LATENT ENTITY NEGOTIATION FOR FEDERATED LEARNING WITH PRIVATE TAXONOMIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated classifiers normally assume that output coordinate j denotes the same class at every client. This interface breaks when institutions retain taxonomy-local labels that independently permute, merge, or conditionally split latent entities; here "private" means unshared identities, not differential privacy. We introduce FEDLENS, an architecture that factors each private classifier into a shared latent-entity bank, a client-private entity-to-word channel, and an input-dependent residual reader. Clients communicate adaptive micro-prototypes without word identities, and the server negotiates anonymous latent entities through mass-aware semi-unbalanced transport with learned capacities. The residual preserves feature-conditioned refinements, while the entity path transfers structure through the encoder even though the default test-time reader is private. We prove vocabulary-permutation invariance, conditional local identifiability under separated-support, pure-component, and correct-capacity assumptions, and conditional encoder convergence under stochastic-gradient local updates with explicit private-state and bank-value variation terms. Across four controlled benchmarks, negotiation improves five-seed mean accuracy by 5.00–7.62 points over the matched residual-only control; gains over the strong native-compatible FedRoD comparator are smaller, at 0.23–1.05 points.

1 INTRODUCTION

Federated learning (FL) permits data owners to learn a common representation without pooling raw examples (McMahan et al., 2017). Classification protocols nevertheless expose a semantic dependency: clients are expected to share an output dimension and class-to-index convention. Hospitals, manufacturers, and annotation teams instead maintain private ontologies whose names and scope can be sensitive. One site may merge two disorders, another may split a disorder using a feature-dependent rule, and a third may use anonymous, permuted codes. Averaging their classifiers is then not merely difficult under non-IID data; it is undefined because equal indices need not denote equal entities.

Personalized FL keeps local heads but shares encoder gradients induced by unrelated partitions (Collins et al., 2021; Chen & Chao, 2022; Oh et al., 2022). Prototype FL exchanges class centroids (Tan et al., 2022), while class-set alignment may use public label names (Zhang et al., 2023). Names are unavailable here; anonymous centroids cannot reveal whether a word is one entity, a merge, or a conditional split. Recent inconsistent-label and private-set formulations (Breitholtz et al., 2026; He et al., 2026) require server-visible label/category resources and cover category subsets or permutations, not anonymous merges and conditional splits. The architectural question is: *what can be shared when neither output coordinates nor label names are comparable?*

Our answer separates client-private *words* from anonymously negotiated *entities*. Figure 1(a) summarizes permutation, merge, and input-dependent split; Figure 1(b) shows clients retaining local words while contributing anonymous evidence. FEDLENS realizes this interface with a shared encoder/entity bank, private entity-to-word channels, and residual readers (Figure 2). Clients upload shuffled weighted micro-prototypes. Mass-aware semi-unbalanced transport preserves source mass and learns nonuniform target capacity; the residual retains site-specific refinements. At test time the default uses only that private residual: latent entities are a training interface for representation transfer, not a shared prediction taxonomy.

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

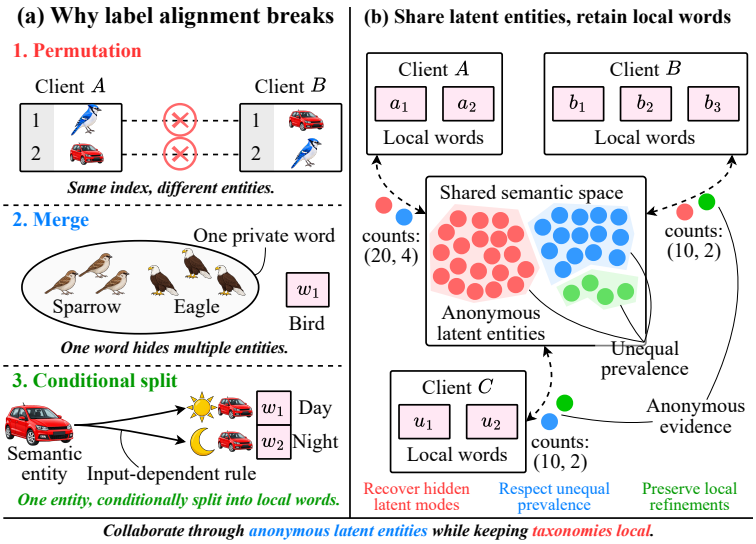


Figure 1: Taxonomy-local interface. (a) Permutations, merges, and input-dependent splits break label-index alignment. (b) FEDLENS negotiates anonymous latent entities while preserving prevalence and keeping words and refinements local. Numbers denote prototype counts.

The design has four consequences. Vocabulary permutations affect only private parameters. The overall channel $A_m(y \mid z, x)$ is input-dependent through the residual reader, covering merges and feature-dependent splits. Capacity $\widehat{K}$ is practitioner-specified; ground-truth $K^\star$ is evaluation-only, and sensitivity covers misspecification. Finally, because the server alternates a barycenter with encoder aggregation, our analysis treats this operator rather than fictitious joint SGD.

Our contributions are: (1) a taxonomy-local FL formulation and vocabulary-invariant latent-entity architecture; (2) adaptive micro-prototypes and mass-aware semi-unbalanced transport with learned capacities; (3) correct-capacity conditional local identifiability plus convergence of the alternating server protocol under stated stochastic-gradient, tracking, and bank-value-variation conditions, without extending exact recovery to the non-oracle budget; and (4) a controlled mechanism study with four benchmarks and 11 comparison mechanisms, including native-compatible methods, anonymous adaptations, a pooled-clustering baseline, a privileged anchor, a matched control, ablations, and stress tests.

2 Related Work

Heterogeneous and personalized FL. FedRep separates representations from personal heads (Collins et al., 2021); FedBABU trains a common body before personalization (Oh et al., 2022); FedRoD couples generic and personalized predictors (Chen & Chao, 2022). FedDisco reweights discrepancy (Ye et al., 2023), FedETF fixes classifier geometry (Li et al., 2023), and FedLWS shrinks layers adaptively (Shi et al., 2025). CrossKT-FRA transfers representations in semi-supervised healthcare VFL (Yin et al., 2024), while calibration and stability-based generalization remain active ICLR topics (Chu et al., 2025; Liu et al., 2025). These methods assume aligned entities or class-compatible supervision; our setting withholds cross-client label correspondence.

Non-identical label spaces. FedProto presumes class identity (Tan et al., 2022); FedAlign anchors sets with public names (Zhang et al., 2023). Private-Set FL shares each client's label set with the server and treats it as a subset of global categories (Breitholtz et al., 2026); CrowdFed uses server-side samples and a global categorical embedding index to match local categories, and evaluates permutation-style representation skew (He et al., 2026). Those interfaces are important restricted cases, but neither is available when the server receives no word identities and one local word may merge entities or conditionally split one entity. FEDLENS instead learns the latent layer from anonymous evidence and analyzes its invariance, identifiability, and alternating optimization.

Transport. Entropic transport supports differentiable matching (Cuturi, 2013); generalized diagonal scaling extends it to relaxed marginals (Chizat et al., 2018). Balanced assignments prevent collapse in centralized self-supervision (Caron et al., 2020), but uniform marginals are inappropriate when prevalence varies. FEDLENS preserves source mass and learns the target marginal.

3 PROBLEM FORMULATION

Client m owns $D_m = \{(x_i, y_i)\}$ with taxonomy-local vocabulary $\mathcal{Y}_m$. The honest-but-curious server receives encoder updates and shuffled prototype/count pairs, and hence observes vector geometry, masses, and message cardinality; it is not explicitly given word identities, $|\mathcal{Y}_m|$, or word–prototype grouping. We provide no differential-privacy or size-hiding guarantee: geometry and counts may reveal taxonomy structure unless combined with padding, secure aggregation, or formal noise. Let $z \in [K^\star]$ be an evaluator-only entity:

$$p_m(y \mid x) = \sum_{z=1}^{K^\star} A_m(y \mid z, x) p(z \mid x). \tag{1}$$

A word shared by several z is a merge; dependence on x permits a semantic entity to be locally split. Training receives no z or A_m. The practitioner supplies capacity $\widehat{K}$ like a representation width; it need not equal $K^\star$.

4 FEDLENS

4.1 OVERVIEW: NEGOTIATION AS A TRAINING INTERFACE

FedLENS does not require every client to deploy the same classifier. Its shared semantic layer is instead a *training interface*: clients negotiate latent entities that shape the common encoder, then decode through an entity path or private residual. Figure 2 exposes one complete round. Clients form anonymous micro-prototypes, consume the returned bank in dual-path decoding, and upload encoder updates plus shuffled vectors and counts. The server performs mass-aware negotiation and volume-tempered encoder aggregation, then broadcasts the updated bank and encoder. The swapped local word names are intentionally noncanonical: their spelling conveys no cross-client semantics.

4.2 SHARED ENTITY PATH AND PRIVATE DECODER

The shared state is an encoder $f_\theta : \mathcal{X} \to \mathbb{S}^{d-1}$ and normalized entity bank $C = [c_1, \ldots, c_{\widehat{K}}]$. Client m retains residual head B_m and word–entity logits Ψ_m. Given $h = f_\theta(x)$, the entity posterior and column-normalized entity-to-word channel are

$$p_\theta(k \mid x) = \mathrm{softmax}_k(c_k^\top h/\tau), \qquad S_m(y \mid k) = \mathrm{softmax}_y(\Psi_m[y, k]). \tag{2}$$

Their composition is $q_m^{\mathrm{ent}}(y \mid x) = \sum_k S_m(y \mid k) p_\theta(k \mid x)$. The unadjusted residual is $q_m^{\mathrm{res}}(y \mid x) = \mathrm{softmax}_y(B_m h)$, and prediction uses

$$q_m(y \mid x) = \rho_m q_m^{\mathrm{res}}(y \mid x) + (1 - \rho_m) q_m^{\mathrm{ent}}(y \mid x). \tag{3}$$

During training, both paths update the shared encoder. At deployment, ρ_m can be chosen locally; our default $\rho_m = 1$ uses the residual reader because it can express arbitrary site-specific refinements. Thus entity negotiation contributes representation shaping rather than requiring a universal inference taxonomy. No vocabulary-shaped tensor is uploaded.

4.3 ANONYMOUS ADAPTIVE MICRO-PROTOTYPES

For word y, client m collects normalized embeddings H_{my}. A sample-size rule proposes $s_{my} = 2, 3,$ or 5 spherical clusters for $|H_{my}| \geq 8, 18,$ or 40, respectively. Farthest-point seeding followed by four cosine k-means steps gives centroids $p_{my\ell}$. We retain the split only when its within-cluster distortion is below 70% of the one-centroid distortion; otherwise the word contributes its single mean. The

client attaches only the raw local count $n_{my\ell}$; after pooling, the server assigns anonymous index j and forms the source marginal:

$$n_{my\ell} = |H_{my\ell}|, \qquad a_j = n_j \Big/ \sum_{j'} n_{j'}. \tag{4}$$

The client uploads the shuffled multiset $\{(p_{my\ell}, n_{my\ell})\}$ without y, an explicit word count, or prototype grouping. Message cardinality can nevertheless reveal bounds on vocabulary scope; our claim is vocabulary invariance, not size privacy. This adaptive decomposition is necessary because a private word may mix several semantic entities. The conservative gain test avoids turning every noisy word into five messages and bounds each word's communication by five 64-dimensional vectors.

4.4 Mass-aware entity negotiation

For pooled prototypes p_j, set $D_{jk} = 1 - p_j^\top c_k^t$. Rather than impose uniform columns, the server solves

$$\Gamma^t = \arg\min_{\Gamma \geq 0,\ \Gamma \mathbf{1} = a} \langle \Gamma, D \rangle + \varepsilon \mathrm{KL}(\Gamma \| a(b^t)^\top) + \lambda_{\mathrm{col}} \mathrm{KL}(\Gamma^\top \mathbf{1} \| b^t). \tag{5}$$

Here a is the measured prototype mass and $b^t \in \Delta^{\widehat{K}}$ is an adaptive entity capacity. Every row is conserved, while columns are only softly capacity-matched. Thus a rare entity is not forced to absorb $1/\widehat{K}$ of the federation.

$$c_k^{t+1} = \frac{\sum_j \Gamma_{jk}^t p_j}{\| \sum_j \Gamma_{jk}^t p_j \|_2}, \qquad b^{t+1} = (1-\beta) b^t + \beta \Gamma^{t\top} \mathbf{1}. \tag{6}$$

We set $\varepsilon = 0.04$ and $\lambda_{\mathrm{col}} = 0.00706$, giving $\gamma = \lambda_{\mathrm{col}}/(\lambda_{\mathrm{col}} + \varepsilon) = 0.15$. For each of 12 barycenter steps, the generalized-Sinkhorn solver (Chizat et al., 2018) forms the full-reference kernel

$$K_{jk} = a_j b_k^t e^{-D_{jk}/\varepsilon}, \quad u \leftarrow a \oslash (Kv), \quad v \leftarrow \left(b^t \oslash K^\top u\right)^\gamma, \quad \Gamma^t = \mathrm{diag}(u) K \, \mathrm{diag}(v). \tag{7}$$

Scaling vectors, rather than an in-place coupling correction, are iterated to fixed-point tolerance; a final u update enforces $\Gamma^t \mathbf{1} = a$. Only after solving Equation 5 do we update $b^{t+1} = 0.7 b^t + 0.3 \Gamma^{t\top} \mathbf{1}$. Appendix A.4 derives these updates, gives pseudocode, and validates the implementation against a generic convex optimizer. Bank initialization uses anonymous farthest-point seeding. The default $\widehat{K} = \max(8, 2M)$ depends only on public client count M; it never reads $K^\star$, a private vocabulary size, or canonical labels. We test 0.75–1.5 times this value without retuning.

4.5 Objective and protocol

The private channel logits $\Psi_m \in \mathbb{R}^{|\mathcal{Y}_m| \times \widehat{K}}$ define $S_m(y \mid k)$ in Equation 2 and $R_m(k \mid y) = \mathrm{softmax}_k(\Psi_m[y,k])$. At round t, client m computes $\bar{p}_{my}$ on its training samples using the broadcast encoder θ^t; geometry gives $G_m^t[y,k] = \bar{p}_{my}^\top c_k^t / \tau$. The initialized map is $\bar{\Psi}_m^t = G_m^0$ for $t = 0$, otherwise $0.65 \Psi_m^{t-1} + 0.35 G_m^t$. Local optimization fixes (C^t, b^t); afterward, new micro-prototypes from the local endpoint form the $t{+}1$ bank. Let π_m be the smoothed word frequency, $\widetilde{q}_m^{\mathrm{res}} = \mathrm{softmax}(B_m h + \log \pi_m)$, and $u_m = \mathbb{E}_x p_\theta(\cdot \mid x)$. The implemented local objective is

$$\begin{aligned} \mathcal{L}_m = {} & \mathrm{CE}(\widetilde{q}_m^{\mathrm{res}}, y) + \lambda_{\mathrm{lex}} \mathrm{CE}(q_m^{\mathrm{ent}}, y) + \lambda_e \mathrm{CE}(\mathrm{sg}[R_m(\cdot \mid y)], p_{\theta, \mathrm{sg}[C^t]}(\cdot \mid x)) \\ & + \lambda_h \frac{1}{\widehat{K}} \sum_k H(S_m(\cdot \mid k)) + \lambda_{\mathrm{use}} \mathrm{KL}(u_m \| b^t) + \lambda_c \| \Psi_m - \bar{\Psi}_m^t \|_F^2. \end{aligned} \tag{8}$$

We use $(\lambda_{\mathrm{lex}}, \lambda_e, \lambda_h, \lambda_{\mathrm{use}}, \lambda_c) = (0.01, 0.002, 0.02, 0.001, 0.03)$. Here sg is stop-gradient: R_m is the geometric teacher and the round-t entity bank C^t is fixed during local optimization, so the entity CE differentiates through the posterior only into the encoder θ. The private map Ψ_m is instead learned by the lexical, entropy, and proximal terms, avoiding target–prediction co-adaptation; C^{t+1} is produced only by the server barycenter. The first term is genuine logit adjustment rather than a constant added outside the softmax. Channel entropy prevents diffuse word decoders; capacity matching suppresses collapse without imposing uniform use; and the proximal cycle term discourages abrupt round-to-round map changes. All arguments of entropy and KL are normalized distributions.

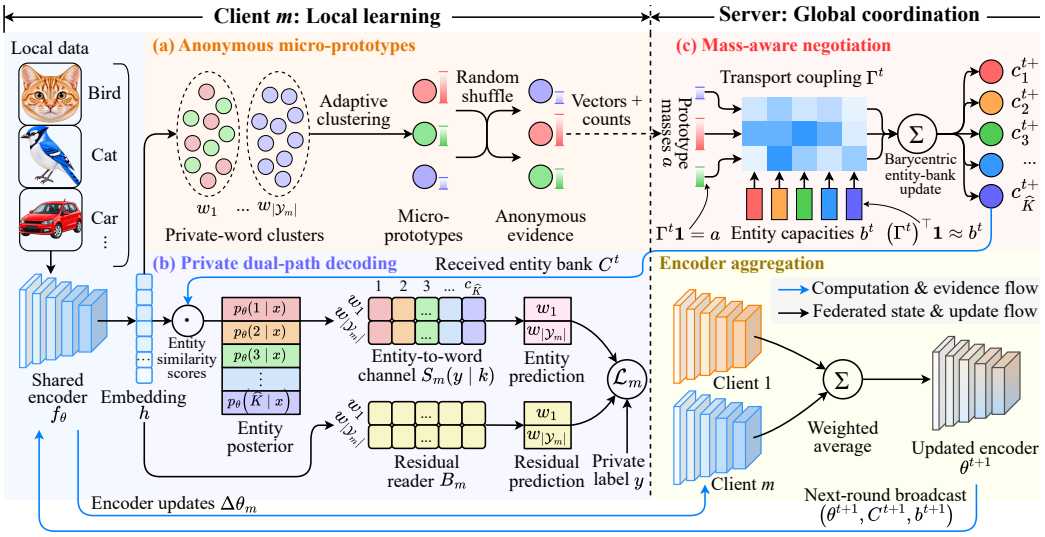


Figure 2: One FedLENS round. Clients hold (C^t, b^t) fixed during local learning, then upload encoder updates and shuffled prototype/count pairs. The server solves for Γ^t, updates $(\theta^{t+1}, C^{t+1}, b^{t+1})$, and broadcasts them only to the next round. Posterior entries are probabilities $p_\theta(k \mid x)$, not entity vectors c_k.

Algorithm 1 FEDLENS training

Require: clients, rounds T, local epochs E; set $\widehat{K} = \max(8, 2M)$

initialize trainable encoder θ^0 with the common random initialization; set $b^0 = \mathbf{1}/\widehat{K}$
clients form initial anonymous prototypes under f_{θ^0}; server farthest-point seeds C^0
for $t = 0, \ldots, T-1$ **do**
  for all selected clients m **in parallel do**
    under θ^t, compute $\bar{p}_{my}$ and $\bar{\Psi}^t_m$; hold (C^t, b^t) fixed; optimize Equation 8
    adaptively split each word using Equation 4
    upload $\Delta\theta_m$ and a shuffled prototype/count multiset
  end for
  $\theta^{t+1} \leftarrow \sum_m \sqrt{|D_m|}\theta_m / \sum_m \sqrt{|D_m|}$
  solve Equation 5 by Equation 7; update (C^{t+1}, b^{t+1}) by Equation 6
end for
clients refit residual heads for two epochs; keep (B_m, Ψ_m) private

After E local epochs, server encoder aggregation uses $w_m \propto \sqrt{|D_m|}$. We treat this volume-tempered weighting as a fixed engineering choice, not a contribution; Appendix B.5 compares it directly with standard sample weighting. The entity bank is not averaged from client-shaped heads: it is replaced by Equations 6–7. At deployment the default $\rho_m = 1$ reads the private residual, while the entity path has already shaped the encoder throughout training. For a controlled comparison, every baseline with a trainable private head receives the same two-epoch final head refit; FedETF retains its fixed head by design.

5 EXPERIMENTS

5.1 PROTOCOL AND BASELINES

We use MNIST (LeCun et al., 1998), CIFAR-10, CIFAR-100 coarse, and **CIFAR-100-Mixed** (Krizhevsky, 2009). The latter is a CIFAR-100 native-hierarchy mixed-granularity benchmark: clients retain coarse, fine, or hybrid native labels, then independently permute their word indices. Canonical fine identities are used only for benchmark construction and evaluator-side metrics; they never enter optimization inputs, server messages, or federation updates. Across benchmarks we

independently vary Dirichlet quantity skew, missing entities, many-to-one merges, and feature-conditioned splits; Appendix B.1 gives every default severity parameter and the exact taxonomy generator used for Table 1. CIFAR images use frozen ImageNet ResNet-18 features; all methods train the same 128–64 MLP projection. We fix 12 full-participation rounds, one local epoch, batch 256, and identical AdamW budgets; Appendix A analyzes the alternating server operator using an SGD-style local-update abstraction rather than claiming an AdamW convergence theorem.

We compare exactly **11 baselines**: FedRep (Collins et al., 2021), FedBABU (Oh et al., 2022), FedRoD (Chen & Chao, 2022), FedProto-style (Tan et al., 2022), FedETF-style (Li et al., 2023), FedDisco-style (Ye et al., 2023), FedAlign-style (Zhang et al., 2023), FedLWS (Shi et al., 2025), OracleAlign, AnonKMeans, and ResidualOnly. The "-style" suffix explicitly marks an anonymous-vocabulary adaptation rather than a faithful execution under unavailable shared labels or names. AnonKMeans is a simple anonymous baseline that pools shuffled private-word prototypes and applies nearest-centroid attraction without transport or learned mass. OracleAlign receives evaluator semantic identities and is a privileged semantic reference, not a utility upper bound or an anonymous competitor. ResidualOnly is the matched control used to isolate latent negotiation while holding the federated optimization pipeline, private reader, aggregation, and residual refit fixed.

Closest-method compatibility. Private-Set FL sends client label sets to the server and assumes global-category subsets; CrowdFed requires server-side samples and a global categorical index, and evaluates permutation-style skew. These interfaces cannot represent anonymous merges or conditional splits without information unavailable in our setting. We therefore do not relabel either as a full-setting baseline; comparison on a shared restricted benchmark remains future work.

The four adapted families retain respectively anonymous prototype attraction, fixed private classifier geometry, local discrepancy weighting, and name-free representation alignment; their labels never imply faithful shared-class reproduction. All trainable private heads receive exactly two final refit epochs; FedETF-style remains fixed. The default $\rho = 1$ is applied consistently: latent negotiation is the training interface, not a privileged test-time ensemble.

Primary endpoints are mean and worst-client private-label accuracy. Evaluator-only NMI measures representation semantics; ARI and recovery are secondary. We report five paired seeds for FEDLENS, FedRoD, and ResidualOnly, and three fixed seeds for the other methods. This concentrates added computation on the strongest native-compatible comparator and matched mechanism control while preserving the complete 12-method grid. Table 2 reports paired effect sizes and 95% intervals; these intervals are not multiplicity-adjusted significance claims. Bold/underline denote best/second observed means only.

5.2 MAIN RESULTS

Table 1 reports the designated primary endpoints. FEDLENS has the highest observed mean in all eight cells. Across MNIST, CIFAR-10, CIFAR-100, and Mixed, its five-seed means are $89.19/80.74$, $76.54/57.16$, $61.65/48.32$, and $47.31/32.04$ for mean/worst accuracy. The mean-accuracy gains over five-seed FedRoD are 1.05, 0.84, 0.51, and 0.23 points; gains over ResidualOnly are 5.13, 5.00, 5.81, and 7.62 points. We therefore distinguish a modest improvement over the strongest personalized comparator from the larger matched-control evidence for negotiated representation transfer.

The FedRoD comparison shows small, consistently positive mean effects and uncertain tail effects; we do not claim broad statistical superiority over this strong comparator. The materially larger ResidualOnly effects isolate the value of latent negotiation within the otherwise matched pipeline. Accordingly, Table 2, rather than aggregate ranks, is the primary quantitative evidence.

Every comparative panel includes all 12 methods. Figure 3(a) separates early representation transfer from late head fitting, while (c)–(d) expose dispersion rather than only means; panel (e) is descriptive and does not replace the paired effects in Table 2. Mixed fine-label NMI is not a headline because coarse clients are never trained to resolve all fine descendants; Appendix B.6 reports the complete common-seed semantic comparison.

Table 1: Taxonomy-local utility (%). Five seeds for FEDLENS, FedRoD, and ResidualOnly; three otherwise. "-style" marks anonymous-vocabulary adaptations; OracleAlign is privileged.

Method	MNIST		CIFAR-10	
	Mean ↑	Worst ↑	Mean ↑	Worst ↑
FedRep	83.85 ± 1.78	69.85 ± 4.08	69.93 ± 4.43	44.58 ± 7.61
FedBABU	83.31 ± 1.93	68.06 ± 4.54	69.15 ± 4.40	43.39 ± 7.67
FedRoD	88.14 ± 2.10	77.52 ± 4.46	75.70 ± 3.00	53.48 ± 6.93
FedProto-style	81.85 ± 1.90	63.37 ± 3.94	67.43 ± 4.88	42.72 ± 7.26
FedETF-style	73.96 ± 10.67	44.21 ± 34.06	63.62 ± 8.68	33.26 ± 18.26
FedDisco-style	83.54 ± 1.90	69.54 ± 4.36	69.31 ± 4.52	44.05 ± 8.38
FedAlign-style	80.65 ± 2.59	62.45 ± 3.77	66.54 ± 4.83	43.08 ± 7.75
FedLWS	83.35 ± 1.84	68.59 ± 4.64	69.33 ± 4.54	44.23 ± 8.39
OracleAlign	85.73 ± 1.34	72.82 ± 4.27	71.79 ± 3.58	50.43 ± 5.57
AnonKMeans	79.01 ± 2.35	59.36 ± 4.65	65.94 ± 2.91	47.47 ± 3.11
ResidualOnly	84.06 ± 2.58	71.91 ± 5.75	71.54 ± 3.48	47.70 ± 4.02
FedLENS	**89.19 ± 1.76**	**80.74 ± 3.24**	**76.54 ± 2.85**	**57.16 ± 5.46**

Method	CIFAR-100		C100-Mixed	
	Mean ↑	Worst ↑	Mean ↑	Worst ↑
FedRep	53.25 ± 1.90	35.35 ± 4.99	37.99 ± 1.10	24.65 ± 0.99
FedBABU	52.05 ± 1.71	34.94 ± 5.42	36.80 ± 1.31	21.67 ± 1.85
FedRoD	61.14 ± 3.77	46.98 ± 7.70	47.08 ± 1.48	30.86 ± 2.58
FedProto-style	49.07 ± 1.83	33.89 ± 7.80	25.59 ± 2.75	13.28 ± 3.08
FedETF-style	45.39 ± 0.88	17.27 ± 8.29	32.10 ± 1.37	15.63 ± 4.08
FedDisco-style	52.42 ± 1.95	35.72 ± 4.79	37.21 ± 1.39	23.11 ± 2.20
FedAlign-style	46.57 ± 1.95	31.52 ± 7.37	21.80 ± 2.24	10.13 ± 4.39
FedLWS	52.18 ± 1.93	34.94 ± 5.42	36.69 ± 1.44	21.79 ± 1.47
OracleAlign	56.77 ± 2.78	41.43 ± 7.41	43.73 ± 1.52	29.73 ± 5.23
AnonKMeans	47.71 ± 1.79	31.46 ± 8.04	29.02 ± 2.34	15.17 ± 3.78
ResidualOnly	55.85 ± 2.05	38.87 ± 5.46	39.69 ± 0.98	25.41 ± 2.09
FedLENS	**61.65 ± 3.53**	**48.32 ± 7.45**	**47.31 ± 1.45**	**32.04 ± 2.51**

Table 2: Paired utility gain of FEDLENS (percentage points; 95% CI, five seeds).

Dataset	Comparator	Mean accuracy Δ	Worst-client Δ
MNIST	FedRoD	**+1.05** [+0.14, +1.96]	**+3.22** [-0.44, +6.87]
MNIST	ResidualOnly	**+5.13** [+3.93, +6.32]	**+8.83** [+1.50, +16.17]
CIFAR-10	FedRoD	**+0.84** [+0.00, +1.68]	**+3.67** [+0.02, +7.33]
CIFAR-10	ResidualOnly	**+5.00** [+2.88, +7.13]	**+9.46** [+1.34, +17.57]
CIFAR-100	FedRoD	**+0.51** [+0.19, +0.83]	**+1.34** [-0.78, +3.46]
CIFAR-100	ResidualOnly	**+5.81** [+3.61, +8.00]	**+9.45** [+3.09, +15.82]
C100-Mixed	FedRoD	**+0.23** [+0.04, +0.42]	**+1.18** [-0.09, +2.45]
C100-Mixed	ResidualOnly	**+7.62** [+6.15, +9.10]	**+6.63** [+3.44, +9.82]

5.3 ABLATION AND MECHANISM EVIDENCE

Table 3 removes or replaces each design choice on the flagship native-hierarchy task; Appendix B.3 gives the matched CIFAR-100 table and targeted structural probes. Removing the residual causes the expected collapse under locally refined labels. The full model has the strongest observed worst-client accuracy, while single-centroid and pooled-k-means variants are statistically close on mean accuracy and NMI. We do not claim that every module improves every aggregate metric: the targeted probes instead show where adaptive splitting and mass-aware transport matter, and the near-ties delimit the empirical value of added complexity. Figure 4 evaluates all methods under increasing Dirichlet skew, merged-entity fraction, and conditional-split fraction. Panel (a) reports a transparent within-condition

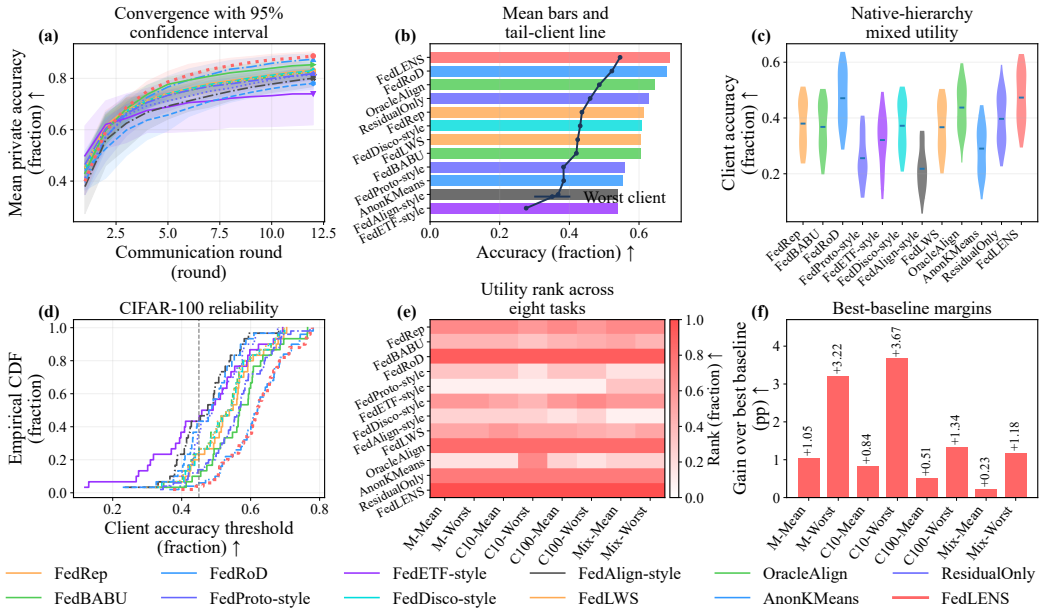


Figure 3: Main diagnostics with all 12 methods: (a) convergence with 95% CIs; (b) mean/tail utility; (c) Mixed client distributions; (d) accuracy ECDF; (e) rank over eight primary tasks; (f) FEDLENS margin over the strongest non-oracle baseline. Four "-style" methods are anonymous adaptations; OracleAlign is privileged.

Table 3: Complete C100-Mixed primary-task ablation (three paired seeds). Upload is the incremental shuffled prototype/count payload; bold/underline mark best/second observed means among variants.

Variant	Mean acc. ↑	Worst acc. ↑	NMI ↑	Upload (MB/round) ↓
Full	47.78 ± 1.64	**33.60 ± 1.87**	52.81 ± 0.10	3.82 ± 0.01
Single centroid	**48.04 ± 1.62**	33.32 ± 2.21	52.93 ± 0.33	**3.70 ± 0.01**
Fixed capacity	47.85 ± 1.61	32.95 ± 2.04	52.86 ± 0.36	3.82 ± 0.01
No entity path	47.49 ± 1.60	32.67 ± 3.14	52.72 ± 0.14	3.82 ± 0.01
No cycle	47.72 ± 1.67	33.00 ± 1.69	**52.98 ± 0.35**	3.82 ± 0.01
Sample weighting	47.88 ± 1.39	32.16 ± 2.31	52.96 ± 0.08	3.82 ± 0.01
Pooled k-means	47.92 ± 1.63	33.57 ± 1.64	52.82 ± 0.20	3.82 ± 0.01
No residual	10.64 ± 3.47	3.57 ± 2.26	45.74 ± 1.45	3.88 ± 0.02

score combining semantics with mean/tail utility; panels (b)–(c) retain the unnormalized primary units. The quantile and success-rate panels test distributional robustness, while panel (f) prevents cherry-picking by ranking every method over the full stress grid. FEDLENS is the top observed method on these stress summaries. Appendix B.7 varies temperature, entropy weight, ρ, and $\widehat{K}$ without retuning. The $\rho = 0$ endpoint deliberately degrades under splits, confirming that the residual is an expressivity component rather than a cosmetic ensemble. Budget over-specification is safer than under-specification because unused columns can lose capacity instead of forcing separated entities together.

5.4 SCOPE AND VALIDITY

The study is a controlled mechanism analysis, not evidence of institutional deployment. MNIST/CIFAR taxonomies are synthetic; C100-Mixed uses a native hierarchy but assigns granularities to simulated clients. Frozen features isolate taxonomy negotiation but do not establish end-to-end scaling or transfer to naturally evolved ontologies. Shuffling gives vocabulary invariance,

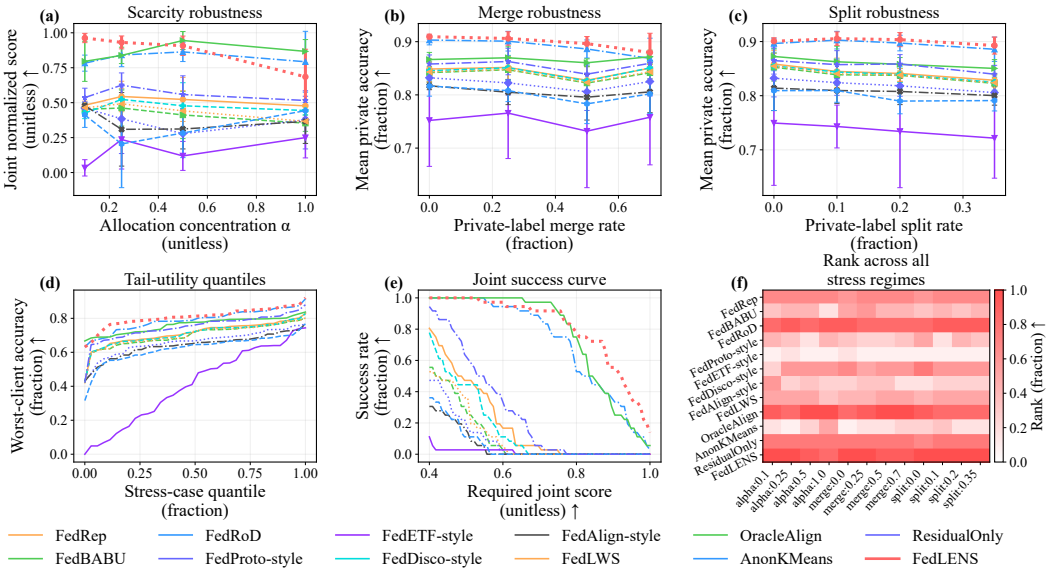


Figure 4: Stress diagnostics with all 12 methods: (a) joint scarcity score; (b) merges; (c) conditional splits; (d) tail quantiles; (e) joint success curves; (f) stress rank. The unitless score weights normalized NMI by $1/2$ and normalized mean/tail utility by $1/4$ each; "-style" marks anonymous adaptations and OracleAlign is privileged.

not differential privacy; message counts and masses can leak vocabulary scope. Semantic scores use evaluator-only matching and never enter training. Healthcare and manufacturing therefore motivate the interface rather than constitute validated application claims.

5.5 Mechanistic interpretation

The results separate representation negotiation from local decision rules. ResidualOnly controls for the encoder, optimizer, logit adjustment, volume-tempered aggregation, private reader, and final refit; its consistent mean-accuracy deficit therefore localizes the remaining gain to the negotiated entity pathway during representation learning. Conversely, the entity-only $\rho = 0$ sensitivity endpoint collapses under conditional splits, showing why the shared semantics cannot replace the private reader. The two paths are complementary rather than interchangeable: the entity bank supplies cross-client structure, while the residual supplies client-conditioned resolution.

Merge, split, and scarcity stress the micro-prototype, residual, and capacity mechanisms, respectively. Targeted interventions localize the corresponding failure modes; aggregate rank is only a descriptive summary. This is mechanism-consistent evidence in the controlled simulator, not a claim that every component dominates every homogeneous-label setting.

6 Conclusion

We introduced taxonomy-local FL and FedLENS, which negotiates entities without shared label identities. Controlled experiments show large matched-control gains, smaller positive effects over FedRoD, and targeted mechanism evidence. The appendix proves invariance and conditional local recovery, and gives convergence with explicit private-state and bank-value-variation conditions. The evidence remains controlled, software-only, and non-DP.

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

REFERENCES

Adam Breitholtz, Edvin Listo Zec, and Fredrik D. Johansson. Federated learning with heterogeneous and private label sets. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, volume 2841 of *Communications in Computer and Information Science*, pp. 472–491. Springer Nature Switzerland, 2026. doi: 10.1007/978-3-032-19102-1_29.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020.

Hong-You Chen and Wei-Lun Chao. On bridging generic and personalized federated learning for image classification. In *ICLR*, 2022.

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018. doi: 10.1090/mcom/3303.

Yun-Wei Chu, Dong-Jun Han, Seyyedali Hosseinalipour, and Christopher G. Brinton. Unlocking the potential of model calibration in federated learning. In *ICLR*, 2025.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *ICML*, 2021.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, 2013.

Yunlong He, Fei Chen, Hanlin Zhang, and Jia Yu. Crowdsourced federated learning with inconsistent label representation. *Information Fusion*, 131:104194, 2026. doi: 10.1016/j.inffus.2026.104194.

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

Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Zexi Li, Xinyi Shang, Rui He, Tao Lin, and Chao Wu. No fear of classifier biases: Neural collapse inspired federated learning with synthetic and fixed classifier. In *ICCV*, 2023.

Yingqi Liu, Qinglun Li, Jie Tan, Yifan Shi, Li Shen, and Xiaochun Cao. Understanding the stability-based generalization of personalized federated learning. In *ICLR*, 2025.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, 2017.

Jaehoon Oh, Sangmook Kim, and Se-Young Yun. FedBABU: Toward enhanced representation for federated image classification. In *ICLR*, 2022.

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

Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. FedProto: Federated prototype learning across heterogeneous clients. In *AAAI*, 2022.

Rui Ye, Mingkai Xu, Jianyu Wang, Chenxin Xu, Siheng Chen, and Yanfeng Wang. FedDisco: Federated learning with discrepancy-aware collaboration. In *ICML*, 2023.

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

Jiayun Zhang, Xiyuan Zhang, Xinyang Zhang, Dezhi Hong, Rajesh K. Gupta, and Jingbo Shang. Navigating alignment for non-identical client class sets: A label name-anchored federated learning framework. In *KDD*, pp. 3297–3308, 2023.

A Detailed Theory and Proofs

This section analyzes the alternating server operator used by FedLENS under an SGD-style abstraction of local encoder optimization: stochastic local encoder updates are followed by a non-gradient transport-barycenter replacement. Experiments use AdamW; the results below do not model its moment states, coordinate-wise preconditioning, or decoupled weight decay and therefore are not an AdamW convergence theorem. The identifiability theorem characterizes the correctly specified case $\widehat{K} = K^\star$; the practical non-oracle rule is evaluated under misspecification and is not claimed to satisfy exact semantic-recovery conditions. The results are local because a latent mixture is not globally identifiable without an initialization or anchor condition.

A.1 Notation and assumptions

Let ν_{mk} denote the population distribution of pure normalized micro-prototypes generated by entity k at client m, and let $a_{mk} > 0$ when the entity is observed. For two banks, define the permutation-invariant distance $d(C, C') = \min_{P \in \mathcal{P}} \|C - C'P\|_F$. Let $\mathcal{B}_r(C^\star) = \{C : d(C, C^\star) \le r\}$.

A1 (separated supports). There are unit population directions $C^\star = (c_1^\star, \dots, c_{K^\star}^\star)$ such that $\mathrm{supp}(\nu_{mk}) \subseteq B(c_k^\star, r)$ and $\|c_k^\star - c_\ell^\star\|_2 \ge \Delta > 4r$ for $k \ne \ell$.

A2 (observable incidence). Adaptive splitting returns at least one positive-mass pure component for every locally observed entity, and the bipartite client–entity incidence graph is connected. Entity masses satisfy $b_k^\star \ge b_{\min} > 0$.

A3 (local initialization and transport accuracy). $\widehat{K} = K^\star$ and $C^0 \in \mathcal{B}_{r_0}(C^\star)$ for $r_0 < (\Delta - 4r)/4$. Within this compact basin, $b_k^t, q_k^t \ge b_{\min}$ for the capacity and exact coupling column marginal $q^t = \Gamma^{t\top}\mathbf{1}$, so $M_b = \sup_{t,k} |\log(q_k^t / b_k^t)| < \infty$. The generalized Sinkhorn solver has additive coupling error δ_t, while empirical prototypes contribute error $\xi_t = O_p(N_t^{-1/2})$.

A4 (profiled encoder regularity). Write the private state as $\phi_m = (B_m, \Psi_m)$ and let $\bar{\phi}_m(\theta, C)$ be a locally stable stationary response of the private objective with C fixed throughout each client local phase. The profiled population objective $\widetilde{\mathcal{F}}(\theta, C) = \sum_m \omega_m \mathbb{E}\mathcal{L}_m(\theta, \bar{\phi}_m(\theta, C), C)$ is lower bounded and L-smooth in θ, with $\omega_m = \sqrt{n_m} / \sum_j \sqrt{n_j}$. Stochastic-gradient variance is bounded by σ^2, weighted client drift by ζ^2, and encoder gradients are L_C-Lipschitz in C and L_ϕ-Lipschitz in private state.

A5 (private-state tracking). During local step e of round t, the jointly updated private state $\phi_{m,e}^t$ has conditional tracking error

$$r_t^2 = \frac{1}{qE} \sum_{m \in S_t} \sum_{e<E} \mathbb{E}\|\phi_{m,e}^t - \bar{\phi}_m(\theta^t, C^t)\|^2 < \infty. \tag{9}$$

This term is not assumed to vanish unless the private optimizer stabilizes; it explicitly measures the bias introduced by updating (B_m, Ψ_m) together with the encoder.

A6 (bank-replacement value variation). The transport barycenter is not a gradient step on $\widetilde{\mathcal{F}}$. We therefore separately define its conditional positive objective variation

$$v_t = \mathbb{E}_t\left[\left(\widetilde{\mathcal{F}}(\theta^{t+1}, C^{t+1}) - \widetilde{\mathcal{F}}(\theta^{t+1}, C^t)\right)_+\right], \qquad \bar{v}_T = T^{-1}\sum_{t<T} v_t, \tag{10}$$

and assume $v_t < \infty$ in the local basin. Exact stationarity will require $\bar{v}_T = o(\eta)$; otherwise the result retains the explicit neighborhood term $\bar{v}_T/(\eta E)$. This is an objective-value condition on bank replacement, not a consequence of the encoder-gradient Lipschitz condition in A4.

Assumptions A1–A3 are sufficient conditions, not claims about every neural representation. A2 assumes that adaptive splitting has already exposed positive-mass pure components, which is precisely one difficult empirical step rather than a consequence of the theorem. A3 is the basin condition required by permutation-symmetric latent models; farthest-point initialization is only a heuristic that seeks this basin. In particular, C100-Mixed uses $\widehat{K} = \max(8, 2M)$ and does not generally satisfy $\widehat{K} = K^\star$; Theorem 1 is therefore not invoked as an exact-recovery explanation for that benchmark.

A.2 Vocabulary invariance

Proposition 1 (Vocabulary-permutation invariance)**.** *For arbitrary permutations P_m of the private vocabularies, replacing each y_m by $P_m y_m$ and applying P_m to the rows of (B_m, Ψ_m) leaves the distribution of every uploaded message and every server iterate (θ^t, C^t, b^t) unchanged.*

Proof. Residual cross-entropy and the entity loss are equivariant to a common permutation of targets and output rows. $S_m(y \mid k)$ and $R_m(k \mid y)$ are permuted in the word dimension, so their scalar entropy, KL, and consistency penalties are unchanged. The word-conditioned sample sets are merely renamed; Equation 4 therefore produces the same multiset of centroids and masses. Shuffling removes the remaining order. Encoder updates coincide after summing over minibatches. Induction over rounds gives identical server inputs, transport coupling, bank, capacity, and aggregated encoder. Only private row order changes. □

This result concerns relabeling invariance. It does not claim differential privacy: prototype values may still leak information, which motivates secure aggregation or formal noise in future work.

A.3 Semi-unbalanced transport

Lemma 1 (Well-posed coupling)**.** *For strictly positive a, b, ε, and λ_{col}, Equation 5 has a unique coupling Γ. Its row marginal is exactly a; its column marginal is positive but need not equal b.*

Proof. The row-feasible polytope is nonempty and compact. The linear cost is continuous, while relative entropy against the full-support reference $ab^\top$ is lower-semicontinuous and strictly convex on the feasible affine set. The column KL term is convex, hence the sum has a unique minimizer. The equality $\Gamma\mathbf{1} = a$ is a constraint; no corresponding column equality appears. □

A.4 Generalized-Sinkhorn derivation and solver audit

The implementation solves exactly the objective in Equation 5; in particular, the target-capacity factor b_k is retained in the kernel because the column marginal is penalized rather than fixed. Define $K = ab^\top \odot \exp(-D/\varepsilon)$, $r = K^\top u$, and $\gamma = \lambda_{\mathrm{col}}/(\lambda_{\mathrm{col}} + \varepsilon)$.

Proposition 2 (Scaling form of the semi-unbalanced optimum)**.** *The unique minimizer of Equation 5 has $\Gamma = \mathrm{diag}(u) K \,\mathrm{diag}(v)$, where the positive scaling vectors satisfy*

$$u = a \oslash (Kv), \qquad v = (b \oslash K^\top u)^\gamma. \tag{11}$$

Consequently $\Gamma\mathbf{1} = a$ exactly, while $q = \Gamma^\top\mathbf{1}$ is generally different from b.

Proof. Let α_j be the multiplier for row j. With generalized KL defined by $\mathrm{KL}(x\|y) = \sum_i [x_i \log(x_i/y_i) - x_i + y_i]$, stationarity at the positive optimum gives

$$D_{jk} + \varepsilon \log \frac{\Gamma_{jk}}{a_j b_k} + \lambda_{\mathrm{col}} \log \frac{q_k}{b_k} + \alpha_j = 0.$$

Absorbing $\exp(-\alpha_j/\varepsilon)$ into u_j yields $\Gamma_{jk} = u_j K_{jk} v_k$ with $v_k = (b_k/q_k)^{\lambda_{\mathrm{col}}/\varepsilon}$. The row constraint gives $u = a \oslash (Kv)$. Since $q_k = v_k(K^\top u)_k$, substitution and collection of the power of v_k gives $v_k = [b_k/(K^\top u)_k]^\gamma$. Strict convexity from Lemma 1 makes this stationary point the unique optimum. □

Algorithm 2 Semi-unbalanced transport used in every server update

Require: $a, b, D, \varepsilon, \lambda_{\mathrm{col}}$; tolerance $\eta_{\mathrm{fp}} = 2 \times 10^{-7}$; maximum $R = 200$
stabilize each row of $-D/\varepsilon$ by subtracting its maximum
$K \leftarrow ab^\top \odot \exp(-D/\varepsilon)$; $v \leftarrow \mathbf{1}$; $\gamma \leftarrow \lambda_{\mathrm{col}}/(\lambda_{\mathrm{col}} + \varepsilon)$
for $r = 1, \ldots, R$ **do**
  $u \leftarrow a \oslash (Kv)$
  $v^+ \leftarrow \big(b \oslash (K^\top u)\big)^\gamma$
  if $\| \log v^+ - \log v \|_\infty \le \eta_{\mathrm{fp}}$ **then**
    $v \leftarrow v^+$; **break**
  end if
  $v \leftarrow v^+$
end for
$u \leftarrow a \oslash (Kv)$ ▷ final exact-row correction
$\Gamma \leftarrow \mathrm{diag}(u) K \, \mathrm{diag}(v)$; return Γ and residuals

We independently checked a fixed 3×3 positive instance against a generic constrained SLSQP optimizer. The solver attained objective 0.13430602636 versus 0.13430602658 (absolute gap 2.2×10^{-10}), coupling ℓ_1 difference 5.7×10^{-8}, exact-row residual 0, and fixed-point residual 6.0×10^{-14}. Its column deviation $\|q - b\|_1 = 0.60$ is intentionally nonzero because columns are KL-penalized, not constrained. Across all extended runs, the maximum observed row residual was 5.33×10^{-8} and the maximum fixed-point residual was 1.79×10^{-7}.

For unit vectors, cosine cost obeys $1 - u^\top v = \|u - v\|_2^2/2$. Under A1 and A3, a prototype is at most $r + r_0$ from its correct initialized column and at least $\Delta - r - r_0$ from any incorrect column. Define the cost gap

$$g = \tfrac{1}{2}\big[(\Delta - r - r_0)^2 - (r + r_0)^2\big] > 0. \tag{12}$$

Lemma 2 (Cross-entity leakage)**.** *Under A1–A3, the mass assigned outside the correct entity satisfies*

$$\mathsf{Leak}(\Gamma^t) \le b_{\min}^{-1} e^{-g/\varepsilon} + c_1 \lambda_{\mathrm{col}}/\varepsilon + c_2(\delta_t + \xi_t). \tag{13}$$

for constants c_1, c_2 independent of t.

Proof. Set $\lambda_{\mathrm{col}} = 0$ first. Rowwise entropic optimality and Equation 12 bound the wrong-to-correct mass ratio by $b_{\min}^{-1} e^{-g/\varepsilon}$. Restoring the column term perturbs the row-entropic objective by a gradient of size at most $\lambda_{\mathrm{col}} M_b$. Because its strong-convexity modulus is proportional to ε, standard minimizer sensitivity contributes $O(\lambda_{\mathrm{col}} M_b/\varepsilon)$, absorbed into $c_1 \lambda_{\mathrm{col}}/\varepsilon$. Approximate Sinkhorn and empirical support add $O(\delta_t + \xi_t)$ by the same stability argument. This direction is also consistent with the limiting cases: removing the column penalty cannot force cross-entity mass, whereas an increasingly hard mismatched marginal can. □

The lemma explains why a learned column marginal matters. With a fixed uniform column constraint, prevalence mismatch may force nonzero cross-entity mass even as $\varepsilon \to 0$; Equation 5 pays a finite KL penalty instead.

A.5 Local identifiability

Theorem 1 (Conditional local identifiability)**.** *Under A1–A3, suppose $e^{-g/\varepsilon}/b_{\min} + \lambda_{\mathrm{col}}/\varepsilon + \delta_t + \xi_t$ is smaller than the basin margin. Then the population transport-barycenter operator maps $\mathcal{B}_{r_0}(C^\star)$ into itself. Every limit fixed point C^∞ that remains in this basin satisfies, up to one global permutation,*

$$d(C^\infty, C^\star) = O\left(r + e^{-g/\varepsilon}/b_{\min} + \lambda_{\mathrm{col}}/\varepsilon + \sup_t(\delta_t + \xi_t)\right). \tag{14}$$

Proof. Lemma 2 implies that each column barycenter is dominated by prototypes from its corresponding support. A normalized positive combination of vectors within radius r stays within that support up to the leakage perturbation, proving basin invariance. Positive entity mass activates every support. Whenever one client observes multiple entities, within-client geometry fixes their relative assignment; connectedness of the incidence graph propagates this assignment between clients. Since $\widehat{K} = K^\star$, the injective correspondence is a bijection, excluding a client-specific permutation. Bounding the normalized barycenter perturbation of any in-basin limit fixed point by its incorrect mass gives Equation 14. Coupling uniqueness for fixed C is not used to claim fixed-point uniqueness. □

The theorem does not assume equal entity prevalence or use private word names. It is a qualitative sufficient-condition result with unknown constants; we do not claim that the experimental hyperparameters numerically certify this conservative basin condition. It also clarifies failure modes: an initialization outside the basin can select a different local partition; a disconnected incidence graph allows independent component permutations; and a word never decomposed into pure modes need not be identifiable.

A.6 Tracking the alternating server operator

Let $\theta^t_{m,E}$ be client m's encoder after its E local steps and let $\theta^{t+1} = \sum_m \omega_m \theta^t_{m,E}$. The actual bank message is constructed under the heterogeneous endpoints $\{\theta^t_{m,E}\}_m$. Define their weighted dispersion

$$\chi_t^2 = \sum_m \omega_m \|\theta^t_{m,E} - \theta^{t+1}\|^2. \tag{15}$$

Let $\mathcal{B}(\theta, C)$ denote one exact population transport-barycenter update when all messages use encoder θ, and let $C^\star(\theta)$ be its fixed point. Within the basin, assume

$$d(\mathcal{B}(\theta, C), \mathcal{B}(\theta, C')) \le \kappa d(C, C'), \quad \kappa < 1, \tag{16}$$

$$d(C^\star(\theta), C^\star(\theta')) \le L_B \|\theta - \theta'\|. \tag{17}$$

These local properties follow from the positive transport gap and a nonsingular normalized barycenter; we state them separately because neural encoders can violate them outside the identified basin. Unlike Theorem 1, Equation 16 is the additional condition that guarantees a unique local fixed point by contraction. We further assume that the prototype map and normalized barycenter are locally Lipschitz, so for the empirical heterogeneous-encoder update $\widehat{\mathcal{B}}_t$,

$$d\Big(\widehat{\mathcal{B}}_t(C), \mathcal{B}(\theta^{t+1}, C)\Big) \le e_t + L_P \chi_t =: \widetilde{e}_t, \tag{18}$$

where e_t contains finite-prototype and finite-Sinkhorn error. For SGD-style local steps with bounded second moments, standard local-drift expansion gives

$$\mathbb{E}\chi_t^2 \le c_\chi \eta^2 E^2 \Big(\mathbb{E}\|\nabla_\theta \widetilde{\mathcal{F}}(\theta^t, C^t)\|^2 + \sigma^2 + \zeta^2 + L_\phi^2 r_t^2\Big). \tag{19}$$

Thus local-encoder prototype drift is explicit and is controlled by the same local-optimization quantities as client drift; it is not hidden inside sampling error ξ_t.

Lemma 3 (Bank tracking). *Let $d_t = d(C^t, C^\star(\theta^t))$ and let $\widetilde{e}_t = e_t + L_P\chi_t$ be defined by Equation 18. Then*

$$d_{t+1} \le \kappa d_t + L_B \|\theta^{t+1} - \theta^t\| + \widetilde{e}_t, \tag{20}$$

$$\frac{1}{T}\sum_{t<T} d_t^2 \le \frac{2d_0^2}{T(1-\kappa^2)} + \frac{4}{(1-\kappa)^2 T}\sum_{t<T}\left(L_B^2\|\Delta\theta_t\|^2 + \widetilde{e}_t^2\right). \tag{21}$$

Proof. Insert $\mathcal{B}(\theta^{t+1}, C^t)$ between the heterogeneous-encoder empirical update and $C^\star(\theta^{t+1})$. Equation 18 contributes $\widetilde{e}_t$, Equation 16 contributes κd_t, and Equation 17 contributes $L_B\|\Delta\theta_t\|$. Unrolling the scalar recursion and applying Cauchy–Schwarz to the geometric convolution yields Equation 21. □

A.7 Conditional encoder convergence with bank and private-state tracking

This subsection uses SGD-style updates $\theta^t_{m,e+1} = \theta^t_{m,e} - \eta g^t_{m,e}$; it isolates the alternating server mechanism and does not analyze AdamW's adaptive state. Volume-tempered weights change the optimized population mixture from sample weighting to ω_m. Conditional on the round filtration, A4–A5 decompose $g^t_{m,e}$ into the profiled encoder gradient, stochastic noise, client drift, and a private-response bias bounded by $L_\phi\|\phi^t_{m,e} - \bar{\phi}_m\|$. Write q for participating clients and E for local steps.

Lemma 4 (One-round encoder descent at a fixed bank)**.** *If $\eta \le (8LE)^{-1}$, then*

$$\mathbb{E}\widetilde{\mathcal{F}}(\theta^{t+1}, C^t) \le \widetilde{\mathcal{F}}(\theta^t, C^t) - \frac{\eta E}{4}\mathbb{E}\|\nabla_\theta \widetilde{\mathcal{F}}(\theta^t, C^t)\|^2 + c_3\eta^2 EL\sigma^2/q + c_4\eta^3 L^2E^3\zeta^2 + c_6\eta E L_\phi^2 r_t^2. \tag{22}$$

Proof. Apply L-smoothness to each local step and condition on the round filtration. The envelope theorem identifies the profiled encoder gradient at the stationary response. Decompose $g^t_{m,e}$ into that gradient, stochastic noise, drift from local encoder iterates, and the private-response error. Weighted independence yields the σ^2/q term; smoothness and E accumulated steps give the $E^2\zeta^2$ drift; A5 and Young's inequality give $L_\phi^2 r_t^2$. The same inequality absorbs gradient cross-terms when $\eta \le (8LE)^{-1}$. This argument holds C^t fixed and makes no claim about the subsequent barycenter replacement. □

Theorem 2 (Conditional encoder convergence under stochastic-gradient local updates)**.** *Under A1–A6 and Equations 16–17, while the iterates remain in the local basin,*

$$\frac{1}{T}\sum_{t<T}\mathbb{E}\|\nabla_\theta \widetilde{\mathcal{F}}(\theta^t, C^t)\|^2 \le \frac{4(\widetilde{\mathcal{F}}_0 - \widetilde{\mathcal{F}}_{\inf})}{\eta ET} + O\left(\frac{\eta L\sigma^2}{q} + \eta^2 L^2E^2\zeta^2\right) + O\big(L_\phi^2 \bar{r}_T^2\big) + \frac{4\bar{v}_T}{\eta E}, \tag{23}$$

Here $\bar{r}_T^2 = T^{-1}\sum_t r_t^2$; independently, $\bar{d}_T^2$ and $\bar{\bar{e}}_T^2 = T^{-1}\sum_t (e_t + L_P\chi_t)^2$ obey Lemma 3. If $\eta = \Theta(T^{-1/2})$, E is bounded, $\bar{r}_T^2 = o(1)$, and $\bar{v}_T = o(\eta)$, then the average profiled encoder-gradient norm is $O(T^{-1/2}) + o(1)$. If additionally the bank-approximation average vanishes, the bank tracks its moving fixed point with factor $(1-\kappa)^{-2}$; Equation 19 makes the local-encoder contribution to $\bar{\bar{e}}_T^2$ vanish under these conditions. Without vanishing positive bank-value variation or private-state error, the theorem gives the corresponding explicit neighborhood rather than exact stationarity.

Proof. By Equation 10, $\mathbb{E}_t\widetilde{\mathcal{F}}(\theta^{t+1}, C^{t+1}) \le \mathbb{E}_t\widetilde{\mathcal{F}}(\theta^{t+1}, C^t) + v_t$. Add this inequality to Equation 22, sum, and telescope the lower-bounded profiled objective. Division by $\eta ET/4$ yields $4\bar{v}_T/(\eta E)$ in addition to the fixed-bank terms; then substitute Lemma 3. With $\eta = \Theta(T^{-1/2})$, Equation 19 absorbs local-encoder prototype dispersion into the usual optimization, variance, drift, and private-state terms; finite transport/prototype estimation remains in e_t. The proof neither freezes (B_m, Ψ_m) nor treats the barycenter as a gradient step, and exact stationarity is explicitly conditional on vanishing positive bank-value variation. □

A.8 Semantic-budget misspecification

Proposition 3 (Capacity mismatch)**.** *Within the separated basin, if $\widehat{K} < K^\star$, at least two entities share one column. If $\widehat{K} > K^\star$, at least $\widehat{K} - K^\star$ columns are unsupported or duplicate an occupied support; letting unsupported capacity approach zero and coalescing within-support duplicates recovers the occupied partition.*

Proof. The first two statements follow by the pigeonhole principle. The positive gap in Equation 12 prevents a supported column from mixing separated components except for Lemma 2's leakage. Therefore excess columns can only be unsupported or share one support. The learned capacity update permits the former to decay, unlike a fixed uniform marginal. □

Table 4: Default construction for the main private-taxonomy benchmarks. "Dir." means absence induced only by finite classwise Dirichlet allocation. Mixed uses the native CIFAR-100 hierarchy rather than synthetic merge/split.

Benchmark	M	Train/Test	α	Missing	Merge	Split	Fixed train/test?
MNIST	8	8k/2k	0.25	Dir.	$p = .28$, pairs	$p = .18$, median	Yes
CIFAR-10	8	8k/2k	0.25	Dir.	$p = .28$, pairs	$p = .18$, median	Yes
CIFAR-100	10	12k/3k	0.35	Dir.	$p = .22$, pairs	$p = .12$, median	Yes
C100-Mixed	12	12k/3k	0.35	Dir.	–	native 4/4/4	Yes

B Additional Experimental Detail

CIFAR images are resized to 128×128, ImageNet-normalized, and encoded by frozen pretrained ResNet-18. MNIST pixels are scaled to $[0, 1]$. The shared projection is Linear–LayerNorm–GELU–Linear with widths input–128–64 and normalized output. All methods use AdamW, learning rate 2×10^{-3}, weight decay 10^{-4}, batch 256, one local epoch, 12 rounds, and full participation. FedLENS retains at most five micro-prototypes per word, uses entity temperature $\tau = 0.10$, and uses transport temperature 0.04.

B.1 Private-taxonomy benchmark construction

Table 4 specifies the exact default severity used for every entry of Table 1. For each canonical class, we first draw its training allocation from one Dirichlet vector. The test allocation reuses that vector after zeroing client–class pairs with zero training count and renormalizing; hence every test entity was observed by that client in training, while missing entities arise from finite training allocation rather than an additional deletion rate. Each private taxonomy is then sampled solely from the client's training support. We permute those entities, scan the order, and merge the current entity with its successor with probability p_{merge}; every accepted merge contains exactly two canonical entities. The resulting groups are shuffled to define anonymous base-word indices. Independently, each training-observed entity is selected for conditional splitting with probability p_{split}; split-derived indices are appended. Its second private word is assigned iff the first input-feature coordinate exceeds that entity's client-local training median, and the same threshold is used at test time. Thus neither realized test labels nor test support can change a training allocation, vocabulary, merge partner, split decision, threshold, or private-word index. The complete mapping is held fixed across train and test.

CIFAR-100-Mixed uses a fine-class Dirichlet allocation. Four clients retain native coarse labels, four retain native fine labels, and four are hybrid: for each hybrid client, half of its observed native superclasses are expanded into their fine descendants and the remainder stay coarse. Their order is shuffled independently in each run. The resulting vocabulary is independently permuted per client and held fixed between train and test. No additional synthetic merge or split is applied. Canonical fine labels are used only to construct this benchmark and compute evaluator-side metrics; they never enter model inputs, optimization, or server messages.

Table 5: CIFAR-100-Mixed private-label accuracy by client granularity (%, mean $\pm$ standard deviation over five paired seeds). Best/second observed means are bold/underlined.

Method	Coarse ↑	Fine ↑	Hybrid ↑
FedRoD	<u>57.53 ± 1.43</u>	<u>38.41 ± 4.37</u>	<u>45.31 ± 2.27</u>
ResidualOnly	49.15 ± 1.40	30.95 ± 2.84	38.98 ± 1.36
FedLENS	**57.67 ± 1.17**	**38.86 ± 4.03**	**45.41 ± 2.29**

FedLENS retains the highest observed mean for coarse, fine, and hybrid clients. Its gains over FedRoD are $+0.14$, $+0.45$, and $+0.10$ points, respectively, showing that the Mixed result is not driven solely by easier coarse clients; all three margins are small and should be interpreted as consistency rather than statistical separation.

B.2 BASELINE ADAPTATIONS

The four class/name-aware families are explicitly suffixed "-style" because their original semantic interface is unavailable. FedProto-style pools private-word centroids without word IDs and pulls features toward the nearest anonymous server prototype. FedETF-style constructs a separate fixed equiangular classifier sized to each client's $|\mathcal{Y}_m|$; no classifier coordinate is shared. FedDisco-style computes discrepancy weights from the normalized entropy of each private-word histogram, so class statistics stay local. FedAlign-style retains representation alignment but receives neither public names nor a cross-client map. AnonKMeans pools shuffled word means, selects $\max(8, 2M)$ centers, and applies nearest-center attraction without mass-aware transport. OracleAlign alone receives canonical evaluator identities to aggregate entity prototypes and is reported as a privileged reference, never as an anonymous method.

The 168 main runs contain the complete four-dataset $\times$ 12-method $\times$ three-seed grid plus seeds 3–4 for FEDLENS, FedRoD, and ResidualOnly. Recomputing all eight utility rankings on common seeds $\{0, 1, 2\}$ leaves FEDLENS first in every cell. The 600 extended runs contain 60 targeted ablations, 48 complete primary-task ablations, 15 temperature, 15 entropy, 18 mixture, 12 budget, and 432 all-method stress runs. Thus every stress comparison contains all 12 algorithms rather than a selected subset.

B.3 COMPLETE PRIMARY AND TARGETED ABLATIONS

Table 6: Complete CIFAR-100 primary-task ablation (three paired seeds). Upload is the incremental prototype/count payload.

Variant	Mean acc. ↑	Worst acc. ↑	NMI ↑	Upload (MB/round) ↓
Full	61.66 ± 4.99	48.78 ± 10.40	48.24 ± 1.52	3.08 ± 0.01
Single centroid	61.66 ± 5.07	**48.95 ± 9.77**	48.20 ± 1.62	**3.01 ± 0.00**
Fixed capacity	**61.73 ± 5.15**	48.92 ± 9.95	48.11 ± 1.45	3.08 ± 0.01
No entity path	61.64 ± 5.07	48.80 ± 9.56	47.80 ± 1.40	3.08 ± 0.01
No cycle	61.69 ± 5.04	48.92 ± 9.95	48.04 ± 1.30	3.08 ± 0.01
Sample weighting	61.40 ± 5.40	48.19 ± 10.78	**48.39 ± 1.03**	3.08 ± 0.01
Pooled k-means	61.65 ± 5.04	48.77 ± 10.06	48.22 ± 1.47	3.08 ± 0.01
No residual	18.73 ± 2.58	7.97 ± 1.24	27.54 ± 8.36	3.04 ± 0.05

The CIFAR-100 reversals among Full, fixed capacity, and single-centroid variants are at most 0.07 mean-accuracy points against standard deviations near five points. We treat them as ties rather than evidence that every component improves an aggregate endpoint. The matched removals are nevertheless diagnostic: eliminating the entity path lowers NMI, standard sample weighting slightly lowers mean/tail utility, and removing the residual destroys private-label utility. This calibration is consistent with the stronger regime-specific interventions below.

Table 7: Targeted mechanism interventions (%, three paired seeds). Each endpoint is chosen before comparing Full with the corresponding removal.

Component	Target regime	Metric	Full	Without	Gain
Micro-prototypes	merge	Mode recall ↑	**92.26 ± 2.35**	50.00 ± 0.00	**+42.26**
Micro-prototypes	merge	Held-out recovery ↑	**96.75 ± 0.50**	77.46 ± 3.59	**+19.29**
Mass-aware transport	merge	Entity NMI ↑	**70.18 ± 0.71**	67.94 ± 3.09	**+2.24**
Private residual	split	Mean acc. ↑	**90.89 ± 2.64**	26.95 ± 9.86	**+63.94**

Adaptive micro-prototypes raise severe-merge mode recall from 50.00% to 92.26% and held-out recovery from 77.46% to 96.75%. Mass-aware transport adds 2.24 entity-NMI points under severe merges, while the private residual adds 63.94 mean-accuracy points under conditional splits. We omit a positive claim for the cycle term because its small aggregate effect changes sign across seeds and tasks; its role is proximal stabilization, not a separately established recovery mechanism.

B.4 Measured communication, computation, and numerical residuals

Table 8: Measured FedLENS negotiation overhead per round (three seeds). Upload excludes the encoder delta common to all methods; times are wall-clock milliseconds on the stated software simulator.

Dataset	Upload MB	Client cluster ms	Server OT ms	Row ℓ_1	Fixed-point ℓ_∞
CIFAR-100	3.08	971.6	48.6	4.0e-08	1.2e-07
C100-Mixed	3.82	1173.2	27.2	3.1e-08	9.7e-08

Each word transmits at most five 64-dimensional float vectors plus counts. Table 8 measures the realized payload and separately accounts for client clustering and server transport. These are implementation timings rather than hardware-independent complexity claims. With J uploaded prototypes, one scaling step costs $O(J\widehat{K})$ and stores $O(J\widehat{K})$ coupling entries; prototype formation is linear in local examples apart from the bounded four-step spherical k-means loop.

B.5 Aggregation control

We isolate the nonstandard encoder weight on the default MNIST construction while keeping the model, optimizer, protocol, and seeds fixed. Table 9 compares the implemented exponent $1/2$ with standard sample weighting (exponent 1) over paired seeds $\{0, 1, 2\}$. These six default-setting runs are reported separately and are not included in the 600-run extended grid counted above. Square-root weighting is modestly better on both utility endpoints, but it is an auxiliary fixed choice rather than a claimed algorithmic contribution. More importantly, ResidualOnly uses the identical exponent in the main study, so the matched-control gap cannot be attributed to aggregation.

Table 9: Aggregation control on default MNIST (%, three paired seeds).

Encoder aggregation	Mean accuracy ↑	Worst-client accuracy ↑
Standard sample ($w_m \propto \lvert D_m \rvert$)	$\underline{89.31 \pm 2.19}$	$\underline{79.70 \pm 5.43}$
Volume-tempered ($w_m \propto \sqrt{\lvert D_m \rvert}$)	$\mathbf{90.09 \pm 1.81}$	$\mathbf{82.64 \pm 3.87}$

B.6 Full semantic results

Table A1: Semantic NMI (%) on the common three-seed grid used by every method.

Method	MNIST	CIFAR-10	CIFAR-100
	NMI ↑	NMI ↑	NMI ↑
FedRep	70.43 ± 1.05	55.60 ± 1.38	45.14 ± 2.34
FedBABU	69.58 ± 1.22	53.63 ± 2.98	44.52 ± 1.87
FedRoD	72.69 ± 0.72	58.00 ± 1.43	$\underline{48.12 \pm 1.11}$
FedProto-style	69.09 ± 2.48	55.26 ± 0.92	44.56 ± 2.41
FedETF-style	69.80 ± 1.95	56.49 ± 0.92	47.52 ± 1.57
FedDisco-style	70.23 ± 1.46	54.86 ± 2.27	44.47 ± 2.19
FedAlign-style	68.27 ± 2.50	55.28 ± 1.93	44.47 ± 1.92
FedLWS	70.02 ± 1.51	54.13 ± 2.05	44.52 ± 2.40
OracleAlign	$\mathbf{74.41 \pm 1.76}$	$\underline{58.64 \pm 1.08}$	47.91 ± 1.09
AnonKMeans	66.80 ± 4.74	53.40 ± 1.97	44.76 ± 1.63
ResidualOnly	70.43 ± 1.05	55.60 ± 1.38	45.14 ± 2.34
FedLENS	$\underline{73.75 \pm 0.62}$	$\mathbf{58.95 \pm 1.35}$	$\mathbf{48.24 \pm 1.52}$

B.7 Sensitivity

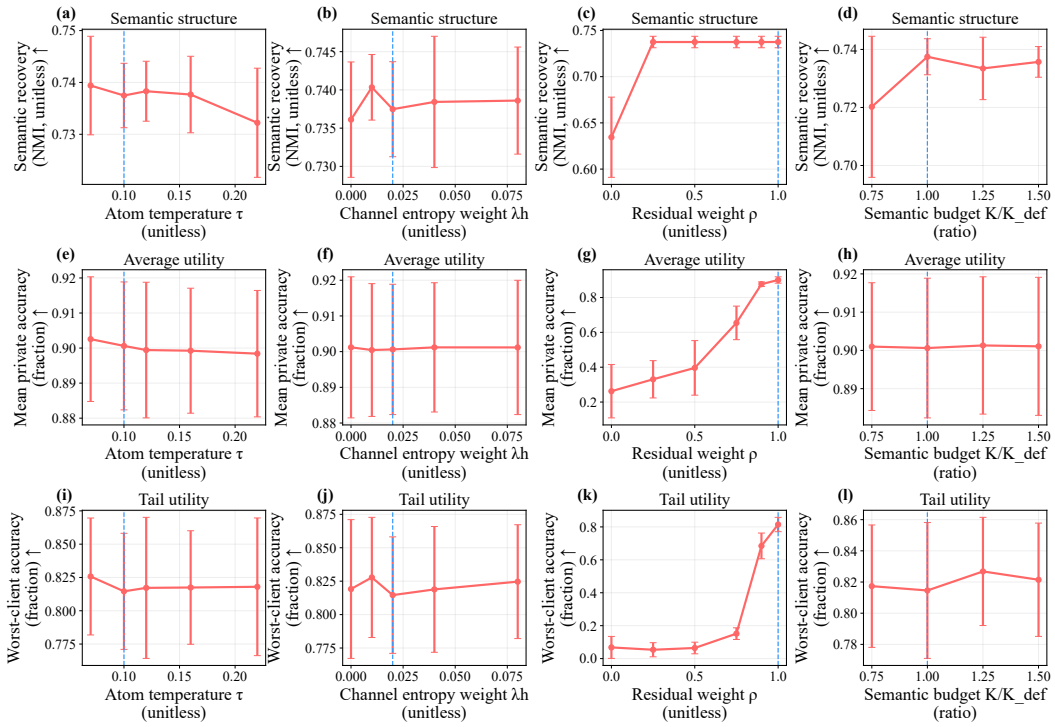


Figure A1: Sensitivity to transport temperature, channel entropy, readout mixture ρ, and non-oracle semantic budget $\widehat{K}$. Columns report NMI, mean utility, and worst-client utility; error bars are standard deviations over three seeds. Temperature and entropy primarily affect semantic sharpness; $\rho = 0$ cannot express arbitrary conditional refinements; over-specified budgets permit unsupported columns to lose mass.