TRUST WITHOUT AGGREGATION: ROBUST EXPERT POOLS FOR PERSONALIZED FEDERATED PROMPT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalized federated prompt learning can preserve client heterogeneity by retaining uploaded prompts as separately retrievable experts instead of averaging them into a single model. However, this design worsens robustness: an expert pool delivers each prompt intact to its recipients, so a malicious prompt that remains geometrically close to benign ones can pass selection while encoding targeted behavior. We formulate this setting as *robust expert-pool learning* and propose PFEDREP, a staged trust-governance mechanism that matches each defense to the evidence available at its stage, i.e., neighborhood-relative screening at cold start, recipient-directed counterfactual utility and distinct-reporter reputation as interaction evidence accumulates, diversity-aware retrieval, and prompt-independent capped routing that bounds reliance on any selected expert. Under fixed retrieval and ordering, we prove that replacing one normalized non-local expert shifts each class logit by at most a cap-controlled constant, yielding a margin condition for top-1 stability. Empirically, across multiple benchmarks, the trust governance of PFEDREP does not sacrifice clean accuracy under heterogeneous settings. More importantly, under the primary attack, clean and worst-decile accuracy change by only 0.11 and 0.08 percentage points. In addition, a 20-client study isolates the individual trust mechanisms: relative to geometric screening alone, the full method reduces targeted exposure from 65.3% to 1.3%; and relative to nearest-prompt retrieval, it reduces severe negative transfer from 17.3% to 0.6%. These results support staged trust and bounded reliance as complementary principles for governing unaggregated expert pools. The code will be open-sourced.

1 INTRODUCTION

Prompt learning (PL) adapts frozen vision–language models (VLMs) such as CLIP (Radford et al., 2021; Jia et al., 2021) through a small set of trainable parameters (Zhou et al., 2022; Khattak et al., 2023a;b). Its compactness enables a distinctive form of federated prompt learning (FPL) (Luo et al., 2025; Wang et al., 2025; Feng et al., 2023; Guo et al., 2023a; Li et al., 2024; Guo et al., 2023b). Instead of averaging all prompts into one model, pFedMoAP (Luo et al., 2025) lets the server retain them as separate experts and retrieve a client-specific subset at test time. This mixture-of-prompt-experts design supports personalization without collapsing heterogeneous client knowledge.

Retaining unaggregated experts necessitates a robustness governance mechanism. Traditional aggregation distills potentially corrupted submissions into a single shared update, whereas an expert-pool server preserves each submission and may deliver it intact to a recipient client. A malicious prompt can therefore remain close enough to benign prompts in parameter space to pass selection while still encoding targeted behavior such as backdoors (Cai et al., 2022; Bagdasaryan et al., 2020). Robust expert pools consequently require defenses that govern the membership, selection, and influence of identifiable experts, rather than only the construction of an aggregate.

Moreover, the evidence needed for such governance arrives in stages, and each stage supports different decisions. Geometric properties of a prompt are observable at upload but can not predict its behavior on a recipient's data. Behavioral utility becomes measurable only after interaction, and population-level feedback can inform cold-start decisions only when reporter identities and report

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.

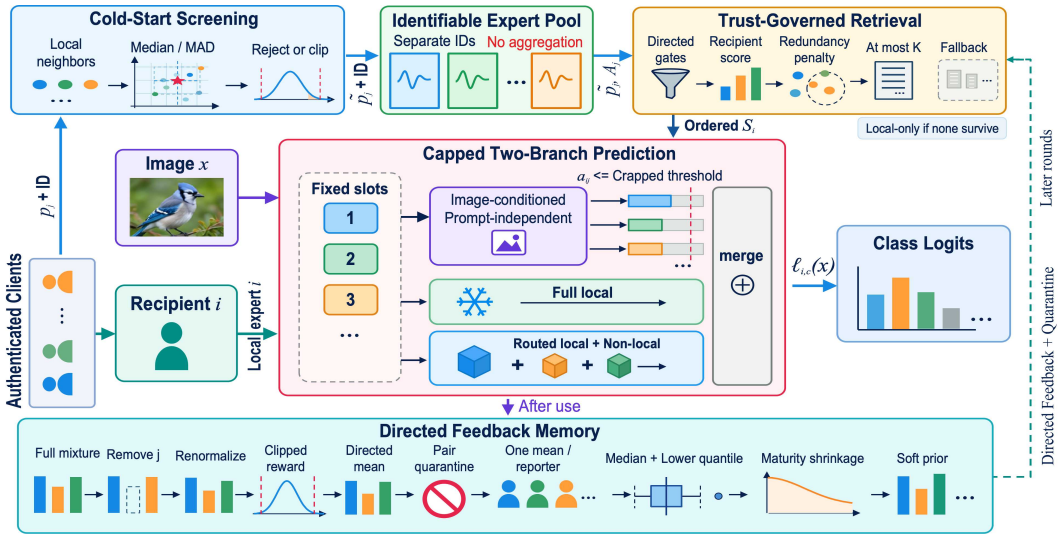

Figure 1: Overview of PFEDREP.

contamination are handled explicitly. Because no stage renders expert selection infallible, robust expert-pool learning must maintain *directed* trust that is updated as evidence accumulates, while bounding reliance on any expert whose trustworthiness remains uncertain.

To bridge these gaps, we propose PFEDREP, a trust-governance mechanism for unaggregated pools of identifiable prompts. Specifically, at cold start, when only geometry is observable, neighborhood-relative robust statistics screen outliers without forcing heterogeneous prompts toward a single global center. As interaction yields behavioral evidence, each client computes leave-one-expert-out loss differences to update directed utility and to quarantine repeatedly harmful recipient–expert pairs. The server then converts these local judgments into a population-level signal. Specifically, the server maintains for every expert a reputation score defined over the set of *distinct* client reporters, where each client that repeatedly flags the same expert as harmful contributes a single negative report, so that no individual client (including honest or malicious) can amplify its influence on an expert's reputation through repetition. This aggregation lets a recipient client benefit from harm observed elsewhere before integrating it locally. Retrieval then scores each candidate expert by combining four complementary signals: geometric affinity between the candidate and the querying client's prompt, the recipient's own directed utility estimate for that expert, the expert's population-level reputation, and a diversity penalty that discourages selecting near-duplicate experts so that the retrieved set covers complementary knowledge instead of concentrating on one region of the pool. Finally, a prompt-independent capped router bounds the routing weight assigned to any selected expert. Together, these stages form a fallback chain: geometric screening controls cold-start risk, behavioral evidence refines recipient-specific trust, and capped routing limits the consequence of any selection error. Figure 1 summarizes the pipeline.

Analytically, under fixed retrieval and ordering, the capped router bounds the class-logit change caused by replacing one normalized non-local expert and yields a margin condition for top-1 stability. Empirically, multiple clean benchmarks measure accuracy under heterogeneous data, federated attack experiments jointly evaluate prediction and pool-level behavior, and a controlled study isolates targeted exposure, negative transfer, cold-start behavior, and retrieval redundancy.

Our contributions are summarized as follows:

1. **Problem formulation.** We formulate *robust expert-pool learning*, where a server retains identifiable prompts rather than aggregating them, and show that this setting demands defenses over expert screening, selection, and influence instead of over a single aggregation.
2. **Staged trust governance.** We propose PFEDREP, which matches each defense to the evidence available at its stage: neighborhood-relative screening when only geometry is observable, directed counterfactual utility and distinct-reporter reputation as behavioral evidence accumulates, and diversity-aware retrieval with capped routing at deployment. The stages form a fallback chain that preserves heterogeneous experts while bounding reliance on any one of them.

3. **Theoretical guarantee.** We prove that, under fixed retrieval and ordering, substituting one normalized non-local expert shifts each class logit by at most a cap-controlled constant, and we derive a margin condition under which top-1 predictions are provably unchanged.
4. **Empirical validation.** Under federated attacks, pFEDREP keeps targeted attack success near the unattacked control at negligible cost to clean accuracy. In addition, a controlled study isolates each trust mechanism, showing that the full method reduces targeted exposure from 65.3% to 1.3% relative to geometric screening alone.

2 PROBLEM FORMULATION

We consider N authenticated clients coordinated by a central server. At round t, client j maintains a local prompt $p_j^t \in \mathbb{R}^d$ and may upload it to the expert pool $\mathcal{P}^t$. A recipient i requests at most K non-local experts, indexed by $S_i^t \subseteq [N] \setminus \{i\}$. The server returns their identified and sanitized prompts without averaging, so that every downloaded prompt remains an identifiable contributor to the recipient's prediction.

For image x and class c, the prediction combines a private local branch with a routed branch over the local and selected non-local experts. Let $z_{\text{full}}(x)$ and $v_{i,c}^{\text{full}}$ be the full-dimensional image and class-text representations retained by client i, and let $z(x)$ be the pooled image representation and $v_{j,c}$ the class-text representation induced by expert j. These representations are unit-normalized. With CLIP logit scale $s > 0$ and local-branch coefficient $\lambda \geq 0$, the class logit is

$$\ell_{i,c}(x) = \underbrace{\lambda s \langle z_{\text{full}}(x), v_{i,c}^{\text{full}} \rangle}_{\text{local branch}} + \underbrace{s \langle z(x), \sum\nolimits_{j \in \{i\} \cup S_i^t} a_{i,j}(x)\, v_{j,c} \rangle}_{\text{routed expert branch}}, \tag{1}$$

where $a_{i,j}(x) \geq 0$ and $\sum_{j \in \{i\} \cup S_i^t} a_{i,j}(x) = 1$. The first term is independent of routed content, and the second enables cross-client transfer but exposes the recipient to selected experts. Including expert i in the routed mixture permits local fallback. We assume an honest server, authenticated client identities, and an uncompromised frozen backbone. Appendix B specifies the adversary and separates these system assumptions from the narrower conditions of our influence analysis.

3 METHOD

Overview. pFEDREP follows the order in which evidence becomes available. It uses geometry for cold-start upload screening, recipient-directed utility and distinct-reporter reputation as interaction evidence accumulates, and a prompt-independent cap after retrieval. This separates trust estimation from exposure control: selection uses the strongest available evidence, while routing limits reliance on any selected expert.

3.1 NEIGHBORHOOD-RELATIVE GEOMETRIC SCREENING

A cold-start signal must operate before behavioral evidence exists without treating domain heterogeneity as malicious deviation. We therefore score each prompt relative to its neighborhood. Let $\mathcal{N}_j$ contain p_j and its h nearest prompts in flattened parameter space, define the coordinate-wise median $m_j = \text{med}\{p_k : k \in \mathcal{N}_j\}$, and set $d_j = \|p_j - m_j\|_2$. Across available prompts, we compute

$$\widetilde{d} = \text{med}_j\, d_j, \qquad \sigma_d = 1.4826\, \text{med}_j\, |d_j - \widetilde{d}| + \epsilon, \qquad A_j = \max\left\{0, \frac{d_j - \widetilde{d}}{\sigma_d}\right\}, \tag{2}$$

where $1.4826 \approx 1/\Phi^{-1}(3/4)$ is the Gaussian consistency factor for the median absolute deviation and $\epsilon > 0$ prevents a zero denominator. The server rejects $A_j > \tau_A$ and otherwise retains

$$\widetilde{p}_j = m_j + (p_j - m_j) \min\left\{1, \frac{\widetilde{d} + \tau_C \sigma_d}{\|p_j - m_j\|_2}\right\}, \tag{3}$$

where $\tau_C \leq \tau_A$ is a clipping radius on the same robust scale. Prompts inside the radius pass unchanged; more distant accepted prompts are shrunk toward m_j. Neighborhood-relative screening accommodates separated honest domains, while later behavioral evidence handles malicious prompts that remain geometrically plausible.

3.2 DIRECTED COUNTERFACTUAL UTILITY AND QUARANTINE

Geometry can not determine whether an expert benefits a particular recipient, and mixture loss can not assign credit to individual members. We therefore compare the observed mixture with a leave-one-expert-out counterfactual. For x in a training batch B_i^t, let $r_{j,c}(x) = s\langle z(x), v_{j,c}\rangle$ and

$$r_c(x) = \sum\nolimits_{e\in\{i\}\cup S_i^t} a_{i,e}(x)\, r_{e,c}(x)$$

be the routed-branch logit. Removing selected expert $j \in S_i^t$ and renormalizing the remaining weights gives

$$r_c^{-j}(x) = \frac{r_c(x) - a_{i,j}(x)\, r_{j,c}(x)}{1 - a_{i,j}(x)}, \tag{4}$$

where capped routing keeps $1 - a_{i,j}(x)$ bounded away from zero. Writing $r(x) = \big(r_c(x)\big)_c$ and $r^{-j}(x) = \big(r_c^{-j}(x)\big)_c$, the client releases

$$\widehat{u}_{ij}^t = \mathrm{clip}\Big(\mathcal{L}_{B_i^t}\big(\lambda\ell_i^{\mathrm{local}} + r^{-j}\big) - \mathcal{L}_{B_i^t}\big(\lambda\ell_i^{\mathrm{local}} + r\big), [-R, R]\Big), \tag{5}$$

where ℓ_i^{local} collects the local logits of Eq. 1. A positive reward means that retaining j lowered recipient i's batch loss. The running mean is

$$u_{ij}^t = \frac{1}{|\mathcal{T}_{ij}^t|} \sum\nolimits_{\tau\in\mathcal{T}_{ij}^t} \widehat{u}_{ij}^\tau,$$

where $\mathcal{T}_{ij}^t$ contains rounds in which j was routed to i. Both $\widehat{u}_{ij}^t$ and u_{ij}^t are directed, preserving recipient-specific transfer that a global score would erase.

To stop repeated harmful transfer, client i quarantines j for T_q rounds when either its last P rewards all fall below τ_u or its running mean remains below τ_u after at least P observations. The two triggers capture consecutive and persistent harm, respectively. Quarantine remains directed, suppressing an expert for the affected recipient without imposing a global veto.

3.3 DISTINCT-REPORTER REPUTATION

Directed utility is sparse for experts that a recipient has not used. To transfer behavioral evidence without letting frequent reports dominate, each client that has used j contributes one running mean u_{ij}^t. Let

$$\mathcal{U}_j^t = \big\{u_{ij}^t : \mathcal{T}_{ij}^t \neq \emptyset\big\}$$

collect one value per distinct reporter. We further define

$$\rho_j^t = \frac{M_j^t}{M_j^t + \kappa}\Big[(1-\gamma)\,\mathrm{med}\big(\mathcal{U}_j^t\big) + \gamma\, Q_q\big(\mathcal{U}_j^t\big)\Big], \tag{6}$$

where $M_j^t = |\mathcal{U}_j^t|$ and Q_q denotes an empirical quantile. The maturity factor $M_j^t/(M_j^t + \kappa)$ shrinks sparse reputations toward zero. Repeated reports update one reporter's mean but do not increase M_j^t, preventing event-count inflation under the authenticated-identity assumption.

The median captures broad transfer quality, while Q_q emphasizes poor-transfer reports. Their combination is a soft ranking prior rather than a default veto: population statistics alone can not distinguish a harmful expert from a legitimate specialist, whereas directed evidence can preserve specialists for matching recipients.

Proposition 1 (Bounded influence of Byzantine (malicious) reporters)**.** *Fix an expert j and round t, and suppose an adversary controls at most $\lfloor\beta M_j^t\rfloor$ of the M_j^t distinct reporter identities and may set their reported means in $\mathcal{U}_j^t$ to arbitrary values. Let $\mathcal{H}_j^t \subseteq \mathcal{U}_j^t$ denote the reports of honest reporters. Then for every q with $\beta < q < 1 - \beta$, the empirical q-quantile of the full report set is bracketed by quantiles of the honest reports:*

$$Q_{\frac{q-\beta}{1-\beta}}\big(\mathcal{H}_j^t\big) \;\le\; Q_q\big(\mathcal{U}_j^t\big) \;\le\; Q_{\frac{q}{1-\beta}}\big(\mathcal{H}_j^t\big). \tag{7}$$

Corollary 1 (Reputation localization)**.** *Under Proposition 1 with $\beta < q \le \frac{1}{2}$, both the median and lower-tail terms in Eq. 6 remain within quantile ranges determined by honest reporters. Consequently, arbitrary Byzantine report values can not move ρ_j^t outside the corresponding maturity-scaled combination of those ranges.*

The explicit interval and proofs are given in Appendix C.

3.4 SAFETY SCORING AND DIVERSITY-AWARE RETRIEVAL

Retrieval combines evidence instead of using any single score. Candidates first pass anomaly screening, the directed-utility test, and recipient-specific quarantine. Notably, reputation remains a soft ranking prior unless a conservative override is enabled. Each surviving expert candidate receives

$$B_{ij}^t = w_u u_{ij}^t + w_r \rho_j^t + w_s \cos(p_i, \widetilde{p}_j) - w_A \min\{A_j, \tau_A\}. \tag{8}$$

By Corollary 1, malicious reporters can move ρ_j^t only within honest quantile ranges. For a recipient without directed evidence, the shrunk reputation in Eq. 6 supplies a population prior.

Starting from an empty set S_i^t, the server greedily adds the expert candidate

$$j^* = \arg\max_{j \notin S_i^t} \left[B_{ij}^t - w_D \max_{k \in S_i^t} \max\{0, \cos(\widetilde{p}_j, \widetilde{p}_k)\} \right]. \tag{9}$$

The maximum-redundancy penalty prevents a tight cluster of near-duplicate prompts from filling S_i^t and leaves room for heterogeneous experts. It is a retrieval prior instead of a standalone guarantee. If all candidates fail the directed gates, the server returns no non-local prompt and the recipient falls back to its local expert. Complexity and implementation details are given in Appendix F.

3.5 PROMPT-INDEPENDENT CAPPED ROUTING

Retrieval can still admit a harmful expert, so routing independently limits the consequence of selection errors. A content-dependent gate would let a prompt influence both its representation and its mixture weight. Instead, pFEDREP maps the image representation to logits over $E = |S_i^t| + 1$ fixed rank slots: slot 0 holds the local expert and the remaining slots follow retrieval order. Thus a slot's weight is independent of the prompt occupying it. Let $b_i(x) = \mathrm{softmax}(g(z(x))) \in \Delta^{E-1} = \{w \in \mathbb{R}^E : w_j \geq 0,\ \sum_{j=1}^E w_j = 1\}$. Because a per-slot cap below $1/E$ is infeasible, we use the effective cap $\bar{\alpha} = \max\{\alpha, 1/E\}$ for a target $\alpha \in (0, 1]$ and define

$$\eta = \frac{1 - \bar{\alpha}}{1 - 1/E}, \qquad a_i(x) = (1 - \eta)\, b_i(x) + \eta\, \mathbf{1}/E. \tag{10}$$

For $E > 1$, Eq. 10 tightly satisfies $a_{i,j}(x) \leq \bar{\alpha}$ for every slot. For $E = 1$, routing is local and trivial.

Theorem 1 (Conditional single-expert logit influence)**.** *Fix the image x, the retrieved set S_i^t and its ordering, and the gate parameters. Assume $\|z(x)\|_2 \leq 1$ and $\|v_{j,c}\|_2, \|v'_{j,c}\|_2 \leq 1$ for all classes c. Replacing one non-local expert j by any normalized expert j' changes every class logit by at most*

$$\left|\ell'_{i,c}(x) - \ell_{i,c}(x)\right| \leq 2s\bar{\alpha}. \tag{11}$$

The full-dimensional local term in Eq. 1 is unchanged.

Corollary 2 (Conditional top-1 invariance)**.** *Let $\widehat{c} = \arg\max_c \ell_{i,c}(x)$ and $m(x) = \ell_{i,\widehat{c}}(x) - \max_{c \neq \widehat{c}} \ell_{i,c}(x)$. Under the conditions of Theorem 1, if $m(x) > 4s\bar{\alpha}$, then replacing one non-local expert can not change the predicted class.*

Proofs are in Appendix C. The guarantee conditions on fixed retrieval, ordering, and gate parameters, and covers replacement of one non-local expert. Retrieval and routing are therefore complementary: directed trust governs which experts enter the set, while capped routing bounds the influence of a selected slot.

4 EXPERIMENTS

Our evaluation asks three research questions (RQs): **RQ1:** whether trust governance of pFEDREP imposes a clean-accuracy sacrifice, **RQ2:** how predictions and pool behavior respond to malicious experts (*this is our primary focus*), and **RQ3:** which mechanisms account for the observed behavior. We present three-seed results across multiple datasets, primary attacks with matched controls, and complementary real-data and controlled ablations. Prior work does not report the pool-level attack metrics used here, so that comparisons under attack use matched pFEDREP controls while clean comparisons retain the published baselines.

Table 1: Almost no loss of few-shot client clean accuracy (%), sharing settings with pFedMoAP.

Method	Flowers102	OxfordPets	Food101	Caltech101	DTD
ZS-CLIP	62.17±0.12	84.47±0.01	75.27±0.05	85.14±0.24	40.21±0.12
CoOp	70.14±0.76	83.21±1.30	70.43±2.42	87.37±0.44	44.23±0.63
PromptFL	72.80±1.14	90.79±0.61	77.31±1.64	89.70±1.99	54.11±0.22
PromptFL + FT	72.31±0.91	91.23±0.50	77.16±1.56	89.70±0.25	53.74±1.36
Prompt + FedPer	72.11±1.35	89.50±1.62	71.29±1.87	86.72±1.45	50.23±0.82
Prompt + FedProx	66.40±0.29	89.24±0.41	76.24±1.94	89.41±0.55	44.26±1.11
Prompt + FedAMP	69.10±0.13	80.21±0.44	74.48±1.71	87.31±1.60	47.16±0.92
pFedPrompt	86.46±0.15	91.84±0.41	92.26±1.34	96.54±1.31	77.14±0.09
FedOTP	96.23±0.44	98.82±0.11	92.73±0.15	97.02±0.36	87.64±0.70
pFedMoAP ($\lambda = 0$)	97.27±0.45	94.01±0.27	86.47±0.57	94.98±0.59	85.47±0.42
pFedMoAP ($\lambda = 0.5$)	98.48±0.51	98.88±0.15	92.76±0.22	97.68±0.29	90.56±0.27
pFedMMA (ViT-B/16)	99.25±0.06	99.29±0.09	96.65±0.04	99.03±0.16	92.66±0.50
pFedMMA (RN50)	98.56±0.19	98.12±0.65	93.20±0.05	97.89±0.34	90.87±0.51
pFedREP (ViT-B/16)	99.48±0.01	98.37±1.29	96.64±0.13	99.02±0.18	92.68±0.51
pFedREP (RN50)	98.78±0.26	98.42±0.35	93.18±0.07	97.61±0.09	91.08±0.24
Δ=pFedREP-Baseline	0.23	-0.92	-0.01	-0.01	0.02

4.1 Experimental setup

Clean protocol. We follow the pFedMoAP protocol (Luo et al., 2025): five 10-client few-shot datasets with disjoint labels. Baselines include zero-shot CLIP (Radford et al., 2021), CoOp (Zhou et al., 2022), PromptFL (Guo et al., 2023b), pFedPrompt (Guo et al., 2023a), FedOTP (Li et al., 2024), and pFedMMA (Ghiasvand et al., 2026). In addition, baselines also include traditional federated techniques on prompt learning, such as FT (Cheng et al., 2021), FedProx (Li et al., 2020), FedPer (Arivazhagan et al., 2019), and FedAMP (Huang et al., 2021). We rerun pFedMoAP at $\lambda \in \{0, 0.5\}$ and evaluate pFedMMA and pFedREP with RN50 and ViT-B/16. Our experiments use seeds 1–3 and five local SGD epochs at learning rate 0.002. In particular, for pFedREP, we employ ten expert slots, $\lambda = 0.5$, and cap 0.35. Experiment results are generally presented as mean $\pm$ sample standard deviation over three seeds. Full schedules and implementation details are in Appendix D.

Attack and metrics. The primary setting controls 20% of clients throughout training, poisons 20% of their examples with a 4×4 bottom-right patch relabeled to class 0, and uses proximity-constrained uploads. Matched variants change the contamination, upload, poison, trigger, target, or duration. We present clean accuracy (CA), worst-decile accuracy (W10), targeted attack success rate (ASR) on triggered non-target images, malicious-expert exposure (MExp), beneficial honest-expert rejection (BRej), local-only abstention (Abst.), and exposure before quarantine (Pre-Q). MExp is the fraction of downloaded non-local slots owned by malicious identities. Results are mean $\pm$ standard deviation.

4.2 RQ1: Clean accuracy evaluations

Table 1 evaluates whether trust governance imposes a clean-accuracy loss, since screening, quarantine, and capped routing could each penalize honest and useful experts. Fortunately, it does not. pFedREP is statistically indistinguishable from the strongest baseline on four of five datasets (differences of at most 0.23 points, within-seed variance). Two design choices explain this near-zero cost. First, because the few-shot protocol assigns disjoint labels, honest prompts legitimately occupy separated regions of parameter space. Meanwhile, the neighborhood-relative statistics of Eq. 2 score each prompt against its local cluster instead of a global center and thus avoid rejecting or over-clipping benign heterogeneous prompts. Second, with $\lambda = 0.5$, a substantial share of each logit flows through the local full-dimensional branch, which is unaffected by screening, retrieval, or the routing cap, so governing the expert pool perturbs only the incremental routed branch.

Table 2 confirms that pFedREP's macro-level parity extends to the weakest clients. Against the best baseline, the W10 deltas are $+0.15$ (Flowers102), $+0.12$ (Food101), and -0.20 (Caltech101), together with a clear $+2.22$-point gain on DTD. The lone negative of -6.09 on OxfordPets is driven by the ± 8.11 standard deviation, where the same instability is visible at the macro level in Table 1. Overall, trust governance almost does not cause degradation for tail clients as well.

Table 2: Worst-decile client accuracy (W10, %) on the five few-shot datasets.

Method	Flowers102	OxfordPets	Food101	Caltech101	DTD
pFedMoAP ($\lambda = 0$)	93.83±0.74	82.12±1.26	80.77±2.56	88.74±2.46	73.15±1.79
pFedMoAP ($\lambda = 0.5$)	94.30±1.50	93.20±0.91	88.74±0.41	92.64±0.37	80.00±1.11
pFedMMA (ViT-B/16)	97.50±0.23	96.39±0.77	94.34±0.19	96.58±1.54	85.00±3.09
pFedMMA (RN50)	94.80±0.64	92.08±2.27	88.33±0.35	93.07±0.37	81.30±0.64
pFedREP (ViT-B/16)	97.65±0.57	90.30±8.11	94.46±0.05	96.38±0.98	87.22±2.42
pFedREP (RN50)	96.28±0.00	91.73±2.22	89.18±0.41	92.42±0.37	84.07±1.40
Δ=pFedREP-Baseline	0.15	-6.09	0.12	-0.20	2.22

Table 3: Overall evaluation under attacks for pFedREP.

Setting	CA ↑	W10 ↑	ASR ↓	MExp ↓	Pre-Q ↓	BRej ↓	Abst. ↓
Clean control	93.20±0.26	88.91±0.52	0.06±0.02	0.00±0.00	n/a	11.58±4.99	0.00±0.00
Proximity + patch	93.09±0.31	88.83±1.13	0.07±0.02	23.83±1.36	23.83±1.36	0.00±0.00	0.00±0.00

Table 4: Attack and contamination sensitivity of pFedREP, with the protocol of Table 3.

Upload/trigger attack	Mal.	CA ↑	ASR ↓	MExp ↓	BRej ↓	Abst. ↓
Proximity + patch	10%	93.34±0.29	0.07±0.01	12.21±0.19	3.76±6.51	0.00±0.00
Proximity + patch	20%	93.09±0.31	0.07±0.02	23.83±1.36	0.00±0.00	0.00±0.00
Proximity + patch	30%	93.32±0.05	0.09±0.03	34.70±1.05	3.02±3.89	0.00±0.00
Model replacement + patch	20%	38.19±10.17	0.03±0.04	0.00±0.00	12.37±9.18	0.00±0.00
Sign flip + patch	20%	93.01±0.28	0.06±0.01	20.63±6.73	8.78±7.41	0.00±0.00
Patch only	20%	93.03±0.35	0.08±0.01	23.68±3.33	8.28±4.40	0.00±0.00
Upload only	20%	93.16±0.17	n/a	23.82±1.35	0.00±0.00	0.00±0.00

4.3 RQ2: Robustness under malicious attacks

Table 3 compares the primary 20% proximity-plus-patch attack against an unattacked pFedREP under the same protocol. Evaluating the trigger patch on the clean control establishes the background rate at which the patch elicits the target class without any poisoned training, so that attack effect can be separated from patch salience. The attack shifts clean accuracy by only -0.11 points and W10 by -0.08 points, and its pooled ASR of $52/71{,}100 = 0.073\%$ is statistically indistinguishable from the clean control's patch-salience baseline of $55/90{,}000 = 0.061\%$. Notably, MExp remains at 23.83% and equals its Pre-Q value, meaning malicious prompts continue to be retrieved yet can not steer predictions. The defense therefore operates at the prediction level by capped routing and the dominant local branch instead of by detection alone, which is precisely the regime a proximity-mimicking attacker is designed to force. BRej and abstention are both zero under attack, so no benign expert is rejected and no client falls back to local-only inference.

Table 4 separates predictive behavior from pool exposure across contamination levels and upload attacks. Under proximity-constrained, sign-flip, and patch-only attacks, clean accuracy remains near 93% and ASR remains below 0.1%. Exposure tracks the malicious-client fraction under the proximity attack, showing that the low ASR is achieved while geometrically evasive experts remain retrievable. Model replacement shows that geometry removes all malicious slots, but clean accuracy falls to $38.19\pm10.17\%$, so that its low ASR does not ensure a useful accuracy–robustness trade-off.

Figure 2 explains why low ASR and high MExp coexist. Directed evidence matures for every recipient by round 3, but exposure remains near 24% across all rounds. The utility estimator differentiates recipient–expert pairs, yet all 48 observed means for the two malicious experts remain positive (seed-wise means 0.0012–0.0017), so neither directed gating nor quarantine is activated.

Table 5 shows that robustness does not hinge on precise controller calibration: across all nine configurations, clean accuracy varies by at most 0.11 points and ASR stays within 0.06–0.08%. The MAD threshold trades screening strictness against honest-expert cost, with the loosest setting (4.5) admitting the occasional beneficial-expert rejection (7.77 ± 7.26) while other settings incur none. The

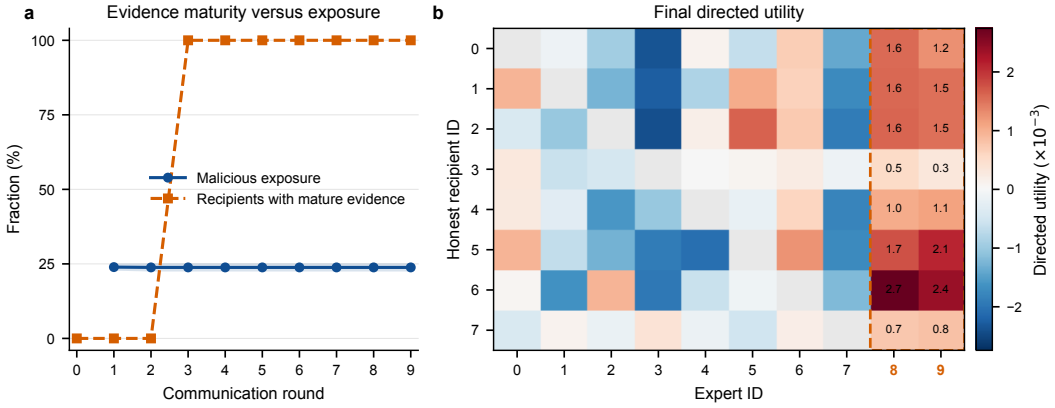

Figure 2: Behavioral evidence under the proximity-plus-patch attack. (a) Malicious-expert exposure and the fraction of recipients with mature directed evidence over communication rounds. Curves and bands are the mean and sample standard deviation over three seeds. Exposure is undefined at round 0 because no non-local expert is selected. (b) Final directed-utility means from honest recipients to expert identities, averaged over available seed-level means. The dashed columns are malicious experts 8 and 9, whose displayed utilities are all positive.

Table 5: Controller sensitivity under attacks.

Control	Value	CA ↑	ASR ↓	MExp ↓	BRej ↓
MAD threshold	2.5	93.20±0.12	0.07±0.03	24.66±0.07	0.00±0.00
	3.5	93.09±0.31	0.07±0.02	23.83±1.36	0.00±0.00
	4.5	93.09±0.33	0.06±0.01	23.49±1.13	7.77±7.26
Utility threshold τ_u	−0.10	93.09±0.31	0.07±0.02	23.83±1.36	0.00±0.00
	−0.30	93.09±0.31	0.07±0.02	23.83±1.36	0.00±0.00
	−0.50	93.09±0.31	0.07±0.02	23.83±1.36	0.00±0.00
Router cap $\bar{\alpha}$	0.20	93.14±0.10	0.06±0.02	23.83±0.99	1.59±2.75
	0.35	93.09±0.31	0.07±0.02	23.83±1.36	0.00±0.00
	0.50	93.17±0.10	0.08±0.03	23.77±1.04	0.96±1.67

utility-threshold rows are identical because observed rewards stay far closer to zero than any tested τ_u, so quarantine never activates in attacks. Router cap $\bar{\alpha}$ changes the conditional per-logit influence bound, leaving MExp near 24% throughout, which is consistent with the design intent that the cap governs influence, not admission.

Table 6 reports the exact attack results behind the factor ranges. Each block varies one property of the primary proximity-plus-patch attack while holding the malicious-client fraction at 20%. The default configuration (0.20 poisoned-example fraction, 4×4 trigger, target label 0) is repeated within each block for comparison.

Table 6: Sensitivity to poisoned-example fraction, trigger size, and target label, presented as mean ± standard deviation.

Factor	Value	CA ↑	W10 ↑	ASR ↓	MExp ↓	BRej ↓
Poison fraction	0.10	93.01±0.15	88.80±0.56	0.07±0.03	23.30±1.19	3.01±5.22
	0.20	93.09±0.31	88.83±1.13	0.07±0.02	23.83±1.36	0.00±0.00
	0.30	93.20±0.12	89.13±0.47	0.07±0.02	23.84±1.37	0.00±0.00
Trigger size	2	93.17±0.13	89.28±0.23	0.07±0.03	24.10±0.49	8.30±7.44
	4	93.09±0.31	88.83±1.13	0.07±0.02	23.83±1.36	0.00±0.00
	8	93.16±0.18	89.28±0.42	0.08±0.02	23.82±0.80	4.56±7.90
Target label	0	93.09±0.31	88.83±1.13	0.07±0.02	23.83±1.36	0.00±0.00
	1	93.08±0.16	88.66±0.60	0.14±0.03	23.54±0.96	3.68±3.20
	2	93.11±0.17	88.94±0.71	0.00±0.00	24.66±1.69	5.39±5.00

Containment is insensitive to attack strength. The poisoned-example fraction from 0.10 to 0.30 leaves ASR fixed at 0.07% and moves MExp by only half a point (23.30% to 23.84%), and quadrupling the trigger size from 2 to 8 pixels changes ASR by at most 0.01 points and MExp by less than 0.3 points; clean accuracy and W10 are flat throughout. The one factor with a visible effect is target label: ASR doubles to 0.14±0.03% for label 1 and drops to zero for label 2, presumably reflecting how easily each target class is reached from typical patched inputs. Even so, the variation is modest in absolute terms and, crucially, MExp stays near 24% in every row. The ASR differences therefore reflect intrinsic attack efficacy under a fixed influence budget, not differential removal of malicious experts: containment continues to operate at the prediction level regardless of attack configuration. Finally, the BRej column should be read with caution. Its large standard deviations (e.g., 8.30±7.44 for the 2×2 trigger) indicate that rejections are rare, seed-dependent events, so that the three-seed means are not stable estimates of a rejection rate.

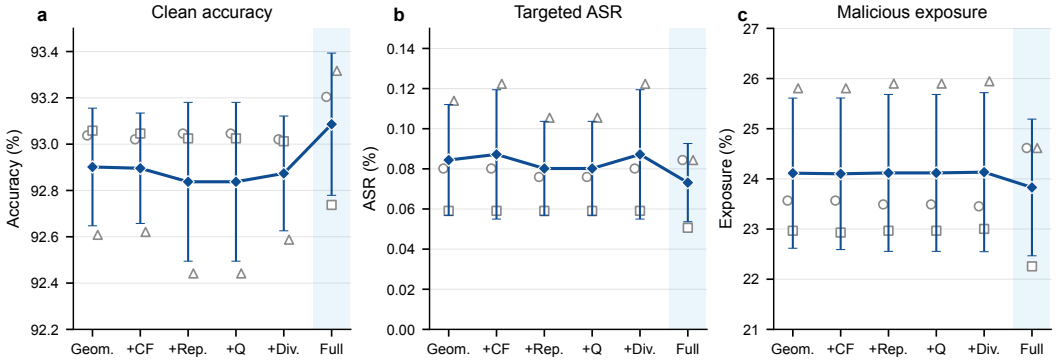

Figure 3: Incremental pFedREP mechanisms under attacks. Open markers show the three seeds, and diamonds and bars show the mean and standard deviation. CF, Rep., Q, and Div. denote counterfactual feedback, reputation, quarantine, and diversity. The shaded column marks full method.

Table 7: Retrieval study with 20% malicious clients. *Severe* refers to the severe negative transfer rate, i.e., the fraction of retrieved experts that significantly hurt the recipient client instead of helping it.

Method	Directed Utility	Severe	Target Exp.	Early → Late	Redund.
Nearest	$-.043 \pm .030$	$.173 \pm .021$	$.453 \pm .061$	$.453 \rightarrow .453$	$.491 \pm .069$
Geometry only	$-.110 \pm .024$	$.245 \pm .017$	$.653 \pm .045$	$.653 \rightarrow .653$	$.594 \pm .102$
+ directed	$.110 \pm .015$	$.006 \pm .000$	$.016 \pm .000$	$.079 \rightarrow .000$	$.335 \pm .091$
+ reputation	$-.015 \pm .047$	$.088 \pm .044$	$.215 \pm .104$	$.239 \rightarrow .209$	$.291 \pm .068$
All, w/o diversity	$.112 \pm .018$	$.006 \pm .000$	$.016 \pm .001$	$.077 \rightarrow .000$	$.336 \pm .111$
pFedREP	$.069 \pm .015$	$.006 \pm .000$	$.013 \pm .001$	$.066 \rightarrow .000$	$.065 \pm .117$

4.4 RQ3: Mechanism ablations

Incremental variant ablation. Figure 3 tests whether each stage receives an actionable signal under the image attack. All incremental variants stay within 0.25 points in clean accuracy, 0.02 points in ASR, and 0.30 points in MExp. The seed-level overlap is consistent with the positive malicious-expert utilities in Figure 2. This ablation establishes that prediction-level containment is not driven by quarantine or reputation activation. Full numerical ablations are in Appendix D.

Controlled 20-client simulation. Table 7 presents the controlled simulation to supply evidence of component-level attacks. Geometry attack alone is actively harmful under proximity imitation: it *amplifies* targeted exposure from 45.3% to 65.3% and severe negative transfer from 17.3% to 24.5%, confirming similarity as an attack surface. Adding directed utility cuts targeted exposure to 1.6% and severe negative transfer to 0.6% while turning mean utility positive (-0.110 to $+0.110$). Reputation alone yields only a softer population-level prior (21.5% exposure), consistent with its role as a cold-start ranking signal rather than a recipient-specific filter. Diversity serves to decorrelate the retrieved set, reducing redundancy from 0.336 to 0.065 while marginally improving targeted exposure (1.6% to 1.3%). The temporal columns show the evidence mechanisms sharpening as behavioral signal accumulates, with exposure falling from 6.6% early to 0.0% late. In summary, directed utility provides personalized safety, reputation provides cold-start ranking, and diversity decorrelates retrieval. More details of this control study are provided in Appendix E.

5 Conclusion

We studied attacks in federated prompt expert sharing. Our proposed pFedREP layers directed utility, reputation, diversity, and capped routing into a single evidence-driven controller, and we proved a structural per-logit influence bound that holds independently of retrieval outcomes. Across attack configurations, the system holds low ASR at no clean-accuracy cost, and controlled ablations attribute this to distinct, complementary roles for each component: personalized safety from directed utility, cold-start ranking from reputation, and decorrelated retrieval from diversity. Extending the framework to adaptive, colluding adversaries is a natural direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools for refining the phrasing, clarity, and organization of author-written text. We have not used generative AI tools for the conception or production of the problem formulation, methods, theoretical results, experiments, or analyses. Other tasks, such as generating research ideas or producing code are not applicable to this work. Additionally, we used generative AI tools for language polishing and editorial assistance. We have reviewed all AI-assisted work. For example, all AI-assisted text was reviewed, verified, and edited by the authors to ensure accuracy and clarity before inclusion in the manuscript. 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 studies poisoning and backdoor attacks in federated prompt expert sharing with the goal of building and evaluating defenses. All attacks implemented in this paper are standard techniques from the existing literature, evaluated in controlled simulations on publicly available benchmark datasets (e.g., Food101), and no human subjects, private data, or deployed systems were involved. While the detailed attack configurations we report could in principle inform adversaries, we believe the defensive insights and the proposed containment mechanisms outweigh this risk, as the attack primitives are already well documented. We adhere to the ICLR Code of Ethics and are aware of no conflicts of interest or sponsorship concerns related to this work.

REPRODUCIBILITY STATEMENT

We have taken several steps to ensure the reproducibility of our results. The threat model, defense components, and controller are formally specified in Section 3, and complete proofs of theorem, proposition, and corollaries are provided in Appendix C. Full experimental details, including datasets and preprocessing, client partitioning, attack configurations, controller hyperparameters (MAD threshold, utility threshold τ_u, router cap $\bar{\alpha}$), and evaluation metrics, are given in Section 4 and Appendix D. Sensitivity analyses over attack factors and controller settings are reported in Section D (Tables 5 and 6). All reported results are means over three random seeds with sample standard deviations. An anonymized implementation of pseudo-code is included in the supplementary material.

REFERENCES

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

Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 2938–2948, 2020.

Jiawang Bai, Kuofeng Gao, Shaobo Min, Shu-Tao Xia, Zhifeng Li, and Wei Liu. BadCLIP: Trigger-aware prompt learning for backdoor attacks on CLIP. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 24239–24250, 2024.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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.

Xiangrui Cai, Haidong Xu, Sihan Xu, Ying Zhang, and Yuan Xiaojie. Badprompt: Backdoor attacks on continuous prompts. In *Advances in Neural Information Processing Systems*, volume 35, pp. 37068–37080, 2022.

Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336, 1998.

Gary Cheng, Karan Chadha, and John Duchi. Fine-tuning is fine in federated learning. *arXiv preprint arXiv:2108.07313*, 2021.

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.

Chengying Fang, Wenke Huang, Guancheng Wan, Yihao Yang, and Mang Ye. FedPHA: Federated prompt learning for heterogeneous client adaptation. In *International Conference on Machine Learning*, pp. 15960–15975, 2025.

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*, 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–federated learning in age of foundation model. *IEEE Transactions on Mobile Computing*, 23(5):5179–5194, 2023b.

Wenwen He, Wenke Huang, Bin Yang, Shukan Liu, and Mang Ye. SPMC: Self-purifying federated backdoor defense via margin contribution. In *International Conference on Machine Learning*, pp. 22422–22433, 2025.

Wenke Huang, Mang Ye, Zekun Shi, Guancheng Wan, He Li, and Bo Du. Parameter disparities dissection for backdoor defense in heterogeneous federated learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

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.

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 (CVPR)*, pp. 19113–19122, June 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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15144–15154. IEEE, 2023b.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

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

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

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.

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.

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.

Yuwei Niu, Shuo He, Qi Wei, Zongyu Wu, Feng Liu, and Lei Feng. Test-time multimodal backdoor detection by contrastive prompting. In *International Conference on Machine Learning*, pp. 46629–46648, 2025.

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.

Yuan Wang, Huazhu Fu, Renuga Kanagavelu, Qingsong Wei, Yong Liu, and Rick Siow Mong Goh. An aggregation-free federated learning for tackling data heterogeneity. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 26233–26242, 2024.

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.

Chulin Xie, Minghao Chen, Pin-Yu Chen, and Bo Li. CRFL: Certifiably robust federated learning against backdoor attacks. In *International Conference on Machine Learning*, pp. 11372–11382, 2021.

Jiahao Xu, Zikai Zhang, and Rui Hu. Detecting backdoor attacks in federated learning via direction alignment inspection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 20654–20664, 2025.

Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, pp. 5650–5659, 2018.

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, 2022.

APPENDIX CONTENTS

A	Related work	13
B	Adversary, trust assumptions, and goals	14
C	Detailed proofs	14
D	Additional experimental results	16
E	Controlled mechanism study	20
F	Implementation-matched selection procedure	21

A RELATED WORK

Federated adaptation of vision–language models. Federated prompt learning adapts frozen vision–language models with a small communicated parameter set, but recent methods encode personalization in different places. PromptFL averages locally trained prompts (Guo et al., 2023b), while pFedPrompt retains local visual personalization around shared text prompts (Guo et al., 2023a). FedOTP couples global and local prompts through unbalanced optimal transport to handle simultaneous feature and label shift (Li et al., 2024). More recent designs broaden the shared–private decomposition: FedPHA combines a fixed-length global prompt with variable-length local prompts and alignment projections (Fang et al., 2025), whereas pFedMMA aggregates only the shared projection of a multimodal adapter and keeps modality-specific projections local (Ghiasvand et al., 2026). pFedMoAP downloads multiple uploaded prompts as distinct experts instead of reducing them to one prompt (Luo et al., 2025). These methods address heterogeneity and parameter-efficient adaptation, while PFEDREP instead asks how a retained expert pool should be governed when some uploaded prompts are adversarial.

Nonaggregated collaboration and expert selection. Avoiding update aggregation does not produce an expert pool. FedAF shares client-condensed data and trains one server model from those summaries (Wang et al., 2024), while pFedMoAP preserves prompt identities and learns a client-side mixture over retrieved experts. Once identities remain visible, the server must decide which experts a recipient should consume without collapsing legitimate specialists into a single center. Our retrieval stage follows the relevance–redundancy principle of maximum marginal relevance (Carbonell & Goldstein, 1998), but augments similarity with neighborhood screening, recipient-specific counterfactual utility, distinct-reporter reputation, and directed quarantine. The diversity term therefore controls redundancy among selected prompts; it is not evidence that their prediction errors are independent.

Backdoor defense under federated heterogeneity. Most Byzantine-robust FL defenses protect a global aggregate. Krum chooses an update using neighbor distances (Blanchard et al., 2017), and coordinate-wise median or trimmed mean limits coordinate-level contamination (Yin et al., 2018). Recent defenses incorporate richer signals to separate malicious behavior from benign heterogeneity: FDCR clusters and rescales updates using Fisher-based parameter importance (Huang et al., 2024), SPMC estimates marginal collaborative contribution and purifies suspicious gradients (He et al., 2025), and AlignIns examines update directions and significant-parameter signs (Xu et al., 2025). These methods filter, reweight, or repair client updates before or during aggregation. In contrast, an expert pool deliberately preserves uploaded parameters, so trust must be attached to an expert identity and, for personalized harm, to a recipient–expert pair. CRFL certifies a smoothed global model against bounded backdoor perturbations (Xie et al., 2021); our result is narrower and conditions on fixed retrieval and ordering while bounding the contribution of one retained non-local expert.

Backdoors in prompt-tuned vision–language models. Prompts are themselves an attack surface. BadPrompt implants targeted behavior in continuous prompts (Cai et al., 2022), and BadCLIP learns a visual trigger together with trigger-aware prompts that affect CLIP image–text alignment (Bai et al., 2024). Defenses for centralized multimodal models act at other stages: BDetCLIP, for example, detects triggered inputs at inference time through contrastive benign and perturbed text descriptions (Niu et al., 2025). Our setting instead keeps the CLIP encoders frozen while clients upload prompts that may later be delivered intact to other clients. PFEDREP governs pool membership and exposure through staged server-side and recipient-side evidence, then caps reliance on any selected expert. Encoder repair and test-time trigger detection remain outside its scope.

B ADVERSARY, TRUST ASSUMPTIONS, AND GOALS

The adversary controls up to f Byzantine clients and may coordinate their behavior across rounds. A malicious client may poison triggered local examples, optimize a targeted prompt backdoor, upload an arbitrary prompt, project that prompt into a small neighborhood of the honest pool, or release arbitrary utility reports about consumed experts. We assume that the server is honest, that client identities are authenticated, and that the frozen CLIP backbone is uncompromised. The server observes prompts, identifiers, and released scalar feedback; it does not observe client examples.

Authenticated identities are necessary for interpreting population feedback: distinct-reporter statistics prevent repeated reports from a single identity from increasing its representation, but they offer no defense against an adversary capable of creating many authenticated identities. We likewise exclude server compromise and backbone replacement from the threat model. The influence result in Section 3.5 rests on the additional, narrower assumption that retrieval and expert ordering remain fixed between the compared predictions.

Robust expert-pool learning must balance four objectives that need not coincide. First, upload screening should reject conspicuous parameter attacks without collapsing minority domains. Second, recipient-level transfer safety should reduce targeted exposure, attack success, and severe negative transfer. Third, personalization should retain experts that benefit heterogeneous clients. Fourth, bounded reliance should limit the effect of any single selected non-local expert. Geometry alone can not establish transfer safety, because nearby prompts may behave differently on a rare trigger distribution. Table 8 summarizes the evidence available at each decision stage and the failure modes that remain outside its scope.

Table 8: Evidence sources, their roles in expert-pool governance, and the failures that remain outside each decision.

Stage	Evidence	Decision	Residual failure
Geometry	Current prompts	Reject/clip outliers	Proximity-constrained backdoor
Directed utility	Recipient loss	Score or quarantine pair (i, j)	Cold-start exposure; noisy credit
Reputation	Distinct reporters	Soft population prior	Sybils; contamination above tolerance
Diversity	Prompt similarity	Penalize redundant set	Parameter diversity is not error independence
Capped router	Image and slot	Bound expert weight	Retrieval remains uncertified
Fallback	Empty safe set	Use local expert only	Loss of collaboration

C DETAILED PROOFS

C.1 BYZANTINE-REPORTER QUANTILE BOUND

Detailed proof of Proposition 1. Write $M = M_j^t$, let B be the number of Byzantine reporters, and let $H = M - B$ be the number of honest reporters. The premise gives $B \le \lfloor \beta M \rfloor$, hence the realized Byzantine fraction $\rho = B/M$ satisfies $0 \le \rho \le \beta$. Since $\beta < q < 1 - \beta$, we have $\beta < 1/2$, so $H > 0$.

Let F, F_H, and F_B be the empirical CDFs of the full, honest, and Byzantine report multisets, respectively. If $B = 0$, take F_B to be any CDF, since its coefficient below is zero. The full empirical CDF decomposes exactly as

$$F(x) = (1 - \rho)F_H(x) + \rho F_B(x). \tag{12}$$

Because $0 \le F_B(x) \le 1$, Eq. 12 first yields

$$(1 - \rho)F_H(x) \le F(x) \le (1 - \rho)F_H(x) + \rho.$$

The inequality $\rho \le \beta$ then gives the uniform bounds

$$(1 - \beta)F_H(x) \le F(x) \le (1 - \beta)F_H(x) + \beta. \tag{13}$$

For the upper inequality, note explicitly that

$$(1-\rho)F_H(x)+\rho=(1-\beta)F_H(x)+\beta-(\beta-\rho)(1-F_H(x)),$$

whose final term is non-positive.

For a finite multiset $\mathcal{A}$ with empirical CDF $F_{\mathcal{A}}$, define its empirical quantile by the generalized inverse

$$Q_r(\mathcal{A})=\inf\{x\in\mathbb{R}:F_{\mathcal{A}}(x)\geq r\},\qquad 0<r<1.$$

Set

$$r_-=\frac{q-\beta}{1-\beta},\qquad r_+=\frac{q}{1-\beta}.$$

The condition $\beta<q<1-\beta$ ensures $0<r_-<r_+<1$.

To prove the lower bound, take any $x<Q_{r_-}(\mathcal{H}_j^t)$. By the generalized-inverse definition, $F_H(x)<r_-$. The upper half of Eq. 13 therefore implies

$$F(x)\leq(1-\beta)F_H(x)+\beta<(1-\beta)r_-+\beta=q.$$

Thus the full CDF can not reach level q before the honest CDF reaches level r_-, and

$$Q_{r_-}(\mathcal{H}_j^t)\leq Q_q(\mathcal{U}_j^t).$$

For the upper bound, let $x_+=Q_{r_+}(\mathcal{H}_j^t)$. Since the honest multiset is finite, its empirical CDF attains the quantile level, so $F_H(x_+)\geq r_+$. The lower half of Eq. 13 gives

$$F(x_+)\geq(1-\beta)F_H(x_+)\geq(1-\beta)r_+=q.$$

Hence $Q_q(\mathcal{U}_j^t)\leq x_+$. Combining the two inequalities proves Eq. 7. The argument uses no restriction on the Byzantine report values beyond their number; their empirical CDF enters only through the universal bounds $0\leq F_B\leq 1$. □

C.2 Localization of the reputation score

Explicit interval. Corollary 1 gives

$$\rho_j^t\in\frac{M_j^t}{M_j^t+\kappa}\Big[(1-\gamma)Q_{\frac{1/2-\beta}{1-\beta}}(\mathcal{H}_j^t)+\gamma Q_{\frac{q-\beta}{1-\beta}}(\mathcal{H}_j^t),\ (1-\gamma)Q_{\frac{1/2}{1-\beta}}(\mathcal{H}_j^t)+\gamma Q_{\frac{q}{1-\beta}}(\mathcal{H}_j^t)\Big].$$

Detailed proof of Corollary 1. The intended reputation hyperparameters satisfy $0\leq\gamma\leq 1$ and $\kappa\geq 0$. Under the corollary's premise $\beta<q\leq 1/2$, both q and $1/2$ lie strictly between β and $1-\beta$. Proposition 1 can therefore be applied at both quantile levels.

For $r\in\{q,1/2\}$, define the honest lower and upper endpoints

$$L_r=Q_{\frac{r-\beta}{1-\beta}}(\mathcal{H}_j^t),\qquad U_r=Q_{\frac{r}{1-\beta}}(\mathcal{H}_j^t).$$

Proposition 1 gives

$$L_q\leq Q_q(\mathcal{U}_j^t)\leq U_q,\qquad L_{1/2}\leq\mathrm{med}(\mathcal{U}_j^t)\leq U_{1/2}.$$

Let $c=M_j^t/(M_j^t+\kappa)$. Since $c\geq 0$ and the coefficients $1-\gamma$ and γ are non-negative, multiplying the two lower bounds by their respective coefficients and adding them preserves the inequality. The same is true for the upper bounds. Substitution into Eq. 6 yields

$$c\big[(1-\gamma)L_{1/2}+\gamma L_q\big]\leq\rho_j^t\leq c\big[(1-\gamma)U_{1/2}+\gamma U_q\big],$$

which is exactly the interval stated in Corollary 1.

Finally, every honest u_{ij}^t is an average of clipped rewards $\widehat{u}_{ij}^{\tau}\in[-1,1]$, so all honest quantile endpoints are finite and lie in $[-1,1]$. Also $0<c\leq 1$ whenever $M_j^t>0$. Byzantine reporters may move the reputation within the displayed honest-quantile interval, but arbitrary report magnitudes can not move it outside that interval. □

C.3 CONDITIONAL SINGLE-EXPERT INFLUENCE

Detailed proof of Theorem 1. The theorem concerns a non-local expert, so the routed mixture contains at least two slots and $E > 1$. Since $1/E \le \bar{\alpha} \le 1$, Eq. 10 also gives $0 \le \eta \le 1$. The weight of slot j is

$$a_{i,j}(x) = (1-\eta)b_{i,j}(x) + \frac{\eta}{E}.$$

Since $b_{i,j}(x) \le 1$ and $\eta = (1-\bar{\alpha})/(1-1/E)$,

$$\begin{aligned} a_{i,j}(x) &\le 1 - \eta + \frac{\eta}{E} \\ &= 1 - \eta\left(1 - \frac{1}{E}\right) \\ &= 1 - (1-\bar{\alpha}) = \bar{\alpha}. \end{aligned}$$

Fixing the image, retrieved set, slot ordering, and gate parameters fixes every mixture weight. In particular, replacing the expert stored in slot j does not change $a_{i,j}(x)$ because the gate depends on $z(x)$ and the slot index rather than on the prompt value. The local term in Eq. 1 is unchanged, as are all routed terms outside slot j. Subtracting the original class-c logit from the replaced logit therefore gives the exact identity

$$\ell'_{i,c}(x) - \ell_{i,c}(x) = s\, a_{i,j}(x)\left\langle z(x), v'_{j,c} - v_{j,c}\right\rangle.$$

Taking absolute values and applying Cauchy–Schwarz followed by the triangle inequality yields

$$\begin{aligned} \left|\ell'_{i,c}(x) - \ell_{i,c}(x)\right| &\le s\, a_{i,j}(x)\|z(x)\|_2\|v'_{j,c} - v_{j,c}\|_2 \\ &\le s\, a_{i,j}(x)\|z(x)\|_2\left(\|v'_{j,c}\|_2 + \|v_{j,c}\|_2\right) \\ &\le 2s\, a_{i,j}(x) \\ &\le 2s\bar{\alpha}. \end{aligned}$$

This proves Eq. 11 for every class c. The constant 2 accounts for the largest possible distance between two unit-norm class representations; no distributional assumption on the replacement expert is used. □

C.4 CONDITIONAL TOP-1 INVARIANCE

Detailed proof of Corollary 2. Let

$$\delta_c = \ell'_{i,c}(x) - \ell_{i,c}(x).$$

Theorem 1 gives $|\delta_c| \le 2s\bar{\alpha}$ for every class. For any competitor $c \ne \widehat{c}$, the post-replacement pairwise margin satisfies

$$\begin{aligned} \ell'_{i,\widehat{c}}(x) - \ell'_{i,c}(x) &= \ell_{i,\widehat{c}}(x) - \ell_{i,c}(x) + \delta_{\widehat{c}} - \delta_c \\ &\ge m(x) - |\delta_{\widehat{c}}| - |\delta_c| \\ &\ge m(x) - 4s\bar{\alpha}. \end{aligned}$$

If $m(x) > 4s\bar{\alpha}$, the final expression is strictly positive for every $c \ne \widehat{c}$. Thus $\widehat{c}$ remains the unique maximizer after the replacement, and the predicted class can not change. The strict margin condition also excludes an original top-1 tie. □

D ADDITIONAL EXPERIMENTAL RESULTS

This section reports completed measurements that are stored in the experiment summaries but are not tabulated in the main paper. All entries are repository-measured means and sample standard deviations over seeds 1, 2, and 3. The clean results cover pFedMoAP at both tested values of λ and both backbone variants of pFedMMA and PFEDREP. In the main clean tables, pFedMoAP uses RN50 on the five few-shot datasets and CIFAR10/100 and ViT-B/16 on DomainNet and Office-Caltech10. Comparisons between different backbones are descriptive, and no significance claim is made from three seeds.

D.1 PROTOCOL AND IMPLEMENTATION DETAILS

We use the multiple datasets and partitions from pFedMoAP. Flowers102 (Nilsback & Zisserman, 2008), OxfordPets (Parkhi et al., 2012), Food101 (Bossard et al., 2014), Caltech101 (Fei-Fei, 2004), and DTD (Cimpoi et al., 2014) are partitioned among ten clients with disjoint class assignments; the primary comparison uses 16 examples per class. CIFAR10/100 (Krizhevsky, 2009; Krizhevsky et al., 2010) use a Dirichlet(0.5) label split, while DomainNet (Peng et al., 2019) and Office-Caltech10 (Gong et al., 2012) combine domain shift with a Dirichlet(0.3) label split.

Table 9: Dataset and federation settings inherited from pFedMoAP (Luo et al., 2025). Participation is the fraction of clients sampled in each round. The backbone column describes the published baselines and pFedMoAP; pFedMMA and PFEDREP are each evaluated with both ViT-B/16 and RN50.

Dataset group	Clients	Rounds	Participation	Backbone	Heterogeneity
Five CLIP datasets	10	10	100%	RN50	Pathological label split
CIFAR10/100	100	120	10%	RN50	Dirichlet(0.5) labels
DomainNet	30	25	25%	ViT-B/16	Feature + Dirichlet(0.3) label
Office-Caltech10	20	25	50%	ViT-B/16	Feature + Dirichlet(0.3) label

All federated methods use five local epochs and SGD with learning rate 0.002; local CoOp trains for 25 epochs. The pFedMoAP gating network uses learning rate 0.01 and eight heads. All repository-measured configurations use fp32 precision and seeds 1–3. Prompt length is 16, and pFedMoAP and PFEDREP use ten total expert slots. For PFEDREP, $\lambda = 0.5$, the configured target cap weight is 0.35, the median/MAD threshold is 3.5, the clipping radius is three robust scales, the directed negative-utility threshold is -0.30, and quarantine lasts five rounds. These values are fixed before attack evaluation.

We implement PFEDREP by extending the pFedMoAP prompt-learning and gating stack while keeping both CLIP encoders frozen. The runtime indexes private prompts and gating parameters by client identity and persists the sanitized pool, directed utility observations, quarantine timers, distinct-reporter counts, and random-number-generator state. Before an update, the server restores the client's private state, screens uploaded prompts, retrieves robustly selected non-local experts, and assigns them to fixed rank slots for the prompt-independent capped router. Clipped leave-one-expert-out rewards then update directed trust. The original campaign uses RN50 for the five few-shot datasets and CIFAR10/100 and ViT-B/16 for DomainNet and Office-Caltech10; the complementary campaign reverses these backbone assignments.

Our pFedMMA implementation uses parallel multi-modal adapters in frozen CLIP encoders. Each adapter has local modality-specific down/up projections and a shared 32-dimensional middle projection averaged by participating-client sample count. ViT-B/16 uses paired adapters in blocks 10–12 for the five few-shot datasets and blocks 5–12 otherwise. For RN50, the text adapter consumes layer-normalized transformer states and the visual adapter operates channel-wise on selected bottlenecks; text layers 10–12 pair with bottlenecks 14–16, and text layers 5–12 pair with bottlenecks 9–16. Both campaigns use the template "a photo of a $\{$class$\}$.", residual scale 0.001, and the optimization settings above. The two backbone rows are matched-protocol measurements, not a controlled backbone ablation or a reproduction of pFedMMA's native schedule.

The primary attack controls 20% of clients throughout training, poisons 20% of their local examples with a 4×4 bottom-right patch, and relabels them to class 0. We pair the trigger with proximity-constrained, model-replacement, and sign-flip prompt uploads. Sensitivity experiments vary the malicious fraction, poison fraction, trigger size, target class, and attack interval. Clean accuracy is macro-averaged across honest clients; W10 averages their lowest accuracy decile; targeted ASR is computed on triggered non-target examples; MExp is the fraction of downloaded non-local slots assigned to malicious identities. Beneficial honest rejection counts removed candidates with positive observed directed mean, and local-only abstention records selections in which every non-local candidate is rejected.

D.2 TAIL-CLIENT ACCURACY

Figure 4 visualizes the within-seed difference between macro accuracy and W10. The gap is below 12 points for every method on the first seven datasets except pFedMoAP at $\lambda = 0$ on DTD and CIFAR10. Feature-and-label shift produces the largest disparities: every DomainNet gap exceeds 15 points and every Office-Caltech10 gap exceeds 20 points. The largest cell is 32.2 points for pFedMoAP at $\lambda = 0.5$ on Office-Caltech10. Thus a method can retain moderate macro accuracy while leaving the lowest-performing clients near failure, which motivates reporting W10 beside the average.

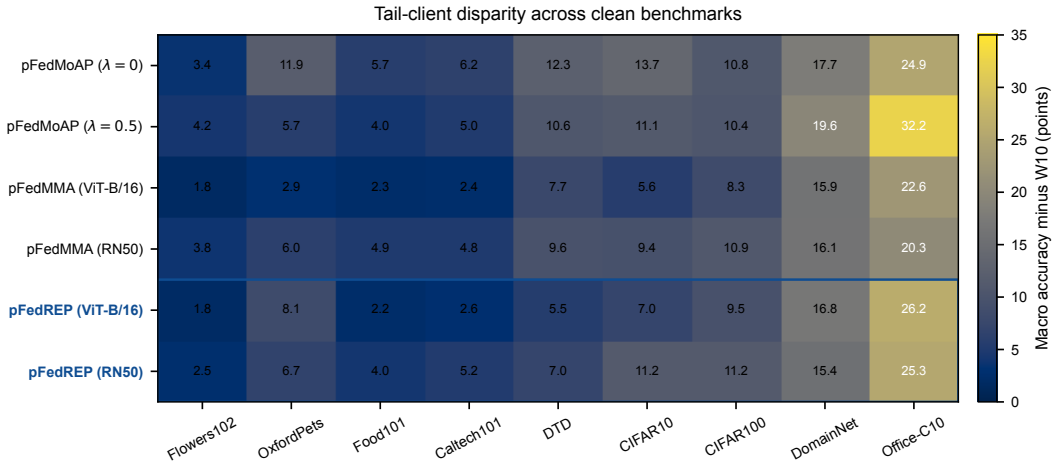

Figure 4: Tail-client disparity across completed clean benchmarks. Each cell is the mean over three seeds of the within-seed difference between macro client accuracy and W10, in percentage points; larger values indicate a wider tail gap, not necessarily lower macro accuracy. The blue outline marks the two pFEDREP rows. pFedMoAP uses the protocol backbones described above.

D.3 DOMAIN-LEVEL BEHAVIOR UNDER ATTACK

Table 10 resolves the attacked DomainNet and Office-Caltech10 results by honest-client domain. The table reports accuracy and beneficial honest-expert rejection because triggered ASR is aggregated at the dataset level in the stored summaries.

Table 10: Domain-level results for honest-client cohorts under the 20% proximity-plus-patch attack. CA and BRej are percentages reported as mean ± sample standard deviation over three seeds. DomainNet has no honest Sketch client, and Office-Caltech10 has one honest Webcam client.

Dataset	Honest-client domain	CA ↑	BRej ↓
DomainNet	Clipart	44.82±0.93	0.00±0.00
	Infograph	47.52±0.43	0.00±0.00
	Painting	30.63±0.18	0.00±0.00
	Quickdraw	37.77±0.10	0.00±0.00
	Real	32.21±0.52	0.00±0.00
Office-Caltech10	Amazon	31.31±2.93	7.64±3.98
	Caltech	33.60±2.64	8.57±3.06
	DSLR	39.33±5.93	9.20±3.94
	Webcam	75.00±0.00	9.82±3.80

Across the honest clients retained by the attack protocol, DomainNet has 38.86±0.29% macro accuracy, 20.79±0.26% W10, and 6.60±2.26% ASR. Its beneficial rejection rate is zero in every observed domain, even though domain accuracy ranges from 30.63±0.18% on Painting to 47.52±0.43% on Infograph. Office-Caltech10 has 37.26±2.40% macro accuracy, 13.00±2.02% W10, and 10.73±2.96% ASR, with nonzero BRej in all four observed domains. The Webcam accuracy of 75.00±0.00% is based on one honest client and must not be read as evidence of low domain variance. These

cohort-specific results show that average rejection and accuracy can mask different domain behavior, but the deterministic malicious identities prevent a balanced six-domain comparison on DomainNet.

D.4 Real-data ablations and model sensitivity

Table 11 supplies the exact values summarized by Figure 3. Across the incremental variants, clean accuracy, ASR, and MExp differ by at most 0.25, 0.02, and 0.30 percentage points, respectively. Together with the positive malicious-expert utilities in Figure 2, this shows that the primary attack does not activate the directed gate or quarantine; the controlled study in Section E is therefore the appropriate source for isolating their roles.

Table 11: Incremental component ablation under the primary attacks. CF is counterfactual feedback, Rep is reputation, Q is quarantine, Div is diversity, Med is robust-median cold start, and Cap is capped routing.

Variant	Geom.	CF	Rep	Q	Div	Med	Cap	CA	ASR	MExp	BRej
Geometry only	✓							92.90±0.25	0.08±0.03	24.11±1.50	n/a
+ directed CF	✓	✓						92.90±0.24	0.09±0.03	24.10±1.51	2.92±2.59
+ reputation	✓	✓	✓					92.84±0.34	0.08±0.02	24.12±1.56	3.41±1.96
+ quarantine	✓	✓	✓	✓				92.84±0.34	0.08±0.02	24.12±1.56	3.41±1.96
+ diversity	✓	✓	✓	✓	✓			92.87±0.25	0.09±0.03	24.13±1.59	3.42±2.20
Full pFedREP	✓	✓	✓	✓	✓	✓	✓	93.09±0.31	0.07±0.02	23.83±1.36	0.00±0.00

Table 12: Focused trust, reputation, router, and initialization ablations under the primary attacks. Full pFedREP is repeated as the reference.

Variant	CA ↑	W10 ↑	ASR ↓	MExp ↓	BRej ↓
Full pFedREP	93.09±0.31	88.83±1.13	0.07±0.02	23.83±1.36	0.00±0.00
Mixture-level feedback	93.20±0.24	89.07±0.73	0.06±0.02	23.96±0.93	n/a
Global utility	93.14±0.05	89.30±0.42	0.08±0.02	23.49±1.11	0.50±0.86
Global reputation	93.18±0.17	89.00±0.60	0.06±0.02	23.85±1.32	0.16±0.28
Raw-report maturity	93.09±0.23	88.74±0.90	0.07±0.01	23.83±1.36	0.00±0.00
Median-only reputation	93.12±0.28	88.73±0.86	0.08±0.02	23.75±1.30	0.00±0.00
Hard reputation veto	93.09±0.31	88.83±1.13	0.07±0.02	23.83±1.36	0.00±0.00
Prompt-content router	92.87±0.25	89.08±0.57	0.09±0.03	24.13±1.59	3.42±2.20
Uncapped router	93.11±0.11	89.19±0.57	0.05±0.02	23.82±1.18	0.54±0.94
Mean cold start	93.09±0.31	88.83±1.13	0.07±0.02	23.83±1.36	0.00±0.00

The focused variants yield similar outcomes for the same activation reason. The uncapped router reaches $0.05\pm0.02\%$ ASR versus $0.07\pm0.02\%$ for the capped router, but it removes the premise of the influence theorem. Thus the cap's contribution is the post-selection structural bound rather than an empirical ASR gain on this attack.

Table 13: Data and model sensitivity under the primary attacks. Defaults are 10 total experts, 16 shots, $\lambda = 0.5$, router dimension 128, and five quarantine rounds.

Factor	Value	CA ↑	W10 ↑	ASR ↓	MExp ↓
Total experts	3	93.19±0.26	89.08±0.68	0.11±0.02	52.08±0.69
	5	93.19±0.06	88.48±0.27	0.07±0.04	37.62±3.50
	10	93.09±0.31	88.83±1.13	0.07±0.02	23.83±1.36
Shots	1	80.06±1.69	68.24±5.54	0.08±0.04	22.29±0.07
	2	86.63±1.08	78.51±5.52	0.09±0.03	22.29±0.12
	4	90.25±0.24	84.76±0.80	0.05±0.01	22.37±0.26
	8	91.98±0.36	86.77±0.82	0.11±0.04	23.01±1.08
	16	93.09±0.31	88.83±1.13	0.07±0.02	23.83±1.36
λ	0	78.82±6.00	54.60±32.38	0.21±0.12	23.27±0.99
	0.5	93.09±0.31	88.83±1.13	0.07±0.02	23.83±1.36
	1.0	92.44±0.28	88.78±0.62	0.06±0.03	23.54±1.19
	2.5	92.08±0.13	88.14±0.59	0.07±0.02	23.68±1.29
	5.0	91.88±0.34	88.34±0.81	0.05±0.02	23.16±1.41
Router dimension	32	93.28±0.12	89.54±0.51	0.08±0.02	24.55±0.12
	64	93.25±0.25	88.84±0.49	0.08±0.01	23.75±1.11
	128	93.09±0.31	88.83±1.13	0.07±0.02	23.83±1.36
	256	93.09±0.25	89.06±0.65	0.11±0.04	24.22±0.55
	512	93.06±0.35	88.58±0.76	0.07±0.02	23.93±1.08
	1024	92.99±0.32	89.14±0.17	0.06±0.01	23.90±0.56
Quarantine rounds	0	93.09±0.31	88.83±1.13	0.07±0.02	23.83±1.36
	3	93.09±0.31	88.83±1.13	0.07±0.02	23.83±1.36
	5	93.09±0.31	88.83±1.13	0.07±0.02	23.83±1.36
	10	93.09±0.31	88.83±1.13	0.07±0.02	23.83±1.36

More data improves clean accuracy monotonically from $80.06\pm1.69\%$ at one shot to $93.09\pm0.31\%$ at 16 shots, while ASR stays between 0.05% and 0.11%. Setting $\lambda = 0$ removes the full-dimensional local branch and reduces accuracy to $78.82\pm6.00\%$; $\lambda = 0.5$ gives the best accuracy among the tested values. Router dimension has little effect. More total slots dilute exposure from 52.08% with three experts to 23.83% with ten, whereas changing the quarantine duration has no effect because quarantine is not activated in these runs.

E Controlled mechanism study

E.1 Questions and setup

We use a controlled simulator to answer four diagnostic questions: (Q1) does geometry detect a proximity-evasive malicious cluster; (Q2) which historical signal suppresses targeted exposure; (Q3) how much cold-start harm occurs before feedback matures; and (Q4) what utility is paid for diversity?

The simulator contains 20 clients in five heterogeneous prompt groups, 50 rounds, five seeds, and five non-local slots. Four clients are Byzantine. Their prompts imitate an honest robust center plus a small perturbation. One third of honest clients are targeted: a malicious expert has utility -0.75 for them and $+0.08$ otherwise. Byzantine reporters praise malicious experts and defame honest experts. Honest pair utilities combine prompt similarity with an independent recipient-preference/expert-trait interaction; feedback adds Gaussian noise with standard deviation 0.06.

We compare nearest retrieval; geometry only; geometry plus directed utility/quarantine; geometry plus reputation; all evidence without diversity; and full pFedREP. "Early" denotes the first ten rounds and "late" the remaining forty. Severe harm is utility below -0.4. Honest-benefit exclusion measures the fraction of honest candidates with positive ground-truth utility removed by hard gates. The default global median-reputation veto is disabled. All methods select up to five experts; local-only abstention is recorded as zero transfer utility and zero exposure.

This study is deterministic given seeds and exercises the exact pool manager, state updates, and influence-cap calculation. It does *not* measure CLIP clean accuracy or image-trigger ASR.

E.2 Results

The evaluated results are summarized in Table 7.

Geometry is necessary but insufficient. The attackers occupy a dense central region, so geometry alone is worse than nearest retrieval: targeted exposure rises from 45.3% to 65.3%. Local robust statistics reject isolated uploads; they do not identify a behaviorally malicious cluster.

Directed evidence supplies the principal safety signal. Adding recipient-specific utility and quarantine reduces targeted exposure to 1.6% and severe harm to 0.6%. Reputation alone is weaker (21.5% exposure) because a targeted expert helps the non-victim majority and Byzantine reporters manipulate the tail. This component result supports the design choice that global reputation should not replace personalized evidence.

The defense learns after exposure. Full pFEDREP reduces target exposure from 6.6% in the first ten rounds to 0.00% later. The overall 1.3% rate therefore contains a concentrated cold-start cost. A deployment should report exposure before detection and attack start and stop times, not only a final ASR.

Diversity trades utility for decorrelation. Relative to the no-diversity ablation, full pFEDREP reduces parameter-space redundancy from 0.336 to 0.065 and target exposure from 1.6% to 1.3%, but mean transfer utility falls from 0.112 to 0.069. Diversity is therefore not a free accuracy improvement. Its value is protection against correlated selection, and w_D should be chosen on a clean utility–tail-risk Pareto curve.

E.3 CONTAMINATION SENSITIVITY

At Byzantine fractions of 10%, 20%, and 30%, full pFEDREP attains targeted exposures of 0.49%, 1.3%, and 2.34%, respectively; severe-harm rates are 0.27%, 0.64%, and 0.95%. Late exposure is zero at 10% and 30%, and 0.03% at 20%. The 30% setting violates the lower-quartile premise of Proposition 1, yet directed utility still contains harm in this simulation. This is evidence for defense in depth, not a guarantee above the reputation breakdown point.

F IMPLEMENTATION-MATCHED SELECTION PROCEDURE

For M available prompts, pairwise geometry costs $O(M^2 d)$, reputation construction from cached reporter summaries costs at most $O(MN)$, and greedy selection costs $O(KMd)$. Prompt text features are computed only for selected experts; approximate nearest-neighbor search and cached summaries can reduce server cost for large pools.

For recipient i at round t, the pool manager executes:

1. enumerate available, shape-compatible prompts and compute local anchors, anomaly scores, and sanitized prompts;
2. remove i and apply anomaly, directed-quarantine, directed-mean-utility, and optional median-reputation gates;
3. compute equation 8 for every survivor and greedily retrieve at most K experts using equation 9;
4. fail closed to zero non-local prompts if every candidate is rejected;
5. route the local and selected normalized text features using equation 10;
6. compute equation 5 for each selected non-local expert, update running pair means and quarantine streaks, and persist the state.

Client identifiers, rather than positions in a filtered list, are returned throughout. Reputation maturity counts distinct nonzero reporter–expert entries; repeated batches from one reporter can not satisfy the minimum-reporter threshold.