NOMAD: FEDERATED GENERALIZED CATEGORY DISCOVERY ACROSS UNSEEN DOMAINS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalized Category Discovery (GCD) classifies known categories while discovering novel ones from unlabeled data. Practical deployments add two constraints rarely studied together: source data are distributed across institutions that cannot share raw samples, and the model must generalize to an unseen domain. Federated GCD rarely evaluates held-out-domain transfer, and domain-generalized GCD assumes centralized data. We formalize their intersection as **Federated Domain Generalization for Generalized Category Discovery** (FEDDG-GCD): decentralized clients train a model that must recognize known and discover novel categories in an unseen target domain without target-domain access. To our knowledge this setting is unexplored, and no evaluation protocol exists for it. We propose **NOMAD** (**N**on-IID **O**pen-world **M**odel for **A**daptive **D**iscovery), which couples a raw-data-free *Global Style Bank* for cross-client appearance diversification with *perturbation-stability aggregation* that replays the *same* perturbations to down-weight clients whose representations are unstable under them, rather than treating augmentation and client weighting as independent add-ons. *Adaptive Prototypical Alignment* anchors known-class representations under this shared perturbation family and separates confusable prototype pairs. A conditional target-risk decomposition formalizes this coupling, relating it to federated estimation error, residual style discrepancy, client heterogeneity, prototype error, and open-world routing and clustering. On PACS, Office-Home, and Mini-DomainNet under Dirichlet heterogeneity ($\beta \in \{0.2, 0.05\}$), **NOMAD** improves overall accuracy by **+2.5 to +4.0 points** over the strongest compared method in each setting, while remaining competitive on both old and novel classes.

1 INTRODUCTION

A deployed recognition system will inevitably encounter categories that were absent from its training data. In high-stakes settings such as medical diagnostics Zhou et al. (2024), autonomous perception Gupta et al. (2022), and large-scale surveillance Liao et al. (2014), neither silently misclassifying such samples nor merely flagging them as unknown is sufficient. **Generalized Category Discovery (GCD)** Vaze et al. (2022); Zhang et al. (2023c) formalizes this challenge: given limited supervision over *known* categories, a model must classify known instances while discovering coherent groups among *novel* examples in unlabeled data. Despite substantial progress Wen et al. (2023); Chiaroni et al. (2023); Kang et al. (2023), most GCD methods remain evaluated under assumptions that are difficult to satisfy in real multi-institutional deployments.

Two assumptions are particularly restrictive: *centralized training* and *distributional consistency*. Centralized training is infeasible when data are held by privacy-sensitive institutions, autonomous fleets, or distributed surveillance systems that cannot exchange raw samples, motivating **Federated Learning (FL)** Li et al. (2021); McMahan et al. (2017). Distributional consistency fails when a model trained across source sites is deployed at a new site with different hardware, acquisition protocols, environments, or population characteristics. These constraints are generally addressed separately. **Fed-GCD** Pu et al. (2024); Kim et al. (2025) enables collaborative discovery from decentralized data, but typically does not consider inductive transfer to a completely held-out target domain. Conversely, **DG-GCD** Rathore et al. (2025a;b) studies discovery under unseen-domain shift, but generally assumes centralized access to source data; related work such as HiLo Wang et al. (2024a) similarly relies on pooled source data. Figure 1 summarizes this gap: existing paradigms

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.

address pairs of federated learning, domain generalization, and open world discovery, but not their joint setting.

We formalize this intersection as **Federated Domain Generalization for Generalized Category Discovery** (FEDDG-GCD). During training, clients collaboratively learn a global shared encoder through federated aggregation without exchanging raw data. At test time, the global model is deployed at a new site whose visual domain was not seen during training. It must classify target samples from known categories and discover coherent groups among novel categories without target-domain adaptation. This joint requirement distinguishes FEDDG-GCD from standard federated GCD, federated domain generalization, and centralized DG-GCD settings.

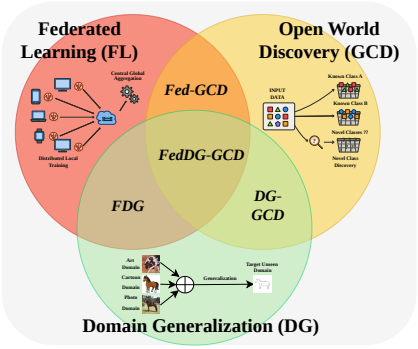

Figure 1: **Comparison of learning settings.** FEDDG-GCD combines decentralized source training, generalization to an unseen target domain, and known–novel category discovery.

The setting is challenging because its sources of error interact. Non-IID client data can fragment the shared representation McMahan et al. (2017); Li et al. (2021); appearance bias can entangle class semantics with source-specific styles Zhou et al. (2021b); Chen et al. (2021; 2023); and strong prototype alignment can stabilize known classes while making ambiguous samples more likely to be assigned to known prototypes. We capture these interactions through a target-risk decomposition (Sec. A) involving residual *style discrepancy*, *client representation heterogeneity*, *prototype drift*, and *novel-cluster separation*. This decomposition motivates a method that jointly addresses appearance variation, cross-client semantic consistency, and open-world geometry.

To address FEDDG-GCD, we propose **NOMAD** (**N**on-IID **O**pen-world **M**odel for **A**daptive **D**iscovery). Rather than pairing an off-the-shelf style augmenter with an independent client-weighting rule, **NOMAD** routes both through one Style-Bank perturbation family. A raw-data-free *Global Style Bank* Chen et al. (2021); Zhong et al. (2022) aggregates channel-wise feature statistics across clients to diversify local appearance without transmitting raw images. *Adaptive Prototypical Alignment* anchors known-class representations to shared prototypes under this perturbation family and penalizes confusable prototype pairs. *Perturbation-stability aggregation* replays the same perturbations to measure saliency consistency, down-weighting clients whose representations are unstable under the appearance shift the model is trained to resist. This shared family links local diversification and server-side aggregation without requiring access to the target. **In summary,**

- We formulate FEDDG-GCD, combining federated source training, open-world category discovery, and generalization to a target domain unavailable during training. The target task requires known-class recognition and novel-category discovery under domain shift.
- We propose **NOMAD**, which routes style diversification and client-reliability weighting through one shared perturbation family rather than designing them independently. A conditional target-risk decomposition formalizes this coupling, separating weighted federated estimation, residual domain and client shift, prototype error, and open-world routing and clustering terms.
- We investigate our experiments on three benchmarks i.e. PACS Li et al. (2017), Office-Home Venkateswara et al. (2017), and Mini-DomainNet Zhou et al. (2021a), under moderate (β=0.2) and extreme (β=0.05) non-IID regimes,[1] and **NOMAD** consistently improves over the evaluated FL, Fed-GCD, and DG-GCD baselines, achieving the highest overall accuracy across all dataset and heterogeneity settings.

2 RELATED WORK

GCD under domain shift. GCD recognizes known categories while clustering novel examples from unlabeled data under limited supervision Vaze et al. (2022); Pu et al. (2023). Most early methods Zhang et al. (2023c); Wen et al. (2023); Chiaroni et al. (2023); Kang et al. (2023); Choi

[1]The concentration parameter β of the Dirichlet distribution controls label heterogeneity across clients.

et al. (2024); Wang et al. (2024b); Hao et al. (2023); Liu & Han (2025); He et al. (2026) assume a shared train–test distribution. Cross-domain discovery methods Rongali et al. (2024); Wang et al. (2024d); Wen & Brbic (2024); Wang et al. (2024a); Feng & Ge (2025), including frequency-domain approaches such as FREE Feng & Ge (2025), use unlabeled target data during training and are therefore transductive. DG-GCD methods, including DG^{2}CD-Net Rathore et al. (2025a) and HIDISC Rathore et al. (2025b), instead generalize to unseen domains without target-domain access, but assume centralized source data.

Federated GCD. Fed-GCD enables category discovery when data remain decentralized. FedoSSL Zhang et al. (2023a) distinguishes locally and globally unseen classes, whereas AGCL Pu et al. (2024) and FedLPA Kim et al. (2025) address heterogeneous federated discovery through contrastive or graph-based alignment. These approaches do not explicitly address inductive generalization to a target domain absent from federated training.

Federated domain generalization. FDG trains from decentralized sources and evaluates on unseen target domains. Existing methods pursue invariant representation learning (e.g., FedADG Zhang et al. (2021), FedSR Nguyen et al. (2022), FedIIR Guo et al. (2023)) or domain/style diversification (e.g., CCST Chen et al. (2023), FedAlign Gupta et al. (2025), FedCCRL Wang et al. (2024c), and FedGA Zhang et al. (2023b)). They generally assume a fixed closed-set label space and therefore do not model novel categories at test time. We compare against FedADG, FedAlign, and FedGA; FedIIR and FedCCRL require additional assumptions incompatible with our decentralized label-heterogeneous protocol.

3 PROBLEM FORMULATION

We consider a federated learning system with K decentralized clients and M source domains $\mathcal{D}_S = \{d_1, d_2, \ldots, d_M\}$, where $K \geq M$. Each client $k \in \{1, \ldots, K\}$ is associated with a single source domain $d_{\phi(k)} \in \mathcal{D}_S$ through a surjective mapping $\phi : \{1, \ldots, K\} \to \{1, \ldots, M\}$, and holds a private dataset $\mathcal{D}_k = \{(x_i^{(k)}, y_i^{(k)})\}_{i=1}^{N_k}$ sampled from $P_{d_{\phi(k)}}(X, Y)$. Clients communicate model updates and lightweight summary statistics, but never share raw samples.

The global label space is partitioned into disjoint known and novel sets, $\mathcal{Y} = \mathcal{Y}_{\text{known}} \cup \mathcal{Y}_{\text{novel}}$, with $\mathcal{Y}_{\text{known}} \cap \mathcal{Y}_{\text{novel}} = \emptyset$. During training, clients observe labels only from $\mathcal{Y}_{\text{known}}$. Label heterogeneity is induced by sampling client class priors $\mathbf{p}_k \sim \text{Dir}(\beta \mathbf{1})$, where a smaller β corresponds to stronger non-IID label skew.

After training, the global model is evaluated on an unlabeled target dataset $\mathcal{D}_{\text{test}}$ from a held-out domain $d_{\text{test}} \notin \mathcal{D}_S$. The target differs from the observed source domains in its class-conditional appearance statistics:

$$P_{d_{\text{test}}}(X \mid Y = y) \neq P_{d_{\phi(k)}}(X \mid Y = y), \qquad \forall y,\ k. \tag{1}$$

The target label space spans $\mathcal{Y}_{\text{known}} \cup \mathcal{Y}_{\text{novel}}$. Consequently, inference must jointly handle domain shift and an open label space: known-class samples should remain identifiable with respect to $\mathcal{Y}_{\text{known}}$, whereas samples from $\mathcal{Y}_{\text{novel}}$ should form semantically coherent groups. The objective is to learn, from decentralized source data alone, a shared encoder $f_\theta : \mathcal{X} \to \mathbb{R}^d$ that supports this inductive generalized-category-discovery objective on the unseen target domain.

4 NOMAD: METHOD AND DESIGN

We propose **NOMAD**, a federated framework for FEDDG-GCD built around a common perturbation family shared by local diversification and global aggregation. The setting contains four coupled difficulties: unseen target-domain appearance, client-induced representation fragmentation, inconsistent known-class anchors across clients, and insufficient separation for novel groups at deployment. These difficulties interact. For example, indiscriminate appearance perturbation may destabilize well-formed semantic anchors, while strong prototype alignment stabilizes known-class geometry but widens the region that the known-class decision rule accepts, making ambiguous target samples more likely to be absorbed as known. **NOMAD** therefore combines cross-client style diversification, prototype alignment, and stability-aware aggregation in a single training procedure.

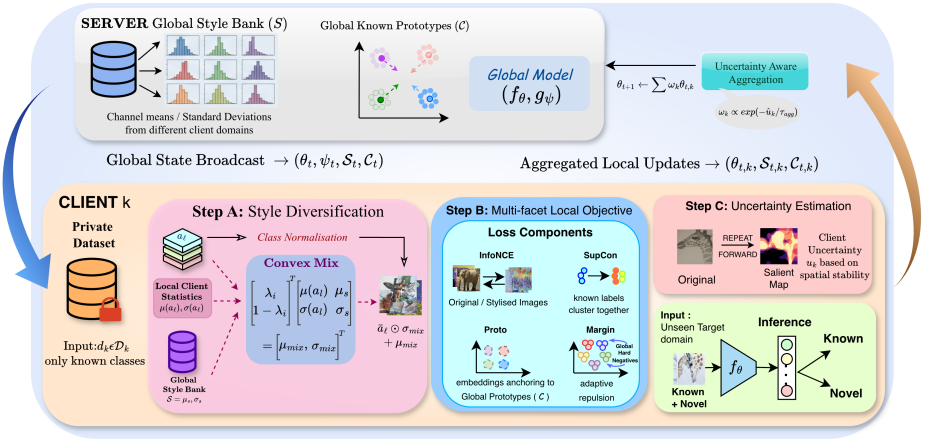

Figure 2: **Overview of the NOMAD framework. NOMAD** integrates cross-client style diversification through a Global Style Bank (Sec. 4.1), adaptive prototypical alignment with a margin loss (Sec. 4.2), and perturbation-stability-weighted aggregation (Sec. 4.3) for generalized category discovery on unseen target domains.

We describe each component in Secs. 4.1–4.4, and relate them to the target-risk decomposition in Sec. 4.5 and Appendix A.

As shown in Fig. 2, the server maintains an encoder f_θ, a training-time projection head g_ψ that is discarded at inference, a Global Style Bank $\mathcal{S} = \{(\mu_s, \sigma_s)\}_{s=1}^{|\mathcal{S}|}$, and EMA-updated known-class prototypes $\mathcal{C} = \{c_y\}_{y \in \mathcal{Y}_{\text{known}}}$. At each round, clients receive $(\theta, \psi, \mathcal{S}, \mathcal{C})$, conduct style-diversified local training, and return their model updates together with the statistics required to update the global state. The server aggregates local models using perturbation stability and refreshes $\mathcal{S}$ and $\mathcal{C}$ via EMA. Hence, the bank defines a common family of cross-client appearance perturbations for both representation learning and client reliability estimation.

4.1 Raw-Data-Free Style Diversification via Global Style Bank

Since client k is associated with a source domain $d_{\phi(k)}$, local optimization can entangle class semantics with domain-specific appearance cues. MixStyle Zhou et al. (2021b) partially alleviates this issue within a client, but cannot expose a client to appearance modes observed only at other sites. We therefore introduce a *Global Style Bank* $\mathcal{S}$, a server-side repository of channel-wise feature statistics aggregated across clients. Channel-wise first- and second-order statistics (μ, σ) encode coarse appearance factors such as texture, illumination, contrast, and sensor characteristics while retaining the spatial layout of the feature map Chen et al. (2021); Zhong et al. (2022). The bank does not transmit raw images and is compatible with the data-locality constraint of federated learning. It does not, by itself, provide a formal privacy guarantee.

For an intermediate feature map $a_\ell(x) \in \mathbb{R}^{C \times H \times W}$, let $\mu(a_\ell(x)) \in \mathbb{R}^C$ and $\sigma(a_\ell(x)) \in \mathbb{R}^C$ denote its channel-wise mean and standard deviation, and let

$$\bar{a}_\ell(x) = \frac{a_\ell(x) - \mu(a_\ell(x))}{\sigma(a_\ell(x)) + \epsilon} \tag{2}$$

be the channel-normalized map. We synthesize a style-diversified view by re-injecting a convex mixture of local and banked statistics:

$$\tilde{a}_\ell(x) = \bar{a}_\ell(x) \odot \sigma_{\text{mix}} + \mu_{\text{mix}}, \quad \mu_{\text{mix}} = \lambda_i \mu(a_\ell(x)) + (1-\lambda_i)\mu_s, \quad \sigma_{\text{mix}} = \lambda_i \sigma(a_\ell(x)) + (1-\lambda_i)\sigma_s, \tag{3}$$

where $(\mu_s, \sigma_s) \sim \mathcal{S}$ and $\lambda_i \in [0, 1]$ is a sample-adaptive mixing coefficient. This operation broadens the support of the source style distribution $\mathcal{T}_\# \bar{P}_\omega$, where $\mathcal{T}(x) = (\mu(a_\ell(x)), \sigma(a_\ell(x)))$ and $\bar{P}_\omega = \sum_k \omega_k P_{d_{\phi(k)}}$ is the stability-weighted source mixture, with ω_k defined in Eq. 7. When the bank-induced style support contains appearance variations closer to a held-out target than those seen by an individual client, it can reduce the residual style discrepancy term in Eq. 9. We empirically visualize this effect in Appendix B, Fig. 6.

The mixing coefficient is sample-adaptive. Let $q_i = \max_{y \in \mathcal{Y}_{\text{known}}} p(y \mid f_\theta(x_i^{(k)}); \mathcal{C})$ be the maximum prototype affinity of $x_i^{(k)}$. We set $\lambda_i = q_i^\rho$ with $\rho \geq 1$ ($\rho = 2$ in all experiments; Appendix H). Computing q_i requires an initial unperturbed pass through f_θ before the style-mixed pass of Eq. 3; both passes are included in the per-round compute budget of Appendix D. High-confidence samples retain more of their local appearance, whereas low-confidence samples receive stronger cross-client style replacement. This focuses style diversification on samples whose representations are less confidently supported by the shared known-class prototype geometry.

4.2 Adaptive Prototypical Alignment for Open-World Geometry

Style diversification alone does not resolve cross-client semantic drift. Under non-IID partitions induced by $\mathbf{p}_k \sim \text{Dir}(\beta \mathbf{1})$, embeddings of the same known class can occupy inconsistent regions across clients. Strongly aligning representations with global prototypes stabilizes this geometry, but also widens the region the known-class rule accepts, making ambiguous target samples more likely to be absorbed as known. Each client therefore optimizes

$$\mathcal{L}_{\text{local}}^{(k)} = \lambda_{\text{nce}} \mathcal{L}_{\text{InfoNCE}} + \lambda_{\text{sup}} \mathcal{L}_{\text{SupCon}} + \lambda_{\text{proto}} \mathcal{L}_{\text{Proto}} + \lambda_{\text{mar}} \mathcal{L}_{\text{Margin}}, \tag{4}$$

with $\lambda_{\text{nce}} = 1.0$, $\lambda_{\text{sup}} = 0.35$, $\lambda_{\text{proto}} = 0.5$, and $\lambda_{\text{mar}} = 0.25$. $\mathcal{L}_{\text{InfoNCE}}$ and $\mathcal{L}_{\text{SupCon}}$ operate on the projection $z = g_\psi(f_\theta(x)) \in \mathbb{R}^{768}$, whereas $\mathcal{L}_{\text{Proto}}$ operates directly on the encoder representation used at inference; see Appendix F. The first three terms pull representations together but do not control how close distinct known-class anchors become. We therefore add an adaptive margin that responds to local class confusion. Let $m_{k,y} = |\mathcal{B}_k^y|^{-1} \sum_{x \in \mathcal{B}_k^y} f_\theta(x)$ be the *batch-local* prototype of known class y, computed from the known-class samples $\mathcal{B}_k^y \subseteq \mathcal{D}_k^y$ present in the current mini-batch at client k, so that gradients flow through it at every step – distinct from the per-round class means $\mathcal{P}_k$ (Algorithm 1) used only to refresh $\mathcal{C}$ via EMA (Sec. 4.3), which are detached statistics. Let $s_{y,y'} = \cos(m_{k,y}, c_{y'})$ be its cosine similarity to a competing global prototype $c_{y'} \in \mathcal{C}$. We define

$$\delta_y = \max\{0,\ \delta_0 + \alpha \max_{y' \neq y} s_{y,y'}\}, \qquad \delta_y \in [0,\ \delta_0 + \alpha], \tag{5}$$

where $\delta_0 = 0.1$ is a base margin and $\alpha = 0.5$ controls the response to the hardest competing prototype. The required separation is therefore capped at $\delta_0 + \alpha$, so the term never demands near-orthogonality between known-class anchors. The outer clamp is a safeguard rather than an operating regime: it activates only when every competitor already satisfies $s_{y,y'} < -\delta_0/\alpha$, in which case class y is well separated and no margin is needed. We minimize

$$\mathcal{L}_{\text{Margin}} = \frac{1}{|\mathcal{Y}_{\text{local}}|} \sum_{y \in \mathcal{Y}_{\text{local}}} \log\left(1 + \sum_{y' \neq y} \exp\left(\frac{s_{y,y'} - (1 - \delta_y)}{\tau_m}\right)\right), \tag{6}$$

where $\mathcal{Y}_{\text{local}}$ is the set of known classes present in the current mini-batch at client k and τ_m is a temperature. The unit term inside the logarithm keeps the objective non-negative and lets it saturate near zero once every competitor falls below $1 - \delta_y$, so the loss rewards adequate separation rather than unbounded repulsion. A large similarity to a competing prototype raises δ_y and thus the penalty on that pair, emphasizing confusable known-class relations without explicit hard-negative mining, and supporting a more reliable known-versus-rejected decision at inference. Sensitivity to δ_0, α, and the loss weights is reported in Appendix H.

4.3 Perturbation-Stability Weighted Aggregation

FedAvg McMahan et al. (2017) weights clients according to local dataset size N_k. Under FedDG-GCD, however, a large client can still produce an update that is sensitive to domain-specific appearance cues. We therefore estimate the stability of each locally trained model under Style-Bank perturbations, the same perturbation family used in Eq. 3 during local training. REPEAT Wickstrøm et al. (2025) is designed to diagnose a single trained model's representational uncertainty; we repurpose its saliency-stability signal as a *federated* client-reliability score, computed under the very Style-Bank perturbations $\mathcal{L}_{\text{local}}^{(k)}$ is simultaneously trained to resist (Fig. 3), rather than as an independent post-hoc diagnostic. Let $p_{ij}^{(k)} \in [0, 1]$ denote the empirical probability that spatial location (i, j) is salient for client k over repeated Style-Bank perturbations; details are provided in Appendix G.

The per-client uncertainty and aggregation weight are

$$u_k = \frac{1}{HW} \sum_{i,j} p_{ij}^{(k)} \left(1 - p_{ij}^{(k)}\right), \qquad \omega_k = \frac{\exp(-\hat{u}_k/\tau_{\text{agg}})}{\sum_{k'} \exp(-\hat{u}_{k'}/\tau_{\text{agg}})}, \tag{7}$$

where $\hat{u}_k = (u_k - \min_{k'} u_{k'})/(\max_{k'} u_{k'} - \min_{k'} u_{k'} + \epsilon)$ is the min–max normalization of $\{u_k\}$ over the clients active in the round, and $\tau_{\text{agg}} = 0.5$ controls the selectivity of the weighting, therefore limiting the influence of updates whose representations shift most under cross-client appearance variation.

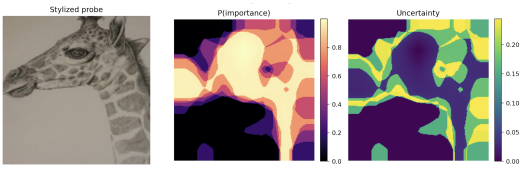

Figure 3: REPEAT Wickstrøm et al. (2025) estimates per-client aggregation weights from spatial importance and uncertainty maps under a stylized probe. The importance map identifies attended regions, while the uncertainty map captures their variability across perturbations.

This closes the training loop: $\mathcal{S}$ defines the perturbation family, $\mathcal{L}_{\text{local}}^{(k)}$ promotes invariance to it, and $\{\omega_k\}$ favor clients with stable responses to it. The server then updates (θ, ψ) through reliability-weighted aggregation and refreshes $\mathcal{S}$ and $\mathcal{C}$ through EMA. The shared perturbation family does not presume access to the target domain; it provides a common mechanism for learning and evaluating robustness to style variations available across the federation.

The full training algorithm is given in Appendix D.

The Global Style Bank sets the perturbation family, perturbation-stability aggregation scores clients by *it, and Adaptive Prototypical Alignment anchors geometry* in *it – remove any one, and the others lose the shared signal they depend on.*

4.4 Inference without Target Adaptation

At inference, g_ψ is discarded and f_θ operates directly in the prototype space defined by $\mathcal{C} \subset \mathbb{R}^d$ – the same encoder-output space in which $\mathcal{L}_{\text{Proto}}$ and $\mathcal{L}_{\text{Margin}}$ were optimized (Sec. 4.2). For each target sample $x_j^{(t)} \in \mathcal{D}_{\text{test}}$, we compute $r_j^{(t)} = f_\theta(x_j^{(t)})$ and evaluate its maximum affinity to the known-class prototypes. The sample is assigned to $\mathcal{Y}_{\text{known}}$ if

$$\max_{y \in \mathcal{Y}_{\text{known}}} p(y \mid r_j^{(t)}; \mathcal{C}) \geq \tau_{\text{inf}}, \tag{8}$$

where τ_{inf} is selected on source-held-out data by the protocol in Appendix I; the target pool is never used for tuning. Samples not accepted by a known-class prototype are clustered into $\widehat{K}_{\text{novel}}$ groups by k-means in the f_θ embedding space, where $\widehat{K}_{\text{novel}}$ is estimated from the rejection set itself (Appendix E.2). No gradient-based target adaptation is performed.

$\mathcal{L}_{\text{Proto}}$ makes known-class regions identifiable through $\mathcal{C}$, while $\mathcal{L}_{\text{Margin}}$ promotes separation among competing known-class anchors. Under the standard clusterability condition that rejected novel samples form compact groups separated from one another and from the known-class prototype regions, clustering the rejection set produces coherent novel-category partitions.

4.5 Theoretical Grounding: A Conditional Target-Risk Decomposition

We provide a conditional decomposition of target generalized-category-discovery risk for the setting with no target-domain access during training. Let $\bar{P}_\omega = \sum_{k=1}^K \omega_k P_{d_{\phi(k)}}$ denote the stability-weighted source mixture, where $\{\omega_k\}_{k=1}^K$ are fixed aggregation weights for the analysis. Let $\bar{P}_\omega^S$ denote its Style-Bank-augmented counterpart, and let $\widehat{R}_\omega^{S,\text{rep}}$ be the corresponding weighted empirical representation risk on known-class source samples. Under bounded-loss and finite-complexity conditions, a bank-supported representation-transfer assumption, prototype concentration, and a tar-

get rejection–clusterability condition, with probability at least $1-\delta$,

$$R^{\mathrm{GCD}}_{\mathrm{test}}(f_\theta) \le \widehat{R}^{S,\mathrm{rep}}_{\omega}(f_\theta) + \mathcal{O}\left(\mathfrak{R}_{\omega,S}\big(\ell_{\mathrm{rep}} \circ \mathcal{F}\big) + \sqrt{\frac{\log(1/\delta)}{n_{\mathrm{eff}}}}\right) + \lambda_s \Delta_{\mathrm{style}}(f_\theta) + \lambda_h \Gamma_{\mathrm{het}}(f_\theta) + \lambda_p \left(\bar{\epsilon}_{\mathrm{proto}}(f_\theta)^{1/2} + \Psi_{\mathrm{proto}}(f_\theta)^{1/2}\right) + R_{\mathrm{rej}}(f_\theta) + R_{\mathrm{clust}}(f_\theta), \tag{9}$$

$$n_{\mathrm{eff}} = \left(\sum_{k=1}^{K} \frac{\omega_k^2}{N_k}\right)^{-1}. \tag{10}$$

Here, $R_{\mathrm{rej}} = R^{\mathrm{novel}\to\mathrm{known}}_{\mathrm{test}} + R^{\mathrm{known}\to\mathrm{reject}}_{\mathrm{test}}$ collects known–novel routing errors, while R_{clust} denotes the conditional clustering error on correctly rejected novel target samples.

The first line bounds augmented weighted source representation risk and its federated estimation error. The second line captures residual target-style mismatch and cross-client representation heterogeneity. The third line accounts for the error introduced when known-class decisions are formed with global prototypes, together with the open-world routing and clustering errors. The result is conditional: the target-style and clusterability terms involve the unseen target distribution and are analysis quantities rather than training-time observable objectives. The Global Style Bank, Adaptive Prototypical Alignment, and stability-aware aggregation are designed to reduce source-side proxies for the respective terms. Appendix A gives the complete definitions, assumptions, and proof.

5 Experimental Evaluations

5.1 Setting up the Experiments

(a) Datasets and Protocol. We evaluate on PACS Li et al. (2017), Office-Home Venkateswara et al. (2017), and Mini-DomainNet Zhou et al. (2021a) under a **Leave-One-Domain-Out (LODO)** protocol. Following *Fed-GCD* Pu et al. (2024), source data is partitioned across K=5 clients via $\mathrm{Dir}(\beta\mathbf{1})$ at two levels: Normally Heterogeneous (NH, β=0.2) and Extremely Heterogeneous (EH, β=0.05). The known/novel class split ranges from 4/3 on PACS, 33/32 on Office Home, and 64/62 on Mini-DomainNet. Full dataset statistics are in the **Appendix** C.

(b) Evaluation Metrics. All methods follow the inference procedure of Sec. 4.4, with no target adaptation or knowledge of the target label-space size. Accepted samples carry their routed known label; clusters on the rejection set are matched to ground-truth categories by a Hungarian permutation computed once over that set. *Old*, *New*, and *All* are accuracies over the known-class samples, the novel-class samples, and the full pool, and therefore include routing errors in both directions. Results are averaged over three seeds and all leave-one-domain-out splits; baseline adaptations are in Appendix E.1.

(c) Implementation Details. We use a DINO-v1-pretrained Caron et al. (2021) ViT-B/16 Dosovitskiy et al. (2021), fine-tuning only the final transformer block Vaze et al. (2022); Rathore et al. (2025a;b). The projection head g_ψ is a 3-layer MLP with 768-dimensional output. Training follows a two-stage curriculum: an *Alignment Phase* of T_{align}=5 rounds using standard FedAvg at η=$1e{-}3$, followed by a 45-round *Discovery Phase* with uncertainty-weighted aggregation and cosine-annealed learning rate ($\eta_{\max}$=0.01). Local training runs for E=2 epoch per round with τ_{agg}=0.5. The Global Style Bank maintains $|\mathcal{S}| = K$ unique entries, one per client. The hyperparameter details are provided in the **Appendix** H.

5.2 Comparison to the Literature

5.2.1 Quantitative Comparison

Table 1 reports *All*/*Old*/*New* target accuracy under moderate (NH, β=0.2) and extreme (EH, β=0.05) heterogeneity.

Overall accuracy. NOMAD achieves the highest *All* accuracy in every dataset–heterogeneity setting. On Office-Home NH it improves over FedADG from 55.12% to 58.92% (**+3.80** points), and on

Table 1: Performance under **moderate** ($\beta = 0.2$, NH) and **extreme** ($\beta = 0.05$, EH) label heterogeneity. Accuracy (%) is reported for Old and New classes in the target domain as mean ± std. over three seeds. **Bold**: best; underline: second-best.

	PACS						Office-Home						Mini-DomainNet					
	NH (β=0.2)			EH (β=0.05)			NH (β=0.2)			EH (β=0.05)			NH (β=0.2)			EH (β=0.05)		
Methods	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
GCD Baselines																		
ViT [ICLR'21]	58.90	57.33	62.20	54.55	53.10	55.66	51.42	59.02	42.73	54.06	62.34	44.50	47.48	50.56	44.47	47.58	49.94	45.29
GCD [CVPR'22]	66.58	66.60	68.08	68.65	64.72	72.62	53.12	61.17	43.94	52.91	60.57	44.19	52.28	57.95	46.71	51.73	59.65	43.96
SimGCD [ICCV'23]	48.59	72.96	18.86	48.57	72.03	20.71	45.60	69.17	18.81	45.38	66.87	20.89	49.88	63.83	36.27	49.46	64.07	35.16
PCL [ICLR'21]	25.77	23.31	27.44	25.61	24.70	25.04	8.66	8.84	8.48	8.91	8.93	8.90	21.72	22.56	20.9	21.72	22.97	20.5
GPC [ICCV'23]	63.89	63.33	65.49	65.12	63.31	67.35	52.89	58.72	46.19	52.40	57.71	46.30	47.75	51.18	44.39	47.72	51.38	44.15
PAL-GCD [AAAI'25]	51.05	70.72	26.99	53.45	72.05	31.54	47.78	**70.08**	22.33	47.05	**68.07**	23.07	45.49	**64.53**	26.73	45.23	**65.67**	25.09
Fed-GCD Baselines																		
FedLPA [NeurIPS'25]	55.69	72.04	36.22	61.99	75.83	42.00	50.25	62.30	36.38	50.38	61.81	37.27	34.68	45.32	24.23	32.87	44.67	21.22
FedAvg+AGCL [CVPR'24]	66.18	63.59	70.58	67.16	62.82	71.83	51.44	56.07	46.07	53.12	59.44	45.66	31.38	42.87	20.16	29.56	38.66	20.66
FedProx+AGCL [CVPR'24]	56.91	62.53	50.08	62.78	66.07	58.72	50.75	55.55	45.45	51.08	56.10	45.37	29.16	37.77	20.71	29.93	39.72	20.36
FDG Baselines																		
FedAlign [CVPRW'25]	50.76	62.94	34.17	46.94	51.07	43.68	50.62	63.87	35.29	50.65	63.91	35.56	42.82	51.12	34.66	43.84	52.83	34.93
FedADG [Arxiv'21]	73.64	73.86	73.41	72.61	72.98	72.18	55.12	59.97	49.28	55.29	61.66	47.85	50.09	53.15	47.06	50.80	54.36	47.30
GA [CVPR'23]	58.68	63.73	52.45	52.16	63.63	40.62	55.11	60.07	49.66	54.45	59.79	48.19	47.98	50.09	45.91	45.77	47.82	43.76
DG-GCD Baselines																		
DG^{2}CD-Net [CVPR'25]	73.34	73.55	73.12	72.67	72.68	72.59	54.01	57.16	50.19	52.66	56.13	48.52	46.79	49.62	44.07	46.14	48.15	44.18
HiDISC [NeurIPS'25]	74.85	73.65	76.80	73.06	71.58	75.22	52.57	56.76	47.75	50.73	54.29	46.61	47.45	49.32	45.33	44.29	48.86	39.78
NOMAD (Ours)	**77.97** ±0.08	**77.74** ±0.03	**77.35** ±0.15	**77.09** ±0.09	**76.78** ±0.11	**77.81** ±0.06	**58.92** ±0.50	66.16 ±0.66	**50.50** ±0.55	**58.52** ±0.60	67.31 ±0.03	**48.54** ±0.29	**54.79** ±0.11	62.32 ±0.96	**47.36** ±0.11	**54.94** ±0.49	62.48 ±0.27	**47.63** ±0.54
Δ	**+3.12**	**+3.88**	**+0.55**	**+4.03**	**+0.95**	**+2.59**	**+3.80**	**-3.92**	**+0.31**	**+3.23**	**-0.76**	**+0.02**	**+2.51**	**-2.21**	**+0.30**	**+3.21**	**-3.19**	**+0.33**

Mini-DomainNet EH over GCD from 51.73% to 54.94% (**+3.21** points). The largest gain is **+4.03** points on PACS EH (HiDISC, 73.06% to 77.09%), and gains are consistently larger under extreme than moderate heterogeneity on PACS and Mini-DomainNet.

Known–novel trade-off. On Office-Home NH, PAL-GCD reaches 70.08% *Old* but only 22.33% *New*. **NOMAD** obtains 66.16% *Old* and 50.50% *New*, trading 3.92 points of known-class accuracy for **+28.17** points on novel categories. The same pattern holds on Mini-DomainNet, where the strongest *Old* baselines are those that assign almost nothing to novel classes. This is consistent with a representation balanced between known recognition and rejected-sample discovery rather than optimized for either alone.

Efficiency. On Office-Home, **NOMAD** requires 15.81 PFLOPs, versus 16.53 for HiDISC and 33.06 for GCD under our common counting procedure, showing that the proposed components do not dominate training cost. Extended results and cluster-cardinality analysis: Appendices E.2 & E.3.

5.2.2 Qualitative Comparison

We visualize t-SNE Van der Maaten & Hinton (2008) embeddings of held-out PACS **Sketch** samples after training on the remaining source domains (Fig. 4). We sample $N = 40$ images per class with perplexity 30. **NOMAD** shows more compact local neighborhoods and less visible category overlap than the displayed baselines, consistent with its higher silhouette score computed in the original embedding space. Because t-SNE can distort global distances, we treat this visualization as qualitative support for the quantitative results.

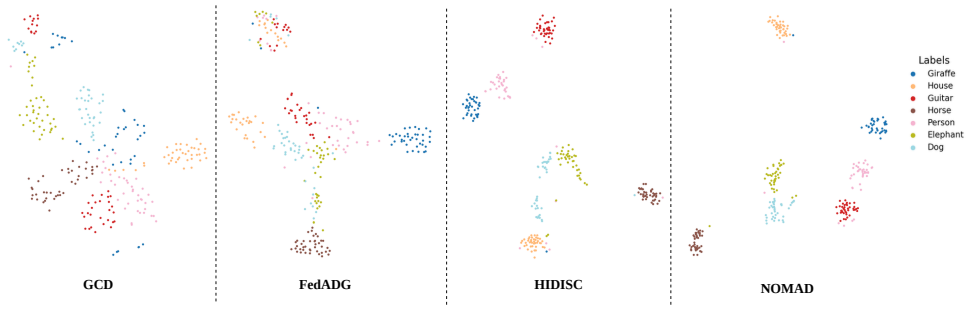

Figure 4: **Target-domain t-SNE visualization.** Embeddings on the held-out PACS Sketch domain for GCD, FedADG, HiDISC, and **NOMAD**. Silhouette scores Rousseeuw (1987) are 0.128, 0.140, 0.151, and **0.295**, respectively.

6 Ablation Analysis

Component efficacy and style diversification. Table 2 evaluates the contributions of the local objective and the Global Style Bank on PACS. All configurations C-0–C-5 and **NOMAD** train on Global-Style-Bank-diversified views (Sec. 4.1), so $\mathcal{L}_{\text{InfoNCE}}$ is well-defined wherever it is enabled; the style *source* itself is ablated separately in the right panel. Among the individual additions, prototypical alignment provides the largest improvement after the contrastive baseline: C-3→C-4 improves NH *All* by **+2.78** points. Combining all terms yields **77.97%** NH and **77.09%** EH *All* accuracy, and improves NH *New* by **+4.93** points over C-3.

The style ablation shows that the Global Style Bank is more effective than local MixStyle in this setting: NOMAD-Full improves NH *All* from 72.68% to **77.97%** (**+5.29** points). Adaptive mixing $\lambda_i = q_i^\rho$ also outperforms fixed mixing ($\lambda = 0.5$: 75.12%) and random mixing (74.60%). In contrast, aggressive fixed mixing ($\lambda = 0.1$) lowers *Old* accuracy to 65.20%, indicating that uniform strong perturbation can disrupt known-class recognition.

Table 2: **Ablation studies on PACS. Left:** Contributions of the NOMAD loss components under NH ($\beta = 0.2$) and EH ($\beta = 0.05$); all rows train on Style-Bank-diversified views (Sec. 4.1), isolating the loss terms rather than the augmentation source. **Right:** Effects of style source and mixing strategy under NH. All configs. use the complete NOMAD objective.

Cfg.	$\mathcal{L}_{\text{SupCon}}$	$\mathcal{L}_{\text{InfoNCE}}$	$\mathcal{L}_{\text{Proto}}$	$\mathcal{L}_{\text{Margin}}$	NH All	NH Old	NH New	EH All	EH Old	EH New
C-0	✗	✗	✗	✗	58.90	57.33	62.20	54.55	53.10	55.66
C-1	✓	✗	✗	✗	70.10	69.53	70.21	67.32	66.62	68.24
C-2	✗	✓	✗	✗	67.13	64.14	70.62	66.05	66.42	65.06
C-3	✓	✓	✗	✗	71.37	71.02	72.42	68.73	71.82	64.70
C-4	✓	✓	✓	✗	74.15	71.62	76.60	69.48	72.36	67.42
C-5	✓	✓	✗	✓	72.45	72.62	71.77	70.12	71.90	68.72
NOMAD	✓	✓	✓	✓	**77.97**	**77.74**	**77.35**	**77.09**	**76.78**	**77.81**

Configuration	Style Source	Mixing Strategy	All	Old	New
No Style Aug.	–	Standard views	71.53	70.10	71.50
Gaussian Aug	Input-level	Fixed (σ)	72.45	72.18	73.10
Local MixStyle	Client-local	Fixed ($\lambda = 0.5$)	72.68	72.30	73.42
NOMAD-Aggressive	Global Bank	Fixed ($\lambda = 0.1$)	68.42	65.20	74.88
NOMAD-Fixed	Global Bank	Fixed ($\lambda = 0.5$)	75.12	74.88	75.40
NOMAD-Random	Global Bank	Beta(α, α)	74.60	73.55	76.12
NOMAD-Full	**Global Bank**	**Adaptive (q_i^ρ)**	**77.97**	**77.74**	**77.35**

Sensitivity and stability. Figure 5 summarizes sensitivity on PACS; panels (a), (b), and (d) use NH unless noted otherwise. The alignment phase is important: setting $T_{\text{align}} = 0$ decreases *New* accuracy by **5.25** points. Perturbation-stability aggregation outperforms FedAvg McMahan et al. (2017) and FedProx Li et al. (2020b) under both NH and EH, and is competitive with FedNOVA Wang et al. (2020); further comparisons to uncertainty-aware aggregation methods are provided in Appendix J.2.

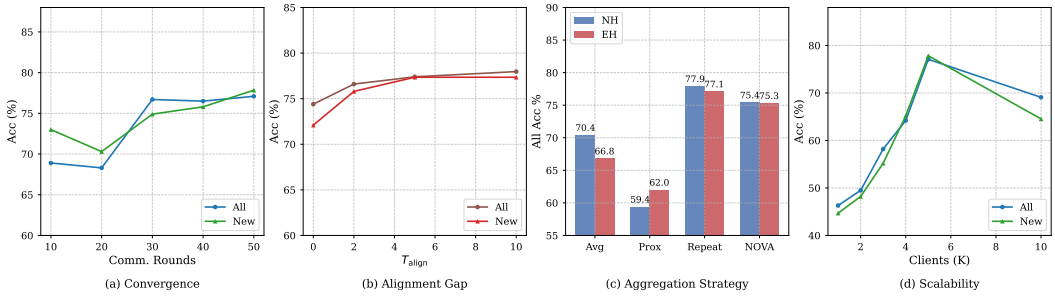

Figure 5: **Sensitivity and stability.** (a) Target accuracy over communication rounds for *All* and *New*. (b) Effect of alignment warmup length T_{align}. (c) Aggregation comparison under NH and EH. (d) Accuracy as the number of clients K increases.

7 Takeaways

We introduce FedDG-GCD setting, combining decentralized source training with generalization to an unseen target domain under an open label space. Our proposed framework **NOMAD**, couples a raw-data-free **Global Style Bank**, **Adaptive Prototypical Alignment**, and **REPEAT**-based stability-aware aggregation through a target-risk decomposition covering style discrepancy, client heterogeneity, prototype error, and open-world discovery. Experiments on PACS, Office-Home, and Mini-DomainNet show consistent gains under moderate and severe non-IID heterogeneity.

Limitations. **NOMAD** relies on accurate novel-category cardinality estimation and incurs additional client-side computation from multi-view forward passes. Exchanging style statistics and prototypes preserves data locality but does not provide formal differential privacy or protection against inference attacks.

Ethics Statement

This work uses only publicly available vision benchmarks (PACS, Office-Home, Mini-DomainNet) containing no personally identifying information, and involves no human subjects, no new data collection, and no release of new data. We have read and adhere to the ICLR Code of Ethics.

Two aspects of the setting warrant comment. First, the Global Style Bank transmits channel-wise feature statistics and known-class prototypes rather than raw images. This satisfies the data-locality constraint of federated learning but provides no formal privacy guarantee, and we do not claim one: membership or reconstruction attacks on these summary statistics are not evaluated here, and deployments with real privacy requirements would need differential privacy or secure aggregation. Second, the method outputs clusters of novel categories without semantic labels. In application domains where such a system might plausibly be deployed, including surveillance and clinical settings, treating an unvalidated cluster as a semantic category could cause harm. We regard human validation of discovered clusters as a prerequisite for deployment. A fuller discussion appears in Appendix L.

Reproducibility Statement

The method is specified in Sec. 4, with the full training procedure in Algorithm 1 and the inference procedure in Algorithm 2. Loss definitions are given in Appendix F and the uncertainty estimator in Appendix G. Hyperparameters, including all temperatures, loss weights, and the two-phase schedule, are listed in Sec. 5.1 with sensitivity analyses in Appendix H. Dataset splits, the Dirichlet partitioning procedure, and per-client statistics are in Appendix C; the evaluation protocol, including cluster-count estimation (Appendix E.2) and threshold selection (Appendix I), is common to all methods. Baseline adaptations and their tuning grids are documented in Appendix E.1. For the theoretical result, Appendix A states the assumptions explicitly, including those that involve unobserved target quantities, and gives the complete proof. Source code and configuration files will be released upon acceptance.

AI use statement

We used a large language model to assist with editing and rewriting portions of the manuscript text, and to check internal consistency of notation, cross-references, and reported numbers. All research ideas, experimental design, implementation, and results are our own. We have reviewed and verified all AI-assisted work and take responsibility for the final content of this paper, including text, claims, code, and artifacts.

References

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Jinhua Chen, Yuning Qiu, Keping Yu, and Qibin Zhao. UFL: Uncertainty-driven federated learning, 2025. URL `https://openreview.net/forum?id=PSGBG2rBbg`.

Junming Chen, Meirui Jiang, Qi Dou, and Qifeng Chen. Federated domain generalization for image recognition via cross-client style transfer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 361–370, 2023.

Yang Chen, Yu Wang, Yingwei Pan, Ting Yao, Xinmei Tian, and Tao Mei. A style and semantic memory mechanism for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9164–9173, 2021.

Florent Chiaroni, Jose Dolz, Ziko Imtiaz Masud, Amar Mitiche, and Ismail Ben Ayed. Parametric information maximization for generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1729–1739, 2023.

Sua Choi, Dahyun Kang, and Minsu Cho. Contrastive mean-shift learning for generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23094–23104, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Wei Feng and Zongyuan Ge. Generalized category discovery under domain shift: A frequency domain perspective. In *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Yaming Guo, Kai Guo, Xiaofeng Cao, Tieru Wu, and Yi Chang. Out-of-distribution generalization of federated learning via implicit invariant relationships. In *International Conference on Machine Learning*, pp. 11905–11933. PMLR, 2023.

Akshita Gupta, Sanath Narayan, KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Ow-detr: Open-world detection transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9235–9244, 2022.

Sunny Gupta, Vinay Sutar, Varunav Singh, and Amit Sethi. Fedalign: Federated domain generalization with cross-client feature alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1801–1810, 2025.

Shaozhe Hao, Kai Han, and Kwan-Yee K Wong. Cipr: An efficient framework with cross-instance positive relations for generalized category discovery. *arXiv preprint arXiv:2304.06928*, 2023.

Zhenqi He, Yuanpei Liu, and Kai Han. Seal: Semantic-aware hierarchical learning for generalized category discovery. *Advances in Neural Information Processing Systems*, 38:160765–160794, 2026.

Wenke Huang, Zekun Shi, Mang Ye, He Li, and Bo Du. Self-driven entropy aggregation for Byzantine-robust heterogeneous federated learning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20096–20110. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/huang24u.html.

Dahyun Kang, Piotr Koniusz, Minsu Cho, and Naila Murray. Distilling self-supervised vision transformers for weakly-supervised few-shot classification & segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19627–19638, 2023.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Geeho Kim, Jinu Lee, and Bohyung Han. FedLPA: Local prior alignment for heterogeneous federated generalized category discovery. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=QzOBE4mi2N.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020a.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks, 2020b. URL https://arxiv.org/abs/1812.06127.

Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623*, 2021.

Shengcai Liao, Zhipeng Mo, Jianqing Zhu, Yang Hu, and Stan Z. Li. Open-set person re-identification, 2014. URL https://arxiv.org/abs/1408.0872.

Yuanpei Liu and Kai Han. Debgcd: Debiased learning with distribution guidance for generalized category discovery. *https://arxiv.org/pdf/2504.04804*, 2025.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. Pmlr, 2017.

A Tuan Nguyen, Philip Torr, and Ser Nam Lim. Fedsr: A simple and effective domain generalization method for federated learning. *Advances in Neural Information Processing Systems*, 35:38831–38843, 2022.

Nan Pu, Zhun Zhong, and Nicu Sebe. Dynamic conceptional contrastive learning for generalized category discovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7579–7588, 2023.

Nan Pu, Wenjing Li, Xingyuan Ji, Yalan Qin, Nicu Sebe, and Zhun Zhong. Federated generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28741–28750, June 2024.

Vaibhav Rathore, Shubhranil B, Saikat Dutta, Sarthak Mehrotra, Zsolt Kira, and Biplab Banerjee. When domain generalization meets generalized category discovery: An adaptive task-arithmetic driven approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a.

Vaibhav Rathore, Divyam Gupta, and Biplab Banerjee. Hidisc: A hyperbolic framework for domain generalization with generalized category discovery. In *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS)*, 2025b.

Sai Bhargav Rongali, Sarthak Mehrotra, Ankit Jha, Shirsha Bose, Tanisha Gupta, Mainak Singha, Biplab Banerjee, et al. Cdad-net: Bridging domain gaps in generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2616–2626, 2024.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7492–7501, 2022.

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*, pp. 5018–5027, 2017.

Hongjun Wang, Sagar Vaze, and Kai Han. Hilo: A learning framework for generalized category discovery robust to domain shifts, 2024a. URL https://arxiv.org/abs/2408.04591.

Hongjun Wang, Sagar Vaze, and Kai Han. Sptnet: An efficient alternative framework for generalized category discovery with spatial prompt tuning. *arXiv preprint arXiv:2403.13684*, 2024b.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization, 2020. URL https://arxiv.org/abs/2007.07481.

Menglin Wang, Zhun Zhong, and Xiaojin Gong. Prior-constrained association learning for fine-grained generalized category discovery, 2025.

Xinpeng Wang, Yongxin Guo, and Xiaoying Tang. Fedccrl: Federated domain generalization with cross-client representation learning. *arXiv preprint arXiv:2410.11267*, 2024c.

Yicheng Wang, Feng Liu, Junmin Liu, Zhen Fang, and Kai Sun. Exclusive style removal for cross domain novel class discovery, 2024d. URL https://arxiv.org/abs/2406.18140.

Shuo Wen and Maria Brbic. Cross-domain open-world discovery. *arXiv preprint arXiv:2406.11422*, 2024.

Xin Wen, Bingchen Zhao, and Xiaojuan Qi. Parametric classification for generalized category discovery: A baseline study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16590–16600, 2023.

Kristoffer K Wickstrøm, Thea Brüsch, Michael C Kampffmeyer, and Robert Jenssen. Repeat: Improving uncertainty estimation in representation learning explainability, 2025.

Haoran Xu, Jiaze Li, Jianzhong Ju, and Zhenbo Luo. Federated joint learning for domain and class generalization, 2026. URL https://arxiv.org/abs/2601.12253.

Jie Zhang, Xiaosong Ma, Song Guo, and Wenchao Xu. Towards unbiased training in federated open-world semi-supervised learning. In *International Conference on Machine Learning*, pp. 41498–41509. PMLR, 2023a.

Liling Zhang, Xinyu Lei, Yichun Shi, Hongyu Huang, and Chao Chen. Federated learning with domain generalization. *arXiv preprint arXiv:2111.10487*, 2021.

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, 2023b.

Sheng Zhang, Salman Khan, Zhiqiang Shen, Muzammal Naseer, Guangyi Chen, and Fahad Shahbaz Khan. Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3479–3488, 2023c.

Bingchen Zhao, Xin Wen, and Kai Han. Learning semi-supervised gaussian mixture models for generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16623–16633, 2023.

Zhun Zhong, Yuyang Zhao, Gim Hee Lee, and Nicu Sebe. Adversarial style augmentation for domain generalized urban-scene segmentation. *Advances in neural information processing systems*, 35:338–350, 2022.

Jiaying Zhou, Yang Liu, and Qingchao Chen. Novel class discovery in chest x-rays via paired images and text, 2024.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain adaptive ensemble learning. *IEEE Transactions on Image Processing*, 30:8008–8018, 2021a. ISSN 1941-0042. doi: 10.1109/tip.2021.3112012. URL http://dx.doi.org/10.1109/TIP.2021.3112012.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021b.

A APPENDIX

Appendix: NOMAD

APPENDIX CONTENTS

§A Theoretical Analysis
- Setup, Assumptions, and Four Error Terms
- Theorem 1 (FEDDG-GCD Generalization Bound)
- Term Coupling, Joint Control, and Style Manifold Coverage

§B Qualitative Analysis
- Style Manifold Visualization (t-SNE analysis)

§C Datasets and Benchmarks
- Benchmark Descriptions and Federated Heterogeneity
- Dataset Statistics and LODO Partitions

§D Detailed Training Algorithm
- The **NOMAD** Algorithm (Algorithm 1)
- Computational Efficiency and Style Bank Construction

§E Experimental Results
- Baseline Adaptations for FEDDG-GCD
- Extended Results across all Benchmark Splits

§F Mathematical Definitions of Loss Components
- Definitions for $\mathcal{L}_{\text{InfoNCE}}$, $\mathcal{L}_{\text{SupCon}}$, and $\mathcal{L}_{\text{Proto}}$

§G Detailed Uncertainty Estimation via REPEAT
- Importance as a Bernoulli Random Variable
- Generating Realizations via the Global Style Bank
- Optimal Thresholding and Weighting
- Interpreting High Uncertainty

§H Hyperparameter Optimization
- Sensitivity Analysis (ρ, λ, τ)
- Impact of Alignment Phase Duration (T_{align})

§I Selection and Sensitivity of the Rejection Threshold

§J Extended Ablation Analysis
- Effectiveness across architectures
- Uncertainty Aggregation Analysis.

§K Limitations

§L Potential Societal Impact

A THEORETICAL ANALYSIS

This appendix presents a conditional target-risk decomposition for FEDDG-GCD. Its purpose is to characterize the distinct error sources addressed by **NOMAD**, rather than to establish a distribution-free guarantee for arbitrary unseen target domains. In particular, the target-style, routing, and clusterability terms involve the unobserved target distribution and are used only for analysis.

A.1 SETUP AND NOTATION

Consider K clients with source distributions $\{P_k\}_{k=1}^K$, where $P_k = P_{d_{\phi(k)}}$ denotes the distribution associated with client k. For the analysis, we fix aggregation weights $\omega = (\omega_1, \dots, \omega_K)$ such that $\omega_k \geq 0$ and $\sum_{k=1}^K \omega_k = 1$. The theorem therefore applies to fixed exogenous weights, or to a fixed aggregation history when the corresponding conditional concentration argument is valid. It does not, by itself, establish concentration for adaptively estimated REPEAT weights over the complete optimization trajectory.

The stability-weighted source mixture is

$$\bar{P}_\omega = \sum_{k=1}^K \omega_k P_k. \tag{11}$$

Let A_ξ denote the stochastic Style-Bank transformation of Eq. 3, where ξ includes the sampled Style-Bank entry and the sample-adaptive mixing coefficient. The Style-Bank-augmented source distribution is

$$\bar{P}_\omega^S = (A_\xi)_\# \bar{P}_\omega, \tag{12}$$

where the pushforward includes the randomness of ξ.

Remark (feature-level vs. input-space perturbation). A_ξ as defined in Sec. 4.1, Eq. 3, acts on the intermediate feature map $a_\ell^\theta(x)$, not on x directly. For the pushforward in Eq. 12 to be well-defined as a distribution over the input space $\mathcal{X}$, we write the encoder as a composition $f_\theta = f_{\text{top}} \circ A_\xi \circ f_{\text{bottom}}$, where f_{bottom} maps x to $a_\ell^\theta(x)$ and f_{top} maps the perturbed feature map to the final embedding. A_ξ appearing in Eq. 12 is then a conceptual input-space surrogate for this feature-level operation. This does not change the quantity being bounded, since $\ell_{\text{rep}}(f_\theta(A_\xi(x)), y)$ in Eq. 13 is evaluated through the same composed encoder used at training time.

Remark (exogeneity of ξ). The mixing coefficient $\lambda_i = q_i^\rho$ inside ξ is itself a function of f_θ through the prototype affinity q_i (Eq. 3), so $\bar{P}_\omega^S$ is, strictly, hypothesis-dependent rather than fixed. Theorem A.1 treats ξ's generating distribution as exogenous at the encoder f_θ where the bound is evaluated: Assumption (A2) is stated for that fixed encoder, and $\mathfrak{R}_{\omega,S}(\ell_{\text{rep}} \circ \mathcal{F})$ is computed by first fixing ξ's law at f_θ and then taking the supremum over $\mathcal{F}$. This sidesteps rather than removes the dependency; a fully adaptive treatment in which λ_i tracks f_θ across the optimization trajectory would require a uniform, rather than per-iterate, complexity argument, which we leave to future work. Throughout, $\mathfrak{R}_{\omega,S}(\cdot)$ denotes the *empirical* weighted Rademacher complexity, computed on the realized federated sample $S = \bigcup_k \mathcal{D}_k$ under weights ω_k/N_k; this is the convention under which the constant 3 (rather than 2) multiplies the confidence term in Eq. 25, following standard empirical Rademacher generalization bounds.

Let $\ell_{\text{rep}}(f_\theta(x), y) \in [0, 1]$ be a bounded representation-level surrogate for known-class learning. It is intentionally defined before applying the global prototype decision rule. The weighted empirical representation risk on Style-Bank-augmented known-class source samples is

$$\widehat{R}_\omega^{S,\text{rep}}(f_\theta) = \sum_{k=1}^K \omega_k \frac{1}{N_k} \sum_{i=1}^{N_k} \mathbb{E}_\xi \left[\ell_{\text{rep}} \left(f_\theta(A_\xi(x_i^{(k)})), y_i^{(k)} \right) \right]. \tag{13}$$

In practice, the expectation in Eq. 13 is approximated by the Style-Bank perturbations sampled during local training.

Since every example at client k receives weight ω_k/N_k, the effective sample size for the weighted empirical process is

$$n_{\text{eff}} = \left(\sum_{k=1}^K \frac{\omega_k^2}{N_k} \right)^{-1}. \tag{14}$$

For sample-proportional aggregation, $\omega_k = N_k / \sum_{j=1}^{K} N_j$, Eq. 14 equals the total number of source samples. Reliability-weighted aggregation can reduce n_{eff} when it concentrates mass on a small subset of clients, reflecting the statistical cost of selective aggregation.

A.2 TARGET-RISK DECOMPOSITION

Let P_t be the distribution of the held-out target domain. We decompose target generalized-category-discovery risk as

$$\begin{aligned} R_{\text{test}}^{\text{GCD}}(f_\theta) = R_{\text{test}}^{\text{known}}(f_\theta) + R_{\text{test}}^{\text{novel}\to\text{known}}(f_\theta) \\ + R_{\text{test}}^{\text{known}\to\text{reject}}(f_\theta) + R_{\text{test}}^{\text{cluster}|\text{reject}}(f_\theta). \end{aligned} \tag{15}$$

The first term is the error of classifying target known samples using the global prototype decision rule. The second and third terms are routing errors: novel samples incorrectly accepted as known and known samples incorrectly rejected, respectively. The final term measures clustering error only among novel samples that are correctly routed to the rejection set. To keep Eq. 15 an exact partition of disjoint events rather than a bound, $R_{\text{test}}^{\text{known}}$ is defined only on the subset of known target samples correctly *accepted* by the rejection rule (Eq. 8); known samples incorrectly rejected are counted once, in $R_{\text{test}}^{\text{known}\to\text{reject}}$, so no sample contributes to more than one term.

Let $a_\theta^\ell(x)$ be the intermediate feature map used by the Style Bank and define the style map

$$T_\theta(x) = \left(\mu(a_\theta^\ell(x)), \sigma(a_\theta^\ell(x))\right). \tag{16}$$

We use the following analysis quantities:

$$\Delta_{\text{style}}(f_\theta) = W_1\left(T_{\theta\#}\bar{P}_\omega^S, T_{\theta\#}P_t\right), \tag{17}$$

$$\Gamma_{\text{het}}(f_\theta) = \sum_{k=1}^{K} \omega_k D\left(f_{\theta\#}P_k, f_{\theta\#}\bar{P}_\omega\right), \tag{18}$$

$$\Psi_{\text{proto}}(f_\theta) = \sum_{k=1}^{K} \omega_k \frac{1}{|\mathcal{Y}_k|} \sum_{y\in\mathcal{Y}_k} \left\| m_y^{(k)} - c_y \right\|_2^2, \tag{19}$$

$$\bar{\epsilon}_{\text{proto}}(f_\theta) = \sum_{k=1}^{K} \omega_k \frac{1}{|\mathcal{Y}_k|} \sum_{y\in\mathcal{Y}_k} \mathbb{E}_{x\sim P_{k,y}} \left[\left\| f_\theta(x) - m_y^{(k)} \right\|_2^2 \right]. \tag{20}$$

Here, D is a bounded integral probability metric over normalized representations, $m_y^{(k)}$ is the local prototype of known class y at client k, and c_y is the associated global prototype. The target-style quantity Δ_{style} depends on the unseen target domain and is not available to the training algorithm; it measures residual target-style mismatch after source-side Style-Bank diversification.

Remark (label skew vs. representation shift in Γ_{het}). Because D in Eq. 18 is evaluated on the marginal representation distributions $f_{\theta\#}P_k$, it conflates two distinct effects: differences in the class-conditional representation $P_k(X \mid Y)$ across clients, and differences in the label prior $P_k(Y)$ induced by $\mathbf{p}_k \sim \text{Dir}(\beta\mathbf{1})$. Under severe label skew, Γ_{het} can be large even when class-conditional representations are well aligned, loosening the bound. A class-conditional refinement, $\Gamma_{\text{het}}^{\text{cond}}(f_\theta) = \sum_k \omega_k \sum_y P_k(y)\, D(f_{\theta\#}P_{k,y}, f_{\theta\#}\bar{P}_{\omega,y})$, would isolate representation shift from label-prior mismatch. We retain the simpler marginal form in Eq. 18 because it is directly estimable without per-client class-conditional statistics, and note the conditional form as a tighter but more data-hungry alternative.

Remark (weighting inconsistency). The empirical representation risk $\widehat{R}_\omega^{S,\text{rep}}$ (Eq. 13) averages over samples, so under label skew each class contributes in proportion to its local sample frequency. Ψ_{proto} and $\bar{\epsilon}_{\text{proto}}$ instead average uniformly over the classes present at each client ($1/|\mathcal{Y}_k|$), so a class with one sample and a class with hundreds contribute equally to the prototype terms. This is a deliberate choice – prototype quality for rare classes matters for novel-vs-known routing regardless of source frequency – but it means the sample-weighted and class-weighted terms in Eq. 25 are not directly commensurable; reconciling the two weightings under extreme skew is left as an open direction.

A.3 ASSUMPTIONS

(A1) Bounded augmented representation loss. For every $f \in \mathcal{F}$,

$$\ell_{\mathrm{rep}}(f(A_\xi(x)), y) \in [0, 1]. \tag{21}$$

(A2) Weighted complexity. The loss class $\ell_{\mathrm{rep}} \circ \mathcal{F}$ has finite weighted Rademacher complexity under example weights ω_k/N_k, evaluated at the fixed encoder f_θ per the exogeneity remark above.

(A3) Bank-supported representation transfer. There exist constants $\lambda_s, \lambda_h > 0$ such that

$$R^{\mathrm{rep}}_{\mathrm{test}}(f_\theta) \le R^{\mathrm{rep}}_{\bar{P}^S_\omega}(f_\theta) + \lambda_s \Delta_{\mathrm{style}}(f_\theta) + \lambda_h \Gamma_{\mathrm{het}}(f_\theta). \tag{22}$$

This conditional assumption connects the augmented source representation risk to the representation-level target risk. It does not claim that Style-Bank mixing universally reduces target discrepancy; it states that the residual transfer gap is controlled by target-style mismatch and cross-client representation disagreement.

(A4) Prototype-to-decision stability. There exists a constant $L_p > 0$ – corresponding to the coefficient denoted λ_p in Eq. 9 of the main paper – such that

$$R^{\mathrm{known}}_{\mathrm{test}}(f_\theta) \le R^{\mathrm{rep}}_{\mathrm{test}}(f_\theta) + L_p \left(\bar{\epsilon}_{\mathrm{proto}}(f_\theta)^{1/2} + \Psi_{\mathrm{proto}}(f_\theta)^{1/2} \right). \tag{23}$$

This assumption formalizes the effect of replacing local class-centered representations by global prototype decisions. It follows, for example, when the known-class decision surrogate is Lipschitz in the representation-to- prototype distance.

(A4′) Bounded semantic drift. Assumption (A4) implicitly requires that target representations do not undergo arbitrary semantic drift from the source global prototypes beyond what Δ_{style} captures: since Δ_{style} (Eq. 17) is defined via the channel-wise style map T_θ, it bounds first- and second-order appearance statistics but does not by itself preclude a target-to-source shift in representation content. We make this explicit: there exists $\kappa \ge 0$ such that, for every known class y, $\|\mathbb{E}_{x \sim P_{t,y}}[f_\theta(x)] - \mathbb{E}_{x \sim \bar{\omega}, y}[f_\theta(x)]\|_2 \le \kappa$ whenever $\Delta_{\mathrm{style}}(f_\theta)$ is small – i.e., residual style alignment implies bounded prototype drift between source and target. Assumption (A4) is understood to hold under (A4′).

(A5) Open-world routing and conditional clusterability. The target rejection rule induces routing errors $R^{\mathrm{novel}\to\mathrm{known}}_{\mathrm{test}}$ and $R^{\mathrm{known}\to\mathrm{reject}}_{\mathrm{test}}$. Conditional on correctly rejecting novel target samples, the clustering procedure satisfies

$$R^{\mathrm{cluster|reject}}_{\mathrm{test}}(f_\theta) \le \mathcal{C}_{\mathrm{cluster}} \left(r_{\mathrm{novel}}, \Delta_{\mathrm{novel}}, \widehat{K}_{\mathrm{novel}} \right), \tag{24}$$

where r_{novel} is a within-cluster radius, Δ_{novel} is the minimum separation between distinct novel clusters, and $\widehat{K}_{\mathrm{novel}}$ is the estimated number of novel clusters. For idealized center-based clustering, a sufficient recovery condition can require $\Delta_{\mathrm{novel}} > 4r_{\mathrm{novel}}$, correct cluster cardinality, and convergence to a global clustering optimum. The function $\mathcal{C}_{\mathrm{cluster}}$ explicitly retains these dependencies rather than treating clustering error as a function of separation alone.

A.4 CONDITIONAL TARGET-RISK BOUND

Theorem A.1 (Conditional FEDDG-GCD target-risk decomposition)**.** *Fix aggregation weights ω independently of the source samples used to evaluate Eq. 13. Under Assumptions (A1)–(A5) (with A4 read under A4′), with probability at least $1 - \delta$ over the federated source samples, every $f_\theta \in \mathcal{F}$ satisfies*

$$\begin{aligned} R^{\mathrm{GCD}}_{\mathrm{test}}(f_\theta) \le \widehat{R}^{S,\mathrm{rep}}_{\omega}(f_\theta) + 2\mathfrak{R}_{\omega,S}\left(\ell_{\mathrm{rep}} \circ \mathcal{F}\right) + 3\sqrt{\frac{\log(2/\delta)}{2n_{\mathrm{eff}}}} \\ + \lambda_s \Delta_{\mathrm{style}}(f_\theta) + \lambda_h \Gamma_{\mathrm{het}}(f_\theta) \\ + L_p \left(\bar{\epsilon}_{\mathrm{proto}}(f_\theta)^{1/2} + \Psi_{\mathrm{proto}}(f_\theta)^{1/2} \right) \\ + R^{\mathrm{novel}\to\mathrm{known}}_{\mathrm{test}}(f_\theta) + R^{\mathrm{known}\to\mathrm{reject}}_{\mathrm{test}}(f_\theta) \\ + \mathcal{C}_{\mathrm{cluster}} \left(r_{\mathrm{novel}}, \Delta_{\mathrm{novel}}, \widehat{K}_{\mathrm{novel}} \right), \end{aligned} \tag{25}$$

where n_{eff} is defined in Eq. 14.

Proof. By Assumptions (A1)–(A2), weighted Rademacher generalization on the Style-Bank-augmented source distribution gives, with probability at least $1-\delta$,

$$R^{\text{rep}}_{\bar{P}^S_\omega}(f_\theta) \le \widehat{R}^{S,\text{rep}}_\omega(f_\theta) + 2\mathfrak{R}_{\omega,S}\left(\ell_{\text{rep}} \circ \mathcal{F}\right) + 3\sqrt{\frac{\log(2/\delta)}{2n_{\text{eff}}}}. \tag{26}$$

Applying Assumption (A3) to Eq. 26 yields

$$\begin{aligned} R^{\text{rep}}_{\text{test}}(f_\theta) \le \widehat{R}^{S,\text{rep}}_\omega(f_\theta) + 2\mathfrak{R}_{\omega,S}\left(\ell_{\text{rep}} \circ \mathcal{F}\right) + 3\sqrt{\frac{\log(2/\delta)}{2n_{\text{eff}}}} \\ + \lambda_s \Delta_{\text{style}}(f_\theta) + \lambda_h \Gamma_{\text{het}}(f_\theta). \end{aligned} \tag{27}$$

Next, Assumption (A4), read under (A4′), converts this representation-level target bound into the known-class decision bound:

$$R^{\text{known}}_{\text{test}}(f_\theta) \le R^{\text{rep}}_{\text{test}}(f_\theta) + L_p\left(\bar{\epsilon}_{\text{proto}}(f_\theta)^{1/2} + \Psi_{\text{proto}}(f_\theta)^{1/2}\right). \tag{28}$$

Finally, substituting Eq. 28 into the decomposition in Eq. 15 and applying Assumption (A5) to the conditional clustering term yields Eq. 25. □

A.5 Connection to **NOMAD**

The theorem should be interpreted as a structured explanation of the proposed design, not a distribution-free guarantee: the exogeneity remark above scopes the complexity term to a fixed encoder, and (A4′) makes explicit an assumption the bound otherwise leaves implicit. Within that scope, the Global Style Bank broadens source-side appearance variation and is intended to reduce a source-side proxy for residual target-style discrepancy, Δ_{style}. Adaptive Prototypical Alignment targets within-client prototype concentration, $\bar{\epsilon}_{\text{proto}}$, and local-to-global prototype drift, Ψ_{proto}. Perturbation-stability aggregation seeks to reduce the influence of updates associated with unstable representations under Style-Bank perturbations and thereby improve a source-side proxy for Γ_{het}.

The theorem does not imply that these target quantities are observable during training, nor that reducing a source-side proxy necessarily reduces the corresponding target quantity. Instead, under the stated transfer, prototype-stability, routing, and clusterability assumptions, it identifies the error sources that must be jointly controlled for generalized category discovery on an unseen target domain. This conditional decomposition is summarized by Eq. 9 in the main paper.

B Qualitative Analysis: Style Manifold Visualization

Visualization of Style Manifold Overlap. To investigate why **NOMAD** excels at zero-shot transfer, we visualize the manifold of feature statistics (μ, σ) in Fig. 6. We observe that the source domains form isolated clusters, representing the "appearance gaps" that typically cause model failure under domain shift. However, by sampling from the Global Style Bank, our framework generates a continuous distribution of *Synthesized* styles (teal) that fill these inter-domain voids.

Most notably, the synthesized styles exhibit a high degree of structural alignment with the held-out *Art Painting* target (grey). This confirms that our sampling strategy (Eq. 3) allows the model to "pre-experience" the target domain's appearance statistics during the decentralized training phase. By regularizing the encoder on this interpolated manifold, **NOMAD** learns to discard domain-specific nuisance factors, ensuring that the resulting embedding space is governed by semantic consistency rather than appearance bias.

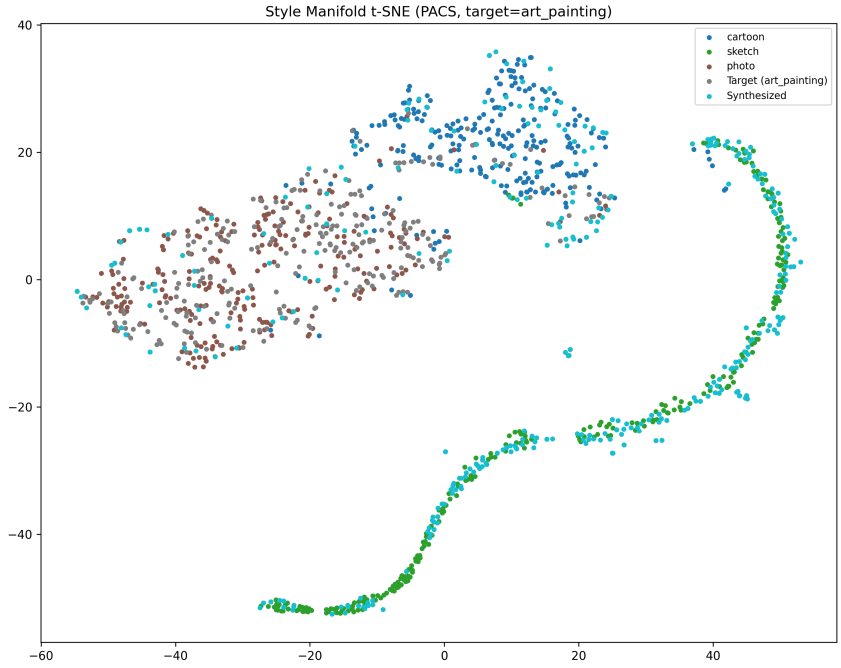

Figure 6: **Empirical Validation of the Global Style Bank.** We present a t-SNE visualization of the channel-wise feature statistics (μ, σ) on the PACS benchmark (d_{test} = Art Painting). While source domains (Cartoon, Sketch, Photo) occupy disjoint clusters, our *Synthesized* styles (teal) effectively populate the convex hull of the appearance manifold. Crucially, the synthesized distribution exhibits significant spatial overlap with the unseen target domain (grey), providing a dense "appearance bridge" that enables zero-shot category discovery.

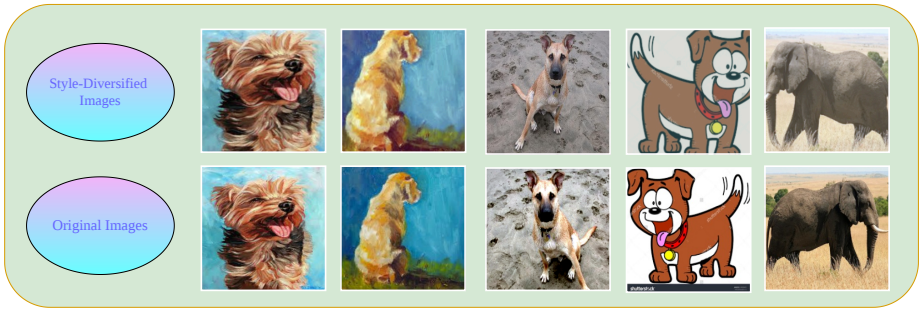

Figure 7: **Qualitative examples of style diversification via the Global Style Bank $\mathcal{S}$.** The bottom row displays original images from various source domains, while the top row shows their corresponding style-diversified counterparts synthesized using Eq. 3. By re-injecting channel-wise feature statistics (μ_s, σ_s) aggregated from the federation, our framework generates novel appearance variations—altering texture, illumination, and contrast—while strictly preserving the spatial layout and semantic identity of the objects. This diversification effectively broadens the style-manifold support of the source distribution, reducing the domain gap Δ_{style} to unseen target environments.

C Datasets and Benchmarks

To rigorously evaluate **NOMAD**under the proposed FedDG-GCD setting, we utilize three distinct benchmarks representing varying levels of domain shift, category scale, and data scarcity. We adopt the **Leave-One-Domain-Out (LODO)** protocol: the model is trained on $N-1$ source domains and evaluated on a strictly held-out target domain containing both known ($\mathcal{Y}_{old}$) and novel ($\mathcal{Y}_{new}$) classes.

C.1 BENCHMARK DESCRIPTIONS

We evaluate on three standard DG benchmarks with progressively increasing scale and heterogeneity:

- **PACS Li et al. (2017):** A high-shift benchmark (Art, Cartoon, Photo, Sketch) with 7 classes. We adopt an aggressive split ($|\mathcal{Y}_{old}| = 4, |\mathcal{Y}_{new}| = 3$) to simulate discovery in low-data regimes. The *Sketch* domain serves as a structural bottleneck, testing robustness to extreme abstraction.
- **Office-Home Venkateswara et al. (2017):** A medium-scale benchmark with 65 categories ($\sim$15.5K images) across 4 domains. We partition the label space into 33 known and 32 novel categories, introducing significant label skew to evaluate mechanisms that prevent dominant-class bias.
- **Mini-DomainNet Zhou et al. (2021a):** A large-scale subset of DomainNet, encompassing 126 classes. With a 64/62 known-to-novel split, it provides a massive inference challenge to assess scalability and cluster fidelity across diverse visual styles.

C.2 FEDERATED HETEROGENEITY

To simulate realistic decentralized conditions, we partition source data across $K = 5$ clients using a Dirichlet distribution Dir(β) following Pu et al. (2024). We evaluate two levels of difficulty:

1. **Normally Heterogeneous (NH, $\beta = 0.20$):** Clients observe moderately imbalanced subsets of known classes, representing standard FL conditions.
2. **Extremely Heterogeneous (EH, $\beta = 0.05$):** A highly skewed setting where clients typically possess samples from only a few unique classes. In PACS, this creates a "sparsity bottleneck" where some clients receive as few as 17 images, stressing the robustness of global aggregation.

C.3 DATASET STATISTICS

Table 3 details the class and image distributions across clients under the most challenging heterogeneity settings.

Table 3: **Client-level Statistics for Federated Heterogeneity ($K = 5$).** We report the min/max variance in class and image counts per client. Shared Known indicates the number of classes present in at least 2 clients.

Dataset	β	# Known Classes		# Train Images		Shared Known	Target Domain (Test)	
		Max	Min	Max	Min	(≥ 2 clients)	# Known	# Novel
PACS	0.05	4	2	2873	17	4	759–2873	911–1056
	0.20	4	2	2873	139	4		
Office-Home	0.05	33	15	2265	670	33	1433–2291	994–2238
	0.20	33	20	2291	582	33		
Mini-DomainNet	0.05	64	45	32346	2983	64	5474–6400	5583–6200
	0.20	64	45	32346	2983	64		

Table 4: **Target-domain testing statistics.** Known/Novel denotes the number of known and novel classes or images, respectively.

Dataset	Target Domain	# Images	# Classes (K/N)	# Images (K/N)
PACS	Photo	1,670	4 / 3	759 / 911
	Art Painting	2,048	4 / 3	1,103 / 945
	Cartoon	2,344	4 / 3	1,327 / 1,017
	Sketch	3,929	4 / 3	2,873 / 1,056
Office-Home	Art	2,427	33 / 32	1,433 / 994
	Clipart	4,365	33 / 32	2,127 / 2,238
	Product	4,439	33 / 32	2,265 / 2,174
	Real World	4,357	33 / 32	2,291 / 2,066
Mini-DomainNet	Clipart	11,224	64 / 62	5,641 / 5,583
	Painting	11,479	64 / 62	5,629 / 5,850
	Real	12,600	64 / 62	6,400 / 6,200
	Sketch	11,418	64 / 62	5,474 / 5,944

D DETAILED TRAINING ALGORITHM FOR **NOMAD**

Algorithm 1 NOMAD: training under FEDDG-GCD

Require: Clients $\{\mathcal{D}_k\}_{k=1}^{K}$; rounds T; alignment rounds T_{align}; local epochs E; mixing power ρ; temperature τ_{agg}; EMA momentum γ; probe repetitions M.
Ensure: $f_{\theta_T}, \mathcal{C}_T, \mathcal{S}_T$.

Initialize f_{θ_0}, head g_{ψ_0}, bank $\mathcal{S}_0$ (K entries), prototypes $\mathcal{C}_0$.
for $t = 0$ to $T-1$ **do**
  Broadcast $(\theta_t, \psi_t, \mathcal{S}_t, \mathcal{C}_t)$ to active clients $\mathcal{K}_t$.
  for all $k \in \mathcal{K}_t$ **in parallel do**
    $(\theta_{t,k}, \psi_{t,k}, u_{t,k}, \mu_{t,k}, \sigma_{t,k}, \mathcal{P}_{t,k}) \leftarrow$ CLIENTUPDATE$(k, \cdot)$.
  end for
  $\omega_{t,k} \leftarrow N_k / \sum N_{k'}$ if $t < T_{\text{align}}$, else Eq. 7 with min–max normalized $\hat{u}_{t,k}$.
  $(\theta_{t+1}, \psi_{t+1}) \leftarrow \sum_k \omega_{t,k}(\theta_{t,k}, \psi_{t,k})$.
  EMA-update bank entry k with $(\mu_{t,k}, \sigma_{t,k})$ for $k \in \mathcal{K}_t$; others unchanged.
  EMA-update each c_y with the mean, over clients holding y, of the *detached* per-client class mean $\mathcal{P}_{t,k}[y]$; renormalize. Classes absent this round are unchanged.
end for
return $(\theta_T, \psi_T, \mathcal{S}_T, \mathcal{C}_T)$.

function CLIENTUPDATE$(k, \theta, \psi, \mathcal{S}, \mathcal{C})$
  for E epochs, each mini-batch $\mathcal{B} \subset \mathcal{D}_k$ **do**
    Unperturbed pass: $q_i \leftarrow \max_y p(y \mid f_\theta(x_i); \mathcal{C})$, $\lambda_i \leftarrow q_i^{\rho}$ ▷ first forward pass
    Stylized pass: sample $(\mu_s, \sigma_s) \sim \mathcal{S}$ and stylize by Eq. 3 ▷ second forward pass
    Compute the *batch-local*, gradient-carrying prototype $m^{\mathcal{B}}_{k,y} \leftarrow |\mathcal{B}_k^y|^{-1} \sum_{x \in \mathcal{B}_k^y} f_\theta(x)$ for each known class $y \in \mathcal{Y}_{\text{local}}$ present in $\mathcal{B}$
    Compute δ_y using $m^{\mathcal{B}}_{k,y}$ (Eq. 5) and descend $\mathcal{L}^{(k)}_{\text{local}}$ (Eq. 4)
  end for
  After local training: $(\mu_k, \sigma_k) \leftarrow$ channel statistics over $\mathcal{D}_k$; $\mathcal{P}_k \leftarrow$ **detached** per-class means $\{m_{k,y}\}_{y \in \mathcal{Y}_k}$ over the *full* local dataset $\mathcal{D}_k$ – distinct from the batch-local $m^{\mathcal{B}}_{k,y}$ above; $\mathcal{P}_k$ feeds only the server-side EMA update of $\mathcal{C}$, not local gradients.
  Stylize a fixed probe M times, threshold each saliency map at its mean, set $p_{ij}^{(k)}$ to the salient fraction, and $u_k \leftarrow \frac{1}{HW} \sum_{i,j} p_{ij}^{(k)}(1 - p_{ij}^{(k)})$.
  return $(\theta, \psi, u_k, \mu_k, \sigma_k, \mathcal{P}_k)$.
end function

Algorithm 2 NOMAD: inductive inference (no target adaptation)

Require: $f_{\theta_T}, \mathcal{C}_T$, target pool $\mathcal{D}_{\text{test}}$, threshold τ_{inf}, estimated count $\widehat{K}_{\text{novel}}$.

Discard g_ψ; compute ℓ_2-normalized $r_j = f_{\theta_T}(x_j)$.
Accept x_j with label $\arg\max_y p(y \mid r_j; \mathcal{C}_T)$ if $\max_y p(y \mid r_j; \mathcal{C}_T) \geq \tau_{\text{inf}}$; otherwise add to rejection set $\mathcal{R}$.
Cluster $\{r_j : x_j \in \mathcal{R}\}$ into $\widehat{K}_{\text{novel}}$ groups by k-means.
return accepted labels and cluster assignments on $\mathcal{R}$.

Algorithm 1 - 2 details the steps of this method. The Style Bank contains one entry per client ($|\mathcal{S}| = K$). The operator $\text{EMA}(\cdot)$ updates the entry associated with each participating client using decay γ and leaves non-participating entries unchanged. For each class y, $\mathcal{Y}_k$ denotes the known classes present at client k. $m^{\mathcal{B}}_{k,y}$ (batch-local, used only inside CLIENTUPDATE to drive $\mathcal{L}_{\text{Margin}}$) and the entries of $\mathcal{P}_k$ (dataset-wide, detached, used only for the server-side EMA refresh of $\mathcal{C}$) are related but distinct quantities and should not be conflated.

D.1 COMPUTATIONAL EFFICIENCY

Table 5 reports estimated total training cost on Office-Home in Peta-FLOPs (PFLOPs), using the same counting protocol for all methods.

Under this accounting, **NOMAD** requires **15.81 PFLOPs**, compared to **16.53 PFLOPs** for HIDISC and **33.06 PFLOPs** for GCD. Relative to these baselines, **NOMAD** reduces total training compute by approximately **4.4%** versus HIDISC and **52.2%** versus GCD. The larger gap to GCD is consistent with differences in optimization recipe and per-step cost , while the smaller gap to HIDISC reflects that HIDISC incurs non-trivial overhead from its domain-generalized open-world pipeline (including synthetic/augmented data generation in our implementation), whereas **NOMAD** concentrates compute in federated local updates with shared style perturbations and lightweight aggregation-side scoring.

Overall, **NOMAD** remains competitive with the strongest DG-GCD baseline in total PFLOPs while substantially reducing cost relative to a standard centralized GCD baseline under the same dataset and convergence criterion used for Table 5.

Table 5: **Training complexity (Office-Home).** Estimated total training PFLOPs to convergence under a unified counting protocol.

Method	**Total PFLOPs**
GCD Vaze et al. (2022)	33.06
HIDISC Rathore et al. (2025b)	16.53
NOMAD (Ours)	**15.81**

E EXPERIMENTAL RESULTS

E.1 BASELINE ADAPTATIONS FOR FEDDG-GCD

No prior method targets FEDDG-GCD directly, so every baseline requires adaptation along at least one axis. Two principles govern these adaptations. First, each baseline is adapted as strongly as its own structure allows, and any training-side advantage available to **NOMAD** is made available to the baseline wherever the method admits it. Second, every method is evaluated under a single inference protocol in which no method, including ours, receives information about the target domain: not the number of target categories, not the known/novel partition, and not any target sample for tuning. Adapting a baseline to its own native evaluation would in several cases require exactly such information, which would place it outside the setting rather than inside it. Configuration files for every entry in Table 1 will be released with the code.

Common evaluation protocol. All methods are evaluated by the procedure of Sec. 4.4: target features are routed by maximum affinity to known-class prototypes at a threshold selected on source-held-out data, and the rejection set is clustered into $\widehat{K}_{\mathrm{novel}}$ groups from the offline estimator of Appendix E.2. For methods that do not maintain prototypes we use per-class means of the labeled source embeddings; for methods with a parametric known-class head, the head's maximum posterior plays the role of the affinity; for HIDISC the affinity is hyperbolic distance to its aggregated prototypes, so the method is evaluated in its own geometry. The threshold is selected per method by the same source-side rule, so no method is advantaged by a better calibrated threshold.

The standard GCD protocol instead clusters the entire target pool and recovers *Old* and *New* by a post-hoc Hungarian assignment. That is sufficient when the task is to partition a pool, but FEDDG-GCD additionally requires a known-versus-novel decision to be made at deployment, without target supervision. Whole-pool clustering never makes that decision, so it leaves the routing behaviour of a method unmeasured. We therefore evaluate the decision directly and apply the same rule to every method.

Shared training protocol. All methods use a DINOv1-pretrained ViT-B/16 with only the final transformer block fine-tuned, 50 communication rounds, $E = 2$ local epochs, $K = 5$ clients under the same Dirichlet partition and seed, and the same LODO splits. Every federated baseline receives the same two-phase schedule as **NOMAD** ($T_{\mathrm{align}} = 5$ warmup rounds under FEDAVG, then the method's own aggregation), since warmup helped all methods and withholding it would advantage ours. Learning rate and any method-specific coefficient were tuned per baseline .

Centralized GCD baselines. GCD Vaze et al. (2022), SimGCD Wen et al. (2023), PCL Li et al. (2020a), PAL-GCD Wang et al. (2025), and GPC Zhao et al. (2023) assume centralized single-domain data. We federate them with FEDAVG, each client optimizing the method's own local objective on its partition. Two choices matter. Several of these methods maintain global statistics (class prototypes in GCD, classifier logits in SimGCD, GMM parameters in GPC) that are ill-defined per client under label skew; we synchronize these at the server every round rather than letting them drift locally. SimGCD and PAL-GCD carry parametric heads over $|\mathcal{Y}_{\mathrm{known}}|$ classes; we keep the head over the global known label space at every client rather than the locally observed subset, so clients holding few classes are not penalized at aggregation.

Federated GCD baselines. AGCL Pu et al. (2024) and FedLPA Kim et al. (2025) are natively federated and require no federation changes. We replace their ResNet-50 backbones with our DINOv1 ViT-B/16 and re-tune their method-specific coefficients, since their published values were selected for a different backbone and a same-domain test set. Both are evaluated on a domain never seen during training, which is outside their design scope; this is a property of the setting, not of our adaptation.

Federated domain generalization baselines. FedAlign Gupta et al. (2025), FedADG Zhang et al. (2021), and GA Zhang et al. (2023b) are closed-set and have no mechanism for novel categories, so a discovery stage must be supplied. We apply the common protocol without modification, using per-class means of the labeled source embeddings as prototypes. This adds no target information

and no method-specific tuning, and it gives these baselines the same routing mechanism **NOMAD** uses, so the comparison isolates the learned representation rather than the discovery head.

DG-GCD baselines. HIDISC Rathore et al. (2025b) and DG^{2}CD-Net Rathore et al. (2025a) are centralized. For HIDISC we distribute the hyperbolic prototypical learning: each client maintains local hyperbolic prototypes and the server aggregates them on the manifold rather than in the ambient space, since Euclidean averaging of hyperbolic prototypes degraded accuracy in preliminary runs. For DG^{2}CD-Net, whose task-arithmetic procedure assumes multiple source domains at one site, each client performs local style-perturbation and task-vector construction over its own partition and the server aggregates the resulting encoder weights; the method therefore composes task vectors within rather than across domains. This is a structural limitation of the setting rather than a weakened implementation, and it likely costs DG^{2}CD-Net accuracy relative to its centralized form. We state it explicitly rather than leaving it implicit in the numbers.

E.2 Unknown Category Number and Cluster-Count Estimation

Prior GCD and cross-domain GCD methods estimate the category count by running k-means over labelled and unlabelled data jointly and scoring candidates by clustering accuracy on the labelled subset Vaze et al. (2022); Rathore et al. (2025a); Rongali et al. (2024). This requires labelled samples to be co-located with the target pool, which FEDDG-GCD does not permit: at deployment the server holds $(f_{\theta_T}, \mathcal{C}_T)$ and no source images. We therefore estimate the cluster count directly on the rejection set, using only its own geometry.

Given the routing decision of Eq. 8, let $\mathcal{R} = \{r_j : x_j \in \mathcal{D}_{\text{test}}, \max_y p(y \mid r_j; \mathcal{C}_T) < \tau_{\text{inf}}\}$ be the set of rejected target representations. For each candidate $K \in [1, K_{\max}]$ we run k-means on $\mathcal{R}$, obtaining clusters $\{\mathcal{G}_j\}_{j=1}^{K}$ with ℓ_2-normalized centroids v_j, and score the partition by

$$S(K) = \underbrace{\frac{1}{K}\sum_{j=1}^{K}\Big(1 - \max_{j' \neq j} \cos(v_j, v_{j'})\Big)}_{\text{separation}} + \underbrace{\frac{1}{|\mathcal{R}|}\sum_{j=1}^{K}\sum_{r \in \mathcal{G}_j} \cos(r, v_j)}_{\text{compactness}}, \qquad \widehat{K}_{\text{novel}} = \arg\max_K S(K). \tag{29}$$

The two terms pull in opposite directions. Compactness rises monotonically with K, since finer partitions place every point nearer its centroid, and would alone select $K_{\max}$. Separation falls as K grows, because splitting a category produces near-duplicate centroids with high mutual cosine. Their sum is therefore maximized at a partition that is tight without being fragmented. Both terms lie in $[0, 1]$ whenever the relevant cosines are non-negative, so no relative weighting is required, and the estimate is obtained without target labels, raw source data, or additional communication. Because $\mathcal{R}$ is produced by the routing rule, the estimate inherits any routing error: known-class samples pushed into $\mathcal{R}$ can inflate $\widehat{K}_{\text{novel}}$.

For comparability with methods that report a total category count, we quote $\widehat{K}_{\text{tot}} = \widehat{K}_{\text{novel}} + |\mathcal{Y}_{\text{known}}|$ in Table 6.

Accuracy of the estimate. Table 6 compares estimated counts against ground truth. **NOMAD** recovers the count exactly on PACS and Office-Home and deviates by two on the 126-class Mini-DomainNet benchmark, reducing the over-clustering seen in the compared methods. We attribute this to prototypical alignment and the margin term, which produce tighter known-class anchors and better separated groups in the rejection set, so spurious splits are less likely to register as extra clusters.

Effect on downstream accuracy. Table 7 repeats inference with the ground-truth novel count in place of the estimate, on the two hardest PACS targets. Routing is unaffected by the cluster count, so *Old* is identical across rows and any change in *All* comes from clustering alone. Forcing the correct count raises novel-class accuracy by 8.6 to 11.0 points for the two baselines, since over-clustering splits true novel categories that Hungarian matching then leaves unmatched; the effect on *All* scales with each pool's novel fraction. **NOMAD** estimates the count exactly on PACS and pays no such penalty.

Table 6: **Estimated number of categories.** Exact estimates in green, deviations of at most 3 in orange, larger deviations in red.

Method	PACS (7)	Office-Home (65)	Mini-DomainNet (126)
Ground Truth	7	65	126
FedLPA	12	70	135
FedAlign	10	68	136
HIDISC	7	66	131
NOMAD (Ours)	**7**	**65**	**128**

Table 7: **Ground-truth vs. estimated novel-cluster count** (PACS, NH, $\beta = 0.2$). Rejection-set clustering uses either the ground-truth novel count $K_{\text{novel}} = 3$ or the offline estimate $\widehat{K}_{\text{novel}}$. Main-paper numbers correspond to the $\widehat{K}_{\text{novel}}$ rows. **NOMAD** estimates the category count exactly on PACS, so its two rows coincide.

			Cartoon			Sketch		
Method	**#Clusters**	$\widehat{K}_{\text{tot}}$	All	Old	New	All	Old	New
FedLPA	K_{novel} (GT)	7	47.45	**53.58**	39.46	48.47	**50.99**	41.60
	$\widehat{K}_{\text{novel}}$ (est.)	12	42.66	53.58	28.42	46.02	50.99	32.48
	Δ*All*			−4.79			−2.45	
FedAlign	K_{novel} (GT)	7	45.19	**61.42**	24.01	45.98	**52.63**	27.87
	$\widehat{K}_{\text{novel}}$ (est.)	10	41.34	61.42	15.14	43.68	52.63	19.32
	Δ*All*			−3.85			−2.30	
NOMAD (Ours)	K_{novel} (GT)	7	**68.77**	**68.65**	**68.93**	**52.46**	**54.79**	**46.12**
	$\widehat{K}_{\text{novel}}$ (est.)	7	**68.77**	**68.65**	**68.93**	**52.46**	**54.79**	**46.12**
	Δ*All*			**0.00**			**0.00**	

Table 8: Comparison of methods on PACS $\beta = 0.2$ All / Old / New accuracy (%).

Method	Photo			Art Painting			Cartoon			Sketch		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New
ViT	93.65	91.70	95.28	49.85	42.16	58.84	51.49	59.76	40.71	40.60	35.68	53.98
GCD	77.90	75.63	79.80	89.40	84.22	95.45	53.46	66.54	36.38	45.56	39.99	60.70
AGCL	70.30	67.06	73.00	85.79	82.68	89.42	65.44	65.71	65.09	43.19	38.91	54.83
FedProx+AGCL	81.98	99.47	67.40	63.53	64.91	61.90	42.24	41.37	43.36	39.88	44.38	27.65
PCL	26.17	17.39	33.48	20.17	16.86	24.02	22.10	23.29	20.55	34.64	35.71	31.72
GPC	68.68	75.36	63.12	86.96	83.23	91.32	56.14	51.54	62.14	43.78	43.20	45.36
SiMGCD	55.87	97.10	21.51	53.37	78.51	24.02	47.27	68.05	20.16	37.85	48.17	9.75
PAL-GCD	56.23	90.78	27.44	56.64	77.52	32.28	50.26	64.51	31.66	41.05	50.05	16.57
FedLPA	76.77	99.74	57.63	57.32	83.86	26.35	42.66	53.58	28.42	46.02	50.99	32.48
DG^{2}CD-Net	99.58	99.47	99.67	88.62	86.40	91.22	58.62	62.25	53.88	46.53	46.08	47.73
HiDISC	99.16	99.34	99.01	88.18	84.13	92.91	65.36	65.79	64.80	46.70	45.32	50.47
FedAlign	76.23	97.63	58.40	41.80	40.07	43.81	41.34	61.42	15.14	43.68	52.63	19.32
FedADG	99.22	99.34	99.12	87.65	83.14	92.91	61.35	67.14	53.79	46.35	45.81	47.82
GA	96.35	97.50	95.39	45.61	56.84	32.49	45.99	53.43	36.28	46.75	47.16	45.64
NOMAD	**99.64**	**99.87**	**99.45**	**91.02**	**87.67**	**94.92**	**68.77**	**68.65**	**68.93**	**52.46**	**54.79**	**46.12**

Table 9: Comparison of methods on PACS $\beta = 0.05$ All / Old / New accuracy (%).

Method	Photo			Art Painting			Cartoon			Sketch		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New
ViT	70.30	64.95	74.75	52.69	45.24	61.38	46.63	53.96	37.07	48.56	48.24	49.43
GCD	84.55	75.63	91.99	75.54	67.45	84.97	68.47	70.23	66.18	46.04	45.56	47.35
AGCL	85.69	75.36	94.29	86.87	83.41	90.90	49.53	47.02	52.80	46.53	45.49	49.34
FedProx+AGCL	98.02	97.23	98.68	61.52	61.11	62.01	43.69	59.01	23.70	47.87	46.92	50.47
PCL	22.87	18.45	26.56	19.68	15.23	24.87	22.53	24.72	19.67	37.36	40.41	29.07
GPC	70.12	66.67	73.00	87.11	83.41	91.43	56.48	57.50	55.16	46.78	45.67	49.81
SiMGCD	55.33	99.60	18.44	56.74	79.78	29.84	43.00	60.36	20.35	39.20	48.38	14.20
PAL-GCD	65.15	99.34	36.66	61.28	80.87	38.41	49.96	63.90	31.76	37.41	44.07	19.32
FedLPA	93.89	99.08	89.57	57.18	81.69	28.57	49.96	66.09	28.91	46.91	56.46	20.93
DG^{2}CD-Net	95.34	95.47	95.23	86.43	83.31	90.07	66.22	69.32	62.18	42.68	42.61	42.88
HiDISC	99.22	99.21	99.23	88.72	85.13	92.91	57.55	56.29	59.19	46.73	45.70	49.53
FedAlign	60.72	81.42	43.47	43.90	44.88	42.75	42.45	36.40	50.34	40.67	41.59	38.16
FedADG	99.04	99.47	98.68	85.84	83.23	88.89	59.47	63.45	54.28	46.07	45.77	46.88
GA	72.57	97.76	51.59	42.87	53.85	30.05	48.21	58.10	35.30	45.00	44.80	45.55
NOMAD	**99.34**	**99.47**	**99.23**	**91.16**	**87.49**	**95.45**	**70.48**	**73.85**	**66.27**	**47.37**	**46.29**	**50.28**

E.3 Extended Results

Table 10: Comparison of methods on Office Home $\beta = 0.2$ All / Old / New accuracy (%).

Method	Art			Clipart			Product			Real World		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New
ViT	47.05	53.94	37.12	30.33	35.21	25.69	63.44	70.99	55.57	64.84	75.95	52.52
GCD	50.43	58.34	39.03	35.01	42.13	28.24	63.71	71.13	55.98	63.32	73.07	52.52
AGCL	47.92	53.31	40.14	31.55	35.03	28.24	64.38	69.09	59.48	61.90	66.83	56.44
FedProx+AGCL	45.90	49.55	40.64	32.12	36.67	27.79	65.10	73.73	56.12	59.88	62.24	57.26
PCL	8.61	8.58	8.65	8.73	8.98	8.49	9.75	10.86	8.60	7.53	6.94	8.18
GPC	51.05	57.85	41.25	32.03	38.22	26.14	66.95	72.05	61.64	61.51	66.74	55.71
SiMGCD	47.59	68.39	17.61	28.36	43.21	14.25	52.58	81.81	22.13	53.87	83.28	21.25
PAL-GCD	48.62	69.71	18.21	31.96	47.48	17.20	55.46	80.09	29.81	55.08	83.02	24.10
FedLPA	45.98	58.62	27.77	35.05	44.24	26.32	62.00	72.49	51.06	57.98	73.85	40.37
DG^{2}CD-Net	52.69	57.82	45.31	34.27	37.35	31.34	65.47	63.84	67.17	63.61	69.64	56.93
HiDISC	48.99	53.24	42.86	32.07	34.84	29.45	65.49	69.58	61.22	63.71	69.36	57.45
FedAlign	48.21	62.81	27.16	35.01	44.62	25.87	60.26	71.96	48.07	59.01	76.08	40.08
FedADG	51.67	59.11	40.95	36.56	40.76	32.57	67.61	67.42	67.80	64.63	72.59	55.81
GA	51.38	54.71	46.58	34.43	39.92	29.22	66.52	72.67	60.12	68.12	72.98	62.73
NOMAD	**56.08**	**65.32**	**42.76**	**38.58**	**44.33**	**33.11**	**69.66**	**78.23**	**60.72**	**71.36**	**76.74**	**65.39**

Table 11: Comparison of methods on Office Home $\beta = 0.05$ All / Old / New accuracy (%).

Method	Art			Clipart			Product			Real World		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New
ViT	52.12	61.27	38.93	33.93	38.74	29.36	64.25	75.36	52.67	65.94	73.99	57.02
GCD	49.24	56.32	39.03	31.94	37.14	26.99	64.43	74.66	53.77	66.01	74.16	56.97
AGCL	49.94	57.99	38.33	33.97	37.75	30.38	63.33	66.58	59.94	65.25	75.43	53.97
FedProx+AGCL	49.98	53.45	44.97	32.42	34.51	30.43	64.18	68.43	59.75	57.72	68.01	46.32
PCL	8.49	8.09	9.05	9.05	8.60	9.47	10.30	11.35	9.20	7.78	7.68	7.89
GPC	49.28	54.36	41.95	34.68	39.02	30.56	63.30	65.39	61.13	62.34	72.06	51.55
SiMGCD	46.85	65.32	20.22	29.14	42.31	16.62	52.56	74.30	29.90	52.95	85.55	16.80
PAL-GCD	47.59	66.85	19.82	31.34	44.90	18.45	55.40	78.90	30.91	53.87	81.62	23.09
FedLPA	45.49	56.11	30.18	33.68	39.87	27.79	61.52	72.63	49.95	60.84	78.61	41.14
DG^{2}CD-Net	49.72	54.39	43.00	33.19	34.68	31.78	64.61	66.29	62.85	63.11	69.14	56.43
HiDISC	46.07	50.24	40.04	32.71	36.67	28.95	64.11	64.68	63.52	60.04	65.56	53.92
FedAlign	47.71	60.22	29.68	35.76	47.39	24.71	58.68	70.64	46.23	60.43	77.39	41.63
FedADG	54.84	63.01	43.06	36.54	42.17	31.19	62.76	67.68	57.64	67.00	73.77	59.49
GA	51.82	58.60	42.06	33.46	36.94	30.14	66.40	71.89	60.68	66.11	71.72	59.89
NOMAD	**56.24**	**64.90**	**43.76**	**38.49**	**45.93**	**31.41**	**68.93**	**79.87**	**57.54**	**70.42**	**78.52**	**61.42**

Table 12: Comparison of methods on Mini-DomainNet for $\beta = 0.2$ All / Old / New accuracy (%).

Method	Clipart			Painting			Real			Sketch		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New
ViT	35.19	40.06	30.28	52.21	53.99	50.49	68.79	70.22	67.31	33.72	37.98	29.81
GCD	43.29	55.13	31.33	56.69	62.43	51.18	71.56	72.69	70.4	37.59	41.56	33.93
AGCL	25.11	39.78	10.28	31.37	52.14	11.38	52.3	55.44	49.06	16.73	24.13	9.91
FedProx+AGCL	24	37.99	9.87	29.36	47.02	12.36	49.59	50.72	48.42	13.7	15.33	12.2
PCL	21.27	22.56	20.97	22.06	22.7	21.61	22.86	23.45	21.94	20.67	21.47	18.86
GPC	37.05	45.36	28.66	52.7	55.66	49.86	68.37	67.47	69.29	32.87	36.24	29.76
SiMGCD	44.01	62.06	25.77	49.25	67.12	32.05	68.53	74.75	62.11	37.72	51.39	25.13
DG^{2}CD-Net	41.96	46.58	37.46	48.56	49.36	47.78	56.77	58.82	54.66	39.87	43.72	36.37
HiDISC	42.27	46.56	37.81	48.15	47.19	49.13	57.51	59.13	55.94	41.87	44.39	38.45
FedAlign	37.38	46.59	28.07	41.26	48.3	34.48	61.13	69.77	52.23	31.49	39.81	23.84
FedADG	40.10	46.79	33.33	54.87	54.18	55.53	69.246	72.71	65.66	36.15	38.92	33.60
GA	36.38	41.25	31.45	54.31	56.19	52.5	67.98	67.86	68.11	33.24	35.06	31.56
NOMAD	**49.41**	**60.08**	**38.64**	**56.94**	**61.47**	**52.58**	**70.46**	**77.62**	**63.06**	**42.34**	**50.13**	**35.16**

Table 13: Comparison of methods on Mini-DomainNet for $\beta = 0.05$ All / Old / New accuracy (%).

Method	Clipart			Painting			Real			Sketch		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New
ViT	35.19	40.06	30.28	52.64	51.49	53.75	68.79	70.22	67.31	33.72	37.98	29.81
GCD	43.95	56.66	31.11	56.65	63.01	50.02	69.43	74.64	64.66	36.89	44.30	30.06
AGCL	23.36	37.55	9.03	29.28	45.09	14.07	50.56	51.75	49.34	15.02	20.24	10.21
FedProx+AGCL	24.13	38.5	9.6	30.12	48.45	12.48	49.85	51.31	48.34	15.61	20.61	11
PCL	22.06	21.56	22.58	22.86	22.94	20.84	21.59	22.5	20.15	20.38	24.85	18.56
GPC	36.89	42.17	31.56	52.48	56.21	48.89	69.06	71.08	66.97	32.46	36.04	29.16
SiMGCD	43.84	62.08	25.42	48.74	66.92	31.25	67.56	77.03	57.79	37.71	50.24	26.18
DG^{2}CD-Net	38.98	41.8	36.12	46.75	47.7	45.83	59.95	61.05	58.83	38.86	42.06	35.92
HiDISC	39.52	44.38	34.52	44.29	45.31	43.23	56.86	62.5	51.03	36.51	43.23	30.32
FedAlign	38.91	48.08	29.64	43.75	52.46	35.37	60.71	72.64	48.4	31.98	38.14	26.3
FedADG	39.02	45.33	32.65	55.6	57.24	54.02	72.01	74.5	69.44	36.57	40.35	33.08
GA	26.97	29.07	24.84	53.82	54.56	53.11	69.56	71.67	67.39	32.71	35.97	29.71
NOMAD	**50.2**	**60.62**	**39.68**	**59.41**	**65.36**	**53.7**	**67.78**	**73.54**	**61.94**	**42.38**	**50.38**	**35.16**

F Detailed Mathematical Definitions of Loss Components

We provide the formal definitions of the loss components in $\mathcal{L}_{\text{local}}^{(k)}$ from Section 4.2. Let

$$h_i = \frac{f_\theta(x_i)}{\|f_\theta(x_i)\|_2} \in \mathbb{R}^{768}, \qquad z_i = \frac{g_\psi(f_\theta(x_i))}{\|g_\psi(f_\theta(x_i))\|_2} \in \mathbb{R}^{768} \tag{30}$$

denote the normalized encoder and projection-head representations, respectively. InfoNCE and supervised contrastive learning operate on z_i, whereas prototypical alignment operates on h_i, consistent with the projection-head-free inference geometry.

F.1 Instance-Level Invariance with InfoNCE

We use InfoNCE van den Oord et al. (2019) to enforce invariance between original and style-diversified views. Given a mini-batch of N images, we form $2N$ views containing $\{x_1, \ldots, x_N\}$ and $\{\tilde{x}_1, \ldots, \tilde{x}_N\}$. For each $i \in \{1, \ldots, 2N\}$, let $j(i)$ denote the index of its corresponding positive view. The loss is

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{2N} \sum_{i=1}^{2N} \log \frac{\exp\big(\cos(z_i, z_{j(i)})/\tau_{\text{nce}}\big)}{\sum\limits_{\substack{a=1 \\ a \neq i}}^{2N} \exp(\cos(z_i, z_a)/\tau_{\text{nce}})}, \tag{31}$$

where τ_{nce} is the temperature. This objective encourages semantic consistency while suppressing style-specific appearance variations.

F.2 Supervised Contrastive Learning

We use supervised contrastive learning Khosla et al. (2020) to reduce within-class scatter among labeled known-category samples. Let I_ℓ denote the labeled indices in the augmented batch, and define

$$P(i) = \{p \in I_\ell \setminus \{i\} : y_p = y_i\}. \tag{32}$$

The supervised contrastive loss is

$$\mathcal{L}_{\text{SupCon}} = -\frac{1}{|I_\ell|} \sum_{i \in I_\ell} \frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\cos(z_i, z_p)/\tau_{\text{sup}})}{\sum\limits_{a \in I_\ell \setminus \{i\}} \exp(\cos(z_i, z_a)/\tau_{\text{sup}})}, \tag{33}$$

where anchors with $P(i) = \varnothing$ are omitted and τ_{sup} is the supervised contrastive temperature. This term encourages compact known-class clusters under client-specific label and style heterogeneity.

F.3 Prototypical Alignment

We align encoder representations with the server-maintained known-class prototypes

$$\mathcal{C} = \{c_y \in \mathbb{R}^{768}\}_{y \in \mathcal{Y}_{\text{known}}}, \qquad \|c_y\|_2 = 1. \tag{34}$$

The alignment is performed directly in the encoder space used at inference:

$$\mathcal{L}_{\text{Proto}} = -\frac{1}{|I_\ell|} \sum_{i \in I_\ell} \log \frac{\exp(\cos(h_i, c_{y_i})/\tau_{\text{p}})}{\sum\limits_{y' \in \mathcal{Y}_{\text{known}}} \exp(\cos(h_i, c_{y'})/\tau_{\text{p}})}, \tag{35}$$

where y_i is the ground-truth label and τ_{p} is the prototypical temperature. This objective mitigates cross-client semantic drift while maintaining consistency between the training and inference geometries.

The remaining component, $\mathcal{L}_{\text{Margin}}$, is defined in Eqs. 5 and 6. It also operates directly in the encoder space through batch-local and global prototypes.

F.4 Summary of Hyperparameters

We set $\tau_{\text{nce}} = 0.07$ and $\tau_{\text{sup}} = \tau_{\text{p}} = \tau_m = 0.1$. The loss weights are $\lambda_{\text{nce}} = 1.0$, $\lambda_{\text{sup}} = 0.35$, $\lambda_{\text{proto}} = 0.5$, and $\lambda_{\text{mar}} = 0.25$. Sensitivity to these temperatures and loss weights is reported in Section H.

G Uncertainty Estimation for Client Weighting

This appendix explains how we compute the per-client scores u_k that appear in Eq.7 and how they become aggregation weights ω_k. The idea is simple: *we measure how much the model's saliency map wiggles when we change only the style of the same probe image.* Clients with steadier saliency get higher weight; clients with jittery saliency get down-weighted.

G.1 What we measure at each pixel

Fix one probe image x_{pr} (the same protocol for every client). For each spatial location (i, j), we ask a yes/no question after each style perturbation:

"Is this location important for the model's explanation map?"

We answer it M times by repeating the following steps:

1. Sample style statistics from the Global Style Bank $\mathcal{S}$ and build a stylized version of x_{pr} using the same style-mixing rule as training (Eq.3).
2. Run a fixed attribution method (as in REPEAT Wickstrøm et al. (2025)) to obtain a nonnegative importance score $\bar{R}_{ij}^{(m)}$ for realization $m \in \{1, \ldots, M\}$.
3. Convert scores to a binary indicator $I_{ij}^{(m)} \in \{0, 1\}$ by thresholding (next subsection).

If the answers are often split between 0 and 1, the location is "uncertain" under style changes. If the answers are almost always 0 or almost always 1, the location is stable.

G.2 Thresholding (mean rule)

For each realization m, we set a threshold to the mean map intensity:

$$\tau^{(m)} = \frac{1}{HW} \sum_{i,j} \bar{R}_{ij}^{(m)}. \tag{36}$$

We then define

$$I_{ij}^{(m)} = \begin{cases} 1, & \bar{R}_{ij}^{(m)} \geq \tau^{(m)}, \\ 0, & \text{otherwise.} \end{cases} \tag{37}$$

This is a lightweight, reproducible binarization rule that matches the REPEAT-style pipeline used in our implementation.

G.3 From many yes/no draws to one score per pixel

We estimate the frequency of "important" answers by averaging:

$$\widehat{p}_{ij} = \frac{1}{M} \sum_{m=1}^{M} I_{ij}^{(m)}. \tag{38}$$

We then use the familiar Bernoulli uncertainty proxy

$$\widehat{U}_{ij} = \widehat{p}_{ij} \left(1 - \widehat{p}_{ij}\right), \tag{39}$$

which is largest near $\widehat{p}_{ij} \approx 0.5$ (lots of flipping) and small near 0 or 1 (stable).

Default choice for M. We use a fixed M for all clients and all methods (see Appendix H for the value). Larger M makes the estimate smoother but costs more compute.

G.4 From pixels to a client weight

We summarize spatial uncertainty with a single scalar for client k:

$$u_k = \frac{1}{HW} \sum_{i,j} \widehat{U}_{ij}, \tag{40}$$

which matches Eq. (6) after identifying $p_{ij}^{(k)}$ with $\widehat{p}_{ij}$. The server converts $\{u_k\}$ to nonnegative weights $\{\omega_k\}$ using the softmax mapping in Eq. (6).

G.5 What this score is (and is not)

What it is. A practical stability signal: "under style perturbations from $\mathcal{S}$, does the model's explanation concentrate consistently?" This aligns with training because local updates already see the same style family from $\mathcal{S}$.

What it is not. It is not a calibrated probability of correct prediction, and it is not a formal privacy guarantee. It also depends on reasonable choices (probe image, M, attribution backend, thresholding); we keep these fixed across comparisons and report sensitivity in Appendix H.

Note on weighted variants. Some drafts reweight indicators by raw saliency magnitudes. Unless weights are carefully normalized, the resulting "probability" need not lie in $[0, 1]$, and the $\widehat{p}(1 - \widehat{p})$ interpretation stops being clean. Our default is the unweighted $\widehat{p}_{ij}$ above.

H HYPERPARAMETER OPTIMIZATION

Sensitivity to Mixing Strength ρ. We evaluate the effect of varying ρ on performance under both NH and EH settings. As shown in Fig. 8, the performance varies only marginally across different values of ρ, indicating that the proposed method is largely insensitive to this hyperparameter. While slightly stronger mixing provides minor gains in more challenging settings, the overall variation remains small, demonstrating the robustness of our adaptive style diversification.

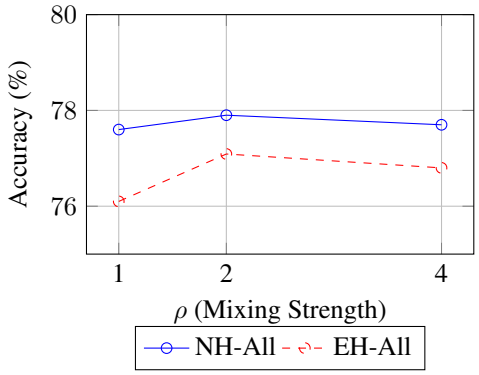

Figure 8: **Sensitivity to mixing strength ρ.** Performance varies only marginally across ρ, indicating robustness to the choice of mixing intensity.

Sensitivity Analysis and Loss Balancing. As summarized in Table 14, **NOMAD** maintains high performance stability on the PACS benchmark ($\beta = 0.2$) across a wide range of coefficient variations. We specifically set $\lambda_{\text{proto}} = 0.5$ to ensure that the global semantic coordinate system remains a primary anchor for feature organization without over-compressing the embedding space. This selection balances the structural signal of the prototypes with the high-frequency instance discrimination provided by $\lambda_{\text{nce}} = 1.0$. Our sensitivity analysis corroborates this design: while the framework is highly robust to variations in contrastive and supervised weights, reducing λ_{proto} to 0.1 leads to a significant performance degradation in both *Old* and *New* categories. This confirms that a sufficient prototypical anchor is essential for stable cluster formation and category discovery in the FEDDG-GCD setting.

Table 14: **Sensitivity Analysis of Loss Weights.** We evaluate the stability of our model by varying the weights $\lambda_{\text{nce}}, \lambda_{\text{sup}}, \lambda_{\text{proto}}$, and λ_{mar} relative to the optimal configuration established in our main paper. Experiments are conducted on PACS under the NH ($\beta = 0.2$) setting.

Setting	λ_{nce}	λ_{sup}	λ_{proto}	λ_{mar}	All	Old	New
NOMAD (Ours)	**1.00**	**0.35**	**0.50**	**0.25**	**77.97**	**77.74**	**77.35**
$\mathcal{L}_{\text{InfoNCE}}$ Var.	0.50	0.35	0.50	0.25	76.42	76.10	75.88
	2.00	0.35	0.50	0.25	77.30	77.12	76.95
$\mathcal{L}_{\text{SupCon}}$ Var.	1.00	0.10	0.50	0.25	75.88	75.72	75.14
	1.00	0.70	0.50	0.25	77.15	76.98	76.92
$\mathcal{L}_{\text{Proto}}$ Var.	1.00	0.35	0.10	0.25	73.10	72.85	72.40
	1.00	0.35	1.00	0.25	77.40	77.15	77.22
$\mathcal{L}_{\text{Margin}}$ Var.	1.00	0.35	0.50	0.10	76.24	76.10	75.98
	1.00	0.35	0.50	0.50	77.05	76.92	76.84

Sensitivity to Temperature Hyperparameters. We evaluate the sensitivity of **NOMAD** to the temperature scales $\{\tau_{\text{nce}}, \tau_{\text{sup}}, \tau_{\text{p}}, \tau_{\text{m}}\}$ which govern the sharpness of the probability distributions in the embedding space. As shown in Table 15, the model remains largely robust to variations within the $[0.05, 0.2]$ range. We observe that setting $\tau_{\text{nce}} = 0.07$ provides the optimal balance for instance-level discrimination; increasing this value to 0.2 overly smoothes the distribution, leading to a marginal drop in "New" class discovery as the contrastive signal becomes less discriminative. Similarly, the uniformity across $\tau_{\text{sup}}, \tau_{\text{p}}$, and τ_{m} at 0.1 ensures that cluster-level and geometric-margin constraints operate on a comparable logit scale. The stability across these values confirms that **NOMAD** effectively organizes the feature manifold without requiring hyper-specific temperature tuning, a critical trait for zero-shot transfer to unseen domains.

Table 15: **Temperature Hyperparameter Sensitivity (Supplementary).** Evaluation of temperature scales on the PACS benchmark ($\beta = 0.2$). The base configuration utilizes $\tau_{\text{nce}} = 0.07$ and $\tau_{\text{sup}} = \tau_{\text{p}} = \tau_{\text{m}} = 0.1$.

Setting	τ_{nce}	τ_{sup}	τ_{p}	τ_{m}	All	Old	New
NOMAD (Ours)	**0.07**	**0.10**	**0.10**	**0.10**	**77.97**	**77.74**	**77.35**
τ_{nce} (InfoNCE)	0.05	0.10	0.10	0.10	77.42	77.10	77.05
	0.20	0.10	0.10	0.10	76.15	76.02	75.80
τ_{sup} (SupCon)	0.07	0.05	0.10	0.10	77.21	77.05	76.84
	0.07	0.20	0.10	0.10	76.58	76.40	76.22
τ_{p} (Proto)	0.07	0.10	0.05	0.10	77.10	76.85	76.70
	0.07	0.10	0.20	0.10	76.45	76.22	76.05
τ_{m} (Margin)	0.07	0.10	0.10	0.05	77.02	76.80	76.65
	0.07	0.10	0.10	0.20	76.32	76.15	75.92

Adaptive-margin parameters. Table 16 varies δ_0 and α, whose sum caps the separation the loss can demand. Accuracy is highest at a cap of 0.60 and falls by roughly two points at 0.30 and 0.90, so the term needs enough margin to act but is harmed by demanding near-orthogonality between known-class anchors. Holding the cap fixed and removing the adaptive response ($\alpha = 0$) costs 2.29 points on *All* and 3.40 on *New*, more than halving the benefit of the term; a purely adaptive margin ($\delta_0 = 0$) is better but still below the default, indicating that both a floor and a confusion-dependent response contribute. With $\delta_0 = \alpha = 0$ the loss demands only $s_{y,y'} < 1$ and is effectively inactive, matching the configuration with the margin term removed.

Table 16: **Sensitivity to the adaptive-margin parameters** (PACS, NH $\beta = 0.2$). Values report the mean over four LODO targets across three seeds. δ_0 sets the base margin and α the response to the hardest competing prototype, capping the required separation at $\delta_0 + \alpha$. The upper block evaluates cap scale, the middle redistributes a fixed cap (0.60), and the lower shows degenerate baselines. The shaded row denotes our default configuration.

δ_0	α	$\delta_0+\alpha$	All	Old	New
Varying the cap					
0.05	0.25	0.30	75.82	74.90	76.88
0.10	0.50	0.60	**77.97**	**77.74**	**77.35**
0.15	0.75	0.90	76.14	76.85	74.82
Fixed cap, redistributed					
0.00	0.60	0.60	76.45	75.80	77.10
0.30	0.30	0.60	77.20	77.30	76.92
0.60	0.00	0.60	75.68	76.52	73.95
Degenerate cases					
0.00	0.00	0.00	74.22	71.80	76.55
no margin term ($\lambda_{\text{mar}} = 0$)			74.15	71.62	76.60

I Selection and Sensitivity of the Rejection Threshold

Selection protocol. The threshold τ_{inf} in Eq. equation 8 is chosen without any access to the target domain. After the final communication round, each participating client evaluates a held-out split of its local known-class data against the aggregated global prototypes $\mathcal{C}_T$ and computes local affinity scores. To minimize communication overhead without exposing raw features, each client uploads a lightweight 1D similarity histogram (or local quantile estimates) to the server. The server aggregates these distributions and sets τ_{inf} to the value at which a fixed fraction κ of source known samples are rejected, with $\kappa = 0.05$ across all experiments. This procedure calibrates the boundary strictly on federated source data and never inspects $\mathcal{D}_{\text{test}}$.

Sensitivity. Table 17 evaluates sensitivity as τ_{inf} is varied across a wide range on PACS. As τ_{inf} increases, the two routing error rates cross predictably: novel-to-known leakage drops from 23.85% to 1.90% while known-sample rejection rises from 1.75% to 23.60%. Performance forms a broad, stable plateau between $\tau = 0.40$ and 0.55, with our source-calibrated threshold ($\tau^\star = 0.50$) lagging behind the unconstrained target oracle by merely 0.38% on overall accuracy.

Table 17: **Sensitivity to the rejection threshold** (PACS, NH $\beta = 0.2$, averaged over the four LODO targets). The starred row ($\tau = 0.50$) is selected via the source-side calibration protocol ($\kappa = 0.05$) and corresponds to reported paper numbers. The two error rates represent percentages of their respective pools: $R^{\text{novel}\to\text{known}}$ denotes novel samples falsely accepted as known classes, and $R^{\text{known}\to\text{reject}}$ denotes known samples erroneously rejected into the discovery pool. The oracle row tunes τ directly on the unlabelled target pool to bound source-side selection cost.

τ_{inf}	All	Old	New	$R^{\text{novel}\to\text{known}}$ (%)	$R^{\text{known}\to\text{reject}}$ (%)
0.30	76.25	81.40	66.82	23.85	1.75
0.40	77.42	80.12	72.35	14.70	3.55
0.50*	**77.97**	77.74	77.35	8.15	6.90
0.60	75.58	72.90	80.60	3.90	13.25
0.70	70.35	64.15	82.20	1.90	23.60
oracle ($\tau = 0.48$)	78.35	78.50	76.10	9.40	5.60

J Extended Ablation Analysis

J.1 Effectiveness across Architectures.

As shown in Fig. 9, stability-weighted aggregation (Repeat) consistently outperforms standard averaging (Avg) across GCD, GPC, and HiDISC baselines. By prioritizing clients with spatial consistency under global style perturbations, this mechanism effectively mitigates local domain bias, proving that style-robust updates are a universally beneficial strategy for decentralized shifts regardless of the underlying local objective.

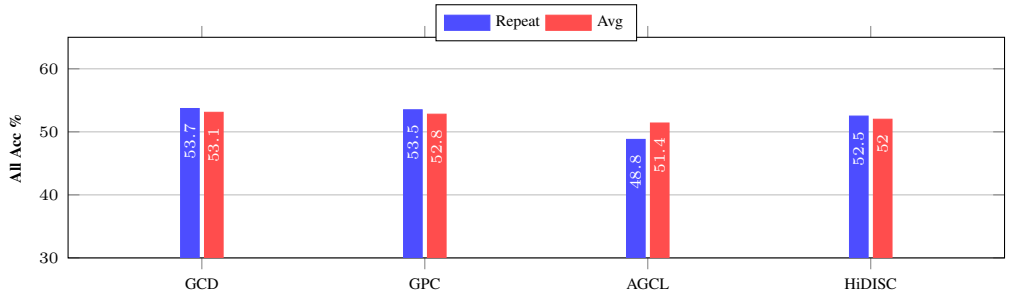

Figure 9: Stability of Repeat vs. Averaging across different baselines.

J.2 Uncertainty Aggregation Analysis

On the PACS (NH) benchmark, FedREPEAT establishes a new state-of-the-art across all metrics, surpassing contemporary uncertainty-aware frameworks. We characterize the reliability of client updates using different proxy metrics: SDEA Huang et al. (2024) employs *predictive entropy* to measure output randomness, UFL Chen et al. (2025) utilizes *MC-Dropout* to quantify model stochasticity, and FedDCG Xu et al. (2026) relies on *similarity scores* to gauge distributional alignment. While these methods demonstrate competitive capability in novel class discovery, our **spatial consistency-driven** estimation yields a significant **+1.52%** improvement in overall accuracy. This advantage confirms that quantifying semantic stability through multi-view variance—measuring how representations fluctuate under spatial transformations—provides a more robust signal for mitigating domain shift than purely statistical or similarity-based proxies.

Table 18: **Aggregation Analysis on PACS (NH).** Evaluation of REPEAT against contemporary uncertainty-aware aggregation strategies. Accuracy (%) is reported for All, Known (Old), and Novel (New) classes. **Bold** and underline denote best and second-best performance, respectively.

	Uncertainty	**PACS (NH) Acc. (%)**		
Method	**Metric**	**All**	**Old**	**New**
SDEA Huang et al. (2024) (PMLR 2024)	Predictive Entropy	75.82	75.10	76.90
UFL Chen et al. (2025) (2025)	MC-Dropout	<u>76.45</u>	<u>76.20</u>	76.85
FedDCG Xu et al. (2026) (ICASSP 2026)	Similarity Score	76.10	75.85	<u>77.05</u>
FedREPEAT	**Spatial Consistency**	**77.97**	**77.74**	**77.35**

K LIMITATIONS

While **NOMAD** successfully bridges the gap between federated learning and zero-shot category discovery, its integrated approach involves certain trade-offs. The use of multi-view forward passes for perturbation-stability aggregation increases the local computational footprint, which may require further optimization for deployment on lower-power edge devices. Finally, the framework's performance remains partially coupled with the accuracy of the initial cluster-count estimation, suggesting that future iterations could benefit from more autonomous cardinality-estimation techniques to further enhance its *in-the-wild* flexibility. These factors represent promising avenues for refinement as federated systems move toward greater autonomy

Our comparison covers representative baselines across GCD, Fed-GCD, FDG, and DG-GCD, but is not exhaustive. As FEDDG-GCD is a new setting, we encourage future work to adapt and evaluate additional methods under the same protocol to further strengthen the benchmark.

L POTENTIAL SOCIETAL IMPACT

Our work studies Federated Domain Generalization for Generalized Category Discovery (FEDDG-GCD), where models learn from decentralized and heterogeneous source clients and are evaluated on an unseen target domain containing both known and novel categories. This setting is relevant to applications such as healthcare, autonomous systems, environmental monitoring, and surveillance, where data may be difficult to centralize and deployment conditions can differ from training environments.

The proposed framework may improve robustness to changing environments while reducing the need to exchange raw client data or repeatedly annotate new categories. However, federated learning alone does not guarantee privacy, and information may still be exposed through model updates or intermediate statistics. Biases and imbalances across clients may also propagate to the global model, while domain shift can amplify spurious correlations and affect novel-category discovery.

Discovered clusters should not be treated as reliable semantic labels without human validation, especially in safety-critical applications. Our experiments use standard vision benchmarks and are intended to study the learning problem rather than support direct deployment. Future work should consider stronger privacy guarantees, fairness evaluation, uncertainty estimation, and human oversight.