Controllable Federated Prompt Learning at Test Time

Rui Zhu[1*] Liang Bai[1*] Yanming Guo[1] Yirun Ruan[1] Tianyuan Yu[1†] Zhihe Lu[2†]
[1]National University of Defense Technology [2]Hamad Bin Khalifa University

Abstract

Federated Prompt Learning (FPL) has recently attracted increasing attention for its ability to leverage large-scale vision-language models such as CLIP within federated learning frameworks. While existing studies have advanced FPL through personalization strategies to enhance client-specific performance, personalized models often suffer severe degradation when deployed across unseen domains due to distribution shifts. In this paper, we take the first step toward exploring Test-Time FPL (TTFPL), aiming to bridge the cross-domain performance gap with minimal effort, requiring only unlabeled target-domain data. We propose COTE, a tri-prompt controllable TTFPL framework that dynamically balances three complementary prompts: the global prompt from standard FPL, the local prompt from personalized FPL, and the frozen CLIP prompt. Specifically, we introduce a novel confidence-guided Model-Data Alignment (MoDA) metric in COTE that quantifies alignment at both macro and micro levels, capturing the consistency between model predictions and data distributions. By integrating MoDA with model confidence, COTE adaptively adjusts the contribution of each prompt at test time, enabling robust generalization across heterogeneous clients and unseen domains without requiring labeled data. Extensive experiments on multiple benchmark datasets demonstrate that our method consistently improves target-domain performance, setting a new direction for adaptive FPL.

1. Introduction

Federated Learning (FL) [34] has recently gained prominence as a scalable and privacy-preserving solution for training models across decentralized edge devices. Unlike conventional centralized learning that requires aggregating raw data to a single server, FL enables local devices to collaboratively optimize a global model while keeping data on-device. This paradigm is particularly valuable in IoT-driven scenarios such as smart cameras, autonomous vehicles, and industrial sensing [10, 35, 49], where data is sensitive. By leveraging distributed computation and reducing reliance on data transfer, FL offers a practical path toward deploying intelligent systems at scale.

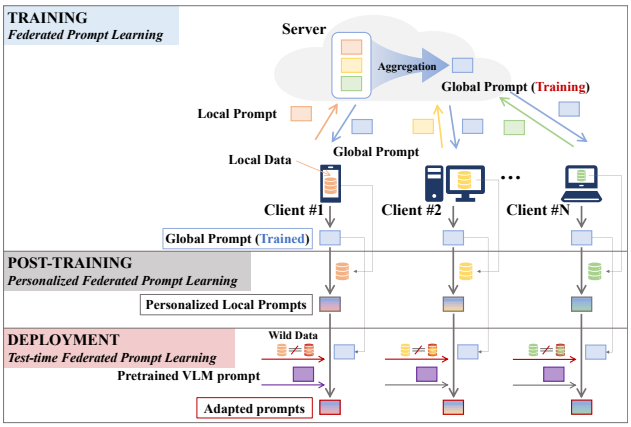


Figure 1. Illustration of three federated learning paradigms — Federated Prompt Learning (FPL), Personalized Federated Prompt Learning (PFPL), and Test-Time Federated Prompt Learning (TTFPL) — all centered on local data adaptation. FPL leverages the strong generalization of Vision-Language Models (VLMs), while PFPL utilizes local training data to fine-tune prompts prior to deployment. TTFPL focuses on enabling models to adapt to each client's test data distribution while preserving cross-client global knowledge.

However, a fundamental limitation remains: once deployed, FL models are typically static and struggle to adapt to distribution shifts [17, 20, 28] in local data, even when the task and label space remain unchanged. Such shifts frequently occur in real-world visual domains [24, 29, 39, 54], where factors like lighting, background, or viewpoint changes alter data characteristics over time. To address FL's rigidity, recent efforts have turned to vision-language models (VLMs) [23, 25, 30, 31, 40, 51, 52], such as CLIP [40], whose strong generalization and modular prompt-based adaptation provide a new foundation under federated settings, named Federated Prompt Learning (FPL). While FPL methods [1, 4, 26] deliver robust cross-client generalization for global model, they still struggle to adapt to client-specific local data distributions. To mitigate this shortcoming, Personalized Federated Prompt Learning (PFPL) [14, 22] has emerged, aiming to tailor prompt-based

*Equal contribution
†Corresponding authors

models to individual client data distributions in the post-training phase before deployment. Despite effectively mitigating inter-client heterogeneity, PFPL relies solely on historical training data, lacking the ability to adapt to emerging, unseen domain variations during inference. Therefore, existing FPL and PFPL approaches [5, 19, 44, 48] remain limited in handling domain shifts at test time, highlighting the need for federated prompt learning frameworks capable of dynamic, post-deployment adaptation.

In this paper, we address the largely unexplored problem of **Test-Time Federated Prompt Learning (TTFPL)**, a novel paradigm for dynamic adaptation of federated prompt learners to distribution shifts using only unlabeled test data, as shown in Figure 1. Unlike PFPL, which tunes prompts with local training data, TTFPL enables edge devices to adapt prompts during inference, which is a key requirement for IoT vision scenarios where post-deployment labeling is infeasible. This overlooked task introduces two non-trivial challenges inherent to the federated VLM deployment: (i) how to adapt locally personalized prompts to shifted visual distributions in unlabeled test data, without accessing historical training samples; (ii) How to preserve the global cross-client knowledge encoded in federated prompts, ensuring adaptation does not degrade generalization across diverse clients. These dual challenges create a key dilemma: how to strategically leverage distinct prompt types during adaptation to balance local distribution alignment and global knowledge retention, an issue unaddressed by existing prompt learning or test-time adaptation (TTA) frameworks.

To tackle these issues, we propose **COtrollable TEst-time federated prompt learning (COTE)**, a framework that leverages a tri-prompt design and a novel Model-Data Alignment (MoDA) metric to enable controllable, data-aware adaptation. The tri-prompt design integrates three complementary prompt types, each addressing a distinct knowledge level: (1) the original prompt (from pre-trained VLMs) for data-agnostic generalization; (2) the global prompt (aggregated across federated clients) for task-specific, cross-domain shared semantics; (3) the local prompt (personalized to each client) for fine-grained sensitivity to local visual patterns. Rather than statically combining prompts, COTE dynamically adjusts their utilization based on real-time data characteristics, embodying the core of controllable adaptation and marking a novel direction for prompt-based test-time adaptation in federated settings.

To effectively govern the tri-prompt usage, we introduce the novel Model-Data Alignment (MoDA) metric, a unified measure that quantifies the consistency between model predictions and the underlying data distribution on each client, laying the foundation for our controllable adaptation. MoDA captures both macro-level and micro-level alignment: the macro-level assesses how a model organizes its predictions across the global semantic space, reflecting overall sparsity and diversity, while the micro-level measures local stability and balance among activated classes. Unlike conventional single-faceted metrics, MoDA provides a stable and sensitive estimation of client distribution heterogeneity, enabling dynamic and data-aware prompt selection during test-time adaptation. Specifically, MoDA is integrated with pseudo-label confidence to enable controllable prompt selection. Confidence scores from the local prompt–based model are first used to separate samples into high-confidence and low-confidence groups. Within each group, MoDA evaluates the alignment between model priors and local data structures to guide decision-making. For highly aligned samples, the global prompt is used in the high-confidence group to leverage shared cross-client knowledge, while the original prompt is adopted in the low-confidence group to enhance generalization under uncertainty. Conversely, for imbalanced or weakly aligned distributions, the local prompt is chosen by default to better capture domain-specific nuances. This data-aware, controllable prompt selection mechanism enables each client to perform robust, self-supervised adaptation without external labels, achieving a controllable balance between local personalization and global generalization.

In summary, our main contributions are as follows:

- We are the first to explore Test-Time Federated Prompt Learning (TTFPL), addressing domain shift and generalization challenges in federated vision–language systems.
- We propose a tri-prompt controllable TTFPL framework (COTE) that integrates original, global, and local prompts based on their complementary strengths, enabling adaptive knowledge utilization across general, shared, and personalized levels.
- We introduce the Model-Data Alignment (MoDA) metric, a unified measure that quantifies the consistency between model predictions and local data distributions on each client, enabling dynamic and data-aware prompt and sample selection during adaptation.
- Extensive experiments on multiple benchmark datasets demonstrate that our method consistently enhances target-domain performance, establishing a new paradigm for adaptive and generalizable federated prompt learning.

2. Related Work

2.1. Federated Prompt Learning

Federated Learning (FL) [34] enables decentralized model training without sharing raw data, offering a privacy-preserving paradigm for collaborative learning. Prompt learning has emerged as an effective way to adapt large Vision-Language Models (VLMs), such as CLIP [40], to downstream tasks. The idea has been extended to decentralized settings, giving rise to federated prompt learn-

ing (FPL), which allows clients to exchange and optimize prompts instead of full model weights. PromptFL [14] first demonstrates this formulation, and FedAPT [22] enhances global consistency via a class-aware prompt generator guided by a global label embedding. Despite leveraging well-trained prompt initializations for strong generalization across diverse clients, existing FPL methods still struggle to adapt to dynamic local data characteristics.

2.2. Personalized Federated Prompt Learning

To enable FL models to prioritize and adapt to local data characteristics, personalized federated learning (PFL) [1] adapts shared global knowledge to individual clients. Prior works address this from different perspectives: parameter decoupling for independent local optimization [1, 4, 26], knowledge distillation for cross-client information transfer [2, 27, 55], and adaptive model interpolation to balance global generalization and local specialization [8, 33]. To further enhance generalization while retaining personalization, recent PFL works have integrated prompt learning, giving rise to personalized federated prompt learning (PFPL) methods. For instance, pFedPrompt [13] learns linguistic consensus while adapting to client-specific visual distributions; pFedPG [50] generates personalized visual prompts on the server; FedOTP [22] jointly learns global and local prompts with optimal-transport alignment; and pFedMoAP [32] refines personalization using attention-based gating over multiple prompt experts. Despite advancing training-time generalization to heterogeneous local data, existing PFPL methods remain limited in deployment scenarios: they fail to adapt to unseen domain shifts when only unlabeled test data is available.

2.3. Test-time Adaptation (TTA)

TTA improves robustness under distribution shifts by adapting models to unlabeled target data during inference. Early approaches such as TENT [46] minimize prediction entropy, while CoTTA [47], MEMO [53], and EATA [37] enhance stability via consistency or selective regularization. More recent methods expand this paradigm through contrastive self-regularization (SAR [38], AdaContrast [6]), prompt-based tuning for VLMs (TPT [42]), and continual or energy-based adaptation (TTAC [45], ECoTTA [43]). While effective in centralized settings, these approaches assume full data access and a single shared model-assumptions incompatible with federated constraints on privacy, communication, and heterogeneity. Extending TTA to FL therefore demands lightweight, modular adaptation strategies, where prompt-based representations offer a natural and privacy-preserving interface. To our knowledge, only FedTHE [18] explores test-time personalization in FL by tuning or ensembling classifier heads using unlabeled client data. However, this approach modifies backbone models, rendering it incompatible with well-optimized FL models during deployment. Besides, the method relies on explicit classification heads and batch-based updates, which are incompatible with VLMs like CLIP that operate via image-text alignment with frozen backbones. These weaknesses highlight the need for head-free, prompt-level adaptation at deployment, motivating our TTFPL framework.

3. Test-time Federated Prompt Learning

Problem setup. We consider a federated prompt learning system built upon a pretrained vision-language model with frozen image and text encoders $f(\cdot)$ and $g(\cdot)$. After the federated training process, each client $i \in \{1, \dots, N\}$ holds three pretrained prompts: a global prompt $\mathbf{P}^g$ aggregated across clients to capture shared semantics, a personalized prompt $\mathbf{P}_i^l$ initialized from $\mathbf{P}^g$ and refined on local data, and a model-agnostic prompt $\mathbf{P}^c$ inherited from the original VLM as a domain-neutral prior. During deployment, client i receives an unlabeled test set $\mathcal{D}_i^{test}$ drawn from an unknown distribution that may differ from its training data $\mathcal{D}_i^{train}$. Note that training datasets, test data labels, and cross-client communication are not available at this stage.

Definition. Given a set of pretrained prompts $\Pi_i = \{\mathbf{P}^g, \mathbf{P}_i^l, \mathbf{P}^c\}$ for each client $i \in \{1, \dots, N\}$ and its unlabeled test data $\mathcal{D}_i^{test}$, *TTFPL* aims to obtain a test-time adapted prompt $\mathbf{P}_i^a$ through a local adaptation operator $\mathcal{A}_i$:

$$\mathbf{P}_i^a = \mathcal{A}_i(\Pi_i, \mathcal{D}_i^{test}). \tag{1}$$

Without test labels, $\mathcal{A}_i$ is optimized using a general *unsupervised test-time objective* $\mathcal{U}$ applied to the prediction distribution:

$$\min_{\mathcal{A}_i} \mathbb{E}_{x \sim \mathcal{D}_i^{test}} \Big[\mathcal{U}\big(p(\hat{y} \mid x; \mathbf{P}_i^a)\big) \Big], \tag{2}$$

where x denotes an input sample and $\hat{y}$ is the model's predicted label. To obtain this distribution, we follow a similarity-based vision-language matching formulation. For each class $k \in \{1, \dots, C\}$, the text input $t_k(\mathbf{P}_i^a)$ is formed by concatenating the adapted prompt $\mathbf{P}_i^a$ with the class name. The logit and prediction probability for class k are then defined as:

$$\text{logit}^{(k)} = \text{sim}(f(x),\, g(t_k(\mathbf{P}_i^a))), \tag{3}$$

$$p(\hat{y} = k \mid x; \mathbf{P}_i^a) = \frac{\exp(\text{logit}^{(k)}/\tau)}{\sum_{j=1}^{C} \exp(\text{logit}^{(j)}/\tau)}, \tag{4}$$

where τ is a temperature parameter.

4. Methodology

As illustrated in Figure 2, we propose **Co**ntrollable Federated Learning at **Te**st Time (**COTE**), a framework designed

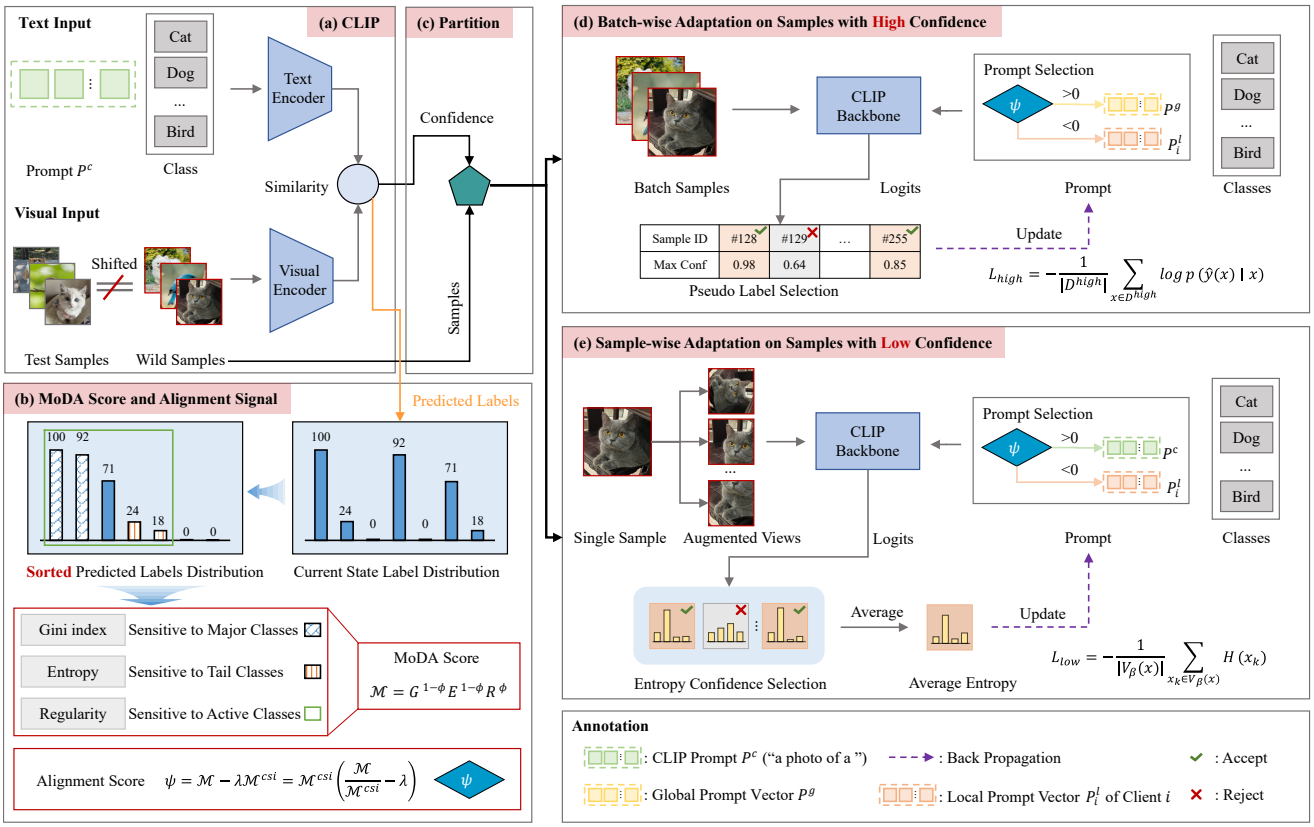


Figure 2. Overview of the **Controllable Federated Learning at Test Time (COTE)** method. The process begins with visual and text inputs encoded by the (a) CLIP model to predict domain-neutral labels, whose distribution is then analyzed using the (b) **MoDA score** and compared with a client-specific ideal reference to compute the alignment score. This guides the **controllable test-time adaptation workflow**, where the model adapts its prompts in two branches with samples (c) partitioned by CLIP confidence: for *high-confidence samples*, (d) batch-wise adaptation is performed, refining prompts with high-confidence pseudo labels; for *low-confidence samples*, (e) sample-wise adaptation is applied by selecting high-entropy augmented views for prompt refinement.

to adapt models to shifted unlabeled data. The approach consists of two main modules: the *Model-Data Alignment* (MoDA) metric and the *controllable test-time adaptation workflow*. MoDA quantifies the consistency between model predictions and the observed data distribution, while the adaptation workflow adjusts model prompts based on alignment score. The two modules work together to dynamically balance global generalization and local specialization, ensuring effective adaptation to real-world data shifts.

4.1. MoDA Metric

4.1.1. Components of MoDA

For each client, we analyze the distribution of pseudo-labels predicted on its unlabeled test set $\mathcal{D}^{\text{test}}$. Let the global label space be S, and let $S_{\text{obs}} \subseteq S$ denote the subset of classes actually predicted by the deployed model on this client. The empirical class-frequency vector is defined as

$$p_k = \frac{n_k}{\sum_{j\in S} n_j}, \quad k \in S, \tag{5}$$

where n_k is the number of samples assigned to class k.

Macro-level statistics. We first extract *macro-level* statistics that describe how the model's predictions are distributed with respect to the entire label space S, including unobserved categories. These statistics measure how globally balanced or sparse the prediction pattern is:

$$G = \frac{1 - \sum_{k\in S} p_k^2}{1 - \frac{1}{|S|}}, \qquad E = -\frac{1}{\log|S|}\sum_{k\in S} p_k \log p_k. \tag{6}$$

The normalized **Gini index** G emphasizes concentration on dominant classes, while the normalized **entropy** E rewards dispersion across tail classes. Normalization by $|S|$ accounts for unobserved categories, allowing G and E to reflect global coverage of the semantic space.

Micro-level statistics. We compute *micro-level* statistics restricted to the actually active subset S_{obs}, which captures

how evenly the model distributes its predictions among the classes that are effectively used on this client. Define the deviation from local uniformity as

$$D = \sum_{k \in S_{\text{obs}}} \left(p_k - \frac{1}{|S_{\text{obs}}|} \right)^2. \tag{7}$$

and the local regularity index as

$$R = 1 - \frac{D}{D_{\max}}, \qquad D_{\max} = \frac{1}{|S_{\text{obs}}|}. \tag{8}$$

A low R indicates concentration on a few active classes, suggesting bias or instability, while a high R reflects smoother adaptation within the local support.

4.1.2. Unified MoDA Formulation

We integrate the macro- and micro-level statistics into a unified *Model-Data Alignment* (MoDA) score that quantifies the coherence between the model's predictions and the client's local data structure.

$$\mathcal{M} = G^{1-\phi} E^{1-\phi} R^{\phi}, \tag{9}$$

where the weighting factor ϕ adaptively controls the contribution of macro- and micro-level evidence:

$$\phi = \frac{|S_{\text{obs}}|}{|S_{\text{obs}}| + |S|}. \tag{10}$$

Intuitively, ϕ reflects the confidence in the observed prediction structure. When few classes are activated ($|S_{\text{obs}}| \ll |S|$), $\mathcal{M}$ relies on global regularities, while as more classes are observed, it emphasizes local regularity R. This allows a smooth transition between global calibration and local specialization. Empirically, higher $\mathcal{M}$ values indicate stable, general-domain predictions, while lower values suggest divergence, signaling a need for stronger personalization.

4.1.3. Client-specific Ideal Reference

To interpret the MoDA score relative to an ideal configuration, we introduce a client-specific reference where the model distributes predictions uniformly over the observed label subset S_{obs}, reflecting maximal alignment with the available data.

$$p_k^{\text{csi}} = \frac{1}{|S_{\text{obs}}|}, \quad k \in S_{\text{obs}}. \tag{11}$$

Based on this definition, the associated macro- and micro-level statistics become

$$\begin{aligned} G^{\text{csi}} &= \frac{1 - \sum_{k \in S} (p_k^{\text{csi}})^2}{1 - \frac{1}{|S|}}, \\ E^{\text{csi}} &= -\frac{1}{\log |S|} \sum_{k \in S} p_k^{\text{csi}} \log p_k^{\text{csi}}, \\ R^{\text{csi}} &= 1 - \frac{\sum_{k \in S_{\text{obs}}} \left(p_k^{\text{csi}} - \frac{1}{|S_{\text{obs}}|} \right)^2}{D_{\max}} = 1. \end{aligned} \tag{12}$$

This reflects a perfectly calibrated model-data relation, with no overconfidence or bias toward any class subset. The corresponding ideal MoDA score is defined as

$$\mathcal{M}^{\text{csi}} = \left(G^{\text{csi}}\right)^{1-\phi} \left(E^{\text{csi}}\right)^{1-\phi} \left(R^{\text{csi}}\right)^{\phi}. \tag{13}$$

4.1.4. Alignment Signal

To measure how far a client's predictions deviate from the ideal state, we compute an *alignment signal*:

$$\psi = \mathcal{M} - \lambda \mathcal{M}^{\text{csi}} = \mathcal{M}^{\text{csi}} \left(\frac{\mathcal{M}}{\mathcal{M}^{\text{csi}}} - \lambda \right), \tag{14}$$

where $\lambda \in (0, 1)$ is a stability coefficient. A higher ψ indicates that the model's predictions align more closely with a model trained on *a diverse range of data*, suggesting that the data distribution is more compatible with such a model's characteristics. Conversely, a lower ψ implies that the data distribution is more aligned with a model that may exhibit a *stronger bias* towards certain data characteristics, signaling the need for further adaptation. Therefore, ψ thus serves as a soft guidance term for the controllable adaptation workflow (Section 4.2), enabling the model to adjust its strategy based on the alignment with the underlying data structure.

4.2. Controllable Test-time Adaptation Workflow

To operationalize the alignment signal ψ, we design a controllable test-time adaptation framework that adjusts model behavior based on client-specific alignment. Guided by ψ, the framework adapts prompts through three steps: confidence-based sample partition, high-confidence adaptation, and low-confidence prompt tuning.

4.2.1. Confidence-based Sample Partition.

For each client, we obtain pseudo-labels and confidence scores from the CLIP model, which serves as a frozen domain-neutral reference for assessing prediction reliability. Following the definitions in Equation 3 and 4, the prediction confidence for a sample x is computed as

$$c(x) = \max_{y \in S} p(y \mid x), \tag{15}$$

where $p(y \mid x)$ denotes the normalized probability predicted by the model.

Samples are partitioned into two subsets according to a confidence threshold τ_c:

$$\begin{aligned} \mathcal{D}^{\text{high}} &= \{x \mid c(x) > \tau_c\}, \\ \mathcal{D}^{\text{low}} &= \{x \mid c(x) \leq \tau_c\}. \end{aligned} \tag{16}$$

High-confidence samples provide reliable pseudo supervision for prompt refinement, while low-confidence ones are reserved for uncertainty-aware adaptation in the subsequent process.

4.2.2. Adaptation of High-confidence Samples

For high-confidence samples, we refine the most reliable prompt using pseudo-label supervision guided by the alignment score ψ. Let the selected prompt be

$$\mathbf{P}^\star = \begin{cases} \mathbf{P}^g, & \psi > 0, \\ \mathbf{P}_i^l, & \psi < 0, \end{cases} \tag{17}$$

where a positive ψ indicates global-aligned behavior and favors the global prompt, while a negative ψ corresponds to locally skewed behavior and thus relies on the local prompt. The prediction probability for class k under the selected prompt follows Equation 4, and the pseudo-label is defined as

$$\hat{y}(x) = \arg\max_{k \in S} p(k \mid x). \tag{18}$$

The prompt is then refined using high-confidence samples $\mathcal{D}^{\text{high}}$ with the loss

$$\mathcal{L}_{\text{high}} = -\frac{1}{|\mathcal{D}^{\text{high}}|} \sum_{x \in \mathcal{D}^{\text{high}}} \log p(\hat{y}(x) \mid x). \tag{19}$$

This update reinforces the prompt consistent with the client's alignment condition, ensuring stable adaptation with reliable pseudo supervision.

4.2.3. Adaptation of Low-confidence Samples

For low-confidence samples, direct pseudo-label refinement is unreliable. We therefore employ a sample-wise *test-time prompt tuning* (TPT) [42] strategy guided by ψ. The initialization of the adaptive prompt is selected as

$$\mathbf{P}_{\text{init}} = \begin{cases} \mathbf{P}^c, & \psi > 0, \\ \mathbf{P}_i^l, & \psi < 0, \end{cases} \tag{20}$$

where $\mathbf{P}^c$ denotes the CLIP-derived prompt providing a domain-neutral initialization, and $\mathbf{P}_i^l$ captures client-specific priors. For each uncertain sample x, we generate K_a augmented views $\{x_k\}_{k=1}^{K_a}$ and compute their prediction entropies:

$$\mathcal{H}(x_k) = -\sum_{k \in S} p(k \mid x_k) \log p(k \mid x_k). \tag{21}$$

The top-β fraction of high-entropy views is denoted as $\mathcal{V}_\beta(x)$, and the prompt is optimized by maximizing their average entropy:

$$\mathcal{L}_{\text{low}} = -\frac{1}{|\mathcal{V}_\beta(x)|} \sum_{x_k \in \mathcal{V}_\beta(x)} \mathcal{H}(x_k). \tag{22}$$

This procedure enables each uncertain sample to adjust its representation toward the most suitable initialization (CLIP or local), enhancing robustness under distribution shifts during inference.

5. Experiments

5.1. Setup

Datasets. We evaluate our method on five popular benchmarks: CIFAR100 [12], ImageNet [7] and its variants (ImageNet-A [16] ImageNet-V2 [41], ImageNet-R [15]), Flowers-102 [36], Caltech-101 [11], and Food-101 [3]. These datasets span diverse domains and granularities, covering general object and fine-grained recognition tasks.

Following the benchmark protocol of BRFL [18], we simulate various test-time distribution shifts to assess model robustness. For CIFAR100, Flowers-102, Caltech-101, and Food-101, we construct four test settings: (1) **Ori** (original local test set), (2) **Corr** (corrupted data with common perturbations), (3) **OoC** (out-of-client data from other users), and (4) **Mix** (a combination of all the above to simulate realistic deployment). For **ImageNet**, we use ImageNet-A, ImageNet-V2, and ImageNet-R to represent real-world domain shift. Further details of the dataset construction and shift simulation are provided in Appendix Section A.1.

Setting. To simulate practical scenarios driven by the Internet of Things, we construct a strongly non-IID federated environment using a Dirichlet partition with concentration parameter $\alpha = 0.01$, which induces extreme label heterogeneity across clients. All other details of the federated setup, training process, and evaluation protocol are included in the Appendix Section A.2.

Baselines. We compare our method with two groups of representative approaches, covering both federated prompt learning (FPL) and test-time adaptation (TTA) paradigms. (1) FPL methods include PromptFL [14], PromptFL [14] + FT, pFedMoAP [32], and FedOTP [22]. PromptFL learns a global prompt collaboratively across clients, while PromptFL + FT fine-tunes local data for personalization. pFedMoAP and FedOTP further improve personalization and robustness within federated systems. (2) TTA methods include PL [21], TENT [46], and TPT [42]. These are adapted to the federated setting by applying test-time optimization on the PromptFL + FT personalized model to assess adaptability under distribution shifts.

Implementation Details. All methods are implemented in PyTorch with a frozen CLIP ViT-B/16 backbone [9], optimizing only the prompt parameters, where each prompt consists of 16 learnable context tokens. Local prompt training uses SGD for one local epoch per round with FedAvg aggregation, after which we retain the global prompt $\mathbf{P}^g$, local prompts $\mathbf{P}_i^l$, and the CLIP prompt $\mathbf{P}^c$ for test-time adaptation. During inference, prompts remain the only learnable components: samples are partitioned using a confidence threshold $\tau_c = 0.7$, low-confidence adaptation uses $K_a = 64$ augmented views with the top-$\beta = 0.1$ entropy samples, optimized with one SGD step, and the stability coefficient in Equation 14 is set to $\lambda = 0.9$. Additional imple-

Table 1. Test accuracy on **CIFAR100** and **ImageNet** datasets under heterogeneous client data partitioning across different test distributions. The **best** and second-best results for each column are highlighted in bold and underlined, respectively.

Methods	CIFAR100					ImageNet						
	Ori	Corr	OoC	Mix	Avg	Ori	A	V2	R	OoC	Mix	Avg
Federated Prompt Learning												
PromptFL [14]	62.49	31.58	**62.30**	52.12	52.12	37.08	30.71	71.83	57.85	**37.13**	49.05	47.28
Personalized Federated Prompt Learning												
PromptFL [14] + FT	95.68	77.36	22.07	65.04	65.04	89.42	77.48	97.02	88.62	5.60	87.23	74.23
FedOTP [22]	94.90	76.62	9.89	60.21	60.41	87.44	66.75	95.30	84.43	2.78	82.72	69.90
pFedMoAP [32]	94.22	72.74	11.07	59.24	59.32	87.41	68.43	93.52	82.14	3.96	81.44	69.48
Test-time Adaptation												
PL [21]	95.68	77.08	21.59	64.83	64.80	89.35	79.48	97.68	90.39	4.77	88.20	74.98
TENT [46]	94.88	50.47	9.39	53.23	51.99	87.91	81.59	**97.73**	**92.99**	2.48	87.32	75.00
TPT [42]	96.15	77.25	19.24	64.05	64.17	90.06	79.42	97.24	89.37	6.87	88.32	75.21
COTE (Ours)	**96.16**	**79.13**	42.34	**66.67**	**71.08**	**93.10**	**81.92**	97.69	89.27	36.60	**90.05**	**81.44**

Table 2. Test accuracy on **Caltech101**, **Flowers102**, and **Food101** datasets under heterogeneous client data partitioning across different test distributions.

Methods	Caltech101					Flowers102					Food101				
	Ori	Corr	OoC	Mix	Avg	Ori	Corr	OoC	Mix	Avg	Ori	Corr	OoC	Mix	Avg
Federated Prompt Learning															
PromptFL [14]	90.76	88.68	**92.76**	90.71	90.73	88.59	76.27	**88.23**	84.37	84.36	74.47	61.42	**74.80**	70.23	70.23
Personalized Federated Prompt Learning															
PromptFL [14] + FT	98.98	98.28	83.85	93.70	93.70	99.68	97.93	54.87	84.19	84.17	97.88	94.65	34.60	75.71	75.71
FedOTP [22]	98.97	98.38	28.89	75.85	75.52	99.24	97.49	2.98	65.98	66.42	96.00	92.17	7.78	65.17	65.28
pFedMoAP [32]	**99.30**	**98.86**	76.02	91.66	91.46	99.56	**98.69**	13.40	70.23	70.47	**99.61**	**98.46**	10.92	68.97	69.49
Test-time Adaptation															
PL [21]	98.95	98.56	86.26	**94.19**	94.49	99.65	97.96	56.62	84.25	84.62	98.00	95.25	35.15	**75.95**	76.09
TENT [46]	99.05	98.37	88.67	94.10	95.05	99.59	98.09	51.30	81.23	82.55	98.25	94.62	29.40	72.33	73.65
TPT [42]	99.05	98.50	81.23	92.98	92.94	**99.85**	98.12	46.55	81.74	81.56	97.88	95.10	28.70	73.67	73.84
COTE (Ours)	99.05	98.38	88.99	94.14	**95.14**	99.77	98.33	62.27	**85.60**	**86.49**	97.97	94.62	58.50	75.86	**81.74**

mentation details and hyperparameters are provided in the Appendix Section A.3.

5.2. Main Results

Results on CIFAR100 and ImageNet. Table 1 presents results on CIFAR100 and ImageNet under heterogeneous client partitions. Our method achieves the highest average accuracy on both benchmarks, obtaining **71.08%** on CIFAR100 and **81.44%** on ImageNet, outperforming the strongest baseline (TPT) by over **6%** on average. The improvement is particularly pronounced under the most challenging OoC and Mix settings, where existing fine-tuning or pseudo-labeling strategies fail to maintain robustness. Specifically, our method raises OoC accuracy from 19.2% to 42.3% on CIFAR100 and from 6.9% to 36.6% on ImageNet, demonstrating a substantial enhancement in generalization across unseen client distributions. Meanwhile, the performance on in-distribution and natural-shift variants (ImageNet-A, V2, and R) remains comparable or superior to all baselines, confirming that the proposed alignment-guided mechanism achieves both stability and adaptability under diverse distributional shifts. Notably, representative PFPL methods achieve suboptimal performance, as their personalization mechanisms heavily align prompts with the biased local training distribution that poorly matches the test data distribution in our setting.

Results on Caltech101, Flowers102, and Food101. Table 2 reports results on fine-grained recognition datasets. Across all three benchmarks, our approach consistently attains the best overall performance. The advantages are most evident under the OoC configuration, where data distributions deviate significantly from those seen during training. Compared with the strongest baseline, our method achieves absolute accuracy gains of 2.7%, 5.7%, and 23.4% on the three datasets, respectively. These results highlight the efficacy of our COTE in facilitating stable prompt refinement

Table 3. Performance comparison of different alignment metrics on CIFAR100 dataset.

Metrics	Ori	Corr	OoC	Mix	Avg
Gini	92.89	66.19	**42.34**	**67.14**	67.14
Entropy	**96.16**	79.00	**42.34**	66.04	70.89
Regularity	93.00	66.15	**42.34**	67.12	67.15
MoDA (Ours)	**96.16**	**79.13**	**42.34**	66.67	**71.08**

Table 4. Effect of sample partition strategy on CIFAR100 and ImageNet datasets

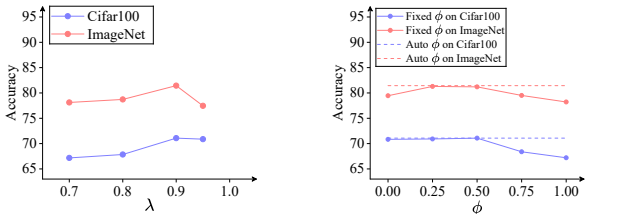
Datasets	Variations	Methods		
		High-conf	Low-conf	COTE (Ours)
CIFAR100	Ori	95.68	96.15	**96.16**
	Corr	77.36	77.25	**79.13**
	OoC	22.07	19.24	**42.34**
	Mix	65.04	64.05	**66.67**
	Avg	65.04	64.17	**71.08**
ImageNet	Ori	89.42	90.06	**93.10**
	A	77.48	79.42	**81.92**
	V2	97.02	97.24	**97.69**
	R	88.62	**89.37**	89.27
	OoC	5.60	6.87	**36.60**
	Mix	87.23	88.32	**90.05**
	Avg	74.23	75.21	**81.44**

for clients with distinct or long-tailed data distributions, while preserving strong performance on in-distribution and corrupted settings.

5.3. Ablation Results

Impact of alignment metric. Table 3 compares the proposed MoDA score with single-factor metrics (*Gini*, *entropy*, and the *local regularity* index) within the same adaptation pipeline on CIFAR100. Among the alignment choices, MoDA attains the best overall average (**71.08%** on CIFAR100) while entropy is the second best. This suggests that pure uncertainty cues (entropy) are strong on large-scale, high-cardinality label spaces, whereas MoDA's integration of macro (G, E) and micro (R) statistics yields more reliable behavior across heterogeneous settings, including smaller or more imbalanced client distributions.

Effect of sample partition strategy. Table 4 presents the contribution of confidence-based sample partitioning within the adaptation procedure. When all samples are processed exclusively by high- or low-confidence method, the overall performance decreases markedly, particularly under the OoC and Mix distribution shifts. On CIFAR100, omitting the partitioning mechanism lowers accuracy from 71.08% to 65.04%, and on ImageNet from 81.44% to 75.21%. These observations indicate that distinguishing between high- and low-confidence samples is essential. The partitioning facilitates accurate pseudo-label adaptation for confident predictions, while enabling more conservative entropy-based updates for uncertain cases, thereby improving both stability and generalization.



(a) Average accuracy of different λ (b) Average accuracy of different ϕ

Figure 3. Further analysis of coefficient λ and ϕ

5.4. Further Analysis

Effect of the stability coefficient λ. Figure 3a evaluates the effect of the stability coefficient λ, which controls the influence of the ideal alignment reference in computing the alignment signal ψ in Equation 14. Increasing λ from 0.7 to 0.9 yields steady gains on both CIFAR100 and ImageNet, with the best average accuracy at $\lambda = 0.9$ (**71.08%** and **81.44%**). A moderate λ achieves a good balance between sensitivity to client-specific shifts and stability against noise.

Effect of the weighting factor ϕ. Figure 3b reports the effect of the weighting factor ϕ, which balances global and local statistics in the MoDA metric in Equation 10. Extreme settings (ϕ =0 or 1) lead to clear performance drops, indicating that relying solely on global or local evidence causes over-generalization or overfitting. Mid-range values ($\phi \in [0.25, 0.50]$) deliver the strongest results, highlighting the importance of jointly leveraging global calibration and local consistency. The adaptive strategy (*auto*), which adjusts (ϕ) based on the number of observed classes per client, performs on par with or slightly better than the best fixed value. This demonstrates that the adaptive mechanism effectively balances generality and specificity, ensuring stable performance under diverse client distributions.

6. Conclusion

We introduced *Test-Time Federated Prompt Learning (TTFPL)* setting, a new problem setting that targets post-deployment domain shifts in federated vision-language systems using only unlabeled target data. To address this setting, we proposed *Controllable Federated Learning at Test Time (COTE)*, a tri-prompt controllable framework that adaptively balances the global prompt, local prompt, and frozen CLIP prompt during deployment. Extensive experiments on five benchmarks demonstrate that COTE consistently improves performance across diverse and challenging federated scenarios, establishing a new direction for adaptive and generalizable FPL.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (General Program, Grant No. 72571281), the Excellent Young Scientist Fund of Hunan Province (Grant No. 2025JJ40066).

References

[1] Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019. 1, 3

[2] Ilai Bistritz, Ariana Mann, and Nicholas Bambos. Distributed distillation for on-device learning. *Advances in Neural Information Processing Systems*, 33:22593–22604, 2020. 3

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision*, pages 446–461. Springer, 2014. 6

[4] Duc Bui, Kshitiz Malik, Jack Goetz, Honglei Liu, Seungwhan Moon, Anuj Kumar, and Kang G Shin. Federated user representation learning. *arXiv preprint arXiv:1909.12535*, 2019. 1, 3

[5] Ying Chang, Xiaohu Shi, Xiaohui Zhao, Zhaohuang Chen, and Deyin Ma. Dual prompt personalized federated learning in foundation models. *Scientific Reports*, 15(1):28026, 2025. 2

[6] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 295–305, 2022. 3

[7] Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017. 6

[8] Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020. 3

[9] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[10] Chen Fang, Yuanbo Guo, Yongjin Hu, Bowen Ma, Li Feng, and Anqi Yin. Privacy-preserving and communication-efficient federated learning in internet of things. *Computers & Security*, 103:102199, 2021. 1

[11] 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. *Computer Vision Image Understanding*, 106(1):59–70, 2007. 6

[12] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2066–2073. IEEE, 2012. 6

[13] Tao Guo, Song Guo, and Junxiao Wang. Pfedprompt: Learning personalized prompt for vision-language models in federated learning. In *Proceedings of the ACM Web Conference 2023*, pages 1364–1374, 2023. 3

[14] 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. 1, 3, 6, 7

[15] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 6

[16] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 6

[17] Wenke Huang, Mang Ye, Zekun Shi, He Li, and Bo Du. Rethinking federated learning with domain shift: A prototype view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16312–16322. IEEE, 2023. 1

[18] Liangze Jiang and Tao Lin. Test-time robust personalization for federated learning. *arXiv preprint arXiv:2205.10920*, 2022. 3, 6

[19] Kun Jin, Tongxin Yin, Zhongzhu Chen, Zeyu Sun, Xueru Zhang, Yang Liu, and Mingyan Liu. Performative federated learning: A solution to model-dependent and heterogeneous distribution shifts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12938–12946, 2024. 2

[20] Huy Q Le, Ye Lin Tun, Yu Qiao, Minh NH Nguyen, Keon Oh Kim, Eui-Nam Huh, and Choong Seon Hong. Mitigating domain shift in federated learning via intra-and inter-domain prototypes. *arXiv preprint arXiv:2501.08521*, 2025. 1

[21] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, International Conference on Machine Learning*, page 896. Atlanta, 2013. 6, 7

[22] Hongxia Li, Wei Huang, Jingya Wang, and Ye Shi. Global and local prompts cooperation via optimal transport for federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12151–12161, 2024. 1, 3, 6, 7

[23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023. 1

[24] 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*, 2021. 1

[25] Xin Li, Dongze Lian, Zhihe Lu, Jiawang Bai, Zhibo Chen, and Xinchao Wang. Graphadapter: Tuning vision-language models with dual knowledge graph. *Advances in Neural Information Processing Systems*, 36:13448–13466, 2023. 1

[26] Paul Pu Liang, Terrance Liu, Liu Ziyin, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 2020. 1, 3

[27] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. *Advances in Neural Information Processing Systems*, 33:2351–2363, 2020. 3

[28] Zhihe Lu, Yongxin Yang, Xiatian Zhu, Cong Liu, Yi-Zhe Song, and Tao Xiang. Stochastic classifiers for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9111–9120, 2020. 1

[29] Zhihe Lu, Da Li, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Uncertainty-aware source-free domain adaptive semantic segmentation. *IEEE Transactions on Image Processing*, 32:4664–4676, 2023. 1

[30] Zhihe Lu, Jiawang Bai, Xin Li, Zeyu Xiao, and Xinchao Wang. Beyond sole strength: Customized ensembles for generalized vision-language models. In *ICML*, 2024. 1

[31] Zhihe Lu, Jiawang Bai, Xin Li, Zeyu Xiao, and Xinchao Wang. Task-to-instance prompt learning for vision-language models at test time. *IEEE Transactions on Image Processing*, 2025. 1

[32] Jun Luo, Chen Chen, and Shandong Wu. Mixture of experts made personalized: Federated prompt learning for vision-language models. *arXiv preprint arXiv:2410.10114*, 2024. 3, 6, 7

[33] Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*, 2020. 3

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

[35] Dinh C Nguyen, Ming Ding, Pubudu N Pathirana, Aruna Seneviratne, Jun Li, and H Vincent Poor. Federated learning for internet of things: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 23(3):1622–1658, 2021. 1

[36] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 6

[37] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International Conference on Machine Learning*, pages 16888–16905. PMLR, 2022. 3

[38] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. *arXiv preprint arXiv:2302.12400*, 2023. 3

[39] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, 32(3):53–69, 2015. 1

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

[41] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 6

[42] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022. 3, 6, 7

[43] Junha Song, Jungsoo Lee, In So Kweon, and Sungha Choi. Ecotta: Memory-efficient continual test-time adaptation via self-distilled regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11920–11929, 2023. 3

[44] Shangchao Su, Mingzhao Yang, Bin Li, and Xiangyang Xue. Federated adaptive prompt tuning for multi-domain collaborative learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 15117–15125, 2024. 2

[45] Yongyi Su, Xun Xu, and Kui Jia. Revisiting realistic test-time training: Sequential inference and adaptation by anchored clustering. *Advances in Neural Information Processing Systems*, 35:17543–17555, 2022. 3

[46] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. 3, 6, 7

[47] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022. 3

[48] Zhihao Wang, Wenke Huang, Tian Chen, Zekun Shi, Guancheng Wan, Yu Qiao, Bin Yang, Jian Wang, Bing Li, and Mang Ye. An empirical study of federated prompt learning for vision language model. *arXiv preprint arXiv:2505.23024*, 2025. 2

[49] Jie Wen, Zhixia Zhang, Yang Lan, Zhihua Cui, Jianghui Cai, and Wensheng Zhang. A survey on federated learning: challenges and applications. *International Journal of Machine Learning and Cybernetics*, 14(2):513–535, 2023. 1

[50] Fu-En Yang, Chien-Yi Wang, and Yu-Chiang Frank Wang. Efficient model personalization in federated learning via client-specific prompt generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19159–19168, 2023. 3

[51] Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10899–10909, 2023. 1

[52] Aiming Zhang, Tianyuan Yu, Liang Bai, Jun Tang, Yanming Guo, Yirun Ruan, Yun Zhou, and Zhihe Lu. Cola: Context-aware language-driven test-time adaptation. *IEEE Transactions on Image Processing*, 2025. 1

[53] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Ad-*

vances in Neural Information Processing Systems, 35: 38629–38642, 2022. 3

[54] Youshan Zhang. A survey of unsupervised domain adaptation for visual recognition. *arXiv preprint arXiv:2112.06745*, 2021. 1

[55] Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In *International Conference on Machine Learning*, pages 12878–12889. PMLR, 2021. 3