pFedVSA with zero-shot CLIP Radford et al. (2021) and representative federated CLIP adaptation methods, including PromptFL Guo et al. (2023), FedPGP Cui et al. (2024), pFedMoAP Luo et al. (2025), FedCLIP-Adapter Gao et al. (2024), FedCLIP-LoRA Zanella & Ben Ayed (2024), and pFedMMA Ghiasvand et al. (2025).

Evaluation label spaces and metrics. For base-to-new evaluation, the class set is divided into disjoint base and novel subsets, denoted by $\mathcal{Y}^{\text{base}}$ and $\mathcal{Y}^{\text{novel}}$. Only base classes are used for federated training, while novel classes are completely held out from all clients. We report three top-1 accuracies under different candidate label spaces. Local evaluates each client on its observed class subset $\mathcal{Y}_i^{\text{obs}}$; Base evaluates over the complete global base-class space $\mathcal{Y}^{\text{base}}$; and Novel evaluates over the held-out novel-class space $\mathcal{Y}^{\text{novel}}$. We summarize the overall balance using $\text{HM} = 3/(\text{Local}^{-1} + \text{Base}^{-1} + \text{Novel}^{-1})$. For analyses focusing on transferable recognition, we additionally report the harmonic mean of Base and Novel accuracy, $H_{\text{BN}} = 2/(\text{Base}^{-1} + \text{Novel}^{-1})$. Detailed label-space definitions and client construction are provided in Appendix.

Implementation details. All methods use frozen CLIP backbones. We adopt ViT-B/16 by default and additionally use ViT-B/32 for the shot-sensitivity study. For the seven base-to-new datasets, we construct $N = 10$ clients with disjoint base-class assignments, and each client samples K images per assigned class. All clients participate for 50 communication rounds with 2 local epochs per round. We use SGD with a learning rate of 0.001 and a training batch size of 32. We set $\lambda_{\text{anc}} = 0.1$

Table 2: Average Base and Novel accuracy (%) over six datasets under different shot settings using ViT-B/32. H_{BN} denotes the harmonic mean of the averaged Base and Novel accuracies. Best and second-best results are shown in bold and underlined, respectively.

	16-shot			8-shot			4-shot			2-shot			1-shot		
Method	Base	Novel	H_{BN}	Base	Novel	H_{BN}	Base	Novel	H_{BN}	Base	Novel	H_{BN}	Base	Novel	H_{BN}
CLIP	75.57	79.04	77.27	75.57	79.04	77.27	75.57	79.04	77.27	75.57	79.04	77.27	75.57	79.04	77.27
PromptFL	**84.92**	76.82	80.67	**84.36**	76.85	80.43	**83.12**	76.73	79.79	80.63	76.92	78.73	79.59	77.53	78.55
FedPGP	70.79	68.89	69.83	80.33	73.60	76.82	81.90	74.50	78.03	**81.44**	77.08	79.20	79.35	80.25	79.80
FedOTP	9.97	16.74	12.50	9.95	19.39	13.15	11.02	21.19	14.50	11.29	23.12	15.17	11.20	20.66	14.53
pFedMoAP	29.82	44.15	35.60	42.53	50.92	46.35	44.14	54.34	48.71	47.14	58.33	52.14	45.53	54.08	49.44
pFedMMA	76.57	79.08	77.81	76.79	79.01	77.88	76.84	79.32	78.06	77.04	79.34	78.18	76.78	79.35	78.04
pFedVSA (Ours)	83.52	**82.60**	**83.06**	82.62	**82.15**	**82.38**	81.98	**82.50**	**82.24**	81.22	**82.83**	**82.02**	**81.69**	**82.97**	**82.32**

Table 3: Average component and personalization ablations over DTD, Caltech101, Flowers102, and OxfordPets. HM denotes the harmonic mean of the average Local, Base, and Novel accuracies.

Variant	Local	Base	Novel	HM
Full pFedVSA	**99.87**	**91.20**	**85.67**	**91.88**
Shared Image Adapter	99.65	89.12	84.47	90.65
w/o Image Adapter	99.42	89.40	84.23	90.59
w/o LVPM M_i	99.41	90.03	75.41	87.14
w/o PST θ_i^{pst}	99.55	80.70	81.37	86.39
w/o Text Adapter θ_i^{text}	97.23	84.23	82.19	87.40
w/o SACS	99.60	53.45	81.91	73.25
w/o $\mathcal{L}_{\mathrm{anc}}$	99.72	90.01	84.02	90.80
w/o $\mathcal{L}_{\mathrm{kd}}$	99.78	89.91	84.24	90.87
w/o $\mathcal{L}_{\mathrm{anc}}$ and $\mathcal{L}_{\mathrm{kd}}$	99.55	89.52	83.92	90.55

and $\lambda_{\mathrm{kd}} = 0.5$. All main results are averaged over three independent runs with different random seeds.

Main Results on Base-to-New Generalization. Table 1 reports the results on seven datasets under the 16-shot setting with ViT-B/16. pFedVSA achieves the highest average HM of 90.13%, outperforming pFedMMA by 5.58 percentage points, and ranks first in HM on all seven datasets. Compared with zero-shot CLIP, pFedVSA improves the average Local, Base, and Novel accuracies from 76.36%, 76.81%, and 81.21% to 98.65%, 87.35%, and 85.46%, respectively. Compared with pFedMMA, it further improves Base and Novel accuracy by 9.95 and 3.97 points while maintaining strong Local performance. These results show that pFedVSA achieves a better balance between client-specific adaptation and transferable Base-to-New generalization.

Robustness under heterogeneous data. Figures 2 and 3 evaluate pFedVSA under label and feature heterogeneity, respectively. On CIFAR-10 and CIFAR-100, pFedVSA maintains strong performance across Dirichlet partitions with different concentration parameters β, including severe label skew with highly imbalanced class coverage across clients. On DomainNet and Office-Caltech10, pFedVSA also achieves consistently strong performance across visually distinct domains. Together, these results show that pFedVSA remains robust under both label shift and cross-domain feature shift, supporting its ability to combine client-specific visual adaptation with transferable federated learning.

Robustness across shot settings. Table 2 summarizes the six-dataset average results with ViT-B/32 from 16-shot to 1-shot. pFedVSA achieves the highest H_{BN} under all five shot settings, reaching 83.06%, 82.38%, 82.24%, 82.02%, and 82.32%, respectively. Notably, Novel accuracy remains above 82% even as local supervision decreases from 16 to 1 shot. These results show that pFedVSA maintains a stable balance between base-class transfer and novel-class generalization under increasingly limited local supervision.

Component and personalization ablations. Table 3 evaluates the main components of pFedVSA. Keeping the image adapter client-specific improves HM by 1.23 and 1.29 points over sharing or removing it, respectively. Removing LVPM, PST, or the text adapter decreases HM by 4.74, 5.49, and 4.48 points, confirming the importance of local visual priors and their visual-to-semantic trans-

Table 4: Analysis of training-time visual-to-semantic transfer. Novel classes always use the standard adapted semantic branch at inference. H_{BN} denotes the harmonic mean of Base and Novel accuracy.

Training Variant	Base	Novel	H_{BN}
w/o LVPM	90.03	75.41	82.07
w/o PST	80.70	81.37	81.03
Shuffled visual priors	84.44	79.45	81.87
Fixed semantic token	87.30	82.31	84.73
Full pFedVSA	**91.20**	**85.67**	**88.35**

Table 5: Effect of SACS under heterogeneous class support. Local support follows $n_{i,c} \in \{4, 8, 16, 24, 28\}$ with an average of 16 samples per class. Results are averaged over SUN397 and DTD. H_{BN} denotes the harmonic mean of Base and Novel accuracy.

Gate Variant	Base	Novel	H_{BN}
Fixed gate	72.12	66.81	69.36
Support only	75.24	68.84	71.90
Similarity only	75.11	68.25	71.52
Learned gate	75.63	69.78	72.59
Full SACS	**77.02**	**71.18**	**73.98**

formation. SACS has the largest effect: removing it reduces Base accuracy from 91.20% to 53.45% and HM from 91.88% to 73.25%, demonstrating the importance of reliability-controlled integration of prototype-conditioned evidence. The anchoring and distillation objectives provide additional complementary gains, with removing both reducing HM to 90.55%.

Training-time visual-to-semantic transfer. Table 4 examines how local visual evidence affects novel-class generalization. Since novel classes never access LVPM or the prototype-conditioned branch at inference, their performance reflects the influence of visual-to-semantic transfer during training. Removing LVPM reduces Novel accuracy by 10.26 points, while removing PST reduces H_{BN} by 7.32 points. Shuffling the class–prior correspondence further lowers H_{BN} to 81.87%, while replacing the visual prior with a fixed semantic token reaches 84.73%, still 3.62 points below the full model. These results show that the Novel gain depends on class-consistent visual evidence being transferred into the shared semantic model rather than on an additional semantic branch alone.

SACS under heterogeneous class support. Table 5 evaluates SACS when class support varies while preserving the same average 16-shot budget. The fixed gate achieves only 69.36% H_{BN}, whereas using support sufficiency or image–prior similarity alone improves it to 71.90% and 71.52%, respectively. The learned gate reaches 72.59%, while Full SACS obtains the best Base, Novel, and H_{BN} results of 77.02%, 71.18%, and 73.98%. Its 4.62-point improvement over the fixed gate and 1.39-point advantage over the learned gate show that explicitly combining support sufficiency with image–prior consistency provides more reliable calibration under uneven local class coverage.

5 Conclusion

We presented pFedVSA, a personalized visual-to-semantic alignment framework for federated CLIP learning. pFedVSA adopts an asymmetric private-visual and shared-semantic design: client-specific image adapters and local visual prior memories remain private, while PST transfers class-level visual evidence into a collaboratively learned semantic space through shared PST and text-adapter parameters. SACS further incorporates the transferred evidence as a reliability-controlled residual, preventing sparse or biased local priors from overriding CLIP's transferable semantic knowledge. Extensive experiments under base-to-new generalization, Dirichlet label shift, and cross-domain heterogeneity demonstrate that pFedVSA achieves a stronger balance among client personalization, cross-client transfer, and robust generalization. This design improves generalization without exposing private visual parameters. Future work will investigate more adaptive visual-prior maintenance and semantic transfer for classes with limited or unavailable local evidence.

REFERENCES

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pp. 446–461. Springer, 2014.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022.

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*, pp. 3606–3613, 2014.

Tianyu Cui, Hongxia Li, Jingya Wang, and Ye Shi. Harmonizing generalization and personalization in federated prompt learning. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

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 *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178. IEEE, 2004.

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.

Sajjad Ghiasvand, Mahnoosh Alizadeh, and Ramtin Pedarsani. pfedmma: Personalized federated fine-tuning with multi-modal adapter for vision-language models. *arXiv preprint arXiv:2507.05394*, 2025.

Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 2066–2073. IEEE, 2012.

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*, 23(5):5179–5194, 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

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*, pp. 4904–4916. PMLR, 2021.

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*, pp. 709–727. Springer, 2022.

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*, pp. 19113–19122, 2023.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.(2009), 2009.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). *URL http://www. cs. toronto. edu/kriz/cifar. html*, 5(4):1, 2010.

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*, pp. 12151–12161, 2024.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10965–10975, 2022.

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10713–10722, 2021a.

Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623*, 2021b.

Wang Lu, Xixu Hu, Jindong Wang, and Xing Xie. Fedclip: Fast generalization and personalization for clip in federated learning. *arXiv preprint arXiv:2302.13485*, 2023.

Jun Luo, Chen Chen, and Shandong Wu. Mixture of experts made personalized: Federated prompt learning for vision-language models. In *International Conference on Learning Representations*, 2025.

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*, pp. 1273–1282. PMLR, 2017.

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*, pp. 722–729. IEEE, 2008.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

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*, pp. 1406–1415, 2019.

Chen Qiu, Xingyu Li, Chaithanya Kumar Mummadi, Madan Ravi Ganesh, Zhenzhen Li, Lu Peng, and Wan-Yi Lin. Text-driven prompt generation for vision-language models in federated learning. *arXiv preprint arXiv:2310.06123*, 2023.

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*, pp. 8748–8763. PmLR, 2021.

Mainak Singha, Subhankar Roy, Sarthak Mehrotra, Ankit Jha, Moloud Abdar, Biplab Banerjee, and Elisa Ricci. Fedmvp: Federated multimodal visual prompt tuning for vision-language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–10. IEEE, 2025.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

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*, pp. 3485–3492. IEEE, 2010.

Jiarong Yang and Yuan Liu. Beyond description: Federated adaptation via semantic-visual prototype alignment. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

Maxime Zanella and Ismail Ben Ayed. Low-rank few-shot adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1593–1603, 2024.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pp. 493–510. Springer, 2022.

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*, pp. 16816–16825, 2022a.

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, 2022b.

Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15659–15669, 2023.

A Appendix

A.1 Evaluation Label Spaces and Disjoint-Client Construction

For each base-to-new dataset, the full class set $\mathcal{Y}$ is divided into two disjoint subsets, $\mathcal{Y}^{\text{base}}$ and $\mathcal{Y}^{\text{novel}}$, satisfying $\mathcal{Y}^{\text{base}} \cap \mathcal{Y}^{\text{novel}} = \emptyset$ and $\mathcal{Y}^{\text{base}} \cup \mathcal{Y}^{\text{novel}} = \mathcal{Y}$. Only base-class images are used for federated training, whereas novel classes are completely held out from all clients and used only for evaluation.

Disjoint-class client construction. For the seven base-to-new datasets, we construct $N = 10$ clients by randomly partitioning the global base-class set into non-overlapping client-observed subsets $\{\mathcal{Y}_i^{\text{obs}}\}_{i=1}^{N}$:

$$\mathcal{Y}_i^{\text{obs}} \cap \mathcal{Y}_j^{\text{obs}} = \emptyset \quad (i \neq j), \qquad \bigcup_{i=1}^{N} \mathcal{Y}_i^{\text{obs}} = \mathcal{Y}^{\text{base}}.$$

If the number of base classes is not divisible by N, the remaining classes are distributed as evenly as possible. Under the K-shot setting, client i samples K training images for each class in $\mathcal{Y}_i^{\text{obs}}$:

$$\mathcal{D}_i^{\text{train}} = \left\{(x, y) \mid y \in \mathcal{Y}_i^{\text{obs}}\right\}, \qquad |\mathcal{D}_{i,c}^{\text{train}}| = K.$$

No image from $\mathcal{Y}^{\text{novel}}$ is assigned to any client. Although clients observe disjoint class subsets, they share the complete base-class vocabulary and perform local classification over $\mathcal{Y}^{\text{base}}$.

Evaluation label spaces. We report three top-1 accuracies with different candidate label spaces.

Local evaluates client i on test images from its observed class set $\mathcal{Y}_i^{\text{obs}}$, with the candidate labels also restricted to $\mathcal{Y}_i^{\text{obs}}$. The reported Local accuracy is averaged across clients and measures client-specific adaptation within the locally observed label space.

Base evaluates recognition over the complete global base-class space $\mathcal{Y}^{\text{base}}$. It therefore includes both classes observed by client i and base classes assigned to other clients, measuring transferable recognition over all training-time categories.

Novel evaluates recognition over the held-out label space $\mathcal{Y}^{\text{novel}}$. These classes never appear in federated training and measure whether the adapted model preserves CLIP's open-vocabulary transfer to unseen categories.

We summarize the balance among the three metrics using

$$\text{HM} = \frac{3}{\text{Local}^{-1} + \text{Base}^{-1} + \text{Novel}^{-1}}.$$

For analyses focusing on transferable recognition, we also use

$$H_{\text{BN}} = \frac{2}{\text{Base}^{-1} + \text{Novel}^{-1}}.$$

All compared methods follow the same training partitions and evaluation label-space definitions.

For pFedVSA, local visual priors are defined only for classes in $\mathcal{Y}_i^{\text{obs}}$. Locally unobserved base classes therefore use only the standard adapted semantic branch. Novel classes are excluded from federated training and have no client-side visual prior; they also use the standard adapted semantic branch at inference.

A.2 Additional Implementation Details

Backbones and optimization. All methods use frozen CLIP backbones. ViT-B/16 is the default backbone for the primary study on seven datasets. The auxiliary ViT-B/32 shot-sensitivity study uses six datasets: OxfordPets, SUN397, DTD, Caltech101, Food101, and UCF101. Thus, the seven-dataset ViT-B/16 average reported in the main paper and the six-dataset ViT-B/32 study in this appendix refer to two different evaluation sets. For the primary study, all ten clients participate in each of the 50 communication rounds, and each client performs two local epochs per round.

We optimize the trainable modules using SGD with a learning rate of 0.001, a training batch size of 32, and a testing batch size of 100. All main-paper and supplementary results are averaged over

three independent runs with different random seeds. All supplementary comparisons use the same data partitions, label spaces, and optimization settings as the corresponding main-paper experiments.

Dataset-specific federated settings. For the seven base-to-new datasets, we construct $N = 10$ clients with disjoint base-class assignments. Each client samples K training images per assigned class, where K is determined by the corresponding shot setting. All clients participate in each of the 50 communication rounds, and every client performs two local epochs per round.

For CIFAR-10 and CIFAR-100, we construct label-heterogeneous clients using Dirichlet partitions and evaluate the methods under different concentration parameters β. A smaller β produces stronger label skew and more imbalanced class coverage across clients.

For DomainNet and Office-Caltech10, the client construction preserves the original visual domains to introduce feature heterogeneity. Dirichlet label partitioning is additionally applied to introduce label-distribution heterogeneity. Unless otherwise specified, the cross-domain experiments use $\beta = 0.5$.

The multi-shot comparisons include CLIP, PromptFL, FedPGP, FedOTP Li et al. (2024), pFedMoAP, and pFedMMA. Complete descriptions and references for these baselines are provided in the main paper. FedCLIP-Adapter and FedCLIP-LoRA are reported in the main-paper 16-shot ViT-B/16 benchmark but are not part of the complete multi-shot and ViT-B/32 sweeps.

Trainable module architecture. At communication round r, the locally trainable parameters of client i are

$$\Theta_i^r = \left\{\theta_{\text{image},i}^r, \theta_{\text{pst},i}^r, \theta_{\text{text},i}^r\right\},$$

where $\theta_{\text{image},i}^r$ denotes the personalized image-adapter parameters, $\theta_{\text{pst},i}^r$ denotes the Prototype-to-Semantic Transfer parameters, and $\theta_{\text{text},i}^r$ denotes the shared text-adapter parameters.

The globally shared parameter set is

$$\Theta_g^r = \left\{\theta_{\text{pst},g}^r, \theta_{\text{text},g}^r\right\}.$$

Only the PST and text-adapter parameters are uploaded and aggregated. The image adapter and the local visual prior memory remain local to each client. Hence, federated optimization learns a shared visual-to-semantic mapping rather than aggregating the class-level visual priors themselves.

The personalized image adapter is inserted into the frozen CLIP image encoder and produces a client-aware normalized image representation

$$v_{i,j}^r = \text{Norm}\left(f_{\text{image}}\left(x_{i,j}; \theta_{\text{image},i}^r\right)\right) \in \mathbb{R}^{512}.$$

The adapter captures client-specific visual characteristics and is optimized only using the corresponding client's data.

The PST module is implemented as a lightweight inversion network with the architecture

$$512 \to 32 \to 512.$$

Given a class-level visual prior $p_{i,c}^r$, PST produces a CLIP-compatible semantic pseudo-token

$$s_{i,c}^r = g_{\text{pst}}\left(p_{i,c}^r; \theta_{\text{pst},i}^r\right).$$

The pseudo-token is inserted into the prototype-conditioned prompt

$$\pi_{i,c}^p = \text{“a photo of a } [s_{i,c}^r]\ [\text{CLASS}].\text{”}$$

The shared text adapter is inserted into the frozen CLIP text encoder. The standard adapted-text branch and the prototype-conditioned branch use the same text-adapter parameters:

$$t_{i,c}^a = \text{Norm}\left(F_{\text{text}}\left(\pi_c^a; \theta_{\text{text},i}^r\right)\right),$$

$$t_{i,c}^p = \text{Norm}\left(F_{\text{text}}\left(\pi_{i,c}^p; \theta_{\text{text},i}^r\right)\right),$$

where $\pi_c^a =$ “a photo of a [CLASS].” The two branches differ only in prompt construction and the presence of the semantic pseudo-token.

Round-wise local visual prior memory. Each client maintains a client-local visual prior memory defined only over its observed classes:

$$\mathcal{M}_i^r = \left\{ \left(p_{i,c}^r, n_{i,c}\right) \mid c \in \mathcal{Y}_i^{\text{obs}} \right\}.$$

The memory is reconstructed at the beginning of every communication round using the current personalized image adapter and the unique local training samples, remains fixed throughout local optimization, and is never uploaded.

For each $c \in \mathcal{Y}_i^{\text{obs}}$, let

$$\mathcal{D}_{i,c}^{\text{tr}} = \left\{ (x_{i,j}, y_{i,j}) \in \mathcal{D}_i^{\text{tr}} \mid y_{i,j} = c \right\}, \qquad N_{i,c} = \left| \mathcal{D}_{i,c}^{\text{tr}} \right|.$$

The corresponding visual prior and support count are

$$p_{i,c}^r = \text{Norm}\left(\frac{1}{N_{i,c}} \sum_{(x_{i,j}, y_{i,j}) \in \mathcal{D}_{i,c}^{\text{tr}}} \text{sg}\left(v_{i,j}^r\right) \right),$$
$$n_{i,c} = N_{i,c},$$

where $\text{sg}(\cdot)$ denotes stop-gradient. Thus, $n_{i,c}$ records unique local support rather than repeated mini-batch exposure. Round-wise reconstruction avoids repeated-count inflation and stale representations generated by earlier image-adapter states.

Reliability gate and gradient flow. For $c \in \mathcal{Y}_i^{\text{obs}}$, SACS computes

$$s_{ij,c}^p = \text{sim}\left(v_{i,j}, t_{i,c}^p\right), \qquad s_{ij,c}^a = \text{sim}\left(v_{i,j}, t_{i,c}^a\right).$$

The reliability of the local visual prior is evaluated using support sufficiency and image–prior consistency:

$$g_{ij,c} = \frac{n_{i,c}}{n_{i,c} + k} \sigma\left(\gamma \left[\text{sim}\left(v_{i,j}, p_{i,c}\right) - \tau_g\right]\right), \qquad c \in \mathcal{Y}_i^{\text{obs}},$$

with $\bar{g}_{ij,c} = \text{clip}(g_{ij,c}, 0, g_{\max})$. The default configuration uses $k = 16$, $\gamma = 10$, $\tau_g = 0.3$, and $g_{\max} = 0.5$.

The reliability gate contains no additional learnable network. The stored prior $p_{i,c}$ is treated as a fixed no-gradient buffer, while gradients can propagate through the current image representation $v_{i,j}$ in the similarity term. The final prediction is

$$\ell_{ij,c} = \begin{cases} \alpha \left[s_{ij,c}^a + \bar{g}_{ij,c} \left(s_{ij,c}^p - s_{ij,c}^a \right) \right], & c \in \mathcal{Y}_i^{\text{obs}}, \\ \alpha s_{ij,c}^a, & c \notin \mathcal{Y}_i^{\text{obs}}, \end{cases}$$

where α is the CLIP logit scale. Thus, locally unsupported base classes use only the standard adapted semantic branch. Novel classes follow the same standard branch at inference.

Local training objective. The local training objective is

$$\mathcal{L}_i = \mathcal{L}_{\text{ce}} + 0.1\, \mathcal{L}_{\text{anc}} + 0.5\, \mathcal{L}_{\text{kd}}.$$

The classification loss is computed over the complete global base-class vocabulary $\mathcal{Y}^{\text{tr}} = \mathcal{Y}^{\text{base}}$:

$$\mathcal{L}_{\text{ce}} = -\frac{1}{|\mathcal{B}_i|} \sum_{(x_{i,j}, y_{i,j}) \in \mathcal{B}_i} \log \frac{\exp(\ell_{ij,y_{i,j}})}{\sum_{c \in \mathcal{Y}^{\text{tr}}} \exp(\ell_{ij,c})}.$$

For the classes supported by the local visual prior memory, the semantic anchoring loss is

$$\mathcal{L}_{\text{anc}} = \frac{1}{|\mathcal{Y}_i^{\mathcal{M}}|} \sum_{c \in \mathcal{Y}_i^{\mathcal{M}}} \left[1 - \text{sim}\left(t_{i,c}^p, \text{sg}(t_{i,c}^a)\right)\right],$$

where

$$\mathcal{Y}_i^{\mathcal{M}} = \{c \mid n_{i,c} > 0\}.$$

Knowledge distillation uses the prediction of the original frozen CLIP model, computed without the image adapter, PST, or text adapter, as the teacher. Let

$$s^0_{ij,c} = \mathrm{sim}\left(v^0_{i,j}, t^0_c\right)$$

denote the original CLIP similarity. The distillation loss is

$$\mathcal{L}_{\mathrm{kd}} = T^2\,\mathrm{KL}\left(\mathrm{sg}\left[\mathrm{softmax}\left(\frac{\alpha s^0}{T}\right)\right] \middle\| \mathrm{softmax}\left(\frac{\ell}{T}\right)\right).$$

Thus, semantic anchoring constrains the prototype-conditioned representation relative to the adapted text semantics, whereas distillation preserves the transferable structure of the original pretrained CLIP model.

Federated communication. At the beginning of communication round r, the server broadcasts

$$\Theta^r_g = \left\{\theta^r_{\mathrm{text},g}, \theta^r_{\mathrm{pst},g}\right\}$$

to the selected clients. Client i initializes

$$\theta^r_{\mathrm{text},i} \leftarrow \theta^r_{\mathrm{text},g}, \qquad \theta^r_{\mathrm{pst},i} \leftarrow \theta^r_{\mathrm{pst},g},$$

combines them with its client-local image adapter $\theta^r_{\mathrm{image},i}$, and reconstructs $\mathcal{M}^r_i$ before local optimization.

During local training, the image adapter, PST, and text adapter are optimized, while the reconstructed local visual prior memory remains fixed. At the end of the round, client i retains its image adapter and local memory and uploads only

$$\Theta^{r+1}_i = \left\{\theta^{r+1}_{\mathrm{text},i}, \theta^{r+1}_{\mathrm{pst},i}\right\}.$$

Let $\mathcal{S}_r$ denote the selected-client set and $N_i = |\mathcal{D}^{\mathrm{tr}}_i|$. For $m \in \{\mathrm{text}, \mathrm{pst}\}$, the server performs sample-weighted aggregation:

$$\theta^{m,r+1}_g = \sum_{i\in\mathcal{S}_r} \frac{N_i}{\sum_{k\in\mathcal{S}_r} N_k}\theta^{m,r+1}_i.$$

Under CLIP ViT-B/16, the shared text adapter contains $3{,}072$ parameters and PST contains $32{,}768$ parameters. Therefore, pFedVSA communicates $35{,}840$ parameters in each direction per participating client and communication round. The personalized image adapter and all class-wise visual priors and support counts remain entirely local.

A.3 Additional Few-Shot and Backbone Results

Robustness to the number of shots. Table A1 reports the detailed ViT-B/16 results on seven datasets under 1-, 2-, 4-, and 8-shot supervision. pFedVSA achieves the highest average HM in all four settings, reaching 87.32%, 88.21%, 88.80%, and 89.23% from 1 to 8 shots, respectively. The corresponding gains over the strongest competing method are 3.85, 4.01, 4.24, and 4.59 percentage points. At the dataset level, pFedVSA obtains the best HM in 27 of the 28 dataset–shot combinations; the only exception is UCF101 at 8-shot, where it is 0.02 points below the best competing result. These results show that the advantage of pFedVSA persists from extremely sparse 1-shot supervision to the denser 8-shot setting.

Results with a second CLIP backbone. Table A2 reports complete dataset-level results with ViT-B/32 under 1-, 2-, 4-, 8-, and 16-shot supervision. pFedVSA achieves the highest HM in all 30 evaluated dataset–shot combinations. Across the five shot settings, the gains remain consistent on OxfordPets, SUN397, DTD, Caltech101, Food101, and UCF101, showing that the observed behavior is not specific to the default ViT-B/16 backbone. The main paper reports the corresponding six-dataset average Base, Novel, and H_{BN} results, while this appendix provides the full per-dataset breakdown.

8-shot		Average on 7 datasets				SUN397				Flowers102				DTD			
Shots	Method	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM
8	CLIP	76.36	76.81	81.21	78.07	69.41	69.38	75.52	71.32	67.89	69.23	76.88	71.12	54.26	54.86	59.18	56.02
	PromptFL	88.24	88.03	77.57	84.31	78.03	78.01	71.95	75.89	95.71	95.63	69.29	84.90	78.29	76.04	46.98	63.55
	FedPGP	93.41	83.36	76.61	83.90	93.95	54.50	57.63	64.73	94.84	92.49	71.07	84.68	85.42	71.47	51.45	66.47
	FedOTP	96.63	24.30	42.92	40.11	93.41	11.90	18.00	19.96	99.73	20.47	45.03	37.00	96.99	23.33	42.48	39.11
	pFedMoAP	97.04	71.31	71.73	78.39	95.16	45.80	49.43	57.06	99.75	66.88	61.99	72.98	94.44	58.11	52.10	63.84
	pFedMMA	96.66	77.86	81.61	84.64	92.71	70.90	76.25	78.94	95.52	72.31	76.43	80.25	95.32	56.30	61.57	67.42
	pFedVSA (Ours)	98.33	86.12	84.46	89.23	94.56	77.65	79.44	83.24	100.00	93.34	81.55	90.98	98.65	72.13	64.12	75.76
	Δ	+1.29	−1.91	+2.85	+4.59	−0.60	−0.36	+3.19	+4.30	+0.25	−2.29	+4.67	+6.08	+1.66	−3.91	+2.55	+8.34
		OxfordPets				Caltech101				Food101				UCF101			
	Method	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM
	CLIP	89.45	89.42	96.81	91.77	96.14	97.22	94.21	95.84	89.40	89.42	90.70	89.84	68.00	68.15	75.18	70.29
	PromptFL	95.53	95.43	97.32	96.09	97.31	98.26	94.43	96.64	90.44	90.46	90.67	90.52	82.39	82.37	72.36	78.75
	FedPGP	96.09	94.93	96.66	95.89	98.13	98.26	93.47	96.57	90.57	90.38	91.04	90.66	94.85	81.48	74.94	82.96
	FedOTP	100.00	15.67	55.72	32.69	99.89	58.68	73.35	73.74	95.08	24.92	40.33	39.77	91.28	15.12	25.52	25.80
	pFedMoAP	99.90	78.13	91.76	89.00	99.56	96.42	93.02	96.26	96.76	80.92	87.45	87.90	93.68	72.88	66.39	76.03
	pFedMMA	99.95	89.35	96.64	95.10	99.85	96.99	94.30	96.99	97.15	89.24	90.73	92.25	96.09	69.94	75.33	78.99
	pFedVSA (Ours)	100.00	95.14	97.86	97.63	100.00	98.01	95.77	97.90	97.99	90.98	93.31	94.00	97.08	75.56	79.15	82.94
	Δ	0.00	−0.29	+0.54	+1.54	+0.11	−0.25	+1.34	+0.91	+0.84	+0.52	+2.27	+1.75	+0.99	−6.81	+3.82	−0.02
4-shot		Average on 7 datasets				SUN397				Flowers102				DTD			
Shots	Method	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM
4	CLIP	76.36	76.81	81.21	78.07	69.41	69.38	75.52	71.32	67.89	69.23	76.88	71.12	54.26	54.86	59.18	56.02
	PromptFL	87.12	86.90	80.06	84.56	77.60	77.57	76.11	77.09	95.05	94.68	69.79	84.72	73.56	71.41	56.88	66.40
	FedPGP	89.80	85.18	78.15	84.10	83.72	71.91	69.78	74.66	94.48	92.43	72.82	85.38	82.69	67.51	51.65	64.84
	FedOTP	95.90	30.70	45.68	46.23	91.88	17.98	27.17	29.04	98.79	19.00	33.61	32.43	95.28	25.12	41.23	40.24
	pFedMoAP	95.89	73.47	73.72	79.78	92.49	59.62	61.08	68.25	99.62	66.14	62.01	72.67	92.41	51.78	49.90	59.79
	pFedMMA	96.09	77.97	81.62	84.54	91.33	70.79	76.12	78.51	95.35	71.89	76.06	79.90	93.75	56.46	61.91	67.37
	pFedVSA (Ours)	97.83	85.76	84.03	88.80	94.12	77.52	79.10	82.95	99.79	92.98	79.85	90.09	97.09	71.24	63.56	74.87
	Δ	+1.74	−1.14	+2.41	+4.24	+1.63	−0.05	+2.98	+4.44	+0.17	−1.70	+2.97	+4.71	+1.81	−0.17	+1.65	+7.50
		OxfordPets				Caltech101				Food101				UCF101			
	Method	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM
	CLIP	89.45	89.42	96.81	91.77	96.14	97.22	94.21	95.84	89.40	89.42	90.70	89.84	68.00	68.15	75.18	70.29
	PromptFL	95.84	95.96	97.76	96.51	97.17	98.00	93.34	96.13	89.93	89.95	90.53	90.14	80.67	80.71	75.99	79.06
	FedPGP	96.55	95.54	97.57	96.55	97.84	97.99	92.34	95.98	90.25	90.24	91.17	90.55	83.09	80.64	71.73	78.17
	FedOTP	99.95	41.83	70.13	62.28	99.73	65.97	77.59	78.79	94.49	29.38	46.57	45.39	91.20	15.60	23.48	25.50
	pFedMoAP	99.75	85.69	94.08	92.81	99.08	95.95	92.73	95.85	95.49	81.68	86.57	87.55	92.40	73.42	69.67	77.33
	pFedMMA	99.90	89.68	96.63	95.21	99.85	96.93	94.40	97.01	96.60	89.34	90.83	92.15	95.87	70.71	75.37	79.28
	pFedVSA (Ours)	99.95	94.79	97.68	97.43	100.00	98.00	95.64	97.85	96.89	90.76	93.33	93.59	97.00	75.06	79.03	82.68
	Δ	0.00	−1.17	−0.08	+0.88	+0.15	0.00	+1.24	+0.84	+0.29	+0.52	+2.16	+1.44	+1.13	−5.65	+3.04	+3.40
2-shot		Average on 7 datasets				SUN397				Flowers102				DTD			
Shots	Method	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM
2	CLIP	76.36	76.81	81.21	78.07	69.41	69.38	75.52	71.32	67.89	69.23	76.88	71.12	54.26	54.86	59.18	56.02
	PromptFL	85.26	84.99	77.30	82.35	75.66	75.65	75.64	75.65	92.71	92.69	65.25	81.30	67.69	65.28	48.43	59.12
	FedPGP	86.24	83.88	78.73	82.83	76.42	74.97	72.34	74.54	90.98	89.38	68.33	81.49	73.66	62.56	56.85	63.63
	FedOTP	94.10	33.34	42.50	46.77	87.53	26.06	31.85	36.95	97.77	18.57	29.49	30.62	91.71	26.94	37.97	40.34
	pFedMoAP	93.11	74.45	73.63	79.46	85.87	70.76	68.06	74.13	98.81	65.00	61.84	71.98	82.64	53.28	46.23	57.14
	pFedMMA	94.57	78.13	81.60	84.20	89.23	71.48	76.46	78.38	94.42	72.67	75.98	79.97	89.72	56.48	61.92	66.67
	pFedVSA (Ours)	97.39	84.98	83.52	88.21	93.54	76.09	78.90	82.18	98.47	91.07	78.44	88.54	96.43	69.75	62.55	73.72
	Δ	+2.82	−0.01	+1.92	+4.01	+4.31	+0.44	+2.44	+3.80	−0.34	−1.62	+1.56	+7.05	+4.72	+4.47	+0.63	+7.05
		OxfordPets				Caltech101				Food101				UCF101			
	Method	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM
	CLIP	89.45	89.42	96.81	91.77	96.14	97.22	94.21	95.84	89.40	89.42	90.70	89.84	68.00	68.15	75.18	70.29
	PromptFL	95.03	94.90	96.92	95.61	96.86	97.61	93.34	95.90	89.93	89.73	89.71	89.79	78.92	79.06	71.82	76.45
	FedPGP	94.88	94.55	97.27	95.55	97.06	97.10	92.99	95.68	89.47	89.46	90.48	89.80	81.22	79.16	72.88	77.59
	FedOTP	100.00	38.63	55.59	55.68	99.64	75.35	81.09	84.18	93.16	27.49	44.35	43.07	88.87	20.34	17.13	25.25
	pFedMoAP	99.69	85.76	90.07	91.48	99.33	96.07	91.76	95.62	94.97	81.21	86.27	87.12	90.46	69.10	71.21	75.82
	pFedMMA	99.64	89.46	96.80	95.10	99.80	97.07	94.14	96.95	95.59	89.40	90.76	91.84	93.60	70.36	75.12	78.52
	pFedVSA (Ours)	99.87	94.46	97.56	97.25	99.89	97.97	95.16	97.63	96.68	90.53	93.10	93.37	96.88	75.02	78.93	82.60
	Δ	−0.13	−0.44	+0.29	+1.64	+0.09	+0.36	+0.95	+0.68	+1.09	+0.80	+2.34	+1.53	+3.28	−4.14	+3.75	+4.08
1-shot		Average on 7 datasets				SUN397				Flowers102				DTD			
Shots	Method	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM
1	CLIP	76.36	76.81	81.21	78.07	69.41	69.38	75.52	71.32	67.89	69.23	76.88	71.12	54.26	54.86	59.18	56.02
	PromptFL	81.99	81.93	78.83	80.89	73.80	73.77	75.11	74.22	83.12	83.67	71.28	78.92	60.51	58.22	50.48	56.06
	FedPGP	82.48	82.06	78.50	80.97	74.36	74.07	76.63	75.00	84.18	82.69	67.23	77.23	61.34	60.01	50.91	57.02
	FedOTP	93.51	32.72	47.55	48.16	86.48	27.51	34.68	39.09	95.88	22.83	45.91	39.47	88.61	32.44	41.28	45.22
	pFedMoAP	88.57	75.82	76.35	79.84	81.90	68.22	68.14	72.21	82.70	70.37	76.55	76.21	75.56	55.00	50.71	58.67
	pFedMMA	92.40	77.66	81.70	83.47	84.42	70.92	76.38	76.85	92.76	70.37	76.55	78.83	86.67	56.31	61.30	65.77
	pFedVSA (Ours)	96.38	84.02	82.80	87.32	91.21	74.99	77.05	80.48	96.66	88.98	77.87	87.14	94.82	66.70	61.67	71.85
	Δ	+2.87	+1.96	+1.10	+3.85	+4.73	+0.92	+0.42	+3.63	+0.78	+5.31	+0.99	+8.22	+6.21	+6.69	+0.37	+6.08
		OxfordPets				Caltech101				Food101				UCF101			
	Method	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM
	CLIP	89.45	89.42	96.81	91.77	96.14	97.22	94.21	95.84	89.40	89.42	90.70	89.84	68.00	68.15	75.18	70.29
	PromptFL	94.89	94.95	97.43	95.74	96.11	97.29	94.43	95.93	89.45	89.46	90.02	89.64	76.05	76.16	73.07	75.07
	FedPGP	94.88	94.75	96.66	95.42	96.30	97.42	92.96	95.52	89.04	89.04	90.91	89.65	77.26	76.47	74.23	75.96
	FedOTP	99.95	23.38	65.26	44.05	99.15	60.79	73.46	74.72	95.88	40.98	50.67	54.98	88.61	21.10	21.59	28.57
	pFedMoAP	98.36	88.83	94.92	93.87	98.40	95.63	92.29	95.37	92.62	83.06	87.32	87.49	90.46	69.62	64.49	73.31
	pFedMMA	98.83	89.64	97.02	94.99	99.36	97.04	94.36	96.88	93.74	89.44	90.83	91.30	91.03	69.91	75.47	77.84
	pFedVSA (Ours)	99.68	94.34	97.49	97.12	99.87	97.89	95.07	97.57	95.82	90.21	93.07	92.98	96.63	75.04	77.40	81.98
	Δ	−0.27	−0.61	+0.06	+1.38	+0.51	+0.47	+0.64	+0.69	−0.06	+0.75	+2.16	+1.68	+5.60	−1.43	+1.93	+4.14

Table A1: Top-1 accuracy (%) of different methods across seven datasets under 8-, 4-, 2-, and 1-shot settings using ViT-B/16 as the backbone. HM denotes the harmonic mean of Local, Base, and Novel accuracy. For Average, Local, Base, and Novel are first averaged across datasets, and HM is then computed from the three reported averages. Δ is the absolute percentage-point difference between pFedVSA and the strongest competing method for the corresponding metric.

16-shot		OxfordPets				SUN397				DTD			
Shots	Method	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM
16	CLIP	86.40	86.87	96.09	89.57	69.81	69.78	72.99	70.83	52.27	53.13	54.47	53.27
	PromptFL	93.78	**93.83**	96.53	94.70	75.38	75.35	68.04	72.75	77.22	**76.85**	57.00	68.96
	FedPGP	95.01	93.01	94.31	94.10	95.33	38.76	47.21	52.20	88.49	61.99	41.74	58.38
	FedOTP	99.90	10.19	32.80	21.64	90.06	9.03	5.79	10.18	**97.70**	10.68	18.48	18.99
	pFedMoAP	**100.00**	30.45	63.43	51.19	94.96	24.98	26.71	34.09	96.66	16.09	24.70	26.55
	pFedMMA	94.05	87.92	95.74	92.45	92.09	70.65	73.77	77.78	72.31	54.62	53.93	59.19
	pFedVSA (Ours)	**100.00**	93.45	**97.83**	**97.02**	**95.40**	**78.82**	**75.88**	**82.54**	96.27	66.68	**57.45**	**70.11**

		Caltech101				Food101				UCF101			
	Method	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM
	CLIP	92.63	94.00	94.00	93.54	84.33	84.27	85.43	84.67	65.16	65.36	71.28	67.15
	PromptFL	95.85	97.29	93.01	95.35	86.29	86.31	86.03	86.21	79.95	**79.89**	60.30	72.10
	FedPGP	98.28	93.63	88.50	93.30	89.65	83.55	85.85	86.28	94.64	53.80	55.73	63.70
	FedOTP	99.38	11.46	23.17	21.36	92.21	9.63	14.03	16.13	88.36	8.84	6.15	10.45
	pFedMoAP	99.92	51.79	65.33	67.23	96.41	31.56	53.51	49.39	95.28	24.08	31.24	35.70
	pFedMMA	98.56	94.74	93.94	95.70	95.31	84.59	85.59	88.24	87.80	66.91	71.53	74.41
	pFedVSA (Ours)	**100.00**	**97.33**	**96.19**	**97.81**	**97.16**	**88.21**	**88.76**	**91.20**	**96.48**	76.63	**79.52**	**83.36**

8-shot		OxfordPets				SUN397				DTD			
Shots	Method	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM
8	CLIP	86.40	86.87	96.09	89.57	69.81	69.78	72.99	70.83	52.27	53.13	54.47	53.27
	PromptFL	93.34	**93.57**	96.03	94.30	75.93	**75.91**	73.61	75.13	75.05	72.34	49.64	63.43
	FedPGP	94.87	91.69	92.94	93.15	93.15	55.78	57.24	65.03	78.66	**72.76**	45.61	62.01
	FedOTP	99.85	11.08	37.61	23.65	87.18	8.78	4.64	8.80	95.84	10.11	19.37	18.64
	pFedMoAP	**100.00**	33.19	63.15	53.61	94.79	26.54	23.41	32.99	95.58	19.78	32.68	32.74
	pFedMMA	93.80	88.65	95.86	92.67	89.77	70.19	73.41	76.91	68.01	55.68	54.34	58.75
	pFedVSA (Ours)	**100.00**	92.43	**96.75**	**96.29**	**95.54**	74.98	**75.11**	**80.82**	**96.00**	68.16	**66.94**	**74.95**

		Caltech101				Food101				UCF101			
	Method	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM
	CLIP	92.63	94.00	94.00	93.54	84.33	84.27	85.43	84.67	65.16	65.36	71.28	67.15
	PromptFL	96.89	**97.93**	92.25	95.63	86.16	86.18	87.63	86.65	80.58	**80.25**	61.93	73.14
	FedPGP	96.46	97.09	91.32	94.89	86.28	86.27	**87.96**	86.83	85.02	78.41	66.56	75.87
	FedOTP	99.03	11.99	27.97	23.21	89.94	9.44	18.90	17.65	82.82	8.31	7.84	11.54
	pFedMoAP	**99.72**	80.74	80.58	86.15	96.21	54.37	68.67	69.21	94.98	40.58	37.06	48.27
	pFedMMA	97.87	94.88	93.64	95.43	95.17	84.73	85.72	88.30	83.06	66.59	71.07	72.94
	pFedVSA (Ours)	**99.72**	94.71	**94.22**	**96.15**	**96.52**	**86.98**	86.61	**89.81**	**98.33**	78.46	**73.26**	**82.05**

4-shot		OxfordPets				SUN397				DTD			
Shots	Method	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM
4	CLIP	86.40	86.87	96.09	89.57	69.81	69.78	72.99	70.83	52.27	53.13	54.47	53.27
	PromptFL	93.73	93.99	96.87	94.84	75.53	**75.50**	72.24	74.39	71.16	**68.87**	46.38	59.84
	FedPGP	93.93	93.23	91.26	92.79	83.31	70.90	66.64	72.97	74.70	68.33	44.37	59.34
	FedOTP	99.80	10.21	31.68	21.50	81.32	8.38	7.93	11.64	**95.72**	10.85	22.77	20.47
	pFedMoAP	**100.00**	33.13	65.34	54.07	**93.34**	38.49	37.64	47.42	94.77	16.43	31.68	29.13
	pFedMMA	93.64	88.23	96.33	92.61	83.44	70.92	73.66	75.64	67.64	54.93	55.06	58.65
	pFedVSA (Ours)	97.10	**94.12**	**98.17**	**96.43**	88.36	72.95	**76.33**	**78.69**	86.53	63.68	**62.00**	**69.14**

		Caltech101				Food101				UCF101			
	Method	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM
	CLIP	92.63	94.00	94.00	93.54	84.33	84.27	85.43	84.67	65.16	65.36	71.28	67.15
	PromptFL	95.44	**96.90**	93.12	95.13	85.59	85.61	87.55	86.24	78.00	77.82	64.20	72.73
	FedPGP	95.50	96.48	90.39	94.05	86.11	86.09	88.01	86.73	81.62	76.40	66.31	74.22
	FedOTP	99.08	18.93	39.24	33.94	89.84	9.39	18.42	17.45	83.32	8.34	7.10	11.00
	pFedMoAP	**99.77**	81.21	82.39	87.02	**95.75**	52.96	65.93	67.43	95.52	42.63	43.05	52.49
	pFedMMA	97.32	94.80	93.84	95.30	94.33	84.15	85.33	87.71	84.53	68.04	71.68	74.11
	pFedVSA (Ours)	99.24	95.03	**94.35**	**96.16**	95.51	**87.76**	**88.35**	**90.41**	**96.31**	**78.33**	**75.81**	**82.55**

Table A2: Top-1 accuracy (%) across six datasets under different shot settings using ViT-B/32 as the backbone. HM denotes the harmonic mean of Local, Base, and Novel accuracy. Best and second-best results within each shot setting are shown in bold and underlined, respectively.

2-shot		OxfordPets				SUN397				DTD			
Shots	Method	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM
2	CLIP	86.40	86.87	96.09	89.57	69.81	69.78	72.99	70.83	52.27	53.13	54.47	53.27
	PromptFL	90.62	91.44	96.87	92.90	74.08	74.06	73.81	73.98	61.20	60.88	47.95	55.95
	FedPGP	92.05	92.41	91.80	92.09	74.78	74.36	73.70	74.28	67.13	**64.93**	54.77	61.78
	FedOTP	99.43	13.86	48.24	29.14	71.89	8.61	8.60	12.18	**92.27**	10.31	22.19	19.62
	pFedMoAP	**100.00**	28.97	67.42	50.55	**89.80**	54.74	57.58	64.14	92.25	21.92	32.49	34.39
	pFedMMA	93.18	88.30	95.73	92.30	79.27	71.52	74.10	74.83	62.18	55.67	54.84	57.38
	pFedVSA (Ours)	95.43	**92.80**	**98.84**	**95.63**	82.79	**78.92**	**80.96**	**80.86**	90.75	62.10	**61.40**	**69.11**

		Caltech101				Food101				UCF101			
	Method	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM
	CLIP	92.63	94.00	94.00	93.54	84.33	84.27	85.43	84.67	65.16	65.36	71.28	67.15
	PromptFL	95.27	**96.45**	91.59	94.39	85.09	85.10	86.97	85.71	75.88	**75.85**	64.31	71.58
	FedPGP	94.62	95.53	90.92	93.65	85.95	85.94	87.81	86.56	79.23	75.44	63.50	72.07
	FedOTP	97.95	14.81	32.70	27.70	88.58	10.78	18.58	19.00	82.00	9.36	8.39	12.59
	pFedMoAP	**99.62**	81.01	80.41	86.16	**95.01**	52.55	68.68	68.00	93.63	43.63	43.39	52.96
	pFedMMA	97.89	94.70	93.73	95.41	92.87	84.98	85.87	87.77	81.02	67.09	71.78	72.85
	pFedVSA (Ours)	99.45	95.05	**94.03**	**96.12**	94.37	**87.17**	**89.32**	**90.19**	**94.06**	71.27	**72.45**	**77.99**

1-shot		OxfordPets				SUN397				DTD			
Shots	Method	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM
1	CLIP	86.40	86.87	96.09	89.57	69.81	69.78	72.99	70.83	52.27	53.13	54.47	53.27
	PromptFL	91.70	**92.29**	97.04	93.62	72.86	72.82	73.56	73.08	56.71	57.29	44.93	52.31
	FedPGP	86.86	88.71	87.98	87.84	74.54	74.34	**87.66**	78.38	60.76	57.93	55.83	58.10
	FedOTP	99.39	11.81	34.94	24.32	70.75	7.69	6.75	10.26	89.47	13.66	20.19	22.40
	pFedMoAP	**100.00**	26.32	52.30	44.70	**85.99**	52.37	54.12	60.97	**89.77**	23.96	31.14	35.30
	pFedMMA	91.55	88.09	96.39	91.88	75.86	71.02	73.87	73.53	59.72	55.31	54.89	56.56
	pFedVSA (Ours)	95.71	91.10	**97.26**	**94.62**	81.50	**78.31**	80.68	**80.14**	80.48	**60.22**	**60.09**	**65.68**

		Caltech101				Food101				UCF101			
	Method	Local	Base	Novel	HM	Local	Base	Novel	HM	Local	Base	Novel	HM
	CLIP	92.63	94.00	94.00	93.54	84.33	84.27	85.43	84.67	65.16	65.36	71.28	67.15
	PromptFL	94.35	**96.26**	92.58	94.37	85.68	85.71	**87.91**	86.42	73.17	73.16	69.17	71.78
	FedPGP	93.66	95.34	91.74	93.56	85.50	85.52	87.66	86.21	74.54	74.24	70.64	73.10
	FedOTP	97.88	14.77	30.07	26.98	85.38	11.43	22.85	20.98	78.14	7.86	9.14	12.03
	pFedMoAP	**99.14**	77.97	77.75	83.86	93.62	56.89	68.39	69.96	90.57	35.69	40.75	47.17
	pFedMMA	94.77	94.18	93.83	94.26	91.67	84.78	85.84	87.33	79.28	67.30	71.29	72.29
	pFedVSA (Ours)	97.24	96.01	**95.48**	**96.24**	**93.89**	**88.00**	86.13	**89.22**	**93.17**	**76.48**	**78.16**	**81.96**

Table A2: Top-1 accuracy (%) across six datasets under different shot settings using ViT-B/32 as the backbone (continued).

Table A3: Ablation of semantic-module sharing strategies. "S" denotes globally shared and "L" denotes client-local. HM is the harmonic mean of Local, Base, and Novel accuracy.

PST / Text	Local	Base	Novel	HM
S / S	**99.87**	**91.20**	**85.67**	**91.88**
L / S	99.81	85.90	82.87	88.95
S / L	99.86	89.39	83.65	90.48
L / L	99.84	84.32	83.76	88.72

Variant	Local	Base	Novel	HM
LVPM lifecycle				
L1	96.64	73.55	69.24	78.15
L2	97.19	76.69	72.30	80.73
SACS gate design				
G1	97.13	75.29	68.93	78.77
G2	97.10	75.33	69.46	79.01
G3	97.22	75.08	70.07	79.20
G4	97.35	75.76	71.64	80.15
Full pFedVSA (L3/G5)	**97.45**	**77.08**	**72.74**	**81.11**

Table A4: Average summary of the LVPM lifecycle and SACS gate over SUN397 and DTD. Local, Base, and Novel accuracies are averaged across the two datasets, and HM is the harmonic mean of the three reported averages. L1 uses historical cumulative prototypes, L2 uses cross-round EMA prototypes, and L3 uses round-wise recomputation. G1–G5 denote fixed, support-only, similarity-only, learned, and full non-parametric gates, respectively. The full pFedVSA adopts L3 and G5. Lifecycle rows fix G5, whereas gate rows fix L3, isolating memory-update and reliability effects.

	SUN397				DTD			
Variant	Local	Base	Novel	HM	Local	Base	Novel	HM
L1	94.17	75.32	76.09	81.00	99.11	71.77	62.39	74.90
L2	95.13	78.38	79.91	83.84	99.24	74.99	64.68	77.18
L3 (Ours)	**95.40**	**78.82**	**80.28**	**84.21**	**99.49**	**75.33**	**65.19**	**77.59**

Table A5: Effect of the LVPM lifecycle on SUN397 and DTD. L1 uses historical cumulative prototypes with exposure counts and online updates; L2 uses cross-round EMA prototypes with unique support counts; and L3 performs round-wise recomputation with unique support counts and fixed in-round memory. HM denotes the harmonic mean of Local, Base, and Novel accuracy.

A.4 Additional Design and Diagnostic Analyses

	SUN397			DTD			Avg.
Variant	Base	Novel	H_{BN}	Base	Novel	H_{BN}	H_{BN}
M1	70.99	71.54	71.26	69.45	61.40	65.18	68.22
M2	75.31	76.56	75.93	73.81	62.31	67.57	71.75
M3	**78.82**	**80.28**	**79.54**	**75.33**	**65.19**	**69.89**	**74.72**

	SUN397				DTD			
Variant	Local	Base	Novel	HM	Local	Base	Novel	HM
G1	95.09	76.65	75.37	81.45	99.17	73.92	62.48	75.73
G2	95.05	76.78	76.49	81.93	99.14	73.88	62.43	75.68
G3	95.24	76.07	76.31	81.63	99.20	74.08	63.82	76.44
G4	95.32	77.33	78.90	83.11	99.37	74.18	64.38	76.77
G5 (Ours)	**95.40**	**78.82**	**80.28**	**84.21**	**99.49**	**75.33**	**65.19**	**77.59**

Table A6: Diagnostic comparison of SACS gate designs on SUN397 and DTD. G1 uses the best fixed gate; G2 uses support sufficiency only; G3 uses image–prior similarity only; G4 uses a learned gate; and G5 is the full non-parametric SACS combining support sufficiency and image–prior consistency. HM denotes the harmonic mean of Local, Base, and Novel accuracy.

Table A7: Comparison with simpler local visual-memory alternatives. M1: direct visual-prior classification; M2: local Tip-Adapter/cache fusion; M3: full pFedVSA. All variants use the same round-wise LVPM and retain the shared adapted branch as the fallback for locally unobserved and novel classes. H_{BN} is the harmonic mean of Base and Novel accuracy and isolates transferable recognition in this diagnostic comparison.

Semantic-module sharing. Table A3 isolates the effect of sharing PST and the text adapter. Sharing both modules gives the best HM of 91.88%. Keeping PST local lowers HM by 2.93 points, keeping the text adapter local lowers it by 1.40 points, and keeping both local reduces HM to 88.72%. Local accuracy changes little across the four variants, while the larger differences occur on Base and Novel recognition. This supports federating the visual-to-semantic mapping and semantic adapter while keeping the image adapter and class-level visual priors client-local.

Summary of LVPM and SACS design choices. Table A4 summarizes the average effect of the LVPM lifecycle and SACS gate over SUN397 and DTD. Cross-round EMA improves HM from 78.15% to 80.73%, while round-wise recomputation further reaches 81.11%. For SACS, fixed and single-cue gates are consistently weaker than the learned gate, and the full non-parametric gate achieves the highest average HM of 81.11%. The following two tables provide the corresponding dataset-level breakdown.

LVPM lifecycle. Table A5 compares three strategies for maintaining the local visual prior memory. Historical cumulative updating (L1) obtains HM values of 81.00% on SUN397 and 74.90% on DTD. Cross-round EMA with unique support counts (L2) improves them to 83.84% and 77.18%, respectively. The proposed round-wise recomputation strategy (L3) performs best on both datasets, reaching 84.21% and 77.59% HM. These results support rebuilding LVPM from unique samples at the beginning of each round rather than accumulating stale representations across model states.

SACS gate design. Table A6 compares fixed, single-cue, learned, and full non-parametric gates. The full SACS gate (G5) reaches 84.21% HM on SUN397 and 77.59% on DTD, outperforming the learned gate (G4) by 1.10 and 0.82 points, respectively. Under the balanced setting, support-only and similarity-only gating perform similarly, whereas combining both cues yields the strongest overall result. The heterogeneous-support experiment in the main paper further tests the support term when class-wise support counts actually vary.

Comparison with simpler alternatives. Table A7 compares pFedVSA with direct visual-prior classification and local cache-based fusion. Both alternatives obtain lower H_{BN} on SUN397 and DTD. The full method improves the average H_{BN} from 68.22% for direct visual-prior classification and 71.75% for cache fusion to 74.72%. This indicates that the gains cannot be explained solely by adding a local prototype or feature cache; transforming local visual evidence into semantic signals and controlling its contribution are both useful.

Method	Trainable Params. ↓	Comm. Params. (Up/Down) ↓	Avg. HM (%) ↑
PromptFL	**8,192**	8,192/8,192	83.90
FedPGP	12,416	8,192/8,192	79.98
pFedMoAP	74,240	8,192/73,728	72.45
pFedMMA	248,832	**3,072/3,072**	84.55
pFedVSA (Ours)	186,368	35,840/35,840	**90.13**

Table A8: Comparison of locally trainable parameters, per-round communication, and predictive performance under CLIP ViT-B/16. Communication parameters are reported as uplink/downlink, and parameter counts exclude bias terms. Average HM is computed over the seven main base-to-new datasets. Communication counts transmitted tensors and excludes client-local image-adapter and LVPM states.

A.5 Parameter and Communication Statistics

Communication and parameter statistics. Table A8 reports locally trainable parameters, per-round communication, and average HM. pFedVSA communicates only the shared text adapter and PST, corresponding to 35,840 parameters in each direction per participating client and round; the image adapter and LVPM remain local. This communication cost is higher than that of PromptFL, FedPGP, and pFedMMA, so we do not claim a communication-efficiency advantage. The table is included to make the additional cost of the shared visual-to-semantic mapping explicit alongside the resulting predictive performance.

A.6 Loss-Weight Sensitivity

Tables A9 and A10 report the sensitivity to the two loss weights. Each result is averaged over three independent runs with different random seeds. In each sweep, all hyperparameters except the evaluated weight remain fixed. A weight of zero removes the corresponding loss, whereas $\lambda_{\text{anc}} = 0.10$ and $\lambda_{\text{kd}} = 0.50$ define the full pFedVSA configuration. HM denotes the harmonic mean of Local, Base, and Novel accuracy.

λ_{anc}	Local	Base	Novel	HM
0	99.72	90.01	84.02	90.80
0.01	99.74	90.67	84.81	91.33
0.05	99.76	89.61	85.02	91.06
0.10	**99.87**	**91.20**	**85.67**	**91.88**
0.20	99.48	88.94	84.17	90.43

Table A9: Sensitivity to the semantic anchoring weight λ_{anc}, averaged over DTD, Caltech101, Flowers102, and OxfordPets. The default value is 0.10.

At the default $\lambda_{\text{anc}} = 0.10$, pFedVSA achieves the highest HM of 91.88. Removing semantic anchoring reduces Novel accuracy from 85.67 to 84.02 and HM from 91.88 to 90.80. Both smaller and larger anchoring weights yield lower HM than the default value.

λ_{kd}	Local	Base	Novel	HM
0	99.78	89.91	84.24	90.87
0.10	99.79	90.33	84.66	91.18
0.50	**99.87**	**91.20**	**85.67**	**91.88**
1.00	99.73	89.48	85.11	91.04
2.00	99.36	88.72	84.02	90.26

Table A10: Sensitivity to the distillation weight λ_{kd}, averaged over DTD, Caltech101, Flowers102, and OxfordPets. The default value is 0.50.

The default $\lambda_{\mathrm{kd}} = 0.50$ also gives the highest HM of 91.88. Setting λ_{kd} to zero reduces Base accuracy from 91.20 to 89.91, Novel accuracy from 85.67 to 84.24, and HM from 91.88 to 90.87. Increasing the weight beyond 0.50 reduces HM to 91.04 at 1.00 and 90.26 at 2.00.

Overall, moderate values of both regularization weights provide the best balance, and performance degrades when either term is removed or overweighted.