WHEN PROMPT EXPERTS DISAGREE: DIAGNOSING OPEN-WORLD DECISIONS IN FEDERATED VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalized federated prompt learning adapts frozen vision–language models (VLMs) by exchanging lightweight prompts across clients, yet prediction accuracy alone reveals little about the internal behavior of the resulting prompt experts, such as whether they agree with one another, whether each is individually confident, and whether a given expert mixture is safe. To make these questions answerable explicitly, we introduce an open-world diagnostic framework (pFEDOPEN). First, pFEDOPEN (i) retains a frozen zero-shot VLM as a semantic anchor along with a geometry regularizer to avoid excessive representation drift, (ii) routes experts via class-aware evidence to prevent unbounded expert misadaptation, and (iii) mixes temperature-controlled expert *probabilities* instead of logits to make each expert's contribution to the final prediction explicit. On top of this design, pFEDOPEN records within-expert uncertainty, inter-expert disagreement, and routing ambiguity as three separate and normalized scores. These diagnostic scores are subsequently used by a client-wise held-out calibration to determine whether to accept the mixture, fall back to the zero-shot expert, or abstain, thereby enhancing transparency of expert assembly. We additionally offer theoretical guarantees. Extensive experiments demonstrate pFEDOPEN's powerful diagnostic capabilities with strong far-OOD discrimination in open-world settings. The code will be released.

1 INTRODUCTION

Prompt tuning offers a communication-efficient path to federated vision–language models (VLMs). In general, clients within this paradigm optimize and exchange prompts while the pretrained VLM backbone remains frozen (Jia et al., 2021; Radford et al., 2021; Wang et al., 2025). Personalized variants further allow each client to exploit learned prompts of beneficial clients for local adaptation. However, beyond deployment accuracy, the system must also express whether a prediction should be trusted, and existing federated prompt mixtures provide no such decision evidence. For instance, pFEDMOAP (Luo et al., 2025) combines a local prompt with identified useful non-local prompts to achieve strong local accuracy, yet its router does not indicate whether those prompts agree, whether their class probabilities are calibrated, etc. Subsequently, pFedMMA (Ghiasvand et al., 2026) reports the gap between local and base-to-novel accuracy, yet without any further decisive signal.

However, making a prompt mixture auditable and explicit requires to solve the following deficiencies. First, once prompts are adapted, originally correlated representations become undetectable in principle, because the pool no longer contains an unchanged reference for its detection. Second, final prediction conflates disagreement among experts with ambiguity within each expert. They actually call for different responses, yet a single scalar can not separate them. Third, routing entropy can not reveal the real decision evidence. Concretely, it could be high when all the candidate experts predict the same class and low when a collapsed router selects unexpected experts.

To probe a prompt mixture, we introduce an open-world diagnostic framework (pFEDOPEN). Specifically, against the missing reference, pFEDOPEN retains a frozen zero-shot CLIP as a semantic anchor, which guarantees a minimum share of every prompt mixture. A geometry regularizer further constrains how far the tuned prediction may drift from its zero-shot counterpart during local adaptation. Against conflated uncertainty, pFEDOPEN routes experts via class-aware evidence and

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.

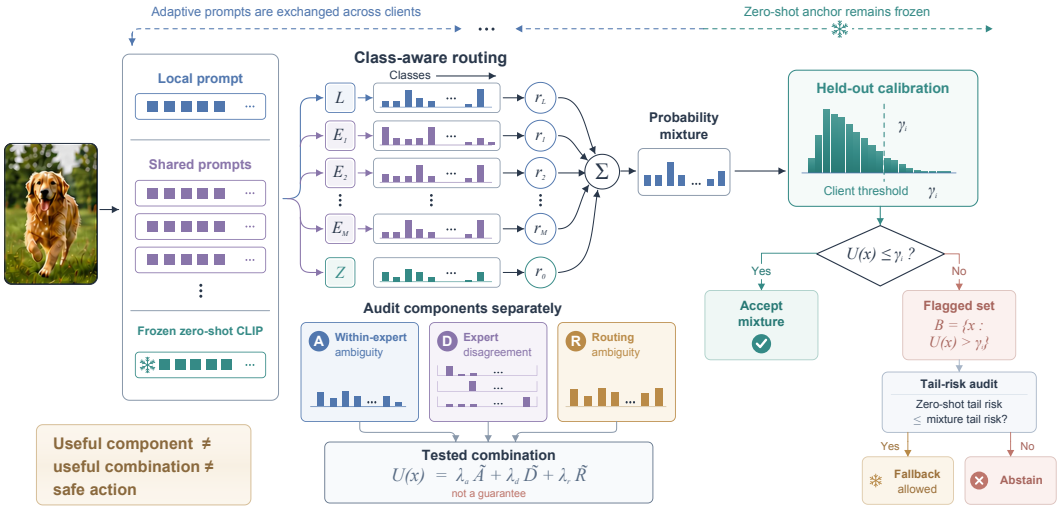


Figure 1: Overview of pFEDOPEN. A client evaluates an input with an adaptive prompt mixture. Class-aware routing assigns weights to the local and non-local experts. It also retains separate scores for within-expert ambiguity A, expert disagreement D, and routing ambiguity R. A held-out calibration determines whether to accept the mixture, fall back to the zero-shot prediction, or abstain.

mixes temperature-scaled expert *probabilities* instead of logits, so that each expert's contribution to the final prediction remains explicit. Meanwhile, it records three separate scores for distinct diagnostic purposes: within-expert ambiguity, indicating whether any single expert classifies the input sharply; inter-expert disagreement, indicating whether locally adapted experts conflict with one another; and routing ambiguity, indicating whether the evidence identifies a competent expert at all. Against the insufficiency of predictive entropy, a held-out calibration adjusts the mixture probabilities and selects a threshold that governs whether the model accepts the mixture prediction, falls back to the zero-shot expert, or abstains. In summary, per-expert predictions, routing weights, uncertainty components, and the final action can each be inspected. Figure 1 summarizes the auditable pipeline.

Theoretically, we guarantee: (i) the predictive entropy of the mixture decomposes exactly into within-expert ambiguity plus inter-expert disagreement. (ii) the zero-shot anchor bounds how far the adapted experts can move a prediction away from the zero-shot anchor. (iii) the held-out calibration guarantees finite-sample marginal coverage at any target level with limited local in-distribution (ID) samples. Empirically, the calibrated mixture (i) raises the harmonic mean of best baseline from 48.81 to 73.53 on the base-to-novel test, showing pFEDOPEN's open-world generalization, and (ii) reveals that useful components in isolation do not guarantee a useful combination.

Our main contributions are highlighted as follows:

- We introduce pFEDOPEN, an auditable federated prompt-expert framework combining zero-shot anchoring, class-aware probability routing, decomposed uncertainty, and held-out calibration.
- We establish an exact uncertainty decomposition, influence bounds, marginal coverage under exchangeability, and a fallback-risk criterion, making the framework's guarantees explicit.
- Through closed-set, base-to-novel, and far-OOD evaluations, we identify pFEDOPEN's powerful auditing and diagnostic capabilities with strong far-OOD discrimination in open-world applications.

2 Problem Setup

Federated prompt pools. Consider M clients, where client i holds private prompt-training, calibration, and test datasets $\mathcal{D}_i^{\text{tr}}$, $\mathcal{D}_i^{\text{cal}}$, and $\mathcal{D}_i^{\text{te}}$. Notably, $\mathcal{D}_i^{\text{cal}}$ is disjoint from $\mathcal{D}_i^{\text{tr}}$. All clients share a frozen CLIP image encoder $f(\cdot)$ and text encoder $g(\cdot)$, which map images and tokenized text into a common d-dimensional embedding space. Neither encoder is updated at any point. Client i learns one local prompt expert on $\mathcal{D}_i^{\text{tr}}$ and may download experts trained by other clients, yielding K_i *adaptive* experts indexed by $k \in \{1, \ldots, K_i\}$. In addition, each client retains the *zero-shot expert* indexed by $k = 0$, whose class prompts use a fixed hand-crafted template (e.g., "a photo of a {class}") in place

of learned context vectors. The zero-shot expert's class embeddings are therefore identical across clients and never change, which is why it is qualified as the semantic anchor.

Class prompts and prompt experts. A *class prompt* of adaptive expert k on client i for class c is a token sequence $\boldsymbol{t}_{ik,c} = [\boldsymbol{v}_{ik,1}, \ldots, \boldsymbol{v}_{ik,L}, \boldsymbol{w}_c]$, consisting of L trainable context vectors $\boldsymbol{v}_{ik,1}, \ldots, \boldsymbol{v}_{ik,L}$ and one class name embedding $\boldsymbol{w}_c$ (Zhou et al., 2022b). Processing $\boldsymbol{t}_{ik,c}$ through the frozen text encoder $g(\cdot)$ and ℓ_2-normalizing yields a class embedding

$$T_{ik,c} = g(\boldsymbol{t}_{ik,c})/\|g(\boldsymbol{t}_{ik,c})\|_2 \in \mathbb{R}^d, \qquad c = 1, \ldots, C. \tag{1}$$

Stacking these class embeddings over the C classes gives the class-embedding matrix $T_{ik} \in \mathbb{R}^{C\times d}$. A prompt is thus fully specified by its context vectors, where only $\{\boldsymbol{v}_{ik,\ell}\}_{\ell=1}^{L}$ are trainable, while $\boldsymbol{w}_c$, $g(\cdot)$, and $f(\cdot)$ remain frozen. Therefore, exchanging an expert between clients amounts to L vectors.

Expert predictions. Given the class embeddings of Eq. 1, adaptive expert k on client i predicts

$$p_{ik}(y = c \mid x) = \mathrm{softmax}_c\big(s\, f(x)^\top T_{ik,c}/\tau_a\big)\,, \tag{2}$$

where $f(x)$ is the ℓ_2-normalized image embedding, s is CLIP's pretrained logit scale, and $\tau_a > 0$ is a temperature shared by all adaptive experts. The zero-shot expert applies the same rule to its frozen embeddings T_{i0} with its own temperature $\tau_0 > 0$. Separate τ_a and τ_0 are required because adapted and zero-shot prompts produce logits on different scales. Mixing the temperature-scaled *probabilities* instead of raw logits keeps each expert's contribution to the final prediction explicit.

Deployment regimes and required outputs. We distinguish three open-world regimes. *Local ID* inputs follow the client distribution and label set used for adaptation. *Novel* inputs belong to held-out classes whose names are supplied only at inference time, so that they test whether adaptation has damaged the pretrained parameters. *OOD* inputs are semantically disjoint from the closed evaluation label set, so no label in the output space is correct and any confident prediction is an error. For each type of input, the system must return a class distribution, an uncertainty score, and exactly one declared action: accept the mixture prediction, fall back to the zero-shot expert, or abstain.

3 METHOD

Overview. Given an image and the client-specific prompt pool, pFEDOPEN first computes the class distributions of all K_i+1 experts together with a class-aware relevance score for each (Section 3.1). It then guarantees the semantic anchor a minimum routing weight, mixes the expert probabilities into the final class distribution, and decomposes the resulting uncertainty into within-expert ambiguity, expert disagreement, and routing ambiguity (Section 3.2). During local training, a geometry regularizer keeps the tuned class distribution close to that of the frozen zero-shot CLIP (Section 3.3). After training, the client-wise held-out calibration based on $\mathcal{D}_i^{\mathrm{cal}}$ calibrates the mixture and selects the threshold that maps the uncertainty score to one of the three actions (Section 3.4).

3.1 IMMUTABLE SEMANTIC ANCHOR AND CLASS-AWARE ROUTING

A fully adaptive expert pool contains no unchanged reference when all its prompts drift together. The frozen zero-shot expert provides this reference by construction. Its class embeddings T_{i0} are obtained from Eq. 1 with the hand-crafted template instead of learned context vectors, computed once per class name and stored as a non-trainable buffer that is never updated. Although its scalar temperature τ_0 may adapt, its semantic representation remains identical to that of zero-shot CLIP for each client.

Routing decides how much each expert contributes to the final prediction. Making this decision on a single averaged class credit can lose evidence when class directions cancel each other, so we instead compute class-aware relevance. Let W_x and W_t denote lightweight local projections applied to image and class embeddings, respectively. For expert $k \in \{0, \ldots, K_i\}$ (including the zero-shot expert), we define

$$q(x) = \mathrm{norm}(W_x f(x)), \qquad z_{ik,c} = \mathrm{norm}(W_t T_{ik,c}), \tag{3}$$

$$e_{ik}(x) = \kappa^{-1} \log \sum\nolimits_{c=1}^{C} \exp\{\kappa\, q(x)^\top z_{ik,c}\}, \tag{4}$$

$$a_{ik}(x) = \mathrm{softmax}_k(e_{ik}(x)/\tau_r)\,, \tag{5}$$

where $\mathrm{norm}(\cdot)$ denotes ℓ_2 normalization, $\kappa > 0$ controls evidence sharpness, and $\tau_r > 0$ controls routing sharpness. The log-sum-exp evidence $e_{ik}(\cdot)$ is a smooth approximation to the best class alignment. By $e_{ik}(\cdot)$, an expert whose several class directions align moderately well with the input also receives routing credit, rather than only the expert with the single closest class.

However, learned routing can eliminate the semantic reference by assigning it vanishing weight. To prevent this, the zero-shot expert receives a guaranteed minimum mass $\rho \in [0, 1)$. Specifically, the routing weights are transformed as

$$\begin{aligned} r_{ik}(x) &= (1-\rho)\, a_{ik}(x), && k = 1, \dots, K_i, \\ r_{i0}(x) &= \rho + (1-\rho)\, a_{i0}(x). \end{aligned} \tag{6}$$

The transformed weight satisfies $r_{i0}(x) \geq \rho$. This floor does not render the anchor correct but rather bounds how far the final distribution can deviate from the zero-shot prediction (Appendix B).

3.2 Probability mixture and decomposed uncertainty

As discussed in Section 2, adapted and zero-shot prompts produce logits on different scales. Summing logits directly would entangle expert preference with scale. We therefore mix the temperature-controlled expert probabilities of Eq. 2 by the transformed routing weight $r_{ik}(x)$ as

$$\bar{p}_i(y \mid x) = \sum\nolimits_{k=0}^{K_i} r_{ik}(x)\, p_{ik}(y \mid x), \tag{7}$$

which makes the expert probability mixture explicit and supports the entropy decomposition below.

The mixture admits a decomposition of uncertainty by source. Using $H(\cdot)$ to denote Shannon entropy and $r_i(x) = (r_{i0}(x), \dots, r_{iK_i}(x))$ for the routing weights of Eq. 6, for client i and input x, we define

$$A_i(x) = \sum\nolimits_{k=0}^{K_i} r_{ik}(x)\, H(p_{ik}(\cdot \mid x)) \qquad \text{(within-expert ambiguity)}, \tag{8}$$

$$D_i(x) = H(\bar{p}_i(\cdot \mid x)) - A_i(x) \qquad \text{(expert disagreement)}, \tag{9}$$

$$R_i(x) = H(r_i(x)) \qquad \text{(routing ambiguity)}. \tag{10}$$

These three terms are large when individual experts are diffuse, when their class distributions conflict, and when the router spreads mass across experts, respectively. To compare fairly, we normalize $\widetilde{A}_i(x) = A_i(x)/\log C$, $\widetilde{D}_i(x) = D_i(x)/\log C$, and $\widetilde{R}_i(x) = R_i(x)/\log(K_i+1)$, and combine them as

$$U_i(x) = \lambda_a \widetilde{A}_i(x) + \lambda_d \widetilde{D}_i(x) + \lambda_r \widetilde{R}_i(x), \qquad (\lambda_a, \lambda_d, \lambda_r) \in \Delta^2, \tag{11}$$

so that each component and $U_i(x)$ itself lie in $[0, 1]$. We deliberately exclude predictive entropy as a fourth term, because it decomposes exactly into $A_i(x) + D_i(x)$ and including it would count disagreement twice. Proposition 1 in Appendix B formalizes this decomposition. We emphasize that the weighted score $U_i(x)$ is a diagnostic hypothesis, and the experiments compare it against each component in isolation.

3.3 Geometry-regularized local optimization

Local prompt adaptation can distort the original class distribution and further damage the final prediction. We use $k = 1$ to index the locally trained expert on client i and $k = 0$ to index the frozen zero-shot expert. Let $T_{i1}, T_{i0} \in \mathbb{R}^{C \times d}$ be their class-embedding matrices derived from Eq. 1. The mean class embedding is given by $\mu(T) = \frac{1}{C}\sum_{c=1}^{C} T_c$, where T denotes either T_{i0} or T_{i1}, and T_c corresponds to the c-th row of T. To prevent generalization loss, we regularize both the mean feature and the pairwise class geometry toward their zero-shot counterparts by

$$\mathcal{L}_{\mathrm{geo},i} = \big\|\mu(T_{i1}) - \mu(T_{i0})\big\|_2^2 + \frac{1}{C^2}\big\|T_{i1}T_{i1}^\top - T_{i0}T_{i0}^\top\big\|_F^2. \tag{12}$$

The first term constrains the global location of the class distribution, while the second term constrains its pairwise angles. This objective penalizes changes to the shared class distribution without forcing every tuned embedding to coincide with its zero-shot counterpart (Sharifdeen et al., 2026).

Client i optimizes the mixture on $\mathcal{D}_i^{\mathrm{tr}}$ via

$$\mathcal{L}_i = \frac{1}{|\mathcal{D}_i^{\mathrm{tr}}|} \sum_{(x,y)\in\mathcal{D}_i^{\mathrm{tr}}} \Big[-\log \bar{p}_i(y \mid x) + \beta_h \big[\eta - \widetilde{R}_i(x)\big]_+^2\Big] + \beta_g \mathcal{L}_{\mathrm{geo},i} + \beta_t \sum\nolimits_{q\in\{r,a,0\}} (\log \tau_q)^2, \tag{13}$$

where $\bar{p}_i(y \mid x)$ is the mixture prediction in Eq. 7 and $\widetilde{R}_i(x)$ is the normalized routing entropy in Eq. 11. The routing penalty discourages collapse only while $\widetilde{R}_i(x)$ falls below the threshold η. The log-temperature penalty prevents extreme scales, and the temperatures are constrained to $[0.05, 10]$. Only the local context vectors $\{\boldsymbol{v}_{i1,\ell}\}_{\ell=1}^{L}$, the routing projections W_x and W_t, and the temperatures τ_r, τ_a, and τ_0 receive gradients. Downloaded prompts and the anchor remain fixed without gradients.

3.4 Held-out calibration and conformal selective decisions

Learning, calibration, and action selection use disjoint data splits, which is why Section 2 requires $\mathcal{D}_i^{\mathrm{cal}}$ to be disjoint from $\mathcal{D}_i^{\mathrm{tr}}$. After minimizing Eq. 13, we freeze all trained parameters. Then we fit an output temperature $\tau_i^{\mathrm{out}} > 0$ by minimizing the negative log-likelihood of $\mathrm{softmax}\big(\log \bar{p}_i(\cdot \mid x)/\tau_i^{\mathrm{out}}\big)$ over $\mathcal{D}_i^{\mathrm{cal}}$, where $\bar{p}_i(\cdot \mid x)$ is the mixture prediction in Eq. 7. This alters the sharpness of the returned prediction. Because $U_i(x)$ is computed from the untempered mixture and routing weights in Eq. 11, $U_i(x)$ and the independently tempered zero-shot expert $p_{i0}(\cdot \mid x)$ remain unchanged.

$\mathcal{D}_i^{\mathrm{cal}}$ is then reused to select the acceptance threshold. This double use is benign because $U_i(x)$ is independent of τ_i^{out}. Let $n_i = |\mathcal{D}_i^{\mathrm{cal}}|$ and let $U_{i,(1)} \le \cdots \le U_{i,(n_i)} \le U_{i,(n_i+1)} = +\infty$ denote the order statistics of the calibration scores $\{U_i(x) : (x, y) \in \mathcal{D}_i^{\mathrm{cal}}\}$. For a target local-ID coverage level $q \in (0, 1)$, we set

$$k_i = \big\lceil (n_i + 1)\, q \big\rceil, \qquad \gamma_i = U_{i,(k_i)}. \tag{14}$$

The three actions of Section 2 are then determined as follows. Whenever $U_i(x) \le \gamma_i$, the system *accepts* and returns the tempered mixture $\mathrm{softmax}\big(\log \bar{p}_i(\cdot \mid x)/\tau_i^{\mathrm{out}}\big)$. Otherwise, it either *falls back* to the zero-shot expert and returns $p_{i0}(\cdot \mid x)$ while recording a zero-shot fallback event, or *abstains*.

The theoretical analyses and proofs of pFedOPEN are provided in Appendix B.

4 Experiments

4.1 Experimental setup

Datasets and baselines. Five CLIP datasets (i.e., Flowers102 (Nilsback & Zisserman, 2008), OxfordPets (Parkhi et al., 2012), Food101 (Bossard et al., 2014), Caltech101 (Fei-Fei, 2004), and DTD (Cimpoi et al., 2014)) form the primary benchmark for calibration, screened far-OOD detection, selective prediction, and few-shot label-shift evaluation. Each dataset is divided among 10 clients with disjoint class sets, producing a non-IID partition. CIFAR100 (Krizhevsky et al., 2009; 2010) use 100 clients and a Dirichlet label shift with $\alpha = 0.5$. DomainNet (Peng et al., 2019) and Caltech10 (Gong et al., 2012) assess joint feature and label shift with $\alpha = 0.3$, giving 30 DomainNet clients and 20 Caltech10 clients. The base-to-novel test uses the seven datasets of pFedMMA: SUN397 (Xiao et al., 2010), OxfordPets, Flowers102, DTD, Caltech101, UCF101 (Soomro et al., 2012), and Food101. Each of them is divided evenly, with half assigned to base and half to novel. Base classes are partitioned into ten clients with disjoint class subsets. Each client is evaluated on its local base classes, other clients' base classes, and novel classes excluded from prompt training and calibration. pFedMoAP and pFedMMA are the principal baselines. We test them on the above datasets and the base-to-novel protocols. Their evaluations share client partitions, communication rounds, participation rates, local epochs, seeds, calibration splits, OOD manifests, and inference data with the proposed pFedOPEN.

Metrics. We use local base, other clients' base, and novel accuracies and their harmonic mean (H.M.) for the base-to-novel generalization test. We further report metrics in four groups to probe different aspects of pFedOPEN. *Calibration and scoring:* expected calibration error (ECE) with 15 equal-width confidence bins, negative log-likelihood (NLL), and the multiclass Brier score. ECE measures the gap between stated confidence and realized accuracy, while NLL and the Brier score are proper scoring rules that jointly penalize miscalibration and inaccuracy. *OOD detection:* the area under the ROC curve (AUROC) and the PR curve (AUPR), treating $U_i(x)$ as the detection statistic for separating ID from OOD inputs. *Selective prediction:* the area under the risk–coverage curve (AURC), which summarizes error rates across all abstention thresholds, and the risk at 80% coverage (R@80), which fixes a single operating point of practical interest. *Conformal validity and deferral behavior:* the realized coverage compared against the target, and the fallback or abstention rate.

4.2 EMPIRICAL DIAGNOSIS

4.2.1 BASE-TO-NOVEL GENERALIZATION

Table 1: Dataset-wise and average base-to-novel transfer results.

Method	Dataset	Local Base ↑	Others' Base ↑	Novel ↑	H.M. ↑
pFedMoAP	SUN397	89.45±1.11	14.30±1.19	21.46±1.31	17.07±1.26
	OxfordPets	99.13±0.72	57.43±2.07	67.50±5.17	61.56±1.92
	Flowers102	**99.75±0.10**	26.04±1.36	22.61±1.44	23.83±1.44
	DTD	93.36±2.88	40.63±2.12	27.65±2.47	32.80±2.38
	Caltech101	98.97±0.51	88.22±0.67	85.53±1.20	86.81±0.82
	UCF101	94.23±1.92	47.37±3.41	47.09±1.40	47.06±2.21
	Food101	96.04±0.68	52.10±2.15	62.08±1.42	56.60±1.82
	Average	95.85±0.48	46.58±1.18	47.70±0.42	46.53±0.75
pFedMMA	SUN397	**93.98±0.15**	30.91±0.70	43.58±0.69	36.14±0.69
	OxfordPets	**99.58±0.09**	58.54±4.08	72.82±2.16	64.59±3.37
	Flowers102	99.66±0.10	28.94±1.14	40.06±1.39	33.36±0.96
	DTD	**96.10±0.72**	21.71±2.42	30.48±4.66	25.19±3.27
	Caltech101	**99.51±0.23**	78.11±2.11	85.13±1.13	81.36±1.27
	UCF101	**95.98±0.38**	38.22±0.57	47.40±0.44	42.12±0.47
	Food101	**96.66±0.10**	54.01±2.34	64.78±1.38	58.89±1.96
	Average	**97.35±0.08**	44.35±1.55	54.89±1.53	48.81±1.52
pFedOPEN (Ours)	SUN397	79.29±1.95	**53.49±2.67**	**56.84±0.58**	**55.06±1.67**
	OxfordPets	89.90±1.34	**89.77±1.46**	**92.69±1.36**	**91.19±0.22**
	Flowers102	79.45±9.67	**76.86±14.91**	**67.74±3.31**	**71.34±6.28**
	DTD	54.80±2.84	**54.70±2.61**	**43.50±3.49**	**48.31±1.18**
	Caltech101	92.04±0.50	**94.60±0.44**	**90.11±0.96**	**92.30±0.72**
	UCF101	71.50±1.41	**72.43±1.26**	**68.23±1.37**	**70.25±0.13**
	Food101	86.24±0.39	**86.33±0.32**	**86.12±0.56**	**86.23±0.44**
	Average	79.03±1.72	**75.46±1.77**	**72.18±1.01**	**73.53±0.86**

Table 1 presents base-to-novel results. Values are reported as mean ± standard deviation over three seeds, where each seed result averages ten clients. H.M. is computed per client and averaged within each seed before the seed-level summary. All methods use ViT-B/16 as the backbone. pFedOPEN uses the calibrated mixture with abstention, and both baselines use their raw settings. The comparison separates local fitting from transfer. pFedMMA has the highest local accuracy at 97.35%, followed by pFedMoAP at 95.85% and pFedOPEN at 79.03%. This local advantage does not transfer to held-out classes. pFedOPEN reaches 75.46% on other clients' base classes, 72.18% novel classes, and 73.53% harmonic-mean accuracy, exceeding pFedMMA by 31.11, 17.29, and 24.72 points, respectively. pFedMMA improves over pFedMoAP by 7.19 novel and 2.28 harmonic-mean points but is 2.23 points lower on other-base accuracy. In summary, pFedOPEN has the highest base-to-novel transfer capabilities. Figure 2 exposes the local fitting and held-out-class transfer behind these averages.

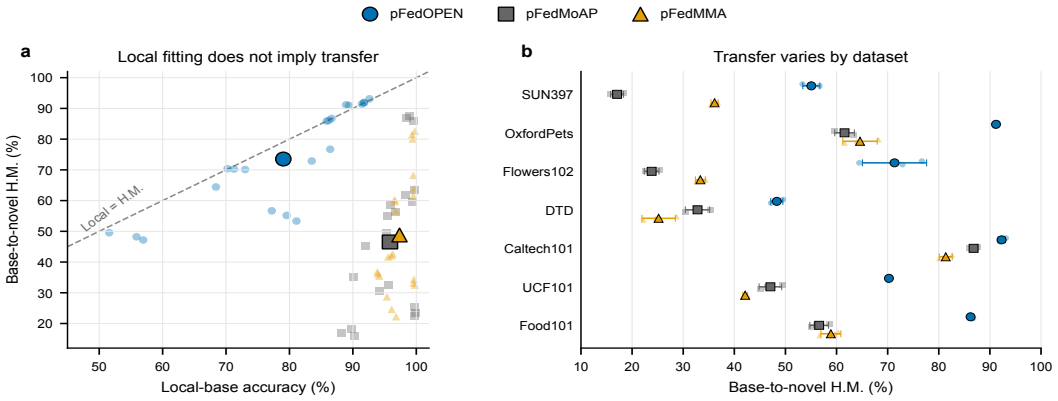


Figure 2: Empirical transfer mechanism under the base-to-novel protocol. **(a)** Small markers are dataset-seed units, and large markers are method means. Notably, the dashed identity denotes equal local-base and harmonic-mean accuracy. **(b)** Faint points are the three seed values for each dataset and method. Larger markers show the mean and whiskers show the sample standard deviation.

4.2.2 PREDICTIVE AND OPEN-WORLD PROFILE

Table 2 profiles pFedOPEN on each of the five CLIP benchmarks. Accuracy, ECE, NLL, AURC, and R@80 are computed from the calibrated mixture prediction. OOD AUROC and the fallback action rate are both driven by $U_i(x)$. ΔFallback isolates the effect of that action, and it is the change in overall accuracy when every flagged mixture prediction is replaced by the zero-shot expert, so that negative values indicate that the replacement is harmful. Two patterns emerge. First, OOD detection is strong but uneven: AUROC lies between 95.98% and 99.74% on four of five datasets, yet drops to 75.51% on DTD, whose AURC 35.68% alone accounts for the most of mean selective risk. Second,

Table 2: pFEDOPEN results on the five CLIP benchmarks.

Dataset	Acc. ↑	ECE ↓	NLL ↓	AUROC ↑	AURC ↓	R@80 ↓	Action	ΔFallback
Flowers102	93.93	5.96	0.870	99.67	1.35	2.28	14.86	−8.12
OxfordPets	88.98	7.25	0.672	99.74	3.17	6.35	10.95	−1.13
Food101	78.78	9.83	1.468	99.16	6.86	12.97	6.61	−0.47
Caltech101	88.25	9.24	0.821	95.98	3.92	6.86	10.26	−0.98
DTD	48.60	14.22	1.936	75.51	35.68	46.22	13.16	−1.30
Mean	79.71	9.30	1.153	94.01	10.20	14.93	11.17	−2.40

and more consequentially, the zero-shot fallback *reduces* accuracy on each dataset, with losses ranging from 0.47 points on Food101 to 8.12 points on Flowers102. The flagged inputs are thus cases the mixture already classifies correctly but the zero-shot expert does not.

4.3 CALIBRATION AND SELECTIVE-PREDICTION ANALYSES

Table 3: Calibration and selective-prediction analyses. Flagged-subset accuracy is computed on each method's own flagged inputs and is therefore not a paired comparison across columns. Notably, all values are percentages.

	pFEDOPEN		pFEDMOAP		pFEDMMA	
Analysis	Mix	Fallback	Raw	+Temp.	Raw	+Temp.
Adaptive-bin ECE	9.19	8.91	4.17	3.31	1.25	2.22
Client-median ECE	8.23	7.84	4.08	2.65	0.98	1.40
Risk at 90% coverage	17.13	18.44	2.07	1.93	0.91	1.23
Realized coverage ($q = .8$)	88.83	88.83	89.77	92.77	88.89	95.08
Flagged-subset accuracy	48.96	33.22	77.90	69.78	78.71	67.90

Table 3 addresses two questions: whether the conformal threshold of pFEDOPEN delivers the coverage derived by Proposition 2, and whether the fallback improves on abstention. We evaluate pFEDOPEN in two variants, i.e., the calibrated mixture with abstention (Mix) and with zero-shot replacement of flagged predictions (Fallback). In addition, we run experiments on pFEDMOAP and pFEDMMA, where each baseline is tested in raw and post-hoc temperature-scaled forms. The baselines provide context on calibration quality and do not implement the fallback mechanism.

Realized coverage matches the finite-sample level, not the nominal target. With a calibration set of $n = 8$ per client and target level $q = 0.8$, the conformal quantile uses rank $k = \lceil (n+1)q \rceil = 8$, so the finite-sample acceptance level is $k/(n+1) = 8/9 \approx 88.9\%$, not 80%. The realized pFEDOPEN coverage of 88.83% therefore agrees with the marginal guarantee of Proposition 2 and it should not be read as overshooting an 80% target. Coverage is identical for the Mix and Fallback variants because both apply the same threshold and differ only in the action taken on flagged inputs.

Fallback vs abstention. Inferring flagged inputs with the zero-shot expert trades a small calibration gain for a selective-risk loss: adaptive-bin ECE improves by 0.28 points ($9.19 \to 8.91$), but risk at 90% coverage worsens by 1.31 points ($17.13 \to 18.44$). The flagged-subset row identifies the cause: the adaptive mixture attains 48.96% accuracy while the zero-shot expert attains only 33.22%.

Baseline flagged accuracies are contextual. pFEDMMA attains the highest flagged-subset accuracy (78.71%), which falls to 67.90% under temperature scaling. pFEDMMA's realized coverage rises to 95.08%. Because each method flags its own subset of inputs, flagged-subset accuracy characterizes each method's own deferral behavior and does not support comparisons across columns.

Table 4: Conformal coverage under registered feature shift. All values are percentages.

Dataset	Target	Realized	Gap	ECE ↓	AURC ↓
DomainNet	80.00	59.19	−20.81	31.99	90.70
Caltech10	80.00	87.19	+7.19	23.94	88.60

Coverage under registered feature shift. Table 4 explores what happens when the exchangeability assumption underlying Proposition 2 is deliberately broken. Each client's threshold is calibrated on its own domain and then applied to a registered cross-domain test split, obtained by shifting every client to the next five-client domain block. The audit confirms that calibration does not transfer across domains, and that the failure is not even directionally consistent: DomainNet under-covers by 20.81 points while Caltech10 over-covers by 7.19 points. Both settings additionally exhibit severe miscalibration (ECE above 23%) and nonsense selective ordering (AURC near 90%). These results are consistent with a failure of exchangeability under domain shift.

4.4 WHERE DOES THE UNCERTAINTY COME FROM?

U is a weighted sum of expert disagreement D, within-expert ambiguity A, and routing ambiguity R. Table 5 asks which of these components carries the reliability signal by evaluating each in isolation.

Table 5: Uncertainty-component analysis. AUROC and AUPR use screened external far-OOD data. Proxy AUROC uses another client's distribution. AUROC, AUPR, AURC, and proxy AUROC are percentages. Mean ID is the average unscaled score on in-distribution inputs.

Score	AUROC ↑	AUPR ↑	AURC ↓	Proxy AUROC ↑	Mean ID
D	82.36	87.02	12.64	50.24	0.089
A	**95.07**	**98.45**	10.69	**55.95**	0.176
R	41.94	66.92	22.67	50.39	0.828
U	94.01	98.02	**10.51**	54.93	0.267

Component-level findings. The three components contribute very unequally. A is the dominant signal. It exceeds U by 1.06 AUROC points (95.07 vs. 94.01) and 0.43 AUPR points on OOD detection, and its AURC of 10.69% is close to U's 10.51%. D is an informative but weaker detector (82.36% AUROC). By contrast, R is *anti-correlated* with OOD (41.94% AUROC) and yields a worse selective ordering (22.67% AURC): far-OOD inputs tend to produce *more* decisive routing than ID inputs, and its high Mean ID 0.828 means it enters U with substantial weight. For the combination, U nevertheless attains the best AURC, indicating D and A jointly compensate for R in selective ordering even though R dilutes the OOD signal. All components remain near 50–56% for proxy AUROC, confirming that none of them separates inter-client shift. In summary, A carries most of the reliability signal, D is secondary, and R is harmful.

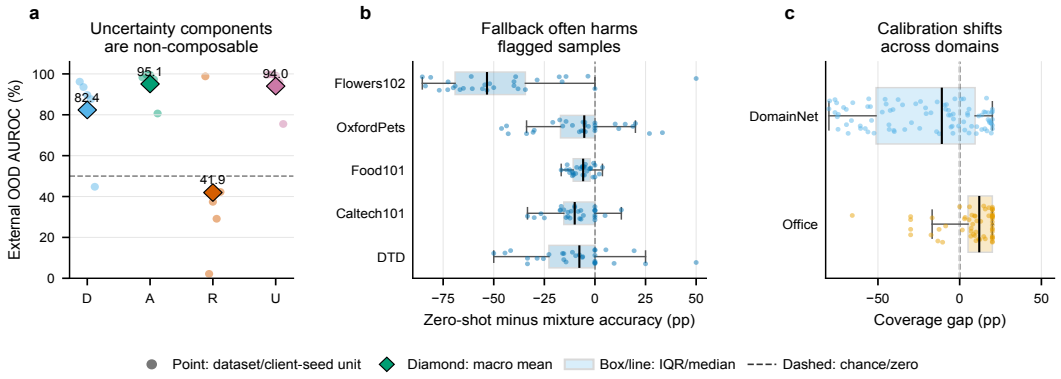


Figure 3: Reliability illustration. **(a)** OOD AUROC per uncertainty: small points are the five dataset-level means for D, A, R, and U, diamonds are macro means, and the dashed line is chance AUROC. **(b)** Zero-shot minus mixture accuracy on each client's flagged subset. Positive values favor fallback. **(c)** Coverage gaps under registered feature shift (90 DomainNet and 60 Caltech10 pairs), where positive values denote over-coverage, boxes denote the interquartile range, center lines denote the median, whiskers extend to 1.5x of the interquartile range, and dashed lines mark zero.

Heterogeneity behind the aggregates. Figure 3 dissects the three reliability mechanisms. Panel (a) shows that the component ranking $A \succ D \succ R$ is stable across datasets rather than an artifact of averaging. Panel (b) plots the difference between zero-shot and mixture accuracy on each client's flagged subset: the mass below zero shows that the fallback deficit is not driven by a few clients but holds for the large majority. Panel (c) shows the coverage gaps under registered feature shift: the DomainNet and Caltech10 distributions lie on opposite sides of zero, confirming that the aggregate under- and over-coverage in Table 4 reflect systematic, dataset-level behavior.

4.5 Ablation Studies

Table 6: Base-to-novel ablations. All values use mixture predictions and are averaged over three seeds and ten clients. The H.M. is computed per client before averaging.

	Food101			DTD		
Variant	Base	Novel	H.M.	Base	Novel	H.M.
Full pFedOPEN	**86.33**	**86.12**	**86.23**	54.70	**43.50**	**48.31**
No zero-shot expert	85.70	85.88	85.78	**55.29**	41.36	47.21
No geometry loss	86.22	86.02	86.12	54.08	43.57	48.23

Component contributions to base-to-novel generalization. Table 6 reflects the ablated component alone. Both contribute positively to the H.M. on both datasets, though with different magnitudes and mechanisms. Removing the zero-shot expert lowers the H.M. by 0.45 points on Food101 and 1.10 points on DTD. The geometry loss contributes more modestly (0.11 and 0.08 H.M. points), with its effect concentrated on base accuracy for DTD (54.70 vs. 54.08) and essentially neutral on novel classes. H.M. gains are small but directionally consistent across both datasets, favoring the inclusion of both components while showing the zero-shot expert is the main driver of novel-class generalization.

Table 7: Component ablations evaluated on Food101. "Fixed threshold" replaces held-out calibration with a prespecified threshold. Every row uses the mixture with abstention yet without fallback, so that the component change is not confounded by zero-shot replacement. All metrics except NLL are percentages and are calculated by averaging client-seed measurements.

Variant	Acc. ↑	ECE ↓	NLL ↓	Proxy AUROC ↑	AURC ↓	Cov.	Action
Full pFEDOPEN	78.78	9.83	1.468	50.64	6.86	93.39	6.61
No zero-shot expert	78.63	**8.58**	1.161	50.13	6.81	91.38	8.62
$\rho = 0$	79.89	11.73	1.689	52.05	6.46	92.14	7.86
Mean-prototype router	79.04	9.31	1.361	50.48	6.70	91.12	8.88
Logit mixing	**81.56**	8.72	1.568	**58.29**	**4.99**	91.82	8.18
No geometry loss	79.35	10.61	1.502	50.85	6.64	92.93	7.07
No router entropy floor	78.96	10.38	1.420	50.40	6.70	93.20	6.80
Fixed expert temperatures	78.99	9.40	1.395	50.60	6.72	92.18	7.82
No output temperature	78.96	12.76	**0.839**	50.40	6.70	93.20	6.80
$\lambda_d = 0$	78.96	10.38	1.420	50.49	6.58	92.69	7.31
$\lambda_r = 0$	78.96	10.38	1.420	50.40	6.70	93.19	6.81
$\lambda_a = 0$	78.96	10.38	1.420	50.20	7.79	92.04	7.96
Fixed threshold	78.96	10.38	1.420	50.40	6.70	100.00	0.00

No configuration is Pareto-dominant. Table 7 does not identify the full configuration as a Pareto optimum: every variant improves at least one metric at the cost of another. Replacing probability averaging with logit mixing raises accuracy by 2.78 points and lowers AURC by 1.87 points, but worsens NLL. Removing the anchor floor gains 1.11 accuracy points at the cost of a 1.90-point ECE increase. Removing the zero-shot expert yields the lowest ECE. Removing the output temperature yields the lowest NLL but simultaneously produces the highest ECE. Finally, the fixed-threshold variant is degenerate rather than competitive: it accepts each ID example and therefore performs no selective action at all, so its apparent coverage advantage is vacuous. Each component is beneficial for some metrics and harmful for others, so none is uniformly beneficial. The variant selection depends on what the deployment prioritizes.

Table 8: Sensitivity ranges. Each row varies one factor with all others at default values. Calibration-size rows use CIFAR100 while all other rows use Food101, so comparisons are valid within rows.

Factor	Values	Acc. range (%)	ECE range (%)	AURC range (%)	Time range (Min.)	Seeds
Shots	$\{1, 2, 4, 8, 16\}$	76.96–78.41	5.92–9.37	6.44–6.92	39.5–49.3	3
Calibration size	$\{25, 50, 100, 200\}$	50.62–53.18	12.92–13.91	23.54–25.98	453.9–563.4	3
Target coverage	$\{.70, .80, .90, .95\}$	77.52–78.79	7.83–10.19	6.72–6.90	49.4–50.7	2–3
Adaptive experts	$\{1, 3, 5, 10\}$	78.41–87.22	6.42–9.37	1.82–6.74	48.4–50.7	2–3
Anchor floor ρ	$\{0, .05, .10, .20, .30\}$	78.05–79.01	8.00–10.25	6.54–6.78	50.7–51.7	3
Geometry β_g	$\{0, .01, .05, .10, .20\}$	78.31–78.63	8.86–9.58	6.66–6.96	50.8–51.2	3
Uncertainty term	$\{D, R, A\}$	78.27–78.38	9.19–9.53	6.62–22.02	50.5–50.9	3
Temperature bounds	$[.05, 5]$, $[.05, 10]$, $[.1, 10]$	78.41–78.41	9.37–9.37	6.74–6.74	49.9–50.7	3

Sensitivity. Table 8 reports the *range* of outcomes of each sweep. These sweeps characterize the sensitivity of the involved components. The sweeps separate into insensitive and sensitive factors. On the insensitive side, the anchor floor and the geometry weight produce narrow ranges in both accuracy (at most 0.96 points) and AURC (at most 0.30 points). In particular, all three intervals in the temperature-bound sweep yield identical metrics, indicating that the bounds are inactive constraints. On the sensitive side, two factors stand out. The number of adaptive experts exhibits the largest ranges (8.81 accuracy points and 4.92 AURC points). As expected, the uncertainty selection leaves accuracy and ECE almost unchanged, since it affects only the selection instead of the classifier. But it widens the AURC range to 6.62–22.02, where the upper end corresponds to the routing-only variant.

5 Conclusion

Reliable federated prompt mixtures require more than accurate routing. pFEDOPEN instruments a prompt pool with an immutable reference, separate uncertainty components, held-out calibration, and explicit abstention or fallback actions. The main empirical finding is the non-composability of components. Within-expert ambiguity has the highest mean OOD AUROC, routing ambiguity can be harmful, and their weighted combination is not uniformly better. The anchor and geometry terms give modest base-to-novel gains. Zero-shot fallback fails its conditional risk test and abstention is preferable on the evaluated data. A reliability component can therefore be useful in one role and harmful after combination or action selection. These findings argue for component-wise reporting and conditional action audits instead of treating a single uncertainty score or semantic anchor as a reliability guarantee. The theory identifies information, influence, marginal coverage, and fallback conditions without claiming OOD correctness.

AI USE STATEMENT

In this work, we used generative AI tools for writing assistance, specifically to polish the phrasing, clarity, and organization of author-written text. We have not used generative AI tools for research ideation, methodology design, theoretical analysis, experimentation, or data analysis. Disclosure regarding AI-generated code is not applicable to this work, as no code was produced with generative AI. All technical content, including the problem formulation, methods, theoretical results, experiments, and analyses, was conceived and carried out by the authors themselves. We have reviewed all AI-assisted work. Every AI-polished passage was checked by the authors against the original text to verify that its technical meaning, claims, and numerical values were preserved, and was further edited where needed. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work follows the ICLR Code of Ethics. It uses public image benchmarks under their original licenses and involves no new data collection or human-subjects research. Federated execution is not a privacy guarantee: exchanged prompts and routing information may expose client properties, and pretrained VLM biases may persist through the zero-shot anchor. Sensitive deployment requires privacy analysis, subgroup-specific coverage and risk audits, and human review of abstentions.

REPRODUCIBILITY STATEMENT

Section 3 specifies the model and objectives. Appendices C.1 and C document datasets, processing, client partitions, calibration, hyperparameters, baselines, and statistical procedures. Appendices C and D provide result provenance and implementation correspondence. The code repository contains implementations, launchers, hashed manifests, resolved configurations, structured outputs, aggregation scripts, metric tests, and the controlled simulator.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27284–27293, 2024a.

Yichen Bai, Zongbo Han, Bing Cao, Xiaoheng Jiang, Qinghua Hu, and Changqing Zhang. ID-like prompt learning for few-shot out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17480–17489, 2024b.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101: Mining discriminative components with random forests. In *European Conference on Computer Vision*, pp. 446–461, 2014.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014.

Tianyu Cui, Hongxia Li, Jingya Wang, and Ye Shi. Harmonizing generalization and personalization in federated prompt learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9646–9661. PMLR, 2024.

Li Fei-Fei. Learning generative visual models from few training examples. In *Workshop on Generative-Model Based Vision, IEEE Proc. CVPR, 2004*, 2004.

Chun-Mei Feng, Bangjun Li, Xinxing Xu, Yong Liu, Huazhu Fu, and Wangmeng Zuo. Learning federated visual prompt in null space for mri reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8064–8073, June 2023.

Sajjad Ghiasvand, Mahnoosh Alizadeh, and Ramtin Pedarsani. pfedmma: Personalized federated fine-tuning with multi-modal adapter for vision-language models. In *International Conference on Learning Representations*, volume 2026, pp. 76213–76240, 2026.

Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2066–2073, 2012.

Tao Guo, Song Guo, and Junxiao Wang. pFedPrompt: Learning personalized prompt for vision-language models in federated learning. In *The Web Conference*, pp. 1364–1374, 2023a.

Tao Guo, Song Guo, Junxiao Wang, Xueyang Tang, and Wenchao Xu. PromptFL: Let federated participants cooperatively learn prompts instead of models. *IEEE Transactions on Mobile Computing*, 2023b.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. Personalized cross-silo federated learning on non-iid data. In *AAAI Conference on Artificial Intelligence*, volume 35, pp. 7865–7873, 2021.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. Negative label guided ood detection with pretrained vision-language models. In *International Conference on Learning Representations*, 2024.

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, 2023a.

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

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. CIFAR-10 and CIFAR-100 Datasets, 2010.

Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

Hongxia Li, Wei Huang, Jingya Wang, and Ye Shi. Global and local prompts cooperation via optimal transport for federated learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.

Tianqi Li, Guansong Pang, Xiao Bai, Wenjun Miao, and Jin Zheng. Learning transferable negative prompts for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17584–17594, 2024b.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 2020.

Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5206–5215, 2022.

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

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In *Advances in Neural Information Processing Systems*, volume 35, pp. 35087–35102, 2022.

Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. LoCoOp: Few-shot out-of-distribution detection via prompt learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 76298–76310, 2023.

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

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, 2019.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *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*, pp. 8748–8763. PmLR, 2021.

Ashshak Sharifdeen, Fahad Shamshad, Muhammad Akhtar Munir, Abhishek Basu, Mohamed Insaf Ismithdeen, Jeyapriyan Jeyamohan, Chathurika Sewwandi Silva, Karthik Nandakumar, and Muhammad Haris Khan. Towards calibrating prompt tuning of vision-language models. *arXiv preprint arXiv:2602.19024*, 2026.

Yang Shu, Xingzhuo Guo, Jialong Wu, Ximei Wang, Jianmin Wang, and Mingsheng Long. CLIPood: Generalizing CLIP to out-of-distributions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31716–31731. PMLR, 2023.

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.

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

Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. CLIPN for zero-shot ood detection: Teaching CLIP to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1802–1812, 2023.

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.

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.

Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6757–6767, 2023.

Hantao Yao, Rui Zhang, and Changsheng Xu. Tcp:textual-based class-aware prompt tuning for visual-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23438–23448, June 2024.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130:2337–2348, 2022b.

APPENDIX CONTENTS

Section	Title	Page
A	Related Work	13
B	Theoretical Analysis	14
C	Extended Experimental Results	18
D	Implementation and Complexity	26

A RELATED WORK

Personalized federated and prompt learning. Federated personalization addresses client heterogeneity through local components, regularized objectives, or client-specific collaboration (Li et al., 2020; Huang et al., 2021). Several parameter-efficient variants replace full-model exchange with prompts. PromptFL communicates shared prompts, pFedPrompt learns personalized prompts, and FedOTP couples global and local prompts through optimal transport (Guo et al., 2023b;a; Li et al., 2024a). FedPR extends this principle to federated MRI reconstruction through null-space visual prompts, whereas FedPGP combines a CLIP-guided global prompt with low-rank personalized adaptation to balance generalization and personalization (Feng et al., 2023; Cui et al., 2024). DiPrompT separates global and latent-domain prompts for federated domain generalization without domain labels, while FedMVP conditions multimodal visual prompts on image and class context to improve transfer to unseen classes and domains (Bai et al., 2024a; Singha et al., 2025). Closest to our setting, PFEDMOAP retains shared prompts as individual experts and learns a local mixture, while pFedMMA studies personalized multimodal adaptation (Luo et al., 2025; Ghiasvand et al., 2026). These methods primarily optimize prediction, transfer, or communication. In contrast, PFEDOPEN treats exchanged prompts as an inspectable expert pool and audits their disagreement, calibration, and downstream selective actions.

Prompt generalization and calibration. CoOp adapts continuous context tokens while retaining CLIP's frozen encoders, but its static context can overfit the classes used for prompt optimization (Radford et al., 2021; Zhou et al., 2022b). CoCoOp generates instance-conditioned tokens, while ProDA represents prompts as a distribution to capture diverse visual contexts (Zhou et al., 2022a; Lu et al., 2022). Knowledge-preserving methods instead constrain adaptation toward the pretrained model: KgCoOp aligns learned and hand-crafted prompt embeddings, and PromptSRC combines frozen-model agreement, trajectory ensembling, and textual diversity (Yao et al., 2023; Khattak et al., 2023b). MaPLe couples hierarchical prompts across the visual and textual branches, and TCP injects class-aware textual knowledge for transfer to unseen domains (Khattak et al., 2023a; Yao et al., 2024). Beyond prompt-only adaptation, CLIPood combines class-adaptive margins with a moving ensemble to limit the loss of out-of-distribution generalization during fine-tuning (Shu et al., 2023). These methods improve transfer, but adaptation can still alter predictive confidence and class

geometry (Sharifdeen et al., 2026). pFEDOPEN addresses both effects inside a federated expert pool through an immutable zero-shot reference, geometry regularization, separate adaptive and zero-shot temperatures, and held-out output calibration.

OOD detection and selective prediction. Temperature scaling adjusts predictive confidence, whereas OOD scores rank unfamiliar inputs (Liu et al., 2020). Recent vision–language OOD methods use textual semantics: MCM performs zero-shot concept matching, CLIPN learns explicit negation semantics, NegLabel draws negative labels from a large vocabulary, and NegPrompt learns negative prompts that transfer to unseen class names (Ming et al., 2022; Wang et al., 2023; Jiang et al., 2024; Li et al., 2024b). Few-shot methods instead adapt the prompt space. LoCoOp treats irrelevant local CLIP features as surrogate OOD evidence, while ID-like prompt learning mines hard outliers near the in-distribution classes (Miyai et al., 2023; Bai et al., 2024b). Selective classification evaluates the risk retained after rejection, split conformal calibration gives finite-sample coverage under exchangeability, and conformal risk control extends this logic to expected monotone losses (Angelopoulos & Bates, 2021; Angelopoulos et al., 2024). Dataset shift can degrade calibration, OOD ranking, and selective risk simultaneously (Ovadia et al., 2019). Rather than proposing another standalone detector, pFEDOPEN decomposes uncertainty within a federated expert mixture and tests whether its calibrated actions remain supported when these criteria disagree.

We emphasize that the individual entropy identity, the temperature fit, and the conformal order statistic are established tools. Our contribution lies in their explicit instrumentation within a federated prompt-expert pool and in the joint audit showing that their rankings and deployment recommendations can conflict rather than improve monotonically.

B Theoretical Analysis

B.1 Uncertainty decomposition

This part fixes client i and input x and suppresses both indices, i.e., denoting $r_k = r_{ik}(x)$, $p_k = p_{ik}(\cdot \mid x)$, $\bar{p} = \bar{p}_i(\cdot \mid x)$, and A, D, R for $A_i(x)$, $D_i(x)$, $R_i(x)$. All statements are conditional on x.

Proposition 1 (Entropy decomposition)**.** *Let $Z \in \{0, \dots, K_i\}$ be an expert index with $\Pr\{Z = k\} = r_k$, and let $Y \in \{1, \dots, C\}$ be a class label with conditional distribution $\Pr\{Y = c \mid Z = k\} = p_k(c)$, so that the marginal law of Y is $\bar{p}$. Then*

$$H(\bar{p}) = A + D, \qquad D = \sum\nolimits_{k=0}^{K_i} r_k \, \mathrm{KL}(p_k \| \bar{p}) = I(Y; Z), \tag{15}$$

and $0 \le D \le \min\{R, H(\bar{p})\}$. The two upper bounds $D \le R$ and $D \le H(\bar{p})$ are the standard inequalities $I(Y;Z) \le H(Z)$ and $I(Y;Z) \le H(Y)$, since $H(Z) = R$ and $H(Y) = H(\bar{p})$.

Proof. Fix the client and input. The construction in Proposition 1 defines the joint probability mass function

$$\Pr\{Y = c, Z = k\} = r_k p_k(c).$$

Marginalizing over the expert index gives

$$\Pr\{Y = c\} = \sum_{k=0}^{K_i} r_k p_k(c) = \bar{p}(c).$$

Consequently, $H(Y) = H(\bar{p})$. Conditional on $Z = k$, the class label has distribution p_k, and hence

$$H(Y \mid Z) = \sum_{k=0}^{K_i} r_k H(p_k) = A.$$

The mutual-information identity $I(Y;Z) = H(Y) - H(Y \mid Z)$ therefore already yields $I(Y;Z) = H(\bar{p}) - A = D$ and $H(\bar{p}) = A + D$.

For completeness, the same identity follows directly from the weighted KL expression. Using $0 \log 0 = 0$ and assigning zero contribution to any term with $r_k = 0$, we obtain

$$\begin{aligned}\sum_{k=0}^{K_i} r_k \,\mathrm{KL}(p_k \| \bar{p}) &= \sum_{k=0}^{K_i} \sum_{c=1}^{C} r_k p_k(c) \log \frac{p_k(c)}{\bar{p}(c)} \\ &= \sum_{k,c} r_k p_k(c) \log p_k(c) - \sum_c \left(\sum_k r_k p_k(c) \right) \log \bar{p}(c) \\ &= -\sum_k r_k H(p_k) - \sum_c \bar{p}(c) \log \bar{p}(c) \\ &= H(\bar{p}) - A = D.\end{aligned}$$

Whenever $r_k > 0$ and $p_k(c) > 0$, the mixture satisfies $\bar{p}(c) \geq r_k p_k(c) > 0$. Moreover, expanding the definition of mutual information under the joint law above gives

$$I(Y; Z) = \sum_{k: r_k > 0} \sum_c r_k p_k(c) \log \frac{r_k p_k(c)}{r_k \bar{p}(c)} = \sum_k r_k \,\mathrm{KL}(p_k \| \bar{p}) = D.$$

Nonnegativity follows either from the nonnegativity of KL divergence or from mutual information. Finally,

$$\begin{aligned}D = I(Y; Z) = H(Z) - H(Z \mid Y) \leq H(Z) = H(r) = R, \\ D = I(Y; Z) = H(Y) - H(Y \mid Z) \leq H(Y) = H(\bar{p}),\end{aligned}$$

because conditional entropy is nonnegative. Combining these inequalities gives $0 \leq D \leq \min\{R, H(\bar{p})\}$, as claimed.

□

The decomposition separates uncertainty within individual experts from conflict among their class distributions, where A is the conditional entropy $H(Y \mid Z)$ and D is the information that the predicted class carries about the selected expert. The upper bounds carry a practical warning. Since $D \leq R$, if an input is routed almost deterministically ($R \approx 0$), then $D \approx 0$ as well even when the experts assign sharply conflicting class distributions to x. This is because when one expert holds nearly all of the routing mass, the mixture never expresses the views of the dissenting experts. Therefore, no conflict is visible in $\bar{p}$. Reliable expert disagreement thus requires the router to retain a minimum level of entropy, which is precisely what the routing-entropy floor in Eq. 13 enforces. The identity $H(\bar{p}) = A + D$ also explains why Eq. 11 omits the predictive entropy $H(\bar{p})$. Because $H(\bar{p})$ decomposes exactly into A plus D, adding a term $\lambda H(\bar{p})$ to the score would be equivalent to raising the weights on A and D by λ each. Predictive entropy thus contributes no signal and including it would only obscure the interpretation of the score's weights.

B.2 Marginal ID coverage

Assumption 1 (ID exchangeability)**.** *Fix a client i. Conditional on the parameters trained on $\mathcal{D}_i^{\mathrm{tr}}$ and on the score function $U_i(\cdot)$, the calibration scores $\{U_i(x) : (x, y) \in \mathcal{D}_i^{\mathrm{cal}}\}$ and the score $U_i(X)$ of a future local-ID input X are exchangeable.*

Proposition 2 (Finite-sample marginal ID coverage)**.** *Under Assumption 1, the threshold γ_i of Eq. 14 satisfies*

$$\Pr\{U_i(X) \leq \gamma_i\} \;\geq\; \frac{k_i}{n_i + 1} \;\geq\; q, \tag{16}$$

where the probability is taken jointly over $\mathcal{D}_i^{\mathrm{cal}}$ and X. If the scores are almost surely distinct, the first inequality holds with equality.

Proof. Condition on the parameters learned from $\mathcal{D}_i^{\mathrm{tr}}$ and on the resulting fixed score function $U_i(\cdot)$. Write the calibration scores as $S_j = U_i(X_j)$ for $j = 1, \ldots, n_i$ and the future local-ID score as $S_{n_i+1} = U_i(X)$. Assumption 1 makes $(S_1, \ldots, S_{n_i+1})$ exchangeable under this conditioning. Recall that

$$k_i = \lceil (n_i + 1) q \rceil, \qquad 1 \leq k_i \leq n_i + 1.$$

First consider the boundary case $k_i = n_i + 1$. We complete the order-statistic rule by the natural convention $\gamma_i = +\infty$, which is the convention used by the implementation. Every future score is then accepted, and

$$\Pr\{S_{n_i+1} \le \gamma_i\} = 1 = \frac{k_i}{n_i + 1} \ge q.$$

It remains to consider $k_i \le n_i$. To account for possible ties, introduce i.i.d. auxiliary variables $V_1, \dots, V_{n_i+1} \sim \mathrm{Unif}(0, 1)$, independent of the scores, and order the pairs (S_j, V_j) lexicographically. The pairs remain exchangeable and are almost surely distinct. Thus the rank $R^\star$ of (S_{n_i+1}, V_{n_i+1}) among the $n_i + 1$ augmented pairs is uniform on $\{1, \dots, n_i + 1\}$, so

$$\Pr\{R^\star \le k_i\} = \frac{k_i}{n_i + 1}.$$

Let $S_{(k_i)}$ be the k_i-th order statistic of the n_i calibration scores. If $S_{n_i+1} > S_{(k_i)}$, at least k_i calibration scores are strictly smaller than S_{n_i+1}, which forces $R^\star > k_i$. Taking the contrapositive gives the event inclusion

$$\{R^\star \le k_i\} \subseteq \{S_{n_i+1} \le S_{(k_i)}\} = \{U_i(X) \le \gamma_i\}.$$

Therefore,

$$\Pr\{U_i(X) \le \gamma_i\} \ge \frac{k_i}{n_i + 1} \ge q,$$

where the last inequality follows from the definition of the ceiling. The conditioning can now be removed because the bound holds for every fixed set of trained parameters satisfying the stated exchangeability assumption.

If the $n_i + 1$ scores are almost surely distinct, the auxiliary tie breakers do not affect their ordering. In that case, $R^\star \le k_i$ holds if and only if $S_{n_i+1} \le S_{(k_i)}$, so the first inequality becomes equality. With ties, weak acceptance may include additional outcomes whose randomized rank exceeds k_i, which is why only the lower bound is asserted.

□

Two aspects of this guarantee deserve emphasis. First, the guarantee remains valid despite using $\mathcal{D}_i^{\mathrm{cal}}$ for both fitting the output temperature τ_i^{out} and selecting the threshold γ_i. Typically, such reuse is problematic because tuning the score function on calibration data would break exchangeability with a fresh test score, invalidating Proposition 2. However, the score function $U_i(\cdot)$ here depends only on parameters trained on $\mathcal{D}_i^{\mathrm{tr}}$, not on τ_i^{out}. Thus, temperature fitting does not alter any score in Eq. 14, the two calibration uses remain independent, and Assumption 1 still holds. Second, the result controls only the accepted local-ID inputs. It makes no claim about the novel or OOD regimes, guarantees neither accuracy, conditional risk, OOD rejection, nor subgroup coverage, and may fail once client drift violates Assumption 1.

B.3 Influence of the semantic anchor

Proposition 3 (Anchor radius and single-expert influence)**.** *Let the routing weights satisfy the floor $r_{i0}(x) \ge \rho$ in Eq. 6. Then for every client i and input x,*

$$\big\|\bar{p}_i(\cdot \mid x) - p_{i0}(\cdot \mid x)\big\|_1 \;\le\; 2\big(1 - r_{i0}(x)\big) \;\le\; 2(1 - \rho). \tag{17}$$

Moreover, fix the routing weight vector $r_i(x) = (r_{i0}(x), \dots, r_{iK_i}(x))$ ($k \ge 1$). Let $q_{ik}(\cdot \mid x)$ be an arbitrary distribution over the C classes and let $\bar{q}_i(\cdot \mid x)$ be the mixture obtained from $\bar{p}_i(\cdot \mid x)$ by replacing the adaptive expert $p_{ik}(\cdot \mid x)$ with $q_{ik}(\cdot \mid x)$. Then

$$\big\|\bar{p}_i(\cdot \mid x) - \bar{q}_i(\cdot \mid x)\big\|_1 \;=\; r_{ik}(x)\,\big\|p_{ik}(\cdot \mid x) - q_{ik}(\cdot \mid x)\big\|_1 \;\le\; 2\,r_{ik}(x). \tag{18}$$

Proof. Fix client i and input x, and suppress the argument x for readability. Because the routing weights sum to one,

$$\begin{aligned}\bar{p}_i - p_{i0} &= r_{i0}p_{i0} + \sum_{k=1}^{K_i} r_{ik}p_{ik} - p_{i0} \\ &= \sum_{k=1}^{K_i} r_{ik}(p_{ik} - p_{i0}).\end{aligned}$$

For any two class-probability vectors $u, v \in \Delta^{C-1}$,

$$\|u - v\|_1 = \sum_{c=1}^{C} |u(c) - v(c)| \le \sum_{c=1}^{C} \big(u(c) + v(c)\big) = 2.$$

Applying the triangle inequality and this diameter bound gives

$$\begin{aligned}\|\bar{p}_i - p_{i0}\|_1 &\le \sum_{k=1}^{K_i} r_{ik}\|p_{ik} - p_{i0}\|_1 \\ &\le 2\sum_{k=1}^{K_i} r_{ik} = 2(1 - r_{i0}) \le 2(1 - \rho),\end{aligned}$$

which proves Eq. 17.

For the single-expert statement, fix $k \ge 1$ and keep the entire routing vector unchanged. Before and after replacing p_{ik} by q_{ik}, respectively, the two mixtures are

$$\begin{aligned}\bar{p}_i &= \sum_{j \ne k} r_{ij}p_{ij} + r_{ik}p_{ik}, \\ \bar{q}_i &= \sum_{j \ne k} r_{ij}p_{ij} + r_{ik}q_{ik}.\end{aligned}$$

All unchanged expert terms cancel, leaving

$$\bar{p}_i - \bar{q}_i = r_{ik}(p_{ik} - q_{ik}).$$

Since $r_{ik} \ge 0$ and both expert outputs are probability vectors,

$$\|\bar{p}_i - \bar{q}_i\|_1 = r_{ik}\|p_{ik} - q_{ik}\|_1 \le 2r_{ik},$$

which proves Eq. 18. The fixed-routing condition is what makes the unchanged terms cancel.

□

The two parts of Proposition 3 play different roles. The anchor-radius bound in Eq. 17 holds pointwise in x and uniformly over the behavior of the adaptive experts. Even if all downloaded prompts fail simultaneously, the mixture $\bar{p}_i(\cdot \mid x)$ remains within an ℓ_1 ball of radius $2(1-\rho)$ around the zero-shot predictive $p_{i0}(\cdot \mid x)$. When zero-shot CLIP is itself wrong on x, anchoring preserves that error rather than correcting it, and the floor ρ therefore encodes a trade-off: a larger ρ tightens the worst-case radius but restricts the mixture's freedom to depart from the anchor on inputs where adaptation would help. The single-expert bound in Eq. 18 quantifies the impact of one downloaded prompt. With routing weights held fixed, replacing expert k changes the mixture by at most $2\,r_{ik}(x)$ in ℓ_1 distance. The fixed-routing condition is a real restriction. A perturbed prompt not only alters its expert's class distribution, but can also shift the routing weights themselves.

B.4 Additional Risk Qualification

The fallback rule replaces the mixture prediction with the zero-shot prediction whenever the uncertainty score exceeds the conformal threshold. The following proposition characterizes exactly when this substitution helps. It is a distribution-free identity: no assumption is placed on the joint law of (X, Y), on the quality of the uncertainty score, or on the relationship between the flagged event and either predictor.

Proposition 4 (Fallback risk decomposition). *Let ℓ be a bounded loss, U be the uncertainty score with client threshold γ_i, and $B = \{U(X) > \gamma_i\}$ denote the flagged event. Write R_{mixture} and R_{fallback} for the risks of the mixture rule and of the rule that predicts with the mixture on B^c and with the zero-shot expert on B. If $\Pr(B) > 0$, let $R_m^B = \mathbb{E}[\ell(\widehat{y}_m(X), Y) \mid B]$ and $R_0^B = \mathbb{E}[\ell(\widehat{y}_0(X), Y) \mid B]$ denote the conditional risks on the flagged event. Then*

$$R_{\text{fallback}} - R_{\text{mixture}} = \Pr(B)\left(R_0^B - R_m^B\right). \tag{19}$$

Consequently, fallback weakly improves total risk if and only if $R_0^B \le R_m^B$. If $\Pr(B) = 0$, the two rules have equal risk.

Proof. Define the sample-wise losses $L_m = \ell(\widehat{y}_m(X), Y)$ and $L_0 = \ell(\widehat{y}_0(X), Y)$. Boundedness of ℓ ensures all expectations below are finite. The fallback rule agrees with the mixture on B^c and with the zero-shot expert on B, so its loss is $L_{\text{fallback}} = \mathbf{1}_{B^c} L_m + \mathbf{1}_B L_0$. Taking expectations and writing the mixture risk over the same partition,

$$R_{\text{fallback}} = \mathbb{E}\big[\mathbf{1}_{B^c} L_m\big] + \mathbb{E}\big[\mathbf{1}_B L_0\big],$$
$$R_{\text{mixture}} = \mathbb{E}\big[\mathbf{1}_{B^c} L_m\big] + \mathbb{E}\big[\mathbf{1}_B L_m\big].$$

The first terms coincide, since both rules use the mixture prediction on B^c. Subtracting gives the unconditional identity

$$R_{\text{fallback}} - R_{\text{mixture}} = \mathbb{E}\big[\mathbf{1}_B\,(L_0 - L_m)\big]. \tag{20}$$

If $\Pr(B) > 0$, then $\mathbb{E}[\mathbf{1}_B L_0] = \Pr(B)\, R_0^B$ and $\mathbb{E}[\mathbf{1}_B L_m] = \Pr(B)\, R_m^B$ by the definition of conditional expectation, and substitution into Eq. 20 yields Eq. 19. Because $\Pr(B) > 0$, the sign of the risk difference equals the sign of $R_0^B - R_m^B$, which proves the equivalence. If $\Pr(B) = 0$, then $\mathbf{1}_B = 0$ almost surely, so Eq. 20 gives $R_{\text{fallback}} = R_{\text{mixture}}$ without requiring conditional risks on a null event to be defined. □

The value of Proposition 4 is operational: whether fallback helps reduces entirely to the sign of $R_0^B - R_m^B$, a quantity estimable from labeled flagged samples. Fallback should therefore be treated as a mechanism whose precondition must be validated per deployment, not as a default. In our flagged-subset analysis (Section 4), zero-shot accuracy on B is below mixture accuracy, so the condition fails, fallback increases risk in accordance with Eq. 19, and abstention with review is the supported action on flagged samples. The synthetic study of Section C.13 exhibits the opposite regime, in which the condition holds by construction and fallback improves both accuracy and selective risk.

C Extended Experimental Results

C.1 Detailed Experimental Setup

Datasets and client partitions. Evaluation proceeds on two tracks. The first inherits the few-shot label-shift protocol of pFedMoAP and covers multiple datasets. Flowers102, OxfordPets, Food101, Caltech101, and DTD are each divided among 10 clients with pairwise disjoint class sets, a pathological non-IID partition in which no class is shared between clients. CIFAR-100 is divided among 100 clients by a Dirichlet label partition with concentration $\alpha = 0.5$. DomainNet and Caltech10 probe simultaneous feature and label shift: each domain is split among five clients with $\alpha = 0.3$, yielding 30 and 20 clients, respectively.

The second track is the seven-dataset base-to-novel protocol of pFedMMA (Ghiasvand et al., 2026), comprising SUN397, OxfordPets, Flowers102, DTD, Caltech101, UCF101, and Food101. For each dataset, the ordered class vocabulary is split evenly into a base half and a novel half, and the base classes are partitioned without overlap across ten clients. Each client is then evaluated on three class populations: its local base classes, base classes assigned to other clients, and novel classes excluded from both prompt training and calibration. The resolved class names and a hash of each split are stored with every run, and both pFedMoAP and pFedMMA are rerun on these frozen manifests.

Training protocol. On the first track, the five CLIP datasets use 16 shots per class, 10 communication rounds, full client participation, and a ResNet-50 (RN50) CLIP backbone; CIFAR-100 uses 120 rounds and 10% participation with the same backbone; and DomainNet and Caltech10 use 25 rounds, participation rates of 25% and 50%, respectively, and a ViT-B/16 backbone. Following pFEDMOAP, all federated methods run five local epochs per round with SGD at learning rate 0.002, and local CoOp trains for 25 epochs. The inherited attention gate uses eight heads, learning rate 0.01, and local-logit coefficient $\lambda_{\mathrm{loc}} = 0.5$. The second track follows the cited 16-shot ViT-B/16 configuration: ten clients, full participation, 50 rounds, two local epochs per round, and learning rate 0.001. The two tracks are configured independently, and results are not comparable across them. Every real-data configuration uses the same three prespecified seeds, with the five exceptions detailed under the artifact inventory below.

pFEDMMA reimplementation and backbone policy. We independently implement pFEDMMA as parallel adapters in blocks 5–12 of the frozen ViT-B/16 image and text transformers. In each block, the 768-dimensional visual or 512-dimensional textual representation is projected to a 32-dimensional bottleneck, passed through a shared 32×32 projection, and projected back to the modality width. The modality-specific down- and up-projections remain client-local, while the shared projections are averaged across participating clients in proportion to their training-set sizes. All runs use the fixed template "a photo of a {class}.", residual scale 0.001, fp32 precision, and a fresh optimizer state for every client update. Because the paired-transformer adapter design requires matched image and text encoders, pFEDMMA uses ViT-B/16 in every run. This induces a backbone asymmetry on part of the first track: the base-to-novel track and the DomainNet and Caltech10 experiments are backbone-matched, whereas the five-dataset open-world track and seven of the multiple closed-set datasets compare ViT-B/16 pFEDMMA against RN50 prompt methods. All mixed-backbone comparisons are reported descriptively and excluded from the confirmatory family defined below.

Calibration and open-world splits. Before any optimization, each client's training partition is divided into a prompt-training subset and a calibration subset. The main study holds out eight calibration examples per client. The calibration subset is used exclusively to fit the scalar output temperature T_i^{out} and the conformal threshold γ_i. All ID, novel-class, and OOD test sets are frozen before model selection. Each OOD manifest contains 1,000 images per ID dataset, selected by automatic RN50 CLIP similarity screening with seed 2027: Flowers102, OxfordPets, Food101, and Caltech101 draw from Places365, and DTD draws from NINCO. Because the screening retains low-similarity examples, this benchmark primarily measures far-OOD behavior and supports no claim about near-OOD or unscreened deployment streams. Experiments that treat another client's distribution as shifted data are labeled "OOD proxy" throughout.

Baselines and uncertainty controls. pFEDMOAP and pFEDMMA are the principal baselines. We run both baselines under the multiple-dataset closed-set and seven-dataset base-to-novel protocols. On the five primary datasets, each baseline yields raw-entropy, post-hoc temperature-scaled entropy, maximum-softmax probability (MSP) (Hendrycks & Gimpel, 2017), and energy (Liu et al., 2020) controls. The runs share client partitions, communication rounds, participation rates, local epochs, seeds, calibration splits, OOD manifests, and test data with pFEDOPEN.

Metrics and inference. The inherited primary endpoint is the mean accuracy evaluated over private client test sets. We evaluate local, base, and novel accuracy, together with their harmonic mean. In addition, the following metrics are used to assess various aspects of pFEDOPEN: (i) calibration quality, assessed via the 15-bin Expected Calibration Error (ECE), the Negative Log-Likelihood (NLL), and the multiclass Brier score; (ii) out-of-distribution (OOD) discrimination, measured by the Area Under the Receiver Operating Characteristic curve (AUROC) and the Area Under the Precision–Recall curve (AUPR); (iii) selective prediction performance, quantified by the Area Under the Risk–Coverage curve (AURC) and the risk at 80% and 90% coverage; and (iv) the realized conformal coverage alongside the fallback or abstention rate. For a pooled test set of size n, with

Table 9: Dataset statistics across the inherited and base-to-novel protocols. "Path." denotes the pathological disjoint-class partition.

Dataset	Train	Test	Classes	Clients	Participation	Heterogeneity
Flowers102	4,093	2,463	102	10	100%	Path.
OxfordPets	2,944	3,669	37	10	100%	Path.
Food101	50,500	30,300	101	10	100%	Path.
Caltech101	4,128	2,465	100	10	100%	Path.
DTD	2,820	1,692	47	10	100%	Path.
SUN397	76,128	21,750	397	10	100%	Path.
UCF101	9,537	3,783	101	10	100%	Path.
Caltech10	2,025	508	10	20	50%	$\mathrm{Dir}(0.3)$
DomainNet	18,278	4,573	10	30	25%	$\mathrm{Dir}(0.3)$
CIFAR100	50,000	10,000	100	100	10%	$\mathrm{Dir}(0.5)$

predicted class probabilities $p_{j,c}$ and corresponding labels y_j, the metrics are defined as follows:

$$\mathrm{ECE} = \sum_{b=1}^{15} \frac{|B_b|}{n} \big|\mathrm{acc}(B_b) - \mathrm{conf}(B_b)\big|,$$

$$\mathrm{NLL} = -\frac{1}{n} \sum_{j=1}^{n} \log p_{j,y_j},$$

$$\mathrm{Brier} = \frac{1}{n} \sum_{j=1}^{n} \sum_{c} \big(p_{j,c} - \mathbf{1}[y_j = c]\big)^2,$$

where B_b are equal-width confidence bins. OOD AUROC treats OOD as the positive class and ranks samples by the uncertainty score $U(x)$. Selective risk at coverage q is the error rate among the $\lceil qn \rceil$ least-uncertain samples, and empirical AURC is the uniform mean of selective risks over the attainable coverages k/n for $k = 1, \ldots, n$. Predictions are pooled for population-level metrics, client-level distributions are retained, and unless stated otherwise, reliability numbers are macro-averaged over client–seed units. The prespecified confirmatory comparison between pFEDOPEN and pFEDMOAP uses paired client-by-seed bootstrap intervals, a paired randomization test, and effect sizes, with Holm correction over the primary comparison family.

Default hyperparameters. The implementation uses anchor floor $\rho = 0.10$, routing-evidence sharpness $\kappa = 10$, and initial temperatures $\tau_r = \tau_a = \tau_0 = 1$ constrained to $[0.05, 10]$. The uncertainty weights are $(\lambda_d, \lambda_r, \lambda_a) = (0.45, 0.20, 0.35)$, the geometry coefficient is $\beta_g = 0.10$, the router-entropy coefficient is $\beta_h = 0.01$ with target $\eta = 0.35$, and the temperature regularization is $\beta_t = 10^{-4}$.

C.2 Inherited dataset statistics

Table 9 records the multiple classification datasets from Luo et al. (2025) and the two additional base-to-novel datasets from Ghiasvand et al. (2026). The calibration and OOD partitions are registered separately so that they do not silently change the inherited test sets.

C.3 Split registration and leakage controls

Prompt and router optimization see only the prompt-training split. Uncertainty weights, regularization coefficients, and global hyperparameters are fixed before test evaluation. The client calibration split then fits only T_i^{out} and γ_i. No test label can change either quantity.

For each of the seven base-to-novel datasets, the manifest stores the ordered class vocabulary, assigns the first half to base and the remainder to novel, and records a content hash. The runner refuses a missing or modified manifest. The main study uses eight calibration examples per client. Flowers102,

OxfordPets, Food101, and Caltech101 pair with CLIP-screened Places365 images; DTD pairs with CLIP-screened NINCO images. Each OOD manifest contains 1,000 images and a content hash. The original client test set remains unchanged; calibration samples come from the training side of the inherited partition.

When an external dataset contains a class with the same semantic label as an ID class, that class is removed according to a rule fixed before prediction scores are inspected. Another client's test distribution is evaluated separately as a shift diagnostic and reported only as proxy AUROC. It is never pooled into the external OOD AUROC or AUPR.

C.4 Baseline adaptation rules

Zero-shot CLIP uses the template "a photo of a {class}." Local CoOp trains one prompt per client without federated communication. PromptFL and its traditional federated adaptations use their published prompt-learning objectives under the matched backbone and communication schedule. pFedPrompt and FedOTP follow their released or published settings when those settings are compatible with the target split. pFedMMA uses the clean-room ViT-B/16 adapter implementation and protocol-matched reruns described in the main experimental setup. pFedMoAP uses the matched multiple-dataset, five-dataset open-world, and seven-dataset base-to-novel reruns reported in this paper.

Temperature scaling fits one scalar on each client's calibration data and does not alter class ranking. MSP uses the negative maximum predicted probability as its uncertainty score. Energy uses the published energy score on unnormalized class logits. All rejection baselines receive the same calibration budget and target coverage as pFedOPEN. For a fair selective comparison, thresholds are chosen on calibration data and evaluated once on the frozen test data. Each open-world pFedMoAP and pFedMMA rerun saves calibration, ID, and OOD logits and derives all four reliability controls from its common predictor. The rerun accuracy tables and reliability metrics are derived only from these completed jobs.

C.5 Statistical analysis

The unit of pairing is a client and seed under the same partition. We report the mean paired difference and a 95% paired bootstrap interval. The primary null hypotheses compare pFedOPEN with pFedMoAP plus temperature scaling for NLL, Brier score, ECE, external OOD AUROC, and AURC under matched client-seed units and training protocols. A paired randomization test accompanies each interval. Holm correction is applied to this primary family. Secondary and ablation tests are labeled exploratory.

Population metrics pool predictions across clients. Client-macro summaries and the 10th percentile are also retained so that a large client cannot conceal poor reliability elsewhere. Every aggregate records the number of clients, seeds, ID examples, OOD examples, and calibration examples.

C.6 Primary paired inference

Table 10 expands the corrected primary tests summarized in the main text. The difference is pFedOPEN with zero-shot fallback minus temperature-scaled pFedMoAP. Positive values therefore favor pFedMoAP for NLL, Brier, ECE, and AURC, but favor pFedOPEN for OOD AUROC.

Every interval excludes zero. The combined pFedOPEN score improves OOD AUROC over temperature-scaled pFedMoAP entropy, but the proper scoring rules and selective ordering favor the baseline.

C.7 Dataset-wise OOD score dependence

The dataset breakdown shows that temperature-scaled entropy is lower than raw entropy for both rerun baselines on every dataset. DTD is the hardest dataset for all pFedOPEN uncertainty components and exposes the largest dependence on score choice. Energy is strongest for pFedMMA on four datasets, while raw entropy is strongest on Caltech101. These patterns explain why a single macro AUROC cannot characterize the reliability interface.

Table 10: Primary paired comparisons over 150 matched client-seed units. Intervals are 95% paired bootstrap intervals with 10,000 resamples and seed 2027. d_p is the paired effect size, p_{raw} is from the paired randomization test, and p_{Holm} is the Holm-adjusted p-value over the five metrics. "pp" denotes percentage points.

Metric	Units	Difference	95% interval	d_p	p_{raw}	p_{Holm}
NLL	150	0.729	$[0.600, 0.864]$	0.88	1.0×10^{-5}	5.0×10^{-5}
Brier	150	0.249	$[0.223, 0.277]$	1.46	1.0×10^{-5}	5.0×10^{-5}
ECE (pp)	150	5.661	$[4.924, 6.429]$	1.19	1.0×10^{-5}	5.0×10^{-5}
OOD AUROC (pp)	150	3.310	$[1.866, 4.768]$	0.36	1.0×10^{-5}	5.0×10^{-5}
AURC (pp)	150	9.606	$[7.526, 11.834]$	0.71	1.0×10^{-5}	5.0×10^{-5}

Table 11: External-OOD AUROC by dataset and score. Entries are client-seed macro-averages over 30 units per dataset. Ent. is predictive entropy, Ent.+T is entropy after temperature scaling, MSP is maximum-softmax uncertainty, and E is energy. pFedMMA uses ViT-B/16 and the other methods use RN50, so its rows are descriptive cross-backbone references.

	pFedOPEN	pFedMoAP				pFedMMA			
Dataset	U	Ent.	Ent.+T	MSP	E	Ent.	Ent.+T	MSP	E
Flowers102	99.67	99.75	95.58	99.31	99.97	99.82	89.40	99.61	**100.00**
OxfordPets	99.74	99.77	89.91	99.24	99.89	99.42	84.31	98.97	**100.00**
Food101	99.16	99.34	92.70	97.98	99.84	99.06	89.26	98.13	**99.96**
Caltech101	95.98	98.49	90.81	97.85	95.03	**99.54**	84.50	99.22	98.57
DTD	75.51	92.01	84.52	89.49	94.38	94.90	83.04	93.34	**98.39**

C.8 Exact sensitivity settings

Tables 12 and 13 replace the ranges in the main paper with the exact evaluated settings. Accuracy, ECE, AURC, coverage, and action rate are client-seed macro-averages. Calibration-size experiments use CIFAR100, and all other rows use Food101. Runtime is the mean end-to-end wall-clock time per run.

The shot sweep gives modest accuracy gains but non-monotone ECE and AURC. Larger CIFAR100 calibration sets hold out more training data and reduce accuracy while leaving realized coverage close to 80%. With only eight calibration scores in the Food101 coverage sweep, target values .90 and .95 accept all test examples and trigger no action. The apparent gains at one, three, and five experts use only two seeds and remain provisional.

Anchor-floor and geometry settings are comparatively flat. The routing-only score is the exception: it raises AURC to 22.02%, compared with 6.62% for within-expert ambiguity. All three temperature intervals yield identical metrics at the shown precision, which is consistent with learned values remaining inside their shared interior rather than being controlled by the tested bounds.

C.9 Seed stability of component ablations

The mean improvements from logit mixing coincide with standard deviations of 1.96 accuracy points and 0.57 AURC points. Removing the anchor floor also has greater seed variation than the full configuration. In contrast, the geometry removal and several uncertainty-weight controls are stable at the shown scale. These values support reporting the ablations as exploratory effects instead of ranked model-selection evidence.

Table 12: Data, calibration, policy, and expert-count sensitivity. "Seeds" is the number of completed runs; Acc., ECE, AURC, Cov., and Action are percentages, and Time is minutes.

Factor	Value	Seeds	Acc.	ECE	AURC	Cov.	Action	Time
Shots	1	3	76.96	6.84	6.92	86.81	13.19	39.5
	2	3	77.44	7.26	6.66	90.84	9.16	41.0
	4	3	77.81	5.92	6.44	88.97	11.03	42.7
	8	3	77.80	6.78	6.45	88.05	11.95	44.7
	16	3	78.41	9.37	6.74	93.20	6.80	49.3
Calibration size	25	3	53.18	13.91	23.54	79.91	20.09	563.4
	50	3	52.97	13.67	23.54	78.91	21.09	525.7
	100	3	51.93	13.61	24.31	79.47	20.53	510.8
	200	3	50.62	12.92	25.98	79.85	20.15	453.9
Target coverage	.70	3	77.52	7.83	6.90	81.16	18.84	50.4
	.80	3	78.41	9.37	6.74	93.20	6.80	50.7
	.90	2	78.79	10.19	6.72	100.00	0.00	49.4
	.95	2	78.79	10.19	6.72	100.00	0.00	49.5
Experts	1	2	87.22	6.89	1.82	90.09	9.91	48.5
	3	2	84.26	7.19	3.75	92.11	7.89	48.4
	5	2	80.82	6.42	5.44	92.31	7.69	48.5
	10	3	78.41	9.37	6.74	93.20	6.80	50.7

Table 13: Mechanism sensitivity on Food101. Notation and units follow Table 12. Single-component rows replace the combined uncertainty with the listed score.

Factor	Value	Seeds	Acc.	ECE	AURC	Cov.	Action	Time
Anchor floor ρ	0	3	79.01	10.25	6.54	92.14	7.86	51.2
	.05	3	78.20	8.69	6.78	91.98	8.02	51.7
	.10	3	78.41	9.37	6.74	93.20	6.80	50.8
	.20	3	78.10	9.04	6.65	91.70	8.30	50.7
	.30	3	78.05	8.00	6.56	92.01	7.99	51.0
Geometry β_g	0	3	78.63	9.44	6.70	92.93	7.07	51.2
	.01	3	78.31	9.58	6.86	92.03	7.97	51.0
	.05	3	78.38	8.86	6.96	92.19	7.81	50.9
	.10	3	78.41	9.37	6.74	93.20	6.80	50.8
	.20	3	78.48	9.28	6.66	92.52	7.48	50.9
Uncertainty score	D	3	78.27	9.53	7.85	92.04	7.96	50.9
	R	3	78.34	9.51	22.02	88.50	11.50	50.8
	A	3	78.38	9.19	6.62	92.67	7.33	50.5
Temperature bounds	$[.05, 5]$	3	78.41	9.37	6.74	93.20	6.80	50.7
	$[.05, 10]$	3	78.41	9.37	6.74	93.20	6.80	49.9
	$[.1, 10]$	3	78.41	9.37	6.74	93.20	6.80	50.4

Table 14: Food101 ablation stability. Values are mean ± sample standard deviation over three seed-level client averages. All metrics are percentages.

Variant	Acc. ↑	ECE ↓	AURC ↓
Full pFEDOPEN	78.78±0.40	9.83±2.32	6.86±0.20
No zero-shot expert	78.63±0.24	**8.58±0.90**	6.81±0.40
$\rho = 0$	79.89±2.17	11.73±2.66	6.46±0.88
Mean-prototype router	79.04±0.26	9.31±1.90	6.70±0.62
Logit mixing	**81.56±1.96**	8.72±1.80	**4.99±0.57**
No geometry loss	79.35±0.17	10.61±1.29	6.64±0.10
No router entropy floor	78.96±0.32	10.38±2.36	6.70±0.04
Fixed expert temperatures	78.99±0.59	9.40±0.95	6.72±0.30
No output temperature	78.96±0.32	12.76±2.83	6.70±0.04
$\lambda_d = 0$	78.96±0.32	10.38±2.36	6.58±0.14
$\lambda_r = 0$	78.96±0.32	10.38±2.36	6.70±0.04
$\lambda_a = 0$	78.96±0.32	10.38±2.36	7.79±0.13
Fixed threshold	78.96±0.32	10.38±2.36	6.70±0.04

C.10 DOMAIN-TRANSITION CALIBRATION SHIFT

Table 15: Calibration under each registered feature-shift transition. Every row contains 15 client-seed units. Realized coverage, gap from the 80% target, ECE, and AURC are percentages.

Dataset	Source → target	Units	Realized	Gap	ECE	AURC
DomainNet	Clipart → Infograph	15	42.10	−37.90	34.55	90.28
	Infograph → Painting	15	62.02	−17.98	36.65	94.74
	Painting → Quickdraw	15	70.93	−9.07	21.35	89.08
	Quickdraw → Real	15	11.45	−68.55	48.31	89.91
	Real → Sketch	15	84.62	+4.62	19.41	86.99
	Sketch → Clipart	15	84.03	+4.03	31.67	93.21
Caltech10	Amazon → Caltech	15	87.30	+7.30	20.63	88.70
	Caltech → DSLR	15	83.67	+3.67	30.22	88.77
	DSLR → Webcam	15	96.05	+16.05	21.74	86.25
	Webcam → Amazon	15	81.74	+1.74	23.17	90.67

The macro DomainNet gap hides qualitatively different transitions. Quickdraw to Real under-covers by 68.55 points, while Real to Sketch and Sketch to Clipart slightly over-cover. All four Caltech10 transitions over-cover, but the gap ranges from 1.74 to 16.05 points. The local calibration result therefore does not transfer uniformly even within one benchmark.

C.11 OOD MANIFEST SCREENING AUDIT

Table 16: Automatic OOD-manifest construction with RN50 CLIP similarity, screening seed 2027. "Review" records whether the selected images received manual semantic review. Full content hashes remain in the saved manifests.

ID dataset	OOD source	Candidates	Selected	Selected score range	Review
Caltech101	Places365	1,600	1,000	[0.150, 0.227]	No
DTD	NINCO	5,879	1,000	[0.139, 0.235]	No
Food101	Places365	1,600	1,000	[0.134, 0.198]	No
Flowers102	Places365	1,200	1,000	[0.140, 0.200]	No
OxfordPets	Places365	1,600	1,000	[0.134, 0.213]	No

C.12 EFFICIENCY ANALYSIS

Computational and communication cost. Table 17 summarizes the resource profile of the Food101 main runs. The method is lightweight on all three axes. The trainable footprint is roughly 41K

Table 17: Efficiency of the three Food101 main runs. Storage and communication quantities are deterministic per configuration. Runtime values report mean and standard deviation over the three runs on the execution system used for the sweep.

Category	Quantity	Measurement
Model size and state	Trainable parameters	40,963
	Persistent gating state	128.02 KiB per client
Communication	Prompt payload	32 KiB per selected client
	Traffic per round	320 KiB up / 3.125 MiB down
Wall-clock time	Client optimization	946.94 ± 21.82 s
	Server aggregation	0.00845 ± 0.00004 s
	Calibration	5.06 ± 0.29 s
	Open-world evaluation	713.98 ± 1.46 s
	End-to-end	3176.95 ± 28.50 s

Table 18: Positive-control study (mean over 20 fixed seeds). Rows 3–4 add components cumulatively to the diagnostic *without* output temperature, while the full configuration combines all components. This is *not* a federated CLIP result. Lower is better for ECE/NLL/Brier/AURC; higher is better otherwise.

Method	Acc.	Novel Acc.	ECE	NLL	Brier	OOD AUROC	AURC	Fallback
pFedMoAP diagnostic	.9846	.9615	.2431	.3682	.1401	.7532	.00137	.000
+ output temperature	.9846	.9615	**.0036**	**.0450**	**.0234**	.7532	.00137	.000
+ immutable zero-shot	.9911	.9778	.2761	.3914	.1425	**.8191**	.00081	.000
+ conformal fallback	.9974	.9935	.2577	.3379	.1131	**.8191**	.00042	.204
pFedOPEN (full)	**.9975**	**.9937**	.0987	.1489	.0620	.8190	**.00040**	.204

parameters, and per-round communication is dominated by the server broadcast (3.125 MiB download versus 320 KiB upload). On the time axis, the cost structure is sharply asymmetric: server aggregation completes in under 10 ms per round, five orders of magnitude below client optimization, so the server is never a bottleneck and the protocol scales with the number of clients essentially at the cost of communication alone. The uncertainty machinery itself is nearly free at deployment (because calibration adds 5.06 s, about 0.16% of the end-to-end time) whereas open-world evaluation accounts for roughly twelve minutes (22% of the total). The latter is an evaluation-protocol cost incurred once per run, not a recurring training cost. We emphasize that these figures describe the present implementation on a single execution system and are intended to establish feasibility; they are not normalized throughput comparisons against other methods, which would require matched hardware and implementation effort.

C.13 Positive-control mechanism study

The simulator serves as a positive control for the implementation: under a data-generating process constructed to satisfy the intended mechanism assumptions, each component should recover its predicted directional effect. The check therefore validates the correctness of the implementation and the metric pipeline. The synthetic model uses ten classes, five adaptive experts, 40% novel samples, correlated prompt drift, and OOD experts that are confident but mutually conflicting. Per seed, $1{,}000$ ID samples fit the output temperature and the conformal threshold, and $6{,}000$ ID and $3{,}000$ OOD samples are held out for evaluation. One substitution is required: because the simulator does not optimize CLIP prompts, geometry regularization is represented by a prespecified reduction in correlated prompt drift rather than by the loss itself.

Each component reproduces its predicted signature. Output temperature improves the proper scoring rules by an order of magnitude (ECE .2431 $\rightarrow$ .0036; NLL and Brier similarly) while leaving accuracy, OOD AUROC, and AURC exactly unchanged, which is the expected behavior of a monotone transformation of the output distribution. The immutable zero-shot anchor improves novel accuracy (.9615 $\rightarrow$ .9778) and OOD ranking (AUROC .7532 $\rightarrow$ .8191) but worsens ECE, as the anchor injects confident probability mass that the uncalibrated mixture cannot absorb. Conformal fallback

fires on 20.4% of samples and improves both accuracy and selective risk (AURC $.00081 \rightarrow .00042$). This gain is expected *by construction*, since the simulated zero-shot expert is designed to satisfy the accuracy condition of Proposition 4. The full configuration recovers most of the temperature calibration benefit (ECE .0987 versus .2577 without it) while retaining the anchor and fallback gains, although no row dominates on every metric: temperature alone remains best on the proper scoring rules, so even under ideal assumptions the components trade off instead of composing additively.

The contrast with the real benchmark delimits what this control establishes. On real data the fallback result reverses, because the actual zero-shot expert is *less* accurate than the mixture on flagged samples and Proposition 4's condition fails. This is the informative outcome of a positive control: agreement under satisfied assumptions confirms that the implementation and metric pipeline behave as designed, while disagreement under violated assumptions confirms that the mechanism's benefit is conditional on those assumptions.

D Implementation and Complexity

Computational complexity. Let $E = K_i + 1$ denote the number of experts at client i (including the immutable zero-shot expert), C the number of classes, d the feature dimension, and B the batch size. Computing expert logits costs $O(BECd)$ per batch. This term is not overhead introduced by our method, since any mixture that exposes per-expert predictions must already incur it. The additional components are cheap by comparison: routing reuses the same class structure in a reduced dimension and is dominated by the logit computation, and geometry matching costs $O(C^2 d)$ per round for the local prompt, which is negligible for the vocabularies considered here ($C \leq 101$). For very large vocabularies the quadratic class dependence would dominate. In that regime we prescribe a class-pair subsample that is pre-registered and reported alongside results, so that the approximation cannot be tuned post hoc. The per-round cost is thus a small constant factor over the base mixture, consistent with the measured runtimes in Table 17.

Metric module and verification. Because several of our reported quantities are nonstandard or easy to implement incorrectly (tie-aware AUROC, empirical AURC, the entropy decomposition identity, conformal thresholds), all evaluation metrics are computed by a single dependency-light module rather than assembled from heterogeneous library calls. The module implements ECE, NLL, Brier score, tie-aware AUROC, empirical AURC, fixed-coverage risk, temperature fitting, the entropy identity of Proposition 1, and conformal threshold fitting. Eleven unit tests verify probability normalization, the disagreement term, exactness of the entropy decomposition, calibration computations, AUROC tie handling, selective ordering, and conformal edge cases (empty and degenerate calibration sets). The same module is used by the training pipeline, the evaluation pipeline, and the simulator of Section C.13, so the positive control exercises the identical metric code path used for the benchmark results. The checked-in simulator and its summary reproduce Table 18 exactly from fixed seeds.