GUARANTEED ENVIRONMENT-ACTION REGULARIZATION FOR FEATURE-LEVEL FEDERATED DOMAIN GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated domain generalization (FDG) trains on isolated source clients but must predict on an unseen domain without target data. Existing feature-statistic augmentation interpolates observed styles or heuristically explores feature statistics, but does not certify that the feature actually consumed by the loss lies outside the observed environment hull. We propose FEDGEAR (Guaranteed Environment-Action Regularization), a target-blind architecture that certifies this *realized* training action. Clients compute shared, class-balanced translation/log-scale sketches; the server learns an effective-rank environment basis, and a step-compensated endpoint makes the re-sketched acted coordinate cross a supporting hyperplane by a positive margin. Clients mine the hardest certified action and directly train the deployed semantic predictor, while a normalized witness is used only during training. We prove the exterior certificate and derive a structural target-risk decomposition that exposes action approximation, candidate coverage, bank drift, predictor consistency, and uncontrolled target mismatch. Across 294 controlled feature-level runs on OfficeHome-65, CIFAR-10C/100C, and a missing-combination benchmark, FEDGEAR has the highest mean accuracy in all seven protocols under a controlled common-interface comparison with thirteen baseline adaptations. Its 0.28-point gain over FedWon is small but consistent: all seven protocol means are positive and a hierarchical bootstrap gives $[0.08, 0.53]$. A matched control keeps every component but omits step compensation: endpoints remain 100% exterior, acted exteriority falls to 2–12%, yet mean accuracy changes by only 0.03 points. The certificate is therefore a verifiable exposure property, not an accuracy guarantee. A separate all-65-class, four-target, three-seed short-budget trainable-ResNet audit finds comparable performance to FedWon (28.08% versus 28.12%; paired 95% CI $[-0.20, 0.13]$), showing end-to-end competitiveness rather than superiority.

1 INTRODUCTION

Federated learning (FL) normally optimizes performance on participating clients, whereas deployment may introduce a client whose acquisition domain was never observed. Federated domain generalization (FDG) addresses this stricter setting: source examples remain local, target examples and statistics are unavailable during training, and a single global predictor must work without test-time adaptation (Nguyen et al., 2022; Bai et al., 2024). The central challenge is therefore not only non-IID optimization. The model must decide where to enforce invariance when the target environment is absent.

Recent FDG methods improve source cooperation without fully resolving missing environment coverage. Representation regularizers suppress client-specific information (Nguyen et al., 2022); normalization and weight reparameterization reduce sensitivity to feature statistics (Li et al., 2021b; Zhuang & Lyu, 2024); server objectives balance gradient agreement, layer drift, or client complementarity (Nguyen et al., 2025; Shi et al., 2025; Yan et al., 2024). These methods can make the observed clients agree, but agreement does not specify which unobserved environments should be traversed. This distinction matters under *compositional shift*, where a target combines nuisance factors in a configuration missing from all source clients.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

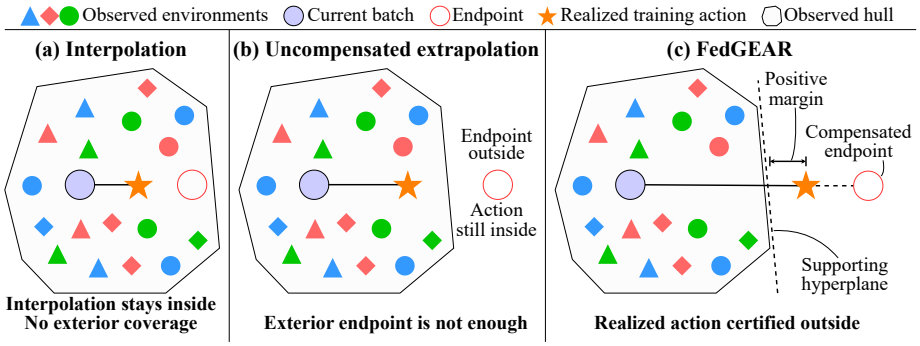


Figure 1: Why exteriority must be certified on the realized training action in coefficient space. (a) Interpolation never covers outside-hull coordinates. (b) A sampled endpoint can be exterior while its blended action remains inside. (c) FEDGEAR step-compensates the endpoint so the re-sketched acted coordinate c_{act} crosses $u^\top c = h_t(u)$ by the certified margin $m\sigma_t(u)$. Symbols encode observed environments; purple and orange denote c_B and c_{act}.

Feature-statistic augmentation offers a more direct route, but its coverage is difficult to audit. MixStyle and FedFA mix source moments (Zhou et al., 2021; Zhou & Konukoglu, 2023); DSU samples uncertain moments (Li et al., 2022); StableFDG explicitly explores new styles (Park et al., 2023); and CSU models channel correlation (Zhang et al., 2024). Interpolation remains inside a convex hull, while Gaussian or heuristic exploration may be outside only at its sampled endpoint. Once that endpoint is blended with a current batch, the feature actually presented to the classifier can fall back inside the hull. Our novelty is therefore not extrapolation per se, but certifying exteriority after the action is realized on the current batch: step compensation solves for an endpoint whose re-sketched feature, consumed by the loss, crosses a supporting hyperplane by a prescribed positive margin. This is an empirical coefficient-hull guarantee; it neither identifies the unknown target direction nor asserts that the target lies in the learned span.

Figure 1 isolates our key insight: environment coverage is a property of the realized training action, not of a sampled endpoint. We therefore construct the action in a low-rank coefficient space learned from shared client sketches. For the empirical hull $\mathcal{H}_t$, a direction u defines the support $h_t(u) = \max_{c\in\mathcal{H}_t} u^\top c = \max_k u^\top c_k$. We choose a compensated endpoint so that the *acted* coordinate—after the prescribed step from the current batch—satisfies $u^\top c_{\text{act}} = h_t(u) + m\sigma_t(u)$. The supporting hyperplane then certifies the re-sketched coordinate of the transformed feature as exterior by a positive, measured margin, without a target sample or assumed target coordinate.

Figure 2 turns this geometric insight into the complete FEDGEAR pipeline. First, clients use the same frozen pre-round model to upload class-balanced translation/log-scale sketches, so the server bank shares coordinates and is not driven by label proportions. Second, the server extracts a low-rank client-supported basis and constructs c_{act}; back-projection gives the environment coordinate $\tilde{s}$, whose feature realization $\tilde{h} = \mathcal{A}_{\tilde{s}}(h)$ is consumed by the loss. Third, each client privately selects its hardest certified action and applies source/acted cross-entropy and semantic prediction consistency directly to the head used at deployment. The normalized witness receives per-view CE and semantic–witness alignment only during training; target-blind inference keeps the semantic predictor unchanged and discards the witness.

Our evaluation combines a 294-run controlled feature-level mechanism study with causal and external-validity audits. Four OfficeHome-65 leave-one-domain-out splits, CIFAR-10C/100C corruption episodes, and a two-factor missing-combination protocol compare thirteen adaptations under one common interface. A compensation-only control isolates the certificate, a margin sweep measures semantic distortion, a protocol-level analysis guards against seed pseudoreplication, and a separate all-65-class short-budget OfficeHome audit trains ResNet-18 end to end. FEDGEAR obtains the highest controlled-study mean in every protocol; paired tests remain significant after Holm correction, although the strongest-comparator effect is small. The end-to-end audit establishes competitiveness with FedWon, not superiority.

Our contributions are:

- a *core realized-action mechanism*: a shared-coordinate bank and supporting-hyperplane construction guarantee that the re-sketched coefficient c_{act} of the transformed training feature, not merely a sampled endpoint, lies outside the observed environment hull;
- deployed-predictor-aligned acted training and a target-risk analysis that separate action coverage, bank drift, predictor consistency, and action misspecification; and
- *supporting controls*—class balancing for label-mixture confounding, an effective-rank bank, a training-only witness, and reliability weighting for aggregation stability—plus controlled, compositional, stress, and end-to-end audits. These controls are not claimed as independent accuracy-improving contributions.

2 Related Work

Federated optimization and multi-domain learning. FedAvg performs sample-weighted local optimization (McMahan et al., 2017); FedProx, SCAFFOLD, and MOON control parameter or representation drift (Li et al., 2020; Karimireddy et al., 2020; Li et al., 2021a). More recent work exploits client complementarity, gradient agreement, layer-wise shrinkage, or cross-institutional representation transfer in vertical healthcare FL (Yan et al., 2024; Nguyen et al., 2025; Shi et al., 2025; Yin et al., 2024). FedBN keeps normalization local, while FedWon removes normalization and standardizes weights for multi-domain FL (Li et al., 2021b; Zhuang & Lyu, 2024). These methods are strong optimization or architecture references. FedGEAR asks a different question: can a client be trained on an environment action that is provably missing from the participating set?

Federated domain generalization. FedSR regularizes local representations (Nguyen et al., 2022); prototype and generalization-adjustment methods reduce source-domain bias (Huang et al., 2023; Zhang et al., 2023). StableFDG explores shared styles (Park et al., 2023), while FedOMG matches normalized gradients without data (Nguyen et al., 2025). Recent work pursues data-centric flatness or token-routed prompt mixtures (Wang et al., 2026; Gong et al., 2026), complementing language- and prompt-based FDG (Yan et al., 2025; Wu et al., 2025). These methods improve optimization or representations; FedGEAR instead certifies empirical exteriority of the realized acted coordinate without target-conditioned selection, following the target-blind ICLR benchmark (Bai et al., 2024).

Feature statistics and domain coverage. Mean and standard deviation encode useful style information. MixStyle interpolates statistics (Zhou et al., 2021); DSU treats them as uncertain variables (Li et al., 2022); FedFA adapts this principle to federated training (Zhou & Konukoglu, 2023); StableFDG explores beyond-source styles (Park et al., 2023); and CSU preserves cross-channel correlation (Zhang et al., 2024). These methods act on sampled or mixed statistics. FedGEAR instead certifies the re-sketched coordinate of the actual transformed feature after blending: an exterior endpoint can produce an interior realized action, whereas Equation 4 compensates for the current batch and action step. The guarantee is limited to empirical coefficient-hull exteriority; moment coordinates need not recover an environment distribution, and arbitrary targets need not lie in their span.

3 FedGEAR

3.1 Setting and round overview

Let source client $k \in [K]$ hold $\mathcal{D}_k \sim P_k$ and let $P_\star$ be an unavailable target distribution. The final semantic predictor is $f_w(x) = g_\phi(h_\theta(x))$, where $h_\theta \in \mathbb{R}^d$ is a trainable bottleneck. The objective of FDG is $R_\star(w) = \mathbb{E}_{P_\star}\ell(f_w(x), y)$, although optimization can access only source data. Figure 2 separates the method into three training modules: (A) shared, class-balanced sketches; (B) a guaranteed realized exterior action; and (C) hard-action selection with direct semantic training and training-only support. Its right panel makes the deployment contract explicit: the target is never observed during training, and inference uses only the unchanged semantic predictor.

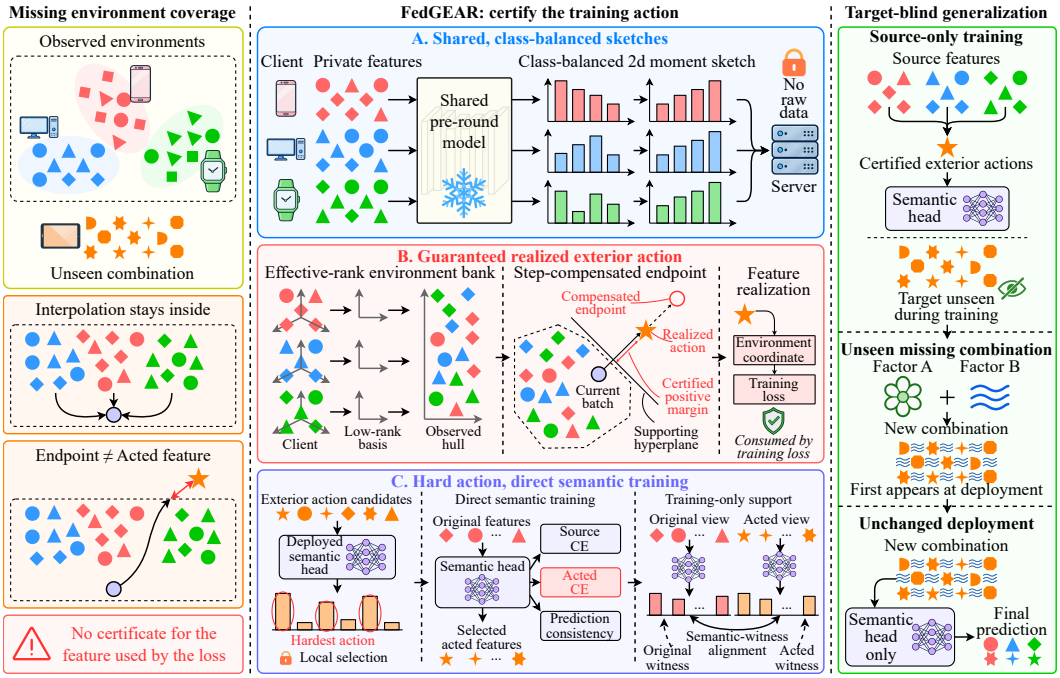


Figure 2: Complete FEDGEAR training-to-deployment pipeline. Left: interpolation and uncompensated endpoints do not certify the coordinate realized by the feature consumed by the loss. Center: (A) a frozen shared pre-round model produces class-balanced sketches without raw-data exchange; (B) an effective-rank bank constructs c_{act}, which maps to $\tilde{s}$ and then to the acted feature $\tilde{h}$; and (C) private hard-action selection directly trains the deployed semantic head. Semantic prediction consistency remains in the direct branch, whereas the normalized original/acted witnesses use per-view semantic–witness alignment only as training support. Right: inference discards the witness and keeps the semantic head unchanged for an unseen factor combination.

3.2 SHARED, CLASS-BALANCED ENVIRONMENT SKETCHES

Environment coordinates are meaningful only when clients use a common representation map. At the start of round t, the server broadcasts w_t. Before local optimization, every selected client evaluates the same deterministic subset with h_{θ_t}, computes a sketch in these shared coordinates, and uploads it for same-round bank construction. Local feature drift therefore cannot rotate one client's coordinate system relative to another.

Client label proportions are a second confounder. Let $\mathcal{Y}_k$ be the classes present on client k, $\mu_{k,y} = \mathbb{E}[h \mid y, k]$, and $q_k(y)$ its empirical label proportion. The ordinary mean $\sum_y q_k(y)\mu_{k,y}$ changes when q_k changes even if all class-conditionals are identical. We instead define

$$a_k = \frac{1}{|\mathcal{Y}_k|} \sum_{y \in \mathcal{Y}_k} \mathbb{E}[h \mid y, k], \tag{1}$$

$$v_k = \frac{1}{|\mathcal{Y}_k|} \sum_{y \in \mathcal{Y}_k} \mathbb{E}[h^{\odot 2} \mid y, k] - a_k^{\odot 2}, \qquad s_k = [a_k, \tfrac{1}{2}\log(v_k + \epsilon)] \in \mathbb{R}^{2d}. \tag{2}$$

This sketch is compact, but it is not asserted to reconstruct P_k. It parameterizes only the diagonal translation/log-scale action used below.

Proposition 1 (Removal of label-mixture confounding)**.** *If two clients have identical class-conditional bottleneck moments and the same set of observed classes, their population sketches in Equation 2 are equal for arbitrary label proportions. Their unconditional sketches need not be equal.*

The proof is direct and appears in Appendix A.3. Missing classes and finite samples introduce estimation error treated in Appendix A and probed by the Dirichlet and participation diagnostics; the unbalanced-sketch replacement separately tests label-proportion confounding.

3.3 EFFECTIVE-RANK BANK AND GUARANTEED EXTERIOR ACTION

The server stacks selected sketches $S_t = [s_k]_{k\in\mathcal{S}_t}$, centers them at $\bar{s}_t$, and computes a thin SVD. Singular values below 10^{-4} of the largest are discarded, and at most $r_{\max}$ right singular vectors form $B_t \in \mathbb{R}^{r_t\times 2d}$. The observed coefficients are

$$c_k = B_t(s_k - \bar{s}_t), \qquad \mathcal{H}_t = \mathrm{conv}\{c_k : k \in \mathcal{S}_t\}. \tag{3}$$

The effective rank r_t is data dependent and at most $|\mathcal{S}_t| - 1$. If none survives ($r_t = 0$), no bank or exterior action is formed and the minibatch uses the source-only branch. Unlike full-dimensional noise, the bank restricts actions to client-supported variation.

To generate a candidate, sample a unit direction $u \in \mathbb{R}^{r_t}$ and compute the support $h_t(u) = \max_{c\in\mathcal{H}_t} u^\top c = \max_k u^\top c_k$ and projected spread $\sigma_t(u) = \mathrm{Std}_k(u^\top c_k)$. On the current labeled minibatch, Equation 2 over its present classes gives s_B; we explicitly define $c_B = B_t(s_B - \bar{s}_t)$. Let $\tau \in (0, 1]$ be the action step and $m > 0$ a margin. We choose the endpoint coefficient

$$c_{\mathrm{end}}(u) = \frac{h_t(u) + m\sigma_t(u) - (1-\tau)u^\top c_B}{\tau}\, u \tag{4}$$

and use $c_{\mathrm{act}} = (1-\tau)c_B + \tau c_{\mathrm{end}}$. Components orthogonal to u may be set to zero because the certificate needs only one separating direction. The corresponding environment coordinate is $\tilde{s} = \bar{s}_t + B_t^\top c_{\mathrm{act}}$. Split $\tilde{s} = [\tilde{a}, \tilde{b}]$, set $\tilde{v} = \exp(2\tilde{b}) - \epsilon$, and apply the same class-balanced convention as Equation 2:

$$\mathcal{A}_{\tilde{s}}(h) = \sqrt{\tilde{v}} \odot \frac{h - a_B}{\sqrt{v_B}} + \tilde{a}. \tag{5}$$

Nonpositive target-variance proposals are discarded; if none is feasible, the minibatch uses only its source loss. Retained coordinates have $v_B > 0$.

Lemma 1 (Realization of the prescribed sketch)**.** *Recomputing Equation 2 on $\mathcal{A}_{\tilde{s}}(h)$ over the same labeled batch yields exactly $\tilde{s}$ on every retained coordinate.*

Proposition 2 (Exterior-action certificate)**.** *For every feasible retained candidate with $m > 0$, $\sigma_t(u) > 0$, $\tilde{v} > 0$, and retained $v_B > 0$, the acted coefficient produced by Equation 4 satisfies*

$$u^\top c_{\mathrm{act}} = h_t(u) + m\sigma_t(u) > h_t(u), \tag{6}$$

and hence $c_{\mathrm{act}} \notin \mathcal{H}_t$. The measured distance to the supporting hyperplane is $m\sigma_t(u)$.

Together with Lemma 1, this statement concerns the re-sketched feature actually consumed by the loss, not merely an action parameter. The client proposes M directions and forms the feasible retained set $\mathcal{C}_t^+$; neither labels nor examples leave the client.

3.4 HARD ACTION SELECTION AND DIRECT SEMANTIC TRAINING

Hard selection and direct supervision make the deployed predictor, rather than an auxiliary branch, learn from the certified action. From $\mathcal{C}_t^+$, the client privately selects the candidate maximizing semantic-head cross-entropy on its current batch. We define semantic logits $z_w(h) = g_\phi(h) = W_s\,\mathrm{GELU}(h)$ and probabilities $p_w(h) = \mathrm{softmax}(z_w(h))$. The witness has logits $z_\psi(h) = W_q\,\mathrm{GELU}(\mathrm{LN}(h))$ and probabilities $q_\psi(h) = \mathrm{softmax}(z_\psi(h))$; it supports training but is never mixed into test-time logits. Thus the theoretical, optimized, and deployed predictor is the same semantic head; we make no exact quotient-invariance claim for the witness.

For an original/acted pair $(h, \tilde{h})$, client k minimizes

$$\begin{aligned}\mathcal{L}_k(w,\psi) = {} & \mathrm{CE}(z_w(h), y) + \lambda_a\,\mathrm{CE}(z_w(\tilde{h}), y) + \lambda_w\Big[\mathrm{CE}(z_\psi(h), y) + \beta\,\mathrm{CE}(z_\psi(\tilde{h}), y)\Big] \\ & + \lambda_c\,\mathrm{SKL}(p_w(h), p_w(\tilde{h})) + \lambda_{cw}\Big[\mathrm{SKL}(p_w(h), q_\psi(h)) + \mathrm{SKL}(p_w(\tilde{h}), q_\psi(\tilde{h}))\Big],\end{aligned} \tag{7}$$

Here $\mathrm{SKL}(p, q) := \frac{1}{2}[\mathrm{KL}(p\|q) + \mathrm{KL}(q\|p)]$, as implemented. The first line directly supervises the final predictor on both environments. The second separates semantic consistency, $p_w(h) \leftrightarrow p_w(\tilde{h})$, from per-view semantic–witness alignment. During the four-round warm-up, the bank follows the

Algorithm 1 One round of FEDGEAR

Sample $\mathcal{S}_t$; broadcast shared (w_t, ψ_t).
Clients compute and upload pre-round class-balanced sketches by Eq. 2.
Server builds and broadcasts $(\bar{s}_t, B_t, \{c_k\})$; warm-up rounds log but do not act on it.
for $k \in \mathcal{S}_t$ in parallel **do**
  if $t < T_{\text{warm}} = 4$ **then**
    On every minibatch optimize $\text{CE}(z_w(h), y) + \lambda_w \text{CE}(z_\psi(h), y)$; set $e_k = 0$.
  else
    for each private minibatch **do**
      Propose M directions by Eq. 4; retain feasible $\mathcal{C}_t^+$.
      if $\mathcal{C}_t^+ \neq \varnothing$ **then**
        Select maximum semantic CE; apply Eq. 5; optimize Eq. 7; record g_B.
      else
        Optimize the same source-only warm-up objective.
      end if
    end for
    Set e_k to mean acted g_B, or 0 if all proposals are infeasible.
  end if
  Upload $(w_{t,k}, \psi_{t,k})$ and semantic prediction gap.
end for
Aggregate both states by Eq. 8; discard ψ only at deployment.

same communication path but is not used for actions, preventing an untrained bottleneck from driving acted training.

For each acted minibatch B, let $g_B = (|B|C)^{-1} \sum_{i,j} |p_w(h_i)_j - p_w(\tilde{h}_i)_j|$, matching the implemented mean absolute probability gap. For client k's acted-batch set $\mathcal{A}_k$, set $e_k = |\mathcal{A}_k|^{-1} \sum_{B \in \mathcal{A}_k} g_B$ when nonempty and $e_k = 0$ otherwise. The latter gives a neutral reliability factor; if all clients have no acted batch, aggregation becomes sample-count weighting. The server uses

$$\alpha_k = \frac{n_k \exp(-\gamma e_k)}{\sum_{j \in \mathcal{S}_t} n_j \exp(-\gamma e_j)}, \qquad (w_{t+1}, \psi_{t+1}) = \sum_{k \in \mathcal{S}_t} \alpha_k (w_{t,k}, \psi_{t,k}). \tag{8}$$

The witness is initialized once on the server, broadcast with w_t, optimized during warm-up using its source-view term and thereafter using both views, and aggregated with the same α_k; there is no client-local persistent witness or re-entry rule. Reliability weighting is intentionally secondary: the ablation removes it while keeping the action and network unchanged. The scalar gap is uploaded with both parameter states; the separate sketch/bank exchange is counted below. The test-time model discards ψ and contains only the encoder and semantic head.

3.5 THEORY ALIGNED WITH THE DEPLOYED PREDICTOR

Appendix A proves the result. This bound is a structural decomposition, not a target-observable guarantee: its purpose is to expose which failure terms remain uncontrolled in target-blind FDG. Let $a_\star$ be the best full-bank target approximation, $\hat{a}$ its closest feasible retained candidate, and $R_{\max} = \max_{c \in \mathcal{C}_t^+} R_{\mathcal{A}_c \# \bar{P}}$. Here δ_{span} measures target mismatch, $\|a_\star - \hat{a}\| \leq \Delta_t + \eta_M$ separates bank and candidate error, and $R_{\max}$ links coverage to hard mining; Appendix Table 2 identifies the corresponding diagnostics and which terms remain proxies.

Theorem 1 (Target-risk decomposition; informal)**.** *Assume bounded loss, action-risk smoothness, target residual ε_{sem}, and uniform predictor consistency $\varepsilon_{\text{pred}}$ over nonempty $\mathcal{C}_t^+$. Then the deployed predictor satisfies*

$$\begin{aligned} R_\star(f_w) &\leq R_{\max}(f_w) + L_s(\delta_{\text{span}} + \Delta_t + \eta_M) + \varepsilon_{\text{sem}} \\ &\leq R_{\text{src}}(f_w) + L_p\sqrt{2\varepsilon_{\text{pred}}} + L_s(\delta_{\text{span}} + \Delta_t + \eta_M) + \varepsilon_{\text{sem}}. \end{aligned} \tag{9}$$

The first, hard-mining-aligned bound has no consistency penalty; the second uses uniform candidate consistency to replace maximum acted risk by source risk. Appendix B gives a separately scoped smooth-surrogate analysis of dynamic-bank drift and reliability bias.

4 EXPERIMENTS

The evaluation asks five questions: (Q1) does certified exterior training improve target-blind accuracy against recent alternatives? (Q2) does it handle missing factor combinations? (Q3) which component causes which behavior? (Q4) what are the robustness and cost boundaries? (Q5) does the mechanism remain viable when the image backbone is trained end to end?

4.1 PROTOCOL AND BASELINES

Benchmarks. **OfficeHome-65** contains all 65 classes across Art, Clipart, Product, and Real World (Venkateswara et al., 2017). We run all four leave-one-domain-out (LODO) tasks, split each source domain into three Dirichlet-0.45 clients, and never use the held-out domain for selection. **CIFAR-10C/100C** use clean source clients and six unseen corruption episodes (Krizhevsky, 2009; Hendrycks & Dietterich, 2019). **CompShift-2F** observes three of four label-preserving factor combinations and tests an inside-hull point, the missing in-span corner, and that corner plus an orthogonal nonlinear factor. An auxiliary **OfficeHome-10 Composition** certificate-control protocol uses Art/Product and clean/blur factors, withholds Product–blur, and adds grayscale as an unsupported factor; it is not an end-to-end external-validity claim.

Model scope. The 294-run controlled grid caches frozen ImageNet ResNet-18 features (He et al., 2016) and trains the complete predictor above them; OfficeHome uses all 15,588 images. Separately, a fixed short-budget raw-pixel audit trains ResNet-18 end to end on all 65 OfficeHome classes over four LODO targets, three seeds, and FedAvg, FedSR, DSU, FedFA, StableFDG, CSU, FedWon, and FEDGEAR. Both regimes exclude target data from training, tuning, banks, and checkpoint selection and report the predeclared final round.

Common-interface baseline adaptations. The comparison includes FedAvg, FedProx, FedBN, and FedSR (McMahan et al., 2017; Li et al., 2020; 2021b; Nguyen et al., 2022); feature-statistic methods MixStyle, DSU, FedFA, StableFDG, and CSU (Zhou et al., 2021; Li et al., 2022; Zhou & Konukoglu, 2023; Park et al., 2023; Zhang et al., 2024); and recent multi-domain/federated methods FedWon, FedOMG, FedLWS, and FedSaC (Zhuang & Lyu, 2024; Nguyen et al., 2025; Shi et al., 2025; Yan et al., 2024). Each is adapted at the same bottleneck: MixStyle interpolates client moments; DSU uses diagonal moment uncertainty; FedFA combines global/local uncertainty; StableFDG shifts beyond a peer style; CSU samples correlated low-rank uncertainty; FedWon uses weight-standardized linear layers; FedOMG weights normalized update agreement; FedLWS shrinks layers toward the pre-round state; and FedSaC combines update similarity with sketch complementarity. FedBN uses source-averaged BN states at an unseen target, never target statistics. Appendix D documents these adaptations and their limitations. The 2026 flatness and prompt-routing methods require defining end-to-end optimization or prompt architectures that this frozen-feature interface cannot preserve, so we treat them as complementary work rather than controlled implementations (Wang et al., 2026; Gong et al., 2026).

Training and metrics. All methods use the same client draws, data caps, initialization seed, SGD minibatch order, and fixed final-round reporting. Dataset-family schedules and one method configuration are frozen from source-only development before target evaluation: each target sees one configuration and a predeclared final checkpoint (Appendix Table 3). FEDGEAR uses effective rank at most four, $M = 3$, margin $m = 0.18$, action step $\tau = 0.32$, and a four-round warm-up. The counterfactual and consistency weights are 0.06 and 0.02. FedWon uses weight-standardized layers in every reported family; FEDGEAR uses the identical implementation except on CIFAR-100C, where source-only development fixed standard Linear/BN in the trainable bottleneck while leaving z_w unchanged. We report accuracy, macro F1, ECE, NLL, Brier score, worst-domain accuracy, source–target gap, runtime, communication, and parameters. Error bars are 95% intervals unless stated otherwise.

4.2 MAIN RESULTS

Table 1 answers Q1. FEDGEAR ranks first in mean accuracy on all four benchmark families and in overall macro F1. Against the strongest common-interface comparator, FedWon, 21 paired units

Table 1: Target-blind performance of FEDGEAR and thirteen common-interface baseline adaptations over three seeds. OfficeHome-65 averages four held-out domains; Comp. denotes CompShift-2F. Accuracy, macro F1, and ECE are percentages. Best and second-best values are bold and underlined.

Method	OH-65 Acc. ↑	C10C Acc. ↑	C100C Acc. ↑	Comp. Acc. ↑	Avg. F1 ↑	Avg. ECE ↓
FedAvg	54.62	50.07	26.54	80.24	51.61	6.71
FedProx	54.62	50.08	26.52	80.27	51.61	6.72
FedBN	54.62	50.07	26.54	80.24	51.61	6.71
FedSR	54.65	50.13	26.55	80.27	51.65	6.74
MixStyle	54.73	49.40	26.57	80.06	51.48	7.28
DSU	54.76	48.68	26.52	79.36	51.34	<u>6.62</u>
FedFA	54.76	48.60	26.49	79.35	51.33	6.63
StableFDG	54.60	48.04	26.24	79.26	51.13	**6.55**
CSU	54.73	49.44	<u>26.67</u>	80.31	51.39	9.14
FedWon	<u>56.07</u>	<u>52.59</u>	25.92	80.77	<u>52.65</u>	7.62
FedOMG	54.52	50.17	26.48	80.15	51.55	6.67
FedLWS	54.50	50.36	26.34	80.33	51.53	6.79
FedSaC	54.73	51.23	26.49	<u>80.86</u>	51.91	6.71
FedGEAR	**56.21**	**52.60**	**26.75**	**81.32**	**53.15**	6.72

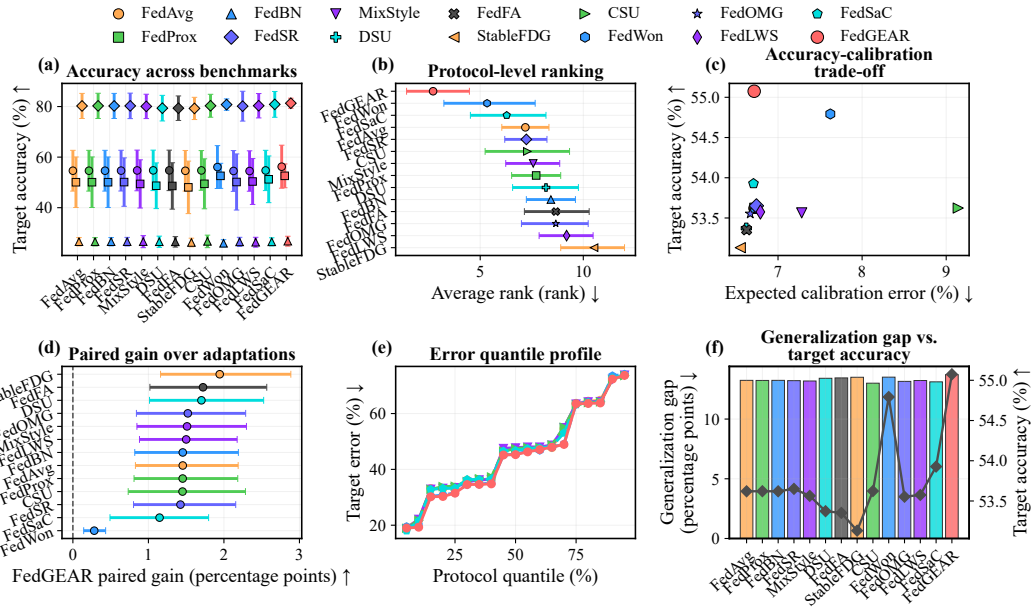


Figure 3: Main evaluation. (a) Accuracy across four benchmark families and fourteen methods with 95% intervals. (b) Protocol-level bootstrap mean-rank intervals. (c) Accuracy–calibration trade-off. (d) Paired FEDGEAR gain over each baseline adaptation. (e) Target-error quantile profile. (f) Generalization gap versus target accuracy. Long method labels are angled; colors follow one fixed palette throughout.

give $+0.28$ points, 95% CI $[0.14, 0.44]$, and $p_{\text{Holm}} = 0.0024$. All seven seed-averaged protocol differences are positive (protocol-first hierarchical CI $[0.08, 0.53]$; Wilcoxon $p = 0.0156$). FedWon is architecture-matched on six protocols, but not CIFAR-100C, where FEDGEAR uses source-selected standard bottleneck Linear/BN and FedWon retains WS (Appendix Table 3). Excluding it and reusing 18 archived blocks, all six matched protocol means remain positive: $+0.19$ points, CI $[0.04, 0.36]$, $p = 0.0313$. Both effects are small; ECE is competitive, not best. These results establish controlled feature-level superiority; they do not by themselves establish end-to-end superiority.

Figure 3 shows that pooling unequal tasks does not drive the result: FEDGEAR has the lowest mean rank, positive paired intervals against every baseline adaptation, and lower error across easy

and difficult quantiles. It improves accuracy without the high ECE of some correlated or weight-standardized alternatives, but is not the best-calibrated method.

4.3 Compositional shift and certificate diagnostics

CompShift-2F addresses Q2 without assuming that domain names reveal factor geometry. Appendix Figure 4 and Table 6 compare an inside-hull target, a missing $(+1, +1)$ corner outside the source hull but inside its span, and an orthogonal nonlinear factor. Selected actions cross the supporting plane with positive margin, verifying Proposition 2. The larger outside-span drop matches δ_{span} in Equation 9; it does not imply arbitrary-shift robustness.

4.4 Causal mechanisms, robustness, and cost

Appendix Figure 5 and Tables 4–7 answer Q3. Re-sketching matches the prescribed action numerically. The compensation-only control keeps endpoints 100% exterior but yields only 1.8–12.1% realized exteriority; full FedGEAR reaches 100%, yet gains only 0.03 accuracy points. Thus the certificate verifies exposure, while hard acted semantic training drives prediction. Unbalanced sketches and convex interpolation are respectively 0.09 and 0.03 points higher in the three-protocol mean, whereas removing the action or witness costs 0.24 or 0.13 points. Class balancing and reliability are therefore controls for geometry and optimization, not independent accuracy contributions. As margin grows, certificate distance rises but semantic agreement and large-margin accuracy decline.

The all-65-class short-budget audit sets the external-validity boundary (Appendix Table 5). Across eight methods, four LODO targets, and three seeds, FedWon and FedGEAR obtain 28.12% and 28.08%, respectively (difference -0.04; paired 95% CI $[-0.20, 0.13]$; Wilcoxon $p = 0.622$). This supports end-to-end competitiveness, not superiority; the broad comparison remains feature-level.

Appendix Figure 6 and Table 8 address Q4 with all fourteen methods. Across twelve stresses, FedGEAR wins eight with mean rank 1.33; FedWon wins four with 1.67. At 30% participation, FedGEAR trails by 0.23 points, exposing the low-sample bank boundary. Participation probes Δ_t, Dirichlet concentration probes sketch estimation, and label noise violates approximate label preservation; no label-noise guarantee is claimed.

Complexity and deployment. For q_t clients and rank r_t, server SVD costs $O(q_t(2d)r_t)$ and action construction $O(Md)$. Extra uplink is $2d + 1$ floats per client; broadcasting $(\bar{s}_t, B_t, \{c_k\})$ adds $2d + 2dr_t + q_t r_t$ floats per recipient and one pre-training synchronization. Appendix D gives bytes. Overhead is training-only; without a network model, we claim no latency advantage.

Statistical interpretation. Each protocol–seed pair is one unit; Holm correction covers all thirteen adaptations. Protocol-first resampling and seed-averaged tests against FedWon support consistency, not a large effect.

5 Conclusion and Limitations

We introduced FedGEAR, whose certificate makes the re-sketched feature consumed by the loss cross a supporting hyperplane outside the empirical environment hull. It leads thirteen common-interface baseline adaptations across seven controlled feature-level protocols; theory and matched diagnostics establish verifiable exposure, not an accuracy or target-coverage guarantee. An all-class raw-pixel audit supports comparable end-to-end performance with FedWon, not superiority. Class balancing, the witness, and reliability weighting are supporting controls, not independent accuracy claims. Limitations include incomplete moment-action coverage, unknown target direction, nonprivate sketches, and same-minibatch selection bias; structured/private actions, split-minibatch validation, and multi-backbone studies remain future work.

Reproducibility statement. Appendix D and the anonymous supplement specify data, protocols, hyperparameters, adaptations, evaluation, and diagnostics; Appendix A gives the assumptions and proofs. Comparisons use predeclared final rounds and matched seeds where applicable.

REFERENCES

Ruqi Bai, Saurabh Bagchi, and David I. Inouye. Benchmarking algorithms for federated domain generalization. In *International Conference on Learning Representations*, 2024.

Shuai Gong, Chaoran Cui, Xiaolin Dong, Xiushan Nie, Lei Zhu, and Xiaojun Chang. Token-level prompt mixture with parameter-free routing for federated domain generalization. *IEEE Transactions on Image Processing*, 35:656–669, 2026. doi: 10.1109/TIP.2026.3652431.

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 Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

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

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the International Conference on Machine Learning*, 2020.

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

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021a.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of MLSys*, 2020.

Xiaotong Li, Yongxing Dai, Yixiao Ge, Jun Liu, Ying Shan, and Ling-Yu Duan. Uncertainty modeling for out-of-distribution generalization. In *International Conference on Learning Representations*, 2022.

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

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 AISTATS*, 2017.

A. Tuan Nguyen, Philip Torr, and Ser Nam Lim. FedSR: A simple and effective domain generalization method for federated learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Trong-Binh Nguyen, Minh-Duong Nguyen, Jinsun Park, Quoc-Viet Pham, and Won Joo Hwang. Federated domain generalization with data-free on-server matching gradient. In *International Conference on Learning Representations*, 2025.

Jungwuk Park, Dong-Jun Han, Jinho Kim, Shiqiang Wang, Christopher G. Brinton, and Jaekyun Moon. StableFDG: Style and attention based learning for federated domain generalization. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Changlong Shi, Jinmeng Li, He Zhao, Dandan Guo, and Yi Chang. FedLWS: Federated learning with adaptive layer-wise weight shrinking. In *International Conference on Learning Representations*, 2025.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Chenyang Wang, Junjun Jiang, Xingyu Hu, Xianming Liu, and Xiangyang Ji. Federated domain generalization via data-centric flatness optimization. *Pattern Recognition*, 174:113023, 2026. doi: 10.1016/j.patcog.2025.113023.

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

Kunda Yan, Sen Cui, Abudukelimu Wuerkaixi, Jingfeng Zhang, Bo Han, Gang Niu, Masashi Sugiyama, and Changshui Zhang. Balancing similarity and complementarity for federated learning. In *Proceedings of the International Conference on Machine Learning*, 2024.

Yunlu Yan, Chun-Mei Feng, Wangmeng Zuo, Salman Khan, Yong Liu, and Lei Zhu. On the importance of language-driven representation learning for heterogeneous federated learning. In *International Conference on Learning Representations*, 2025.

Zilong Yin, Haoyu Wang, Bin Chen, Xin Zhang, Xiaogang Lin, Hangling Sun, Anji Li, and Chenyu Zhou. Federated semi-supervised representation augmentation with cross-institutional knowledge transfer for healthcare collaboration. *Knowledge-Based Systems*, 300:112208, 2024. doi: 10.1016/j.knosys.2024.112208.

Ruipeng Zhang, Qinwei Xu, Jiangchao Yao, Ya Zhang, Qi Tian, and Yanfeng Wang. Federated domain generalization with generalization adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3954–3963, 2023.

Zheyuan Zhang, Bin Wang, Debesh Jha, Ugur Demir, and Ulas Bagci. Domain generalization with correlated style uncertainty. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2000–2009, 2024.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with MixStyle. In *International Conference on Learning Representations*, 2021.

Tianfei Zhou and Ender Konukoglu. Federated feature augmentation. In *International Conference on Learning Representations*, 2023.

Weiming Zhuang and Lingjuan Lyu. FedWon: Triumphing multi-domain federated learning without normalization. In *International Conference on Learning Representations*, 2024.

AI USE STATEMENT

Generative AI tools were used to assist with generating and refining portions of the experimental code and with checking and revising intermediate steps in the mathematical proofs. All AI-assisted code was manually reviewed and tested, and all mathematical derivations and proofs were independently verified by the authors. Generative AI was also used for language polishing and improving the clarity of the manuscript; the research problem, proposed method, experimental design, and interpretation of results were developed by the authors. The authors take full responsibility for the final content of the paper, including its code, theoretical claims, experimental results, and text.

A DETAILED THEORY

This appendix separates four questions around the final semantic predictor: whether an action is outside the empirical hull, whether finite candidates cover the licensed action set, whether that set approximates the target, and whether federated optimization tracks a dynamically changing objective.

A.1 NOTATION AND ASSUMPTIONS

Let $Z = h_\theta(X)$ and let P_s^{ZY} denote the representation–label distribution in environment coordinate s. For coordinate a, let $\mathcal{A}_a \# P_s^{ZY}$ be the push-forward that applies Equation 5 to Z while retaining Y. The deployed semantic predictor induces probabilities $p_w(z)$. Loss $\ell(p, y) \in [0, 1]$ is L_p-Lipschitz in

ℓ_1 probability distance. The risk bounds are stated for this bounded probability-Lipschitz loss class. The cross-entropy terms in Equation 7 are training surrogates and, without additional restrictions, do not themselves satisfy these boundedness and global Lipschitz assumptions.

Assumption 1 (Target-to-full-action approximation)**.** *Let $a_\star$ be the best target approximation inside the action set licensed by the full-client bank. For a source mixture $\bar{P}$,*

$$d_{\mathcal{F}}(P_{s_\star}^{ZY}, \mathcal{A}_{a_\star}\#\bar{P}^{ZY}) \le L_s\delta_{\mathrm{span}} + \varepsilon_{\mathrm{sem}}, \tag{10}$$

where $d_{\mathcal{F}}$ is the integral probability metric induced by the loss class. The residual includes non-affine covariate effects and changes in $P(Y \mid Z)$.

Assumption 2 (Action smoothness)**.** *For all licensed a, a', $|R_{\mathcal{A}_a\#\bar{P}}(f_w) - R_{\mathcal{A}_{a'}\#\bar{P}}(f_w)| \le L_s\|a - a'\|_2$.*

Assumption 3 (Deployed-predictor consistency)**.** *Uniformly over the feasible retained set $\mathcal{C}_t^+$, the expected symmetric KL of the final semantic probabilities satisfies*

$$\sup_{c\in\mathcal{C}_t^+} \mathbb{E}\,\mathrm{SKL}(p_w(Z), p_w(\mathcal{A}_c(Z))) \le \varepsilon_{\mathrm{pred}}. \tag{11}$$

These assumptions are testable only indirectly. CompShift controls δ_{span} by construction; the action/no-action and consistency ablations probe the other terms.

A.2 Finite-sample sketch estimation

The population geometry above is observed through per-class empirical moments. This distinction connects the class-balancing proposition to finite client data rather than treating the uploaded sketch as exact.

Lemma 2 (Balanced-sketch consistency)**.** *Assume that each conditional raw first- and second-moment estimator has uniformly bounded sub-exponential fluctuations, every retained population variance is at least $v_{\min} > 0$, and $n_{k,y} \ge n_{\min}$. For finite K, d, and $|\mathcal{Y}|$,*

$$\max_k \|\hat{s}_k - s_k\|_2 = \widetilde{O}_p\Big(\sqrt{d/n_{\min}}\Big) =: \varepsilon_{\mathrm{sk}}, \tag{12}$$

where hidden constants depend on the moment bounds and $v_{\min}$, and $\widetilde{O}_p$ suppresses logarithmic factors.

Proof. Bernstein concentration and a finite union bound give the usual root-$n_{\min}$ rate, up to logarithmic factors, for all conditional raw first and second moments. Equal-class averaging preserves consistency. The map from these raw moments to (a, v) is continuous, including $\hat{a}^{\odot 2} - a^{\odot 2} = (\hat{a} - a) \odot (\hat{a} + a)$. Consequently $\|\hat{v} - v\|_\infty \le v_{\min}/2$ with probability tending to one, so every retained empirical variance is at least $v_{\min}/2$ and the log map is locally Lipschitz on this event. The continuous-mapping argument and Euclidean aggregation yield Equation 12. Constants are left implicit because they depend on conditional mean and second-moment scales. □

The $n_{\min}^{-1/2}$ dependence explains why sparse classes and partial participation enlarge bank error even though population balancing removes label-mixture weights. The unbalanced-sketch replacement tests confounding, while the Dirichlet and participation diagnostics probe finite-sample and bank-estimation boundaries.

A.3 Sketch and certificate proofs

Proof of Proposition 1. For common class set $\mathcal{Y}$, equal conditional first and second moments imply $a_k = |\mathcal{Y}|^{-1}\sum_y \mu_y = a_j$ and $v_k = |\mathcal{Y}|^{-1}\sum_y m_{2,y} - a_k^{\odot 2} = v_j$, independently of q_k and q_j. By contrast, unconditional means are $\sum_y q_k(y)\mu_y$ and $\sum_y q_j(y)\mu_y$, which differ whenever label proportions and class means differ. □

Proof of Lemma 1. Equal-class averaging gives $a(\mathcal{A}_{\tilde{s}}(h)) = \tilde{a}$. Affine equivariance gives $v(\mathcal{A}_{\tilde{s}}(h)) = \tilde{v}$; therefore its stored log-scale is $\frac{1}{2}\log(\tilde{v} + \epsilon) = \tilde{b}$. Hence the re-sketched coordinate is $\tilde{s}$, and its bank projection is exactly c_{act}. □

Proof of Proposition 2. By construction,

$$\begin{aligned} u^\top c_{\text{act}} &= (1-\tau)u^\top c_B + \tau u^\top c_{\text{end}} \\ &= (1-\tau)u^\top c_B + h_t(u) + m\sigma_t(u) - (1-\tau)u^\top c_B \\ &= h_t(u) + m\sigma_t(u) > h_t(u). \end{aligned}$$

Every $c \in \mathcal{H}_t$ satisfies $u^\top c \le h_t(u)$ by the definition of a support function. Thus the hyperplane $u^\top c = h_t(u)$ strictly separates c_{act} from $\mathcal{H}_t$. Because $\|u\|_2 = 1$, the Euclidean distance to that plane is $m\sigma_t(u)$. □

Proposition 3 (Certificate under bank estimation error)**.** *Let $h(u)$ be the support of the population projected hull and $\hat{h}(u)$ its empirical counterpart. Suppose $|\hat{h}(u) - h(u)| \le \varepsilon_H$ and the population projection of the implemented acted coordinate differs from its empirical coefficient by at most ε_A along u. Then an empirical certificate of margin $\mu = m\hat{\sigma}(u)$ implies*

$$u^\top c_{\text{act}}^{\text{pop}} - h(u) \ge \mu - \varepsilon_H - \varepsilon_A. \tag{13}$$

Thus the certificate transfers to the population hull whenever $\mu > \varepsilon_H + \varepsilon_A$.

Proof. Insert and subtract the empirical acted coefficient and support: $u^\top c_{\text{act}}^{\text{pop}} - h(u) = [u^\top \hat{c}_{\text{act}} - \hat{h}(u)] + u^\top(c_{\text{act}}^{\text{pop}} - \hat{c}_{\text{act}}) + [\hat{h}(u) - h(u)]$. The first bracket is μ and the remaining brackets are lower bounded by $-\varepsilon_A$ and $-\varepsilon_H$. □

Equation 13 separates the deterministic empirical guarantee actually logged by the algorithm from population transfer, which additionally depends on Equation 12 and subspace perturbation. We never report empirical exteriority as a population certificate.

Lemma 3 (Prediction displacement)**.** *If $\mathbb{E}\,\mathrm{SKL}(p_w(Z), p_w(\mathcal{A}_c(Z))) \le \varepsilon_{\text{pred}}$, then $\mathbb{E}\|p_w(Z) - p_w(\mathcal{A}_c(Z))\|_1 \le \sqrt{2\varepsilon_{\text{pred}}}$.*

Proof. Pinsker applied in both directions gives $\mathrm{KL}(p\|q) \ge \frac{1}{2}\|p-q\|_1^2$ and $\mathrm{KL}(q\|p) \ge \frac{1}{2}\|p-q\|_1^2$. Hence $\mathrm{SKL}(p,q) \ge \frac{1}{2}\|p-q\|_1^2$ under our averaged definition. Cauchy–Schwarz then yields $\mathbb{E}\|p - q\|_1 \le \sqrt{\mathbb{E}\|p-q\|_1^2} \le \sqrt{2\varepsilon_{\text{pred}}}$. □

A.4 Coverage and target risk

Condition on a nonempty feasible retained set $\mathcal{C}_t^+$, and let $\hat{a} \in \mathcal{C}_t^+$ be its closest candidate to $a_\star$. We define η_M as the residual approximation error of the retained set after M raw proposals; together with partial-bank displacement,

$$\|a_\star - \hat{a}\|_2 \le \Delta_t + \eta_M, \tag{14}$$

where Δ_t measures set displacement from using $\mathcal{S}_t$. The target-to-full-span error δ_{span} already appears in Assumption 1 and is deliberately absent here. Because feasibility filtering changes the proposal law, η_M is retained explicitly rather than inferred from the raw sphere-cap calculation below.

Define the worst retained-candidate risk

$$R_t^{\max}(f_w) = \max_{c \in \mathcal{C}_t^+} R_{\mathcal{A}_c \# \bar{P}}(f_w). \tag{15}$$

The coverage candidate $\hat{a}$ need not be the candidate selected by maximum CE; its risk is nevertheless upper bounded by Equation 15, which is the population counterpart of hard mining.

Theorem 2 (Target risk for the deployed predictor)**.** *Under Assumptions 1–3,*

$$R_\star(f_w) \le R_t^{\max}(f_w) + L_s(\delta_{\text{span}} + \Delta_t + \eta_M) + \varepsilon_{\text{sem}}, \tag{16}$$

$$R_\star(f_w) \le R_{\bar{P}}(f_w) + L_p\sqrt{2\varepsilon_{\text{pred}}} + L_s(\delta_{\text{span}} + \Delta_t + \eta_M) + \varepsilon_{\text{sem}}. \tag{17}$$

Proof. Assumption 1 first yields $R_\star(f_w) \le R_{\mathcal{A}_{a_\star} \# \bar{P}}(f_w) + L_s\delta_{\text{span}} + \varepsilon_{\text{sem}}$. Action smoothness and Equation 14 give $R_{\mathcal{A}_{a_\star} \# \bar{P}}(f_w) \le R_{\mathcal{A}_{\hat{a}} \# \bar{P}}(f_w) + L_s(\Delta_t + \eta_M)$. Since $\hat{a} \in \mathcal{C}_t^+$,

$R_{\mathcal{A}_{\hat{a}}\#\bar{P}}(f_w) \le R_t^{\max}(f_w)$, proving Equation 16 with δ_{span} appearing exactly once. Uniform candidate consistency, loss Lipschitzness, and the prediction-displacement lemma give $R_{\mathcal{A}_c\#\bar{P}}(f_w) \le R_{\bar{P}}(f_w) + L_p\sqrt{2\varepsilon_{\text{pred}}}$ for every retained c; taking the maximum proves Equation 17. Thus consistency is needed only for the source-risk form. The source and maximum-CE terms in Equation 7 are empirical optimization surrogates for the corresponding risks in the decomposition; the theorem does not assert that minimizing cross-entropy directly minimizes the stated bound. Because selection reuses the minibatch, no unbiasedness claim is made. □

The theorem is deliberately one-way. Proposition 2 ensures nonzero empirical exteriority, but a large margin can increase ε_{sem}; a target outside the learned span increases δ_{span}; and small M increases η_M. Thus "farther outside" is not monotonically better.

A.5 Finite directional coverage

The continuous sphere-cap expression below is stated for $r_t \ge 2$. For $r_t = 1$, the unit direction set is $S^0 = \{-1, +1\}$; the normalized scalar Gaussian in the implementation selects either sign with equal probability. If $r_t = 0$, no bank is formed, no exterior action is generated, and the minibatch follows the source-only branch. Suppose the licensed coefficient set is contained in a radius-R ball in r_t dimensions and candidate directions $U_1, \dots, U_M$ are sampled uniformly on the sphere. For a desired direction v, let

$$p_{r_t}(\epsilon) = \Pr[\angle(U_j, v) \le \epsilon] = \frac{\int_0^\epsilon \sin^{r_t-2}\theta\, d\theta}{\int_0^\pi \sin^{r_t-2}\theta\, d\theta}. \tag{18}$$

Independence gives $\Pr[\min_j \angle(U_j, v) > \epsilon] = (1 - p_{r_t}(\epsilon))^M \le e^{-Mp_{r_t}(\epsilon)}$. This statement concerns the M raw uniformly proposed directions before variance-feasibility rejection. After rejection, the retained directions are not generally i.i.d. uniform and their number may be smaller than M; we claim no post-rejection directional or action-space coverage guarantee. Moreover, converting angular error into η_M requires regularity of the complete map $u \mapsto c_{\text{act}}(u)$, whose radius depends on $h_t(u)$, $\sigma_t(u)$, and $u^\top c_B$. We therefore retain η_M as the explicit feasible-candidate approximation term and make no unsupported Lipschitz-rate claim. Effective-rank truncation improves only the stated pre-filter sphere-cap probability for fixed M.

Candidate construction and maximum-CE selection use the same minibatch, so a fixed, sample-independent finite-class Hoeffding bound does not apply. We make no such generalization claim; the random-candidate control isolates hard selection empirically.

B Optimization of a Smoothed Dynamic-Bank Surrogate

The maximum-CE switch in Algorithm 1 is nonsmooth at candidate ties, so a smooth descent lemma does not apply to its exact iterates. For analysis only, condition on a realized bank, directions, and retained set; for candidate score $s_{t,m}(w)$, replace the one-hot argmax by fixed-temperature weights $\pi_{t,m}^{(\kappa)} = \exp(\kappa s_{t,m}) / \sum_j \exp(\kappa s_{t,j})$ and average the candidate-conditioned terms of Equation 7. The empty set retains the source-only branch. Let $F_t^{(\kappa)}$ be the conditional expected surrogate objective over the coupled client-set, bank, direction, and minibatch draws. Its log-sum-exp score satisfies

$$\max_m s_{t,m} \le \kappa^{-1} \log \sum_m e^{\kappa s_{t,m}} \le \max_m s_{t,m} + \frac{\log M}{\kappa}. \tag{19}$$

This is only a selection-score function-value comparison, not a stationarity transfer to hard iterates. Reliability weighting has bias b_t relative to $\nabla F_t^{(\kappa)}$; round drift is $\zeta_t^{(\kappa)} = |F_{t+1}^{(\kappa)}(w_{t+1}) - F_t^{(\kappa)}(w_{t+1})|$.

Assumption 4 (Fixed-temperature smooth surrogate)**.** *For fixed finite κ, every realized surrogate local objective is L_κ-smooth and $F_t^{(\kappa)} \ge F_{\text{inf}}^{(\kappa)}$. Before local drift and reliability bias, the coupled random round direction is unbiased for $\nabla F_t^{(\kappa)}$ with conditional variance at most $(\sigma^2 + G^2)/q_t$. Client gradient dissimilarity is at most G^2, $\|b_t\| \le B_t$, and $\eta \le 1/(4L_\kappa E)$. These are surrogate assumptions; typically L_κ grows with κ.*

Theorem 3 (Conditional stationarity of the fixed-temperature surrogate)**.** *Under Assumption 4, the iterates of the surrogate procedure satisfy, for universal constants C_1, C_2, C_3,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F_t^{(\kappa)}(w_t)\|^2 \le \frac{4(F_0^{(\kappa)}(w_0)-F_{\inf}^{(\kappa)})}{\eta ET} + \frac{4\sum_{t=0}^{T-1}\mathbb{E}\zeta_t^{(\kappa)}}{\eta ET} + \frac{C_1 L_\kappa \eta E(\sigma^2+G^2)}{\min_t q_t} + C_2 L_\kappa^2 \eta^2 E^2 G^2 + \frac{C_3}{T}\sum_{t=0}^{T-1} B_t^2. \tag{20}$$

Proof. Write the averaged surrogate direction as $g_t = \nabla F_t^{(\kappa)}(w_t) + \xi_t + d_t + b_t$, where $\mathbb{E}_t\xi_t = 0$ and $\mathbb{E}_t\|\xi_t\|^2 \le (\sigma^2+G^2)/q_t$; standard local-SGD drift gives $\mathbb{E}_t\|d_t\|^2 \le C\eta^2 L_\kappa^2 E^2 G^2$. Surrogate smoothness and $w_{t+1} = w_t - \eta E g_t$ imply

$$\begin{aligned}\mathbb{E}_t F_t^{(\kappa)}(w_{t+1}) &\le F_t^{(\kappa)}(w_t) - \eta E\langle \nabla F_t^{(\kappa)}, \mathbb{E}_t g_t\rangle + \tfrac{L_\kappa \eta^2 E^2}{2}\mathbb{E}_t\|g_t\|^2 \\ &\le F_t^{(\kappa)}(w_t) - \tfrac{\eta E}{4}\|\nabla F_t^{(\kappa)}(w_t)\|^2 + C_1' L_\kappa \eta^2 E^2(\sigma^2+G^2)/q_t \\ &\quad + C_2' L_\kappa^2 \eta^3 E^3 G^2 + C_3' \eta E B_t^2,\end{aligned}$$

where Young's inequality absorbs the inner products with d_t and b_t, and $\eta L_\kappa E \le 1/4$ absorbs the remaining gradient term. By definition, $F_{t+1}^{(\kappa)}(w_{t+1}) \le F_t^{(\kappa)}(w_{t+1}) + \zeta_t^{(\kappa)}$. Summing, telescoping, using $F_T^{(\kappa)} \ge F_{\inf}^{(\kappa)}$, and dividing by $\eta ET/4$ proves Equation 20. □

Equation 20 isolates stochastic/local error, bank drift, and reliability bias at fixed κ. As κ grows, Equation 19 shrinks but L_κ and the step-size restriction can worsen. We neither take a cost-free $\kappa \to \infty$ limit nor claim stationarity for Algorithm 1's hard iterates; full nonsmooth dynamic-bank analysis is outside scope. Participation/rank diagnostics probe bank stability, and the no-reliability control probes B_t.

C Theory–Algorithm–Experiment Correspondence

Table 2: Theory–algorithm–experiment correspondence. Every theoretical object is paired with its implemented mechanism and the evidence that can diagnose it; proxy-only rows are named explicitly.

Theory object	Algorithmic realization	Direct experimental evidence
$\varepsilon_{\mathrm{sk}}$, Δ_t	shared-model class-balanced sketches; effective-rank bank	unbalanced-sketch replacement; Dirichlet and participation sweeps; rank traces
$m\sigma(u)$, Lemma 1	class-balanced realization plus step-compensated supporting-plane endpoint	prescribed/re-sketched exterior rates, coordinate error, and no-step-compensation control
η_M	M raw proposals, feasibility filtering, and maximum-CE selection	pre-filter sphere-cap analysis and random-candidate replacement; no post-filter coverage claim
$\varepsilon_{\mathrm{sem}}$, $\varepsilon_{\mathrm{pred}}$	acted semantic CE and symmetric-KL/witness losses	prediction agreement, confidence and acted-label drops versus margin; no-witness/no-consistency controls
δ_{span}	low-rank diagonal moment-action family	in-hull, exterior-in-span, and orthogonal outside-span CompShift targets
B_t, $\zeta_t^{(\kappa)}$	reliability aggregation and round-wise rebuilt bank in the smooth surrogate	no-reliability control; participation stress and certificate/rank dynamics

D Additional Experimental Details

Data and splits. OfficeHome-65 features are extracted once with torchvision ResNet-18 ImageNet-1K V1 weights and stored with image paths, labels, domain IDs, class names, and encoder metadata. Each held-out domain is capped only for evaluation memory; source validation is a deterministic 15% domain-stratified split. CIFAR feature caches contain clean client features, a clean calibration split, and fixed corruption episodes. Every configuration stores its exact selected client indices per round.

Optimization. OfficeHome uses 24 rounds, two local epochs, learning rate 0.018, and at most 900 examples per client. CIFAR-10C uses 20 rounds, two epochs, and rate 0.025; CIFAR-100C uses 22 rounds, two epochs, and rate 0.022; CompShift uses 22 rounds and rate 0.024. Batch size is 128, weight decay 10^{-4}, momentum 0.9, and participation 0.67. Gradients are clipped at norm five. Reported checkpoints are the fixed final communication round.

Unless otherwise stated, Equation 7 uses $\lambda_a = 0.06$, $\lambda_w = 0.03$, $\beta = 0.70$, $\lambda_c = 0.02$, and $\lambda_{cw} = 0.005$; Equation 8 uses $\gamma = 1.25$. For byte scale, the main OfficeHome setting ($d = 64$, $r_t = 4$, $q_t = 7$, float32) adds 129 floats (0.50 KiB) uplink and 668 floats (2.61 KiB) downlink per selected client, or 5,579 floats (21.79 KiB) aggregate per round. We do not convert the extra synchronization barrier to latency because no network is simulated.

Baseline adaptation scope. Original methods span image networks, feature modules, and server algorithms. We therefore compare mechanisms at a common trainable bottleneck and name the adaptation explicitly rather than claiming bitwise reproduction. All feature-statistic methods receive the same shared-model sketch bank and augmentation probability. FedWon uses weight standardization throughout; FEDGEAR shares that implementation except on CIFAR-100C, where standard Linear/BN is used only in the trainable bottleneck and the semantic head z_w is unchanged. FedWon is therefore a matched normalization control on the other families and a strong, but not architecture-matched, comparator on CIFAR-100C. FedBN uses source-averaged running statistics because target-specific BN is unavailable in FDG. Server-only methods operate on identical flattened client updates. Dataset-family schedules are frozen before target evaluation, and every method has one target-evaluated configuration per family; there is no per-target search or checkpoint selection. Table 3 reports the verifiable frozen settings rather than reconstructing an unavailable per-trial source-development history.

Recent data-centric flatness methods require data-space adversarial augmentation and end-to-end optimization, while prompt-mixture systems require prompt-capable backbones outside this controlled feature interface. A shallow surrogate would not faithfully preserve their defining mechanisms; we therefore discuss these lines as related work rather than claiming to outperform them. The raw-pixel audit is separately labeled and supports end-to-end viability and comparability only.

Statistics. The primary unit is a protocol–seed pair ($n = 21$), not an individual target example. We report paired bootstrap intervals with 10,000 resamples and two-sided Wilcoxon tests, with thirteen comparisons corrected by Holm's method. For the strongest common-interface comparator, a 20,000-replicate hierarchical bootstrap resamples seven protocols and then their seeds; a second Wilcoxon test uses only the seven protocol means. The analysis writes all raw statistics and effect sizes to machine-readable files.

E Certificate, Semantic, and External-Validity Audits

Table 4 isolates Proposition 2 and Lemma 1: it reports prescribed exteriority, exteriority after re-sketching the actual transformed feature, and their projected discrepancy. The uncompensated endpoint is exterior by construction, but the convex action step returns roughly nine of ten realized coordinates inside the hull. Accuracy barely changes, so compensation is necessary for verifiable exposure but is not the entire predictive mechanism. The OH-10 row is only an auxiliary certificate-control setting, not a separate end-to-end external-validity claim. Figure 5(c–d) varies m over $\{0, .05, .10, .18, .30, .50, .75\}$ and operationalizes $\varepsilon_{\mathrm{sem}}$ through agreement, confidence, and acted-label degradation.

Table 3: Common-bottleneck baseline protocol. All rows share the dataset schedule, source-only development, client draws, minibatch order, and predeclared final checkpoint. "Target configs" counts configurations evaluated on each held-out target, not source-only development trials.

Method	Common-bottleneck adaptation	Frozen setting or rule	Target configs
FedAvg	Sample-weighted local SGD	No method-specific coefficient	1
FedProx	Proximal local objective	$\mu = 0.01$	1
FedBN	Client batch normalization	Source states averaged; no target BN	1
FedSR	Bottleneck moment regularization	$\lambda_{\mathrm{SR}} = 0.012$	1
MixStyle	Peer-sketch convex interpolation	$\lambda \sim \mathcal{U}(0,1)$; $p_{\mathrm{aug}} = 0.75$	1
DSU	Diagonal sketch uncertainty	strength 0.45; $p_{\mathrm{aug}} = 0.75$	1
FedFA	Global/local diagonal uncertainty	strength 0.45; $p_{\mathrm{aug}} = 0.75$	1
StableFDG	Peer-style extrapolation	strength 0.45; $p_{\mathrm{aug}} = 0.75$	1
CSU	Correlated low-rank sampling	$r_t \leq 4$; $p_{\mathrm{aug}} = 0.75$	1
FedWon	Weight-standardized trainable layers	Enabled for every dataset family	1
FedOMG	Normalized-update agreement	Score coefficient 2.5	1
FedLWS	Layer-wise shrinkage	$\gamma_\ell \in [0.75, 1.00]$	1
FedSaC	Similarity plus sketch complementarity	Complementarity weight 0.25	1
FEDGEAR	Certified hard action plus witness	WS except CIFAR-100C (standard bottleneck Linear/BN; z_w unchanged); $r_t \leq 4$, $M = 3$	1

Table 4: Matched certificate and realization control over three seeds. Both variants share the learned subspace, candidates, action step, losses, and aggregation. "Re-sketched" recomputes Equation 2 on the transformed feature consumed by the loss; coordinate error is its projected distance from the prescribed action.

Protocol	Variant	Prescribed ext. (%) ↑	Re-sketched ext. (%) ↑	Coord. error ↓	Accuracy (%) ↑
OH-65 Art held out	FedGEAR	**100.0**	**100.0**	1.38e-07	**53.83**
OH-65 Art held out	No step comp.	12.1	12.1	**1.32e-07**	53.78
CIFAR-10C	FedGEAR	**100.0**	**100.0**	2.09e-07	**52.60**
CIFAR-10C	No step comp.	11.7	11.7	**1.72e-07**	52.60
CompShift-2F	FedGEAR	**100.0**	**100.0**	1.42e-07	**81.32**
CompShift-2F	No step comp.	10.8	10.8	**1.37e-07**	81.24
OH-10 composition	FedGEAR	**100.0**	**100.0**	1.09e-07	**33.83**
OH-10 composition	No step comp.	1.8	1.8	**0.00e+00**	**33.83**

The raw-image protocol uses 96-pixel inputs, ImageNet-initialized but fully trainable ResNet-18, a trainable 64-dimensional bottleneck, 15 communication rounds, one local epoch, and 67% client participation. Each source domain is split into two clients, capped at 450 training examples per client; each held-out target is capped at 2,000 examples. The four LODO targets, eight methods, and three seeds produce 96 target-blind runs, all evaluated only at the predeclared final round. Reusing the same source partitions, initialization, and minibatch order within each target–seed block supports a paired comparison. This audit supports transfer of the mechanism to trainable features, while the controlled grid remains the basis for the thirteen-method common-interface comparison.

Table 5: All-65-class OfficeHome raw-pixel audit under a fixed short-budget protocol, with an end-to-end trainable ResNet-18 over four LODO targets and three seeds (accuracy, mean±standard deviation in %). The eight-method set includes a FedAvg anchor and strong representatives of feature regularization, statistic perturbation, normalization, and recent FDG. Best and second-best values are bold and underlined.

Method	Art ↑	Clip. ↑	Product ↑	Real ↑	LODO mean ↑
FedAvg	15.77±0.77	18.53±1.56	18.10±1.25	23.67±0.30	19.02±0.79
FedSR	15.58±0.76	18.68±1.57	18.15±1.23	23.85±0.39	19.07±0.79
DSU	15.72±1.07	18.73±1.61	18.22±1.20	23.73±0.38	19.10±0.93
FedFA	15.85±0.87	18.60±1.72	18.25±1.28	23.60±0.33	19.08±0.86
StableFDG	15.53±1.06	18.52±1.70	18.30±1.13	24.10±0.13	19.11±0.95
CSU	16.08±0.56	18.62±1.70	18.33±1.68	23.20±0.48	19.06±0.85
FedWon	**21.85±0.70**	**26.55±1.05**	<u>30.05±0.38</u>	**34.03±0.64**	**28.12±0.34**
FedGEAR	<u>21.73±0.50</u>	<u>26.42±1.08</u>	**30.20±0.52**	<u>33.98±0.36</u>	<u>28.08±0.32</u>

Table 6: Complete CompShift-2F final-round comparison over three seeds. Accuracies and drops are percentage points. "Exterior in span" is the missing observed-factor combination; "outside span" adds an orthogonal nonlinear factor. Best and second-best values are bold and underlined.

Method	Inside hull ↑	Exterior in span ↑	Outside span ↑	Hull drop ↓	Span drop ↓
FedAvg	81.29	79.73	79.69	**1.56**	0.04
FedProx	81.38	79.69	79.73	1.69	-0.04
FedBN	81.29	79.73	79.69	<u>1.56</u>	0.04
FedSR	81.33	79.73	79.73	1.60	0.00
MixStyle	81.47	79.29	79.42	2.18	<u>-0.13</u>
DSU	81.11	78.53	78.44	2.58	0.09
FedFA	81.11	78.49	78.44	2.62	0.04
StableFDG	80.80	78.53	78.44	2.27	0.09
CSU	81.69	79.60	79.64	2.09	-0.04
FedWon	<u>82.36</u>	79.91	80.04	2.44	**-0.13**
FedOMG	81.38	79.51	79.56	1.87	-0.04
FedLWS	81.42	79.78	79.78	1.64	0.00
FedSaC	81.91	<u>80.31</u>	<u>80.36</u>	1.60	-0.04
FedGEAR	**82.80**	**80.58**	**80.58**	2.22	0.00

F Additional Mechanism and Robustness Plots

G Scope of Claims and Evidence

The strongest controlled-grid claim is "highest mean accuracy in all seven protocols," supported by four OfficeHome targets, CIFAR-10C/100C, and CompShift over three seeds. Significance against all baseline adaptations is restricted to the 21-unit Holm-corrected analysis; consistency against FedWon is additionally supported at protocol level and remains positive after excluding the non-architecture-matched CIFAR-100C family. The certificate claim is deterministic and separated from its small, mixed accuracy effect by the matched control. We do not claim best ECE, arbitrary OOD robustness, end-to-end superiority, privacy, or that an unknown target lies in the learned span.

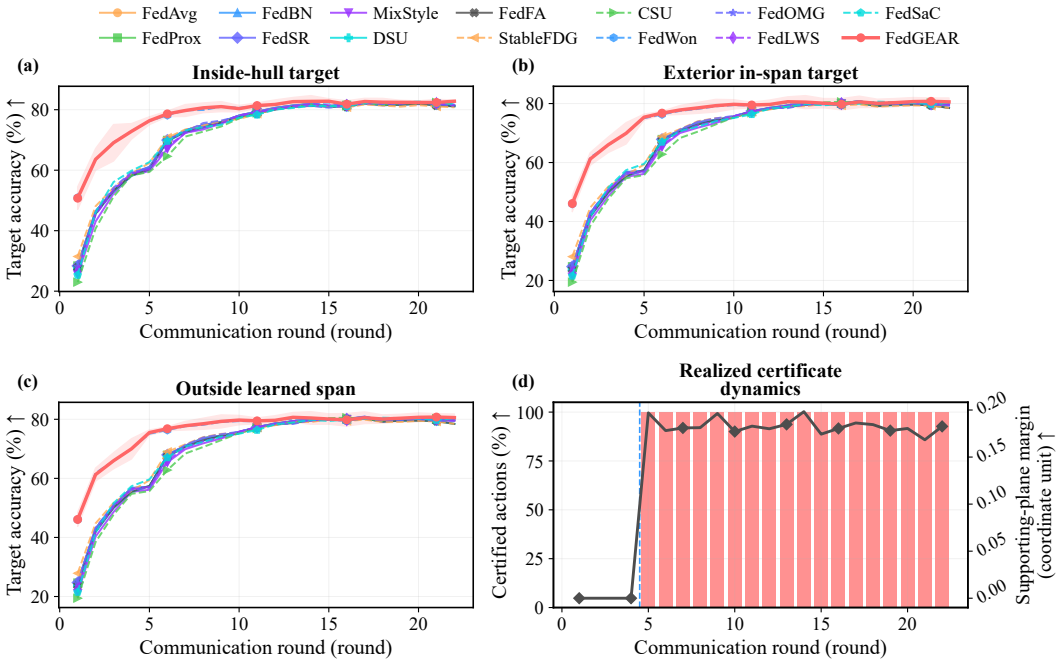


Figure 4: Explicit composition test in a readable 2×2 layout. (a–c) Round-wise accuracy of FEDGEAR and all thirteen common-interface adaptations on an inside-hull target, a missing exterior in-span corner, and an outside-learned-span target; the shaded band is the 95% interval for FEDGEAR. (d) Measured certificate rate and supporting-plane margin. Complete final-round values are reported in Table 6.

Table 7: Complete same-budget mechanism ablation (three seeds; target accuracy, %). The figure emphasizes signed effects and uncertainty; this table carries all dense numerical values. Best and second-best values are bold and underlined.

Controlled variant	OfficeHome-65 Art ↑	CIFAR-10C ↑	CompShift-2F ↑	Mean ↑
FedGEAR	<u>53.83</u>	52.60	<u>81.32</u>	62.58
Unbalanced sketch	**53.84**	<u>52.77</u>	**81.41**	**62.67**
Random subspace	53.77	52.60	81.23	62.53
Full-dimensional noise	53.54	52.72	81.20	62.49
Convex interpolation	53.74	**52.80**	81.29	<u>62.61</u>
No step compensation	53.78	52.60	81.24	62.54
Random candidate	53.80	52.61	81.14	62.52
No witness head	53.83	52.55	80.98	62.45
No environment action	53.52	52.58	80.90	62.34
No prediction consistency	53.67	52.62	81.17	62.49
No reliability weighting	53.83	52.57	81.30	62.57

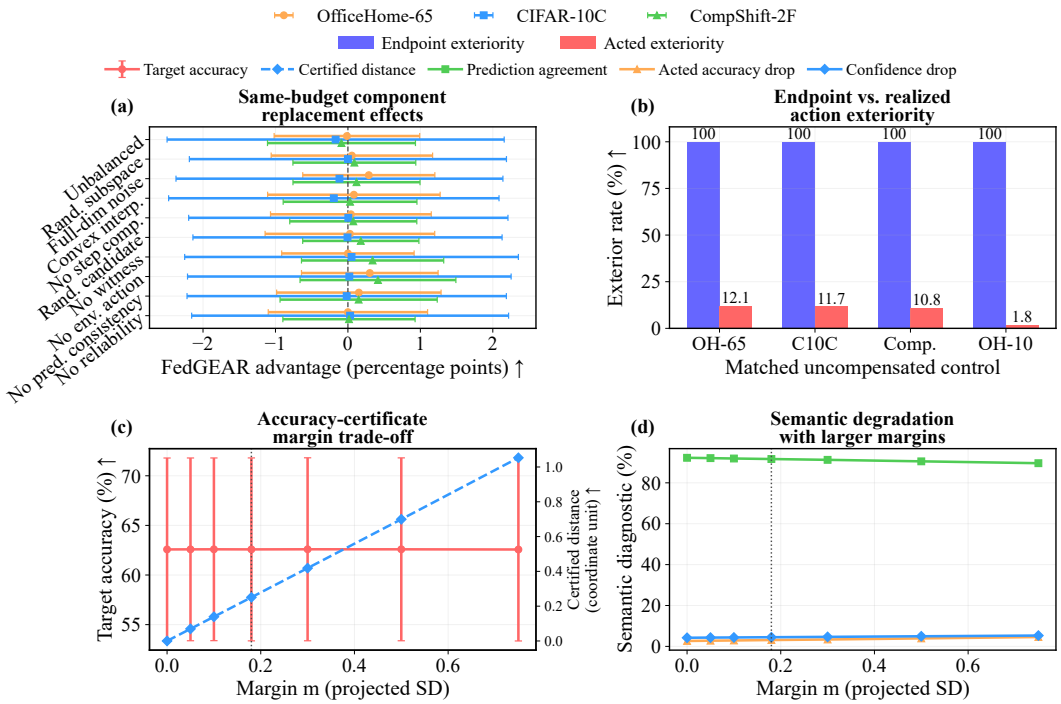


Figure 5: Controlled mechanism study. (a) Horizontal 95% intervals for the FedGEAR advantage over each same-budget replacement on three protocols. (b) Endpoint versus re-sketched acted exteriority for the matched uncompensated control. (c) Target accuracy and achieved certificate distance versus requested margin. (d) Reference-predictor agreement, acted-label accuracy drop, and confidence drop. The dotted line marks $m = .18$; exact ablation values appear in Table 7.

Table 8: Stress-test summary for FEDGEAR and all thirteen common-interface adaptations across four participation fractions, four symmetric-label-noise rates, and four Dirichlet concentrations. "Worst" is the minimum within that stress family; wins and average rank use all twelve settings. Accuracies are percentages. Best and second-best values are bold and underlined.

Method	Part. mean ↑	Part. worst ↑	Noise mean ↑	Noise .30 ↑	Hetero. mean ↑	Hetero. worst ↑	Avg. rank ↓	Wins ↑
FedAvg	46.19	41.71	41.10	37.49	51.17	51.05	10.25	0
FedProx	46.21	41.76	41.14	37.53	51.23	51.05	9.17	0
FedBN	46.19	41.71	41.10	37.49	51.17	51.05	11.42	0
FedSR	46.31	41.88	41.34	37.77	51.35	51.17	7.42	0
MixStyle	46.29	41.34	42.58	39.85	51.74	51.50	6.08	0
DSU	45.34	40.65	41.29	38.28	51.48	51.22	8.83	0
FedFA	45.28	40.63	41.17	38.17	51.47	51.26	9.25	0
StableFDG	44.31	39.92	40.61	38.44	51.26	50.93	11.83	0
CSU	46.39	41.59	42.38	40.49	51.97	51.83	5.50	0
FedWon	**51.95**	**51.62**	<u>49.09</u>	<u>45.71</u>	<u>52.78</u>	<u>52.29</u>	<u>1.67</u>	<u>4</u>
FedOMG	46.03	41.86	41.22	38.07	51.12	50.72	9.83	0
FedLWS	46.80	42.81	41.95	38.71	51.36	51.17	5.92	0
FedSaC	47.66	45.04	43.13	39.51	51.16	50.97	6.50	0
FedGEAR	<u>51.90</u>	<u>51.39</u>	**49.46**	**46.33**	**53.24**	**52.53**	**1.33**	**8**

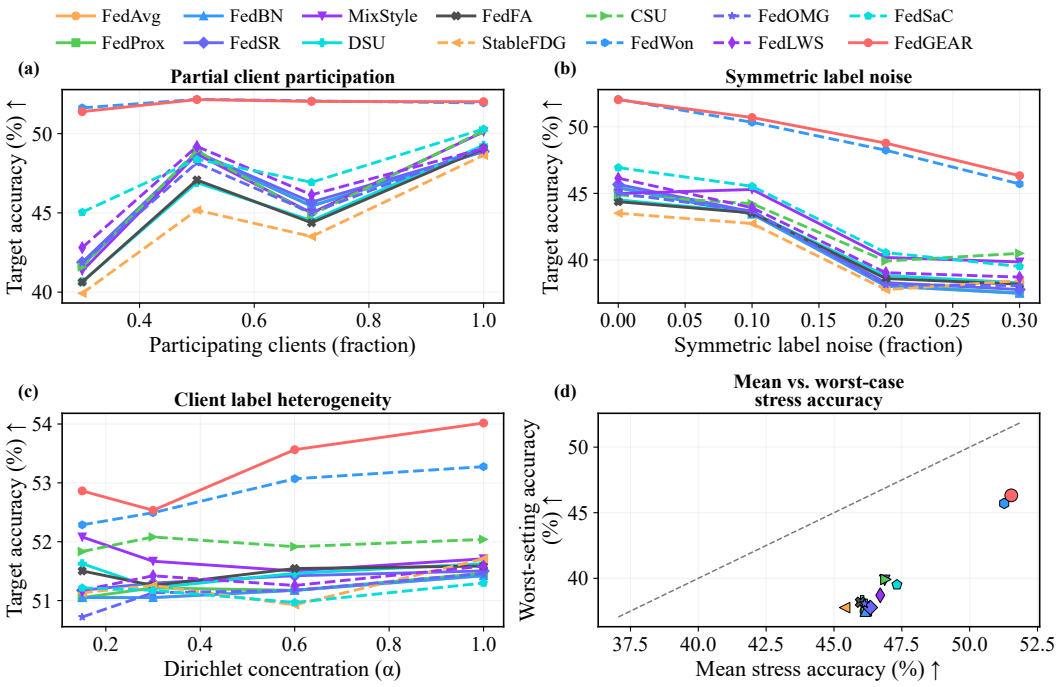


Figure 6: Robustness and failure analysis for FEDGEAR and all thirteen common-interface adaptations in a 2×2 layout. (a) Partial client participation, (b) symmetric label noise, and (c) Dirichlet client-label heterogeneity sweeps. (d) Mean versus worst-case stress accuracy across all twelve settings; the dashed identity line indicates the unattainable no-degradation boundary. Aggregate ranks and wins are reported in Table 8, avoiding long categorical y-axis labels.