FEDOPHD: FEDERATED POST-HOC OOD DETECTION VIA UNLEARNING-INSPIRED PROBING

Anonymous authors
Paper under double-blind review

ABSTRACT

In federated learning (FL), in-distribution (ID) data include both Local-ID (L-ID) from the deployment client and Transferred-ID (T-ID) from other participating clients. Locally unfamiliar T-ID must be distinguished from true out-of-distribution (OOD) inputs, even when only a trained global model and L-ID data are available. We propose FedOPhD, a post-hoc federated OOD detector that probes semantic support through unlearning responses. FedOPhD performs client-conditioned semantic attenuation on a diagnostic model copy, using the client's learned label support to attenuate locally supported semantics. It measures how an input's representation changes align with responses on an L-ID reference subset across multiple layers. These responses offer additional evidence for distinguishing inputs with similar original model prediction confidence. This response evidence is combined with the original model's global recognition support through locally calibrated fusion, accounting for T-ID beyond local semantic coverage. Detector construction requires neither OOD data nor further federated communication, while the deployed classifier remains unchanged. Experiments on CIFAR-10 and CIFAR-100 demonstrate the effectiveness of FedOPhD in distinguishing OOD inputs from both L-ID and T-ID across heterogeneous federated settings.

1 INTRODUCTION

Federated learning (FL) decouples what a client has seen from what its model has learned. A deployment client may encounter inputs absent or scarce in its local data yet supported by semantics learned from other participants (Yu et al., 2023; Liao et al., 2024). This creates a distinctive ambiguity for out-of-distribution (OOD) detection: local unfamiliarity does not imply that an input lies outside the federation's in-distribution (ID) support. We refer to ID samples originating from the current client as Local-ID (L-ID) and those originating from other participating clients as Transferred-ID (T-ID); both constitute valid federation-level ID. The OOD boundary is therefore federation-wide, while the data available to characterize it at an individual client may be only local (Liao et al., 2024). A detector that equates local unfamiliarity with OOD thus risks rejecting precisely the cross-client knowledge that FL is intended to transfer.

This ambiguity becomes particularly difficult once federated training has ended. At deployment, a client may have access only to the trained global model and its own L-ID data. These two remaining sources of information provide complementary but individually insufficient evidence. The global model encodes federation-wide knowledge and can recognize T-ID, yet its predictive scores may remain high on OOD inputs (Hendrycks & Gimpel, 2017; Liu et al., 2020). L-ID data provide a direct reference for ID behavior, but their incomplete semantic coverage can make valid T-ID appear anomalous. Figure 1 illustrates this tension: prediction-based Energy assigns similar scores to L-ID and T-ID but overlaps substantially with OOD, whereas the locally fitted ViM reference causes T-ID itself to overlap with OOD. Existing federated OOD methods largely obtain federation-aware information during federated training, relying on client–server interaction to in-

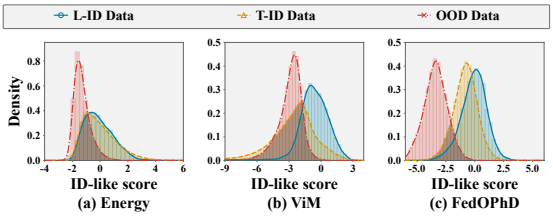

Figure 1: Illustrative score distributions on CIFAR-10 for post-hoc federated OOD detection.

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.

corporate OOD awareness into the learned model or detector (Yu et al., 2023; Liao et al., 2024; Huang et al., 2026). In contrast, we target already-trained federated models for which OOD detection must be added without revisiting federated training: client–server interaction has ended, and no new federation-wide information can be acquired. This leaves a central post-hoc question: how can federation-level semantic support be recovered from only a trained global model and semantically incomplete L-ID data?

Our key observation is that federation-level semantic support need not be reconstructed from federation-wide data; it can instead be probed from the trained global model. The global model already encodes semantics acquired across clients, while L-ID data can be used to interrogate this knowledge rather than define its boundary. Inspired by machine unlearning (MU), we repurpose selective forgetting of learned semantics as a diagnostic operation (Cao & Yang, 2015; Liu et al., 2021; Chen et al., 2023; Zhang et al., 2025). Inputs with similar original-model confidence may depend differently on the model's learned semantic support and therefore respond differently when that support is attenuated. Such responses provide dynamic evidence that is unavailable from the original prediction alone (Granz et al., 2024; Chen et al., 2025; Gu et al., 2025). Importantly, local data serve to induce and calibrate the probe rather than delimit federation-level ID, so T-ID need not belong to a locally observed class to exhibit an ID-like response.

Motivated by this insight, we propose **FedOPhD**, a **Fed**erated **O**OD **P**ost-**h**oc **D**etector that probes federation-wide semantic support through unlearning-inspired interventions. FedOPhD first performs client-conditioned semantic attenuation on a diagnostic copy of the global model, using the client's learned label support to guide a short sequence of updates. It then characterizes each input by its multi-layer representation changes between the original and attenuated models. Responses on an L-ID reference subset provide reference patterns for these changes; FedOPhD models multiple response modes and uses classifier behavior to identify a functionally relevant mode before measuring response alignment. Finally, the resulting response support is combined with the original model's global recognition support through locally calibrated, response-dependent fusion. This allows cross-client recognition to complement the diagnostic evidence without making local semantic coverage the ID boundary. FedOPhD requires neither OOD data nor further federated communication, while the original global classifier remains unchanged for deployment. Figure 1 illustrates the resulting separation of OOD from both L-ID and T-ID.

The main contributions of this work are summarized as follows:

- We study post-hoc federated OOD detection under a restrictive deployment setting where only a trained global model and current-client ID data remain available, while federation-level ID can extend beyond local semantic coverage.
- We propose FedOPhD, which repurposes semantic attenuation as a diagnostic probe and extracts multi-layer response evidence of learned semantic support, then combines it with global recognition through locally calibrated adaptive fusion.
- We extensively evaluate FedOPhD across heterogeneous federated settings, near- and far-OOD distributions, federation scales, and aggregation algorithms. Despite being constructed only after federated training from the trained global model and current-client ID data, FedOPhD remains competitive with training-time federated OOD detectors on near-OOD detection and surpasses them on both far-OOD metrics across all six main settings.

2 Related Work

Post-hoc OOD Detection. Post-hoc OOD detection equips trained models with OOD awareness without modifying the original training procedure (Yang et al., 2022; Lu et al., 2025). Existing methods extract evidence from predictive outputs (Hendrycks & Gimpel, 2017; Liang et al., 2018; Liu et al., 2020), learned representations (Lee et al., 2018; Sun et al., 2021; Wang et al., 2022a; Sun et al., 2022; Müller & Hein, 2025; Demirel et al., 2025), or gradient geometry (Huang et al., 2021; Behpour et al., 2023; Seleznova et al., 2026). Recent approaches further actively probe trained models through input, weight, or feature interventions (Chen et al., 2025; Granz et al., 2024; Gu et al., 2025). These methods provide diverse post-hoc evidence, but typically assume that the available model or ID reference adequately represents the target ID distribution. In federated deployment, however, the global model captures federation-wide semantics whereas a client's post-hoc ID ref-

erence may cover only a local subset. This mismatch motivates evidence that complements global recognition without treating local semantic coverage as the federation-level ID boundary.

Federated OOD Detection. Existing FL-specific OOD methods largely operate during federated training, using client–server collaboration or auxiliary resources to acquire federation- or OOD-aware information. FOSTER exploits cross-client heterogeneity (Yu et al., 2023), FOOGD models decentralized distributions (Liao et al., 2024), and FOCoOp leverages federated prompts (Liao et al., 2025). Other approaches use server-side auxiliary outliers (Jeong & Choi, 2025) or federated representation learning (Huang et al., 2026). In contrast, we study post-hoc federated OOD detection after training, using only the trained global checkpoint and current-client ID data, without further federation access or OOD data.

Machine Unlearning. MU removes designated training information without full retraining (Cao & Yang, 2015; Bourtoule et al., 2021), including recent advances in selective forgetting (Chen et al., 2023; Kurmanji et al., 2023). Federated Unlearning (FU) extends this objective to decentralized contributions (Liu et al., 2021; Fraboni et al., 2024; Wang et al., 2022b; Sheng et al., 2024). Rather than treating forgetting as the desired endpoint, we view selective weakening of learned semantics as a way to expose how an input depends on the model's semantic support. This suggests a diagnostic principle for post-hoc OOD detection: inputs with similar original-model confidence may respond differently when learned support is attenuated, revealing semantic evidence that static predictions alone may not expose.

3 Problem Formulation

FL under client heterogeneity. FL involves K clients indexed by $i \in \{1, \dots, K\}$, which collaboratively train a shared classifier without centralizing their raw data. Client i holds a labeled dataset $\mathcal{D}_i \subseteq \mathcal{X} \times \mathcal{Y}_i$ drawn from a local distribution P_i, where $\mathcal{X}$ is the input space and $\mathcal{Y}_i$ is the client's local label set. Let $\mathcal{Y} = \{1, \dots, C\}$ denote the federation-level label space, with $\bigcup_{i=1}^{K} \mathcal{Y}_i = \mathcal{Y}$. Under class heterogeneity, $\mathcal{Y}_i$ need not cover $\mathcal{Y}$; in particular, the deployment client k may satisfy $\mathcal{Y}_k \subsetneq \mathcal{Y}$. We denote the resulting global classifier by $\mathbf{f}_{\boldsymbol{\theta}_0} : \mathcal{X} \to \mathbb{R}^C$.

Following prior federated OOD work (Liao et al., 2024), we define ID with respect to the federation rather than the deployment client alone. Thus, both L-ID and T-ID are valid federation-level ID, and their distinction reflects client provenance rather than disjoint label supports. In particular, a class $c \in \mathcal{Y} \setminus \mathcal{Y}_k$ remains ID for client k when it is represented by another participating client. Hence, the client's local label coverage $\mathcal{Y}_k$ does not define the federation-level ID boundary.

Post-hoc federated OOD detection. After federated training, we consider detector construction at a deployment client k, with access only to the trained global classifier $\mathbf{f}_{\boldsymbol{\theta}_0} : \mathcal{X} \to \mathbb{R}^C$ and its L-ID data $\mathcal{D}_k$, but not to OOD data, other clients' data, or further client–server communication. The task is semantic OOD detection with respect to the federation-level label space $\mathcal{Y}$: an input is ID if its semantic class belongs to $\mathcal{Y}$, and OOD otherwise. For a fixed client k, we suppress the client index and construct an OOD scoring function $S : \mathcal{X} \to \mathbb{R}$, where larger values indicate stronger OOD evidence. With a threshold τ calibrated solely on L-ID data, a test input x is rejected as OOD if $S(x) > \tau$ and accepted as ID otherwise. An accepted input is classified by the global model as $\hat{y}(x) = \arg\max_{c \in \mathcal{Y}} f_{\boldsymbol{\theta}_0, c}(x)$, where $f_{\boldsymbol{\theta}_0, c}(x)$ denotes the logit for class c.

4 Method

Overview. FedOPhD takes a counterfactual view of federation-wide semantic support: rather than asking whether an input resembles the client's L-ID under a frozen model, we ask whether it exhibits an ID-like response to an unlearning-inspired intervention that selectively weakens locally supported semantics. This turns L-ID data from an incomplete proxy for the federation-wide ID boundary into an intervention for interrogating the knowledge already encoded by the trained global classifier $f_{\boldsymbol{\theta}_0}$, where $\boldsymbol{\theta}_0$ denotes the post-training checkpoint. FedOPhD measures whether the resulting representation response follows locally observed ID response structure, yielding *Response Alignment Support* s_{R}, and complements it with *Global Recognition Support* s_{G} from the unchanged model to retain evidence beyond local semantic coverage. Accordingly, client-conditioned support attenuation constructs the probe, while response alignment and adaptive fusion read out the induced

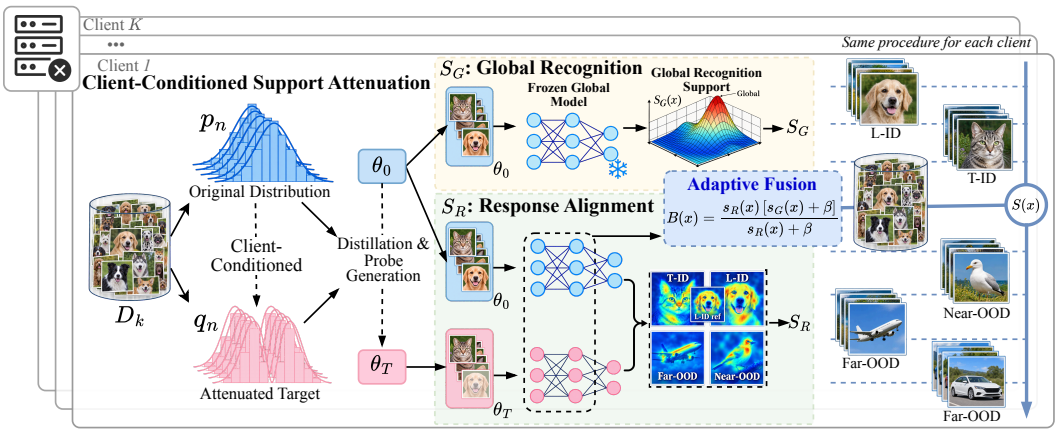

Figure 2: Overview of FedOPhD.

response together with global recognition for OOD scoring. The overall pipeline is summarized in Figure 2.

4.1 Client-Conditioned Support Attenuation

To instantiate the unlearning-inspired probe, we selectively weaken semantics supported by the deployment client and observe how inputs respond. A simple true-class removal is inadequate: uniformly redistributing the removed probability can strengthen other classes already supported by the same client, merely shifting recognition within familiar local semantics. We therefore condition the attenuation target on the client's learned label support, discouraging redistribution among locally supported classes while preserving the remaining structure of each original prediction.

Given an N-sample adaptation subset $\mathcal{D}_{\text{ada}} = \{(x_n, y_n)\}_{n=1}^{N} \subset \mathcal{D}_k$, we freeze the original prediction $\mathbf{p}_n = \text{softmax}(f_{\boldsymbol{\theta}_0}(x_n))$ and its true-label confidence $\omega_n = p_{n,y_n}$. We summarize the semantics supported by the client through the confidence-weighted label profile $\boldsymbol{\pi}_k = (\sum_{n=1}^{N} \omega_n \mathbf{e}_{y_n})/(\sum_{n=1}^{N} \omega_n)$, where $\mathbf{e}_y$ is the one-hot vector for class y. Weighting by ω_n emphasizes labels that are both locally observed and strongly recognized by the trained global model, allowing local data to condition the intervention without treating their label coverage as the federation-wide ID boundary.

For $C \geq 2$, let $\Delta_{-y} = \{\mathbf{q} \geq 0 : \mathbf{1}^\top \mathbf{q} = 1,\ q_y = 0\}$. We construct a fixed attenuation target for each adaptation sample:

$$\mathbf{q}_n = \arg\min_{\mathbf{q} \in \Delta_{-y_n}} \left\{ \tfrac{1}{2} \|\mathbf{q} - \mathbf{p}_n\|_2^2 + \omega_n \boldsymbol{\pi}_k^\top \mathbf{q} \right\}. \tag{1}$$

The first term preserves the sample-specific structure of the original prediction, while the second discourages probability from being assigned to classes with stronger client-level support. Together with $q_{n,y_n} = 0$, the target therefore removes the true-class support without mechanically transferring it to other locally supported classes. The factor ω_n further scales this suppression by the amount of true-class probability being removed, so strongly recognized samples induce correspondingly stronger attenuation.

We then optimize only a diagnostic copy of the model toward these fixed targets:

$$\mathcal{L}_{\text{att}}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{n=1}^{N} D_{\text{KL}}(\mathbf{q}_n \,\|\, \text{softmax}(f_{\boldsymbol{\theta}}(x_n))), \tag{2}$$

where D_{KL} denotes Kullback–Leibler divergence. Starting from $\boldsymbol{\theta}_0$, we apply T updates at attenuation learning rate η with the adaptation inputs and targets fixed, yielding the attenuated endpoint $\boldsymbol{\theta}_T$ while leaving the deployed checkpoint unchanged. Together, $(\boldsymbol{\theta}_0, \boldsymbol{\theta}_T)$ instantiate the unlearning-inspired probe: their representation differences capture how an input responds when locally supported semantics are weakened. The next section develops the corresponding response readout. Target construction and optimization details are provided in Appendix D.2.

4.2 Response Alignment and Adaptive Fusion

Having constructed the attenuation probe, the remaining question is how to turn its induced representation changes into reliable ID evidence. An L-ID input need not respond along a single common direction: different locally supported semantics can produce distinct, yet equally valid, response patterns under the same attenuation. Collapsing them into a single reference direction would therefore obscure this structure. Our key insight is that a useful readout should preserve this response diversity while distinguishing geometrically common changes from those that are actually tied to weakened recognition. We therefore model multiple local response modes and use a two-stage readout: classifier behavior identifies the functionally relevant mode for a query, while its response geometry measures how closely the query follows the corresponding locally observed ID pattern.

Let $\mathcal{D}_{\mathrm{ref}} = \{x_n\}_{n=1}^{n_{\mathrm{ref}}}$ be an L-ID reference subset disjoint from the attenuation subset, and let $\mathbf{h}_t^\ell(x)$ denote the spatially pooled representation at layer ℓ under endpoint $\boldsymbol{\theta}_t$, $t \in \{0, T\}$. We define the attenuation-induced response as $\mathbf{d}_\ell(x) = \mathbf{h}_T^\ell(x) - \mathbf{h}_0^\ell(x)$. To preserve the multiple ID response patterns described above, we fit J unit geometric directions $\boldsymbol{\mu}_{\ell,j}$ by norm-weighted spherical clustering of the reference responses, with assignment $\gamma_n^\ell = \arg\max_j \langle d^\ell(x_n)/\|d^\ell(x_n)\|_2, \mu_{\ell,j}\rangle$ and unit center $\mu_{\ell,j} = \big(\sum_{n:\gamma_n^\ell=j} d^\ell(x_n)\big)/\big\|\sum_{n:\gamma_n^\ell=j} d^\ell(x_n)\big\|_2$. These directions capture recurring local response geometry, but do not yet indicate which geometric mode is most relevant to the attenuation of recognition.

We therefore attach classifier function to the response modes. For each reference input, we fix its original prediction $\hat{y}_n = \arg\max_c f_{\boldsymbol{\theta}_0,c}(x_n)$ and measure the corresponding endpoint margin $m_t(x_n) = f_{\boldsymbol{\theta}_t,\hat{y}_n}(x_n) - (C-1)^{-1}\sum_{c\neq\hat{y}_n} f_{\boldsymbol{\theta}_t,c}(x_n)$, for $t \in \{0, T\}$. A larger decrease $m_0(x_n) - m_T(x_n)$ indicates that the associated representation response is more strongly coupled to the intended weakening of the original recognition. Within each geometric mode, we use these decreases to form the functional routing direction

$$\mathbf{b}_{\ell,j} = \frac{\sum_{n:\gamma_n^\ell=j}\big[m_0(x_n) - m_T(x_n)\big]\mathbf{d}_\ell(x_n)}{\Big\|\sum_{n:\gamma_n^\ell=j}\big[m_0(x_n) - m_T(x_n)\big]\mathbf{d}_\ell(x_n)\Big\|_2}. \tag{3}$$

A query is routed to $j_\ell^*(x) = \arg\max_j \langle \mathbf{d}_\ell(x)/\|\mathbf{d}_\ell(x)\|_2, \mathbf{b}_{\ell,j}\rangle$, but its agreement is evaluated against the corresponding geometric direction $\boldsymbol{\mu}_{\ell,j_\ell^*(x)}$. Thus, classifier behavior determines *which* locally observed response pattern is functionally relevant, while response geometry determines *how closely* the query follows that ID pattern.

Directional agreement alone can nevertheless be misleading when the probe induces only a negligible displacement: the direction of a small response is inherently sensitive to small absolute changes. We therefore require directional evidence to be supported by sufficient response strength. Let $\rho_\ell = [n_{\mathrm{ref}}^{-1}\sum_n \|\mathbf{h}_0^\ell(x_n)\|_2^2]^{1/2}$ denote the reference representation scale at layer ℓ, and $\mathcal{I}$ denote the set of selected response stages. The layer-wise readout and resulting *Response Alignment Support* are

$$\begin{aligned} a_\ell(x) &= \left\langle \frac{\mathbf{d}_\ell(x)}{\max\{\|\mathbf{d}_\ell(x)\|_2, \eta\rho_\ell\}}, \boldsymbol{\mu}_{\ell,j_\ell^*(x)} \right\rangle, \\ s_{\mathrm{R}}(x) &= \left[\prod_{\ell\in\mathcal{I}} \widehat{F}_\ell(a_\ell(x))\right]^{1/|\mathcal{I}|}, \end{aligned} \tag{4}$$

where $\eta > 0$ is the same learning rate used for the attenuation updates and sets the layer-relative normalization scale, and $\widehat{F}_\ell$ is the empirical rank map fitted on $\mathcal{D}_{\mathrm{ref}}$. When the response is weak, its directional evidence is attenuated by its magnitude; once sufficiently excited, a_ℓ reduces to cosine alignment with the selected response mode. The rank maps place different layers on a common local-reference scale before aggregation. Their construction and fitting are detailed in Appendix D.5.

Importantly, s_{R} remains referenced to L-ID behavior and therefore should not by itself define federation-level ID. The unchanged global model provides the complementary evidence missing from this local reference: recognition learned across participating clients. We instantiate *Global Recognition Support* using negative energy (Liu et al., 2020), $a_{\mathrm{G}}(x) = \log\sum_{c=1}^{C}\exp(f_{\boldsymbol{\theta}_0,c}(x))$, and locally calibrate it as $s_{\mathrm{G}}(x) = \widehat{F}_{\mathrm{G}}(a_{\mathrm{G}}(x))$. Thus, s_{R} measures whether the query exhibits an

ID-like response to the probe, whereas s_{G} preserves recognition evidence beyond the deployment client's semantic coverage. The rank transforms provide reference-calibrated support scales rather than ID posterior probabilities.

Response-dependent fusion. The two supports provide complementary but asymmetric evidence. Response alignment reveals whether a query behaves like L-ID under the probe, but its reference remains semantically incomplete; global recognition captures knowledge learned across clients, but can remain high even for OOD inputs. We therefore let response alignment modulate the contribution of global recognition rather than treating the two supports symmetrically. Weak response alignment limits the contribution of s_{G}, whereas a strong ID-like response allows federation-wide recognition to contribute more strongly.

Let $\beta > 0$ be a client-specific scale fitted solely from local reference responses. We define the fused ID support as

$$B(x) = \frac{s_{\mathrm{R}}(x)}{s_{\mathrm{R}}(x) + \beta} s_{\mathrm{G}}(x) + \frac{\beta}{s_{\mathrm{R}}(x) + \beta} s_{\mathrm{R}}(x) = \frac{s_{\mathrm{R}}(x)\,[s_{\mathrm{G}}(x) + \beta]}{s_{\mathrm{R}}(x) + \beta}. \tag{5}$$

The coefficient on global recognition, $s_{\mathrm{R}}(x)/(s_{\mathrm{R}}(x) + \beta)$, makes the gating role of the response explicit: stronger response alignment grants greater influence to s_{G}, while weak alignment suppresses it. Equivalently, $B(x)$ can be written as $s_{\mathrm{G}}(x)^{\lambda(x)} s_{\mathrm{R}}(x)^{1-\lambda(x)}$ for a query-dependent $\lambda(x) \in (0, 1)$; its closed form and the reference fitting of β are given in Appendix D.7.

Finally, we set $S(x) = -B(x)$ so that larger values indicate stronger OOD evidence, consistent with Section 3. All fitted quantities and the deployment threshold use only L-ID data; calibration and numerical details are provided in Appendix D.

4.3 Theoretical Analysis

We characterize why semantic attenuation provides a structured diagnostic excitation relative to forward training: it remains non-vanishing on confident samples and is directed by client-specific semantic support.

Theorem 4.1 (Non-saturating client-conditioned attenuation)**.** *Uniformly redistributing the removed true-class mass produces a logit residual of norm $\sqrt{C/(C-1)}\,\omega_n$. By Lemma A.1, our target preserves this base excitation and, whenever it differs from uniform redistribution, adds an orthogonal residual component aligned with the client profile $\boldsymbol{\pi}_k$. It therefore satisfies*

$$\|\mathbf{p}_n - \mathbf{q}_n\|_2 \geq \sqrt{\frac{C}{C-1}}\,\omega_n, \qquad \|\mathbf{p}_n - \mathbf{e}_{y_n}\|_2 \leq \sqrt{2}(1 - \omega_n). \tag{6}$$

Consequently, as ω_n approaches one, forward cross-entropy forcing vanishes while semantic attenuation remains bounded away from zero; the forward cross-entropy residual under any uniformly bounded scalar sample weight has the same vanishing behavior.

Thus, Theorem 4.1 formalizes the non-saturating and client-profile-directed excitation provided by semantic attenuation relative to forward training. Appendix A further shows that, compared with a norm-matched isotropic parameter perturbation, the attenuation update is concentrated in directions induced by the local adaptation data. The appendix also provides the complete proofs and additional theoretical results for response readout, adaptive fusion, and conditional transfer.

Moreover, in the small-step regime, Appendix A.3 shows that the query response is governed to first order by the cross-Jacobian coupling between the query representation and client-conditioned attenuation gradients. This motivates interpreting response alignment as a readout of whether a query follows semantic tangent directions induced by locally supported ID.

5 Experiments

Datasets. We evaluate semantic OOD detection on CIFAR-10 and CIFAR-100 (Krizhevsky, 2009). Following prior work (Liao et al., 2024), we use SVHN (Netzer et al., 2011), iSUN (Xu et al., 2015), LSUN-Crop, LSUN-Resize (Yu et al., 2015), and Textures (Cimpoi et al., 2014) for

far-OOD evaluation. For near-OOD evaluation, we additionally use the other CIFAR dataset and TinyImageNet with the saved benchmark image lists (Yang et al., 2022).

Comparison Methods. We compare FedOPhD with centralized post-hoc detectors adapted to federated deployment: FedEnergy (Liu et al., 2020), FedViM (Wang et al., 2022a), FedNCI (Liu & Qin, 2025), FedGradPCA (Seleznova et al., 2026), and FedRAS (Guglielmo & Masana, 2026). We also compare with training-time federated OOD methods FOSTER (Yu et al., 2023), FOOGD (Liao et al., 2024), and FedSDWC (Huang et al., 2026). The post-hoc methods share the same FedAvg checkpoint and complete deployment-client L-ID data, whereas the training-time methods use their own trained models. Additional perturbation-based comparisons are provided in Appendix E.2.6.

Evaluation Metrics. We report AUROC and FPR95 using federation-wide held-out ID; higher AUROC and lower FPR95 indicate better performance. With OOD as the positive class, FPR95 is the fraction of negative ID inputs misclassified as OOD at a threshold detecting at least 95% of OOD. Except for explicitly identified single-run diagnostics, results are means over five random seeds with sample standard deviations across seeds. Further evaluation details are provided in Appendix D.9.

Implementation Details. We use CIFAR-style ResNet-18 (He et al., 2016) with 20 clients and class-wise Dirichlet partitioning with concentration α. Unless otherwise stated, FedOPhD uses $J = 16$ reference modes and $T = 5$ semantic-attenuation updates with learning rate $\eta = 10^{-4}$, and reads attenuation-induced responses from intermediate feature-extraction stages, excluding the final representation stage and classification head from the response readout. Complete training and detector settings are provided in Appendix D. Error bars indicate ± 1 sample standard deviation across five seeds. FedOPhD uses disjoint client-local subsets for attenuation and reference fitting, while the deployment threshold is calibrated on a separately held-out L-ID subset; no OOD or test sample is used for detector construction or calibration.

5.1 Experimental Results

Performance comparison across federated settings. Our goal is to approach or surpass the OOD-detection performance of training-time federated OOD detectors using only a trained global model and current-client ID data, without post-training communication or access to other clients' data. Table 1 shows that FedOPhD outperforms all evaluated training-time federated OOD detectors on both metrics in all six far-OOD settings, while remaining competitive with these detectors on near-OOD detection. Among post-hoc methods, FedOPhD leads both far-OOD metrics in all six settings and achieves the highest near-OOD AUROC in all six and the lowest near-OOD FPR95 in five. Full FedOPhD outperforms both the response-only variant FedOPhD-R and the global-recognition baseline FedEnergy across all reported settings.

Table 1: Near- and far-OOD detection results (%; mean with standard deviation in parentheses).

Method	CIFAR-10						CIFAR-100					
	$\alpha = 0.1$		$\alpha = 0.5$		$\alpha = 10$		$\alpha = 0.1$		$\alpha = 0.5$		$\alpha = 10$	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Near-OOD												
FedEnergy	83.73 (1.18)	50.86 (1.42)	86.99 (1.06)	42.74 (1.53)	87.28 (1.15)	46.56 (1.49)	72.05 (1.01)	72.27 (1.29)	75.51 (1.05)	65.40 (1.16)	75.68 (2.08)	64.40 (2.12)
FedViM	55.03 (3.71)	72.60 (3.06)	76.22 (3.78)	63.03 (3.18)	86.89 (2.08)	43.13 (2.24)	61.90 (3.03)	81.85 (3.06)	67.97 (2.03)	73.63 (2.06)	70.09 (2.12)	69.59 (2.10)
FedNCI	75.42 (1.36)	63.90 (2.76)	81.22 (2.08)	50.19 (2.74)	85.96 (2.06)	52.90 (2.40)	70.37 (2.03)	73.08 (2.22)	74.22 (2.04)	66.98 (2.19)	76.79 (2.22)	62.28 (2.22)
FedGradPCA	72.57 (1.87)	65.68 (2.38)	83.48 (2.30)	50.82 (1.41)	89.10 (2.34)	41.21 (2.29)	58.51 (2.61)	83.55 (2.63)	64.28 (2.72)	80.80 (2.47)	69.23 (2.57)	76.14 (2.53)
FedRAS	80.90 (1.67)	59.44 (2.78)	83.77 (1.06)	43.13 (1.13)	86.10 (2.05)	47.85 (1.59)	72.93 (2.04)	65.98 (2.13)	73.54 (2.03)	65.33 (2.24)	73.95 (1.07)	64.63 (1.26)
FOSTER	79.41 (3.13)	52.89 (3.48)	83.70 (2.23)	47.69 (2.54)	85.13 (2.07)	44.13 (2.84)	74.39 (2.18)	72.59 (2.46)	74.58 (2.38)	65.18 (2.77)	77.62 (1.14)	64.34 (1.32)
FOOGD	**86.37** (2.13)	44.01 (1.51)	**90.00** (2.08)	**34.96** (2.38)	89.28 (1.07)	39.93 (1.45)	**77.09** (2.16)	**62.59** (2.67)	**79.07** (2.15)	**60.41** (2.51)	79.30 (2.32)	59.15 (2.41)
FedSDWC	85.57 (2.23)	45.37 (2.23)	89.02 (1.37)	37.33 (1.37)	88.81 (1.16)	39.85 (1.32)	76.27 (2.28)	64.93 (2.45)	76.21 (2.22)	63.62 (1.57)	76.21 (2.23)	62.68 (2.51)
FedOPhD-R	64.71 (1.38)	84.85 (1.75)	65.05 (1.66)	80.16 (2.63)	69.88 (1.73)	78.48 (1.98)	58.15 (1.88)	85.97 (2.44)	58.84 (1.93)	84.88 (2.35)	59.21 (1.75)	84.01 (2.47)
FedOPhD (ours)	86.03 (1.18)	**43.96** (1.31)	89.15 (1.05)	39.30 (1.13)	**89.52** (1.06)	**39.57** (1.31)	76.79 (2.21)	67.45 (1.67)	78.72 (1.12)	61.72 (1.58)	**79.58** (1.09)	**58.71** (1.57)
Far-OOD												
FedEnergy	85.14 (2.36)	45.22 (2.63)	89.09 (1.16)	30.09 (1.31)	90.06 (1.17)	28.69 (1.64)	78.09 (2.12)	60.16 (2.66)	77.94 (1.08)	57.19 (1.29)	78.59 (1.31)	59.44 (1.52)
FedViM	63.13 (2.97)	62.08 (3.23)	87.78 (2.41)	35.90 (2.46)	92.28 (3.09)	26.06 (3.38)	71.06 (2.16)	67.48 (2.31)	71.98 (2.24)	65.33 (2.21)	73.89 (1.23)	61.61 (1.42)
FedNCI	75.88 (2.31)	66.16 (2.59)	90.27 (2.14)	33.06 (2.55)	91.61 (1.10)	29.11 (1.22)	76.45 (2.06)	63.17 (3.33)	79.50 (2.11)	56.13 (2.23)	80.95 (1.26)	55.97 (1.49)
FedGradPCA	75.26 (2.15)	58.79 (2.28)	87.94 (2.34)	32.98 (2.46)	91.06 (2.15)	27.19 (2.36)	71.29 (3.18)	67.15 (2.54)	74.78 (1.16)	64.59 (1.30)	77.77 (1.27)	60.45 (1.00)
FedRAS	82.74 (1.36)	57.38 (1.24)	89.45 (1.29)	32.40 (1.27)	92.28 (1.12)	26.46 (1.53)	80.84 (2.05)	56.71 (1.98)	79.87 (2.12)	54.37 (2.40)	79.77 (1.21)	55.85 (1.31)
FOSTER	90.01 (2.14)	30.64 (2.33)	90.49 (1.13)	29.11 (1.48)	92.33 (1.16)	26.26 (1.22)	83.26 (3.39)	47.85 (3.51)	83.49 (1.26)	47.36 (1.47)	83.88 (1.20)	46.54 (1.33)
FOOGD	91.37 (1.07)	35.67 (1.44)	92.57 (2.04)	27.20 (2.47)	93.72 (1.06)	23.53 (1.51)	84.60 (2.11)	47.07 (3.49)	85.13 (1.18)	45.36 (1.48)	86.96 (1.12)	42.54 (1.56)
FedSDWC	89.41 (2.18)	34.78 (3.54)	90.06 (3.24)	29.24 (2.42)	91.53 (1.26)	27.78 (1.53)	80.20 (2.31)	61.37 (2.63)	83.71 (1.20)	57.01 (1.48)	85.41 (1.15)	54.21 (1.62)
FedOPhD-R	91.79 (1.22)	35.66 (2.78)	94.82 (1.03)	24.86 (2.34)	96.73 (1.08)	16.31 (1.97)	91.04 (1.03)	35.04 (2.28)	92.39 (0.89)	30.16 (1.96)	90.64 (1.34)	38.85 (2.45)
FedOPhD (ours)	**97.13** (0.77)	**13.96** (2.18)	**98.32** (1.07)	**7.66** (2.15)	**98.43** (1.06)	**6.72** (2.04)	**94.06** (1.31)	**23.80** (3.11)	**94.29** (1.51)	**21.87** (3.61)	**93.52** (1.41)	**23.63** (4.14)

Robustness to federated domain shift. To test OOD detection when T-ID differs from L-ID in both class coverage and visual style, we evaluate post-hoc detectors on a 12-client, six-domain DomainNet10 federation (Peng et al., 2019) under FedAvg and FedBN (Li et al., 2021). Figure 3 shows

that FedBN is more challenging for OOD detection than FedAvg, yet FedOPhD achieves the best AUROC and FPR95 among the evaluated post-hoc detectors under both protocols. This supports robustness to combined semantic, appearance, and normalization heterogeneity. Appendix E.3.1 provides details.

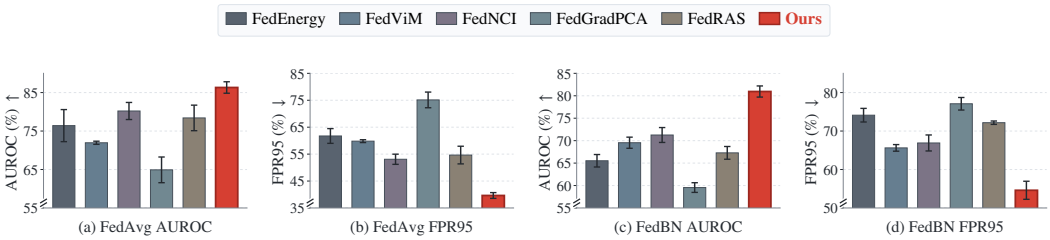

Figure 3: Domain-shift OOD detection on DomainNet10 under FedAvg and FedBN.

Deployment under an L-ID-calibrated threshold. Federation-wide metrics do not directly measure deployment behavior when each client calibrates its threshold using only L-ID. Figure 4 therefore compares T-ID rejection and OOD detection at the locally calibrated threshold under the most heterogeneous setting. FedOPhD preserves T-ID with a small additional rejection while achieving the strongest far-OOD and competitive near-OOD detection; the stronger detection of ViM and GradPCA comes with substantially larger T-ID rejection. Appendix E.1.1 provides the complete protocol and results.

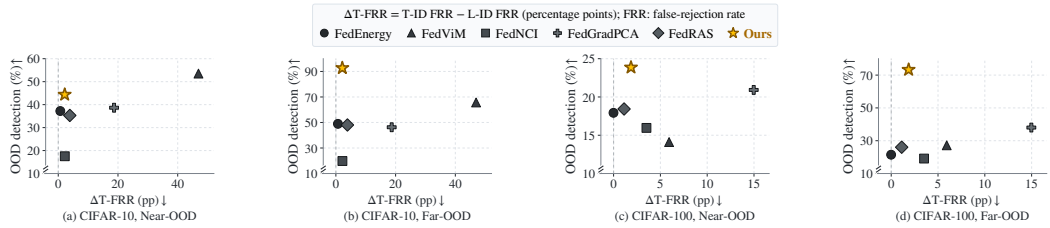

Figure 4: L-ID-calibrated T-ID rejection–OOD detection trade-off at $\alpha = 0.1$.

Fusion-rule comparison. To assess the contribution of the response-dependent fusion design, we compare FedOPhD with equal-weight arithmetic and geometric averaging (AM/GM), GLRT score combining (Reehorst & Schniter, 2025), and the BH-based OOD detection test with conformal p-values (Magesh et al., 2023). Figure 5 evaluates how the fusion rule affects near- and far-OOD detection. FedOPhD achieves the best performance on both metrics in both OOD regimes, supporting its fusion design. Appendix E.2.2 provides the complete results on CIFAR-10 and CIFAR-100.

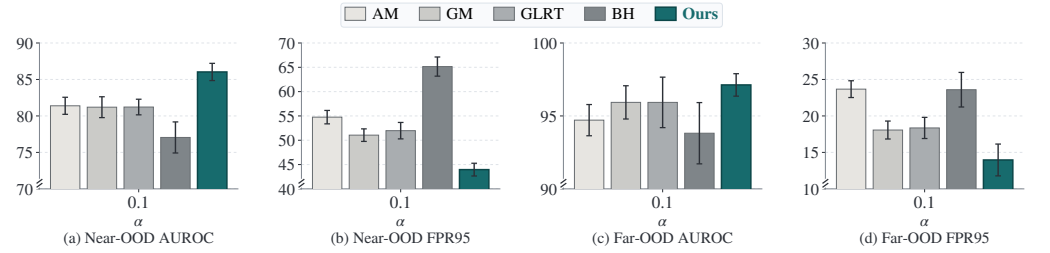

Figure 5: Fusion-rule comparison on CIFAR-10 at $\alpha = 0.1$. Only the fusion rule varies.

Choice of diagnostic intervention. Figure 6 compares semantic attenuation with uniform redistribution, random parameter perturbation, forward training, norm-matched forward, and static-feature controls on CIFAR-10, all within the full fusion pipeline. FedOPhD achieves higher AUROC and lower FPR95 than these controls in the displayed settings, supporting the benefit of semantic attenuation beyond update magnitude and static feature geometry, including after fusion with global recognition. Appendix E.2.8 reports standalone and fused results for the latter controls, with the fused ordering preserved across all three heterogeneity levels.

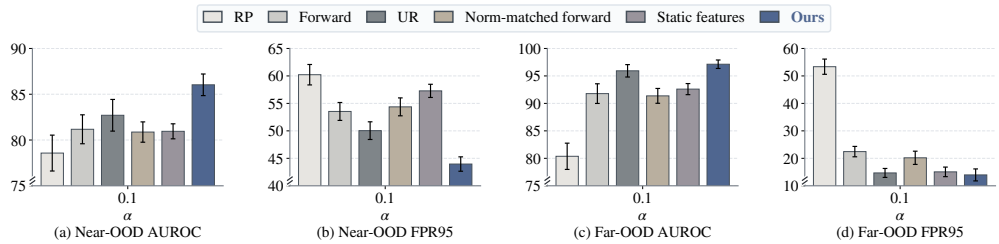

Figure 6: Intervention and static-feature controls on CIFAR-10 at $\alpha = 0.1$. All results include fusion.

Sensitivity to local class coverage. We fix the CIFAR-10 global model trained at $\alpha = 0.1$ and vary the number of locally available classes N_{cls} only during post-training detector construction. Figure 7 shows that FedOPhD generally benefits from broader local class coverage in both OOD regimes, with some non-monotonic variation. It achieves the lowest near-OOD FPR95 throughout and the highest near-OOD AUROC for $N_{\text{cls}} \geq 2$, while remaining strongest on both far-OOD metrics across the tested range. Appendix E.1.3 reports the corresponding deployment statistics. Appendix E provides the numerical results and further sensitivity analyses.

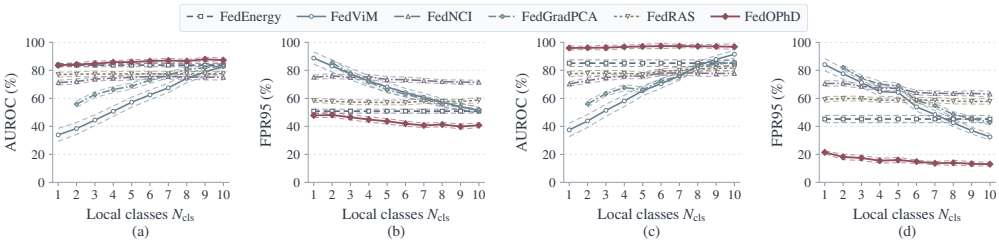

Figure 7: Effect of local class coverage on CIFAR-10 OOD detection at $\alpha = 0.1$. (a,b) Near-OOD; (c,d) Far-OOD.

5.2 Additional Comparisons and Robustness

Beyond the baselines in Table 1, we compare FedOPhD with recent intervention-based post-hoc OOD detectors spanning weight, input, and feature interventions, including WeiPer+KLD (Granz et al., 2024), PRO (Chen et al., 2025), and Gradient Short-Circuit (Gu et al., 2025). FedOPhD achieves the highest mean AUROC and lowest mean FPR95 across both ID datasets, all three heterogeneity levels, and both near- and far-OOD regimes; complete results are reported in Appendix E.2.6.

We further evaluate federation sizes from 10 to 150 clients, reduced client-local data budgets, alternative federated optimization algorithms FedProx and FedAvgM (Li et al., 2020; Hsu et al., 2019), and multiple backbones, including VGG, ResNet, Wide ResNet, and DenseNet (Simonyan & Zisserman, 2015; He et al., 2016; Zagoruyko & Komodakis, 2016; Huang et al., 2017). Finally, we extend the evaluation to TinyImageNet (Le & Yang, 2015) as a 200-class ID dataset, transferring the detector configuration from the CIFAR experiments without OOD-specific retuning. Complete robustness results are provided in Appendix E.3.

6 Conclusion

We presented FedOPhD for post-hoc federated OOD detection using only a trained global model and current-client ID data. FedOPhD combines unlearning-inspired semantic-attenuation responses with global recognition, without OOD data, further federated communication, or changes to the deployed classifier. Our analysis characterizes the client-conditioned excitation, and experiments demonstrate effective near- and far-OOD detection, T-ID preservation, and robustness to domain shift and limited local semantic coverage. These findings support controlled post-hoc interventions as a way to probe distributed knowledge already encoded in a trained model under restricted data and communication access.

AI USE STATEMENT

In this work, we used generative AI tools to edit the manuscript to improve its readability and language. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The manuscript provides detailed training and detector-construction procedures, including data partitions, preprocessing, model architectures, optimization settings, hyperparameters, and random seeds. It also provides the definitions and implementation details of the attenuation targets, response readout, reference fitting, adaptive fusion, and deployment calibration. Evaluation datasets, metrics, aggregation rules, and comparison protocols are described explicitly to support reproduction of the reported experimental results. Theoretical assumptions and derivations are stated together with the corresponding proofs.

REFERENCES

Sima Behpour, Thang Long Doan, Xin Li, Wenbin He, Liang Gou, and Liu Ren. GradOrth: A simple yet efficient out-of-distribution detection with orthogonal projection of gradients. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159. IEEE, 2021.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pp. 463–480. IEEE, 2015.

Min Chen, Weizhuo Gao, Gaoyang Liu, Kai Peng, and Chen Wang. Boundary unlearning: Rapid forgetting of deep networks via shifting the decision boundary. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 7766–7775, 2023.

Wenxi Chen, Raymond A. Yeh, Shaoshuai Mou, and Yan Gu. Leveraging perturbation robustness to enhance out-of-distribution detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4724–4733, 2025. doi: 10.1109/CVPR52734.2025.00445.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

Berker Demirel, Marco Fumero, and Francesco Locatello. Out-of-distribution detection with relative angles. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-3125.

François Fleuret. Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research*, 5:1531–1555, 2004.

Yann Fraboni, Martin Van Waerebeke, Kevin Scaman, Richard Vidal, Laetitia Kameni, and Marco Lorenzi. Sifu: Sequential informed federated unlearning for efficient and provable client unlearning in federated optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 3457–3465. PMLR, 2024.

Maximilian Granz, Manuel Heurich, and Tim Landgraf. Weiper: Ood detection using weight perturbations of class projections. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 35879–35908. Curran Associates, Inc., 2024. doi: 10.52202/079017-1131.

Jiawei Gu, Ziyue Qiao, and Zechao Li. Gradient short-circuit: Efficient out-of-distribution detection via feature intervention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 457–466, 2025.

Gianluca Guglielmo and Marc Masana. Ranked activation shift for post-hoc out-of-distribution detection. In *European Conference on Computer Vision*, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

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.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017. doi: 10.1109/CVPR.2017.243.

Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Zhenyuan Huang, Hui Zhang, Wenzhong Tang, and Haijun Yang. FedSDWC: Federated synergistic dual-representation weak causal learning for OOD. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Gu-Bon Jeong and Dong-Wan Choi. Out-of-distribution detection via outlier exposure in federated learning. *Neural Networks*, 185:107141, 2025. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2025.107141.

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

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. In *Adv. Neural Inf. Process. Syst.*, volume 36, pp. 1957–1987, 2023.

Ya Le and Xuan Yang. Tiny ImageNet visual recognition challenge. CS231N Course Project, Stanford University, 2015. URL `https://cs231n.stanford.edu/reports/2015/pdfs/yle_project.pdf`.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In I. Dhillon, D. Papailiopoulos, and V. Sze (eds.), *Proceedings of Machine Learning and Systems*, volume 2, pp. 429–450, 2020.

Xiaoxiao Li, Meirui JIANG, Xiaofei Zhang, Michael Kamp, and Qi Dou. FedBN: Federated learning on non-IID features via local batch normalization. In *International Conference on Learning Representations*, 2021.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

Xinting Liao, Weiming Liu, Pengyang Zhou, Fengyuan Yu, Jiahe Xu, Jun Wang, Wenjie Wang, Chaochao Chen, and Xiaolin Zheng. Foogd: Federated collaboration for both out-of-distribution generalization and detection. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 132908–132945. Curran Associates, Inc., 2024. doi: 10.52202/079017-4225.

Xinting Liao, Weiming Liu, Jiaming Qian, Pengyang Zhou, Jiahe Xu, Wenjie Wang, Chaochao Chen, Xiaolin Zheng, and Tat-Seng Chua. Focoop: enhancing out-of-distribution robustness in federated prompt learning for vision-language models. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org, 2025.

Gaoyang Liu, Xiaoqiang Ma, Yang Yang, Chen Wang, and Jiangchuan Liu. Federaser: Enabling efficient client-level data removal from federated learning models. In *2021 IEEE/ACM 29th international symposium on quality of service (IWQOS)*, pp. 1–10. IEEE, 2021.

Litian Liu and Yao Qin. Detecting out-of-distribution through the lens of neural collapse. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15424–15433, 2025.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, pp. 21464–21475. Curran Associates, Inc., 2020.

Xixi Liu, Yaroslava Lochman, and Christopher Zach. Gen: Pushing the limits of softmax-based out-of-distribution detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23946–23955, 2023. doi: 10.1109/CVPR52729.2023.02293.

Shuo Lu, Yingsheng Wang, Lijun Sheng, Lingxiao He, Aihua Zheng, and Jian Liang. Out-of-distribution detection: A task-oriented survey of recent advances. *ACM Computing Surveys*, 58 (2):1–39, 2025.

Akshayaa Magesh, Venugopal V. Veeravalli, Anirban Roy, and Susmit Jha. Principled out-of-distribution detection via multiple testing. *Journal of Machine Learning Research*, 24(378):1–35, 2023.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1614–1623. PMLR, 2016.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

Maximilian Müller and Matthias Hein. Mahalanobis++: improving ood detection via feature normalization. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org, 2025.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1406–1415. IEEE, 2019.

Edward T Reehorst and Philip Schniter. Score combining for contrastive ood detection. *arXiv preprint arXiv:2501.12204*, 2025.

Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. In *International conference on machine learning*, pp. 8491–8501. PMLR, 2020.

Mariia Seleznova, Hung-Hsu Chou, Claudio Mayrink Verdun, and Gitta Kutyniok. GradPCA: Leveraging NTK alignment for reliable out-of-distribution detection. In *International Conference on Learning Representations*, 2026.

Xinyi Sheng, Wei Bao, and Liming Ge. Robust federated unlearning. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 2034–2044, 2024.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Yiyou Sun, Chuan Guo, and Yixuan Li. ReAct: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems*, 2021.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International conference on machine learning*, pp. 20827–20840. PMLR, 2022.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. ViM: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022a.

Junxiao Wang, Song Guo, Xin Xie, and Heng Qi. Federated unlearning via class-discriminative pruning. In *Proceedings of the ACM web conference 2022*, pp. 622–632, 2022b.

Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. OpenOOD: Benchmarking generalized out-of-distribution detection. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2022.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Shuyang Yu, Junyuan Hong, Haotao Wang, Zhangyang Wang, and Jiayu Zhou. Turning the curse of heterogeneity in federated learning into a blessing for out-of-distribution detection. In *The Eleventh International Conference on Learning Representations*, 2023.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 87.1–87.12. BMVA Press, September 2016. ISBN 1-901725-59-6. doi: 10.5244/C.30.87.

Chenhao Zhang, Weitong Chen, Wei Emma Zhang, and Miao Xu. Countering relearning with perception revising unlearning. In Vu Nguyen and Hsuan-Tien Lin (eds.), *Proceedings of the 16th Asian Conference on Machine Learning*, volume 260 of *Proceedings of Machine Learning Research*, pp. 1336–1351. PMLR, 2025.

Jinsong Zhang, Qiang Fu, Xu Chen, Lun Du, Zelin Li, Gang Wang, xiaoguang Liu, Shi Han, and Dongmei Zhang. Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KkazG4lgKL.

A THEORY AND PROOFS

This appendix completes the analysis from semantic attenuation to the final detector. We first expose the client-profile component of the attenuation target and prove Theorem 4.1. We then connect the attenuation update to the query response, establish the finite-gain readout, and propagate these properties through fusion and T-ID response coverage.

A.1 CLIENT-PROFILE STRUCTURE OF THE ATTENUATION TARGET

For a sample (x_n, y_n), define the uniform-redistribution target

$$q^{\mathrm{U}}_{n,y_n} = 0, \qquad q^{\mathrm{U}}_{n,c} = p_{n,c} + \frac{\omega_n}{C-1}, \quad c \neq y_n. \tag{7}$$

Lemma A.1 (Profile-aligned lateral attenuation)**.** *The target in Equation 1 is the unique projection* $\mathbf{q}_n = \Pi_{\Delta_{-y_n}}(\mathbf{p}_n - \omega_n \boldsymbol{\pi}_k)$ *and satisfies*

$$\omega_n \boldsymbol{\pi}_k^\top (\mathbf{q}^{\mathrm{U}}_n - \mathbf{q}_n) \geq \|\mathbf{q}_n - \mathbf{q}^{\mathrm{U}}_n\|_2^2. \tag{8}$$

Moreover, the attenuation residual has the orthogonal decomposition

$$\|\mathbf{p}_n - \mathbf{q}_n\|_2^2 = \frac{C}{C-1}\omega_n^2 + \|\mathbf{q}_n - \mathbf{q}^{\mathrm{U}}_n\|_2^2. \tag{9}$$

Thus, uniform redistribution gives the base true-class-removal component, and the client-conditioned projection adds the orthogonal residual component $\mathbf{q}^{\mathrm{U}}_n - \mathbf{q}_n$*, which is aligned with the client profile.*

Proof. Completing the square rewrites Equation 1 as the projection of $\mathbf{p}_n - \omega_n \boldsymbol{\pi}_k$ onto Δ_{-y_n}. The projection is unique because this is a closed convex set and the squared distance is strictly convex.

To compare this target with uniform redistribution, use $\mathbf{q}^{\mathrm{U}}_n$ as a feasible point in the projection inequality:

$$\left\langle \mathbf{q}_n - (\mathbf{p}_n - \omega_n \boldsymbol{\pi}_k), \mathbf{q}^{\mathrm{U}}_n - \mathbf{q}_n \right\rangle \geq 0. \tag{10}$$

Let $\boldsymbol{\delta}_n = \mathbf{q}_n - \mathbf{q}^{\mathrm{U}}_n$. Both targets have zero true-class coordinate and sum to one, so $\boldsymbol{\delta}_n$ sums to zero over the non-true classes. Since $\mathbf{q}^{\mathrm{U}}_n - \mathbf{p}_n$ is constant on those coordinates, their inner product is zero. The inequality therefore reduces to

$$\omega_n \boldsymbol{\pi}_k^\top (\mathbf{q}^{\mathrm{U}}_n - \mathbf{q}_n) \geq \|\boldsymbol{\delta}_n\|_2^2, \tag{11}$$

which is Equation 8. The key point is that the additional residual component $\mathbf{q}^{\mathrm{U}}_n - \mathbf{q}_n$ is positively aligned with the client-support profile.

For the norm identity, decompose the residual as $\mathbf{p}_n - \mathbf{q}_n = (\mathbf{p}_n - \mathbf{q}^{\mathrm{U}}_n) + (\mathbf{q}^{\mathrm{U}}_n - \mathbf{q}_n)$. The same zero-sum argument makes the two terms orthogonal, and the base component has norm

$$\|\mathbf{p}_n - \mathbf{q}^{\mathrm{U}}_n\|_2^2 = \omega_n^2 + (C-1)\left(\frac{\omega_n}{C-1}\right)^2 = \frac{C}{C-1}\omega_n^2. \tag{12}$$

Pythagoras now gives Equation 9: client conditioning adds energy orthogonally to the common true-class-removal component. □

A.2 PROOF OF THEOREM 4.1

Proof. Equation 9 separates the attenuation residual into a base component of norm $\sqrt{C/(C-1)}\,\omega_n$ and the orthogonal client-profile component from Lemma A.1. Keeping only the base component gives the first bound in Equation 6.

For comparison, the forward cross-entropy residual satisfies

$$\|\mathbf{p}_n - \mathbf{e}_{y_n}\|_2^2 = (1-\omega_n)^2 + \sum_{c \neq y_n} p_{n,c}^2 \leq 2(1-\omega_n)^2, \tag{13}$$

where the inequality uses $\sum_{c \neq y_n} p_{n,c} = 1 - \omega_n$. If $\omega_n \geq 1 - \varepsilon$, the semantic-attenuation residual norm is at least $\sqrt{C/(C-1)}(1-\varepsilon)$, while the forward residual norm is at most $\sqrt{2}\varepsilon$. The same vanishing bound holds when the forward cross-entropy residual is multiplied by any scalar sample weight $0 \leq w_n \leq M$ with fixed M. This is the non-saturating effect: confident samples continue to inject an attenuation signal as the ordinary forward residual shrinks.

We next locate this signal in parameter space. Let P be the number of updated parameters and let $J_{\mathbf{f}}(x;\boldsymbol{\theta}) = \partial \mathbf{f}_{\boldsymbol{\theta}}(x)/\partial \boldsymbol{\theta} \in \mathbb{R}^{C \times P}$ denote the logit Jacobian with respect to these parameters. Define the local-data tangent span

$$\mathcal{U}_k = \operatorname{span}\{J_{\mathbf{f}}(x_n;\boldsymbol{\theta}_0)^\top \mathbf{v} : \mathbf{v} \in \mathbb{R}^C,\ \mathbf{1}^\top \mathbf{v} = 0\}_{n=1}^N, \qquad r_k = \dim(\mathcal{U}_k), \tag{14}$$

where $\Pi_{\mathcal{U}_k}$ denotes the orthogonal projector onto $\mathcal{U}_k$ and the samples are from $\mathcal{D}_{\mathrm{ada}}$. Every probability residual $\mathbf{p}_n - \mathbf{q}_n$ sums to zero, so

$$\mathbf{g}_0 = \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{att}}(\boldsymbol{\theta}_0) = \frac{1}{N} \sum_{n=1}^N J_{\mathbf{f}}(x_n;\boldsymbol{\theta}_0)^\top (\mathbf{p}_n - \mathbf{q}_n), \tag{15}$$

and therefore $\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0 = -\eta \mathbf{g}_0 \in \mathcal{U}_k$. Each sample contributes at most $C-1$ independent zero-sum directions, giving $r_k = \dim(\mathcal{U}_k) \leq \min\{P, N(C-1)\}$.

Let $\boldsymbol{\xi}$ be a norm-matched isotropic random step, so that $\|\boldsymbol{\xi}\|_2 = \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\|_2$. Isotropy distributes its expected squared length equally over the P parameter coordinates; projecting onto an r_k-dimensional span therefore gives

$$\mathbb{E}[\|\Pi_{\mathcal{U}_k} \boldsymbol{\xi}\|_2^2 \mid \mathcal{D}_{\mathrm{ada}}] = \frac{r_k}{P} \|\boldsymbol{\xi}\|_2^2. \tag{16}$$

Thus, the first attenuation step lies entirely in directions generated by the client's examples, whereas a norm-matched isotropic step places the fraction r_k/P of its squared norm in this span in expectation. □

A.3 Finite-step response and finite-gain readout

For differentiation, write $\mathbf{h}^\ell(x;\boldsymbol{\theta})$ for the layer representation as a function of the updated parameters, so that $\mathbf{h}^\ell(x;\boldsymbol{\theta}_t) = \mathbf{h}_t^\ell(x)$. Its parameter Jacobian is $J_{\mathbf{h}_\ell}(x;\boldsymbol{\theta}) = \partial \mathbf{h}^\ell(x;\boldsymbol{\theta})/\partial \boldsymbol{\theta}$, with one row per representation coordinate and P columns. Throughout this analysis, the normalization uses the same attenuation learning rate η as the parameter updates.

Proposition A.2 (Semantic tangent response with finite-gain readout)**.** *Assume the attenuation gradient and the parameter Jacobian $J_{\mathbf{h}_\ell}$ of the layer representation are bounded and Lipschitz on a convex neighborhood containing the T-step trajectory, and let $\rho_\ell > 0$. For fixed T and $\eta \to 0$, the response has the cross-Jacobian form*

$$\begin{aligned} \frac{\mathbf{d}_\ell(x)}{\eta} = -\frac{T}{N} \sum_{n=1}^N & J_{\mathbf{h}_\ell}(x;\boldsymbol{\theta}_0) \\ & \cdot J_{\mathbf{f}}(x_n;\boldsymbol{\theta}_0)^\top (\mathbf{p}_n - \mathbf{q}_n) + \mathcal{O}(\eta). \end{aligned} \tag{17}$$

The leading term measures how strongly the query representation couples to attenuation gradients generated by the client's local adaptation examples.

The implemented response normalization further satisfies, for any $\mathbf{d}, \mathbf{d}'$,

$$\left\| \frac{\mathbf{d}}{\max\{\|\mathbf{d}\|_2, \eta\rho_\ell\}} - \frac{\mathbf{d}'}{\max\{\|\mathbf{d}'\|_2, \eta\rho_\ell\}} \right\|_2 \leq \frac{\|\mathbf{d} - \mathbf{d}'\|_2}{\eta\rho_\ell}. \tag{18}$$

For a fixed selected mode, the layer score therefore factors as

$$a_\ell(x) = \min\left\{1, \frac{\|\mathbf{d}_\ell(x)\|_2}{\eta\rho_\ell}\right\} \left\langle \frac{\mathbf{d}_\ell(x)}{\|\mathbf{d}_\ell(x)\|_2}, \boldsymbol{\mu}_{\ell, j_\ell^*(x)} \right\rangle, \tag{19}$$

with the zero-response value defined as zero. Thus, the cross-Jacobian supplies a client-conditioned query response, and the finite-gain readout limits the directional evidence of weak responses while preserving the angular geometry of sufficiently excited responses.

Proof. Let $\mathbf{g}_0 = N^{-1}\sum_{n=1}^{N} J_{\mathbf{f}}(x_n;\boldsymbol{\theta}_0)^\top(\mathbf{p}_n - \mathbf{q}_n)$. Gradient smoothness gives $\boldsymbol{\theta}_T - \boldsymbol{\theta}_0 = -T\eta\mathbf{g}_0 + \mathcal{O}(\eta^2)$: for fixed T, later gradients differ from the initial one by only first-order terms in η. A first-order expansion of the query representation then gives $\mathbf{d}_\ell(x) = J_{\mathbf{h}_\ell}(x;\boldsymbol{\theta}_0)(\boldsymbol{\theta}_T - \boldsymbol{\theta}_0) + \mathcal{O}(\eta^2)$. Substituting the parameter change and dividing by η proves Equation 17; the Jacobian product transfers each local logit residual into the query's representation space.

The normalized vector in Equation 18 is the Euclidean projection of $\mathbf{d}/(\eta\rho_\ell)$ onto the unit ball. Projection onto a closed convex set is non-expansive, which directly gives the finite-gain bound. Finally, separating the projected vector into its norm and direction gives Equation 19. □

The J modes retain multiple local directions in the reference response distribution, providing a compact multimodal representation for routing and scoring. For a fixed mode whose weighted response sum is nonzero, the functional routing direction in Equation 3 is equivalently the unit vector maximizing

$$\sum_{n:\gamma_n^\ell = j} [m_0(x_n) - m_T(x_n)]\langle \mathbf{b}, \mathbf{d}_\ell(x_n)\rangle \quad \text{over } \|\mathbf{b}\|_2 \le 1. \tag{20}$$

Here the sum ranges over reference inputs $x_n \in \mathcal{D}_{\text{ref}}$ assigned to mode j at stage ℓ. If the weighted response sum is zero, this objective is zero for every feasible $\mathbf{b}$. The prescribed choice $\mathbf{b}_{\ell,j} = \boldsymbol{\mu}_{\ell,j}$ is therefore also a maximizer and retains a unit routing direction. Hence $m_0 - m_T$ is a signed semantic attenuation amount: responses producing larger decreases of the fixed original-class margin receive greater influence, while margin increases contribute with the opposite sign. The routing direction $\mathbf{b}_{\ell,j}$ selects the mode, and $\boldsymbol{\mu}_{\ell,j}$ measures geometric agreement within it. Within a fixed route, the selected geometric direction is fixed, so the finite-gain bound applies directly to the layer alignment.

A.4 Response-gated fusion

Proposition A.3 (Bounded response-gated compensation)**.** *When the adaptive fusion in Equation 5 is used, for $s_{\mathrm{G}}(x), s_{\mathrm{R}}(x) \in (0,1]$ and $\beta > 0$,*

$$B(x) = s_{\mathrm{R}}(x) + \frac{s_{\mathrm{R}}(x)}{s_{\mathrm{R}}(x) + \beta}\big[s_{\mathrm{G}}(x) - s_{\mathrm{R}}(x)\big], \tag{21}$$

and

$$\frac{\beta}{1+\beta} s_{\mathrm{R}}(x) \le B(x) \le \left(1 + \frac{1}{\beta}\right) s_{\mathrm{R}}(x). \tag{22}$$

For the deployment threshold τ with $0 < -\tau < 1$, the rejection rule $S(x) > \tau$ is equivalent to $B(x) < -\tau$, and

$$s_{\mathrm{R}}(x) < \frac{-\tau\beta}{1+\beta+\tau} \quad \Longrightarrow \quad B(x) < -\tau. \tag{23}$$

This implication holds for every $s_{\mathrm{G}}(x) \in (0,1]$. The arithmetic form shows the mechanism directly: response support controls how far fusion moves from $s_{\mathrm{R}}(x)$ toward $s_{\mathrm{G}}(x)$. The bounds then limit the relative contribution of global recognition, and Equation 23 gives an explicit rejection condition for weak response support.

Proof. Expanding the numerator in Equation 5 gives Equation 21. Next, divide by $s_{\mathrm{R}}(x)$ and use $s_{\mathrm{G}}(x), s_{\mathrm{R}}(x) \in (0,1]$:

$$\frac{\beta}{1+\beta} \le \frac{s_{\mathrm{G}}(x) + \beta}{s_{\mathrm{R}}(x) + \beta} \le \frac{1+\beta}{\beta}. \tag{24}$$

Multiplying by $s_{\mathrm{R}}(x)$ proves Equation 22. At fixed response support, fusion increases with global support and is at most $s_{\mathrm{R}}(x)(1+\beta)/(s_{\mathrm{R}}(x)+\beta)$. Requiring this upper bound to be below $-\tau$ gives Equation 23. □

A.5 Conditional T-ID stability

We pair a T-ID query with an L-ID anchor and track how their normalized response mismatch propagates to the final support.

Proposition A.4 (Conditional transfer under response coverage)**.** *Fix one client's fitted directions, empirical rank maps and β, and use the adaptive fusion in Equation 5. Consider a joint pairing of an L-ID anchor X_{L} and a T-ID query X_{T}. Let E denote the paired cases in which both inputs use the same route at every selected layer and, for $\epsilon_\ell \geq 0$,*

$$\left\| \frac{\mathbf{d}_\ell(X_{\mathrm{T}})}{\max\{\|\mathbf{d}_\ell(X_{\mathrm{T}})\|_2, \eta\rho_\ell\}} - \frac{\mathbf{d}_\ell(X_{\mathrm{L}})}{\max\{\|\mathbf{d}_\ell(X_{\mathrm{L}})\|_2, \eta\rho_\ell\}} \right\|_2 \leq \epsilon_\ell. \tag{25}$$

Let $G_- : \mathcal{X} \to (0, 1]$ be a fixed measurable function of the L-ID anchor, defined on the entire input space, that lower-bounds the paired T-ID global support on E: $s_{\mathrm{G}}(X_{\mathrm{T}}) \geq G_-(X_{\mathrm{L}})$. This lower-bound condition is an assumption on the pairing; G_- is an auxiliary function for the analysis. Define the response-support lower envelope by

$$R_-(X_{\mathrm{L}}) = \left[\prod_{\ell \in \mathcal{I}} \widehat{F}_\ell(a_\ell(X_{\mathrm{L}}) - \epsilon_\ell) \right]^{1/|\mathcal{I}|}. \tag{26}$$

Then, on E,

$$B(X_{\mathrm{T}}) \geq \frac{R_-(X_{\mathrm{L}})[G_-(X_{\mathrm{L}}) + \beta]}{R_-(X_{\mathrm{L}}) + \beta}. \tag{27}$$

For every deployment threshold τ,

$$\begin{aligned} \Pr[B(X_{\mathrm{T}}) < -\tau] \leq{}& \Pr(E^c) \\ &+ \Pr\left[\frac{R_-(X_{\mathrm{L}})[G_-(X_{\mathrm{L}}) + \beta]}{R_-(X_{\mathrm{L}}) + \beta} < -\tau \right]. \end{aligned} \tag{28}$$

Proof. On E, the shared route fixes the same unit geometric direction for both inputs at each layer. Cauchy–Schwarz converts the normalized-response distance in Equation 25 into the alignment bound $a_\ell(X_{\mathrm{T}}) \geq a_\ell(X_{\mathrm{L}}) - \epsilon_\ell$. Each empirical rank map is nondecreasing, so the layer-wise bounds pass through the rank maps and their geometric aggregation: $s_{\mathrm{R}}(X_{\mathrm{T}}) \geq R_-(X_{\mathrm{L}})$. The fusion function increases with both inputs, which gives the pointwise lower bound on $B(X_{\mathrm{T}})$. Finally, splitting the rejection event into coverage failure and lower-envelope rejection gives Equation 28, making the two sources of T-ID rejection explicit. □

A.6 Summary of the theoretical characterization

Taken together, the preceding results characterize semantic attenuation as a structured diagnostic intervention from several complementary perspectives. Relative to forward cross-entropy updates, its logit-level excitation remains non-vanishing as sample confidence increases. Relative to a norm-matched isotropic parameter displacement, its first update is concentrated in directions induced by the local adaptation data. Furthermore, in the small-step regime, the induced query representation response is governed to first order by the coupling between these client-conditioned attenuation directions and the query representation Jacobian. The finite-gain readout controls this response, while adaptive fusion and conditional T-ID transfer propagate its structure to the final detector. Together, these results establish the non-saturating excitation, local-data directionality, bounded response readout, response-gated fusion, and conditional transfer properties of FedOPhD.

B NOTATION

Table 2 summarizes the notation used in the formulation, method and implementation details.

Table 2: Principal notation used by FedOPhD.

Symbol	Meaning	Symbol	Meaning
K	Number of clients	k	Deployment client
$\mathcal{D}_k$	Client-local training set	$\mathcal{Y}_k$	Locally observed classes
$\mathcal{Y}$	Global ID label set	C	Number of global ID classes
$\boldsymbol{\theta}_0$	Original global checkpoint	$\boldsymbol{\theta}_t$	Checkpoint after step t
T	Number of attenuation updates	η	Learning rate, also used in response normalization
$\mathcal{D}_{\mathrm{ada},k}$	Attenuation subset	$\mathcal{D}_{\mathrm{ref},k}$	Reference-fitting subset
$\mathcal{D}_{\mathrm{cal},k}$	Threshold-calibration subset	N	Number of adaptation samples
$\mathbf{p}_n$	Frozen original prediction	ω_n	True-label confidence
$\boldsymbol{\pi}_k$	Client label-support profile	Δ_{-y}	True-label-suppressed simplex
$\mathbf{q}_n$	Fixed attenuation target	$\mathbf{h}_t^\ell(x)$	Layer-ℓ representation
$\mathbf{d}_\ell(x)$	Attenuation-induced response	$\mathcal{I}$	Selected response stages
J	Directions per response stage	$\boldsymbol{\mu}_{\ell,j}$	Geometric reference direction
$\mathbf{b}_{\ell,j}$	Functional routing direction	γ_n^ℓ	Reference-mode assignment
$j_\ell^*(x)$	Query routing index	$a_\ell(x)$	Norm-scaled layer alignment
ρ_ℓ	Reference representation RMS	$a_{\mathrm{G}}(x)$	Original-model log-sum-exp
$\widehat{F}_\ell$	Layer empirical rank map	$\widehat{F}_{\mathrm{G}}$	Global empirical rank map
$s_{\mathrm{R}}(x)$	Response alignment support	$s_{\mathrm{G}}(x)$	Global recognition support
$s_{\mathrm{R},n}^{\mathrm{ref}}$	Reference response support	n_{ref}	Reference-set size
$Q_{0.1}$	Empirical tenth percentile	β	Client-specific fusion scale
$\lambda(x)$	Geometric global weight	$B(x)$	Fused ID support
$w_{\mathrm{G}}^{\mathrm{arith}}$	Arithmetic global coefficient	$w_{\mathrm{R}}^{\mathrm{arith}}$	Arithmetic response coefficient
$S(x)$	Final OOD score	τ	Local rejection threshold
$\hat{y}(x)$	Original-model prediction	N_{cls}	Locally available class count

C ALGORITHM

FedOPhD separates one-time client-local detector construction from query-time scoring. Algorithms 1 and 2 summarize these two stages; Appendix D specifies the numerical conventions and sparse-client handling.

Algorithm 1 Client-local construction of FedOPhD

Input: deployment client k, fixed global checkpoint $\boldsymbol{\theta}_0$, local subsets $\mathcal{D}_{\mathrm{ada},k}$, $\mathcal{D}_{\mathrm{ref},k}$, and $\mathcal{D}_{\mathrm{cal},k}$, response stages $\mathcal{I}$, attenuation steps T, learning rate η, and nominal direction count J
Output: client detector $\{\boldsymbol{\theta}_T, \mathcal{I}, \eta, \boldsymbol{\mu}_{\ell,j}, \mathbf{b}_{\ell,j}, \rho_\ell, \widehat{F}_\ell, \widehat{F}_{\mathrm{G}}, \beta, \tau\}$

1: Compute $\mathbf{p}_n = \mathrm{softmax}(f_{\boldsymbol{\theta}_0}(x_n))$ and $\omega_n = p_{n,y_n}$ for $(x_n, y_n) \in \mathcal{D}_{\mathrm{ada},k}$.
2: Set $\boldsymbol{\pi}_k \leftarrow (\sum_n \omega_n \mathbf{e}_{y_n})/(\sum_n \omega_n)$.
3: Construct every fixed target $\mathbf{q}_n$ by Equation 1.
4: Initialize the diagnostic copy $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_0$.
5: **for** $t = 0, 1, \ldots, T-1$ **do**
6: $\quad \boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{att}}(\boldsymbol{\theta})$ using the complete $\mathcal{D}_{\mathrm{ada},k}$.
7: **end for**
8: Set the attenuated endpoint $\boldsymbol{\theta}_T \leftarrow \boldsymbol{\theta}$.
9: **for each** response stage $\ell \in \mathcal{I}$ **do**
10: $\quad$ Compute $\mathbf{d}_\ell(x_n) = \mathbf{h}_T^\ell(x_n) - \mathbf{h}_0^\ell(x_n)$ for $x_n \in \mathcal{D}_{\mathrm{ref},k}$.
11: $\quad$ Choose the largest locally supported direction count not exceeding J.
12: $\quad$ Fit $\boldsymbol{\mu}_{\ell,j}$ and assignments γ_n^ℓ by the norm-weighted spherical updates in Equation 31.
13: $\quad$ Fit $\mathbf{b}_{\ell,j}$ by Equation 3, using $\boldsymbol{\mu}_{\ell,j}$ if the weighted sum is zero; compute ρ_ℓ.
14: $\quad$ Evaluate the reference alignments $a_\ell(x_n)$ and fit their empirical rank map $\widehat{F}_\ell$.
15: **end for**
16: Compute $a_{\mathrm{G}}(x_n)$ on $\mathcal{D}_{\mathrm{ref},k}$ and fit $\widehat{F}_{\mathrm{G}}$.
17: Compute $s_{\mathrm{R},n}^{\mathrm{ref}}$ and set $\beta \leftarrow Q_{0.1}(\{s_{\mathrm{R},n}^{\mathrm{ref}}\}_n)$.
18: Compute the fused scores $S(x)$ on $\mathcal{D}_{\mathrm{cal},k}$ from the fitted quantities and set τ by Equation 37.
19: **return** $\{\boldsymbol{\theta}_T, \mathcal{I}, \eta, \boldsymbol{\mu}_{\ell,j}, \mathbf{b}_{\ell,j}, \rho_\ell, \widehat{F}_\ell, \widehat{F}_{\mathrm{G}}, \beta, \tau\}$.

Construction steps. In Algorithm 1, lines 1–3 construct the local profile and fixed targets; lines 4–8 perform semantic attenuation. Within the stage loop (lines 9–15), line 10 extracts responses, line 11 selects the supported direction count, lines 12–13 fit geometric centers, routing directions, and scales, and line 14 calibrates alignments. Lines 16–17 fit global support and the fusion scale. Line 18 calibrates the deployment threshold; line 19 returns the detector.

Algorithm 2 FedOPhD query scoring and deployment decision

Input: query x, fixed checkpoint $\boldsymbol{\theta}_0$, and the client detector returned by Algorithm 1
Output: OOD score $S(x)$, accept/reject decision, and the original-model prediction for an accepted query

1: **for each** response stage $\ell \in \mathcal{I}$ **do**
2: $\quad$ Compute $\mathbf{d}_\ell(x) = \mathbf{h}_T^\ell(x) - \mathbf{h}_0^\ell(x)$.
3: $\quad$ Route x to $j_\ell^*(x)$ with $\mathbf{b}_{\ell,j}$ and compute $a_\ell(x)$ by Equation 4.
4: $\quad$ Set the layer support to $\widehat{F}_\ell(a_\ell(x))$.
5: **end for**
6: Aggregate $s_{\mathrm{R}}(x) \leftarrow [\prod_{\ell \in \mathcal{I}} \widehat{F}_\ell(a_\ell(x))]^{1/|\mathcal{I}|}$.
7: Compute $a_{\mathrm{G}}(x) \leftarrow \log \sum_{c=1}^{C} \exp(f_{\boldsymbol{\theta}_0,c}(x))$ and $s_{\mathrm{G}}(x) \leftarrow \widehat{F}_{\mathrm{G}}(a_{\mathrm{G}}(x))$.
8: Fuse $B(x) \leftarrow s_{\mathrm{R}}(x)[s_{\mathrm{G}}(x) + \beta]/[s_{\mathrm{R}}(x) + \beta]$ and set $S(x) \leftarrow -B(x)$.
9: Compute the unchanged classifier prediction $\hat{y}(x) \leftarrow \arg\max_c f_{\boldsymbol{\theta}_0,c}(x)$.
10: **if** $S(x) > \tau$ **then reject** x; **otherwise accept** x and return $\hat{y}(x)$.
11: **return** $S(x)$ and the deployment decision.

Query-time steps. In Algorithm 2, lines 1–5 compute stage supports: line 2 extracts responses, line 3 routes with functional directions and scores geometric alignment, and line 4 applies the rank map. Lines 6–8 aggregate

response support, compute global support, and fuse them. Line 9 preserves the original prediction; lines 10–11 apply the threshold and return the decision and score.

D IMPLEMENTATION DETAILS

This appendix specifies the two-signal detector defined in Section 4.2. The geometric and functional direction banks, layer scales, marginal references and fusion scale are fitted offline; query-dependent fusion weights are evaluated in closed form. We use J for the number of directions and β for the fusion scale, reserving K for the client count and τ for the deployment threshold.

D.1 FIXED SETTINGS AND ATTENUATION PROTOCOL

The diagnostic states start from the original ResNet-18 checkpoint. The teacher remains fixed, while all student parameters, including normalization affine parameters, are eligible for gradient updates. Evaluation mode fixes batch-normalization running statistics and disables any dropout. Inputs use the dataset's deterministic evaluation preprocessing, including normalization, and the same cached tensors are reused throughout attenuation. The original predictions, confidence-weighted label profile and projected targets are computed once on $\mathcal{D}_{\mathrm{ada}}$. At each of the five steps, the same unweighted mean KL objective in Equation 2 is evaluated over all of $\mathcal{D}_{\mathrm{ada}}$. Every update uses the same complete sample set, equal sample weighting, and fixed targets. For memory efficiency, its gradient may be accumulated over deterministic micro-batches of at most 128 examples; the optimizer steps once after forming the complete empirical gradient. Each of the five diagnostic-state updates therefore uses the complete empirical gradient over $\mathcal{D}_{\mathrm{ada}}$. Scoring retains the original and final diagnostic states. Successful detector construction requires five complete updates with finite parameters.

Table 3: Fixed detector settings.

Setting	Value or rule
Attenuation updates	$T = 5$ full empirical-gradient steps
Attenuation target	Client-support subtraction and probability projection, fixed across updates
Optimizer	SGD; zero momentum and weight decay
Semantic-attenuation learning rate	$\eta = 10^{-4}$ for the diagnostic-copy SGD updates
Gradient accumulation	Deterministic micro-batches of at most 128; one optimizer step per full empirical gradient
Distillation temperature	1
Updated parameters	All model parameters
Forward-pass mode	Evaluation mode
Response layers	Pooled residual stages $1, 2, 3$
Reference directions	Nominally $J = 16$ per stage; the effective count is reduced to the largest locally supported number for sparse clients; norm-weighted spherical clustering
Direction fitting	Deterministic initialization; at most 40 refinement iterations
Query-to-mode routing	Within each mode, weight reference responses by the decrease in their original predicted-class margin; route a query to the most aligned weighted direction
Response normalization	Original-feature RMS scale using the semantic-attenuation step scale $\eta = 10^{-4}$
Layer aggregation	Equal-weight geometric mean
Reference fitting	Complete client-local $\mathcal{D}_{\mathrm{ref},k}$
Fusion scale	β: tenth percentile of local reference response supports
OOD score	$S = -B$ applied directly after adaptive fusion
L-ID acceptance target	95%

D.2 TARGET CONSTRUCTION

The profile $\boldsymbol{\pi}_k$ in Section 4.1 is the original-confidence-weighted empirical distribution of true labels on the adaptation subset. It represents the deployment client's learned label support for target construction. The factor ω_n couples each sample's contribution to this profile with the scale of its target subtraction. Both quantities are frozen before optimization.

Projection computation. The target in Equation 1 is the Euclidean projection of $\mathbf{v}_n = \mathbf{p}_n - \omega_n \boldsymbol{\pi}_k$ onto Δ_{-y_n}. We compute it using the standard sorting-based simplex projection (Martins & Astudillo, 2016) on the non-true-class coordinates. Sort these $C-1$ coordinates as $v_{n,(1)} \geq \cdots \geq v_{n,(C-1)}$ and set

$$m_n = \max\left\{ j \in \{1,\ldots,C-1\} : v_{n,(j)} > \frac{\sum_{r=1}^{j} v_{n,(r)} - 1}{j} \right\}, \qquad \kappa_n = \frac{\sum_{r=1}^{m_n} v_{n,(r)} - 1}{m_n}. \tag{29}$$

The target is $q_{n,y_n} = 0$ and $q_{n,c} = \max\{v_{n,c} - \kappa_n, 0\}$ for $c \neq y_n$. This yields the unique Euclidean projection, with nonnegative entries summing to one. The sorting-based construction costs $O(NC \log C)$ operations and directly produces each fixed probability target before attenuation begins. The projection acts on these target probability vectors and uses the simplex threshold above to determine active coordinates.

Relation to uniform redistribution. Let $q^{\mathrm{U}}_{n,y_n} = 0$ and $q^{\mathrm{U}}_{n,c} = p_{n,c} + \omega_n/(C-1)$ for $c \neq y_n$. When all projected non-true-class entries are positive, the new target satisfies

$$q_{n,c} = q^{\mathrm{U}}_{n,c} - \omega_n \left(\pi_{k,c} - \frac{1 - \pi_{k,y_n}}{C-1} \right), \qquad c \neq y_n. \tag{30}$$

Thus, relative to uniform redistribution, it subtracts each alternative's excess support above the mean of the non-true classes. With active zero coordinates, the projection above remains the defining rule. The target can reduce selected alternatives' original probabilities according to the projected client-support contrast. For $\omega_n > 0$, a strictly changed target also changes the initial logit-gradient direction: both $\mathbf{p}_n - \mathbf{q}_n$ and $\mathbf{p}_n - \mathbf{q}^{\mathrm{U}}_n$ have true-class coordinate ω_n, making the two distinct residuals directionally different. After this logit-level change, the resulting parameter-update and pooled-response directions are determined by the model Jacobians and response readout.

Boundary and numerical cases. Construction uses $N > 0$, $C \geq 2$, and finite logits. Under these conditions, softmax assigns every true class a finite positive confidence ω_n, so the profile-normalization denominator $\sum_n \omega_n$ is positive. Uniform support across the non-true classes recovers $\mathbf{q}^{\mathrm{U}}_n$, covering single-class clients and binary classification within the same projection rule. All target classes remain in the global label set; the sample's true class is fixed at zero and the projection determines the active alternatives. Zero target entries contribute zero to KL; the student term is evaluated through log-softmax for numerical stability.

D.3 FUNCTIONAL ROUTING AND RESPONSE READOUT

At each selected response stage $\ell \in \mathcal{I} = \{1,2,3\}$, form the reference responses $\mathbf{d}_\ell(x_n)$ and normalize each nonzero response by its Euclidean norm for angular assignment. Responses with norm at most 10^{-12} are represented by zero vectors and excluded from direction fitting. We nominally fit $J = 16$ unit centers $\boldsymbol{\mu}_{\ell,j}$ by norm-weighted spherical clustering. When a sparse client's nonzero reference responses cannot populate 16 modes, we reduce the effective direction count to the largest locally supported number before fitting. The fixed-J sensitivity experiment in Appendix E.4.1 deliberately disables this direction-count reduction to evaluate the specified bank sizes directly. In the equations below, J denotes this client-effective count. Writing γ^ℓ_n for the cluster assignment, each refinement uses

$$\gamma^\ell_n = \arg\max_{1 \leq j \leq J} \left\langle \frac{\mathbf{d}_\ell(x_n)}{\|\mathbf{d}_\ell(x_n)\|_2}, \boldsymbol{\mu}_{\ell,j} \right\rangle, \qquad \boldsymbol{\mu}_{\ell,j} \leftarrow \frac{\sum_{n:\gamma^\ell_n = j} \mathbf{d}_\ell(x_n)}{\left\| \sum_{n:\gamma^\ell_n = j} \mathbf{d}_\ell(x_n) \right\|_2}. \tag{31}$$

The center update normalizes the sum of raw displacements, so response norms weight the fit; sufficiently excited query responses are then compared through this angular geometry. Initialization

selects the observed direction closest to the mean response direction, then adds centers by deterministic farthest-angle selection; refinement runs for at most 40 iterations. Each stage is fitted independently from the client-local reference responses.

To attach classifier function to these geometric modes, fix each reference input's original prediction $\hat{y}_n = \arg\max_c f_{\boldsymbol{\theta}_0,c}(x_n)$ and define its endpoint margin as $m_t(x_n) = f_{\boldsymbol{\theta}_t,\hat{y}_n}(x_n) - (C-1)^{-1}\sum_{c\neq\hat{y}_n} f_{\boldsymbol{\theta}_t,c}(x_n)$ for $t \in \{0, T\}$. Within the original geometry assignment γ_n^ℓ, the functional routing direction is $\mathbf{b}_{\ell,j} = \mathrm{unit}(\sum_{n:\gamma_n^\ell=j}[m_0(x_n) - m_T(x_n)]\mathbf{d}_\ell(x_n))$ when the weighted sum is nonzero. If this sum is zero, we instead set $\mathbf{b}_{\ell,j} = \boldsymbol{\mu}_{\ell,j}$; routing then uses the fitted unit geometric direction of that mode, without normalizing a zero vector. Thus, larger decreases of the fixed original-class margin contribute more strongly to the routing pattern while the geometric centers $\boldsymbol{\mu}_{\ell,j}$ remain unchanged.

For a query, $j_\ell^*(x) = \arg\max_j \langle \mathbf{d}_\ell(x)/\|\mathbf{d}_\ell(x)\|_2, \mathbf{b}_{\ell,j}\rangle$ selects the functionally matching mode, and the response is scored against its original geometric direction, with the same zero-response convention used during fitting. Let $\rho_\ell = [n_{\mathrm{ref}}^{-1}\sum_n \|\mathbf{h}_0^\ell(x_n)\|_2^2]^{1/2}$ be the RMS norm of the original-model reference representations. The raw readout is $a_\ell(x) = \langle \mathbf{d}_\ell(x)/\max\{\|\mathbf{d}_\ell(x)\|_2, \eta\rho_\ell\}, \boldsymbol{\mu}_{\ell,j_\ell^*(x)}\rangle$, using the same attenuation learning rate $\eta = 10^{-4}$. Equivalently, a_ℓ is the selected cosine multiplied by $\min\{1, \|\mathbf{d}_\ell(x)\|_2/(\eta\rho_\ell)\}$. This gives weak responses proportionate directional evidence and exactly recovers the original angular readout once their norm reaches the shared layer scale. A zero response has readout zero and remains in the subsequent reference-score calibration. The response readout uses stages 1–3. Both models retain the complete architecture, but global recognition is computed only from the original model; logits from both endpoints are used to compute reference-margin changes for functional routing.

D.4 Diagnostic-only population cosine summary

This subsection defines the reporting-only mechanism diagnostic in Table 22 from the deployed detector's routed responses. For a query x at client k and residual stage $\ell \in \{1, 2, 3\}$, functional routing first selects one of the $J = 16$ client-local reference modes, $j_\ell^*(x)$, exactly as in Section 4.2. We then retain the pure geometric cosine with the selected mode,

$$c_{k,\ell}(x) = \left\langle \frac{\mathbf{d}_\ell(x)}{\|\mathbf{d}_\ell(x)\|_2}, \boldsymbol{\mu}_{\ell,j_\ell^*(x)} \right\rangle, \qquad c_k(x) = \frac{1}{3}\sum_{\ell=1}^{3} c_{k,\ell}(x), \tag{32}$$

using the same zero-response convention as the response readout. Thus, each client–query pair contributes one observation obtained by equal-weight averaging across the three stages. This diagnostic retains the pure routed geometric cosine; the deployed response support s_{R} additionally incorporates response excitation, marginal rank transforms, and geometric aggregation of stage-wise supports.

Query responses and routes are fixed first, followed by grouping the resulting observations as L-ID, T-ID, near-OOD, and far-OOD. Samples receive equal weight within each client–source block, client aggregates are weighted in proportion to their complete own-test sample counts, and each constituent source receives equal total weight within the near- and far-OOD groups. Table 22 reports the weighted mean, weighted sample standard deviation, and weighted median of the resulting $c_k(x)$ distribution, together with its observed minimum and maximum. Min and Max are the observed extrema of the per-observation three-stage averages $c_k(x)$.

D.5 Marginal rank transforms

For reference values $v_1, \ldots, v_n$, we use an empirical rank transform with an add-one adjustment:

$$\widehat{F}(v; v_1, \ldots, v_n) = \frac{1 + \sum_{i=1}^{n} \mathbb{I}\{v_i \le v\}}{n+1}. \tag{33}$$

Here n is the reference count and $\mathbb{I}$ is the indicator function; the weak inequality assigns ties their right-side rank. The transform takes values in $(0, 1]$. The global support uses the ascending rank of the positive log-partition defined in Section 4.2. Each stage's readouts are ranked against reference scores computed with the same geometric centers, functional routing directions and original-feature RMS scale. The response support is the equal-weight geometric mean of the three stage-wise ranks,

evaluated by averaging their logarithms and exponentiating. The geometric mean is the final response support, and adaptive fusion operates directly on this value. Marginal ranks provide a local percentile scale, with values beyond the reference range sharing the corresponding endpoint rank.

D.6 Data roles, local partitioning, and reference fitting

Table 4: Data roles and fixed partitions used by FedOPhD.

Data role	Source and size	Fixed partition	Use
FL training pool	45,000 images from the official training set	Existing 20-client Dirichlet partition; union of all $\mathcal{D}_{\mathrm{ada},k}$ and $\mathcal{D}_{\mathrm{ref},k}$	Federated training and the resulting global checkpoint
ID validation	Remaining 5,000 official training images	Preallocated before FL training to form $\bigcup_k \mathcal{D}_{\mathrm{cal},k}$; retained locally	Global-checkpoint selection and client-local deployment calibration
$\mathcal{D}_{\mathrm{ada},k}$	Approximately half of client k's local FL training samples	Per-class half split paired with $\mathcal{D}_{\mathrm{ref},k}$	Original FL training; then client profile, fixed attenuation targets, and semantic-attenuation updates
$\mathcal{D}_{\mathrm{ref},k}$	Remaining half of client k's local FL training samples	Per-class half split paired with $\mathcal{D}_{\mathrm{ada},k}$	Original FL training; then geometric and functional response modes, RMS scales, G/R empirical rank maps, and fusion scale β
$\mathcal{D}_{\mathrm{cal},k}$	Client allocation from the 5,000-image ID validation set	Assigned before FL training using fixed class proportions; retained by client k	L-ID deployment threshold after detector construction
ID test	Complete official 10,000-image test set	Assigned before FL training using the α-induced class-wise client ratios; fixed for L-ID/T-ID grouping	All-ID metrics, L-ID/T-ID analysis, and locally calibrated acceptance evaluation
OOD evaluation	Complete fixed near- and far-OOD test sources	Shared query sources; five far-OOD datasets receive equal weight	AUROC, FPR95, and OOD detection under the deployed local threshold

Before federated training, the official training set is divided in a class-stratified manner into a 45,000-image FL training pool and a 5,000-image ID validation set. The FL training pool retains the existing 20-client Dirichlet partition and supplies every $\mathcal{D}_{\mathrm{ada},k}$ and $\mathcal{D}_{\mathrm{ref},k}$. For each client k and locally represented class c, a fixed seed permutes the $n_{k,c}$ local training indices and divides them as evenly as possible between the two subsets. Even class counts split equally; odd remainders alternate between the two sides so their client-level totals differ by at most one, and a singleton class contributes its sample to one side. This exhaustive per-class split assigns every local FL training sample exactly once and preserves the client's original class distribution.

The 5,000-image ID validation set provides checkpoint validation and the client calibration subsets. For validation class c, define the allocation proportion from the existing FL partition as

$$p(k \mid c) = \frac{n_{k,c}}{\sum_{j=1}^{K} n_{j,c}}. \tag{34}$$

Before federated training begins, largest-remainder allocation converts these proportions to integer client counts and assigns all validation images exactly once. The resulting mutually exclusive subsets $\mathcal{D}_{\mathrm{cal},k}$ form a complete partition of ID validation and retain its class-wise client proportions. Each client retains its preassigned $\mathcal{D}_{\mathrm{cal},k}$ locally and uses it to calibrate its deployment threshold after detector construction; no validation data are redistributed or accessed across clients at deployment. The training/validation split and class-wise allocation are completed before FL training; deployment-time calibration uses each client's preassigned local subset.

Reference statistics and fusion scale. For a fixed client, the complete $\mathcal{D}_{\mathrm{ref}}$ fits the geometric centers, functional routing directions, original-feature RMS scales, layer-wise empirical rank maps, and the global-energy rank map. Applying these fitted reference statistics to the same local reference samples yields response supports $\{s_{\mathrm{R},n}^{\mathrm{ref}}\}_{n=1}^{n_{\mathrm{ref}}}$. For a nondegenerate reference, define

$$\beta = Q_{0.1}\left(\{s_{\mathrm{R},n}^{\mathrm{ref}}\}_{n=1}^{n_{\mathrm{ref}}}\right), \tag{35}$$

where $n_{\mathrm{ref}} = |\mathcal{D}_{\mathrm{ref}}|$ and $Q_{0.1}$ is the empirical tenth percentile with linear interpolation. This lower-tail reference statistic sets the client-specific transition scale of the response-dependent fusion rule.

The experiment records the sample indices and per-client class counts for every role and verifies the stated partition relationships before detector fitting. FedOPhD uses $J = 16$ reference directions and selected response stages $\mathcal{I} = \{1, 2, 3\}$ throughout the main experiments; the corresponding sensitivity analyses vary one fixed configuration at a time. The client profile, response modes, empirical ranks, fusion scale, and deployment threshold are all determined from the assigned L-ID subsets.

D.7 Adaptive fusion

Taking logarithms of Equation 5 gives $\log B = \log s_{\mathrm{R}} + \log(s_{\mathrm{G}} + \beta) - \log(s_{\mathrm{R}} + \beta)$, and the implementation evaluates its rational expression directly, including at equal supports. For positive supports and $\beta > 0$, the geometric representation $B(x) = s_{\mathrm{G}}(x)^{\lambda(x)} s_{\mathrm{R}}(x)^{1-\lambda(x)}$ has the explicit weight

$$\lambda(x) = \begin{cases} \dfrac{\log[(s_{\mathrm{G}}(x) + \beta)/(s_{\mathrm{R}}(x) + \beta)]}{\log[s_{\mathrm{G}}(x)/s_{\mathrm{R}}(x)]}, & s_{\mathrm{G}}(x) \neq s_{\mathrm{R}}(x), \\ \dfrac{s_{\mathrm{G}}(x)}{s_{\mathrm{G}}(x) + \beta}, & s_{\mathrm{G}}(x) = s_{\mathrm{R}}(x). \end{cases} \tag{36}$$

The unequal-support expression is a secant slope of $t \mapsto \log(e^t + \beta)$, whose derivative is $e^t/(e^t + \beta) \in (0, 1)$; hence $0 < \lambda(x) < 1$. The equal-support expression is its continuous extension, although any exponent yields the same fused support at equality.

The equivalent arithmetic coefficients are $w_{\mathrm{G}}^{\mathrm{arith}} = s_{\mathrm{R}}/(s_{\mathrm{R}} + \beta)$ and $w_{\mathrm{R}}^{\mathrm{arith}} = \beta/(s_{\mathrm{R}} + \beta)$; the geometric representation uses the corresponding exponents λ and $1 - \lambda$. Consequently B lies between its two input supports and increases with either input when the other is fixed. This monotone, response-dependent scoring rule lets global recognition complement response support while preserving response-gated influence.

The scale in Equation 35 is used directly to compute the fusion weights. The final OOD score is the direct fused-support negation $S = -B$.

D.8 Threshold calibration and inference

After fixing the detector, sort the $n_{\mathrm{cal}} = |\mathcal{D}_{\mathrm{cal},k}|$ local calibration scores as $S_{(1)} \leq \cdots \leq S_{(n_{\mathrm{cal}})}$. Set

$$k_{\mathrm{cal}} = \min\left\{n_{\mathrm{cal}}, \lceil 0.95(n_{\mathrm{cal}} + 1)\rceil\right\}, \qquad \tau = S_{(k_{\mathrm{cal}})}. \tag{37}$$

Reject x when $S(x) > \tau$; equality is accepted as ID. The equivalent acceptance condition is $B(x) \geq -\tau$. This finite-sample quantile implements the 95% empirical L-ID acceptance convention; T-ID acceptance is evaluated separately with the transferred-ID protocol. FPR95 and T-FPR95 use the OOD-detection operating point and their respective negative ID evaluation populations described in Appendix D.9.

Inference. At inference, compute original-model recognition and initial-to-final representation changes, route each response with the fitted functional directions, score the selected geometric direction under the stored RMS-scaled normalization, and apply the marginal ranks and fusion rule to obtain $s_{\mathrm{G}}(x)$, $s_{\mathrm{R}}(x)$, $B(x)$ and $S(x)$. The original model supplies the class returned for accepted inputs. All stored detector quantities remain fixed during test-time scoring.

D.9 Experimental Protocol and Evaluation Details

Datasets. We evaluate federation-wide semantic OOD detection on CIFAR-10 and CIFAR-100 (Krizhevsky, 2009), containing 10 and 100 ID classes, respectively. Each dataset provides 50,000 training images and 10,000 test images at 32×32 resolution. Before federated training, we make a class-stratified split of the official training set: 45,000 images form the FL training pool and the remaining 5,000 form the ID validation set. The FL training pool follows the fixed 20-client class-wise Dirichlet partition with concentration $\alpha \in \{0.1, 0.5, 10\}$. Smaller α induces stronger label heterogeneity and more uneven local class coverage. For near-OOD evaluation, we additionally

use the other CIFAR dataset and TinyImageNet (Le & Yang, 2015), following the saved benchmark image lists from OpenOOD (Yang et al., 2022). For TinyImageNet, we use the project's fixed image lists. Far-OOD evaluation uses LSUN-Crop, LSUN-Resize, SVHN, iSUN, and Textures. Textures denotes the complete 5,640-image DTD collection and contributes once to the five-source far-OOD average. All inputs follow the corresponding ID dataset's evaluation preprocessing.

Score-distribution visualization. Figure 1 uses CIFAR-10 as ID with $\alpha = 0.1$ and 20 clients, and the complete set of 8,925 iSUN images as OOD. For visualization only, all scores are oriented so that larger values indicate stronger ID evidence. We apply a logit transform to the bounded FedOPhD fusion support B and retain the ID-oriented raw scores for Energy and ViM. For each method and client, the resulting scores are standardized using the mean and population standard deviation estimated from the same client-specific calibration subset drawn from the official training set. The fitted transformation is shared across L-ID, T-ID, and OOD samples, so the plots show densities in method-specific standardized display coordinates. AUROC and FPR95 are computed from the original OOD-oriented detection scores before these visualization transformations.

Comparison Methods and Evaluations. We consider post-training detectors and federated training-time methods as separate comparison groups. The selected post-hoc group comprises FedEnergy (Liu et al., 2020), FedViM (Wang et al., 2022a), FedNCI (Liu & Qin, 2025), FedGradPCA (Seleznova et al., 2026), and FedRAS (Guglielmo & Masana, 2026), listed in the same order as Table 1. The perturbation-based post-hoc group comprises WeiPer+KLD (Granz et al., 2024); PRO-MSP, PRO-ENT, and PRO-GEN from Perturbation Rectified OOD detection (Chen et al., 2025); and Gradient Short-Circuit (GSC) (Gu et al., 2025). These methods and FedOPhD start from the same FedAvg checkpoint for each setting and receive the same complete deployment-client local FL training dataset. Each reference-dependent baseline fits its method-specific statistics using all of these local samples. FedOPhD alone partitions the same complete local dataset into $\mathcal{D}_{\mathrm{ada},k}$ and $\mathcal{D}_{\mathrm{ref},k}$ for its two internal construction roles. FedViM uses a 256-dimensional principal subspace. FedGradPCA fits its gradient class-mean principal subspace using the deployment client's L-ID data. FedRAS fits its mean rank-ordered activation curve on local training features. FedNCI uses its overall mean and the original classifier weights, with the coefficient fixed at the author's configured default of 10^{-4}. We evaluate these centralized scoring rules as local-reference adaptations under the common federated setting. The perturbation-based implementations and their complete five-seed results are specified in Appendix E.2.6. The training-time comparison group comprises FOSTER (Yu et al., 2023), FOOGD (Liao et al., 2024), and FedSDWC (Huang et al., 2026), which use their own trained models and method-specific scores.

Evaluation metrics. All scores are oriented so that larger values indicate stronger OOD evidence. ACC is classification accuracy on the complete federation-level ID test set; post-hoc detectors and FedOPhD inherit it from their shared FedAvg warm-up, whereas FOSTER, FOOGD, and FedSDWC use their own model ACC. For each client, AUROC and FPR95 distinguish OOD from the complete official ID test set, whereas T-AUROC and T-FPR95 use test samples owned by the other clients as the ID population. Client provenance defines this grouping, with all locally held classes and prediction outcomes retained. AUROC is the probability that an OOD sample scores above an ID sample, with ties receiving half credit. With OOD as the positive class and ID as the negative class, the FPR95 threshold is the largest threshold at which at least 95% of OOD is detected; FPR95 is the fraction of the relevant negative ID population whose score exceeds this threshold and is therefore incorrectly rejected as OOD. T-FPR95 uses the same OOD-detection operating point and replaces the complete federation-wide ID population with T-ID when computing the negative-class false-positive rate. These evaluation thresholds differ from the deployment threshold calibrated on L-ID. Near- and far-OOD are evaluated separately: constituent datasets receive equal weight within each group for every client, followed by client aggregation in proportion to the fixed own-test sample counts.

Table 5 reports the classification accuracies omitted from the main OOD-comparison tables.

Table 5: Federation-level ID classification accuracy (%; mean ± sample standard deviation). The post-hoc detectors, FedOPhD-R, and FedOPhD share the corresponding FedAvg checkpoint; training-time methods use their own trained models.

ID dataset	Model or checkpoint	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 10$
CIFAR-10	Shared FedAvg checkpoint	88.94 ± 0.73	**92.40** ± 0.74	**93.12** ± 0.81
	FOSTER	**90.22** ± 1.31	90.93 ± 1.39	89.63 ± 1.07
	FOOGD	85.86 ± 0.76	91.96 ± 0.81	92.98 ± 0.78
	FedSDWC	88.47 ± 0.73	91.70 ± 0.67	92.32 ± 0.69
CIFAR-100	Shared FedAvg checkpoint	67.36 ± 0.76	69.40 ± 0.57	**73.66** ± 0.63
	FOSTER	66.27 ± 0.94	69.88 ± 1.19	73.53 ± 0.86
	FOOGD	66.96 ± 0.60	69.40 ± 0.73	71.32 ± 0.51
	FedSDWC	**68.58** ± 0.75	**70.31** ± 0.31	72.88 ± 0.45

Data partitions. For every client k, its fixed local FL training indices are divided per class and as evenly as possible into $\mathcal{D}_{\mathrm{ada},k}$ and $\mathcal{D}_{\mathrm{ref},k}$ using a saved random seed. The split is exhaustive, preserves the existing client class distribution, and assigns odd per-class remainders so the two client totals differ by at most one. This split is specific to FedOPhD; the post-hoc baselines use the corresponding complete client-local index set before this internal split. The 5,000-image ID validation set supports global-checkpoint selection and is allocated exhaustively across clients before federated training to form $\{\mathcal{D}_{\mathrm{cal},k}\}_{k=1}^{K}$. For each class c, largest-remainder allocation follows the existing FL training proportion $p(k \mid c) = n_{k,c} / \sum_j n_{j,c}$, producing mutually exclusive client calibration subsets whose union is the complete ID validation set. These subsets remain local to their assigned clients; post-training calibration uses only the client's own preallocated data. This allocation is an experimental partitioning choice, not deployment-time distribution of data from a global pool. **Pre-training assignment of ID test samples.** Client ownership of the CIFAR test samples is constructed by our experimental partitioning protocol, not supplied by the original dataset. Before federated training begins, we partition the complete official 10,000-image ID test set across clients class by class, using the same class-wise client allocation ratios induced by the training partition at the corresponding Dirichlet concentration α. The resulting mutually exclusive client test subsets cover the complete ID test set, and their sample-to-client assignments remain fixed throughout training and evaluation. For a receiving client k, samples assigned to k define L-ID, while those assigned to all other clients define T-ID; these assignments also determine the client ID-test counts used for evaluation weighting. The assignment is solely an evaluation partition: test samples are not used for model training, detector fitting, calibration, or model selection. The complete fixed near- and far-OOD sources form the OOD evaluation queries. Appendix D.6 records the source, partition, and use of every data role.

Implementation Details. We use a CIFAR-style ResNet-18 (He et al., 2016), with a 3×3 stride-one stem and no initial max-pooling. The shared FedAvg (McMahan et al., 2017) checkpoint is trained for 200 communication rounds with all 20 clients participating. Each selected client performs five local epochs with batch size 128. The backbone optimizer is SGD with initial learning rate 0.1, five warmup rounds followed by cosine decay, momentum 0.9, and weight decay 5×10^{-4}. Training uses random cropping and horizontal flipping; checkpoint selection uses accuracy on the 5,000-image class-stratified ID validation set. Except for explicitly identified single-run diagnostic distributions, all experiments use the five global random seeds $\{42, 43, 44, 45, 46\}$. Reported results are means across the five seed-level results, and their standard deviations are sample standard deviations computed after within-run aggregation. FedOPhD constructs its diagnostic endpoint using five SGD attenuation updates at learning rate $\eta = 10^{-4}$, zero momentum, and zero weight decay. The client label-support profile and projected targets are computed once, and the same complete empirical KL objective is used for every step and every class. Deterministic micro-batches of at most 128 are used only for gradient accumulation. All parameters are eligible for updates, while batch-normalization running statistics remain fixed in evaluation mode. The response readout fits 16 geometric directions at each of residual stages 1–3, routes queries with margin-drop-weighted functional directions, and scores the selected geometry using original-feature-RMS-scaled normalization. The locally ranked layer scores are aggregated by an equal-weight geometric mean. The complete $\mathcal{D}_{\mathrm{ref},k}$ fits the geometric and functional directions, layer scales, marginal ranks, and client-specific fusion scale. After detector construction, $\mathcal{D}_{\mathrm{cal},k}$ determines the L-ID deployment threshold. The data roles, adaptive fusion, and fixed detector configuration are specified in Appendix D.

Hardware. We use an NVIDIA GeForce RTX 5080 GPU and an AMD Ryzen 9 9950X3D CPU.

E Additional Experiments

E.1 Deployment and Federation-Level ID Preservation

E.1.1 Deployment with an L-ID-only calibrated threshold

Motivation. Federation-wide AUROC and FPR95 characterize ranking and a federation-wide ID operating point. We additionally measure L-ID and T-ID acceptance together with near- and far-OOD detection at the deployed threshold calibrated from the current client's $\mathcal{D}_{\mathrm{cal},k}$, with particular attention to additional T-ID rejection in non-IID settings.

Protocol and metrics. For every method, random seed, and client k, we fix the trained model and fitted detector, then calibrate $\tau_{\mathrm{L},k}$ on $\mathcal{D}_{\mathrm{cal},k}$ from the 5,000-image ID validation partition. The finite-sample ID quantile in Equation 37 produces a fixed deployment threshold before scoring the official ID test set and OOD evaluation sources. On these evaluation populations, we compute

$$\begin{aligned} \text{L-FRR}_k &= \Pr_{X_{\mathrm{L}}}[S_k(X_{\mathrm{L}}) > \tau_{\mathrm{L},k}]\,, \\ \text{T-FRR}_k &= \Pr_{X_{\mathrm{T}}}[S_k(X_{\mathrm{T}}) > \tau_{\mathrm{L},k}]\,, \\ \Delta\,\text{T-FRR}_k &= \text{T-FRR}_k - \text{L-FRR}_k, \\ \text{DR}_{k,g} &= \Pr_{X_{\mathrm{O}}\sim P_g}[S_k(X_{\mathrm{O}}) > \tau_{\mathrm{L},k}]\,, \qquad g \in \{\text{Near}, \text{Far}\}. \end{aligned} \tag{38}$$

Thus, lower false-rejection rates and higher OOD detection rates are preferable. Within each client, constituent datasets receive equal weight within the near- and far-OOD groups; clients are weighted by their complete own-test sample counts, and the resulting metrics are averaged over five random seeds. These detection rates characterize the local deployment operating point, while FPR95 uses the separate evaluation threshold at which at least 95% of OOD is detected.

Figure 4 in the main text visualizes the preservation–detection trade-off in the most heterogeneous setting. Table 6 reports the exact five-seed means underlying the figure and the complete results for both ID datasets.

Table 6: L-ID-calibrated deployment on CIFAR-10 and CIFAR-100 (%; five-seed means). L/T-FRR denote L/T-ID false-rejection rates, ΔT-FRR is their difference, and DR is the OOD detection rate. L-FRR has a calibration target of 5% and is not ranked. Bold marks the lowest T-FRR and ΔT-FRR and the highest DR within each dataset and α.

		CIFAR-10					**CIFAR-100**				
α	Method	L-FRR	T-FRR↓	ΔT-FRR↓	Near-DR↑	Far-DR↑	L-FRR	T-FRR↓	ΔT-FRR↓	Near-DR↑	Far-DR↑
0.1	FedEnergy	4.62	**5.31**	**+0.69**	37.21	48.93	5.13	**5.12**	**−0.01**	17.93	21.46
	FedViM	5.31	52.26	+46.95	**53.52**	65.70	4.79	10.72	+5.93	14.11	27.20
	FedNCI	5.12	7.31	+2.19	17.52	19.80	4.27	7.78	+3.51	15.96	19.12
	FedGradPCA	4.76	23.45	+18.69	38.65	46.27	5.29	20.26	+14.97	20.92	38.02
	FedRAS	4.67	8.54	+3.87	35.29	48.05	4.33	5.45	+1.12	18.43	26.04
	FedOPhD (ours)	5.06	7.22	+2.16	44.36	**92.70**	5.25	7.11	+1.86	**23.86**	**73.29**
0.5	FedEnergy	4.68	4.93	+0.25	40.56	63.11	5.56	**5.50**	**−0.06**	18.66	24.31
	FedViM	4.65	16.48	+11.83	46.07	70.11	5.03	6.26	+1.23	11.36	17.81
	FedNCI	5.24	**3.61**	**−1.63**	30.02	50.04	5.58	5.67	+0.09	19.92	20.89
	FedGradPCA	4.64	11.48	+6.84	39.22	51.33	5.63	11.22	+5.59	15.37	29.17
	FedRAS	4.67	6.90	+2.23	40.67	68.81	5.75	5.78	+0.03	19.69	22.34
	FedOPhD (ours)	5.08	6.79	+1.71	**46.34**	**93.86**	4.99	5.58	+0.59	**26.77**	**76.93**
10	FedEnergy	4.97	**5.00**	**+0.03**	39.12	52.88	5.14	5.17	+0.03	16.77	21.87
	FedViM	4.95	5.09	+0.14	39.32	63.97	5.47	5.28	**−0.19**	11.32	16.49
	FedNCI	5.34	5.42	+0.08	40.15	54.52	5.18	5.30	+0.12	18.06	19.71
	FedGradPCA	4.76	**5.00**	+0.24	44.45	53.19	4.78	**4.87**	+0.09	11.87	18.59
	FedRAS	5.27	5.37	+0.10	40.97	57.25	5.19	5.25	+0.06	18.03	20.01
	FedOPhD (ours)	5.37	5.49	+0.12	**45.65**	**93.01**	5.35	5.21	−0.14	**25.98**	**73.61**

Results. *Calibration fidelity.* Across all methods and settings, the realized L-ID false-rejection rate remains close to the intended operating point, confirming that the comparison uses similarly calibrated local thresholds.

Preservation of federation-level ID. Under strong heterogeneity, FedOPhD introduces only a modest additional rejection of T-ID, and this gap contracts toward zero as the client partitions become less

heterogeneous. Although FedEnergy and several other baselines sometimes produce an even smaller T-ID gap, they do so with substantially weaker OOD detection. By contrast, ViM's sole near-OOD advantage over FedOPhD on CIFAR-10 and the comparatively strong detection of GradPCA relative to several baselines under strong heterogeneity are accompanied by sharply increased T-ID rejection. Their apparent detection strength therefore partly reflects treating legitimate inputs from other clients as OOD.

OOD detection at the deployed threshold. FedOPhD achieves the highest far-OOD detection in every dataset–heterogeneity combination and the highest near-OOD detection in all but one, combining consistently high far-OOD detection with competitive near-OOD detection. Taken together, these results establish the desired deployment behavior: FedOPhD preserves T-ID while retaining strong OOD rejection.

E.1.2 T-ID-ONLY EVALUATION

To complement the L-ID-calibrated deployment evaluation in Appendix E.1.1, we provide a separate ranking-based analysis that isolates T-ID. Federation-wide AUROC and FPR95 evaluate OOD against the union of L-ID and T-ID. Because the central federated challenge is to preserve ID inputs contributed by other clients, we additionally isolate T-ID as the ID evaluation population while retaining the same fixed detector scores and OOD test sets. For a client k, let X_{T} be drawn from the official ID test samples assigned to the other clients before federated training using the α-induced class-wise client ratios (Appendix D.9), and let X_{O} be drawn from the relevant OOD evaluation source. With larger S indicating stronger OOD evidence, we define

$$\text{T-AUROC} = \Pr[S(X_{\mathrm{O}}) > S(X_{\mathrm{T}})] + \tfrac{1}{2}\Pr[S(X_{\mathrm{O}}) = S(X_{\mathrm{T}})]\,. \tag{39}$$

Let τ_{95} be the largest evaluation threshold that detects at least 95% of OOD, i.e., $\Pr[S(X_{\mathrm{O}}) > \tau_{95}] \geq 0.95$. We then define

$$\text{T-FPR95} = \Pr[S(X_{\mathrm{T}}) > \tau_{95}]\,, \tag{40}$$

the fraction of negative T-ID inputs incorrectly rejected as OOD at that operating point. Thus, T-AUROC and T-FPR95 use T-ID and OOD for evaluation, while the original AUROC and FPR95 use the complete federation-wide ID test set and OOD. The client-local subsets determine the fitted detector and deployment threshold τ; τ_{95} is a separate evaluation operating point determined by the OOD scores.

Table 7 shows that T-ID-only performance closely tracks the federation-wide results across both datasets. Near-OOD separation generally improves as client partitions become less heterogeneous, and far-OOD remains easier than near-OOD throughout. Far-OOD performance improves with increasing α on CIFAR-10. On CIFAR-100, both evaluation populations attain their highest far-OOD AUROC and lowest far-OOD FPR95 at $\alpha = 0.5$. Overall, the close agreement between the two metric families indicates that the federation-wide behavior also extends to inputs transferred from other clients.

Table 7: Federation-wide and T-ID-only OOD results for FedOPhD (%; five-seed means). T-AUROC and T-FPR95 use only T-ID as the ID evaluation population.

OOD	α	CIFAR-10				CIFAR-100			
		AUROC↑	FPR95↓	T-AUROC↑	T-FPR95↓	AUROC↑	FPR95↓	T-AUROC↑	T-FPR95↓
Near-OOD	0.1	86.03	43.96	85.69	44.38	76.79	67.45	76.64	67.66
	0.5	89.15	39.30	89.08	39.46	78.72	61.72	78.69	61.75
	10	89.52	39.57	89.51	39.60	79.58	58.71	79.57	58.72
Far-OOD	0.1	97.13	13.96	97.00	14.41	94.06	23.80	93.95	24.22
	0.5	98.32	7.66	98.27	7.87	94.29	21.87	94.20	21.99
	10	98.43	6.72	98.40	6.78	93.52	23.63	93.55	23.67

E.1.3 LOCAL CLASS COVERAGE AFTER FEDERATED TRAINING

To quantify the effect of post-training local class coverage, we report the per-coverage numerical results underlying Figure 7. For each seed in $\{42, 43, 44, 45, 46\}$, we fix its corresponding global checkpoint trained on CIFAR-10 with $\alpha = 0.1$ and 20 clients, then vary the number of classes N_{cls}

available to each client for post-training detector construction. Within each seed, the checkpoint remains unchanged across coverage levels; α describes the FL training partition.

Controlled data allocation. A class-stratified split of the official 50,000-image training set uses the fixed split seed 20260905 to reserve 5,000 validation images, leaving 45,000 training images with 4,500 per class. For this experiment, we repartition the complete 45,000-image pool into disjoint detector-construction subsets of exactly 2,250 images per client, rather than removing classes from the original Dirichlet client subsets used in Table 1. For each experimental seed, a NumPy `default_rng` generator draws permutations of the ten classes, twenty clients, and ten class offsets, in that order. Writing the class and offset permutations as $c_0, \ldots, c_9$ and $o_0, \ldots, o_9$, the client at position i in the client permutation receives classes $\{c_{(i+o_j) \bmod 10} : j = 0, \ldots, N_{\text{cls}} - 1\}$. Thus, each client has exactly N_{cls} classes and each class appears at exactly $2N_{\text{cls}}$ clients. Each selected client–class pair initially receives $\lfloor 2250/N_{\text{cls}} \rfloor$ images; an integer maximum-flow allocation distributes the remainder while preserving exactly 2,250 images per client and 4,500 images per class globally. At $N_{\text{cls}} = 10$, each client consequently receives 225 images per class. Validation images are allocated for local calibration only to clients assigned the corresponding class, with allocation weights proportional to their detector-construction sample counts for that class. This keeps calibration within each client's selected class support while holding the total detector-construction budget fixed across coverage levels.

At each N_{cls}, every post-hoc method receives the same complete selected client-local subset; FedOPhD partitions that subset internally into its attenuation and reference-fitting roles. Tables 8 and 9 provide the near- and far-OOD means; the figure additionally shows one standard deviation over five random seeds. Table 10 complements these ranking metrics with deployment behavior under an L-ID-calibrated threshold. FedGradPCA requires a positive-dimensional gradient subspace; under its class-count-based rank choice, $N_{\text{cls}} = 1$ gives rank $N_{\text{cls}} - 1 = 0$ and does not define the required subspace. We therefore mark this setting as unavailable in Tables 8–10. In this controlled post-training setting, FedOPhD generally improves as local class coverage increases, with non-monotonic fluctuations in both AUROC and FPR95 for near- and far-OOD detection. FedOPhD achieves the lowest near-OOD FPR95 at every coverage level and the highest near-OOD AUROC for $N_{\text{cls}} \geq 2$; at $N_{\text{cls}} = 1$, FedEnergy has a slightly higher AUROC (83.73% versus 83.63%). For far-OOD detection, FedOPhD remains strongest on both metrics across all reported coverage levels. These results indicate that FedOPhD remains effective when a client's data cover only a small part of the federation-wide label space.

Table 8: Near-OOD results (%) for controlled post-training class coverage on CIFAR-10. Each entry reports AUROC / FPR95; – denotes an unavailable result.

N_{cls}	FedEnergy	FedViM	FedNCI	FedGradPCA	FedRAS	FedOPhD
1	**83.73** / 50.86	33.91 / 88.76	71.24 / 75.10	–	76.98 / 58.49	83.63 / **47.81**
2	83.73 / 50.86	38.45 / 83.17	72.09 / 76.01	55.87 / 85.43	77.29 / 57.96	**84.20 / 48.13**
3	83.73 / 50.86	44.46 / 77.47	73.75 / 74.99	62.70 / 77.68	77.35 / 57.33	**84.79 / 46.50**
4	83.73 / 50.86	50.78 / 72.33	74.43 / 75.38	66.30 / 71.71	77.43 / 57.16	**85.84 / 44.89**
5	83.73 / 50.86	57.22 / 68.09	75.00 / 73.51	68.71 / 66.10	77.50 / 57.08	**85.93 / 43.63**
6	83.73 / 50.86	62.51 / 64.03	75.51 / 73.35	72.97 / 63.02	77.53 / 57.08	**86.61 / 41.96**
7	83.73 / 50.86	67.50 / 60.66	75.57 / 72.67	76.54 / 59.58	77.56 / 57.98	**87.01 / 40.67**
8	83.73 / 50.86	74.30 / 57.22	75.21 / 71.96	80.72 / 55.79	77.61 / 57.88	**86.63 / 41.34**
9	83.73 / 50.86	78.79 / 54.12	75.15 / 71.63	81.92 / 53.89	77.61 / 57.86	**87.90 / 39.87**
10	83.73 / 50.86	82.90 / 51.31	75.02 / 71.52	83.20 / 52.02	77.63 / 58.75	**87.26 / 40.79**

Seen- and unseen-class ID preservation. For each controlled coverage level, we additionally partition held-out federation-level ID by whether its class is included in the client's selected N_{cls} classes. Using the same locally calibrated threshold as the deployment experiment, Table 10 reports false rejection for seen-class ID and unseen-class ID together with near- and far-OOD detection. S-FRR and U-FRR denote seen- and unseen-class false-rejection rates, respectively; $\Delta\text{FRR} = \text{U-FRR} - \text{S-FRR}$ measures additional rejection in percentage points, and Near-DR and Far-DR denote OOD detection rates. Since all CIFAR-10 classes are locally covered at $N_{\text{cls}} = 10$, the unseen-class population is empty and U-FRR and ΔFRR are undefined; Table 10 therefore reports $N_{\text{cls}} = 1, \ldots, 9$, while the full-coverage ranking results remain available in Tables 8–9.

Table 9: Far-OOD results (%) for controlled post-training class coverage on CIFAR-10. Each entry reports AUROC / FPR95; – denotes an unavailable result.

N_{cls}	FedEnergy	FedViM	FedNCI	FedGradPCA	FedRAS	FedOPhD
1	85.14 / 45.22	37.49 / 84.08	70.33 / 70.56	–	77.57 / 59.24	**95.97 / 21.43**
2	85.14 / 45.22	43.86 / 77.59	72.70 / 70.74	56.07 / 82.07	77.93 / 59.90	**96.11 / 18.18**
3	85.14 / 45.22	51.44 / 71.29	74.81 / 68.62	63.33 / 74.43	77.88 / 59.80	**96.21 / 17.42**
4	85.14 / 45.22	58.14 / 64.83	75.81 / 67.31	67.67 / 70.12	77.94 / 58.68	**96.77 / 15.51**
5	85.14 / 45.22	65.45 / 64.25	75.94 / 67.13	66.20 / 69.06	77.04 / 58.72	**97.06 / 15.96**
6	85.14 / 45.22	70.48 / 53.93	77.99 / 64.17	72.44 / 58.42	79.08 / 58.71	**97.31 / 14.81**
7	85.14 / 45.22	75.97 / 48.51	78.23 / 63.62	75.81 / 54.81	80.13 / 58.67	**97.31 / 13.64**
8	85.14 / 45.22	83.40 / 42.58	78.02 / 63.44	83.14 / 48.51	80.88 / 57.97	**97.19 / 13.97**
9	85.14 / 45.22	87.84 / 37.01	77.75 / 63.55	84.34 / 45.55	81.39 / 57.77	**97.07 / 13.28**
10	85.14 / 45.22	91.50 / 32.49	77.83 / 63.08	85.81 / 42.90	81.72 / 57.91	**96.83 / 12.97**

FedOPhD maintains far-OOD detection above 90% throughout the tested range while its additional unseen-class rejection remains moderate and generally contracts as coverage broadens. FedViM and FedGradPCA exhibit substantially larger unseen-class rejection, showing that their apparent OOD sensitivity is closely coupled to incomplete local class coverage.

Table 10: Local-class-coverage deployment on CIFAR-10 (%; five-seed means). S-FRR has a calibration target of 5% and is not ranked. Bold marks the lowest U-FRR and ΔFRR and the highest Near-DR and Far-DR across methods at each N_{cls}.

Method	Metric	Local class coverage N_{cls}								
		1	2	3	4	5	6	7	8	9
FedEnergy	S-FRR	4.49	4.89	5.07	4.61	5.13	4.74	5.04	4.91	4.95
	U-FRR↓	**6.36**	**6.03**	**6.46**	**5.28**	**5.68**	**5.36**	**5.09**	**5.04**	**5.05**
	ΔFRR↓	**+1.87**	**+1.14**	**+1.39**	**+0.67**	**+0.55**	**+0.62**	**+0.05**	**+0.13**	**+0.10**
	Near-DR↑	41.49	39.97	40.11	36.42	38.13	36.36	37.05	36.35	36.55
	Far-DR↑	52.73	51.45	51.54	48.33	49.89	48.36	49.07	48.41	48.61
FedViM	S-FRR	5.10	4.83	5.03	4.98	5.09	4.43	4.42	4.48	5.74
	U-FRR↓	93.65	87.96	84.95	81.98	74.84	72.02	69.50	55.19	57.04
	ΔFRR↓	+88.55	+83.13	+79.92	+77.00	+69.75	+67.59	+65.08	+50.71	+51.30
	Near-DR↑	**86.20**	**68.83**	**59.68**	**54.85**	**44.85**	38.21	32.92	28.72	32.16
	Far-DR↑	91.02	78.67	71.89	70.24	64.33	58.97	56.03	54.90	59.83
FedNCI	S-FRR	5.44	5.92	5.25	5.05	4.94	4.94	5.28	5.03	5.08
	U-FRR↓	7.37	7.96	8.19	7.35	7.65	8.00	8.13	7.92	7.40
	ΔFRR↓	+1.93	+2.04	+2.94	+2.30	+2.71	+3.06	+2.85	+2.89	+2.32
	Near-DR↑	4.24	13.67	18.54	20.43	25.28	27.50	29.39	27.93	28.57
	Far-DR↑	4.63	16.56	22.13	24.10	30.56	32.85	35.46	34.05	34.47
FedGradPCA	S-FRR	–	5.71	5.12	5.05	5.26	5.19	5.08	4.56	4.88
	U-FRR↓	–	35.86	33.83	33.00	24.95	19.63	14.17	11.99	10.06
	ΔFRR↓	–	+30.15	+28.71	+27.95	+19.69	+14.44	+9.09	+7.43	+5.18
	Near-DR↑	–	34.07	37.60	35.29	34.66	33.45	31.94	28.84	29.29
	Far-DR↑	–	35.00	38.60	38.86	39.07	37.50	36.15	35.67	36.11
FedRAS	S-FRR	4.83	4.53	5.16	4.43	4.87	4.71	5.06	5.00	5.11
	U-FRR↓	9.52	7.66	7.53	7.41	6.60	7.55	6.20	6.52	6.04
	ΔFRR↓	+4.69	+3.13	+2.37	+2.98	+1.73	+2.84	+1.14	+1.52	+0.93
	Near-DR↑	48.09	41.03	42.70	36.67	38.15	37.11	37.89	38.11	38.19
	Far-DR↑	56.65	50.65	51.81	45.81	47.73	46.89	47.79	48.13	48.22
FedOPhD (ours)	S-FRR	5.05	4.95	5.44	4.79	5.02	5.09	4.69	4.86	5.54
	U-FRR↓	9.36	9.21	8.37	8.03	7.10	7.17	6.75	7.23	6.95
	ΔFRR↓	+4.31	+4.26	+2.93	+3.24	+2.08	+2.08	+2.06	+2.37	+1.41
	Near-DR↑	54.07	53.70	48.79	46.86	41.96	**42.54**	**39.08**	**39.10**	**42.54**
	Far-DR↑	**91.99**	**91.02**	**92.50**	**93.24**	**91.64**	**92.22**	**91.49**	**90.90**	**91.51**

E.2 Mechanism and Component Analyses

E.2.1 Information-theoretic diagnostic

To quantify the complementary evidence contributed by the two supports, we report a conditional-information diagnostic together with its estimation and aggregation protocol.

Signals and samples. The diagnostic examines G, the local-reference ECDF support of the original global model's negative energy, and R, the response support obtained from the multi-direction readout and equal-weight layer aggregation in Equation 4. Only these two supports enter the diagnostic. We use the complete official ID test set and the complete fixed OOD evaluation sources for this diagnostic. The ID population contains all official test samples from the participating clients, including both L-ID and T-ID, and retains every sample regardless of local class coverage or prediction correctness. Each client and each OOD dataset are evaluated separately before near-/far-OOD summarization. Both supports are constructed and calibrated using current-client ID data only; query ID/OOD annotations are used exclusively to compute this diagnostic.

Equal-prior information scale. Let $Y = 0$ denote ID and $Y = 1$ denote OOD. For this diagnostic, each population receives total probability $1/2$, giving $H(Y) = 1$ bit before observing either support. This equal-prior weighting leaves the sample set unchanged and is used only for the diagnostic; AUROC and FPR95 follow the evaluation protocol above. The one-bit value is the entropy of the balanced label prior, and the captured information is determined by the estimates below.

Binned estimates. Six equal-width intervals on $[0, 1]$ discretize the supports into $\widetilde{G}$ and $\widetilde{R}$. For $Z \in \{\widetilde{G}, \widetilde{R}, (\widetilde{G}, \widetilde{R})\}$, let $n_{b,y}$ count samples in bin b with label y and $n_y = \sum_b n_{b,y} > 0$. We use the empirical joint distribution

$$\widehat{P}(Z = b, Y = y) = \frac{n_{b,y}}{2n_y}, \qquad \widehat{I}(Y; Z) = 1 - \widehat{H}(Y \mid Z), \tag{41}$$

where $\widehat{H}(Y \mid Z)$ is conditional entropy under $\widehat{P}$, computed with base-two logarithms and zero contribution from zero-probability cells. The raw estimates lie in $[0, 1]$ bit and quantify uncertainty reduction at the chosen six-bin resolution.

Conditional permutation correction. Conditional information quantifies information about a label remaining beyond another observed variable (Fleuret, 2004). The raw response increment and its corrected diagnostic are

$$\begin{aligned} \widehat{I}_{R|G} &= \widehat{I}(Y; \widetilde{G}, \widetilde{R}) - \widehat{I}(Y; \widetilde{G}), \\ \Delta^{\mathrm{corr}}_{R|G} &= \widehat{I}_{R|G} - \frac{1}{M} \sum_{m=1}^{M} \widehat{I}^{(m)}_{R|G}, \qquad M = 199. \end{aligned} \tag{42}$$

For permutation m, we shuffle $\widetilde{R}$ across samples within each $\widetilde{G}$ bin, leaving the conditioning bin and label of each sample fixed. This preserves the within-bin marginal counts while breaking the observed response–label pairing; the same prior-weighted estimator is recomputed after each shuffle. Swapping G and R gives $\Delta^{\mathrm{corr}}_{G|R}$. Subtracting the permutation mean centers the plug-in increment on a finite-sample conditional-permutation reference.

Aggregation. For each seed, corrected increments are averaged equally across OOD datasets within each near-/far-OOD group and then equally across clients; Table 11 reports the mean across five seeds. This client-equal diagnostic differs from the ID-size-weighted AUROC/FPR95 aggregation in Appendix D.9 and is separate from the within-pair ID/OOD prior weighting.

Table 11: Permutation-corrected conditional information increments (bits; five-seed means). $G \mid R$ denotes the increment from G given R, and conversely.

α	CIFAR-10				CIFAR-100			
	Near-OOD		Far-OOD		Near-OOD		Far-OOD	
	$G \mid R$	$R \mid G$	$G \mid R$	$R \mid G$	$G \mid R$	$R \mid G$	$G \mid R$	$R \mid G$
0.1	0.2937	0.0948	0.2614	0.2591	0.1394	0.0823	0.1547	0.2916
0.5	0.3536	0.0762	0.2810	0.2490	0.1530	0.0736	0.1381	0.3122
10	0.2958	0.1050	0.1915	0.3042	0.1604	0.0906	0.1361	0.2955

Interpretation. Table 11 reports positive corrected increments in both directions for every reported setting. The reported quantity is the plug-in increment centered by the mean conditional-permutation increment; chain-rule identities apply to the corresponding raw estimates. At the chosen six-bin resolution, the results provide consistent descriptive evidence that global recognition and response alignment contribute complementary ID/OOD information. The detector-level effect of combining these signals is evaluated directly in the fusion comparisons.

E.2.2 Numerical results for fusion strategies

To provide the exact values underlying the main-text fusion comparison, we report all three heterogeneity levels in Tables 12 and 13. The CIFAR-10 table supplements the $\alpha = 0.1$ values visualized in Figure 5 with $\alpha = 0.5$ and $\alpha = 10$, while the CIFAR-100 table checks whether the same comparison pattern extends to a second ID dataset. Across both ID datasets and all reported settings, FedOPhD yields the highest mean AUROC and the lowest mean FPR95, while the strongest alternative varies across OOD regime and heterogeneity level.

Table 12: Complete numerical results for the CIFAR-10 fusion comparison (mean ± standard deviation, %).

α	Fusion strategy	Near-OOD		Far-OOD	
		AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
0.1	Arithmetic mean	81.39 ± 1.17	54.75 ± 1.38	94.71 ± 1.07	23.67 ± 2.06
	Geometric mean	81.20 ± 1.13	51.05 ± 1.29	95.93 ± 1.14	18.07 ± 2.17
	GLRT score combining	81.22 ± 1.07	51.97 ± 1.39	95.93 ± 1.23	18.35 ± 2.05
	BH-based conformal test	77.05 ± 1.33	65.15 ± 1.54	93.82 ± 1.07	23.59 ± 1.92
	FedOPhD	**86.03** ± 1.18	**43.96** ± 1.31	**97.13** ± 0.77	**13.96** ± 2.18
0.5	Arithmetic mean	84.80 ± 1.04	49.98 ± 1.13	95.02 ± 1.19	20.25 ± 2.02
	Geometric mean	85.46 ± 1.06	48.12 ± 1.20	96.38 ± 1.12	10.80 ± 2.03
	GLRT score combining	85.77 ± 1.33	49.31 ± 1.21	96.03 ± 1.15	14.62 ± 2.10
	BH-based conformal test	85.89 ± 1.28	51.03 ± 1.29	96.75 ± 1.27	12.73 ± 1.91
	FedOPhD	**89.15** ± 1.05	**39.30** ± 1.13	**98.32** ± 1.07	**7.66** ± 2.15
10	Arithmetic mean	85.78 ± 1.08	47.82 ± 1.17	95.55 ± 1.08	20.21 ± 2.01
	Geometric mean	85.03 ± 1.15	44.56 ± 1.22	96.06 ± 1.13	11.33 ± 1.95
	GLRT score combining	85.39 ± 1.03	42.78 ± 1.27	95.54 ± 1.43	21.98 ± 2.07
	BH-based conformal test	87.01 ± 1.08	45.22 ± 1.32	96.00 ± 0.95	13.91 ± 1.96
	FedOPhD	**89.52** ± 1.06	**39.57** ± 1.31	**98.43** ± 1.06	**6.72** ± 2.04

Table 13: Numerical results for the CIFAR-100 fusion comparison (mean ± standard deviation, %).

α	Fusion strategy	Near-OOD		Far-OOD	
		AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
0.1	Arithmetic mean	65.71 ± 1.17	80.27 ± 1.58	92.61 ± 1.11	34.28 ± 1.40
	Geometric mean	72.09 ± 1.20	73.49 ± 1.64	92.02 ± 1.06	30.54 ± 1.33
	GLRT score combining	67.42 ± 1.78	82.18 ± 2.37	92.86 ± 1.95	25.69 ± 2.04
	BH-based conformal test	70.62 ± 2.15	76.96 ± 1.70	89.54 ± 2.13	32.92 ± 2.09
	FedOPhD	**76.79** ± 2.21	**67.45** ± 1.67	**94.06** ± 1.31	**23.80** ± 3.11
0.5	Arithmetic mean	66.56 ± 1.19	73.19 ± 1.45	93.42 ± 1.11	29.74 ± 1.39
	Geometric mean	74.23 ± 1.23	70.84 ± 1.66	91.64 ± 1.06	32.23 ± 1.48
	GLRT score combining	70.67 ± 2.45	75.32 ± 2.39	92.40 ± 1.98	27.43 ± 2.36
	BH-based conformal test	74.33 ± 1.71	69.50 ± 2.39	85.68 ± 2.09	38.06 ± 1.87
	FedOPhD	**78.72** ± 1.12	**61.72** ± 1.58	**94.29** ± 1.51	**21.87** ± 3.61
10	Arithmetic mean	66.55 ± 1.18	73.20 ± 1.39	93.10 ± 1.07	29.26 ± 1.22
	Geometric mean	75.01 ± 1.13	68.03 ± 1.44	90.98 ± 1.03	34.01 ± 1.30
	GLRT score combining	71.46 ± 1.77	74.79 ± 1.82	91.64 ± 1.63	28.91 ± 1.72
	BH-based conformal test	75.31 ± 2.04	68.60 ± 1.71	84.77 ± 2.06	40.10 ± 2.28
	FedOPhD	**79.58** ± 1.09	**58.71** ± 1.57	**93.52** ± 1.41	**23.63** ± 4.14

E.2.3 FUNCTIONAL ROUTING AND GEOMETRIC SCORING

FedOPhD assigns complementary roles to the two direction banks: $\mathbf{b}_{\ell,j}$ routes a query according to the response patterns most strongly coupled to decreased original-class recognition, while $\boldsymbol{\mu}_{\ell,j}$ scores its agreement with the corresponding reference-response geometry. To isolate these roles, Table 14 compares the default $\mathbf{b} \to \boldsymbol{\mu}$ readout with functional routing and scoring ($\mathbf{b} \to \mathbf{b}$) and geometric routing and scoring ($\boldsymbol{\mu} \to \boldsymbol{\mu}$).

Holding functional routing fixed, replacing $\mathbf{b}$-based scoring with $\boldsymbol{\mu}$-based scoring improves every reported mean metric. Holding geometric scoring fixed, replacing $\boldsymbol{\mu}$-based routing with $\mathbf{b}$-based routing yields the same consistent improvement. The default $\mathbf{b} \to \boldsymbol{\mu}$ construction consequently achieves the highest AUROC and lowest FPR95 in every evaluated dataset, OOD regime, and heterogeneity setting, supporting the intended division between function-aware mode selection and geometric agreement scoring.

Table 14: Functional-routing and geometric-scoring ablation (%; five-seed means). $\mathbf{b} \to \boldsymbol{\mu}$ is FedOPhD; $\mathbf{b} \to \mathbf{b}$ uses functional directions for routing and scoring; $\boldsymbol{\mu} \to \boldsymbol{\mu}$ uses geometric directions for both.

ID dataset	Routing → scoring	α	Near AUROC↑	Near FPR95↓	Far AUROC↑	Far FPR95↓
CIFAR-10	$\mathbf{b} \to \boldsymbol{\mu}$ (FedOPhD)	0.1	**86.03**	**43.96**	**97.13**	**13.96**
		0.5	**89.15**	**39.30**	**98.32**	**7.66**
		10	**89.52**	**39.57**	**98.43**	**6.72**
	$\mathbf{b} \to \mathbf{b}$	0.1	85.53	48.06	94.00	17.09
		0.5	87.74	41.19	95.23	15.93
		10	88.50	44.10	96.32	11.31
	$\boldsymbol{\mu} \to \boldsymbol{\mu}$	0.1	85.41	45.39	93.93	24.75
		0.5	87.09	43.69	95.14	14.08
		10	87.47	43.07	95.81	12.59
CIFAR-100	$\mathbf{b} \to \boldsymbol{\mu}$ (FedOPhD)	0.1	**76.79**	**67.45**	**94.06**	**23.80**
		0.5	**78.72**	**61.72**	**94.29**	**21.87**
		10	**79.58**	**58.71**	**93.52**	**23.63**
	$\mathbf{b} \to \mathbf{b}$	0.1	74.51	68.03	90.33	37.29
		0.5	77.00	64.18	91.28	34.09
		10	77.33	64.26	89.76	42.31
	$\boldsymbol{\mu} \to \boldsymbol{\mu}$	0.1	73.15	67.51	92.39	29.06
		0.5	75.58	66.32	92.04	31.35
		10	76.28	63.83	91.20	35.21

E.2.4 DIAGNOSTIC-INTERVENTION COMPARISON

To assess the contribution of the client-conditioned intervention, Tables 15 and 16 compare semantic attenuation with uniform redistribution, random parameter perturbation, and forward training within the complete FedOPhD pipeline.

These experiments use controlled comparisons to evaluate the choice of diagnostic intervention within the FedOPhD framework. We evaluate the resulting response-based detection scores and their fusion with global recognition. Across the evaluated settings, the results favor semantic attenuation overall among the tested alternatives.

Uniform redistribution removes the true-class probability and distributes it equally over the remaining classes, omitting the client-support term in Equation 1. Random parameter perturbation draws five independent Gaussian directions over all trainable parameters and rescales each direction so that its global ℓ_2 update norm matches the corresponding fresh semantic-attenuation step for the same client; all model buffers, including normalization statistics, remain fixed. The comparison matches the stepwise global norm, while endpoint displacement and functional change follow the sampled directions. Conditional on the fixed client data, which determine the matched radii, these directions are isotropic and independent of client examples and labels. Forward training minimizes true-label cross-entropy on the same adaptation subset using five full empirical-gradient steps, learning rate 10^{-4}, SGD with zero momentum and zero weight decay, and the same trainable parameter set as semantic attenuation. The semantic-attenuation entries are taken directly from Table 1.

Across the two ID datasets, three heterogeneity levels, and both OOD regimes, semantic attenuation leads twenty-three of the twenty-four mean AUROC/FPR95 comparisons; uniform redistribution attains the highest CIFAR-100 far-OOD AUROC at $\alpha = 10$. Overall, the results show a broadly favorable trade-off for client-conditioned semantic attenuation across the evaluated settings.

Table 15: Full FedOPhD diagnostic-intervention results on CIFAR-10 (%); only the diagnostic intervention varies. Each entry is AUROC ↑ / FPR95 ↓; Figure 6 visualizes $\alpha = 0.1$.

α	Diagnostic intervention	Near-OOD	Far-OOD
0.1	Uniform redistribution	82.70 / 50.04	95.93 / 14.68
	Random parameter perturbation	78.57 / 60.22	80.37 / 53.36
	Forward training	81.17 / 53.53	91.78 / 22.45
	Semantic attenuation	**86.03 / 43.96**	**97.13 / 13.96**
0.5	Uniform redistribution	88.01 / 41.09	96.28 / 7.81
	Random parameter perturbation	85.79 / 42.13	82.61 / 45.37
	Forward training	85.23 / 41.65	88.54 / 24.67
	Semantic attenuation	**89.15 / 39.30**	**98.32 / 7.66**
10	Uniform redistribution	89.47 / 40.42	98.24 / 8.33
	Random parameter perturbation	85.55 / 48.58	89.74 / 31.21
	Forward training	85.49 / 49.84	87.51 / 26.94
	Semantic attenuation	**89.52 / 39.57**	**98.43 / 6.72**

Table 16: Full FedOPhD diagnostic-intervention results on CIFAR-100 (%); only the diagnostic intervention varies. Each entry is AUROC ↑ / FPR95 ↓.

α	Diagnostic intervention	Near-OOD	Far-OOD
0.1	Uniform redistribution	72.84 / 74.81	92.59 / 33.62
	Random parameter perturbation	69.02 / 74.31	68.38 / 75.06
	Forward training	71.50 / 77.38	72.49 / 63.87
	Semantic attenuation	**76.79 / 67.45**	**94.06 / 23.80**
0.5	Uniform redistribution	76.84 / 63.45	93.84 / 37.15
	Random parameter perturbation	73.97 / 65.98	74.39 / 72.09
	Forward training	72.69 / 67.30	70.45 / 67.38
	Semantic attenuation	**78.72 / 61.72**	**94.29 / 21.87**
10	Uniform redistribution	79.02 / 59.65	**93.62** / 40.70
	Random parameter perturbation	74.41 / 63.76	73.37 / 74.03
	Forward training	73.70 / 65.64	76.49 / 54.87
	Semantic attenuation	**79.58 / 58.71**	93.52 / **23.63**

Response-only comparison across diagnostic interventions. To isolate the discriminative evidence contributed through the response channel, we repeat the intervention comparison using only s_{R} while retaining the same intervention-specific endpoints and evaluation protocol. Semantic attenuation in this analysis is FedOPhD-R, the response-only ablation reported in the main tables.

Tables 17 and 18 show that semantic attenuation attains the highest response-only AUROC and lowest FPR95 across the evaluated settings, with a tie in CIFAR-100 near-OOD AUROC at $\alpha = 10$. Uniform redistribution is the closest control overall; its far-OOD results on CIFAR-10 are particularly close to semantic attenuation. The comparison supports semantic attenuation as an effective response-generating intervention without implying a large advantage in every setting. Together, the response-only and fused-score results characterize intervention-dependent response evidence before and after fusion with global recognition.

Table 17: Response-only diagnostic-intervention results on CIFAR-10 (%; five-seed means), reported as AUROC ↑ / FPR95 ↓.

α	Diagnostic intervention	Near-OOD	Far-OOD
0.1	Uniform redistribution	60.94 / 85.04	91.66 / 35.93
	Random parameter perturbation	49.74 / 94.60	43.54 / 92.24
	Forward training	52.77 / 94.36	84.34 / 55.75
	Semantic attenuation	**64.71 / 84.85**	**91.79 / 35.66**
0.5	Uniform redistribution	63.79 / 87.77	94.76 / 25.09
	Random parameter perturbation	50.84 / 94.53	55.79 / 87.08
	Forward training	53.72 / 94.54	73.79 / 73.34
	Semantic attenuation	**65.05 / 80.16**	**94.82 / 24.86**
10	Uniform redistribution	69.77 / 83.67	96.65 / 16.58
	Random parameter perturbation	54.49 / 93.09	60.13 / 85.01
	Forward training	52.95 / 94.29	72.05 / 74.06
	Semantic attenuation	**69.88 / 78.48**	**96.73 / 16.31**

Table 18: Response-only diagnostic-intervention results on CIFAR-100 (%; five-seed means), reported as AUROC ↑ / FPR95 ↓.

α	Diagnostic intervention	Near-OOD	Far-OOD
0.1	Uniform redistribution	53.04 / 88.68	89.97 / 39.19
	Random parameter perturbation	50.65 / 93.34	52.92 / 86.58
	Forward training	45.12 / 96.17	68.46 / 74.78
	Semantic attenuation	**58.15 / 85.97**	**91.04 / 35.04**
0.5	Uniform redistribution	54.77 / 86.95	91.10 / 36.32
	Random parameter perturbation	50.54 / 93.14	60.47 / 80.21
	Forward training	46.18 / 95.21	65.55 / 74.39
	Semantic attenuation	**58.84 / 84.88**	**92.39 / 30.16**
10	Uniform redistribution	**59.21** / 84.96	89.49 / 40.10
	Random parameter perturbation	51.01 / 92.92	60.07 / 81.59
	Forward training	47.38 / 94.89	65.27 / 73.77
	Semantic attenuation	**59.21 / 84.01**	**90.64 / 38.85**

E.2.5 STATIC-REPRESENTATION TRANSFER

Existing post-hoc OOD detectors can extract OOD evidence directly from static representations of a trained classifier. A natural question is whether such detectors remain deployable when their reference data are restricted to a single client under strong federated heterogeneity, and whether transferring their representation readout can recover the response-side discrimination provided by FedOPhD. We therefore compare FedOPhD-R with four representative static-representation detectors: SHE (Zhang et al., 2023), Mahalanobis (Lee et al., 2018), Deep kNN (Sun et al., 2022), and Gram (Sastry & Oore, 2020).

Protocol. We evaluate CIFAR-10 and CIFAR-100 under the strongly heterogeneous $\alpha = 0.1$ setting with 20 clients and five random seeds. For each baseline, the *Original* variant follows the detector definition and implementation choices of the corresponding paper, using the authors' public implementation when available, with the trained FedAvg checkpoint and deployment-client ID data supplied under our common federated protocol. We additionally evaluate a *Transferred* variant that moves the same method-specific detector to the intermediate representation stages used by FedOPhD-R while retaining its original scoring principle and fitting procedure. No response intervention or global-recognition fusion is introduced in either variant. This comparison therefore evaluates whether static representation evidence, either in its original form or after representation transfer, can substitute for attenuation-induced response evidence under restricted client-local deployment.

Table 19: Static-representation transfer at $\alpha = 0.1$ (20 clients; five-seed means, %). No global-recognition fusion is applied.

		Near-OOD		Far-OOD	
Method	**Setting**	AUROC↑	FPR95↓	AUROC↑	FPR95↓
		CIFAR-10			
SHE	Original	58.46	86.74	77.47	49.16
SHE	Transferred	58.02	90.56	61.20	75.62
Mahalanobis	Original	40.26	87.05	47.50	79.20
Mahalanobis	Transferred	55.19	90.10	79.26	60.31
Deep kNN	Original	42.94	89.94	50.54	73.77
Deep kNN	Transferred	54.73	88.23	82.06	57.88
Gram	Original	53.42	86.84	74.54	52.42
Gram	Transferred	53.73	89.12	76.01	54.59
FedOPhD-R	Original	**64.71**	**84.85**	**91.79**	**35.66**
		CIFAR-100			
SHE	Original	52.91	97.95	51.63	78.14
SHE	Transferred	53.04	89.55	58.51	70.43
Mahalanobis	Original	44.28	93.73	53.07	87.76
Mahalanobis	Transferred	45.50	95.00	80.34	54.78
Deep kNN	Original	39.23	91.40	48.08	84.55
Deep kNN	Transferred	43.76	96.34	80.73	58.59
Gram	Original	54.91	92.53	55.99	67.88
Gram	Transferred	49.97	96.43	65.88	68.56
FedOPhD-R	Original	**58.15**	**85.97**	**91.04**	**35.04**

Results. The static-representation detectors exhibit substantially different transfer behavior under strongly heterogeneous federated deployment. Transferring Mahalanobis and Deep kNN to the intermediate representations considerably improves their far-OOD AUROC on both CIFAR-10 and CIFAR-100, whereas SHE degrades on CIFAR-10 and Gram exhibits dataset-dependent changes: its far-OOD AUROC improves modestly on CIFAR-10 and more substantially on CIFAR-100, while its FPR95 does not improve on either dataset. Thus, access to intermediate representations can strengthen some static detectors, but the benefit is neither universal nor sufficient to match the response-based detector.

FedOPhD-R achieves the highest near- and far-OOD AUROC on both datasets and substantially lower far-OOD FPR95 than all evaluated static-representation detectors. In particular, the strongest

transferred static detector reaches far-OOD AUROC of 82.06 on CIFAR-10 and 80.73 on CIFAR-100, compared with 91.79 and 91.04 for FedOPhD-R, respectively. These results show that the response-side advantage of FedOPhD cannot be explained solely by transferring established static OOD detectors to alternative representations of the frozen global model; the attenuation-induced representation changes provide additional discriminative evidence under strong federated heterogeneity.

E.2.6 Comparison with perturbation-based post-hoc detectors

Recent post-hoc detectors actively perturb classifier weights, input images, or intermediate features to expose OOD sensitivity. We compare FedOPhD with representative methods from these three families to quantify the benefit of client-conditioned semantic intervention in federated deployment. All methods start from the same frozen FedAvg checkpoints and are evaluated using the federation-level protocol in Appendix D.9; all scores are oriented so that larger values indicate stronger OOD evidence.

WeiPer+KLD (Granz et al., 2024) extracts penultimate-layer features and forms 100 fixed projections by adding norm-scaled Gaussian perturbations to the classifier weights. In our federated adaptation, each client uses its complete local FL training dataset to construct reference histograms for the original features and perturbed projections. The reproduced score combines the symmetric KL discrepancies of the two query histograms from their local references with the mean MSP across perturbed classifiers. **PRO-MSP**, **PRO-ENT**, and **PRO-GEN** instantiate Perturbation-Rectified OOD detection (Chen et al., 2025) with different softmax-based confidence functions. PRO-MSP takes one signed input-gradient step of size 0.003 at temperature one to reduce MSP and uses the negative minimum MSP over the original and perturbed inputs. PRO-ENT takes two signed steps of size 0.0014 to reduce negative predictive entropy and uses the largest entropy along the resulting trajectory. PRO-GEN uses $E_{\mathrm{GEN}} = \sum_{j \in \mathrm{Top}\text{-}M} p_j^{\gamma}(1-p_j)^{\gamma}$ with $\gamma = 0.1$ and $M = 100$ (or all classes when $C < 100$) (Liu et al., 2023), applies three signed steps of size 10^{-4}, and uses the largest generalized entropy along the trajectory. **Gradient Short-Circuit (GSC)** (Gu et al., 2025) computes the predicted-class logit gradient with respect to the penultimate feature, zeros the 5% most sensitive coordinates (25 of 512 here), updates the logits for the linear classifier, and applies negative LogSumExp to the intervened logits. The three PRO variants and GSC operate directly on each query, while WeiPer+KLD incorporates the fitted L-ID reference histograms.

Tables 20 and 21 report the complete five-seed means. FedOPhD achieves the highest mean AUROC and lowest mean FPR95 for both near- and far-OOD in every dataset–heterogeneity combination. The distinction is most important at $\alpha = 0.1$: this strongly non-IID regime produces fragmented client-level semantic coverage and therefore most closely reflects the client heterogeneity that motivates federated OOD detection. Under this setting, WeiPer+KLD performs substantially worse on CIFAR-10 than at the larger α values, whereas FedOPhD remains consistently effective, with especially large advantages on far-OOD. On CIFAR-100, GSC improves on both near-OOD metrics as α increases, while its far-OOD metrics are best at $\alpha = 0.5$. FedOPhD retains the leading results throughout. These results establish the value of semantically directed intervention for robust FL OOD detection: FedOPhD suppresses client-supported knowledge in a controlled manner and combines the resulting response with the global model's recognition evidence.

Table 20: Comparison with perturbation-based post-hoc detectors on CIFAR-10 (%). Entries are means over five random seeds.

Method	α	Near AUROC↑	Near FPR95↓	Far AUROC↑	Far FPR95↓
WeiPer+KLD	0.1	65.87	78.23	68.65	67.51
	0.5	87.19	49.32	90.53	32.15
	10	87.31	41.16	92.01	26.78
PRO-MSP	0.1	82.40	49.81	83.46	52.21
	0.5	83.11	43.58	90.32	31.33
	10	83.05	46.91	88.24	33.53
PRO-ENT	0.1	83.54	47.75	85.94	48.99
	0.5	84.31	43.09	91.79	29.23
	10	84.10	44.91	89.39	33.10
PRO-GEN	0.1	83.18	48.14	88.14	41.67
	0.5	85.07	41.38	93.80	22.85
	10	83.96	46.51	90.99	32.97
Gradient Short-Circuit	0.1	74.04	70.45	83.73	49.08
	0.5	81.52	60.55	86.88	40.90
	10	79.61	71.28	82.19	56.95
FedOPhD (ours)	0.1	**86.03**	**43.96**	**97.13**	**13.96**
	0.5	**89.15**	**39.30**	**98.32**	**7.66**
	10	**89.52**	**39.57**	**98.43**	**6.72**

Table 21: Comparison with perturbation-based post-hoc detectors on CIFAR-100 (%). Entries are means over five random seeds.

Method	α	Near AUROC↑	Near FPR95↓	Far AUROC↑	Far FPR95↓
WeiPer+KLD	0.1	71.09	72.77	76.68	66.69
	0.5	74.13	65.29	75.62	65.04
	10	76.99	60.38	77.33	64.11
PRO-MSP	0.1	72.58	70.23	74.28	72.15
	0.5	73.68	64.55	74.29	66.19
	10	73.26	67.89	73.95	65.83
PRO-ENT	0.1	73.25	69.46	76.93	67.12
	0.5	74.51	63.20	76.40	63.96
	10	74.25	65.87	76.02	65.04
PRO-GEN	0.1	72.91	70.13	79.44	60.65
	0.5	75.80	63.43	79.37	57.60
	10	74.93	65.84	79.21	57.85
Gradient Short-Circuit	0.1	65.03	80.03	71.49	70.23
	0.5	67.17	77.97	73.02	65.15
	10	68.28	74.86	71.90	68.17
FedOPhD (ours)	0.1	**76.79**	**67.45**	**94.06**	**23.80**
	0.5	**78.72**	**61.72**	**94.29**	**21.87**
	10	**79.58**	**58.71**	**93.52**	**23.63**

E.2.7 POPULATION-WISE RESPONSE GEOMETRY UNDER SEMANTIC ATTENUATION

The following mechanism diagnostic resolves the response geometry of L-ID, T-ID, near-OOD, and far-OOD queries under semantic attenuation. We therefore examine the selected-mode cosine alignment before the excitation scaling, marginal rank transforms, and fusion used by FedOPhD. This diagnostic tests whether T-ID follows the locally learned ID response geometry while OOD inputs exhibit weaker agreement. Appendix D.4 specifies its routing, layer aggregation, and population-weighting protocol as a separate analysis of the deployed detector's responses. This diagnostic uses the run with random seed 42; its mean, standard deviation, median, and extrema describe the weighted client–query distribution within that run.

Table 22: Population-wise selected-mode cosine alignment under semantic attenuation on CIFAR-10 using the fixed diagnostic seed 42. SD is the weighted sample standard deviation across client–query observations. Higher values indicate closer agreement with local semantic-attenuation response geometry.

α	Role	Min	Max	Median	Mean $\pm$ SD
0.1	L-ID	0.8557	0.9908	0.9641	0.9623 $\pm$ 0.0137
	T-ID	0.7717	0.9902	0.9484	0.9446 $\pm$ 0.0211
	Near	0.6247	0.9871	0.9253	0.9202 $\pm$ 0.0232
	Far	0.4901	0.9841	0.9039	0.8950 $\pm$ 0.0441
0.5	L-ID	0.8678	0.9869	0.9546	0.9529 $\pm$ 0.0134
	T-ID	0.7702	0.9861	0.9492	0.9469 $\pm$ 0.0155
	Near	0.6610	0.9825	0.9232	0.9189 $\pm$ 0.0201
	Far	0.5050	0.9723	0.9018	0.8940 $\pm$ 0.0412
10	L-ID	0.8380	0.9828	0.9557	0.9537 $\pm$ 0.0130
	T-ID	0.8016	0.9838	0.9556	0.9536 $\pm$ 0.0131
	Near	0.7121	0.9824	0.9245	0.9200 $\pm$ 0.0207
	Far	0.5540	0.9751	0.8986	0.8919 $\pm$ 0.0392

Across every evaluated heterogeneity level, both mean and median alignment follow the same ordering: L-ID is highest, followed by T-ID, near-OOD, and far-OOD. T-ID remains closer to L-ID than to near-OOD, and the L-ID–T-ID gap narrows as the client partitions become less heterogeneous, supporting response geometry that transfers beyond the current client's own ID samples. Far-OOD has the lowest central tendency and the largest dispersion throughout, indicating weaker and less consistent agreement with the locally learned response modes. The complete FedOPhD score combines this geometric mechanism with response strength, local rank calibration, and global recognition.

E.2.8 UPDATE-MAGNITUDE AND STATIC-FEATURE CONTROLS

These controls address two successive questions: whether static feature geometry can substitute for intervention responses, and whether matching the update magnitude makes forward training an effective substitute for semantic attenuation. Together, they examine the role of intervention-induced responses and the choice of intervention within the response branch. We evaluate the controls both as standalone scores and within the complete fusion pipeline, separating response-side discrimination from its contribution after combination with global support.

Matched update norms. The control retains the forward cross-entropy direction but matches the global parameter-update norm of each corresponding semantic-attenuation step.

For each seed and client, let $\boldsymbol{\theta}_t^{\mathrm{A}}$ denote the semantic-attenuation trajectory and initialize the matched forward trajectory at the same checkpoint, $\boldsymbol{\theta}_0^{\mathrm{F}} = \boldsymbol{\theta}_0^{\mathrm{A}} = \boldsymbol{\theta}_0$. For $t = 0, \ldots, T-1$, define

$$
\begin{aligned}
r_t &= \|\boldsymbol{\theta}_{t+1}^{\mathrm{A}} - \boldsymbol{\theta}_t^{\mathrm{A}}\|_2, \qquad \mathbf{g}_t^{\mathrm{F}} = \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{CE}}(\boldsymbol{\theta}_t^{\mathrm{F}}; \mathcal{D}_{\mathrm{ada}}), \\
\boldsymbol{\theta}_{t+1}^{\mathrm{F}} &= \boldsymbol{\theta}_t^{\mathrm{F}} - r_t \frac{\mathbf{g}_t^{\mathrm{F}}}{\|\mathbf{g}_t^{\mathrm{F}}\|_2}, \\
\|\boldsymbol{\theta}_{t+1}^{\mathrm{F}} - \boldsymbol{\theta}_t^{\mathrm{F}}\|_2 &= r_t = \|\boldsymbol{\theta}_{t+1}^{\mathrm{A}} - \boldsymbol{\theta}_t^{\mathrm{A}}\|_2.
\end{aligned}
\tag{43}
$$

The norm is taken jointly over the common trainable parameter set, and the normalized update is defined for $\|\mathbf{g}_t^{\mathrm{F}}\|_2 > 0$.

Both interventions use the same checkpoint, client data split, adaptation and reference subsets, and $T = 5$ updates, with the remaining evaluation settings unchanged. For each intervention, we inde-

pendently refit the geometric centers, functional routing directions, and empirical rank maps from its own responses on the same reference subset. The standalone evaluation uses only the response support s_{R}, with $-s_{\mathrm{R}}$ as the OOD-oriented score. Panel (a) of Table 23 reports means over global random seeds 42–46 for both ID datasets and $\alpha \in \{0.1, 0.5, 10\}$. The ordinary forward entries are shared with the conventional response-only comparison.

Static-feature control. Using the frozen ResNet-18 checkpoints for the same five global random seeds 42–46, we replace $h_T^\ell(x) - h_0^\ell(x)$ with the original global-average-pooled feature $h_0^\ell(x)$ from stages 1–3, with no parameter updates. On the same client-specific training-pool reference samples, we independently fit $J = 16$ directional prototypes per layer. Both routing and scoring use geometric prototypes $\boldsymbol{\mu}_{\ell,j}$: the maximum-cosine prototype is selected and its cosine similarity is scored. We apply a local-reference ECDF to each layer's score and aggregate the three calibrated scores by an equal-weight geometric mean. The resulting standalone static score is evaluated on the complete official ID test set and the same Near/Far OOD sources and aggregation protocol; no test samples enter reference fitting or calibration. Unlike the $\boldsymbol{\mu} \to \boldsymbol{\mu}$ ablation in Table 14, this control uses original features rather than intervention responses.

Table 23: Additional controls for response-side detection (%; AUROC ↑ / FPR95 ↓). Panel (a) reports five-seed means; ordinary forward and semantic attenuation reproduce Tables 17–18. Panel (b) reports the standalone static-feature control over five random seeds 42–46. No scores are fused with global support. Bold in (a) marks the best mean within each setting.

(a) Stepwise update-norm control: response-only scores

ID dataset	α	Intervention	Near-OOD	Far-OOD
CIFAR-10	0.1	Forward training	52.77 / 94.36	84.34 / 55.75
		Norm-matched forward	51.98 / 95.02	85.78 / 53.06
		Semantic attenuation	**64.71 / 84.85**	**91.79 / 35.66**
CIFAR-10	0.5	Forward training	53.72 / 94.54	73.79 / 73.34
		Norm-matched forward	54.92 / 93.59	86.73 / 54.12
		Semantic attenuation	**65.05 / 80.16**	**94.82 / 24.86**
CIFAR-10	10	Forward training	52.95 / 94.29	72.05 / 74.06
		Norm-matched forward	55.05 / 94.44	86.96 / 51.55
		Semantic attenuation	**69.88 / 78.48**	**96.73 / 16.31**
CIFAR-100	0.1	Forward training	45.12 / 96.17	68.46 / 74.78
		Norm-matched forward	53.61 / 89.47	81.87 / 61.38
		Semantic attenuation	**58.15 / 85.97**	**91.04 / 35.04**
CIFAR-100	0.5	Forward training	46.18 / 95.21	65.55 / 74.39
		Norm-matched forward	54.35 / 89.50	79.78 / 63.57
		Semantic attenuation	**58.84 / 84.88**	**92.39 / 30.16**
CIFAR-100	10	Forward training	47.38 / 94.89	65.27 / 73.77
		Norm-matched forward	56.95 / 87.52	79.10 / 64.79
		Semantic attenuation	**59.21 / 84.01**	**90.64 / 38.85**

(b) Static-feature control: geometric routing and scoring

ID dataset	α	Near-OOD	Far-OOD
CIFAR-10	0.1	57.37 / 87.00	87.90 / 41.42
CIFAR-10	0.5	60.35 / 89.01	90.89 / 35.20
CIFAR-10	10	62.17 / 80.35	91.81 / 33.91
CIFAR-100	0.1	53.14 / 90.12	88.06 / 42.62
CIFAR-100	0.5	54.36 / 89.38	87.13 / 43.41
CIFAR-100	10	54.32 / 89.30	86.07 / 44.38

Mechanistic finding. In panel (a), semantic attenuation retains higher mean AUROC and lower mean FPR95 than norm-matched forward in both OOD regimes across all six settings, supporting the importance of the intervention beyond stepwise update magnitude alone. Panel (b) shows that the static-feature control has lower mean AUROC and higher mean FPR95 than semantic attenuation throughout, supporting the role of intervention-induced responses beyond static feature geometry.

Full-pipeline evaluation. We additionally insert each alternative response score into the same fusion pipeline while keeping global support unchanged and fitting all variant-specific quantities on the same client-local reference subset. For the static-feature control, the attenuation-induced response $\mathbf{d}_\ell(x)$ is replaced by the original pooled feature $h_0^\ell(x)$, and both routing and scoring use the geometric directions $\mu_{\ell,j}$; the remaining reference calibration, layer aggregation, global support, adaptive fusion, and evaluation procedure are kept identical to FedOPhD. Table 24 reports five-seed means of the resulting complete detection scores, separately from the standalone scores in Table 23; Figure 6 visualizes $\alpha = 0.1$ on CIFAR-10. FedOPhD attains higher AUROC and lower FPR95 than both fused alternatives in every reported dataset–heterogeneity setting and OOD regime. This comparison isolates the contribution of semantic-attenuation responses within the same downstream fusion pipeline.

Table 24: Full-pipeline update-norm and static-feature controls (%; five-seed means). Each entry is AUROC ↑ / FPR95 ↓. All variants use the same global support and fusion rule with variant-specific local-reference fitting. FedOPhD entries reproduce Table 1; bold marks the best mean within each setting.

ID dataset	α	Variant	Near-OOD	Far-OOD
CIFAR-10	0.1	Norm-matched forward + fusion	80.87 / 54.37	91.37 / 20.19
		Static features + fusion	80.95 / 57.28	92.60 / 15.07
		FedOPhD	**86.03 / 43.96**	**97.13 / 13.96**
CIFAR-10	0.5	Norm-matched forward + fusion	85.19 / 41.14	92.90 / 14.07
		Static features + fusion	85.68 / 41.88	95.29 / 7.75
		FedOPhD	**89.15 / 39.30**	**98.32 / 7.66**
CIFAR-10	10	Norm-matched forward + fusion	85.23 / 46.96	93.92 / 15.66
		Static features + fusion	86.07 / 47.21	94.27 / 23.65
		FedOPhD	**89.52 / 39.57**	**98.43 / 6.72**
CIFAR-100	0.1	Norm-matched forward + fusion	72.27 / 74.33	84.24 / 48.69
		Static features + fusion	71.20 / 74.88	87.08 / 43.34
		FedOPhD	**76.79 / 67.45**	**94.06 / 23.80**
CIFAR-100	0.5	Norm-matched forward + fusion	73.78 / 66.62	84.07 / 48.03
		Static features + fusion	74.08 / 66.30	88.70 / 39.72
		FedOPhD	**78.72 / 61.72**	**94.29 / 21.87**
CIFAR-100	10	Norm-matched forward + fusion	73.81 / 65.74	85.28 / 47.45
		Static features + fusion	74.61 / 64.77	87.98 / 42.16
		FedOPhD	**79.58 / 58.71**	**93.52 / 23.63**

E.3 Robustness, Scalability, and Efficiency

E.3.1 Robustness to cross-domain federated deployment

The label-skew experiments isolate heterogeneous semantic coverage, while realistic federations may additionally contain substantial appearance shifts across participating domains. We therefore evaluate DomainNet10, a ten-class subset of DomainNet (Peng et al., 2019), using the ID categories bird, feather, headphones, ice cream, teapot, tiger, whale, windmill, wine glass, and zebra from all six domains: Real, Clipart, Infograph, Painting, Quickdraw, and Sketch. The federation contains 12 clients, with two clients per domain; within each domain, the two clients hold complementary non-overlapping sets of five classes whose union is the complete ten-class ID space. The data split contains 16,478 FL-training images, a disjoint 1,800-image validation subset held out from the training pool, and the complete 4,573-image ID test set. The validation subset supports ID-only checkpoint selection. Each post-hoc detector is fitted from the receiving client's local training-pool ID data according to its own estimator; FedOPhD uses its local data for attenuation, reference fitting, and score calibration. The ID test set and all external OOD sources are reserved for evaluation and do not participate in training, detector fitting, or model selection.

We evaluate both FedAvg and FedBN (Li et al., 2021). Shared parameters are aggregated in proportion to client training-set size. Under FedBN, each client retains its own batch-normalization parameters and running statistics, and every query is evaluated with the receiving client's BN state, including queries from other domains and external OOD sources; no test-domain identity is used to switch BN states.

For receiver i, cross-domain transferred ID is

$$\mathcal{D}_{\mathrm{T},i}^{\mathrm{domain}} = \{x \in \mathcal{D}_{\mathrm{ID,test}} : \mathrm{domain}(x) \neq \mathrm{domain}(i)\}\,. \tag{44}$$

This population contains ID from the other five participating domains and excludes all test images from the receiver's domain, including those owned by the second client in that domain; it is not filtered by class or prediction correctness. The complete ID test set defines AUROC and FPR95, whereas Equation 44 defines T-AUROC and T-FPR95. For each receiver and each OOD source, the threshold is chosen to detect at least 95% of OOD, and FPR95 is the fraction of the corresponding negative ID population rejected as OOD. Metrics are first computed for every receiver–source pair, receivers are aggregated in proportion to their owned ID-test counts, and the eight OOD sources—CIFAR-100, TinyImageNet, CIFAR-10, SVHN, Textures, LSUN-Crop, LSUN-Resize, and iSUN—are then macro-averaged with equal weight. Before OOD evaluation, we exclude CIFAR-10 bird samples and CIFAR-100 tiger and whale samples. For TinyImageNet, we exclude six overlapping classes (teapot, ice cream, goose, black stork, king penguin, and albatross) and nine additional neighboring classes (tabby, Persian cat, Egyptian cat, cougar, lion, dugong, ice lolly, beaker, and water jug), retaining 185 classes. Dataset names in the domain-shift figures and tables refer to these filtered OOD evaluation sets.

Figure 3 visualizes the all-ID results, and Table 25 reports their exact values. FedOPhD achieves the highest mean AUROC and lowest mean FPR95 under both aggregation protocols. All evaluated detectors find the receiver-BN cross-domain protocol more challenging than FedAvg, yet FedOPhD retains a clear advantage on both metrics, showing that its intervention response remains informative when label-support heterogeneity is compounded by domain-dependent appearance and normalization statistics.

Table 25: DomainNet10 domain-shift results (%; five-seed mean ± sample standard deviation). All eight external OOD sources are macro-averaged. Bold marks the best mean within each aggregation protocol.

	FedAvg		FedBN	
Method	AUROC↑	FPR95↓	AUROC↑	FPR95↓
FedOPhD (ours)	**86.32 ± 1.49**	**39.64 ± 1.07**	**80.95 ± 1.24**	**54.60 ± 2.35**
FedEnergy	76.41 ± 4.17	61.74 ± 2.74	65.52 ± 1.39	74.12 ± 1.77
FedViM	71.95 ± 0.41	59.82 ± 0.53	69.54 ± 1.24	65.62 ± 0.86
FedNCI	80.21 ± 2.22	53.09 ± 1.89	71.26 ± 1.66	66.90 ± 2.08
FedGradPCA	64.91 ± 3.34	75.13 ± 2.91	59.55 ± 1.08	77.09 ± 1.63
FedRAS	78.41 ± 3.32	54.65 ± 3.28	67.28 ± 1.42	72.17 ± 0.44

The macro averages establish the overall ranking, but they can conceal source-specific strengths and failures. Tables 26 and 27 report the available per-source results under FedAvg and FedBN, respectively. Their matrices separate AUROC and FPR95; the final columns reproduce the equal-weight eight-source averages in Table 25. Because the complete-ID comparison can also conceal whether performance extends to transferred ID from unseen domains, panel (c) of each table reports the corresponding FedOPhD AUROC, FPR95, T-AUROC, and T-FPR95 for every source; the caption additionally records the shared checkpoint's ID accuracy.

Table 26: Per-source DomainNet10 results under FedAvg (%; five-seed means, with sample standard deviations where available; shared-checkpoint ID ACC $= 65.33 \pm 0.43$). Panels (a)–(b) compare methods; panel (c) reports FedOPhD on complete ID and cross-domain T-ID. C-100/C-10 denote CIFAR-100/CIFAR-10, Tiny-IN denotes TinyImageNet, and LSUN-C/LSUN-R denote LSUN-Crop/LSUN-Resize. Avg. is the equal-weight eight-source mean; – denotes an unavailable result. Bold in (a)–(b) marks the best reported mean in each column.

(a) AUROC ↑

Method	C-100	Tiny-IN	C-10	SVHN	Textures	LSUN-C	LSUN-R	iSUN	Avg.
FedEnergy	77.16 ± 5.31	76.53 ± 1.44	74.74 ± 5.67	76.76 ± 8.00	69.41 ± 2.51	83.73 ± 3.87	78.20 ± 3.99	74.74 ± 4.29	76.41 ± 4.17
FedViM	68.87 ± 0.93	68.61 ± 0.37	63.53 ± 0.91	81.19 ± 1.57	**88.71** ± 0.58	77.67 ± 1.12	62.15 ± 0.67	64.91 ± 0.36	71.95 ± 0.41
FedNCI	82.68 ± 2.97	74.82 ± 0.49	80.77 ± 3.57	89.28 ± 3.20	81.13 ± 0.85	89.48 ± 1.55	71.29 ± 3.13	72.25 ± 2.67	80.21 ± 2.22
FedGradPCA	68.00 ± 4.44	58.31 ± 0.80	64.25 ± 5.21	71.44 ± 4.64	62.03 ± 2.70	77.30 ± 3.45	59.97 ± 3.02	57.95 ± 3.43	64.91 ± 3.34
FedRAS	79.84 ± 3.89	77.39 ± 1.11	75.82 ± 4.54	81.46 ± 7.88	83.22 ± 0.87	81.04 ± 4.13	74.51 ± 3.55	73.96 ± 3.21	78.41 ± 3.32
FedOPhD	**88.83** ± 2.13	**81.04** ± 0.34	**87.59** ± 2.22	**92.90** ± 2.19	84.48 ± 0.54	**90.55** ± 1.21	**82.51** ± 2.10	**82.67** ± 2.06	**86.32** ± 1.49

(b) FPR95 ↓

Method	C-100	Tiny-IN	C-10	SVHN	Textures	LSUN-C	LSUN-R	iSUN	Avg.
FedEnergy	59.90 ± 3.66	62.66 ± 1.08	62.65 ± 2.33	59.83 ± 10.02	74.02 ± 2.19	43.71 ± 5.64	64.09 ± 1.68	67.08 ± 0.79	61.74 ± 2.74
FedViM	64.34 ± 0.90	61.71 ± 0.53	67.61 ± 1.01	58.24 ± 4.35	**43.95** ± 0.92	52.44 ± 2.18	65.91 ± 0.51	64.31 ± 0.10	59.82 ± 0.53
FedNCI	50.43 ± 4.05	66.54 ± 1.48	53.14 ± 3.87	30.28 ± 5.80	50.71 ± 0.74	32.73 ± 2.32	71.41 ± 1.69	69.47 ± 1.27	53.09 ± 1.89
FedGradPCA	71.67 ± 5.05	84.04 ± 1.01	72.92 ± 5.30	61.35 ± 4.21	82.62 ± 2.00	62.73 ± 4.04	82.03 ± 2.15	83.65 ± 2.43	75.13 ± 2.91
FedRAS	50.23 ± 3.94	61.57 ± 0.62	57.55 ± 3.49	40.09 ± 13.53	47.84 ± 3.01	44.22 ± 7.04	68.37 ± 1.77	67.29 ± 0.51	54.65 ± 3.28
FedOPhD	**36.67** ± 3.13	**50.50** ± 1.53	**37.02** ± 2.34	**23.24** ± 4.53	47.89 ± 0.62	**28.12** ± 2.22	**47.37** ± 0.84	**46.28** ± 0.40	**39.64** ± 1.07

(c) FedOPhD complete-ID and cross-domain T-ID breakdown

OOD source	AUROC↑	FPR95↓	T-AUROC↑	T-FPR95↓
CIFAR-100	88.83 ± 2.13	36.67 ± 3.13	87.47 ± 2.38	39.80 ± 3.08
TinyImageNet	81.04 ± 0.34	50.50 ± 1.53	79.55 ± 0.28	54.01 ± 1.86
CIFAR-10	87.59 ± 2.22	37.02 ± 2.34	86.09 ± 2.48	40.24 ± 2.18
SVHN	92.90 ± 2.19	23.24 ± 4.53	91.62 ± 2.60	26.35 ± 4.95
Textures	84.48 ± 0.54	47.89 ± 0.62	82.81 ± 0.64	51.65 ± 0.55
LSUN-Crop	90.55 ± 1.21	28.12 ± 2.22	89.46 ± 1.39	30.73 ± 2.35
LSUN-Resize	82.51 ± 2.10	47.37 ± 0.84	81.08 ± 2.19	50.84 ± 1.18
iSUN	82.67 ± 2.06	46.28 ± 0.40	81.09 ± 2.18	49.81 ± 0.50
Macro avg.	86.32 ± 1.49	39.64 ± 1.07	84.90	42.93

Under FedAvg, FedOPhD gives the strongest macro-average result and leads both metrics among the reported results on every source except Textures, where FedViM's feature-space construction is particularly effective. Several methods exhibit comparatively large seed variation on SVHN and LSUN-Crop.

Table 27: Per-source DomainNet10 results under FedBN (%; five-seed means, with sample standard deviations where available; shared-checkpoint ID ACC = 85.74 ± 0.69). Panels (a)–(b) compare methods; panel (c) reports FedOPhD on complete ID and cross-domain T-ID. C-100/C-10 denote CIFAR-100/CIFAR-10, Tiny-IN denotes TinyImageNet, and LSUN-C/LSUN-R denote LSUN-Crop/LSUN-Resize. Avg. is the equal-weight eight-source mean; – denotes an unavailable result. Bold in (a)–(b) marks the best reported mean in each column.

(a) AUROC ↑

Method	C-100	Tiny-IN	C-10	SVHN	Textures	LSUN-C	LSUN-R	iSUN	Avg.
FedEnergy	63.40 ± 2.79	64.66 ± 1.89	63.87 ± 3.03	70.25 ± 3.19	58.05 ± 0.80	73.04 ± 0.57	66.27 ± 1.20	64.59 ± 1.14	65.52 ± 1.39
FedViM	67.09 ± 0.69	64.47 ± 0.90	61.78 ± 0.66	81.06 ± 2.49	**87.05** ± 0.49	75.82 ± 2.50	57.36 ± 2.61	61.74 ± 2.22	69.54 ± 1.24
FedNCI	71.26 ± 2.98	65.81 ± 1.75	70.35 ± 3.40	81.88 ± 3.30	70.52 ± 1.06	79.84 ± 0.18	64.88 ± 1.48	65.52 ± 1.49	71.26 ± 1.66
FedGradPCA	61.09 ± 0.62	54.95 ± 0.81	61.05 ± 0.82	64.46 ± 3.71	52.55 ± 0.41	68.58 ± 1.97	57.58 ± 1.22	56.13 ± 1.27	59.55 ± 1.08
FedRAS	67.43 ± 1.47	64.01 ± 1.08	64.84 ± 2.25	72.42 ± 2.44	73.46 ± 0.30	69.72 ± 1.57	62.67 ± 2.09	63.69 ± 1.75	67.28 ± 1.42
FedOPhD	**83.17** ± 2.22	**73.99** ± 1.67	**82.59** ± 2.56	**91.02** ± 1.90	79.66 ± 0.11	**86.48** ± 0.67	**74.37** ± 1.81	**76.35** ± 1.67	**80.95** ± 1.24

(b) FPR95 ↓

Method	C-100	Tiny-IN	C-10	SVHN	Textures	LSUN-C	LSUN-R	iSUN	Avg.
FedEnergy	76.81 ± 3.34	77.76 ± 1.30	74.40 ± 3.85	63.29 ± 3.65	87.60 ± 0.94	65.47 ± 0.68	71.90 ± 1.41	75.72 ± 1.38	74.12 ± 1.77
FedViM	70.98 ± 1.15	70.28 ± 0.60	72.80 ± 0.72	57.03 ± 2.81	**50.71** ± 0.74	62.07 ± 2.97	71.28 ± 1.17	69.78 ± 1.24	65.62 ± 0.86
FedNCI	67.98 ± 3.73	76.07 ± 1.25	68.26 ± 4.16	46.19 ± 5.70	69.55 ± 1.68	55.17 ± 0.74	76.01 ± 2.14	75.95 ± 1.61	66.90 ± 2.08
FedGradPCA	74.10 ± 1.28	84.87 ± 0.20	73.59 ± 1.89	63.69 ± 3.30	89.78 ± 0.38	68.82 ± 3.29	79.80 ± 1.71	82.09 ± 1.46	77.09 ± 1.63
FedRAS	71.62 ± 1.07	79.13 ± 0.53	74.21 ± 2.06	59.13 ± 1.00	70.88 ± 0.72	70.06 ± 1.83	75.66 ± 0.33	76.66 ± 0.10	72.17 ± 0.44
FedOPhD	**56.47** ± 3.31	**67.27** ± 1.85	**54.62** ± 4.04	**28.47** ± 5.25	63.59 ± 1.04	**43.82** ± 1.30	**61.21** ± 1.59	**61.32** ± 1.73	**54.60** ± 2.35

(c) FedOPhD complete-ID and cross-domain T-ID breakdown

OOD source	AUROC↑	FPR95↓	T-AUROC↑	T-FPR95↓
CIFAR-100	83.17 ± 2.22	56.47 ± 3.31	81.33 ± 2.51	61.23 ± 3.65
TinyImageNet	73.99 ± 1.67	67.27 ± 1.85	71.97 ± 1.77	72.17 ± 1.90
CIFAR-10	82.59 ± 2.56	54.62 ± 4.04	80.89 ± 2.90	58.87 ± 4.50
SVHN	91.02 ± 1.90	28.47 ± 5.25	89.34 ± 2.24	32.26 ± 5.62
Textures	79.66 ± 0.11	63.59 ± 1.04	77.26 ± 0.15	69.21 ± 1.14
LSUN-Crop	86.48 ± 0.67	43.82 ± 1.30	84.74 ± 0.75	48.30 ± 1.39
LSUN-Resize	74.37 ± 1.81	61.21 ± 1.59	72.59 ± 1.87	65.41 ± 1.64
iSUN	76.35 ± 1.67	61.32 ± 1.73	74.54 ± 1.73	65.76 ± 1.85
Macro avg.	80.95 ± 1.24	54.60 ± 2.35	79.08	59.15

Under the receiver-BN protocol, FedOPhD remains strongest on both macro-average metrics and again leads the reported per-source comparisons except Textures. FedBN improves the shared checkpoint's ID classification accuracy, while every evaluated detector has lower macro-average AUROC and higher macro-average FPR95 than under FedAvg. This contrast shows that higher closed-set ID accuracy and stronger post-hoc OOD separation are distinct properties under cross-domain deployment. Within each aggregation protocol, cross-domain T-ID is consistently more difficult than complete ID: every source exhibits lower T-AUROC and higher T-FPR95 than its corresponding all-ID metric, while the aggregate trends remain closely aligned. The source-wise ordering is also coherent across protocols: SVHN and LSUN-Crop are comparatively well separated, whereas TinyImageNet, Textures, LSUN-Resize, and iSUN are more challenging. Together, the per-source and T-ID breakdowns support a federation-wide improvement that is not attributable to a single external dataset and remains visible when evaluation is restricted to ID from unseen domains.

E.3.2 Sensitivity to the number of clients and L-ID data budget

To assess FedOPhD across federation sizes, we keep the CIFAR-10 FL training pool fixed at 45,000 images and vary the number of clients from 10 to 150 under each heterogeneity level. This design makes federation scaling a direct L-ID data-budget stress test: as K grows, the mean number of local samples decreases exactly from $4{,}500$ to 300, while the Dirichlet concentration controls how unevenly those samples and semantic classes are distributed. Table 28 reports the corresponding per-dataset detection performance. This analysis reports AUROC and federation-wide FPR95 from the held-out evaluation populations; it therefore evaluates the fitted OOD score without using $\mathcal{D}_{\mathrm{cal},k}$ or a

client-local deployment threshold. The local calibration subsets are reserved for the fixed-threshold deployment analysis in Appendix E.1.1.

The nominal direction-bank size $J = 16$ serves as a ceiling in this scaling experiment. For a data-sparse client whose local reference samples and represented categories cannot support 16 nonempty response modes, detector construction uses the largest locally supported effective direction count before fitting the spherical clusters.

FedOPhD remains broadly stable through moderate federation sizes, especially for $\alpha \in \{0.5, 10\}$, even though the mean L-ID budget decreases substantially. The clearest degradation occurs in the largest, strongly heterogeneous federations. Near-OOD detection is more sensitive to federation scaling than far-OOD detection; among the far-OOD sources, Textures remains the most challenging, while the LSUN variants and SVHN are generally more stable. Together, these detection results characterize FedOPhD's sensitivity to federation size and the associated L-ID data budget.

Table 28: Per-dataset FedOPhD results under different numbers of FedAvg clients and Dirichlet heterogeneity levels on CIFAR-10 (%). All entries are means over five random seeds. The 20-client setting is the default; bold marks the best value across client counts at each heterogeneity level.

Clients K	α	CIFAR-100		TinyImageNet		LSUN-Crop		LSUN-Resize		Textures		SVHN		iSUN	
		AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
10	0.1	**86.81**	49.31	88.17	41.52	97.05	9.77	97.20	10.68	92.93	33.32	96.83	14.51	97.26	**11.10**
	0.5	**89.21**	**43.66**	**91.53**	**34.20**	98.33	**5.47**	98.43	5.91	**97.77**	**12.00**	98.54	7.67	98.42	6.28
	10	**88.70**	48.28	**91.11**	**31.90**	**98.19**	**6.15**	**99.06**	**3.72**	**98.15**	**8.52**	98.78	3.99	**98.89**	**4.42**
20 (default)	0.1	85.17	**46.81**	86.89	**41.11**	**97.58**	**9.64**	97.59	11.27	**94.97**	**27.72**	**98.13**	**9.11**	**97.36**	12.07
	0.5	87.97	43.82	90.33	34.79	**98.57**	**5.47**	**98.62**	**5.08**	97.01	17.33	**98.93**	**4.37**	**98.46**	**6.07**
	10	88.30	**42.48**	90.74	36.66	98.07	7.53	98.69	5.14	97.65	12.57	**99.14**	**2.95**	98.62	5.42
50	0.1	85.16	51.33	**88.93**	41.75	95.85	18.89	**97.70**	**10.14**	91.62	40.98	96.90	12.80	97.32	12.67
	0.5	87.21	46.47	90.40	35.65	98.32	7.54	98.46	6.34	96.97	16.23	97.62	11.40	98.16	7.87
	10	86.93	47.44	89.54	39.39	97.76	9.28	98.34	6.75	97.20	15.60	98.57	5.62	98.10	7.76
75	0.1	83.17	54.53	87.52	45.70	94.59	22.86	94.58	23.75	89.63	44.90	96.42	15.71	93.99	26.03
	0.5	84.57	52.14	87.96	40.98	98.08	8.60	97.17	12.99	94.50	24.90	97.67	8.70	96.83	15.12
	10	87.62	42.95	89.44	37.23	97.91	9.47	97.82	9.88	96.58	19.30	99.05	3.40	97.46	12.13
100	0.1	82.09	58.40	85.27	50.94	96.49	15.91	96.04	17.84	89.16	47.25	97.50	9.37	95.07	22.84
	0.5	84.34	54.47	87.57	43.32	97.78	10.44	97.09	13.88	94.94	23.97	98.10	6.71	96.93	14.98
	10	86.97	44.72	89.42	38.48	97.94	9.57	97.89	10.15	96.30	19.47	98.57	5.07	97.46	12.24
150	0.1	80.30	66.99	83.73	55.96	92.23	31.17	90.81	36.26	87.13	49.89	89.57	28.61	89.95	39.62
	0.5	82.51	58.95	85.57	49.39	96.98	14.41	94.95	22.13	91.82	34.53	97.89	7.00	94.42	24.49
	10	84.77	49.69	87.26	43.45	97.58	12.28	96.90	14.44	95.37	21.68	97.44	8.86	96.62	15.76

E.3.3 Sensitivity to the local data budget

To examine robustness to limited client-local data, we evaluate FedOPhD with 75%, 50%, and 25% of the local data budget on CIFAR-10 and CIFAR-100. This experiment complements the federation-size analysis by directly varying the available local budget.

Reduced-budget protocol. We keep the trained FL checkpoint fixed and vary only the local data available for detector construction. Starting from each client's fixed, disjoint adaptation and reference sets, which approximately split its local training data equally, we retain 25%, 50%, 75%, or 100% of each set separately, rounding sample counts down. Subsets are nested across budgets and constructed using class-prioritized sampling: samples are shuffled within each class, one sample per class is retained whenever feasible, and the remaining slots are allocated toward the original class proportions. The per-client calibration set, drawn from the held-out official-training validation pool, remains identical across budgets, as do all ID and OOD test samples. Reference directions, empirical score distributions, and reference-derived fusion parameters are rebuilt using the reduced reference set. Calibration scores and thresholds are recomputed for each fitted detector using the same calibration samples.

Results. Table 29 reports near- and far-OOD results. Overall ranking performance remains broadly stable across these budgets: from 75% to 25%, AUROC changes by at most 3.08 percentage points across the reported settings, with far-OOD AUROC remaining above 91%. The results support effective detection even with a reduced local budget.

FPR95 measures a tail operating point rather than ranking over the full score distribution. At 95% OOD recall, small changes in scores near the threshold can move the decision boundary across a region containing appreciable ID probability mass, producing a larger change in ID rejection than in

AUROC. This sensitivity is especially relevant to the far-OOD results, where high overall AUROC can coexist with more variable tail behavior. The same threshold sensitivity can also amplify seed-to-seed variation in FPR95; it does not require a comparable change in AUROC.

Table 29: FedOPhD under reduced local data budgets (%). Bold marks the best value among the three budgets for each setting and metric.

α	Budget	CIFAR-10				CIFAR-100			
		Near-OOD		Far-OOD		Near-OOD		Far-OOD	
		AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
0.1	75%	**85.04**	**46.79**	**97.13**	**13.65**	**75.81**	**69.56**	**93.53**	**26.88**
	50%	84.91	46.85	96.97	14.13	75.17	70.26	92.12	34.79
	25%	84.31	49.02	96.68	14.77	72.73	72.97	91.64	33.22
0.5	75%	**88.96**	**40.13**	**98.25**	**7.85**	**76.68**	**63.68**	**93.86**	**27.09**
	50%	88.24	40.98	98.18	8.49	75.77	64.53	92.90	33.76
	25%	88.65	40.88	97.95	10.38	75.48	65.53	92.62	33.50
10	75%	**89.51**	**39.30**	**98.26**	**7.99**	**78.01**	60.55	**92.95**	32.81
	50%	88.62	41.52	98.07	9.88	77.86	**60.33**	92.87	**32.17**
	25%	88.54	42.20	97.32	15.37	75.63	64.43	92.16	33.98

E.3.4 Sensitivity to the federated aggregation algorithm

To assess FedOPhD across federated optimization paths, we compare FedAvg (McMahan et al., 2017), FedProx (Li et al., 2020), and FedAvgM (Hsu et al., 2019) in the same 20-client CIFAR-10 setting across all reported heterogeneity levels and OOD datasets in Table 30. FedAvg serves as the default reference for directly inspecting changes associated with the aggregation algorithm. All three algorithms train a CIFAR-style ResNet-18 for 200 communication rounds with all 20 clients participating and five local epochs per round, at $\alpha \in \{0.1, 0.5, 10\}$. Local training uses SGD with momentum 0.9, weight decay 5×10^{-4}, batch size 128, and initial learning rate 0.1, with five warmup rounds followed by cosine decay. FedProx uses a proximal coefficient of 0.01; FedAvgM uses server momentum 0.9, server learning rate 1, and server-side Nesterov momentum.

Across the three aggregation algorithms, far-OOD detection remains consistently strong on most sources, whereas near-OOD performance varies more with both the aggregation rule and partition heterogeneity. FedProx and FedAvgM improve on FedAvg in some settings; FedAvg is stronger on both near-OOD datasets at $\alpha = 10$. At this heterogeneity level, FedProx and FedAvgM show higher near-OOD FPR95 while maintaining strong far-OOD separation, indicating that the learned representation affects the two regimes differently. These results show FedOPhD's behavior under all three evaluated aggregation algorithms and indicate that detection quality varies with the learned representation.

Table 30: Per-dataset FedOPhD results with different federated aggregation algorithms on CIFAR-10 with 20 clients (%). Each entry is AUROC ↑ / FPR95 ↓; FedAvg is the default.

Algorithm	α	CIFAR-100 (Near)	TinyImageNet (Near)	LSUN-Crop (Far)	LSUN-Resize (Far)	Textures (Far)	SVHN (Far)	iSUN (Far)
FedAvg	0.1	85.17 / 46.81	86.89 / 41.11	97.58 / 9.64	97.59 / 11.27	94.97 / 27.72	98.13 / 9.11	97.36 / 12.07
	0.5	87.97 / 43.82	90.33 / 34.79	98.57 / 5.47	98.62 / 5.08	97.01 / 17.33	98.93 / 4.37	98.46 / 6.07
	10	88.30 / 42.48	90.74 / 36.66	98.07 / 7.53	98.69 / 5.14	97.65 / 12.57	99.14 / 2.95	98.62 / 5.42
FedProx	0.1	87.57 / 44.41	90.54 / 36.15	96.96 / 10.88	97.98 / 9.56	94.57 / 28.30	97.27 / 13.59	97.90 / 10.23
	0.5	88.80 / 42.26	90.92 / 35.85	98.58 / 5.26	98.51 / 6.16	97.31 / 15.96	99.19 / 2.89	98.41 / 6.43
	10	87.15 / 47.19	89.67 / 43.80	98.13 / 6.95	98.56 / 4.95	97.36 / 13.54	99.02 / 3.38	98.60 / 4.84
FedAvgM	0.1	87.03 / 42.30	89.06 / 43.48	96.81 / 10.93	97.57 / 10.17	95.14 / 25.02	96.13 / 18.48	97.39 / 11.71
	0.5	88.81 / 40.12	91.24 / 32.37	98.62 / 5.53	98.66 / 5.53	97.52 / 13.98	98.76 / 4.73	98.65 / 5.75
	10	86.50 / 48.79	88.84 / 46.95	98.13 / 5.93	98.77 / 4.28	96.94 / 14.24	98.51 / 4.98	98.66 / 4.95

E.3.5 Sensitivity to the model backbone

To evaluate whether the response-based detector transfers across representation architectures, we apply FedOPhD to VGG11-BN (Simonyan & Zisserman, 2015), ResNet-50 (He et al., 2016), WRN-40-2 (Zagoruyko & Komodakis, 2016), and DenseNet-BC-100 (Huang et al., 2017) checkpoints

under 20-client FedAvg on CIFAR-10. For VGG11-BN, ResNet-50, and WRN-40-2, s_{R} aggregates responses from intermediate feature-extraction stages, excluding the final representation stage and the classification head. For DenseNet-BC-100, we spatially average the cumulative concatenated features at each bottleneck output before transition processing, excluding the final bottleneck output (`block3_unit16`) and the classification head. Table 31 includes the ResNet-18 main-result means as the common reference.

Training configuration. The official training set is split in a class-stratified manner into 45,000 training and 5,000 internal validation images; the official 10,000-image test set is used for testing. All models are trained from scratch without pretrained weights, with all 20 clients participating in every round and five local epochs per round. ResNet-18 uses 200 communication rounds; the other backbones use 150. For each $\alpha \in \{0.1, 0.5, 10\}$, random seeds are $\{42, 43, 44, 45, 46\}$. Training uses standard cross-entropy and SGD with batch size 128, momentum 0.9, weight decay 5×10^{-4}, and Nesterov disabled. The initial learning rate is 0.1, with five communication rounds of linear warmup followed by cosine decay to 0.001. Inputs are padded by four pixels, randomly cropped to 32×32, and horizontally flipped with probability 0.5, then normalized with mean $(0.4914, 0.4822, 0.4465)$ and standard deviation $(0.2470, 0.2435, 0.2616)$. Training uses BF16 mixed precision, without Mixup, CutMix, label smoothing, or dropout.

Architectures. Both ResNets use a 3×3 stride-one stem without initial max-pooling. VGG11-BN retains eight convolutional layers with batch normalization and five max-pooling operations, replacing the ImageNet classifier with global average pooling and a $512 \to 10$ linear head. WRN-40-2 uses a 16-channel stem and three stages of widths 32, 64, and 128, each containing six pre-activation residual blocks. DenseNet-BC-100 uses growth rate 12, compression factor 0.5, a 24-channel stem, and three dense blocks with 16 bottlenecks each; block-output widths are 216, 300, and 342, followed by global average pooling and a linear classification head.

ResNet-50 closely tracks ResNet-18 throughout the evaluated heterogeneity range, particularly on far-OOD detection. VGG11-BN improves steadily as the client partitions become more balanced, whereas WRN-40-2 attains its strongest detection results at $\alpha = 0.5$ and partially declines at $\alpha = 10$. Because each backbone produces a classifier checkpoint with different ACC, part of the corresponding AUROC and FPR95 variation may reflect differences in the learned classifier as well as its representation architecture. Across this range of checkpoint accuracies, FedOPhD consistently provides effective near- and far-OOD separation, supporting strong cross-backbone generalization; far-OOD separation remains stronger, while the exact sensitivity to heterogeneity depends on the learned representation.

Table 31: Backbone sensitivity on CIFAR-10 under 20-client FedAvg (%; five-seed means). The ResNet-18 entries reproduce the corresponding means from the main experiment.

Backbone	α	ACC↑	Near AUROC↑	Near FPR95↓	Far AUROC↑	Far FPR95↓
ResNet-18	0.1	88.94	86.03	43.96	97.13	13.96
	0.5	92.40	89.15	39.30	98.32	7.66
	10	93.12	89.52	39.57	98.43	6.72
VGG11-BN	0.1	85.08	74.16	66.31	86.29	47.65
	0.5	89.95	85.10	52.58	92.77	31.83
	10	91.15	87.42	47.30	95.25	24.47
ResNet-50	0.1	86.38	85.99	46.35	96.67	15.15
	0.5	92.87	88.01	40.49	98.00	8.67
	10	94.33	88.91	39.15	98.30	7.64
WRN-40-2	0.1	85.08	81.16	54.26	95.20	21.25
	0.5	91.52	86.42	44.74	97.47	11.48
	10	93.19	85.92	54.50	96.84	16.07
DenseNet-BC-100	0.1	85.79	83.62	54.52	96.41	16.29
	0.5	92.08	89.57	38.92	98.25	7.72
	10	93.39	89.22	41.29	98.57	6.89

E.3.6 EVALUATION WITH TINYIMAGENET AS THE ID DATASET

To assess FedOPhD on a more diverse and fine-grained ID domain, we provide per-dataset results with TinyImageNet as the ID dataset under 20-client FedAvg in Table 32. This evaluation retains the near-/far-OOD reporting structure used for the CIFAR experiments.

Training protocol. All 200 TinyImageNet classes are ID, with 64×64 inputs. We split the official training set in a class-stratified manner into 90,000 training and 10,000 internal validation images; the separate official labeled validation set of 10,000 images serves as the ID test set. We train ResNet-18 from scratch without pretrained weights, using a 3×3 stride-one stem and no initial max-pooling. Class-wise Dirichlet partitioning uses $\alpha \in \{0.1, 0.5, 10\}$; the five runs use global random seeds $\{42, 43, 44, 45, 46\}$. Training comprises 200 communication rounds with all 20 clients participating and five local epochs per round. Local SGD uses batch size 256, momentum 0.9, weight decay 10^{-4}, and initial learning rate 0.2, with five communication rounds of linear warmup followed by cosine decay to 0.002. Training augmentation uses padding by eight pixels, random cropping to 64×64, and random horizontal flipping. Channel-wise normalization uses mean $(0.4802, 0.4481, 0.3975)$ and standard deviation $(0.2770, 0.2691, 0.2821)$.

Importantly, the FedOPhD detector configuration is transferred unchanged from the CIFAR experiments: we retain $T = 5$, $\eta = 10^{-4}$, nominal $J = 16$ with the same support-aware reduction rule, response stages $\mathcal{I} = \{1, 2, 3\}$, equal-weight geometric layer aggregation, and fusion quantile $q = 0.1$. No TinyImageNet OOD data are used to select or retune these detector settings.

Near-OOD filtering. We filter the official CIFAR test sets by manually mapping their classes to the TinyImageNet ID classes. Identical semantic classes and explicit class-containment relationships lead to exclusion of the entire CIFAR class; visual similarity or a shared superclass alone does not. For CIFAR-10, we exclude automobile, bird, cat, dog, frog, and truck, retaining airplane, deer, horse, and ship (4,000 images). For CIFAR-100, we exclude aquarium_fish, bear, bee, beetle, bottle, bridge, bus, butterfly, camel, cattle, chair, chimpanzee, cockroach, elephant, keyboard, lawn_mower, lion, lobster, mountain, mushroom, orange, snail, snake, spider, sweet_pepper, table, telephone, tractor, and train, leaving 71 classes (7,100 images). The CIFAR-10 and CIFAR-100 entries in Table 32 refer to these filtered Near-OOD sets.

Near-OOD separation remains consistently strong across the reported heterogeneity levels. For far-OOD, the results remain strongest on LSUN-Crop and SVHN, while LSUN-Resize, Textures, and iSUN exhibit more dataset-dependent behavior; in particular, Textures becomes more challenging as the partition becomes more balanced. Taken together, these results extend FedOPhD beyond CIFAR-scale ID training to a 200-class, 64×64 setting, while also demonstrating effective cross-dataset transfer of a fixed detector configuration without TinyImageNet-specific OOD tuning. The per-source variation further shows that this behavior is not tied to a single OOD benchmark or heterogeneity regime.

Table 32: FedOPhD results with TinyImageNet as ID under 20-client FedAvg (%; AUROC ↑ / FPR95 ↓).

OOD dataset	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 10$
CIFAR-10 (Near)	96.63 / 8.28	97.33 / 6.66	96.76 / 9.01
CIFAR-100 (Near)	96.74 / 8.66	97.27 / 7.35	97.17 / 7.91
LSUN-Crop (Far)	98.62 / 4.53	99.04 / 3.38	99.35 / 2.42
LSUN-Resize (Far)	94.93 / 11.87	95.21 / 11.48	95.56 / 10.80
Textures (Far)	94.17 / 24.45	93.50 / 31.56	92.99 / 33.50
SVHN (Far)	98.18 / 6.04	99.78 / 0.83	99.55 / 1.52
iSUN (Far)	95.55 / 11.29	95.81 / 10.60	96.06 / 9.93

E.3.7 PER-OOD-DATASET FEDOPHD RESULTS

To assess how the aggregate near- and far-OOD findings extend across individual benchmarks, we provide FedOPhD results for every constituent OOD dataset under all reported heterogeneity levels in Tables 33 and 34. The breakdown shows a shared overall pattern across both near-OOD benchmarks and several visually distinct far-OOD sources. Far-OOD separation is generally stronger than near-OOD separation, while dataset difficulty and sensitivity to heterogeneity vary across the in-

dividual benchmarks. Together, these cross-dataset results support the aggregate trends and reveal benchmark-specific variation.

Table 33: Per-OOD-dataset FedOPhD results (%) with CIFAR-10 as the ID dataset. Each entry is AUROC ↑ / FPR95 ↓.

OOD dataset	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 10$
CIFAR-100 (Near)	85.17 / 46.81	87.97 / 43.82	88.30 / 42.48
TinyImageNet (Near)	86.89 / 41.11	90.33 / 34.79	90.74 / 36.66
LSUN-Crop (Far)	97.58 / 9.64	98.57 / 5.47	98.07 / 7.53
LSUN-Resize (Far)	97.59 / 11.27	98.62 / 5.08	98.69 / 5.14
Textures (Far)	94.97 / 27.72	97.01 / 17.33	97.65 / 12.57
SVHN (Far)	98.13 / 9.11	98.93 / 4.37	99.14 / 2.95
iSUN (Far)	97.36 / 12.07	98.46 / 6.07	98.62 / 5.42

Table 34: Per-OOD-dataset FedOPhD results (%) with CIFAR-100 as the ID dataset. Each entry is AUROC ↑ / FPR95 ↓.

OOD dataset	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 10$
CIFAR-10 (Near)	74.13 / 71.02	76.04 / 66.26	78.47 / 61.36
TinyImageNet (Near)	79.45 / 63.88	81.40 / 57.18	80.69 / 56.06
LSUN-Crop (Far)	93.56 / 23.18	93.48 / 22.65	93.11 / 23.10
LSUN-Resize (Far)	93.83 / 23.19	94.10 / 21.91	91.74 / 27.64
Textures (Far)	92.82 / 34.49	94.02 / 29.48	94.29 / 27.03
SVHN (Far)	96.25 / 13.60	95.64 / 13.12	96.53 / 10.86
iSUN (Far)	93.83 / 24.53	94.20 / 22.20	91.94 / 29.53

E.3.8 RUNTIME OVERHEAD

To quantify computational overhead relative to FedAvg training, we report setup and full inference times in Table 35 for the same $K = 20$ clients as in Table 1. Setup comprises additional training, reference computation, and fitting, as applicable to each method. For training-time methods, additional training denotes the increment over the corresponding FedAvg training time. The FedAvg training cost is excluded from setup for all methods; full inference time is reported separately as an absolute runtime. All timings are measured on our own hardware (NVIDIA GeForce RTX 5080 GPU and AMD Ryzen 9 9950X3D CPU). Client-local baseline costs and FedOPhD inference are scaled from per-client timings by K; federated training times use our own recorded measurements.

FedOPhD has a combined reported setup and inference cost close to FedRAS, with additional overhead relative to the other post-hoc baselines. Compared with the training-time federated OOD methods, its lower setup time yields a lower total over the two reported stages despite higher inference time. FedOPhD can be added to an existing federated checkpoint while preserving the original training algorithm.

Table 35: Setup and full inference times (s) for 20 clients. Additional training is incremental to FedAvg; the FedAvg training cost is excluded for all methods. Full inference time is absolute.

Dataset	Method	Additional Training + Reference & Fitting	Full Inference Time
CIFAR-10	FedEnergy	**0.00**	**571.80**
	FedViM	64.80	582.80
	FedNCI	22.00	593.20
	FedGradPCA	28.80	624.40
	FedRAS	36.80	1215.60
	FedOPhD	107.37	1167.40
	FOOGD	3267.13	754.80
	FOSTER	2764.63	771.20
	FedSDWC	2481.52	760.80
CIFAR-100	FedEnergy	**0.00**	691.60
	FedViM	68.40	**564.40**
	FedNCI	22.20	612.00
	FedGradPCA	122.20	740.40
	FedRAS	39.40	1251.20
	FedOPhD	108.83	1132.20
	FOOGD	3379.86	748.80
	FOSTER	2779.23	769.20
	FedSDWC	2448.46	769.00

E.4 Design and Hyperparameter Sensitivity

E.4.1 Sensitivity to the number of reference directions

Motivation and protocol. The number of client-local reference directions J controls the granularity with which the response readout represents distinct local response modes. Too few directions can merge different semantic response patterns, whereas an unnecessarily large direction bank can fragment limited local reference support and need not provide additional benefit. We therefore vary the per-stage value over $J \in \{1, 2, 4, 8, 16, 32, 64\}$ on CIFAR-10 while retaining all three selected response stages, $\mathcal{I} = \{1, 2, 3\}$. **For this sensitivity experiment, automatic reduction of the direction count is deliberately disabled: each tested configuration uses the specified J, rather than treating it as a ceiling.** This isolates the effect of the requested direction-bank size; the remaining settings follow the main protocol. Each result is averaged over the same five random seeds used in the main experiments. The default $J = 16$ is reproduced directly from Table 1; under $\alpha = 0.1$, local reference support is insufficient for the fixed $J \in \{32, 64\}$ configurations, so these settings are reported as not applicable rather than replaced by results at a smaller effective direction count.

Results. As α increases, each client generally observes more diverse local semantic content, and a richer direction bank can capture response modes that are merged at very small J. This benefit is clearest under the less heterogeneous partitions, where performance generally improves up to the default setting. The trend is not monotone without limit: increasing J beyond the default gives only metric-specific near-OOD changes and can degrade far-OOD separation, while the largest fixed-size banks are not applicable in the most heterogeneous setting under this no-reduction protocol.

Default choice. The nominal count $J = 16$ is motivated by a balance between local response diversity and reference-sample support. A small bank may merge distinct response patterns, while a large bank distributes the finite local reference samples among more directions and may yield less reliable estimates. A moderate bank size is therefore a reasonable design choice for representing multiple modes with limited local data. In the main protocol, the support-aware reduction described in Appendix D further limits the effective count when local samples are scarce.

Table 36: Fixed-J sensitivity on CIFAR-10 (%; five-seed means; automatic direction-count reduction disabled). Bold marks the best value for each setting and metric; – denotes an unsupported fixed J.

α	OOD	Metric	$J=1$	$J=2$	$J=4$	$J=8$	$J=16$ (default)	$J=32$	$J=64$
0.1	Near	AUROC↑	85.78	86.06	**86.08**	86.02	86.03	–	–
		FPR95↓	43.07	**42.89**	43.85	43.69	43.96	–	–
	Far	AUROC↑	96.72	96.70	96.98	96.86	**97.13**	–	–
		FPR95↓	14.01	14.07	14.04	14.16	**13.96**	–	–
0.5	Near	AUROC↑	85.09	85.73	86.73	87.85	89.15	**89.24**	88.97
		FPR95↓	46.88	44.77	43.15	41.89	39.30	**38.90**	39.77
	Far	AUROC↑	95.78	97.23	97.67	98.03	**98.32**	98.28	98.02
		FPR95↓	13.52	10.37	9.20	8.35	**7.66**	8.49	10.23
10	Near	AUROC↑	86.41	86.33	87.31	88.67	89.52	89.21	**89.75**
		FPR95↓	45.16	45.10	43.62	41.40	**39.57**	39.62	40.92
	Far	AUROC↑	96.69	98.13	98.20	98.26	**98.43**	98.01	97.12
		FPR95↓	11.91	8.15	8.02	7.71	**6.72**	11.10	15.45

E.4.2 Sensitivity to response-readout stages

FedOPhD measures the representation changes induced by semantic attenuation, and the selected readout stages determine which levels of this response are retained. Because different residual stages encode visual information at different levels of abstraction, this experiment examines whether the default stage set is necessary for capturing transferable ID response structure and whether the deepest residual stage provides additional OOD evidence. On CIFAR-10 with CIFAR-style ResNet-18, we compare response supports constructed from each individual residual stage, $\mathcal{I} \in \{\{1\},\{2\},\{3\},\{4\}\}$, as well as the aggregated configurations $\mathcal{I} = \{1,2,3\}$ and $\mathcal{I} = \{1,2,3,4\}$. We fix the number of reference directions to $J = 16$ and retain every other training, detector-construction, and evaluation setting from the main protocol. Each entry is averaged over the same five random seeds used in the main experiments. Changing $\mathcal{I}$ changes only the stages read by the response support; the complete network, including the fourth residual stage and classifier, remains active in the forward pass and global support. The Dirichlet concentration α also controls the local semantic coverage available to the client-conditioned detector: smaller α produces more label-skewed partitions and generally narrower per-client semantic support, whereas larger α produces more balanced partitions in which each client observes a broader set of classes. Comparing the readout choices across α therefore tests whether their usefulness depends on the diversity of locally available semantics.

Table 37: Response-stage sensitivity on CIFAR-10 (%; five-seed means; $J = 16$). Bold marks the best value within each setting and metric.

Response stages $\mathcal{I}$	$\alpha=0.1$				$\alpha=0.5$				$\alpha=10$			
	Near-OOD		Far-OOD		Near-OOD		Far-OOD		Near-OOD		Far-OOD	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
$\{1\}$	85.59	**43.48**	**97.16**	**12.63**	87.43	40.71	97.08	12.08	88.01	40.25	96.44	16.78
$\{2\}$	85.50	43.56	96.13	13.90	88.80	40.88	97.76	9.79	87.65	40.59	97.20	14.09
$\{3\}$	84.94	44.73	96.66	16.40	**90.38**	**38.40**	96.59	14.41	86.85	40.18	97.80	11.41
$\{4\}$	69.04	68.36	74.85	56.50	81.47	48.11	88.50	31.99	87.99	43.70	90.90	27.99
$\{1,2,3\}$ (default)	**86.03**	43.96	97.13	13.96	89.15	39.30	**98.32**	**7.66**	**89.52**	39.57	**98.43**	**6.72**
$\{1,2,3,4\}$	80.44	55.39	95.73	19.99	86.42	42.66	97.98	9.46	88.48	**39.38**	98.11	9.78

The first three residual stages provide substantially more stable response evidence than the fourth stage, particularly under strong heterogeneity. A plausible explanation is that the fourth stage is closest to the classifier and has the largest receptive field and lowest spatial resolution. Its representation is consequently more strongly organized around closed-set, class-discriminative semantics, while lower- and intermediate-level visual distinctions that can help separate OOD inputs are increasingly suppressed. Under locally conditioned semantic attenuation, an OOD input that is confidently mapped to a known class can therefore exhibit a high-level response resembling that of ID inputs,

reducing the separability of the stage-four response geometry. This effect is strongest at small α, where limited local class coverage makes the client-specific high-level response modes particularly narrow. As α increases and local semantic coverage broadens, the standalone fourth-stage results improve, especially for near-OOD. Nevertheless, its far-OOD separation remains substantially weaker. Adding stage four to the first three stages reduces near-OOD AUROC and worsens both far-OOD metrics at every evaluated heterogeneity level; near-OOD FPR95 also worsens except at $\alpha = 10$, where it improves slightly from 39.57 to 39.38. This comparison evaluates final-stage perturbation responses as the response-side evidence. No individual stage uniformly dominates every metric in the most heterogeneous setting, indicating that useful response evidence is distributed across representation depths. At $\alpha = 0.1$, stage 1 leads three of the four metrics: near-OOD FPR95 is 43.48 versus 43.96 for the default, and far-OOD AUROC/FPR95 is $97.16/12.63$ versus $97.13/13.96$; the default achieves higher near-OOD AUROC (86.03 versus 85.59). Stage 3 leads both near-OOD metrics at $\alpha = 0.5$. At $\alpha = 10$, stages 1–3 outperform every individual-stage readout on all four metrics.

Default choice. The choice $\mathcal{I} = \{1, 2, 3\}$ is motivated by the restricted semantic coverage of client-local data in federated learning. The final residual stage is closest to the classifier and is expected to encode more class-specific semantics, but a highly heterogeneous client's samples need not cover the global model's full semantic repertoire. Locally fitted response directions at that stage may therefore characterize only a narrow portion of its semantic structure. This consideration motivates combining earlier and intermediate stages to retain multi-scale response information under limited local semantic coverage. The final stage remains active in the model and in global recognition.

E.4.3 Sensitivity to response-layer aggregation

After selecting response stages $\mathcal{I} = \{1, 2, 3\}$, FedOPhD combines their locally rank-calibrated response supports by an equal-weight geometric mean. To test whether this soft agreement across stages is preferable to a hard extreme, we replace only the layer-aggregation operator with the maximum or minimum while retaining CIFAR-10, CIFAR-style ResNet-18, $J = 16$, and every other setting from the main protocol. The maximum allows one strongly aligned stage to dominate the response support, whereas the minimum lets the weakest stage determine it; the geometric mean lies between these extremes and rewards agreement without reducing the score to a single stage.

Table 38: Response-layer aggregation on CIFAR-10 (%; five-seed means) with $\mathcal{I} = \{1, 2, 3\}$ and $J = 16$. Bold marks the best value for each setting and metric.

α	Aggregation	Near-OOD		Far-OOD	
		AUROC↑	FPR95↓	AUROC↑	FPR95↓
0.1	Maximum	81.62	48.30	95.98	15.29
	Minimum	78.89	53.70	94.83	18.65
	Geometric mean (default)	**86.03**	**43.96**	**97.13**	**13.96**
0.5	Maximum	86.48	41.69	97.42	10.82
	Minimum	88.42	42.27	97.92	10.37
	Geometric mean (default)	**89.15**	**39.30**	**98.32**	**7.66**
10	Maximum	87.40	44.91	97.85	9.87
	Minimum	88.74	44.80	97.41	13.83
	Geometric mean (default)	**89.52**	**39.57**	**98.43**	**6.72**

The geometric mean attains the highest mean AUROC and lowest mean FPR95 for both OOD regimes at every evaluated heterogeneity level. Neither extreme operator is consistently preferable to the other: the maximum can preserve spuriously high support from one stage, while the minimum can be dominated by a weak or noisy stage response. These results support the geometric mean as a stable aggregation rule for complementary multi-stage response evidence.

E.4.4 Sensitivity to the attenuation learning rate

The learning rate η controls the magnitude and finite-step trajectory of semantic attenuation. We vary η on CIFAR-10 with CIFAR-style ResNet-18 while fixing $T = 5, \mathcal{I} = \{1, 2, 3\}$, $J = 16$, geometric

layer aggregation, and every other setting from the main protocol. Following the response definition in Section 4.2, the same η is used in the original-feature-RMS normalization for each configuration.

Remark E.1 (Learning-rate normalization and update count)**.** *Under the assumptions of Proposition A.2, define*

$$\Phi_{\eta,\ell}(\mathbf{d}) = \frac{\mathbf{d}}{\max\{\|\mathbf{d}\|_2, \eta\rho_\ell\}}, \qquad \widetilde{\mathbf{d}}_\ell = \mathbf{d}_\ell + \boldsymbol{\zeta}_\ell, \tag{45}$$

where $\boldsymbol{\zeta}_\ell$ represents an absolute response error. For fixed $\eta > 0$ and $\rho_\ell > 0$, Equation 18 gives

$$\left\|\Phi_{\eta,\ell}(\widetilde{\mathbf{d}}_\ell) - \Phi_{\eta,\ell}(\mathbf{d}_\ell)\right\|_2 \le \frac{\|\boldsymbol{\zeta}_\ell\|_2}{\eta\rho_\ell}. \tag{46}$$

Thus, cancellation of the leading learning-rate scale does not ensure uniform stability to absolute response errors as η decreases. Finite-precision updates and representation differencing may introduce errors that do not decrease proportionally with η, providing a possible contributor to small-step sensitivity.

Let $\mathbf{g}_0$ denote the initial attenuation gradient, as in the proof of Proposition A.2. For each fixed finite T, Equation 17 implies

$$\Phi_{\eta,\ell}(\mathbf{d}_\ell(x)) = \frac{-T J_{\mathbf{h}_\ell}(x;\boldsymbol{\theta}_0)\mathbf{g}_0}{\max\left\{T\left\|J_{\mathbf{h}_\ell}(x;\boldsymbol{\theta}_0)\mathbf{g}_0\right\|_2, \rho_\ell\right\}} + \mathcal{O}(\eta). \tag{47}$$

Unlike η, the finite update count T remains explicitly in the leading normalized response. Within the first-order regime, increasing T increases the response amplitude while $T\|J_{\mathbf{h}_\ell}(x;\boldsymbol{\theta}_0)\mathbf{g}_0\|_2 < \rho_\ell$; once this quantity reaches ρ_ℓ, the leading response enters the unit-norm regime. This characterizes saturation of the first-order amplitude effect, not necessarily of detection performance, which also depends on routing, rank transforms, fusion, and finite-step trajectory changes.

Table 39: Learning-rate sensitivity on CIFAR-10 (%; five-seed means). We fix $T = 5$; bold marks the best value for each setting and metric.

η	α	Near-OOD		Far-OOD	
		AUROC↑	FPR95↓	AUROC↑	FPR95↓
10^{-5}	0.1	80.99	52.89	92.23	28.79
	0.5	83.74	47.26	95.91	15.73
	10	84.00	50.92	93.08	25.87
5×10^{-5}	0.1	83.15	46.62	96.71	15.13
	0.5	85.91	41.34	98.12	8.51
	10	86.53	42.09	97.59	11.97
10^{-4} (default)	0.1	86.03	43.96	**97.13**	**13.96**
	0.5	**89.15**	**39.30**	98.32	7.66
	10	89.52	**39.57**	**98.43**	**6.72**
5×10^{-4}	0.1	**86.11**	**43.49**	97.02	14.15
	0.5	89.06	39.51	98.29	7.71
	10	**89.55**	40.31	98.29	8.03
10^{-3}	0.1	85.06	44.08	96.97	14.16
	0.5	88.94	39.64	**98.36**	**7.43**
	10	89.03	39.71	98.32	7.79

Empirically, very small learning rates yield poorer separation, while the default 10^{-4} provides strong performance across the three heterogeneity levels and both OOD groups. Further increases yield small, metric-specific fluctuations and can degrade the far-OOD operating point despite comparable AUROC.

Default choice. We use $\eta = 10^{-4}$ to balance the numerical resolution of the response with a modest perturbation magnitude. With a fixed update budget, excessively small steps can make the absolute representation displacement more susceptible to finite-precision effects, as discussed in Remark E.1. Conversely, large steps can move the diagnostic endpoint farther from the trained checkpoint and introduce stronger nonlinear trajectory effects. This scale is intended to support numerically stable response readout along a short attenuation trajectory.

E.4.5 SENSITIVITY TO THE NUMBER OF ATTENUATION UPDATES

The update count T controls the length of the semantic-attenuation trajectory: too few steps may under-excite the response, while additional steps increase detector-construction cost. We vary T on CIFAR-10 with $\eta = 10^{-4}$, $\mathcal{I} = \{1, 2, 3\}$, and $J = 16$; all other settings follow the main protocol. Remark E.1 characterizes the first-order role of T in the normalized response.

Table 40: Update-count sensitivity on CIFAR-10 (%; five-seed means). We fix $\eta = 10^{-4}$; bold marks the best value for each setting and metric.

T	α	Near-OOD		Far-OOD	
		AUROC↑	FPR95↓	AUROC↑	FPR95↓
1	0.1	83.10	50.34	95.36	20.07
	0.5	84.05	43.40	96.69	12.23
	10	85.06	44.48	95.99	18.59
3	0.1	85.03	45.57	95.99	14.29
	0.5	87.89	41.33	97.23	8.05
	10	86.57	42.09	97.74	11.22
5 (default)	0.1	86.03	43.96	**97.13**	**13.96**
	0.5	**89.15**	39.30	**98.32**	7.66
	10	89.52	39.57	**98.43**	**6.72**
10	0.1	**86.10**	43.81	97.11	14.13
	0.5	89.11	**39.28**	98.28	7.82
	10	**89.54**	39.53	98.28	8.06
20	0.1	86.09	**43.58**	97.03	14.17
	0.5	89.13	39.42	98.31	**7.63**
	10	89.53	**39.33**	98.30	7.94

One or three updates produce weaker separation. At $T = 5$, the results are broadly comparable to those with ten or twenty updates, which provide no consistent benefit and can weaken the far-OOD operating point.

Default choice. We use $T = 5$ to balance response stabilization over multiple updates with detector-construction time. Each additional update requires another gradient computation, whereas too short a trajectory may provide limited excitation under the finite-gain readout. A short five-update trajectory keeps the client-side construction budget modest while allowing the response to accumulate over multiple steps.

E.4.6 SENSITIVITY TO THE FUSION-SCALE QUANTILE

The quantile level q determines the client-specific fusion scale through $\beta = Q_q(\{s^{\mathrm{ref}}_{\mathrm{R},n}\})$, generalizing Equation 35. We vary $q \in \{0.05, 0.1, 0.2, 0.5\}$ on CIFAR-10 while retaining the complete FedOPhD pipeline and the remaining default settings. Table 41 includes the default $q = 0.1$ results from Table 1 for comparison.

The results show a trade-off between near- and far-OOD detection as the quantile varies. At $\alpha = 0.5$ and 10, $q = 0.05$ improves both near-OOD metrics over the default but worsens both far-OOD metrics. Increasing q to 0.2 or 0.5 reduces near-OOD performance relative to the default across all three heterogeneity levels, while $q = 0.2$ gives slightly better far-OOD metrics at $\alpha = 10$. Thus, the quantile affects the balance of the fused evidence, with the larger shifts particularly apparent in near-OOD detection.

Default choice. We use $q = 0.1$ as a fixed ID-side convention for setting the reference-derived regularization scale β. Since the arithmetic fusion weights are equal at $s_R = \beta$, the quantile determines how much of the local-reference ID distribution lies below this balance point. A substantially smaller quantile, such as 0.05, anchors β to a more extreme portion of the empirical reference distribution and is therefore more dependent on a small number of lower-tail samples, especially for data-sparse clients. Conversely, a substantially larger quantile, such as 0.2, extends stronger

Table 41: Fusion-quantile sensitivity on CIFAR-10 (%). Bold marks the best value within each heterogeneity setting and metric.

α	q	Near-OOD		Far-OOD	
		AUROC↑	FPR95↓	AUROC↑	FPR95↓
0.1	0.05	85.64	49.32	97.05	14.65
	0.1 (default)	**86.03**	**43.96**	**97.13**	**13.96**
	0.2	79.33	57.73	96.61	15.61
	0.5	72.87	65.88	95.45	21.17
0.5	0.05	**89.74**	**38.92**	98.17	9.17
	0.1 (default)	89.15	39.30	**98.32**	**7.66**
	0.2	85.48	43.67	97.96	8.86
	0.5	80.14	50.41	97.25	12.63
10	0.05	**90.73**	**37.90**	97.90	12.25
	0.1 (default)	89.52	39.57	98.43	6.72
	0.2	87.56	43.86	**98.45**	**6.65**
	0.5	84.61	45.16	98.23	8.36

response-side regularization over a broader portion of ordinary L-ID responses, which can unnecessarily limit the contribution of federation-wide recognition. The tenth percentile therefore provides a small but non-extreme lower-tail reference: it regularizes unusually weak response support without making that regularization dominant over a large fraction of normal local-ID behavior. This convention is determined entirely from client-local ID reference scores and is fixed independently of OOD samples or OOD performance.