FOCUS:UNKNOWN-AWARE PROMPTING AND SEMANTIC INJECTION FOR OPEN-SET DOMAIN GENERALIZATION IN FEDERATED SETTINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce Federated Open-Set Domain Generalization (FODG), a novel paradigm that requires decentralized models to reconcile domain shifts while identifying novel categories. While prompt-based federated learning has advanced domain generalization, existing methods predominantly operate under a closed-set assumption and falter in data-scarce regimes, leading to catastrophic misclassification of unknown samples. To address these challenges, we propose FOCUS (**F**ederated **O**pen-Set **C**ross-Domain **U**nknown **S**eparation), an innovative federated prompt-learning framework with two core breakthroughs. Specifically, FOCUS reformulates the task as a unified $|\mathcal{C}| + 1$ prompt learning problem by introducing an additional unknown-aware prompt, enabling federated vision-language models to learn transferable decision boundaries between known and unseen categories across decentralized domains. Furthermore, to combat severe domain shifts in data-scarce federated settings, we develop a cross-attention-based attribute-semantic injection mechanism that dynamically grounds textual attributes into instance-specific visual features. By synergizing this mechanism with a hierarchical multi-modal prompt learner and a class-diversity regularization that enforces orthogonal semantic representations, our framework facilitates robust domain-invariant adaptation and strictly prevents feature entanglement between known and novel categories. Extensive experiments on five benchmark datasets demonstrate that FOCUS consistently outperforms state-of-the-art methods under challenging few-shot federated settings. Code is available at https://anonymous.4open.science/r/FOCUS-2104.

1 INTRODUCTION

Federated learning (FL) has emerged as a machine learning paradigm that enables multiple clients to collaboratively train a global model without exchanging raw data, thus overcoming data isolation while preserving privacy (McMahan et al., 2017). However, in practical applications, traditional FL methods encounter a critical challenge: they typically overlook the prevalent domain shift (Tan et al., 2023) between training data and future test data. Due to significant discrepancies in data distributions, this oversight often leads to severe performance degradation when the model is deployed on completely unseen target domains. To tackle this formidable bottleneck, the principles of Domain Generalization (DG) have been introduced into decentralized settings, naturally giving rise to Federated Domain Generalization (FDG) (Liu et al., 2021)

To address the domain shift problem in FL, some methods have been proposed, establishing the paradigm of FDG. Early methods primarily focus on feature and data level alignment. Specifically, Liu et al. (2021) shares the amplitude spectrum of images among local clients to enrich data diversity. Nguyen et al. (2022) utilizes local regularizers to achieve implicit representation alignment. Zhang et al. (2024) incorporates GANs to perform targeted data augmentation. Gong et al. (2026); Wu et al. (2025) tackle this challenge through prompt tuning by dynamically generating domain-specific soft prompts and employing attention-based mechanisms for global aggregation.

Despite their progress, current methodologies rely on a restrictive closed-set assumption, failing to consider open classes in the target domain. In real-world environments, confronting novel classes in-

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.

evitably causes these models to suffer from severe class mismatch (Shu et al., 2021a). This limitation necessitates a shift towards Federated Open-Set Domain Generalization (FODG). To overcome the severe class mismatch of open-set domain generalization, we propose a novel framework, FOCUS, with dual innovations at both the prompt-architecture and optimization levels. To close the domain gap between decentralized source clients, we avoid updating massive foundation model weights but learn generalizable semantic representations across domains by prompt tuning (Guo et al., 2024). To mitigate large domain shifts, we optimize residual prompts anchored on specific semantics and introduce an Attribute-Semantic Injection Module. To enhance the fine-grained discriminability across different classes, we apply a class-separation loss ($\mathcal{L}_{class}$) on the textual embeddings to explicitly maximize inter-class distances. Ultimately, FOCUS establishes a robust continuous prompt space that embeds the shared knowledge of all source domains. By collaboratively optimizing these lightweight prompts, our framework effectively calibrates predictive uncertainty. This enables the model to accurately classify known categories while rejecting novel classes in completely unseen target domains. In summary, our main contributions are highlighted as follows:

- We highlight a new and practical problem: *Federated Open-Set Domain Generalization (FODG)* , which learns from decentralized source clients with heterogeneous distributions to generalize to an unseen target domain containing novel open classes.
- We enhance semantic generalizability across decentralized domains by designing residual prompts coupled with an Attribute-Semantic Injection Module. Additionally, we introduce a class-diversity loss ($\mathcal{L}_{class}$) at the optimization level to explicitly maximize fine-grained inter-class distances, calibrating the model for robust open-class rejection.
- We perform extensive experiments across five benchmark datasets under challenging, data-scarce federated settings. Empirical results demonstrate that FOCUS establishes a new state-of-the-art, outperforming the strongest baseline by a substantial margin of 12.09% in average H-score, while drastically reducing communication overhead to merely 24K parameters per round.

2 Related Work

VLMs and prompt learning Vision-Language foundation models such as CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), LiT (Zhai et al., 2022), and Florence (Xiao et al., 2024) have made significant strides in image recognition by harnessing large-scale aligned image-text data to construct robust multi-modal feature spaces. While these models demonstrate powerful open-vocabulary recognition, tailoring them to specific target distributions while preserving their broad generalization ability remains a non-trivial optimization problem. To address this, prompt tuning has emerged as a parameter-efficient paradigm to customize these models for specific tasks. Early methods focus on the textual branch, replacing manual prompt engineering with continuous learnable vectors (e.g., CoOp (Zhou et al., 2022b), CoCoOp (Zhou et al., 2022a), and PromptSRC (Li et al., 2025)). Subsequent research has expanded this paradigm to capture richer representations. For instance, VPT (Jia et al., 2022) introduces learnable parameters into deep visual layers, demonstrating the superiority of hierarchical prompting. Extending this to both modalities, MaPLe (Khattak et al., 2023) proposes synchronized multi-modal deep prompting to ensure optimal alignment between the vision and language branches.

(Open)DG DG aims to address domain shifts by training models across diverse source domains, thereby enhancing performance on unseen targets (Blanchard et al., 2011; Kang et al., 2022; Zhou et al., 2020). Methods such as self-supervised learning (Carlucci et al., 2019), ensemble learning (Xu et al., 2014), and meta-learning (Ma & Li, 2022; Li et al., 2017a; 2019a;b; Patricia & Caputo, 2014) have been developed to tackle the DG problem. Furthermore, the concept of domain augmentation (Li et al., 2021; Foley et al., 2025; Kang et al., 2022; Zhou et al., 2020), which generates pseudo-samples to be included in the training set, was introduced to mitigate domain disparity. Subsequently, Shu et al. (2021a) pioneered Open Domain Generalization (ODG) through domain-augmented meta-learning. MEDIC (Wang et al., 2023) further proposed domain-wise and class-wise gradient matching, which aims to address the biased decision boundary in (Shu et al., 2021a). Unlike these conventional models, our work explores Vision-Language Models (VLMs) to solve the ODG task from the perspective of prompt processing, efficiently tackling the associated open-set challenges.

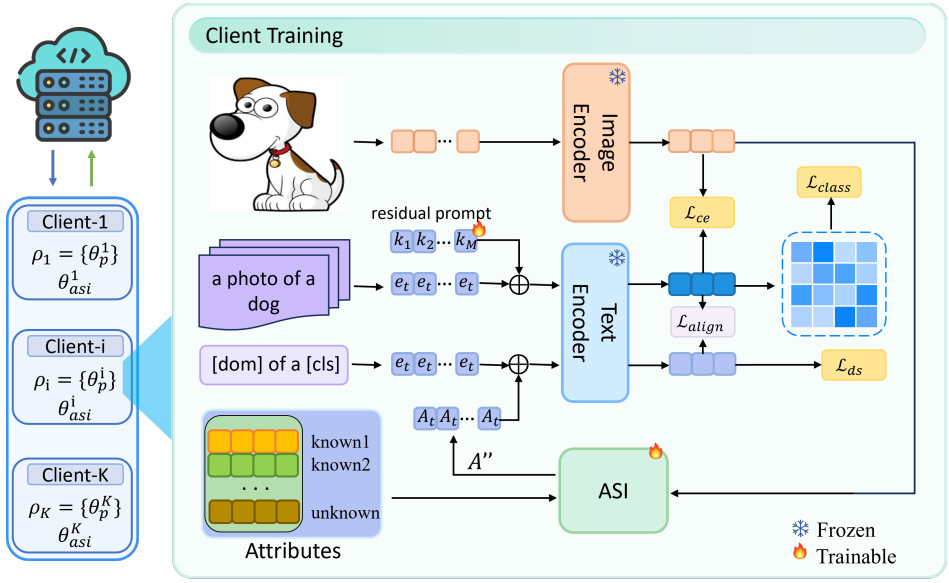

Figure 1: Overview of FOCUS. Clients locally train prompts Θ_p and the ASI module Θ_{asi}. To preserve domain privacy, the server aggregates only Θ_p. Meanwhile, the localized ASI employs cross-attention to fuse visual features with textual attributes, generating dynamic prompts for robust open-set adaptation.

Federated prompt learning and FDG Federated prompt tuning (FPL) has emerged as an efficient and privacy-preserving paradigm to adapt Vision-Language Models across heterogeneous clients by transmitting only lightweight prompts (Guo et al., 2024; Ma et al., 2024; Singha et al., 2025). However, the inherent data heterogeneity across isolated clients inevitably induces severe distribution shifts, propelling the critical challenge of Federated Domain Generalization (Liu et al., 2021). To mitigate this, a growing body of literature has leveraged FPL to address the FDG problem (Gong et al., 2026; Bai et al., 2024; Wu et al., 2025; Zheng et al., 2025). These approaches effectively utilize learnable prompts to capture domain-invariant semantics without exposing local data. Despite this progress, existing FPL-based FDG frameworks fundamentally rely on a strict closed-set assumption, making them vulnerable to novel categories in open-world scenarios. To date, addressing ODG within a federated framework (FODG) remains a critical void.

3 Task Formulation

We consider a Federated Open-Set Domain Generalization (FODG) scenario involving K decentralized clients. Each client $k \in \{1, \ldots, K\}$ possesses a private local dataset $\mathcal{D}_k = \{(\mathbf{x}_i^k, y_i^k)\}_{i=1}^{n_k}$ where $\mathbf{x}_i^k \in \mathcal{X}_k$ represents input images and $y_i^k \in \mathcal{Y}_k$ are the corresponding labels. Due to the inherent heterogeneity in FL, these local datasets are drawn from unique distributions ($\mathcal{P}_k \neq \mathcal{P}_{k'}$ for $k \neq k'$) and exhibit heterogeneous label spaces with a mix of shared and domain-specific classes. The global set of known classes is collectively represented by $\mathcal{C} = \bigcup_{k=1}^{K} \mathcal{Y}_k$. To reflect real-world constraints, each client typically operates in a data-scarce regime for its local classes.During testing, the collaborative model is evaluated on an unlabeled target domain $\mathcal{T}$ with dataset $\mathcal{D}_t = \{\mathbf{x}_j^t\}_{j=1}^{n_t}$, where the target distribution $\mathcal{P}_t$ differs from all local distributions $\{\mathcal{P}_k\}_{k=1}^{K}$. The target label set $\mathcal{Y}_t$ comprises both known classes $\mathcal{Y}_t^{known} = \mathcal{C}$ and novel outlier categories $\mathcal{Y}_t^{novel} = \mathcal{Y}_t \setminus \mathcal{C}$.

To address the challenges of FODG, we reformulate the problem into a unified $|\mathcal{C}| + 1$ classification framework within the CLIP architecture. We design learnable prompts for each of the $|\mathcal{C}|$ known classes and an additional Unknown-class prompt. This framework facilitates the cohesive identification of outlier instances during inference through decentralized collaborative learning, without requiring any access to open-set knowledge during the local training phase.

4 METHODOLOGY

4.1 FOCUS OVERVIEW

As illustrated in Figure 1, FOCUS aims to collaboratively adapt frozen Vision-Language Models for Federated Open-Set Domain Generalization. To achieve this, our framework is driven by two core modules meticulously designed to overcome the inherent bottlenecks of open-world federated learning:

Combating Unknown Collapse via Unknown-Aware Prompting. First, standard federated prompt tuning inherently operates under a closed-set assumption. Consequently, known classes completely dominate the federated prompt space. When confronted with unseen target domains, this closed-set monopoly forces the model into *unknown collapse*, producing over-confident misclassifications for novel outliers. To fundamentally resolve this, we introduce the **Unknown-Aware Prompt Learning** module. By reformulating the task into a unified $|\mathcal{C}|+1$ classification framework and synthesizing a distinct open-set semantic anchor, this module establishes a definitive rejection boundary to actively calibrate predictive uncertainty.

Mitigating Prompt Drift via Attribute-Semantic Injection. Deploying prompt tuning across heterogeneous few-shot clients inevitably induces severe *prompt drift*, crippling the model's generalization on unseen domains. To counteract this, our **Attribute-Semantic Injection** module dynamically grounds textual semantic attributes within instance-specific visual representations. Enriching local prompts with these grounded visual semantics effectively mitigates prompt drift and extracts robust, domain-invariant representations.

These two synergistic components are collaboratively optimized through a selective communication strategy and unified objective functions, preserving domain-specific privacy and minimizing bandwidth overhead while maximizing open-set domain generalization.

4.2 UNKNOWN-AWARE PROMPT LEARNING

To explicitly establish the rejection boundary for novel outliers mentioned above, we propose an unknown-aware prompt learning mechanism. By capturing deterministic knowledge of known classes and synthesizing a distinct open-set anchor, we adapt the frozen Vision-Language Model by optimizing a hierarchical prompt structure.

Shallow Prompt The shallow prompt is integrated into the input layer of the text encoder to provide the initial task-specific context. To explicitly bridge the gap between hand-crafted priors and adaptive learning, the prompt $P_{shallow}$ is decomposed into a fixed semantic base and a learnable residual component:

$$P_{shallow} = [SOS, \mathbf{v}_1, \ldots, \mathbf{v}_M, \mathbf{c}, EOS] \tag{1}$$

where $\mathbf{c}$ denotes the class name embedding. The context tokens are represented as a zero-initialized learnable residual added to a deterministic anchor:

$$[\mathbf{v}_1, \ldots, \mathbf{v}_M] = \text{Emb}(\text{“a photo of a”}) + [\mathbf{k}_1, \ldots, \mathbf{k}_M] \tag{2}$$

Deep Multimodal Prompt To capture complex hierarchical features, we introduce independent learnable tokens into both the textual and visual branches across multiple transformer layers (Khattak et al., 2023). This layer-wise adaptation yields robust representations $\mathbf{z}_{txt}$ and $\mathbf{z}_{vis}$ from the final layer K:

$$\mathbf{z}_{txt} = \text{TextProj}(\mathbf{w}_K^N), \quad \mathbf{z}_{vis} = \text{VisProj}(\mathbf{e}_K^{cls}) \tag{3}$$

Unified Rejection Boundary Lacking access to open-set data during training, we leverage LLMs to generate base semantic attributes ($\mathcal{A}_{known}$) for known classes. To represent disjoint outliers, an open-set anchor ($\mathbf{a}_{unk}$) is synthesized as the negative of the known attributes' global average.

$$\mathbf{a}_{unk} = -\frac{1}{|\mathcal{C}|}\sum_{i=1}^{|\mathcal{C}|}\mathbf{a}_i \tag{4}$$

This expands our prompt space to $|\mathcal{C}|+1$ dimensions. The classification probability is thus computed over this unified space using cosine similarity:

$$p(y=c|\mathbf{x}) = \frac{\exp(\text{sim}(\mathbf{z}_{vis}, \mathbf{z}^c_{txt})/\tau)}{\sum_{i=1}^{|\mathcal{C}|+1} \exp(\text{sim}(\mathbf{z}_{vis}, \mathbf{z}^i_{txt})/\tau)} \tag{5}$$

where τ is a learnable temperature. Target instances lacking the deterministic features of known classes are naturally attracted to the $\mathbf{a}_{unk}$ anchor, actively calibrating predictive uncertainty for robust open-set rejection.

4.3 Attribute-Semantic Injection

In decentralized federated scenarios, clients typically suffer from extreme data scarcity and heterogeneous distributions. Naively tuning local prompts under such conditions inevitably induces severe "prompt drift", resulting in weak domain invariance. To counteract this, we propose the Attribute-Semantic Injection (ASI) module. By dynamically grounding the augmented textual attributes $\mathcal{A}$ within instance-specific visual representations, ASI extracts robust, domain-invariant features.

Visual-Semantic Grounding via Cross-Attention To enrich static domain-specific prompts with fine-grained visual semantics, we employ a cross-attention mechanism. Specifically, using the projected visual features $\mathcal{F}_v(\mathbf{x})$ as Queries and the encoded augmented attributes $\mathcal{F}_t(\mathcal{A})$ as Keys and Values, we compute an attribute-enhanced embedding $\mathcal{A}'(\mathbf{x})$. Rather than isolating modalities, this mechanism explicitly anchors semantic attributes to the instance-specific visual representation, absorbing nuanced, localized visual details to combat domain shift.

Dynamic Prompt Integration For a source domain s, the base prompt is defined as $P_s^y = \text{Emb}(\text{"[Domain] of a [CLS]"})$. To construct the final dynamic prompt, we first derive a class-agnostic attribute encoding $\mathcal{A}''(\mathbf{x})$ by averaging the attribute-enhanced embeddings $\mathcal{A}'(\mathbf{x})$ across all $|\mathcal{C}|+1$ classes. This visual-semantic encoding is then integrated into the base prompt via token-wise addition:

$$P_s^y(\mathbf{x}) = P_s^y + \mathcal{A}''(\mathbf{x}) \tag{6}$$

These dynamic prompts become highly adaptive to local data distributions and subsequently serve as structured targets to regularize the global domain-agnostic prompts during the local optimization process.

4.4 Overall Training Strategy of FOCUS

4.4.1 Optimization Objectives

To ensure the constructed dynamic prompts $P_s^y(\mathbf{x})$ fully absorb fine-grained visual cues, we optimize them using a domain-specific contrastive loss $\mathcal{L}_{ds}$:

$$\mathcal{L}_{ds} = -\log \frac{\exp(\text{sim}(\mathbf{z}_{vis}, \mathcal{F}_t(P_s^y(\mathbf{x})))/\tau)}{\sum_{j=1}^{\mathcal{C}} \exp\big(\text{sim}(\mathbf{z}_{vis}, \mathcal{F}_t(P_s^j(\mathbf{x})))/\tau\big)}, \tag{7}$$

where $\text{sim}(\cdot,\cdot)$ is the cosine similarity and τ is the learnable temperature. To transfer this localized knowledge, we regularize the global domain-agnostic prompt P_{gen} by aligning it with the enriched domain-specific prompt via $\mathcal{L}_{align}$:

$$\mathcal{L}_{align} = 1 - \text{sim}\left(\mathcal{F}_t(P_{gen}), \mathcal{F}_t(P_s^y(\mathbf{x}))\right) \tag{8}$$

Next, we optimize the domain-agnostic prompts for the primary classification task. Crucially, adhering to our unknown-aware formulation, the cross-entropy loss (Radford et al., 2021) $\mathcal{L}_{ce}$ is computed over the unified $|\mathcal{C}|+1$ space:

$$\mathcal{L}_{ce} = -\log \frac{\exp\left(\text{sim}(\mathbf{z}_{vis}, \mathbf{z}^y_{txt})/\tau\right)}{\sum_{i=1}^{|\mathcal{C}|+1} \exp\left(\text{sim}(\mathbf{z}_{vis}, \mathbf{z}^i_{txt})/\tau\right)} \tag{9}$$

where y is the ground-truth label.

Finally, although pre-trained text embeddings offer rich semantics, their high intrinsic correlation often blurs decision boundaries. To explicitly maximize the fine-grained inter-class margin and prevent feature collapse, we introduce a class diversity loss $\mathcal{L}_{class}$. By forcing the text similarity matrix

towards an identity matrix, this objective penalizes off-diagonal correlations, ensuring distinct and orthogonal category representations:

$$\mathcal{L}_{class} = \text{MSE}(\mathbf{Z}_{txt}\mathbf{Z}_{txt}^{\top}, \mathbf{I}) \tag{10}$$

where $\mathbf{Z}_{txt}$ is the aggregated textual feature matrix and $\mathbf{I}$ is the identity matrix. The overall objective function optimized by each local client is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \mathcal{L}_{ds} + \mathcal{L}_{align} + \lambda\mathcal{L}_{class} \tag{11}$$

where λ is hyperparameter ensuring both alignment and consistency.

4.4.2 Federated Optimization Protocol

Our framework structurally decouples parameters into multimodal prompts θ_p and the ASI module θ_{asi}. We selectively transmit only θ_p for global aggregation, keeping θ_{asi} strictly localized to balance global sharing and local specialization. Following the server's distribution of initialized global prompts θ_p^0, the training alternates as follows:

1. **Local Training and Upload:** At communication round r, each client $k \in \{1, \ldots, K\}$ optimizes its joint parameter set $\theta_k^r = \{\theta_{p,k}^r, \theta_{asi,k}^r\}$ over E local epochs using the AdamW optimizer with learning rate η:

$$\theta_{k,\text{local}}^{r+1} = \text{AdamW}(\eta, \theta_k^r, \mathcal{D}_k, \mathcal{L}_{total}) \tag{12}$$

   After local training, client k uploads only the updated prompt parameters $\theta_{p,k}^{r+1}$ to the server (Guo et al., 2024). This strict modality decoupling inherently protects the domain-specific privacy encapsulated within $\theta_{asi,k}$ and bypasses heavy communication bottlenecks.

2. **Server Aggregation and Distribution:** The server aggregates the prompt parameters from all K clients using a weighted average (McMahan et al., 2017) to obtain the new global prompts θ_p^{r+1}:

$$\theta_p^{r+1} = \sum_{k=1}^{K} \frac{|\mathcal{D}_k|}{|\mathcal{D}|} \theta_{p,k}^{r+1} \tag{13}$$

   where $|\mathcal{D}| = \sum_{k=1}^{K} |\mathcal{D}_k|$ is the total number of samples across all clients. The server then distributes θ_p^{r+1} back to each client for the next round.

3. **Iteration:** Repeat Step 1 and Step 2 until the global model reaches convergence or the maximum number of communication rounds is met.

The detailed pseudo-code of our proposed FOCUS pipeline is deferred to the Appendix.

5 Experiments

Datasets. We evaluate FOCUS on five benchmark datasets:OfficeHome (Venkateswara et al., 2017), PACS (Li et al., 2017b), VLCS (Fang et al., 2013), Mini-DomainNet (Peng et al., 2019), following standard known-new class splits (Shu et al., 2021a; Singha et al., 2024), but considering few training samples per class.

Implementation details. For all models including ours, Vit-B / 32 is implemented as the visual backbone F_v, with a Transformer (Vaswani et al., 2017) as a textual encoder F_t.Specifically, we implement deep prompts across 9 transformer layers, while the learnable residual prompt in the shallow layer is configured with a context length of 4 tokens.

Training and evaluation protocols. Training is performed over 10 epochs with the AdamW optimizer (Loshchilov & Hutter, 2017).Batch sizes are set per dataset: 6 for PACS/VLCS, 9 for OfficeHome, Multi-Dataset,and Mini-Domainnet, with each batch incorporating three pseudo-open samples from each source domain. During the synthesis of pseudo-open samples, a threshold of 0.2 on the grey-value entropy is applied to discard irrelevant images.To assess our model's performance, we adhere to the leave-one-domain-out protocol, employing Top-1 accuracy(Acc) for the closed-set

classes and the harmonic mean(H-score) for the combined closed-set and open-set performance. All reported results are averaged across all target domains over three independent random seeds.

Baselines. Given the absence of existing FODG methods in the literature, we design following baselines by adapting state-of-the-art approaches to our setting: 1) the leading centralized prompt tuning methods PromptSRC (Li et al., 2025), MaPLe (Khattak et al., 2023); 2) the foremost native federated prompt frameworks promptFL (Guo et al., 2024), FedMVP (Singha et al., 2025); and 3) the premier centralized ODG methods ODG-CLIP (Singha et al., 2024), OSLoPROMPT (N C et al., 2025) extended to the federated regime to evaluate open-set generalization performance. Furthermore, as demonstrated by (Singha et al., 2024) and (N C et al., 2025), generative data augmentation via Stable Diffusion (SD) (Rombach et al., 2022) can effectively endow models with superior open-set recognition capabilities.To ensure a strictly fair comparison, we apply this SD-based augmentation strategy to synthesize pseudo-open samples for all evaluated baselines, as well as our proposed method, during their local training phases.

Table 1: Comparative analysis across five datasets in a two-shot setting, reporting both the average closed-set accuracy (Acc) and the harmonic mean (H-score) over all domain combinations following a leave-one-domain-out protocol. To account for the inherent trade-off between closed-set accuracy and open-set recognition, the H-score serves as the primary metric to comprehensively evaluate the overall capability of the models (Scheirer et al., 2013; Bendale & Boult, 2015).

Methods	PACS		VLCS		Office-Home		Multi-Dataset		Mini-DomainNet		Average	
	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score
PromptFL	94.09	73.26	77.97	64.24	72.36	4.46	80.03	8.68	80.37	2.29	82.04	15.93
FedpromptSRC	**94.67**	93.79	76.86	76.32	75.09	9.22	82.80	14.06	77.92	12.87	81.47	41.25
FedMaPLe	93.61	93.49	**85.07**	76.00	**81.97**	18.85	**84.87**	31.10	**83.13**	26.37	**85.73**	49.16
FedMVP	95.30	43.54	84.17	46.14	75.70	51.50	80.61	61.59	80.89	18.22	83.33	44.20
FedODG-CLIP	82.65	81.91	67.27	64.26	56.01	41.27	63.00	69.64	54.35	44.62	64.66	60.34
FedOSLoPROMPT	93.54	93.49	79.67	75.07	56.91	41.27	73.45	75.01	78.12	48.75	76.34	66.72
FOCUS(Ours)	94.03	**96.26**	81.42	**80.18**	68.31	**73.43**	79.10	**80.20**	77.06	**64.00**	79.98	**78.81**

5.1 Discussion on the main results

Evaluating the critical trade-off between closed-set accuracy and open-set recognition (Scheirer et al., 2013; Bendale & Boult, 2015) reveals the limitations of current methods, as reported in Table 7. While existing baselines maintain high closed-set accuracy, they severely overfit the known label space and fail drastically on open-set outliers (e.g., PromptFL yields a 0.0 H-score on PACS). FOCUS decisively resolves this performance bottleneck. First, by introducing the unknown-aware anchor ($\mathbf{a}_{unk}$), we break the closed-set monopoly and establish a definitive boundary that actively repels unseen novelties. Simultaneously, our Attribute-Semantic Injection (ASI) module combats the prompt drift caused by data heterogeneity, dynamically grounding textual attributes into robust, domain-invariant visual features. By successfully harmonizing known-class discrimination with robust outlier rejection, FOCUS achieves a state-of-the-art average H-score of 78.81%—outperforming the leading competitor by 12.09%—and demonstrates superior resilience in open-world federated environments.

Table 2: Ablation for Office-Home and Multi-dataset in a 2-shot setting.

Methods	Office-Home			Multi-dataset		
	Acc	Open-acc	H-score	Acc	Open-acc	H-score
w/o Shallow	64.57	75.77	71.69	76.69	77.65	76.96
w/o Deep	67.55	69.60	67.87	75.62	72.07	73.65
w/o a_{unk}	64.50	77.44	69.83	77.19	78.69	77.78
w/o ASI	67.84	76.75	71.63	75.24	77.72	76.22
w/o L_{class}	65.30	78.60	71.12	76.75	76.32	76.36
w/o L_{align}	65.54	77.27	70.91	78.62	79.41	78.84
w/o L_{ds}	66.77	80.30	71.37	75.25	74.74	74.72
Full(ours)	**68.31**	**80.75**	**73.43**	**79.10**	**81.43**	**80.20**

5.2 Ablation analysis

Analysis on core components. We conduct ablation studies on Office-Home and Multi-dataset (Table 2) to validate our dynamic prompt adaptation premise. 1) First, decoupling the hierarchical prompt structure (*w/o Deep*) proves inflexible for complex domain shifts, leading to up to a 5.56% H-score decrease and a 6.55% decrease on Multi-dataset. 2) Next, removing the Attribute-Semantic Injection module (*w/o ASI*) forces optimization of isolated textual prompts without cross-attention visual grounding, noticeably compromising overall recognition. 3) Additionally, omitting specific optimization objectives ($\mathcal{L}_{class}$, $\mathcal{L}_{align}$, $\mathcal{L}_{ds}$) uniformly degrades feature alignment and class separation. In contrast, our full FOCUS systematically resolves these limitations by synergizing hierarchical multimodal prompts with instance-specific visual attributes. Benefiting from this synergy, the final model substantially boosts comprehensive performance, proving that dynamic, visually-grounded prompt adaptation is vital for heterogeneous federated environments.

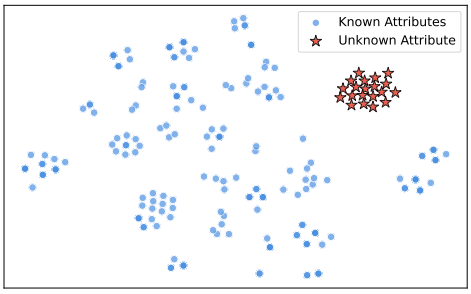

Figure 2: t-SNE (van der Maaten & Hinton, 2008) visualization of the semantic attribute space.

Table 3: Evaluating the impact of different composition strategies for the unknown-aware anchor ($\mathbf{a}_{unk}$) on the Office-Home dataset for robust outlier rejection.

Composition Strategy	Acc	Open-acc	H-score
Orthogonal	63.92	76.82	69.02
Global Mean	59.39	80.70	67.08
Random	63.88	79.65	70.62
Zero	56.79	80.23	65.76
Inverse Mean (Ours)	**68.31**	**80.75**	**73.43**

Impact of the unknown attribute ($\mathbf{a}_unk$). Discarding the open-set anchor ($\mathbf{a}_{unk}$) entirely drops the H-score by 3.60% on Office-Home and 2.4% on Multi-dataset, as the model loses its rejection boundary and confuses novelties with known concepts (Figure 2). Furthermore, to evaluate the composition strategy for $\mathbf{a}_{unk}$, we compare our approach against four alternatives in Table 3. Specifically, the **Zero** strategy provides no semantic guidance, severely collapsing the model's discriminative ability, while a **Random** initialization creates an unstable and arbitrary decision boundary. The **Global Mean** strategy fatally entangles novelties with known categories by averaging known attributes. Alternatively, an **Orthogonal** anchor provides mere passive irrelevance rather than active repulsion. Conversely, our proposed **Inverse Mean** explicitly negates the shared features of known classes to construct a diametrically opposed semantic space. By actively repelling known concepts, our strategy achieves the optimal harmonic balance and yields the highest H-score.

Table 4: Sensitivity analysis of the class diversity loss weight (λ) on the Office-Home dataset.

λ	Acc	Open-acc	H-score
0	61.08	81.41	68.88
0.1	62.96	80.87	68.79
0.25	63.90	81.42	72.17
0.75	64.13	82.53	72.00
1.0	60.93	80.32	69.19
2.0	59.42	79.10	66.38
0.5 (Ours)	**68.31**	**80.75**	**73.43**

Table 5: Comparison of communication overhead per round across different methods. Sharing only lightweight prompts, our methods maximizes communication efficiency.

Methods	Shared Content	Comm. Params
PromptFL	Prompt Only	2048
FedpromptSRC	Prompt Only	46k
FedMaPLe	Prompt + Projector	3.55M
FedMVP	Prompt+Cross-Attn	9.8M
FedODG-CLIP	Prompt + Projector	16.32M
FEDOSLoPROMPT	Prompt + Projector +Cross-Attn	2.8M
Ours	Prompt Only	24K

Sensitivity Analysis of $\mathcal{L}_class$. Table 4 varies the loss weight λ to evaluate its impact on the framework. Intrinsically, $\mathcal{L}_{class}$ effectively enlarges inter-class distances to reduce classification ambiguity, but excessive penalties can disrupt the valuable pre-trained semantic topology. Without

this constraint ($\lambda = 0$), blurred decision boundaries yield a suboptimal 68.88% H-score. As the weight increases, performance steadily improves, peaking at $\lambda = 0.5$ (73.43%) by achieving optimal class separation. However, overly aggressive weights ($\lambda \geq 1.0$) violently over-push feature dispersion. This destroys the foundational visual-textual alignment established during pre-training, causing noticeable performance drops. Thus, setting $\lambda = 0.5$ optimally balances distinct class separation with the preservation of pre-trained semantic knowledge.

5.3 Few-Shot Evaluation

As illustrated in Figure 3, we evaluate the H-score performance under varying training shot configurations (i.e., 2, 4, 8, and 16 shots) across four distinct benchmark datasets: Office-Home, PACS, VLCS, and Multi-Dataset. The results demonstrate that FOCUS consistently outperforms all baseline methods across all evaluated settings, achieving the best overall performance. Furthermore, with the increasing number of shots, our framework exhibits highly stable trajectories across all datasets. This steady capability highlights the model's exceptional adaptability to low-data regimes, ensuring robust open-set domain generalization even under strict data-scarce constraints. This steady capability highlights the model's exceptional adaptability to low-data regimes, ensuring robust open-set domain generalization even under strict data-scarce constraints.

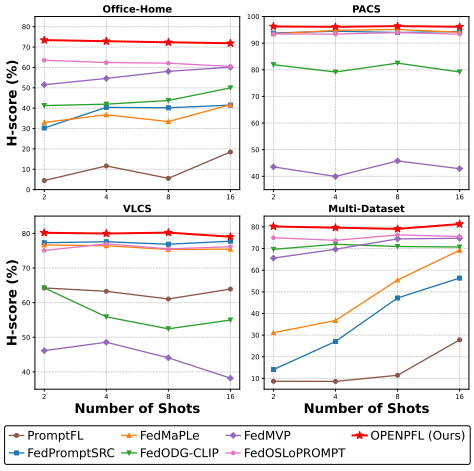

Figure 3: H-score of different methods under 2/4/8/16 shots across four benchmarks.

5.4 Communication Efficiency Analysis

As shown in Table 5, multimodal baselines entangle heavy structural components within their communication payload, transmitting up to 16.32M parameters per round. In contrast, FOCUS employs a strict modality decoupling strategy. By keeping visual operations localized, we transmit **prompt only**, drastically reducing the overhead to merely 24K parameters per round—a 147× to 408× reduction. Crucially, this extreme efficiency directly accelerates optimization. As illustrated in Figure 4, FOCUS exhibits exceptionally rapid convergence and achieves a dominant peak H-score. By collaboratively optimizing only essential semantic anchors, our framework bypasses the arduous alignment of heavy modules, establishing a striking win-win paradigm: state-of-the-art open-set recognition with minimal bandwidth costs.

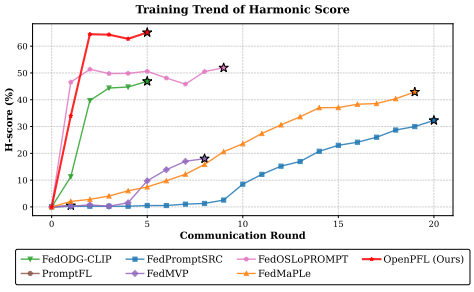

Figure 4: Training trajectories on the Office-Home dataset. H-score trends are shown over training epochs.

6 Conclusion

We propose FOCUS, a parameter-efficient federated prompt learning framework for open-set domain generalization. By localizing the Attribute-Semantic Injection (ASI) module, FOCUS decouples modalities to achieve a strictly **Prompt-Only** communication protocol. Combined with synthetic open-set anchors, this local visual-semantic grounding effectively calibrates predictive uncertainty while bypassing bandwidth bottlenecks. Extensive experiments across five benchmarks demonstrate that FOCUS sets a new state-of-the-art, outperforming baselines by 2.47%–15.25% in H-score under strict privacy and data-scarce constraints.

7 AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs, propose or refine hypotheses, implement methods, and assist with translation. We have not used generative AI tools for other required disclosure tasks, such as generating synthetic data sets, developing theoretical models, designing experiments, or interpreting results, and the rest of the required tasks are not applicable to this work. Additionally, we used generative AI tools to suggest experimental parameters, identify relevant literature, and suggest a structure for the research paper.

REFERENCES

Sikai Bai, Jie Zhang, Song Guo, Shuaicheng Li, Jingcai Guo, Jun Hou, Tao Han, and Xiaocheng Lu. DiPrompT: Disentangled Prompt Tuning for Multiple Latent Domain Generalization in Federated Learning . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27274–27283, Los Alamitos, CA, USA, June 2024. IEEE Computer Society. doi: 10.1109/CVPR52733.2024.02576. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52733.2024.02576.

Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1563–1572, 2015. URL https://api.semanticscholar.org/CorpusID:14240373.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/b571ecea16a9824023ee1af16897a582-Paper.pdf.

Fabio M. Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain Generalization by Solving Jigsaw Puzzles . In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2224–2233, Los Alamitos, CA, USA, June 2019. IEEE Computer Society. doi: 10.1109/CVPR.2019.00233. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2019.00233.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Mark Everingham, S. M. Eslami, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vision*, 111 (1):98–136, January 2015. ISSN 0920-5691. doi: 10.1007/s11263-014-0733-5. URL https://doi.org/10.1007/s11263-014-0733-5.

Chen Fang, Ye Xu, and Daniel N. Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *2013 IEEE International Conference on Computer Vision*, pp. 1657–1664, 2013. doi: 10.1109/ICCV.2013.208.

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. *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pp. 178–178, 2004. URL https://api.semanticscholar.org/CorpusID:2156851.

Sean Foley, Hong Nguyen, Jihwan Lee, Sudarsana Reddy Kadiri, Dani Byrd, Louis Goldstein, and Shrikanth S. Narayanan. Towards disentangling the contributions of articulation and acoustics in multimodal phoneme recognition. *ArXiv*, abs/2505.24059, 2025. URL https://api.semanticscholar.org/CorpusID:279070573.

Shuai Gong, Chaoran Cui, Chunyun Zhang, Wenna Wang, Xiushan Nie, and Lei Zhu. Federated domain generalization via prompt learning and aggregation. *IEEE Transactions on Information Forensics and Security*, 21:1528–1541, 2026. doi: 10.1109/TIFS.2026.3655520.

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, 2024. doi: 10.1109/TMC.2023.3302410.

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 *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pp. 709–727, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-19826-7. doi: 10.1007/978-3-031-19827-4_41. URL https://doi.org/10.1007/978-3-031-19827-4_41.

Juwon Kang, Sohyun Lee, Namyup Kim, and Suha Kwak. Style neophile: Constantly seeking novel styles for domain generalization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7120–7130, 2022. doi: 10.1109/CVPR52688.2022.00699.

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.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI Conference on Artificial Intelligence*, 2017a. URL https://api.semanticscholar.org/CorpusID:1883787.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5543–5551, 2017b. URL https://api.semanticscholar.org/CorpusID:6037691.

Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1446–1455, 2019a.

Jie Li, Yongli Wang, and Bin Song. Self-regulating texture-enhanced prompt learning for vision-language models. In *2025 5th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC)*, pp. 916–922, 2025. doi: 10.1109/ICAIRC68035.2025.11385229.

Pan Li, Da Li, Wei Li, Shaogang Gong, Yanwei Fu, and Timothy M. Hospedales. A simple feature augmentation for domain generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8866–8875, 2021. doi: 10.1109/ICCV48922.2021.00876.

Yiying Li, Yongxin Yang, Wei Zhou, and Timothy Hospedales. Feature-critic networks for heterogeneous domain generalization. In *International conference on machine learning*, pp. 3915–3924. PMLR, 2019b.

Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1013–1023, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Yixiao Ma and Fanzhang Li. Self-challenging mask for cross-domain few-shot classification. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pp. 4456–4463, 2022. doi: 10.1109/ICPR56361.2022.9956314.

Yuting Ma, Lechao Cheng, Yaxiong Wang, Zhun Zhong, Xiaohua Xu, and Meng Wang. Fedhpl: Efficient heterogeneous federated learning with prompt tuning and logit distillation. *ArXiv*, abs/2405.17267, 2024. URL https://api.semanticscholar.org/CorpusID:270063100.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Mohamad Hassan N C, Divyam Gupta, Mainak Singha, Sai Bhargav Rongali, Ankit Jha, Muhammad Haris Khan, and Biplab Banerjee. Osloprompt: Bridging low-supervision challenges and open-set domain generalization in clip. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10110–10120, 2025. doi: 10.1109/CVPR52734.2025.00945.

A. Tuan Nguyen, Philip Torr, and Ser Nam Lim. Fedsr: A simple and effective domain generalization method for federated learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 38831–38843. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/fd946a6c99541fddc3d64a3ea39a1bc2-Paper-Conference.pdf.

Novi Patricia and Barbara Caputo. Learning to learn, from transfer learning to domain adaptation: A unifying perspective. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1442–1449, 2014.

Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

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 International Conference on Computer Vision*, pp. 1406–1415, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:231591445.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, Los Alamitos, CA, USA, June 2022. IEEE Computer Society. doi: 10.1109/CVPR52688.2022.01042. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01042.

Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1):157–173, 2008.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pp. 213–226. Springer, 2010.

Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7): 1757–1772, 2013. doi: 10.1109/TPAMI.2012.256.

Yang Shu, Zhangjie Cao, Chenyu Wang, Jianmin Wang, and Mingsheng Long. Open domain generalization with domain-augmented meta-learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9619–9628, 2021a. URL https://api.semanticscholar.org/CorpusID:233181672.

Yang Shu, Zhangjie Cao, Chenyu Wang, Jianmin Wang, and Mingsheng Long. Open domain generalization with domain-augmented meta-learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9619–9628, 2021b. doi: 10.1109/CVPR46437.2021.00950.

Mainak Singha, Ankit Jha, Shirsha Bose, Ashwin Nair, Moloud Abdar, and Biplab Banerjee. Unknown prompt, the only lacuna: Unveiling clip's potential for open domain generalization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13309–13319, 2024. doi: 10.1109/CVPR52733.2024.01264.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17869–17878, 2025.

Yue Tan, Chen Chen, Weiming Zhuang, Xin Dong, Lingjuan Lyu, and Guodong Long. Is heterogeneity notorious? taming heterogeneity to handle test-time shift in federated learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 27167–27180. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/565f995643da6329cec701f26f8579f5-Paper-Conference.pdf.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5385–5394, 2017. doi: 10.1109/CVPR.2017.572.

Xiran Wang, Jian Zhang, Lei Qi, and Yinghuan Shi. Generalizable Decision Boundaries: Dualistic Meta-Learning for Open Set Domain Generalization . In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11530–11539, Los Alamitos, CA, USA, October 2023. IEEE Computer Society. doi: 10.1109/ICCV51070.2023.01062. URL https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.01062.

Jianhan Wu, Xiaoyang Qu, Zhangcheng Huang, and Jianzong Wang. Federated domain generalization with domain-specific soft prompts generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2366–2375, 2025.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4818–4829, June 2024.

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.

Zheng Xu, Wen Li, Li Niu, and Dong Xu. Exploiting low-rank structure from latent domains for domain generalization. In *European Conference on Computer Vision*, 2014. URL https://api.semanticscholar.org/CorpusID:13777916.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18102–18112, 2022. doi: 10.1109/CVPR52688.2022.01759.

Xunzheng Zhang, Juan Marcelo Parra-Ullauri, Shadi Moazzeni, Xenofon Vasilakos, Reza Nejabati, and Dimitra Simeonidou. Federated analytics with data augmentation in domain generalization

toward future networks. *IEEE Transactions on Machine Learning in Communications and Networking*, 2:560–579, 2024. URL https://api.semanticscholar.org/CorpusID:269392574.

Yubin Zheng, Pak Hei Yeung, Jing Xia, Tianjie Ju, Peng Tang, Weidong Qiu, and Jagath C Rajapakse. Feddeap: Adaptive dual-prompt tuning for multi-domain federated learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5375–5384, 2025.

Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *European conference on computer vision*, pp. 561–578. Springer, 2020.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16795–16804, 2022a. doi: 10.1109/CVPR52688.2022.01631.

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.

A Pseudo-code of FOCUS

In Algorithm 1, we provide a pseudo-code of the full FOCUS. The collaborative training protocol is structurally decoupled into server-side global aggregation and client-side local optimization. Specifically, during the local training phase, each decentralized client dynamically synthesizes pseudo-open outliers and optimizes the text-visual hierarchical prompts via the localized Attribute-Semantic Injection (ASI) module. Crucially, to preserve domain-specific privacy and bypass communication bottlenecks, clients strictly upload only the lightweight text prompt parameters to the server, keeping the heavy visual structures entirely localized.

B Datasets Description

Office-Home Dataset (Venkateswara et al., 2017): Comprising 15,500 images, this dataset is divided into 65 classes across four domains:Art, Clipart, Product, and Real. **PACS** (Li et al., 2017b): The PACS dataset includes 9,991 images, categorized into seven classes and spread over four domains: Art Painting, Cartoon, Photo, and Sketch. **VLCS** (Fang et al., 2013): This dataset amalgamates images from four classification datasets (PASCAL VOC 2007 (Everingham et al., 2015), Caltech101 (Fei-Fei et al., 2004), LabelMe (Russell et al., 2008), and SUN09 (Xiao et al., 2010)) and consists of images across five shared classes: Bird, Car, Chair, Dog, and Person. **Multi-dataset** (Shu et al., 2021b): This dataset combines various public datasets such as Office-31 (Saenko et al., 2010), STL-10 (Coates et al., 2011), and VisDA-2017 (Peng et al., 2017), including four domains from DomainNet. It features 20 open classes not present in the source domains' joint label set. **Mini-DomainNet** (Peng et al., 2019): This dataset features four domains, each comprising images from 125 categories. The specific known and unknown class splits for all five datasets used in our open-set setting are detailed in Table 6, with classes arranged in alphabetical order.

Domain	PACS	VLCS	Office-Home	Multi-dataset	Mini-DomainNet
Source 1	3,0,1	0,1	0-14,21-31	0-30	0-19,40-59
Source 2	4,0,2	1,2	0-8,15-20,32-42	1,31-41	0-9,20-39,80-89
Source 3	5,1,2	2,3	0-2,9-20,43-53	31,33,34,41-47	10-19,40-49,60-79
Target	0-6	0-4	0,3-4,9-10,15-16,21-23, 32-34,43-45,54-64	0,1,5,6,10,11, 14,17,20,26, 31-36,39-43, 45-46,48-67	0-4,8-17,25-34, 43-47,75-79, 83-87,90-125

Table 6: Experimental configuration of cross-domain class distributions. For each evaluation cycle, three domains are designated as localized source clients.

Algorithm 1 Decentralized Training Protocol of FOCUS

Require: Decentralized datasets $\{\mathcal{D}_k\}_{k=1}^{K}$, rounds R, epochs E, weights $\{w_k\}_{k=1}^{K}$, learning rate η, known attributes $\mathcal{A}_{\text{known}}$.
Ensure: Optimized global multimodal prompts θ_p^*.

Initialize global prompts θ_p^0; ▷ Central Server Initialization
Synthesize open-set anchor: $\mathbf{a}_{unk} \leftarrow -\frac{1}{|\mathcal{C}|}\sum_{i=1}^{|\mathcal{C}|}\mathbf{a}_i$; ▷ Synthesize semantic anchor
for each communication round $r = 0, 1, \ldots, R-1$ **do**
  Server distributes θ_p^r to clients;
  for each client $k \in \{1, \ldots, K\}$ **in parallel do**
    Sync local prompts: $\theta_{p,k}^r \leftarrow \theta_p^r$; ▷ Synchronize global model
    Initialize local ASI module $\theta_{asi,k}^r$;
    for local epoch $e = 1, \ldots, E$ **do**
      for each mini-batch $\mathcal{B}_{\text{known}} \sim \mathcal{D}_k$ **do**
        Generate pseudo-open images $\mathcal{B}_{\text{unk}}$ via Stable Diffusion; ▷ Dynamic Outlier Synthesis
        Filter $\mathcal{B}_{\text{unk}}$ by image quality; ▷ Quality Filtering
        Build batch: $\mathcal{B}_{\text{total}} \leftarrow \mathcal{B}_{\text{known}} \cup \text{Filtered}(\mathcal{B}_{\text{unk}})$;
        Augment attributes: $\mathcal{A} \leftarrow \mathcal{A}_{\text{known}} \cup \{\mathbf{a}_{unk}\}$; ▷ Multimodal Injection
        Extract visual features $f(\mathbf{x})$;
        Fuse features via ASI cross-attention (Eq. 6);
        Compute $\mathcal{L}_{ce}$ and $\mathcal{L}_{ds}$ (Eqs. 7, 9); ▷ Local Objective Optimization
        Compute diversity loss $\mathcal{L}_{class}$ (Eq. 10);
        $\mathcal{L}_{total} \leftarrow \mathcal{L}_{ce} + \mathcal{L}_{ds} + \mathcal{L}_{align} + \lambda\mathcal{L}_{class}$;
        Update parameters via AdamW:
        $\{\theta_{p,k}^r, \theta_{asi,k}^r\} \leftarrow \text{AdamW}(\mathcal{L}_{total}, \eta)$;
      end for
    end for
    Keep ASI module $\theta_{asi,k}$ local; ▷ Modality Decoupling & Privacy
    Upload prompts $\theta_{p,k}^{r+1}$ to server;
  end for
  Aggregate global prompts: $\theta_p^{r+1} \leftarrow \sum_{k=1}^{K} w_k \theta_{p,k}^{r+1}$; ▷ Server-Side Federated Aggregation
end for
return θ_p^R;

C Evaluation on Alternative Vision Backbone (ViT-B/16)

To validate its architectural agnosticism, we evaluated FOCUS with a ViT-B/16 backbone in the two-shot setting. FOCUS consistently achieves the highest average H-score (78.81%) across five datasets, successfully avoiding the outlier degradation seen in baselines. This confirms our mechanisms universally balance known-class discrimination and unknown-class rejection.

Methods	PACS		VLCS		Office-Home		Multi-Dataset		Mini-DomainNet		Average	
	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score
PromptFL	96.09	69.99	77.46	48.62	72.56	1.22	83.74	1.45	83.08	0.57	82.04	15.93
FedpromptSRC	96.47	95.94	70.01	68.67	71.69	31.05	82.44	14.56	83.96	35.22	81.47	41.25
FedMaPLe	96.32	94.09	90.23	62.62	75.91	66.02	85.12	30.01	85.29	39.38	**85.73**	49.16
FedMVP	97.32	59.72	85.50	60.88	84.25	27.70	81.17	69.93	82.73	22.66	83.33	44.20
FedODG-CLIP	74.94	58.91	69.57	64.64	38.10	47.72	41.28	50.90	60.08	42.97	64.66	60.34
FedOSLoPROMPT	95.20	95.51	75.04	74.64	59.44	64.17	75.29	78.31	78.17	54.87	76.34	66.72
FOCUS(Ours)	96.19	**97.19**	84.41	**81.30**	73.00	**73.94**	82.68	**80.74**	79.63	**66.02**	79.98	**78.81**

Table 7: Comparative analysis across five datasets in a two-shot setting, reporting both the average closed-set accuracy (Acc) and the harmonic mean (H-score) over all domain combinations following a leave-one-domain-out protocol. (ViT/16)