LEARNING NEW CLASSES FEDERATIVELY WITHOUT A SHARED LABEL SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated few-shot class-incremental learning (FFSCIL) assumes that new classes arrive in one shared global label space. Once clients label the world with private, partially overlapping taxonomies, the server must first decide which client classes denote the same concept, from one to five samples and with no correspondence oracle. We study this concept-identity problem as Taxonomy-Free FFSCIL and build its benchmark from real taxonomies: synonyms, real homonyms, and clients that label at the superclass level. We propose FedTaxo, which encodes each local class as a feature-sharing concept capsule and grows a latent concept graph by re-clustering every novel capsule each session under the constraint that two identifiers of one client are never one concept. For a Gaussian surrogate, a centered threshold identifies concepts at the minimax rate, and constrained re-clustering recovers a session's exact partition under a separating score. Controlled experiments on four datasets show that the assignment procedure matters more than the matching score: re-clustering recovers 85% of an oracle's novel-class gain, against 69% when the same score is assigned greedily, while a learned matcher adds little accuracy but recovers far more of the true sub/super edges. When client taxonomies genuinely cross-cut, even merging by true set overlap falls below not federating. Code, protocol and LLM-written client vocabularies are available at anonymous.4open.science/r/fedtaxo-iclr27-FF2A.

1 INTRODUCTION

Federated few-shot class-incremental learning (FFSCIL) asks for a model that keeps learning new classes from the few examples each client can supply, without pooling raw data and without forgetting the old ones (McMahan et al., 2017; Ma'sum et al., 2025). The setting has moved quickly, from representation augmentation (Jiang et al., 2024) and non-IID decentralized FSCIL (Liu et al., 2025) to prompt-based federated incremental learning (Ma'sum et al., 2025; 2024; Guo et al., 2024). Every one of these pipelines contains a step that is easy to miss: incoming client classes are already indexed into one global label set, so the server grows a head by the number of new global classes and aggregates prototypes per global class ID. This bundles three assumptions: all clients draw labels from one shared vocabulary, a label string means the same thing on every client, and all clients label at the same granularity.

All three fail in the deployments that motivate federated learning. Parties log the same pavement defect as "road crack" and "linear damage" (synonymy), tag one bird as "sparrow" or "bird" (coarse and fine labels), or reuse one string for two concepts (homonymy), and aligning such vocabularies by hand is the centralized coordination federated learning exists to avoid. The server must therefore decide, from one to five samples per class, with no correspondence oracle and no raw data, whether two client-side new classes are the same, related or different concepts. We call this the *concept-identity* problem of Taxonomy-Free FFSCIL (Figure 1). What it should be judged by is each client's accuracy in its own label space, against what a true correspondence between clients would give. An identity error is not a one-off mistake here: a wrongly merged prototype is aggregated into every later session, so alignment error compounds into forgetting.

Recent work makes this problem concrete, but does not yet fully solve it. Text-side alignment merges labels whose embeddings agree, as a CLIP text encoder does implicitly (Radford et al., 2021) and as FedAlign does by anchoring client class representations to their names (Zhang et al., 2023); it needs

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.

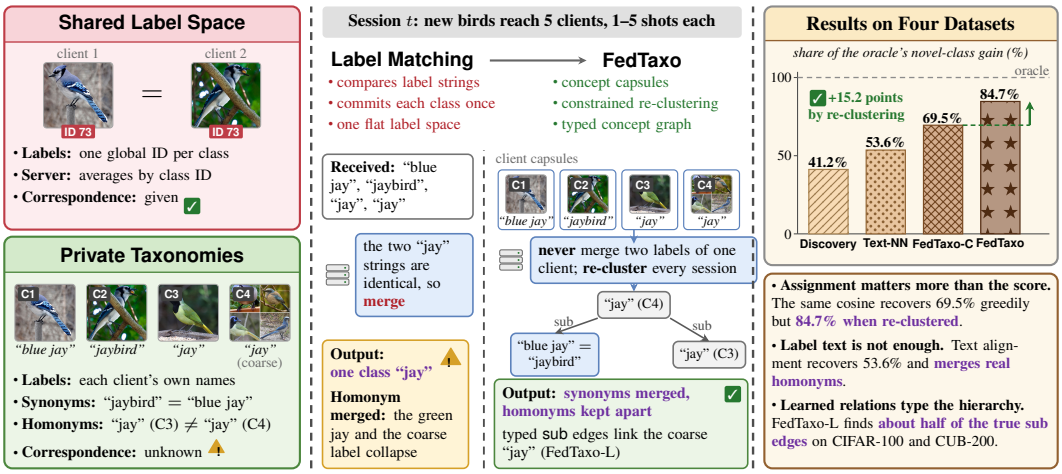

Figure 1: Beyond a shared label space, **Taxonomy-Free FFSCIL** must infer which private client labels name the same concept. Photos from CUB-200 (Wah et al., 2011).

no training and resolves synonymy, but it cannot separate two concepts that share a string. Data-side discovery clusters the samples themselves (Vaze et al., 2022; Pu et al., 2024; Wang et al., 2025) and discards the labels clients do have. Federated methods for mismatched label sets handle subsets of one known label set (Deng et al., 2023) or match categories across crowdsourced annotators (He et al., 2026); here, classes instead arrive incrementally at one to five shots and may subsume one another. We ask a different question: once every method is given the constraints a client already knows, which part of an alignment front-end, the matching score or the assignment procedure that applies it, recovers the gain of a true correspondence?

Motivated by this question, we propose FedTaxo, an alignment front-end for Taxonomy-Free FFSCIL. Each client encodes a local class as a concept capsule of shot embeddings, prototype, description embedding and calibrated covariance, and only the capsule leaves the client. The server scores capsule pairs, by default with a fixed cosine over the visual and textual channels and, in the variant FedTaxo-L, with a learned relation matcher, and assigns them to a latent concept graph by re-clustering every novel capsule each session under two constraints: capsules a client sent under one identifier stay together, and two identifiers of one client never merge. Typed edges connect broader and narrower concepts, and a fixed rule maps the graph back to each client's own labels.

We validate FedTaxo both theoretically and empirically. For an isotropic Gaussian surrogate, a centered threshold identifies concepts at the minimax rate, no monotone text-only rule separates a homonym from a synonym while a fused score can, and constrained re-clustering recovers a session's exact partition under a separating score. Experiments on four datasets, with a training-free and a trained classifier, show that the procedure carries more of the result than the score, that the learned matcher earns its place in typed hierarchies, and where the formulation stops working.

Key contributions.

- We formulate Taxonomy-Free FFSCIL, which separates local class *identifiers* from the *descriptions* clients attach to them, and build its benchmark from each dataset's real taxonomy: synonyms, real homonyms and clients that label at the superclass level.
- We propose FedTaxo, whose assignment is *constrained*, since two identifiers of one client are never one concept, and *revisable*, since every novel capsule is re-clustered each session, with all hyperparameters chosen by a dry run on the base session.
- We prove that a centered threshold identifies concepts at the minimax rate $\sigma d^{1/4}/\sqrt{K}$, that monotone text-only matchers cannot separate a homonym from a synonym, and that constrained re-clustering recovers a session's exact partition under a separating score.
- In a controlled 2×2, re-clustering is worth $+0.70$ Avg to the cosine score while our learned matcher is worth $+0.47$ under greedy assignment but -0.03 under re-clustering; the learned matcher alone

recovers typed hierarchies, and the diagnosis survives a weaker encoder, an unaligned base session and a trained classifier.

2 PROBLEM FORMULATION

We first define the setting and its target. A federation of M clients, coordinated by a server G, learns a sequence of sessions $[0, 1, \ldots, T]$. At session t, client l receives a labeled set $\mathcal{T}_l^t = \{(x_i, y_i)\}$ with images $x_i \in \mathcal{X}$ and labels $y_i \in Y_l^t$ drawn from its *private label space*, a finite set of local class *identifiers*. Each identifier carries a *description*, a string or code, and the two are distinct: a client always knows that two of its own identifiers are different classes even when their descriptions collide (both of its jay species may be described as "jay"). Descriptions are therefore the only cross-client evidence, and any front-end may assume that two identifiers *of one client* denote different concepts, a cannot-link constraint we give every method we compare. As in standard FSCIL, session 0 has abundant data and sessions $t > 0$ are few-shot, with $K \in \{1, 3, 5\}$ shots per new identifier and non-IID clients each holding a subset of the classes present at session t.

The label spaces are not a shared vocabulary. There is a latent set of ground-truth concepts $C^\star$, unknown to every party, and per client a *grounding map* $g_l : Y_l \to \mathcal{P}(C^\star)$. Across clients these maps induce synonymy, homonymy (identical descriptions, $g_a \neq g_b$), coarse and fine subsumption ($g_a(y) \supset g_b(y')$), client-private concepts and evolution across sessions. There is *no correspondence oracle*: neither G nor any client observes g_l, $C^\star$ or any cross-client alignment. Clients share only *concept capsules* (Sec. 3), never raw samples, and G aggregates them into a concept graph G_t, a structural operator over concepts rather than a mean over class IDs.

After session t, each client l must classify a test input into *its own* label space $Y_l^{(\leq t)}$ while benefiting from other clients' data about the same concepts. Let $\Phi_{G_t}(x)$ score concept nodes, let $V_{l,t}$ contain the nodes accessible to client l, and let $A_l : V_{l,t} \to Y_l$ map them to its labels. Writing $\widehat{v}_l(x) = \arg\max_{v \in V_{l,t}} [\Phi_{G_t}(x)]_v$ and $h_l(x) = A_l(\widehat{v}_l(x))$, the goal is

$$\min_{\{A_l\}, G_t} \frac{1}{M} \sum_{l=1}^{M} \mathbb{E}_{(x,y)\sim\mathcal{D}_l^{(\leq t)}} \left[\mathbf{1}(h_l(x) \neq y)\right], \tag{1}$$

subject to the no-oracle and no-raw-data constraints. We measure per-client accuracy in Y_l with the standard FSCIL summaries, and the average precision of predicting "same concept" for cross-client label pairs, the hidden g_l being used only for evaluation. The base session is shared and canonically labeled, as in the FSCIL protocols we build on, so base concepts are excluded from later matching; Sec. 4.4 withdraws that assumption. What we establish is accordingly narrow: alignment of unaligned *novel* classes with client taxonomies simulated from each dataset's own hierarchy, not a general solution for arbitrary private taxonomies. Section 3 defines how FedTaxo builds G_t.

3 FEDTAXO

FedTaxo is our alignment front-end for Taxonomy-Free FFSCIL. Each client uploads one concept capsule per local class; the server groups capsules into concepts under the constraints a client already knows, links broader to narrower concepts with typed edges, and revisits every novel capsule each session so that an earlier assignment can be undone. A fixed rule maps the graph back to each client's labels (Figures 1 and 2). FedTaxo scores pairs by a fixed cosine and FedTaxo-L by a learned relation matcher; Appendix B states every step formally.

3.1 CONCEPT CAPSULES AND PAIRWISE MATCHING

Client m turns the images $\mathcal{X}_{m,k}^t$ it holds at session t under local identifier k, described by the string $\ell_{m,k}$, into

$$S_z = \{\nu(f_\theta(x)) : x \in \mathcal{X}_{m,k}^t\}, \qquad p_z = \nu\Big(\tfrac{1}{n_z} \textstyle\sum_{s\in S_z} s\Big), \qquad e_z = \nu(g_\psi(\ell_{m,k}))\,, \tag{2}$$

with frozen CLIP encoders f_θ, g_ψ (Radford et al., 2021) and $\nu(a) = a/\|a\|_2$: the shot embeddings, their prototype, and the embedding of the *description* rather than of the identifier, whose identity

stays in the metadata (m, k) so that two identical descriptions remain two classes. A diagonal covariance, calibrated from the three nearest base classes in a 64-dimensional PCA space (Yang et al., 2021), completes the capsule $z = [S_z, p_z, e_z, \Sigma_z]$; the default score does not use it. A fine capsule holds K shots, a coarse one pools the shots of several fine classes. What leaves the client is $(n_z{+}1)d + 64$ floats and the label string: in float32 at $d{=}768$, 18 KB for a five-shot fine capsule and 49 KB for a coarse capsule pooling five shots of each of three members. Shot embeddings are deterministic functions of the images, so this is feature sharing without a formal guarantee; we measure perturbation sensitivity, not privacy.

A group of capsules U is summarized by a shot-weighted prototype p_U and an equally weighted text embedding e_U. FedTaxo scores a pair of groups inside each embedding space,

$$s_{\cos}(U, V) = w_v \langle p_U, p_V \rangle + w_s \langle e_U, e_V \rangle, \qquad w_v = 0.6, \quad w_s = 0.4, \tag{3}$$

which is symmetric and needs no training: text connects synonyms, while visual evidence lowers the score of distinct concepts that share a name. FedTaxo-L instead predicts $q_\omega(r \mid U, V)$ over $\{\text{different}, \text{same}, \text{super}, \text{sub}\}$ from 18 pair features (Table 6), among them channel cosines, directed shot-set similarities and lexical and vocabulary-depth cues. It is trained only on the labeled base session, whose classes simulate private vocabularies at the deployment shot count and whose pseudo-coarse capsules pool two to four base classes under a superclass name; clustering uses its symmetrized same-concept probability and the hierarchy its directed ones.

3.2 Constrained, revisable assignment

Every novel capsule seen so far is pooled with the session's arrivals, the base nodes staying fixed. Capsules a client sent under one identifier form a must-linked unit, and two clusters may merge only when they share no client, since two identifiers of one client are two concepts. With $\mathcal{M}(A)$ the clients in cluster A and s either score, constrained average linkage is

$$L(A, B) = \frac{1}{|A||B|} \sum_{U \in A} \sum_{V \in B} s(U, V) \;\text{ if } \mathcal{M}(A) \cap \mathcal{M}(B) = \varnothing, \qquad L(A, B) = -\infty \;\text{ otherwise,} \tag{4}$$

and the best admissible pair is merged while its linkage reaches τ. Each session starts again from the units, so an earlier merge can be split and an earlier split merged as evidence arrives. Committing each capsule on arrival instead is the greedy ablation.

Each node keeps its own prototype, so broader and narrower concepts stay separate, and a directed edge $v \to u$ marks v as a sub-concept of u. FedTaxo-L proposes an edge when a node pair is related above τ and super is its most probable relation; the cosine default must read direction from the spread ratio ($\gamma = 2$) and a head-noun or WordNet cue instead. Both insert edges most confident first, one parent per node and no cycle. A fixed rule then maps the nodes a client can reach to its own identifiers, and the fine descendants of a coarse node inherit its label, so a coarse client can use a finer prototype while predicting its broader one; a query goes to the nearest accessible prototype and is then mapped. No adapter parameters are trained.

The one quantity selected per dataset is the merge threshold τ, drawn for the cosine score from the F_β-optimal thresholds on simulated base-class pairs ($\beta \in \{0.5, 1\}$) plus the fixed values $\{0.3, 0.5, 0.7\}$, and for FedTaxo-L from that grid on its probability. Selection scores assignment and prediction together: the base classes are split into three folds, each in turn playing pseudo-novel under the same corruption on two seeds, and the best mean pseudo-novel accuracy wins ($\tau{=}0.3$ everywhere). No novel label enters selection, and every front-end is calibrated the same way.

3.3 Theoretical guarantees

We now ask when a capsule matcher can recover the true correspondence g_l. We analyze an unnormalized Gaussian surrogate for the normalized prototypes: each concept $c \in C^\star$ induces features $\phi(x) = \mu_c + \xi$ with $\xi \sim \mathcal{N}(0, \sigma^2 I_d)$, p_c is the mean of K such features, and we write $\Delta_{ab} = \|\mu_a - \mu_b\|_2$ and $s^2 = 2\sigma^2/K$. Proofs are in Appendix G.

Theorem 1 (Few-shot identifiability, informal)**.** *Fix a target separation Δ_0 and test H_0, that two prototypes come from one concept, against H_1, that they come from concepts at distance $\Delta_{ab} \geq \Delta_0$, by the* centered *rule that declares* same *iff $\|p_a - p_b\|_2^2 < s^2 d + \Delta_0^2/2$, a threshold fixed by Δ_0 and*

not by the unknown pair. In the regime $\Delta_0^2 \le s^2 d$, the test is correct with probability at least $1-\delta$ under each hypothesis whenever

$$\Delta_0^2 \;\ge\; 8\,s^2\big(\sqrt{d\log(1/\delta)}+\log(1/\delta)\big), \qquad \textit{i.e.} \qquad \Delta_0 = \Omega\Big(\sigma\,\big(d\log(1/\delta)\big)^{1/4}/\sqrt{K}\Big). \tag{5}$$

Conversely, if $\Delta_0 = o(\sigma/\sqrt{K})$ every test errs with probability at least $1/4$. This converse is loose; the minimax rate is of order $\sigma d^{1/4}/\sqrt{K}$.

Theorem 1 explains why the threshold must be centered. Two prototypes of one concept differ by $\mathcal{N}(0, s^2 I_d)$, so $\|p_a - p_b\| \approx \sigma\sqrt{2d/K}$ rather than 0, and thresholding the raw distance at $\Delta_0/2$ splits same-concept pairs almost surely in any moderate dimension (d=64, K=5, σ=1: $\Pr \approx 4{\times}10^{-22}$ of a correct merge). Centering on the same-concept distance, which is what tuning τ on corrupted base pairs estimates, costs a $d^{1/4}$ factor, and that factor is the minimax rate rather than a gap: for this isotropic Gaussian test the optimal sample complexity is $\Theta(\sqrt{d}/\varepsilon^2)$ (Acharya et al., 2020, p. 2), a separation of order $\sigma d^{1/4}/\sqrt{K}$, so the centered rule is rate-optimal up to $\log(1/\delta)$ and only our converse is loose. The bound concerns unnormalized Gaussian means, not normalized CLIP prototypes.

Theorem 2 explains why the visual channel is needed. Any text-only matcher whose decision is monotone in the label-embedding distance, which every text baseline here is, merges every homonym pair once it merges some true synonym pair, whereas a fused score can separate them: some threshold keeps the synonym and rejects the homonym exactly when the homonym's weighted visual distance exceeds the synonym's own visual term plus $w_s\gamma$. This is separability, not rejection at the dry-run threshold. The visual margin is scaled by $\tilde{\rho} = \max(\rho_z\rho_v, \rho_{\min})$, a product of two per-capsule reliabilities that we do not bound but measure: 0.81–0.85 at five shots and 0.59–0.66 at one on CIFAR-100 and CUB-200. The one-shot failure of our Bhattacharyya variants is consistent both with this shrinking margin and with prototype noise (Eq. (5)), which our experiments do not separate. Theorem 3 gives the one guarantee of the split pass used by our hand-designed variants, an exact impurity accounting in which a split helps only when its cut separates concepts cleanly. Proposition 1 covers FedTaxo's own assignment: constrained re-clustering recovers a session's exact concept partition, whatever the arrival order or earlier graph, whenever the score is τ-separating on the session's units.

4 Experiments

In Taxonomy-Free FFSCIL a front-end makes two choices, how to score a pair of capsules and how to assign capsules to concepts, and either could carry the gain of a true correspondence. Our experiments ask which one does and how far the answer extends. We organize them around four research questions (RQs):

(RQ1) Is the gain carried by the assignment procedure or by the matching score?
(RQ2) How much of an oracle correspondence does FedTaxo recover across datasets, shot counts and classifiers?
(RQ3) Can the concept graph recover the hierarchy that coarse labelers induce?
(RQ4) Does the diagnosis hold beyond the simulator and its assumptions?

We use a frozen CLIP ViT-L/14 (Radford et al., 2021) on four FSCIL datasets, CIFAR-100 (Krizhevsky, 2009), *mini*ImageNet (Vinyals et al., 2016), CUB-200 (Wah et al., 2011) and FGVC-Aircraft (Maji et al., 2013). The loop follows UOPP (Ma'sum et al., 2025): 60 base classes and 8×5-way sessions (100 + 10×10 on CUB-200), new classes arriving at 60% of five clients, and a training-free nearest concept-node classifier. Taxonomy corruption is built from each dataset's real taxonomy (Appendix C): **SYNHOM** relabels a concept with a WordNet synonym (Miller, 1995) and collides distinct concepts on a shared name ("jay", "737"), and **HIER** additionally lets two of the five clients label at the superclass level. We report **Avg**, the mean accuracy over sessions, and final **novel**-class accuracy, over 5 seeds; significance is a Holm-corrected one-sided paired t-test at $p < 0.05$. We compare FedTaxo and FedTaxo-L with the following front-ends, all given identical capsules, encoders and constraints (Table 5):

- **independent** (*no federation*) keeps every client's classes apart;

Table 1: Score against assignment on the four datasets (Avg, %).

Shots	Dataset	Cosine score greedy *FedTaxo-C*	Cosine score re-cluster **FedTaxo**	Learned score greedy	Learned score re-cluster **FedTaxo-L**	Learned, re-cluster 4 feat.	Learned, re-cluster no CL	Oracle
5-shot	CIFAR-100	80.4	81.1 ✓	80.6	80.9 ✓	81.0	80.2	81.3
	*mini*ImageNet	93.8	93.9 ✓	94.0	94.0 =	93.9	94.0	94.0
	CUB-200	85.5	85.1 ✗	86.0	86.1 ✓	86.0	85.7	86.2
	Aircraft	59.2	59.3 ✓	58.7	59.3 ✓	59.3	58.7	59.5
Mean	*(4 settings)*	79.73	79.86 ✓	79.81	80.08 ✓	80.06	79.67	80.24
1-shot	CIFAR-100	75.3	76.1 ✓	76.0	75.7 ✗	75.5	75.0	77.0
	*mini*ImageNet	89.4	91.3 ✓	90.5	90.7 ✓	90.3	89.9	91.6
	CUB-200	78.4	80.4 ✓	80.2	80.6 ✓	80.4	79.7	81.5
	Aircraft	56.0	56.4 ✓	55.7	56.1 ✓	55.8	55.7	56.5
Mean	*(4 settings)*	74.77	76.04 ✓	75.64	75.77 ✓	75.50	75.07	76.65
All	*(8 settings)*	*77.25*	*77.95* ✓	*77.73*	*77.92* ✓	*77.78*	*77.37*	*78.44*
HIER sub-edges, rec./prec.		1/67	8/25	39/84	53/83	–	–	100/100

- **text-NN** and **text-strong** (*text alignment*) merge on the CLIP or BGE (Xiao et al., 2024) cosine of the label strings;
- **visual** merges on the prototype cosine, and **discovery** (*GCD-style*) re-clusters prototypes without labels;
- **FedTaxo-C, -B, -A** (*our hand-designed variants*, Appendix E) commit each capsule greedily with a fixed fusion, adding covariance, repair or joint transport;
- **oracle** merges by the true correspondence and serves as the reference.

4.1 RQ1: Assignment against score

We first answer RQ1 by crossing the two choices a front-end makes while holding the classifier, the constraints and the threshold selection fixed. In Table 1 each column is one configuration and each row one dataset-by-shot setting; the *Mean* rows average Avg over four settings and the *All* row over eight, and the last row gives sub-edge recall and precision in the HIER regime on CIFAR-100. A green ✓ or red ✗ marks a re-clustering cell above or below greedy assignment of the same score, and a gray = a tie, at the printed precision.

Over the eight settings, replacing greedy assignment by constrained re-clustering is worth $+0.70$ Avg to the cosine score and the cannot-links a further $+0.55$, while the learned 18-feature matcher is worth $+0.47$ under greedy assignment but -0.03 under re-clustering, four features being within 0.14 of eighteen (differences of unrounded means). The two components are not interchangeable. Scoring every cross-client pair of novel capsules against the truth (Table 10), re-clustering repairs 503 of the greedy cosine's errors over eight settings and the learned matcher 406, but only 78% of the matcher's repairs are also the procedure's, and each breaks pairs the other keeps; on five-shot CUB-200 the learned score leaves 8 errors under re-clustering where the cosine leaves 102, and there FedTaxo-L leads on accuracy. This does not show that a better score cannot help: at one shot our matcher is the weaker ranker, below the cosine on all four datasets and at 0.294 alignment AP against 0.552 on CUB-200 (Table 36).

Takeaway 1

Constrained re-clustering carries more of an oracle's gain than the learned matcher we trained; the two repair overlapping but different errors, so a better score is not ruled out.

4.2 RQ2: Distance to the oracle

We next answer RQ2 with Table 2, which reports Avg and novel accuracy for every front-end, with a mean over the four datasets and an asterisk where FedTaxo beats both text alignment and not federating significantly (mean±std in Table 11).

FedTaxo leaves a mean gap to the oracle of 0.49 Avg against 1.19 for the same score assigned greedily and 3.35 for not federating, and captures 85% of the oracle's novel-class gain against 69%

Table 2: Average session accuracy and final novel-class accuracy (%).

		CIFAR-100		*mini*ImageNet		CUB-200		Aircraft		Mean	
Shots	Front-end	Avg	novel	Avg	novel	Avg	novel	Avg	novel	Avg	novel
5-shot	independent	78.8	55.5	93.0	83.4	84.2	67.8	57.7	36.5	78.4	60.8
	text-NN	80.0	59.7	93.3	85.1	85.2	74.1	58.8	43.7	79.3	65.7
	text-strong	79.9	59.3	93.6	86.8	85.4	75.6	58.7	43.0	79.4	66.2
	visual	80.3	65.8	93.8	89.4	85.8	76.7	58.7	46.5	79.7	69.6
	discovery	80.8	68.5	94.0	**91.1**	86.0	77.6	58.6	44.6	79.8	70.4
	FedTaxo-C	80.4	66.2	93.8	89.6	85.5	75.5	59.2	49.3	79.7	70.2
	FedTaxo-B	80.4	66.7	93.5	88.1	85.7	76.3	58.8	46.6	79.6	69.4
	FedTaxo-A	79.4	58.9	92.1	75.3	85.0	71.8	59.1	47.1	78.9	63.3
	FedTaxo	**81.1***	**71.2**	93.9*	90.6	85.1	73.1	**59.3***	**50.4**	79.9	71.3
	FedTaxo-L	80.9	69.5	**94.0**	90.8	**86.1**	**78.4**	59.3	50.2	**80.1**	**72.2**
	oracle	*81.3*	*72.1*	*94.0*	*90.9*	*86.2*	*78.8*	*59.5*	*51.9*	*80.2*	*73.4*
1-shot	independent	72.3	11.0	85.9	39.2	74.6	21.7	54.4	14.3	71.8	21.5
	text-NN	75.9	32.5	89.7	60.8	79.3	44.3	56.1	24.5	75.3	40.5
	text-strong	75.7	31.4	89.6	59.9	79.8	46.3	56.1	24.1	75.3	40.4
	visual	73.3	17.9	87.6	49.6	76.8	31.9	55.3	20.3	73.3	29.9
	discovery	71.5	7.8	87.3	36.8	75.6	23.2	55.0	17.0	72.3	21.2
	FedTaxo-C	75.3	30.0	89.4	60.3	78.4	40.0	56.0	24.7	74.8	38.8
	FedTaxo-B	73.4	19.5	83.4	21.9	77.4	34.8	54.7	16.6	72.2	23.2
	FedTaxo-A	75.1	25.0	82.2	11.4	77.6	34.4	56.1	24.3	72.7	23.8
	FedTaxo	**76.1***	**35.1**	**91.3***	**71.5**	80.4*	49.2	**56.4**	**26.9**	**76.0**	**45.7**
	FedTaxo-L	75.7	32.1	90.7	67.8	**80.6**	**49.6**	56.1	24.4	75.8	43.5
	oracle	*77.0*	*42.0*	*91.6*	*74.1*	*81.5*	*54.5*	*56.5*	*27.8*	*76.6*	*49.6*

greedy, 54% for text alignment and 41% for discovery. It is the best non-oracle front-end on CIFAR-100 and Aircraft at five shots and on those plus *mini*ImageNet at one (14 of 16 tests survive Holm correction). The exception is CUB-200, where FedTaxo-L leads and FedTaxo does not beat text. Text alignment gains 0.4–4.8 Avg once it too receives the constraints. At 10 and 20 clients every front-end gains up to 1.9 Avg, FedTaxo staying best on CIFAR-100 and lowest of the three leaders on CUB-200 (Table 8).

To rule out the training-free classifier as the explanation, we repeat the study on a trained prompt-based method, a frozen ViT-B/16 with a base-session prompt and prototype aggregation keyed by the front-end's estimated class ID, reimplemented because UOPP's public code (Ma'sum et al., 2025) lacks its backbone file (Table 21). FedTaxo matches the oracle at five shots (86.8 against 86.8, novel 82.1 against 82.2) and reaches 83.7 against 84.2 at one, significantly above every other front-end, FedTaxo-L included ($p = 0.014$ and 0.004). The text rules gain from the constraints (85.5 and 82.6, from 83.5 and 81.4) yet stay 1.3 points below at five shots, and the hand-designed FedTaxo-B and -A collapse at one shot (26.9 and 7.6 novel). Fed-L2P from the UOPP repository, run unchanged with our label maps injected, forgets so heavily that not federating beats every merging front-end, the oracle included, so we draw no claim from it (Table 22).

Takeaway 2

On average, a constrained, revisable assignment recovers most of what a true correspondence gives, with a training-free or a trained classifier; which score leads depends on the dataset.

4.3 RQ3: Recovering real hierarchies

To answer RQ3, two of the five clients label at the superclass level, so a coarse capsule covers several fine concepts and a front-end may keep it apart, merge it into a member, or link it by a sub edge, which requires a rule that predicts the relation. Table 3 reports the accuracy of the coarse and the fine clients, each marked against not federating, and the recall and precision of the true coarse-to-fine edges (Avg and merge rates in Table 19).

Table 3: Hierarchy recovery in the HIER regime at five shots (%).

	CIFAR-100			*mini*ImageNet			CUB-200			Aircraft		
Front-end	coarse	fine	edges	coarse	fine	edges	coarse	fine	edges	coarse	fine	edges
independent	58.0	55.4	0/–	83.3	83.1	0/–	69.5	68.5	0/–	37.8	36.5	0/–
text-NN	60.8✓	57.6✓	0/–	84.9✓	84.3✓	0/–	73.4✓	71.5✓	0/–	46.9✓	44.4✓	0/–
visual	63.9✓	61.3✓	0/–	87.1✓	87.5✓	0/–	74.2✓	73.5✓	0/–	49.6✓	46.9✓	0/–
discovery	53.9✗	**67.6**✓	0/–	87.9✓	**90.6**✓	0/–	67.0✗	**77.3**✓	0/–	46.7✓	43.8✓	0/–
FedTaxo-C	61.0✓	62.8✓	1/67	85.7✓	86.9✓	0/0	73.6✓	72.6✓	0/30	46.7✓	46.3✓	0/28
FedTaxo	62.1✓	62.7✓	8/25	87.2✓	89.4✓	3/0	73.5✓	73.0✓	0/50	**49.8**✓	**48.3**✓	0/2
FedTaxo-L	**66.8**✓	66.3✓	53/83	**88.4**✓	89.4✓	3/50	**80.5**✓	76.1✓	47/84	47.4✓	47.3✓	2/56
oracle	*74.4*✓	*66.8*✓	100/100	*91.6*✓	*88.5*✓	100/100	*83.1*✓	*75.4*✓	100/100	*54.5*✓	*46.3*✓	100/100

FedTaxo-L recovers 53% and 47% of the true coarse-to-fine edges at 83% and 84% precision on CIFAR-100 and CUB-200, where the same procedure with the cosine score recovers 8% and 0%. It serves both client groups at once, with 66.8 coarse and 66.3 fine accuracy on CIFAR-100 and 80.5 and 76.1 on CUB-200, 2.9 and 6.3 coarse points above the best alternative. Merging cannot substitute: discovery merges 84% of those pairs and is the worst front-end for the coarse clients on CIFAR-100 (53.9, below not federating at 58.0, and below it on CUB-200 too), a coarse label in a fine node no longer naming its members.

The benefit is dataset-dependent. A group holds 5.0 classes on CIFAR-100 and 2.9 on CUB-200 but 1.3 and 1.4 on *mini*ImageNet and Aircraft, where most groups hold one class and sub and same-as then designate the same images (Table 32). The cause is partly ours and partly the benchmark's (Table 33): drawing single-member coarse capsules in the simulator lifts *mini*ImageNet recall 3→38 but costs CUB-200 13 points, so it stays out of the frozen recipe, while a non-degenerate coarse level alone lifts it to 94 at 99.5% precision. Aircraft resists every configuration, since a numeric variant code shares no token with "boeing" and WordNet covers neither. At an equal shot budget per coarse label, recall and precision fall while coarse accuracy is unchanged (Table 24).

Takeaway 3

Coarse labels become a typed hierarchy only with a rule that predicts the relation; where groups hold several classes, the learned matcher recovers about half of the true edges and serves coarse and fine clients at once, while merging can leave coarse clients below not federating.

4.4 RQ4: Beyond the simulator

We finally answer RQ4 by removing, one at a time, the assumptions the results so far rest on. Table 4 gathers four such departures at five shots and marks each front-end against not federating; under cross-cutting taxonomies (†) no correspondence exists and the oracle row merges by ground-truth class overlap.

FedTaxo-L's simulator uses the same resources as the test corruption; trained on synonyms or homonyms alone it stays within 0.5 of the default, without pseudo-coarse labelers it types no edge and loses 2–5 coarse points, held-out paraphrases keep both variants within 0.5 of the oracle, and dropping the WordNet features costs CIFAR-100 edge recall but not Avg (Tables 18–26, 9). Vocabularies an LLM writes for five personas keep the ordering but shrink the margins (Table 20): FedTaxo leads five of eight columns, discovery two and visual one, its gap to the oracle widens from 0.5 to 0.0–3.2, and at five shots it slips below not federating on CUB-200 and, by 0.1, on Aircraft (Table 4).

Perturbations separate the two assumptions FedTaxo rests on (Table 28). Isotropic Gaussian noise on every shot embedding costs it what it costs the oracle (-6.7 against -6.9 at σ=1); we tested no other noise model. Duplicate labels cost it 2.1 and 3.3 points at p=0.2 and 0.4 while the oracle gains 0.2 and 0.5, since a duplicate violates exactly the cannot-link FedTaxo exploits. A softer rule, letting a client's two labels merge above $\tau + \sigma$, recovers 38% and 53% of that damage at p=0.4, beats dropping the constraint in all six cells, and changes Avg by -0.69 on CIFAR-100 and $+0.53$ on CUB-200 where every constraint is correct (Table 30). Re-extracting every image feature with CLIP ViT-B/32 lowers Avg by 6.0–19.8 points but moves FedTaxo's share of the oracle gain by at most 6 points, and every merging front-end stays above not federating (Tables 4 and 31); five-shot

Table 4: Avg under four departures from the default protocol at five shots (%).

Front-end	LLM-written vocabularies CIFAR-100	*mini*ImageNet	CUB-200	Aircraft	Unaligned base session CIFAR-100	*mini*ImageNet	CUB-200	Aircraft
independent	78.9	93.1	84.2	57.7	78.0	93.0	84.2	56.3
text-NN	79.4 ✓	93.1 =	84.2 =	57.7 =	64.0 ✗	87.2 ✗	70.6 ✗	54.2 ✗
visual	80.5 ✓	93.7 ✓	85.9 ✓	**58.8** ✓	66.1 ✗	87.7 ✗	75.8 ✗	41.6 ✗
discovery	**80.9** ✓	93.8 ✓	**86.0** ✓	58.5 ✓	80.7 ✓	**94.0** ✓	**86.0** ✓	58.2 ✓
FedTaxo-C	79.1 ✓	93.4 ✓	84.2 =	57.7 =	74.0 ✗	89.4 ✗	79.0 ✗	54.6 ✗
FedTaxo	80.0 ✓	**93.9** ✓	83.7 ✗	57.6 ✗	**81.2** ✓	**94.0** ✓	84.8 ✓	**58.9** ✓
FedTaxo-L	79.7 ✓	93.8 ✓	85.5 ✓	57.7 =	80.6 ✓	**94.0** ✓	85.9 ✓	**58.9** ✓
oracle	*81.3* ✓	*93.9* ✓	*86.2* ✓	*59.6* ✓	*81.3* ✓	*94.0* ✓	*86.2* ✓	*59.0* ✓
	Weaker encoder (ViT-B/32)				Cross-cutting taxonomies†			
independent	66.9	86.6	69.3	38.4	79.2	92.8	83.4	57.8
text-NN	67.5 ✓	87.4 ✓	70.9 ✓	39.0 ✓	77.7 ✗	92.2 ✗	80.4 ✗	56.9 ✗
visual	67.8 ✓	87.5 ✓	71.2 ✓	39.0 ✓	79.6 ✓	92.5 ✗	83.9 ✓	**58.1** ✓
discovery	68.3 ✓	**88.0** ✓	71.7 ✓	39.0 ✓	**79.9** ✓	92.4 ✗	**84.4** ✓	57.9 ✓
FedTaxo-C	68.2 ✓	87.7 ✓	71.2 ✓	39.3 ✓	79.4 ✓	92.7 ✗	83.6 ✓	58.0 ✓
FedTaxo	**68.6** ✓	87.9 ✓	70.6 ✓	**39.5** ✓	79.2 =	92.7 ✗	83.6 ✓	58.0 ✓
FedTaxo-L	68.3 ✓	**88.0** ✓	**71.9** ✓	**39.5** ✓	78.8 ✗	**92.8** =	84.1 ✓	57.8 =
oracle	*68.8* ✓	*88.0* ✓	*72.0* ✓	*39.6* ✓	*78.6* ✗	*91.0* ✗	*81.8* ✗	*57.6* ✗

CUB-200 is again the exception, where the cosine falls 0.2 below string matching while the learned matcher keeps 94% of the gain.

The shared base session gives a front-end two things, which we withdraw in turn. A base class may come back under a private name: FedTaxo routes 90–99% of such capsules to the right base node against 57–97% assigned greedily, though that test keeps the canonical base graph and so shows only matching into an aligned graph (Table 34). With the alignment itself removed, every client holding every base class under its own label, FedTaxo rebuilds the canonical partition, merging 97–100% of cross-client base pairs into one node per class within 0.36 Avg of the aligned protocol, as does label-free discovery, while every greedy rule ends below not federating: on CIFAR-100 the cosine falls to 74.0 and the text and visual rules to 64–66, against 78.0 (Tables 4 and 35). The threshold τ is still chosen on canonically labeled base classes. Re-clustering costs 0.4–1.0 s per session with the cosine and 0.9–12.4 with the learned matcher (Table 29). Where clients do not refine a common structure but each partitions the novel classes its own way, no method helps consistently and merging by ground-truth overlap falls below not federating on all four datasets (Table 4, Appendix D.2).

Takeaway 4

The diagnosis survives a weaker encoder, an unaligned base session and unseen vocabularies; duplicate labels within a client, which break the cannot-link constraint, erode it, and no method helps consistently when client taxonomies cross-cut.

5 Conclusion

Federated incremental learning usually assumes a shared symbol space, an assumption that stays invisible while benchmarks split one labeled dataset across clients. Taxonomy-Free FFSCIL removes it, and FedTaxo assigns client capsules to a concept graph under client constraints, revisably across sessions. The difficulty lies less in the matching score than in the procedure that applies it: a plain cosine under constrained re-clustering recovers 85% of what a true correspondence gives, and our 18-feature matcher adds nothing on average on top of it, though it turns a flat merge into a typed hierarchy. The winning configuration is standard by design; the contribution is the formulation, the benchmark and the diagnosis of which component carries the result, and two of our designs, a calibrated-covariance score and joint optimal-transport assignment, did not survive the protocol.

This remains a study of one setting. Client taxonomies are simulated and the LLM vocabularies are synthetic, so nothing here establishes behaviour on label systems that real institutions maintain; the within-client constraint is wrong whenever a client holds two identifiers for one concept, and a soft form recovers only 38–53% of that loss; hierarchy recovery fails on Aircraft; and capsules are feature-sharing, not private. Evidence from independently labeled data is future work.

Reproducibility Statement

All hyperparameters, splits, corruption seeds, and the dry-run selection procedure that every front-end shares are listed in Appendix C, with the selected values in Table 7; every reported number is a mean over 5 seeds (3 for the 10- and 20-client sweep), and every significance claim is a one-sided paired t-test over those seeds with Holm correction. Proofs of Theorems 1 to 3 are in Appendix G. The anonymous repository anonymous.4open.science/r/fedtaxo-iclr27-FF2A contains the code (the corruption protocol, every front-end, the from-scratch UOPP core, and the soft-constraint, unaligned-base and alternative-coarse-level variants), the scripts that extract the CLIP features of both encoders, the LLM-written vocabularies, the FedL2P mismatch wrapper as a patch against the UOPP repository, every result file with the command that produced it, and scripts that regenerate every table from those files.

Use of Large Language Models

Large language models enter this work in two distinct roles, and we separate them. As an *object of study*, Qwen2.5-7B-Instruct writes the persona vocabularies of the LLMVOC regime (Sec. 4.4 and Appendix C); those prompts, decoding settings and outputs are in the code repository. As an *assistant*, large language models supported the authors in three ways. *Writing:* a large language model drafted and revised text and improved its grammar, word choice and conciseness; no text was included without author review. *Literature:* it helped identify relevant work; the authors independently read every suggested paper and verified each citation against its original source before including it. *Experiments:* Claude Code assisted with designing and running the experiments, writing and refactoring analysis code, analyzing the results, and building tables and figures from the result files; the authors reviewed the code and verified every reported number against the saved outputs, from which the released scripts regenerate all tables. All claims are the authors' own, no AI system is an author, and the authors take full responsibility for the content of this paper, including any remaining errors.

References

Jayadev Acharya, Clément L. Canonne, and Himanshu Tyagi. Distributed signal detection under communication constraints. In *Conference on Learning Theory (COLT)*, volume 125 of *Proceedings of Machine Learning Research*, pp. 41–63, 2020.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26, 2013.

Zhipeng Deng, Luyang Luo, and Hao Chen. Scale federated learning for label set mismatch in medical image classification. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 118–127, 2023. doi: 10.1007/978-3-031-43898-1_12. FedLSM.

Haiyang Guo, Fei Zhu, Wenzhuo Liu, Xu-Yao Zhang, and Cheng-Lin Liu. PILoRA: Prototype guided incremental LoRA for federated class-incremental learning. In *European Conference on Computer Vision (ECCV)*, pp. 141–159, 2024. doi: 10.1007/978-3-031-73650-6_9.

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.

Yalan Jiang, Yang Cheng, Dan Wang, and Bin Song. RRA-FFSCIL: Inter-intra classes representation and relationship augmentation federated few-shot incremental learning. *Neurocomputing*, 597:127956, 2024. doi: 10.1016/j.neucom.2024.127956.

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

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. doi: 10.1214/aos/1015957395.

Cuiwei Liu, Siang Xu, Huaijun Qiu, Jing Zhang, Zhi Liu, and Liang Zhao. Few-shot class-incremental learning with non-IID decentralized data. *IEEE Internet of Things Journal*, 12(22): 48556–48572, 2025. doi: 10.1109/JIOT.2025.3605647. F2SCIL.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Muhammad Anwar Ma'sum, Mahardhika Pratama, Savitha Ramasamy, Lin Liu, Habibullah Habibullah, and Ryszard Kowalczyk. PIP: Prototypes-injected prompt for federated class incremental learning. In *ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 1670–1679, 2024. doi: 10.1145/3627673.3679794.

Muhammad Anwar Ma'sum, Mahardhika Pratama, Lin Liu, Habibullah Habibullah, and Ryszard Kowalczyk. Federated few-shot class-incremental learning. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=ZiPoAlKf9Y. UOPP.

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 (AISTATS)*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282, 2017.

George A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995. doi: 10.1145/219717.219748.

Nan Pu, Wenjing Li, Xingyuan Ji, Yalan Qin, Nicu Sebe, and Zhun Zhong. Federated generalized category discovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28741–28750, 2024. doi: 10.1109/CVPR52733.2024.02715.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12180–12189, 2020. doi: 10.1109/CVPR42600.2020.01220.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7482–7491, 2022. doi: 10.1109/CVPR52688.2022.00734.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016. miniImageNet.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Lixu Wang, Chenxi Liu, Junfeng Guo, Qingqing Ye, Heng Huang, Haibo Hu, and Wei Dong. Federated continuous category discovery and learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2429–2439, 2025. doi: 10.1109/ICCV51701.2025.00234. FC2DL.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general Chinese embeddings. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 641–649, 2024. doi: 10.1145/3626772.3657878. BGE embeddings.

Shuo Yang, Lu Liu, and Min Xu. Free lunch for few-shot learning: Distribution calibration. In *International Conference on Learning Representations (ICLR)*, 2021.

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 *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 3297–3308, 2023. doi: 10.1145/3580305.3599443. FedAlign.

Da-Wei Zhou, Fu-Yun Wang, Han-Jia Ye, Liang Ma, Shiliang Pu, and De-Chuan Zhan. Forward compatible few-shot class-incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9036–9046, 2022. doi: 10.1109/CVPR52688.2022.00884. FACT.

A Related works

A first line of related work studies few-shot class-incremental learning and its federated extensions. Prototype-based FSCIL learns new classes from a handful of examples by classifying against class means, in the original formulation (Tao et al., 2020), through forward-compatible training (Zhou et al., 2022) and with the prototypical classifier itself (Snell et al., 2017), while distribution calibration borrows base-class statistics to stabilize few-shot estimates (Yang et al., 2021). RRA-FFSCIL (Jiang et al., 2024) and F2SCIL (Liu et al., 2025) bring the setting to federated and non-IID decentralized data, UOPP formalizes it with prompt-based federated learning (Ma'sum et al., 2025), and PIP and PILoRA address federated class-incremental learning with prompts and incremental LoRA (Ma'sum et al., 2024; Guo et al., 2024). These methods are natural classifiers for our setting, and Sec. 4.2 repeats our study on a trained prompt-based one. Every one of them, however, assumes that new classes arrive already indexed into one global label set. Our contribution is complementary: we make that indexing step the object of study.

A second line of work relates classes across parties without a shared vocabulary. Generalized category discovery clusters labeled and unlabeled samples into known and novel categories (Vaze et al., 2022), with federated (Pu et al., 2024) and continual (Wang et al., 2025) variants; it outputs a flat partition and uses neither the labels clients do have nor per-client label spaces, which is why discovery serves as a label-free baseline in our experiments. Text-side alignment instead relates labels through their embeddings, as a CLIP or BGE text encoder does implicitly (Radford et al., 2021; Xiao et al., 2024). Our experiments show that both help once they receive the constraints a client already knows, and that neither separates homonyms from synonyms as reliably as a fused score under revisable assignment.

A third line of work addresses non-identical client label sets in federated learning. FedLSM handles clients that annotate different subsets of one known global label set by uncertainty-based pseudo-labeling (Deng et al., 2023): label sets are partial but never conflicting, and identity is given. FedAlign drops the overlap requirement and anchors client class representations to their label names, the closest prior use of text as the alignment medium, but assumes that a name identifies the class, so shared strings and subsumption fall outside it (Zhang et al., 2023). CrowdFed matches categories across crowdsourced annotators with inconsistent label representations (He et al., 2026), the closest setting to ours. Our formulation differs in three respects: classes arrive incrementally at one to five shots, so each identity decision compounds through later sessions; the relation to be predicted includes subsumption and not only identity; and each client keeps predicting in its own label space rather than in a reconciled global one. Our assignment uses standard tools, constrained agglomerative clustering and, in an ablation, entropic optimal transport (Cuturi, 2013); what we contribute is the formulation, the benchmark and the controlled diagnosis of which component recovers the gain of a true correspondence.

B Method details

This appendix states the steps of Sec. 3 formally. Equations (2), (3) and (4) are not repeated. Figure 2 shows every module in one diagram.

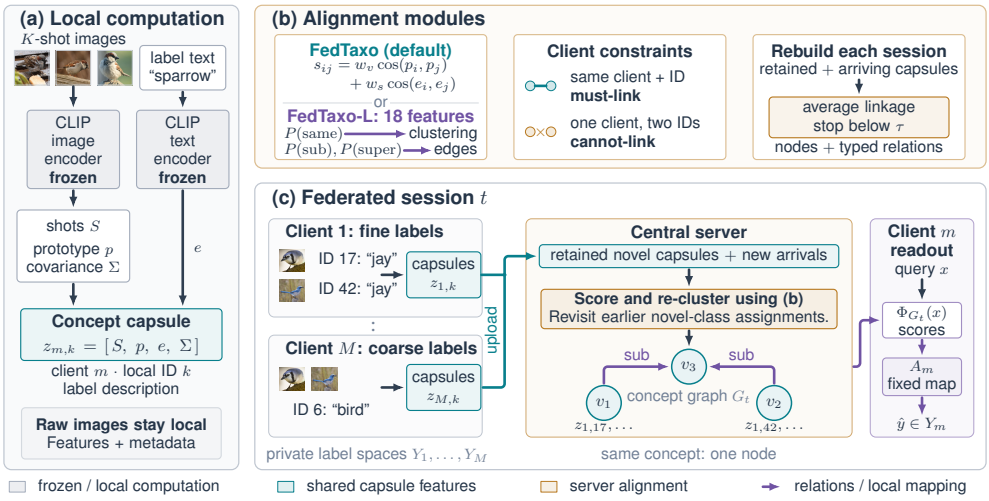

Figure 2: FedTaxo in detail. (a) capsule construction on each client; (b) pairwise scoring and constrained re-clustering on the server; (c) the concept graph of one session with typed edges and each client's readout in its own label space. Bird images are from CUB-200 (Wah et al., 2011).

B.1 Capsule statistics

Let $\mathcal{X}^t_{m,k}$ contain the images that client m receives under local identifier k at session t, with description $\ell_{m,k}$ and size $n_z = |\mathcal{X}^t_{m,k}|$. A fine-class capsule contains K shots; a coarse capsule can pool shots of several fine classes and therefore contain more than K images. Write $\nu(a) = a/\|a\|_2$ for normalization, and let $s_1, \dots, s_{n_z}$ be the shot embeddings of Eq. (2), whose visual and semantic summaries p_z and e_z that equation also defines.

Covariance and spread. Few-shot sample covariance is unreliable, especially for a single shot. We fit a $d_p = 64$ dimensional PCA projection P_0 on base-session image embeddings and write $a_i = P_0(s_i - \bar{s}_0)$, where $\bar{s}_0$ is the base centering vector. The projected mean μ_z and coordinate-wise sample variance $\widehat{v}_z$ are then

$$\mu_z = \frac{1}{n_z}\sum_{i=1}^{n_z} a_i, \qquad \widehat{v}_z = \begin{cases} \dfrac{1}{n_z - 1}\displaystyle\sum_{i=1}^{n_z}(a_i - \mu_z)^{\odot 2}, & n_z > 1, \\ 0, & n_z = 1. \end{cases} \tag{6}$$

Here $\odot 2$ means coordinate-wise squaring. Let $\mathcal{N}_3(\mu_z)$ be the three nearest base-class means in this projected space and let v_b be the diagonal variance vector of base class b. Following distribution calibration (Yang et al., 2021), we use

$$\Sigma_z = \operatorname{diag}\left(\eta\widehat{v}_z + (1-\eta)\frac{1}{3}\sum_{b\in\mathcal{N}_3(\mu_z)} v_b + \epsilon\mathbf{1}\right), \qquad \eta = 0.5, \quad \epsilon = 10^{-4}. \tag{7}$$

The uncalibrated spread $a_z^{\mathrm{sp}} = \mathbf{1}^\top\widehat{v}_z$ is available only when $n_z > 1$. The visual prototype p_z remains in the full encoder space; PCA is used for these auxiliary statistics. Covariance supports the hand-designed ablations in Appendix E; the default cosine score does not use it.

Message and disclosure. We denote the capsule by $z = [S_z, p_z, e_z, \Sigma_z]$, accompanied by its description and client-local key; shot count, projected mean and spread are derived summaries. Only features and metadata leave the client. Raw pixels and all backbone, prompt and head parameters remain local. Because shot embeddings are deterministic functions of the images, this is feature sharing without a formal privacy guarantee; Sec. 4.4 measures perturbation sensitivity, not differential privacy or inversion resistance.

B.2 Group statistics and the learned matcher

The matcher compares capsules or groups of capsules. For any group U, define its normalized prototype and text embedding by

$$p_U = \nu\left(\frac{\sum_{z\in U} n_z p_z}{\sum_{z\in U} n_z}\right), \qquad e_U = \nu\left(\frac{1}{|U|}\sum_{z\in U} e_z\right), \qquad S_U = \biguplus_{z\in U} S_z. \tag{8}$$

Here $\uplus$ concatenates the shot sets, retaining multiplicities. Visual prototypes are weighted by shot count; textual summaries weight capsules equally. These definitions also give the statistics of the final concept nodes.

Learned relation score. FedTaxo-L predicts $q_\omega(r \mid U, V)$ over $\mathcal{R} = \{\text{different}, \text{same}, \text{super}, \text{sub}\}$, where super means that U is broader than V. Its 18-dimensional feature vector $\varphi(U, V)$ contains channel cosines, string equality, spread and shot counts, shot-set structure, and lexical and vocabulary-depth cues (Table 6). A second, frozen BGE text embedding (Xiao et al., 2024) is computed from the same description. For example, the directed shot-set feature is

$$d_{\to}(U, V) = \frac{1}{|S_U|}\sum_{s\in S_U} \max_{s'\in S_V} \langle s, s' \rangle. \tag{9}$$

Both directions are retained: a fine class may be well represented inside a coarse shot set even when the reverse coverage is weaker. The remaining shot-set features measure the cosine separation of a deterministic two-means split of each set.

To train q_ω, the labeled base classes simulate private client vocabularies at the deployment shot count. Fine capsules receive synonyms or shared homonyms; pseudo-coarse capsules pool two to four base classes under a superclass name. Singletons and grouped capsules provide ordered training pairs $\mathcal{D}_0 = \{(U_i, V_i, r_i)\}$. Relation targets use known base-class and superclass identities: equal identities give same, superclass membership gives the directed super/sub target, and unrelated identities give different. The target therefore refers to the declared concept, rather than equality of the particular images sampled into two capsules. Histogram gradient boosting fits the weighted multiclass log-loss

$$\mathcal{L}_{\text{rel}}(\omega) = -\frac{1}{\sum_i \alpha_{r_i}}\sum_i \alpha_{r_i} \log q_\omega(r_i \mid U_i, V_i), \qquad \alpha_r = \sqrt{\frac{N_{\text{different}}}{\max(N_r, 1)}}. \tag{10}$$

Here N_r counts retained training pairs of relation r; implementation settings and pair sampling are in Appendix C. Since the features are directed, we symmetrize the same-concept score used for clustering:

$$s_{\text{L}}(U, V) = \tfrac{1}{2}\big[q_\omega(\text{same} \mid U, V) + q_\omega(\text{same} \mid V, U)\big]. \tag{11}$$

The directed relation probabilities are retained for hierarchy construction. Neither the matcher nor its training targets use novel-class correspondences.

B.3 Constrained re-clustering

At session t, let $\mathcal{Z}_t$ contain the arriving novel capsules and let $\mathcal{H}_t = \mathcal{H}_{t-1} \uplus \mathcal{Z}_t$ retain all novel capsules observed so far. Shared base nodes are excluded from this pool. We first form one *must-linked unit* per client-local identifier:

$$U^t_{m,k} = \{z \in \mathcal{H}_t : (m_z, k_z) = (m, k)\}, \qquad \mathcal{U}_t = \{U^t_{m,k} : U^t_{m,k} \neq \varnothing\}. \tag{12}$$

Repeated messages under one identifier stay together even if they arrive in different sessions. Different identifiers of the same client must stay apart, regardless of whether their descriptions coincide.

Let $\mathcal{P}$ be a partition of $\mathcal{U}_t$, initially its singleton partition. A cluster $A \in \mathcal{P}$ holds units, and $\mathcal{M}(A)$ is their set of client IDs. Two clusters are admissible for merging exactly when $\mathcal{M}(A) \cap \mathcal{M}(B) = \varnothing$. Thus each resulting concept contains at most one local identifier from a given client. With $s = s_{\cos}$ for FedTaxo or $s = s_{\text{L}}$ for FedTaxo-L, constrained average linkage is Eq. (4). While at least two clusters remain, select distinct $(A^\star, B^\star)$ maximizing $L(A, B)$. If their score is at least τ, replace them by $A^\star \cup B^\star$; otherwise stop. Pair scores are computed on the original must-linked units and

remain fixed within this session. After an admissible merge, ordinary average linkage updates the score to another admissible cluster C as

$$L(A \cup B, C) = \frac{|A|L(A,C) + |B|L(B,C)}{|A| + |B|}; \quad \mathcal{M}(A \cup B) = \mathcal{M}(A) \cup \mathcal{M}(B). \tag{13}$$

If the updated client sets overlap, the corresponding score is instead masked by $-\infty$. The weights count units, rather than shots or capsules.

The final partition $\mathcal{P}_t$ defines the novel nodes: $V_t = V_0 \cup \{v_A : A \in \mathcal{P}_t\}$, with member capsules $\mathrm{caps}(v_A) = \biguplus_{U \in A} U$. Every session starts again from the units of Eq. (12), dissolving the previous novel nodes and their edges. An earlier merge can therefore be split, or an earlier split merged, as new evidence arrives. This permits correction without guaranteeing it for an imperfect score; Proposition 1 gives exact recovery under a separating matcher. Greedy assignment is the ablation that commits each capsule on arrival and cannot reconsider its partition.

B.4 Concept graph and directed hierarchy

Each node stores its capsules and the summaries in Eq. (8). Same-concept assignment is represented by membership in one node. A directed edge $(v, u) \in E_t$ instead means that v is a sub-concept of u; its reverse expresses the corresponding super relation. Broader and narrower concepts retain separate prototypes.

FedTaxo-L edges. After re-clustering, the learned matcher scores ordered pairs of distinct novel nodes. A candidate parent u and child v must satisfy

$$1 - q_\omega(\text{different} \mid u, v) \geq \tau, \qquad \operatorname*{arg\,max}_{r \in \{\text{same,super,sub}\}} q_\omega(r \mid u, v) = \text{super}. \tag{14}$$

Candidates are processed in descending $q_\omega(\text{super} \mid u, v)$ order. We insert $v \to u$ only if v has no parent and the edge creates no directed cycle. Consequently the implemented hierarchy is a forest: a node has at most one parent and may have multiple children. The relatedness test above uses the same selected threshold as assignment, but a different probability.

Default edges. Without a learned matcher, candidate node pairs must satisfy $s_{\cos}(u, v) \geq \tau$. Each node's spread is the shot-count-weighted mean of its members' available uncalibrated spreads. If both spreads are available, a ratio exceeding $\gamma = 2$ marks the larger-spread node as broader; a ratio below $1/\gamma$ marks it as narrower. Otherwise, or when the ratio lies between these bounds, a head-noun/suffix or WordNet-hypernym cue determines the direction when available. Only pairs typed as broader-to-narrower propose an edge, processed by decreasing cosine score under the same one-parent and cycle checks. This heuristic does not affect the already computed partition; its limited hierarchy recovery is measured in Sec. 4.3.

B.5 Prediction in each client's private label space

The graph does not impose a shared output vocabulary. For client m, first map each shared base node to its canonical base label and each node containing that client's capsules to their local identifier. Under the hard constraint, all of a client's capsules in one novel node have the same identifier. A node owned by a coarse local label also makes its fine descendants accessible: with the edge direction above, v is a descendant of u when a path $v \to \cdots \to u$ exists. Previously unmapped descendants inherit u's local label. Direct memberships take priority over inherited labels; a fixed traversal order resolves multiple inherited assignments. In ablations permitting conflicting memberships, the first assigned local label is retained.

This constructs a partial map $A_{m,t} : V_t \rightharpoonup Y_m^{(\leq t)}$ and an accessible node set $V_{m,t} = \mathrm{dom}(A_{m,t})$. For a query image x, prediction is nearest-prototype classification followed by this map:

$$v_m^\star(x) = \operatorname*{arg\,max}_{v \in V_{m,t}} \langle \nu(f_\theta(x)), p_v \rangle, \qquad \widehat{y}_m(x) = A_{m,t}\big(v_m^\star(x)\big). \tag{15}$$

A coarse client can therefore use a fine node's more specific visual prototype while predicting its own broader label. The map is recomputed from the graph by a fixed rule; no adapter parameters are trained, and it is a heuristic for the objective in Eq. (1).

B.6 Base-session calibration and the session procedure

With encoder and score weights fixed, the selected scalar for FedTaxo is the merge threshold τ. For cosine matching, candidate rules are the thresholds maximizing F_β on simulated base-class pairs for $\beta \in \{0.5, 1\}$, together with fixed values $\{0.3, 0.5, 0.7\}$. FedTaxo-L uses that fixed probability grid. The selection criterion evaluates assignment and prediction together, rather than pairwise matching alone.

Split the base classes into three folds. In fold f, one fold plays pseudo-novel classes and the others supply pseudo-base covariance statistics and, for FedTaxo-L, matcher training. Keep the PCA projection fitted on the original base session. For each candidate rule h, re-estimate its threshold on the pseudo-base classes when necessary, then run the incremental protocol with the same shot count and corruption on two seeds. If $\mathrm{Acc}^{\mathrm{pn}}_{f,r,t}(h)$ is pseudo-novel accuracy at incremental session t, select

$$h^\star = \arg\max_{h \in \mathcal{H}} \frac{1}{3 \cdot 2} \sum_{f=1}^{3} \sum_{r=1}^{2} \frac{1}{T_f} \sum_{t=1}^{T_f} \mathrm{Acc}^{\mathrm{pn}}_{f,r,t}(h). \tag{16}$$

Finally materialize $h^\star$ and refit the learned matcher, when used, on the full base session. The selected threshold is then fixed during novel sessions. Every front-end uses this base-only dry-run principle; no actual novel-class labels enter selection. The complete candidate grids and simulator settings appear in Appendix C.

Session procedure. Starting from the shared base graph G_0, each novel session executes the following steps:

1. Clients encode their local classes and upload capsules with client-local metadata using Eqs. (2)–(7).
2. The server appends arrivals to $\mathcal{H}_t$, forms must-linked units, and scores every unit pair by Eq. (3) or (11).
3. It re-clusters the complete novel pool with the admissible merges of Eq. (4), keeping the base nodes fixed.
4. It rebuilds node summaries, attaches hierarchy edges, and derives each client's accessible prototypes and fixed label map.
5. Clients classify queries by Eq. (15); the server retains the capsules for revision in the next session.

The federated update aggregates evidence by inferred concept membership. It exchanges feature summaries and label mappings without averaging backbone parameters or learning a global classifier head.

C Hyperparameters and experimental details

Every hyperparameter is either a fixed constant listed here or selected by the dry run on the base session; nothing is tuned on novel-class data. Encoders: CLIP ViT-L/14 (Radford et al., 2021), frozen, for Secs. 4.1 to 4.3 and Appendices D.1 and F; the trained comparison (Sec. 4.2) instead uses a frozen timm ViT-B/16 with 10 learnable prefix prompt tokens, trained on the base session for 3 epochs, batch size 128, Adam, learning rate 0.005, gradient-norm clipping at 1.0. Fed-L2P (Sec. 4.2) runs the UOPP repository code unchanged: ViT-B/16 with a prompt pool of 10 and top-1 selection, 20 clients with 6 sampled per round, 10 rounds per session, 10 local epochs in few-shot sessions, learning rate 0.03, 5 shots, 5 seeds; a sampled client that observes none of a session's classes is skipped for that round. **Matcher** (FedTaxo-L): scikit-learn histogram gradient boosting (200 iterations, learning rate 0.08, 15 leaves, ℓ_2 regularization 1) on 18 pair features (Sec. 3), trained on 300 corrupted fine capsules over M simulated clients plus 40 pseudo-coarse capsules from two coarse clients, grouped into singletons, 1–3-capsule fine units and 1–2-capsule coarse units, every ordered pair labeled from its declared base-class or superclass identity, at most 60,000 "different" pairs kept, classes weighted by the square root of their inverse frequency; $\tau \in \{0.3, 0.5, 0.7\}$ by the dry run (0.3 selected everywhere); WordNet depths from the first three noun synsets, capped shot counts of 40. **Score** (Eq. 17): FedTaxo-C and FedTaxo-B use $w_v = 0.6$, $w_s = 0.4$, and

FedTaxo-B has no reliability floor; FedTaxo-A selects $w_v \in \{0, 0.25, 0.5, 0.75, 1\}$, $\rho_{\min} \in \{0, 1\}$ and $\beta \in \{0.5, 1\}$ by the dry run (three folds, two seeds each, scored by pseudo-novel-class accuracy; Table 7 lists the winners), and the remaining fixed hand-designed scores select β from the same set. **Typed edges**: the hand-designed variants type a match as sub/super when the spread ratio exceeds $\gamma = 2$ or the labels are lexically related (one is a suffix or head-noun token of the other, or a WordNet hypernym of it in any sense); FedTaxo-B and FedTaxo-A apply the same test to node pairs in the merge pass. Every front-end matches new capsules among novel nodes only, and every client's adapter maps each accessible node to exactly one of its labels (the shared base label for base nodes, otherwise the first of its classes to reach the node), so a wrong merge costs the class that lost the node. **Baselines**: text-strong embeds label strings with BGE-large-en-v1.5 (Xiao et al., 2024) (CLS pooling, L2-normalized); discovery runs average-linkage agglomerative clustering of visual prototypes under cosine distance, cut at the number of clusters that maximizes the silhouette over $k = 2, \ldots, n-1$, re-run over the whole novel pool at every session and scored by dendrogram height in the alignment study. **Distribution calibration**: each capsule's covariance blends its own shot variance and the mean covariance of its three nearest base classes with weight $\eta = 0.5$ and diagonal offset 10^{-4} (Eq. (7)), following Yang et al. (2021). **PCA**: auxiliary means μ_z and diagonal covariances are computed in a 64-d PCA space fit on base-class features only; visual prototypes p_z remain in the full encoder space (Sec. B.1). **Threshold tuning**: 200 base-class capsules under synonym and homonym corruption are scored in unordered pairs, with each target represented as a singleton node; the F_β-optimal threshold on these pairs is the candidate for each β, never touching novel-class data. **Protocol**: 5 clients; 40 base classes and the rest novel for the alignment study (Appendix D.1), or the UOPP base/session split for Sec. 4.1 and Appendix F (60 base $+ 8\times5$-way for CIFAR/*mini*, 100 base $+ 10\times10$-way for CUB); each session's new classes are observed by $\lceil 0.6M \rceil$ of the M clients. **Corruption**: synonyms replace a label by another form of the class name, drawn from a curated list of two to three forms per class on CIFAR-100 and from the WordNet synset lemmas on the other datasets, at a random subset of clients; real homonyms are groups of classes that share a terse name, their common head noun or, failing that, their superclass name (CIFAR-100's given superclasses, WordNet hypernyms two levels up for *mini*ImageNet, the genus head noun for CUB-200, the manufacturer family for Aircraft), up to 8 groups per run with at least two novel members, whose members are relabeled with the shared terse name at some clients. **LLM-written vocabularies** (Sec. 4.4; prompt and responses in Appendix C.1): Qwen2.5-7B-Instruct, greedy decoding, is asked for each class name and each of five personas (a warehouse worker who drops every modifier, a domain expert using the technical name, a layperson in plain words, a data-entry clerk using codes of at most eight characters, a hobbyist using the common name with one adjective) to write the label it would use; the reply is lowercased, stripped of quotes, and cut to five words, and client m adopts persona m; no other corruption is applied, and the matcher and τ are still trained and selected on the WordNet-based simulator. **Perturbations**: with probability p a client sends a second capsule of a new class under another form of its name (or the terse or superclass name) as a separate local label; noise adds $\sigma \, \|x\| / \sqrt{d}$ Gaussian noise per coordinate to every shot embedding x (and its PCA projection) before renormalization. **Hierarchy regime**: two of the five clients label every novel class by its superclass name and pool the K shots of each member into one capsule; a coarse client is scored in its own label space, and its concept covers the fine nodes reachable by sub-edges. **Soft within-client constraint** (Sec. 4.4, Table 30): in the re-clustering path a pair of units carrying two labels of one client is admissible when its score exceeds $\tau + m\sigma$, where σ is the standard deviation of the off-diagonal pair scores of the current session and m is the strength, $m{=}1$ throughout; $m{\to}\infty$ is the hard constraint used elsewhere in the paper, and it is the hard constraint that every other result uses, so the soft rule changes nothing outside that table. In the greedy path the same rule is applied to the candidate scores of each arriving capsule. **Base-class reappearance** (Sec. 4.4, Table 34): every client draws a disjoint random half of the base classes, spreads them over the sessions, and re-declares each under a private synonym of its name, with base nodes added to the candidate set for those capsules only. The shared canonical base graph, the canonical base-to-label mapping at evaluation and an explicit flag marking a capsule as a returning base class are all retained, so the regime tests matching into an aligned graph rather than the absence of base alignment. In the re-clustering front-ends such a capsule is matched against the whole graph once, by the same score and threshold, and is not revised afterwards; the novel pool is re-clustered as before. Front-ends that only cluster the novel pool (discovery) have no mechanism for this and are omitted from that table. **Unaligned base session** (Sec. 4.4, Table 35): no canonical base node and no shared base label exist. Every client holds every base class under its own private label, drawn by the same corruption generator (synonyms and real homonyms) with 40 shots per client per class,

so the 200 base samples per class of the standard protocol are split across the five clients. The front-end aligns those capsules with the rule it uses for novel classes, the resulting nodes become the base graph, and each client's base classes are then scored in its own label space exactly like its novel ones, through the node that holds most of its capsules for that label. *merged* is the fraction of cross-client pairs of one base class that end in the same node. What this regime does *not* remove is labeled data at the server: the merge threshold is still selected by the dry run of Appendix C on base classes under simulated private vocabularies. **Single-member coarse units** (Sec. 4.3, Table 33): the matcher's training simulator draws a superclass uniformly from all base superclasses and pools 1 to 4 of its members, instead of drawing only superclasses with at least two base members and pooling 2 to 4, so that single-member coarse labels occur as often as the dataset's own hierarchy implies. **Coarser levels** (Table 32): the alternative coarse level is the WordNet hypernym four levels above the class synset for *mini*ImageNet and the manufacturer for FGVC-Aircraft, in place of two levels and the family. **Weaker encoder** (Table 31): all image features re-extracted with CLIP ViT-B/32 (512-d), everything else unchanged, label-text embeddings included, since the pipeline compares text with text and images with images and never mixes the two. **Repair**: the split pass uses deterministic 2-means (farthest-point initialization) with up to 8 iterations; both merge and split use the same tuned threshold. **OT assignment** (Appendix F): entropic regularization reg $= \max(10^{-3},\ 0.1 \cdot \mathrm{std}(\mathrm{cost}))$, 100 Sinkhorn iterations, reject-column cost fixed at the decision threshold. All results are means over 5 seeds; the 10- and 20-client sweep uses 3.

Table 5 lists the front-ends. Each decides whether an incoming capsule merges into an existing concept node or opens a new one; τ and every other hyperparameter are selected per front-end by the same dry run on the base session. The three hand-designed FedTaxo scores are the variants FedTaxo grew out of. The Text and Visual columns mark which capsule channel a rule reads, and discovery follows the generalized-category-discovery recipe and never reads a label.

Table 5: The front-ends compared.

Front-end	Merge rule for an incoming capsule	Text	Visual	Source
independent	never merges			no federation
string	label strings equal	✓		exact match
text-NN	CLIP text cosine $\geq \tau$	✓		Radford et al. (2021)
text-strong	BGE-large text cosine $\geq \tau$	✓		Xiao et al. (2024)
visual	K-shot prototype cosine $\geq \tau$		✓	Snell et al. (2017)
discovery	re-cluster novel prototypes, silhouette k		✓	GCD (Vaze et al., 2022)
FedTaxo-C	Eq. 17 on cosines, w_v=0.6, greedy, typed edges	✓	✓	this paper, hand-designed
FedTaxo-B	Eq. 17 on Bhattacharyya + reliability, repair	✓	✓	this paper, hand-designed
FedTaxo-A	FedTaxo-B with dry-run weights and joint OT	✓	✓	this paper, hand-designed
FedTaxo	Eq. 17 on cosines, constrained re-clustering, typed edges	✓	✓	this paper
FedTaxo-L	learned 4-way relation matcher, same re-clustering and edges	✓	✓	this paper, variant
oracle	true correspondence g_l			reference

Table 6 lists the features, computed for a capsule or node z against a node v; the last two groups are the shot-set and vocabulary features ablated in Table 9.

Table 6: The 18 pair features of the relation matcher.

Group	Features
Channel similarities	prototype cosine; CLIP label-embedding cosine; BGE label-embedding cosine; string equality
Size and spread	log spread ratio and its availability flag; log shot count of z and of v (capped at 40)
Shot sets	directed set similarities $z \to v$ and $v \to z$ (mean over one set of the maximal cosine to the other); bimodality of each set (cosine separation of a 2-means split)
Labels and vocabularies	lexical cue ($+1$ if z's label is broader by head-noun truncation or WordNet hypernymy, -1 if narrower); WordNet-depth difference and its flag; difference of the two clients' mean vocabulary depths and its flag; token-count difference

Table 7 gives the visual weight w_v (text weight $1 - w_v$), the reliability floor $\rho_{\min}$ and the F_β that sets the threshold, chosen per dataset and shot count by the three-fold dry run on the base session,

together with the dry run's mean pseudo-novel accuracy (%). A floor of 1 disables the reliability weight, and $w_v = 1$ ignores the text channel.

Table 7: Hyperparameters of FedTaxo-A selected by the dry run.

	5-shot				1-shot			
Dataset	w_v	$\rho_{\min}$	β	novel acc.	w_v	$\rho_{\min}$	β	novel acc.
CIFAR-100	1.00	1	1.0	74.9	0.00	0	0.5	38.2
*mini*ImageNet	1.00	1	0.5	91.5	0.75	1	1.0	62.2
CUB-200	1.00	1	1.0	85.4	0.75	0	0.5	46.7
FGVC-Aircraft	0.00	0	1.0	42.6	0.00	0	1.0	19.1
CIFAR-100, trained core	–	–	–	–	–	–	–	–

Table 8 repeats the five-shot SYNHOM loop with 5, 10 and 20 clients, 60% of which observe each new class (mean±std over 3 seeds for 10 and 20 clients, 5 seeds for 5). More clients pool more shots per concept and lift every front-end by up to 1.9 points. The broad picture is stable, but the exact ordering is not: on CIFAR-100 FedTaxo is the best non-oracle front-end at every M, whereas on CUB-200 it trails FedTaxo-L and discovery by 1.0, 0.5 and 0.3 points as M grows.

Table 8: Client-count sweep on CIFAR-100 and CUB-200 (Avg, %).

	CIFAR-100			CUB-200		
Front-end	M=5	M=10	M=20	M=5	M=10	M=20
independent	$78.8_{\pm 0.1}$	$79.0_{\pm 0.1}$	$79.0_{\pm 0.1}$	$84.2_{\pm 0.1}$	$84.3_{\pm 0.1}$	$84.3_{\pm 0.1}$
text-NN (CLIP)	$80.0_{\pm 0.2}$	$81.0_{\pm 0.1}$	$81.3_{\pm 0.1}$	$85.2_{\pm 0.2}$	$85.8_{\pm 0.1}$	$86.2_{\pm 0.1}$
text-strong (BGE)	$79.9_{\pm 0.2}$	$80.7_{\pm 0.1}$	$81.1_{\pm 0.1}$	$85.4_{\pm 0.1}$	$86.2_{\pm 0.1}$	$86.4_{\pm 0.1}$
visual	$80.3_{\pm 0.1}$	$81.3_{\pm 0.1}$	$81.6_{\pm 0.1}$	$85.8_{\pm 0.1}$	$86.5_{\pm 0.1}$	$86.6_{\pm 0.0}$
discovery (GCD)	$80.8_{\pm 0.2}$	$81.7_{\pm 0.1}$	$81.5_{\pm 0.1}$	$86.0_{\pm 0.2}$	$86.7_{\pm 0.2}$	$86.8_{\pm 0.1}$
FedTaxo-C	$80.4_{\pm 0.1}$	$81.4_{\pm 0.1}$	$81.6_{\pm 0.1}$	$85.5_{\pm 0.1}$	$86.2_{\pm 0.1}$	$86.6_{\pm 0.0}$
FedTaxo-B	$80.4_{\pm 0.2}$	$81.4_{\pm 0.2}$	$81.4_{\pm 0.2}$	$85.7_{\pm 0.2}$	$86.4_{\pm 0.1}$	$86.5_{\pm 0.1}$
FedTaxo-A	$79.4_{\pm 0.3}$	–	–	$85.0_{\pm 0.2}$	–	–
FedTaxo	$81.1_{\pm 0.1}$	$82.0_{\pm 0.1}$	$82.1_{\pm 0.1}$	$85.1_{\pm 0.2}$	$86.2_{\pm 0.2}$	$86.5_{\pm 0.0}$
FedTaxo-L	$80.9_{\pm 0.1}$	$82.0_{\pm 0.1}$	$81.9_{\pm 0.1}$	$86.1_{\pm 0.1}$	$86.6_{\pm 0.1}$	$86.7_{\pm 0.1}$
oracle (ref.)	$81.3_{\pm 0.1}$	$82.1_{\pm 0.1}$	$82.2_{\pm 0.1}$	$86.2_{\pm 0.2}$	$86.8_{\pm 0.1}$	$86.9_{\pm 0.1}$

Table 9 reports Avg at five and one shots on the four datasets, and the coarse clients' accuracy and true-edge recall in the hierarchy regime, for FedTaxo-L, for greedy assignment in place of re-clustering, and for the matcher without its shot-set features, its vocabulary-depth cue or its WordNet features. Each feature group is worth at most 0.3 Avg on any dataset yet moves CIFAR-100 edge recall by 3 to 28 points, so the extra features buy structure rather than accuracy. Means over 5 seeds.

Table 9: Ablations of the learned matcher (Avg and hierarchy, %).

	CIFAR-100		*mini*ImageNet		CUB-200		FGVC-Aircraft		HIER: coarse acc. / edge recall			
Variant	K=5	K=1	K=5	K=1	K=5	K=1	K=5	K=1	CIFAR-100	*mini*ImageNet	CUB-200	Aircraft
FedTaxo-L	80.9	75.7	94.0	90.7	86.1	80.6	59.3	56.1	66.8 / 53	88.4 / 3	80.5 / 47	47.4 / 2
greedy assignment	80.6	76.0	94.0	90.5	86.0	80.2	58.7	55.7	64.3 / 39	88.4 / 1	78.0 / 32	46.3 / 12
no set features	80.8	75.7	93.8	90.8	86.0	80.6	59.0	56.0	64.6 / 50	89.1 / 3	80.3 / 49	47.0 / 0
no vocabulary cue	80.8	76.0	93.9	90.5	86.0	80.6	59.1	56.1	65.3 / 30	90.3 / 8	79.6 / 50	48.3 / 8
no WordNet at all	80.8	75.9	93.9	90.6	86.0	80.6	59.1	56.1	63.9 / 25	90.1 / 8	80.5 / 49	48.7 / 11

Table 10 scores every cross-client pair of novel fine capsules against the truth: a pair is an error when two capsules of one concept end in different nodes, or two capsules of different concepts end in the same node. The middle columns count these errors for the four cells of Table 1, and the right columns count the errors of the greedy cosine that constrained re-clustering repairs, that the learned matcher repairs under greedy assignment, and that both repair; the number of true same-concept pairs is given beside the pair count. All eight dataset-by-shot settings are covered, with means over 5 seeds on identical capsules. The two components repair overlapping but distinct sets, 78% of the matcher's repairs being also the procedure's, and each introduces errors the other does not, so they are neither redundant nor strictly ordered: over the eight settings the score introduces 235 errors of

its own and the procedure 272. Five-shot CUB-200 is the only setting in which the score repairs more than the procedure. Under re-clustering, the learned score leaves fewer errors than the cosine in three settings, both CUB-200 ones and one-shot *mini*ImageNet, and the cosine leaves fewer in the other five.

Table 10: Pairwise assignment errors repaired by each component.

			pairwise errors				errors of greedy+cosine repaired by		
Dataset	shots	pairs (same)	greedy+cos	re-cl+cos	greedy+learned	re-cl+learned	procedure	score	both
CIFAR-100	5	5746 (120)	48.4	9.0	46.8	21.0	46.4	30.8	29.4
CUB-200	5	35920 (300)	124.6	102.2	42.0	8.4	93.8	100.4	73.0
miniImageNet	5	5737 (120)	27.8	0.8	6.2	1.2	27.0	25.8	25.0
Aircraft	5	5733 (120)	36.4	11.8	57.8	15.8	31.8	18.8	16.2
CIFAR-100	1	5748 (120)	104.8	108.0	92.4	109.6	72.2	57.4	43.4
CUB-200	1	35963 (300)	195.2	138.0	142.2	111.4	125.4	95.6	64.2
miniImageNet	1	5748 (120)	71.2	30.8	41.2	22.0	63.8	46.6	43.8
Aircraft	1	5738 (120)	64.2	40.2	73.0	50.8	43.0	30.6	21.0

C.1 Client vocabularies: prompt, responses and corruption examples

Figure 3 gives the prompt with which Qwen2.5-7B-Instruct writes the LLM client vocabularies of Sec. 4.4, Figure 4 its responses for example classes of each dataset, and Figure 5 examples of the label corruption used everywhere else. Every entry is copied from the released files. The LLM vocabularies create homonyms of their own: the layperson persona gives 59 distinct labels to the 100 classes of CIFAR-100, and the terse persona 81 to the 200 classes of CUB-200. Some replies are malformed or wrong ("trifle (this seems to be", or "small ornament" for the trifle of *mini*ImageNet); we keep them, since no real client's vocabulary is curated either.

LLM vocabulary prompt (user turn)

```
You are {persona}. You must write the label you would personally use for this category: '{class name}' ({dataset
context}). Reply with the label only, no explanation, no quotes.
```

Placeholders, decoding and post-processing

```
{persona}, client m adopts persona m:
- terse (client 1): a warehouse worker who writes the shortest label possible, usually just the most general noun
and dropping every modifier (for example 'jay' for 'blue jay', '737' for 'Boeing 737-800', 'fish' for 'aquarium
fish')
- expert (client 2): a domain expert who uses the precise technical, scientific, or manufacturer name
- layperson (client 3): an ordinary person describing what they see in plain everyday words (two to four words)
- abbrev (client 4): a data-entry clerk who writes compact codes or abbreviations of at most eight characters
- collector (client 5): a hobbyist who uses the common name with one descriptive adjective

{dataset context}:
- CIFAR-100: an everyday object, animal, plant, vehicle, person, or scene category from a photo dataset
- miniImageNet: an ImageNet object or animal category
- CUB-200: a North American bird species
- FGVC-Aircraft: an aircraft model variant

Decoding: chat template of Qwen2.5-7B-Instruct, greedy (temperature 0), at most 16 new tokens.
Post-processing: keep the first line; drop a leading "label:" or "answer:"; strip surrounding quotes, brackets and
punctuation; lowercase; keep at most five words; use the class name if nothing is left.
```

Figure 3: Prompt with which Qwen2.5-7B-Instruct writes each client's vocabulary.

Responses: CUB-200

class	blue jay	green jay	florida jay	american crow
terse	jay	jay	jay	crow
expert	cyanocitta cristata	green jay	florida jay	corvus brachyrhynchos
layperson	cyan feathered bird	green jay bird	blue thrush finch	black bird
abbrev	bjay	gjayna	fldjay	amcr_ownamerica
collector	skyblue feathered friend	emerald explorer	sunshine jay	urban explorer

Responses: FGVC-Aircraft

class	737-800	cessna 172	f-16a/b
terse	737	cessna	f-16
expert	boeing 737-800	cessna 172 model aircraft	f-16a/b fighting falcon
layperson	boeing passenger jet	small plane	fighter jet variant
abbrev	b738	c172	f16ab
collector	aerospace enthusiast's favorite	skyfarer cessna 172	eagle scout

Responses: CIFAR-100

class	maple tree	sweet pepper	lawn mower	pickup truck
terse	tree	pepper	lawn mower	truck
expert	acer tree	capsicum annuum	lawn_mower	pickuptruck
layperson	plant	vegetable	vehicle	pickup truck
abbrev	maptree	sweetpepper8char	lmr	pttruck
collector	autumnmaple	sweet bell pepper	gardentool	rugged transport

Responses: miniImageNet

class	school bus	hourglass	french bulldog	trifle
terse	bus	hourglass	dog	trifle
expert	school_bus	hourglass_figure	french_bulldog	trifle (this seems to be
layperson	yellow bus	sand timer	french bulldog	small ornament
abbrev	sbus8char	hrsglass	frenbulldog	trlfe
collector	yellowbus	hourglass figure	french bulldog pup	tinyornament

Figure 4: Responses of Qwen2.5-7B-Instruct after post-processing, one column per class.

Synonyms: curated forms (CIFAR-100)

aquarium fish	-> tropical fish, ornamental fish
sweet pepper	-> bell pepper, capsicum
lawn mower	-> mower
girl	-> lass

Real homonyms: classes that share a terse name

"jay"	CUB-200	blue jay, florida jay, green jay
"737"	FGVC-Aircraft	737-200, 737-300, 737-400, 737-500, 737-600, 737-700, 737-800, 737-900
"fish"	CIFAR-100	aquarium fish, flatfish, ray, shark, trout
"working dog"	miniImageNet	komondor, tibetan mastiff, french bulldog, malamute

Coarse labels: what a superclass-level client sends (HIER)

CIFAR-100	apple -> produce	flatfish -> fish	motorcycle -> vehicle
CUB-200	anna hummingbird -> hummingbird	white throated sparrow -> sparrow	
miniImageNet	house finch -> oscine	komondor -> working dog	
FGVC-Aircraft	707-320 -> boeing 707	a340-600 -> a340	

Figure 5: Examples of the label corruption of the benchmark.

D ADDITIONAL RESULTS

Table 11 repeats Table 2 with the standard deviation over the five seeds for every cell; bold marks the best non-oracle front-end in each column and the asterisk a Holm-significant win of FedTaxo over both text alignment and not federating.

Table 11: Table 2 with standard deviations over five seeds (%).

	CIFAR-100		*mini*ImageNet		CUB-200		FGVC-Aircraft	
Front-end	Avg	novel	Avg	novel	Avg	novel	Avg	novel
5-shot								
independent (no fed.)	$78.8_{\pm 0.1}$	$55.5_{\pm 0.6}$	$93.0_{\pm 0.2}$	$83.4_{\pm 0.5}$	$84.2_{\pm 0.1}$	$67.8_{\pm 0.4}$	$57.7_{\pm 0.1}$	$36.5_{\pm 0.3}$
text-NN (CLIP)	$80.0_{\pm 0.2}$	$59.7_{\pm 1.3}$	$93.3_{\pm 0.1}$	$85.1_{\pm 0.7}$	$85.2_{\pm 0.2}$	$74.1_{\pm 0.4}$	$58.8_{\pm 0.3}$	$43.7_{\pm 2.2}$
text-strong (BGE)	$79.9_{\pm 0.2}$	$59.3_{\pm 1.7}$	$93.6_{\pm 0.1}$	$86.8_{\pm 0.6}$	$85.4_{\pm 0.1}$	$75.6_{\pm 0.3}$	$58.7_{\pm 0.3}$	$43.0_{\pm 2.1}$
visual	$80.3_{\pm 0.1}$	$65.8_{\pm 1.2}$	$93.8_{\pm 0.1}$	$89.4_{\pm 0.4}$	$85.8_{\pm 0.1}$	$76.7_{\pm 0.6}$	$58.7_{\pm 0.4}$	$46.5_{\pm 2.3}$
discovery (GCD)	$80.8_{\pm 0.2}$	$68.5_{\pm 1.7}$	$94.0_{\pm 0.2}$	$\mathbf{91.1}_{\pm 0.6}$	$86.0_{\pm 0.2}$	$77.6_{\pm 0.9}$	$58.6_{\pm 0.4}$	$44.6_{\pm 3.0}$
FedTaxo-C (greedy)	$80.4_{\pm 0.1}$	$66.2_{\pm 1.1}$	$93.8_{\pm 0.1}$	$89.6_{\pm 0.5}$	$85.5_{\pm 0.1}$	$75.5_{\pm 0.5}$	$59.2_{\pm 0.3}$	$49.3_{\pm 1.5}$
FedTaxo-B (greedy)	$80.4_{\pm 0.2}$	$66.7_{\pm 0.8}$	$93.5_{\pm 0.4}$	$88.1_{\pm 2.5}$	$85.7_{\pm 0.2}$	$76.3_{\pm 0.9}$	$58.8_{\pm 0.2}$	$46.6_{\pm 0.9}$
FedTaxo-A (greedy)	$79.4_{\pm 0.3}$	$58.9_{\pm 3.8}$	$92.1_{\pm 1.1}$	$75.3_{\pm 7.9}$	$85.0_{\pm 0.2}$	$71.8_{\pm 0.9}$	$59.1_{\pm 0.2}$	$47.1_{\pm 1.5}$
FedTaxo	$\mathbf{81.1}^{*}{}_{\pm 0.1}$	$\mathbf{71.2}_{\pm 1.4}$	$93.9^{*}{}_{\pm 0.2}$	$90.6_{\pm 0.6}$	$85.1_{\pm 0.2}$	$73.1_{\pm 1.2}$	$\mathbf{59.3}^{*}{}_{\pm 0.1}$	$\mathbf{50.4}_{\pm 1.5}$
FedTaxo-L	$80.9_{\pm 0.1}$	$69.5_{\pm 1.4}$	$\mathbf{94.0}_{\pm 0.1}$	$90.8_{\pm 0.4}$	$\mathbf{86.1}_{\pm 0.1}$	$\mathbf{78.4}_{\pm 0.6}$	$59.3_{\pm 0.2}$	$50.2_{\pm 1.7}$
oracle (ref.)	$81.3_{\pm 0.1}$	$72.1_{\pm 1.7}$	$94.0_{\pm 0.1}$	$90.9_{\pm 0.4}$	$86.2_{\pm 0.2}$	$78.8_{\pm 0.8}$	$59.5_{\pm 0.3}$	$51.9_{\pm 1.9}$
1-shot								
independent (no fed.)	$72.3_{\pm 0.3}$	$11.0_{\pm 1.0}$	$85.9_{\pm 0.3}$	$39.2_{\pm 1.9}$	$74.6_{\pm 0.4}$	$21.7_{\pm 1.9}$	$54.4_{\pm 0.2}$	$14.3_{\pm 1.4}$
text-NN (CLIP)	$75.9_{\pm 0.3}$	$32.5_{\pm 1.8}$	$89.7_{\pm 0.3}$	$60.8_{\pm 2.4}$	$79.3_{\pm 0.5}$	$44.3_{\pm 2.4}$	$56.1_{\pm 0.2}$	$24.5_{\pm 1.8}$
text-strong (BGE)	$75.7_{\pm 0.4}$	$31.4_{\pm 2.1}$	$89.6_{\pm 0.3}$	$59.9_{\pm 1.9}$	$79.8_{\pm 0.5}$	$46.3_{\pm 2.4}$	$56.1_{\pm 0.2}$	$24.1_{\pm 1.8}$
visual	$73.3_{\pm 0.5}$	$17.9_{\pm 3.2}$	$87.6_{\pm 0.2}$	$49.6_{\pm 1.7}$	$76.8_{\pm 0.7}$	$31.9_{\pm 2.7}$	$55.3_{\pm 0.1}$	$20.3_{\pm 1.0}$
discovery (GCD)	$71.5_{\pm 0.5}$	$7.8_{\pm 2.1}$	$87.3_{\pm 1.6}$	$36.8_{\pm 17.0}$	$75.6_{\pm 0.5}$	$23.2_{\pm 1.5}$	$55.0_{\pm 0.3}$	$17.0_{\pm 2.1}$
FedTaxo-C (greedy)	$75.3_{\pm 0.5}$	$30.0_{\pm 2.4}$	$89.4_{\pm 0.3}$	$60.3_{\pm 1.8}$	$78.4_{\pm 0.5}$	$40.0_{\pm 2.5}$	$56.0_{\pm 0.2}$	$24.7_{\pm 1.5}$
FedTaxo-B (greedy)	$73.4_{\pm 0.5}$	$19.5_{\pm 1.0}$	$83.4_{\pm 0.3}$	$21.9_{\pm 0.5}$	$77.4_{\pm 0.6}$	$34.8_{\pm 3.1}$	$54.7_{\pm 0.3}$	$16.6_{\pm 1.4}$
FedTaxo-A (greedy)	$75.1_{\pm 0.4}$	$25.0_{\pm 1.7}$	$82.2_{\pm 0.3}$	$11.4_{\pm 1.3}$	$77.6_{\pm 0.7}$	$34.4_{\pm 3.2}$	$56.1_{\pm 0.3}$	$24.3_{\pm 2.0}$
FedTaxo	$\mathbf{76.1}^{*}{}_{\pm 0.3}$	$\mathbf{35.1}_{\pm 1.4}$	$\mathbf{91.3}^{*}{}_{\pm 0.2}$	$\mathbf{71.5}_{\pm 0.9}$	$80.4^{*}{}_{\pm 0.6}$	$49.2_{\pm 2.8}$	$\mathbf{56.4}_{\pm 0.3}$	$\mathbf{26.9}_{\pm 1.8}$
FedTaxo-L	$75.7_{\pm 0.4}$	$32.1_{\pm 2.0}$	$90.7_{\pm 0.3}$	$67.8_{\pm 1.6}$	$\mathbf{80.6}_{\pm 0.7}$	$\mathbf{49.6}_{\pm 3.0}$	$56.1_{\pm 0.4}$	$24.4_{\pm 2.2}$
oracle (ref.)	$77.0_{\pm 0.4}$	$42.0_{\pm 1.9}$	$91.6_{\pm 0.2}$	$74.1_{\pm 1.0}$	$81.5_{\pm 0.7}$	$54.5_{\pm 3.4}$	$56.5_{\pm 0.3}$	$27.8_{\pm 1.3}$

Table 13 gives the fusion ablation and the downstream federation gain of every front-end discussed in Appendix D.1, Table 14 sweeps the shot count for every front-end on the hardest regime, and Table 15 tests FedTaxo against the strong text encoder at each shot count. Figure 6 plots per-session accuracy for every front-end across all four datasets. Table 16 gives the per-session accuracy on CIFAR-100 at five shots underlying the Avg/PD summary of Table 2. Table 17 and Figure 7 report the stability–plasticity view (harmonic mean of base and novel accuracy, and the base/novel split by session).

Table 12 reports the average precision (AP) of predicting whether two labels from different clients denote the same concept, under synonym (SYN), real-homonym (HOM) and combined (SYNHOM) corruption, for the front-ends of Table 5; FedTaxo-B is the calibrated Bhattacharyya variant with fixed weights, FedTaxo-A the one with dry-run weights, and bold marks the best per row. Real homonyms, groups of up to five concepts sharing a name, defeat every text-only front-end, CLIP or BGE, on three datasets, whereas FedTaxo stays at 0.91–1.00 HOM AP; on SYNHOM it is within 0.03 of the best on CIFAR-100 and Aircraft, best on *mini*ImageNet, and 0.15 behind discovery on CUB-200. Means over 5 seeds.

Table 12: Cross-client alignment AP at five shots.

Dataset	Regime	string	text-NN	text-strong	visual	discovery	FedTaxo-C	FedTaxo-B	FedTaxo-A
CIFAR-100	SYN	0.687	0.906	0.920	0.796	0.821	0.965	**0.970**	0.858
	HOM	0.423	0.455	0.425	0.808	0.834	0.914	**0.933**	0.864
	SYNHOM	0.362	0.474	0.417	0.824	0.863	0.881	**0.886**	0.880
FGVC-Aircraft	SYN	0.914	0.938	**0.956**	0.732	0.751	0.929	0.935	0.924
	HOM	0.499	0.503	0.508	0.713	0.735	0.919	**0.927**	0.908
	SYNHOM	0.523	0.519	0.554	0.713	0.733	0.903	**0.908**	0.908
CUB-200	SYN	0.549	0.743	0.848	0.933	**0.942**	0.861	0.833	0.917
	HOM	0.451	0.528	0.563	0.940	0.952	**0.973**	0.972	0.934
	SYNHOM	0.245	0.355	0.373	0.934	**0.939**	0.787	0.758	0.923
*mini*ImageNet	SYN	0.413	0.718	0.872	0.966	0.973	**0.984**	0.944	0.958
	HOM	0.752	0.729	0.867	0.976	0.981	0.998	**0.999**	0.967
	SYNHOM	0.468	0.485	0.623	0.980	0.986	**0.988**	0.872	0.972

D.1 THE ALIGNMENT STUDY: TEXT ALIGNMENT COLLAPSES UNDER REAL HOMONYMS, FROM THE FIRST SHOT

This study scores label pairs with the hand-designed front-ends (the learned matcher is evaluated in the loop and, as a pairwise ranker, in Table 36). Under real homonyms every text-only front-end collapses: string, CLIP text, and BGE text reach 0.42–0.56 HOM AP on CIFAR-100, Aircraft, and CUB-200, because a shared name such as "fish" or "737" joins up to five distinct concepts and no encoder separates identical strings, whereas the cosine fusion FedTaxo-C stays at 0.91–1.00 by rejecting the merge through the visual channel (Table 12); the stronger encoder does not help (0.43 vs. 0.46 on CIFAR-100). Sweeping the shot count (Tables 14 and 15), FedTaxo-C is significantly ahead of the strong encoder at every K on all four datasets, by 0.10–0.17 AP already at one shot: a single fine-grained image, too noisy to separate synonyms, still beats a name that joins five concepts, so the boundary of Theorem 1 sits at one shot rather than at the three to five the bound suggests. In this one-session study merging is hard to turn into accuracy (FedTaxo-C beats local-only training only on Aircraft, Table 13); the incremental loop, where merges compound, is where the gain appears.

The left half of Table 13 gives the SYNHOM alignment AP of three fusion variants computed from the same capsules: the cosine score (FedTaxo), the calibrated Bhattacharyya distance with fixed weights (FedTaxo-B), and the same with dry-run weights (FedTaxo-A). The right half gives per-client test accuracy after merging by each front-end at its best-F_1 threshold, by no federation (local-only) and by the oracle; bold marks the best non-oracle entry. In this one-session study local-only is hard to beat: FedTaxo-C beats it only on Aircraft and ties on CUB-200, discovery leads on *mini*ImageNet, and the text front-end is below local-only on every dataset. Means over 5 seeds.

Table 13: Fusion ablation and downstream federation gain at five shots.

	Alignment AP (ablation)			Downstream per-client accuracy						
Dataset	cosine (FedTaxo-C)	Bhatt., fixed (B)	Bhatt., adaptive (A)	text-strong	discovery	FedTaxo-C	FedTaxo-B	FedTaxo-A	local-only	oracle
CIFAR-100	0.881	0.901	0.886	0.632	0.783	0.744	0.747	0.681	**0.784**	0.839
FGVC-Aircraft	0.903	0.885	0.908	0.595	0.607	**0.640**	0.633	0.628	0.622	0.678
CUB-200	0.787	0.769	0.758	0.839	0.904	0.902	**0.904**	0.898	0.904	0.914
*mini*ImageNet	0.988	0.893	0.872	0.872	**0.950**	0.938	0.914	0.898	0.947	0.961

Table 14 gives the SYNHOM cross-client alignment AP at one, two, three and five shots per class for each front-end on four datasets; bold marks the best per column within a dataset. On the FedTaxo-C row an asterisk marks a significant win over the strong text encoder and a circle a significant loss to it (one-sided paired t-tests over 5 seeds, $p < 0.05$). FedTaxo-C beats the strong text encoder at every shot count on every dataset; among the hand-designed variants the Bhattacharyya score leads at one to three shots on CIFAR-100 and Aircraft and FedTaxo-A on CUB-200 from two shots, while the cosine leads on *mini*ImageNet. Means over 5 seeds.

Table 14: Shot-count sweep of alignment AP under SYNHOM.

Dataset	Front-end	K=1	K=2	K=3	K=5
CIFAR-100	text-NN	0.53	0.48	0.49	0.47
	text-strong	0.38	0.38	0.40	0.42
	visual	0.19	0.38	0.59	0.82
	discovery	0.18	0.42	0.64	0.86
	FedTaxo-C	0.55*	0.66*	0.78*	0.88*
	FedTaxo-B	**0.59**	0.68	0.78	**0.89**
	FedTaxo-A	0.59	**0.72**	**0.80**	0.88
FGVC-Aircraft	text-NN	0.50	0.52	0.48	0.52
	text-strong	0.52	0.54	0.51	0.55
	visual	0.25	0.39	0.56	0.71
	discovery	0.24	0.40	0.58	0.73
	FedTaxo-C	0.62*	0.76*	0.84*	0.90*
	FedTaxo-B	0.68	0.78	0.84	**0.91**
	FedTaxo-A	**0.70**	**0.78**	**0.85**	0.91
CUB-200	text-NN	0.33	0.38	0.46	0.35
	text-strong	0.40	0.45	0.48	0.37
	visual	0.38	0.60	0.81	0.93
	discovery	0.38	0.61	0.83	**0.94**
	FedTaxo-C	0.54*	0.71*	0.80*	0.79*
	FedTaxo-B	**0.56**	0.70	0.77	0.76
	FedTaxo-A	0.55	**0.74**	**0.83**	0.92
*mini*ImageNet	text-NN	0.47	0.49	0.57	0.49
	text-strong	0.62	0.63	0.54	0.62
	visual	0.54	0.83	0.91	0.98
	discovery	0.56	0.85	0.93	0.99
	FedTaxo-C	**0.75***	**0.94***	**0.93***	**0.99***
	FedTaxo-B	0.69	0.80	0.85	0.87
	FedTaxo-A	0.68	0.84	0.92	0.97

Table 15 compares the SYNHOM alignment AP of the BGE text front-end and of the cosine fusion FedTaxo-C as the shots per class K grow from one to five. A check mark denotes a significant FedTaxo-C win, a circled minus a significant loss and a cross no significant difference (one-sided paired t-tests over 5 seeds, $p < 0.05$). FedTaxo-C is significantly ahead of the strong encoder at every shot count on every dataset.

Table 15: FedTaxo-C against the strong text encoder by shot count.

Dataset	K=1	K=2	K=3	K=5
CIFAR-100	0.38 → 0.55 ✓	0.38 → 0.66 ✓	0.40 → 0.78 ✓	0.42 → 0.88 ✓
FGVC-Aircraft	0.52 → 0.62 ✓	0.54 → 0.76 ✓	0.51 → 0.84 ✓	0.55 → 0.90 ✓
CUB-200	0.40 → 0.54 ✓	0.45 → 0.71 ✓	0.48 → 0.80 ✓	0.37 → 0.79 ✓
*mini*ImageNet	0.62 → 0.75 ✓	0.63 → 0.94 ✓	0.54 → 0.93 ✓	0.62 → 0.99 ✓

Figure 6 plots per-client accuracy after each incremental session for each dataset (columns) at five shots (top) and one shot (bottom), with one line per front-end and the oracle as reference. FedTaxo tracks the oracle through every session, and text-NN degrades fastest because homonym over-merges corrupt prototypes that are then carried forward. At one shot the hand-designed scores compress toward text alignment while FedTaxo stays nearest the oracle. Means over 5 seeds.

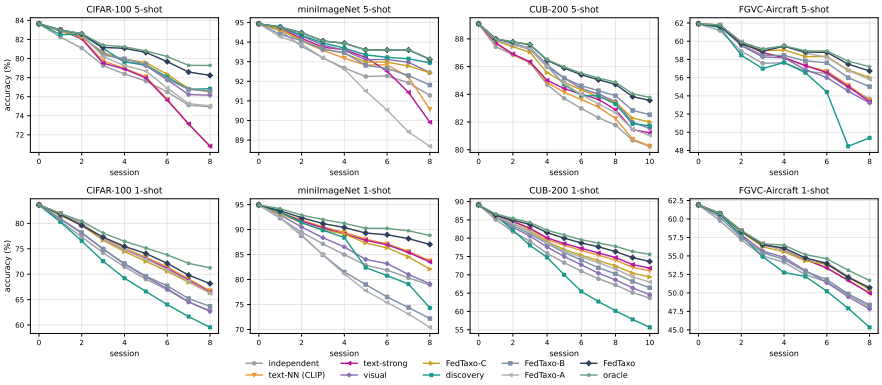

Figure 6: Per-session accuracy of every front-end.

Table 16 gives the accuracy after each of the nine sessions of the UOPP protocol for every front-end, with Avg, PD (the drop from session 0 to session 8) and Gap (oracle Avg minus Avg). Session 0 is identical across rows because the base session is shared, and the rows separate from session 1 onward. By the last session text-NN ends 0.6 points above independent, FedTaxo-C 2.6 and FedTaxo 3.9 above it, against 4.4 for the oracle: text alignment helps once it receives the constraints, but recovers far less of the oracle's gain. Mean±std over 5 seeds.

Table 16: Per-session accuracy on CIFAR-100 at five shots (%).

Front-end	Accuracy in each session (%) ↑									Avg↑	PD↓	Gap↓
	0	1	2	3	4	5	6	7	8			
independent (no fed.)	83.6	82.3	81.1	79.3	78.4	77.6	76.5	75.1	74.9	78.8±0.1	8.7±0.4	2.5
string	83.6	83.0	82.4	80.5	80.2	79.4	77.9	76.5	75.3	79.9±0.2	8.3±0.5	1.4
text-NN (CLIP)	83.6	83.0	82.5	80.7	80.4	79.7	78.2	76.5	75.5	80.0±0.2	8.2±0.6	1.3
text-strong (BGE)	83.6	83.0	82.4	80.5	80.2	79.5	77.9	76.5	75.4	79.9±0.2	8.3±0.5	1.4
visual	83.6	82.9	82.3	80.6	80.2	79.7	78.8	77.4	77.3	80.3±0.1	6.3±0.7	0.9
discovery (GCD)	83.6	82.9	82.5	81.3	81.0	80.5	79.5	78.1	77.9	80.8±0.2	5.7±0.5	0.4
FedTaxo-C	83.6	82.9	82.3	80.6	80.2	79.8	78.9	77.5	77.5	80.4±0.1	6.1±0.6	0.9
FedTaxo-B	83.6	83.0	82.5	80.6	80.2	79.8	78.8	77.6	77.5	80.4±0.2	6.1±0.5	0.9
FedTaxo-A	83.6	83.0	82.5	80.2	79.3	78.7	76.9	75.2	74.9	79.4±0.3	8.7±0.6	1.9
FedTaxo	83.6	83.0	82.6	81.4	81.2	80.8	80.0	78.9	78.8	81.1±0.1	4.8±0.6	0.1
FedTaxo-L	83.6	83.0	82.6	81.1	81.1	80.6	79.7	78.4	78.1	80.9±0.1	5.5±0.5	0.3
FedTaxo-L (greedy)	83.6	82.9	82.3	80.9	80.6	80.1	79.2	78.0	77.7	80.6±0.2	5.9±0.5	0.7
FedTaxo-L (no set features)	83.6	83.0	82.6	81.5	80.9	80.4	79.4	78.0	77.6	80.8±0.1	6.1±0.4	0.5
FedTaxo-L (no vocabulary cue)	83.6	83.0	82.5	81.5	80.9	80.4	79.5	78.2	77.6	80.8±0.2	6.0±0.3	0.5
FedTaxo-L (no WordNet)	83.6	83.0	82.6	81.6	81.0	80.4	79.4	78.0	77.3	80.8±0.2	6.3±0.5	0.5
oracle (ref.)	83.6	83.0	82.6	81.4	81.2	80.8	80.2	79.3	79.3	81.3±0.1	4.4±0.6	0.0

Table 17 gives the harmonic mean of base and novel accuracy by session at five shots, the stability–plasticity summary used by UOPP, with its Avg and its PD from session 1 to session 8 (a negative PD means the harmonic mean rose). Base accuracy is nearly identical across front-ends, so the harmonic mean is driven by novel-class accuracy. FedTaxo gives the best balance among non-oracle rows (Avg 77.5, against 74.4 for FedTaxo-C); the text-only rows start highest, since a shared string aligns most pairs early, and end at 69.0 as homonym merges corrupt their novel prototypes. Means over 5 seeds.

Table 17: Harmonic mean of base and novel accuracy on CIFAR-100 (%).

Front-end	Harmonic-mean acc. by session (%)								Avg	PD
	1	2	3	4	5	6	7	8		
independent (no fed.)	67.2	69.5	66.3	66.3	66.8	66.1	64.8	66.6	66.7	0.5
string	78.0	79.2	73.8	74.3	73.9	71.3	69.7	68.7	73.6	9.3
text-NN (CLIP)	78.0	79.5	75.1	75.2	74.8	72.3	69.9	69.0	74.2	9.0
text-strong (BGE)	78.0	79.2	73.8	74.4	74.0	71.4	69.7	68.8	73.7	9.2
visual	76.2	78.1	74.2	74.5	74.7	73.7	71.9	72.9	74.5	3.3
discovery (GCD)	76.9	79.7	77.6	77.3	77.0	75.6	73.6	74.3	76.5	2.6
FedTaxo-C	76.4	78.0	73.7	74.0	74.5	73.8	71.9	73.3	74.4	3.1
FedTaxo-B	77.7	79.6	74.7	74.4	74.8	73.7	72.5	73.6	75.1	4.1
FedTaxo-A	78.0	79.6	72.7	71.3	71.7	68.7	66.6	67.8	72.0	10.2
FedTaxo	78.0	80.4	77.9	77.8	77.7	76.7	75.5	76.2	77.5	1.9
FedTaxo-L	78.0	80.2	76.9	77.5	77.3	76.2	74.5	75.0	76.9	3.0
FedTaxo-L (greedy)	76.8	78.5	75.5	75.8	75.8	74.7	73.5	73.9	75.6	2.9
FedTaxo-L (no set features)	77.5	79.3	75.8	75.0	75.6	74.6	72.5	72.7	75.4	4.8
FedTaxo-L (no vocabulary cue)	77.6	78.8	75.5	75.2	75.7	74.8	73.0	73.0	75.4	4.6
FedTaxo-L (no WordNet)	77.6	79.4	75.9	75.3	75.8	74.7	72.6	72.3	75.4	5.2
oracle (ref.)	78.0	80.3	77.9	77.9	77.7	77.2	76.2	77.0	77.8	1.0

Figure 7 separates base-class accuracy, novel-class accuracy and their harmonic mean after each session at five shots (top) and one shot (bottom). Base accuracy is stable and nearly identical across front-ends because the base session is shared and its prototypes are never merged. The front-ends separate on novel classes, where correct alignment pools cross-client shots into sharper prototypes and a wrong merge pools shots of different concepts. Means over 5 seeds.

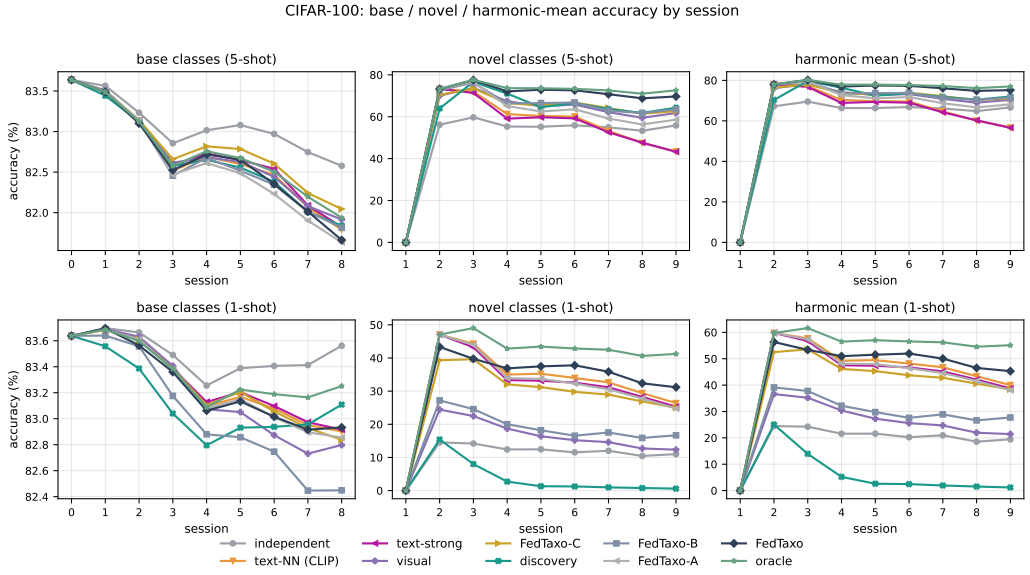

Figure 7: Base, novel and harmonic-mean accuracy by session on CIFAR-100.

Table 18 reports the five-shot Avg of FedTaxo-L when its matcher is trained on synonyms only or on homonyms only, and its coarse-client accuracy and edge recall in the hierarchy regime when it is trained without pseudo-coarse labelers, against the default simulator, the untrained FedTaxo-C and the oracle; 5 seeds.

Table 18: Sensitivity of FedTaxo-L to its training simulator.

	Avg, five shots, SYNHOM				HIER: coarse acc. / edge recall			
Matcher trained with	CIFAR-100	*mini*ImageNet	CUB-200	Aircraft	CIFAR-100	*mini*ImageNet	CUB-200	Aircraft
the full simulator (default)	$80.9_{\pm 0.1}$	$94.0_{\pm 0.1}$	$86.1_{\pm 0.1}$	$59.3_{\pm 0.2}$	66.8 / 53	88.4 / 3	80.5 / 47	47.4 / 2
synonyms only (no homonyms)	$81.1_{\pm 0.1}$	$94.0_{\pm 0.2}$	$86.1_{\pm 0.2}$	$59.4_{\pm 0.2}$	–	–	–	–
homonyms only (no synonyms)	$80.8_{\pm 0.2}$	$93.7_{\pm 0.2}$	$85.6_{\pm 0.1}$	$59.3_{\pm 0.2}$	–	–	–	–
no pseudo-coarse labelers	–	–	–	–	64.0 / 0	88.2 / 0	75.4 / 0	47.1 / 0
FedTaxo-C (no learning)	$80.4_{\pm 0.1}$	$93.8_{\pm 0.1}$	$85.5_{\pm 0.1}$	$59.2_{\pm 0.3}$	61.0 / 1	85.7 / 0	73.6 / 0	46.7 / 0
oracle	$81.3_{\pm 0.1}$	$94.0_{\pm 0.1}$	$86.2_{\pm 0.2}$	$59.5_{\pm 0.3}$	74.4 / 100	91.6 / 100	83.1 / 100	54.5 / 100

Table 19 adds to SYNHOM two of five clients that label at the superclass level, and reports Avg, the coarse clients' accuracy in their own label space, the fine clients' novel-class accuracy, the recall and precision of predicted sub-edges between a coarse node and its member fine nodes, and the percentage of true coarse-to-fine pairs merged into one node instead; 5 seeds. Only a rule that predicts the relation types an edge. Recall on *mini*ImageNet and Aircraft is too small for its precision to be meaningful, which Tables 32 and 33 trace to those benchmarks' own hierarchies and to the training simulator.

Table 19: The hierarchy regime at five shots (%).

	CIFAR-100					*mini*ImageNet					CUB-200					FGVC-Aircraft				
Front-end	Avg	coarse	fine	edge R/P	merged	Avg	coarse	fine	edge R/P	merged	Avg	coarse	fine	edge R/P	merged	Avg	coarse	fine	edge R/P	merged
independent	79.4	$58.0_{\pm 2.3}$	$55.4_{\pm 2.4}$	0/–	0	93.0	$83.3_{\pm 0.9}$	$83.1_{\pm 0.7}$	0/–	0	84.7	$69.5_{\pm 1.3}$	$68.5_{\pm 1.3}$	0/–	0	57.8	$37.8_{\pm 2.8}$	$36.5_{\pm 1.7}$	0/–	0
text-NN (CLIP)	80.1	$60.8_{\pm 1.5}$	$57.6_{\pm 2.1}$	0/–	37	93.3	$84.9_{\pm 1.1}$	$84.3_{\pm 1.3}$	0/–	8	85.1	$73.4_{\pm 1.8}$	$71.5_{\pm 1.6}$	0/–	15	58.9	$46.9_{\pm 3.0}$	$44.4_{\pm 1.7}$	0/–	41
visual	80.4	$63.9_{\pm 1.6}$	$61.3_{\pm 2.3}$	0/–	52	93.6	$87.1_{\pm 0.8}$	$87.5_{\pm 0.9}$	0/–	72	85.5	$74.2_{\pm 0.3}$	$73.5_{\pm 1.0}$	0/–	56	58.9	$49.6_{\pm 3.6}$	$46.9_{\pm 2.9}$	0/–	71
discovery (GCD)	80.7	$53.9_{\pm 5.7}$	$67.6_{\pm 1.4}$	0/–	84	93.9	$87.9_{\pm 0.8}$	$90.6_{\pm 0.4}$	0/–	99	85.6	$67.0_{\pm 2.0}$	$77.3_{\pm 0.6}$	0/–	89	58.6	$46.7_{\pm 4.1}$	$43.8_{\pm 2.7}$	0/–	94
FedTaxo-C	80.3	$61.0_{\pm 2.9}$	$62.8_{\pm 2.0}$	1/67	34	93.4	$85.7_{\pm 0.5}$	$86.9_{\pm 0.7}$	0/0	16	85.3	$73.6_{\pm 1.9}$	$72.6_{\pm 0.8}$	0/30	14	58.9	$46.7_{\pm 3.1}$	$46.3_{\pm 2.3}$	0/28	41
FedTaxo-B	80.1	$58.4_{\pm 2.4}$	$60.8_{\pm 3.2}$	20/78	54	93.1	$81.6_{\pm 4.0}$	$83.3_{\pm 3.1}$	32/72	57	85.2	$71.0_{\pm 1.7}$	$72.6_{\pm 1.1}$	36/76	33	58.3	$46.3_{\pm 2.4}$	$41.6_{\pm 2.8}$	10/28	63
FedTaxo	80.3	$62.1_{\pm 3.9}$	$62.7_{\pm 2.6}$	8/25	64	93.8	$87.2_{\pm 2.4}$	$89.4_{\pm 1.1}$	3/0	89	85.4	$73.5_{\pm 1.9}$	$73.0_{\pm 1.0}$	0/50	14	59.0	$49.8_{\pm 1.4}$	$48.3_{\pm 3.0}$	0/2	84
FedTaxo-L	80.8	$66.8_{\pm 2.9}$	$66.3_{\pm 1.2}$	53/83	27	93.8	$88.4_{\pm 1.2}$	$89.4_{\pm 0.7}$	3/50	77	86.0	$80.5_{\pm 0.6}$	$76.1_{\pm 0.8}$	47/84	42	59.0	$47.4_{\pm 3.1}$	$47.3_{\pm 2.1}$	2/56	43
oracle (ref.)	81.0	$74.4_{\pm 2.2}$	$66.8_{\pm 1.0}$	100/100	0	93.8	$91.6_{\pm 0.5}$	$88.5_{\pm 1.1}$	100/100	0	86.0	$83.1_{\pm 1.3}$	$75.4_{\pm 1.2}$	100/100	0	59.2	$54.5_{\pm 2.6}$	$46.3_{\pm 1.4}$	100/100	0

Table 20 reports Avg when each client labels with the vocabulary an LLM wrote for its persona, a corruption the matcher's simulator never saw; bold marks the best non-oracle front-end, with mean$\pm$std over 5 seeds.

Table 20: LLM-written client vocabularies (Avg, %).

	CIFAR-100		*mini*ImageNet		CUB-200		FGVC-Aircraft	
Front-end	K=5	K=1	K=5	K=1	K=5	K=1	K=5	K=1
independent (no fed.)	$78.9_{\pm 0.1}$	$72.3_{\pm 0.4}$	$93.1_{\pm 0.1}$	$86.1_{\pm 0.4}$	$84.2_{\pm 0.1}$	$74.6_{\pm 0.3}$	$57.7_{\pm 0.2}$	$54.4_{\pm 0.2}$
text-NN (CLIP)	$79.4_{\pm 0.2}$	$73.5_{\pm 0.4}$	$93.1_{\pm 0.1}$	$87.8_{\pm 0.7}$	$84.2_{\pm 0.1}$	$74.7_{\pm 0.3}$	$57.7_{\pm 0.2}$	$54.4_{\pm 0.2}$
visual	$80.5_{\pm 0.3}$	$72.9_{\pm 0.6}$	$93.7_{\pm 0.1}$	$87.8_{\pm 0.6}$	$85.9_{\pm 0.2}$	$76.9_{\pm 0.5}$	$\mathbf{58.8}_{\pm 0.4}$	$55.3_{\pm 0.1}$
discovery (GCD)	$\mathbf{80.9}_{\pm 0.2}$	$71.5_{\pm 0.3}$	$93.8_{\pm 0.2}$	$87.0_{\pm 1.6}$	$\mathbf{86.0}_{\pm 0.2}$	$76.1_{\pm 0.6}$	$58.5_{\pm 0.6}$	$54.9_{\pm 0.3}$
FedTaxo-C	$79.1_{\pm 0.2}$	$73.0_{\pm 0.4}$	$93.4_{\pm 0.2}$	$88.5_{\pm 0.6}$	$84.2_{\pm 0.1}$	$74.8_{\pm 0.3}$	$57.7_{\pm 0.1}$	$54.6_{\pm 0.3}$
FedTaxo-B	$80.2_{\pm 0.5}$	$73.0_{\pm 0.5}$	$92.9_{\pm 0.5}$	$83.5_{\pm 0.4}$	$85.5_{\pm 0.2}$	$77.2_{\pm 0.5}$	$58.4_{\pm 0.2}$	$54.7_{\pm 0.2}$
FedTaxo	$80.0_{\pm 0.2}$	$\mathbf{75.1}_{\pm 0.5}$	$\mathbf{93.9}_{\pm 0.2}$	$\mathbf{91.1}_{\pm 0.4}$	$83.7_{\pm 0.3}$	$\mathbf{78.2}_{\pm 0.3}$	$57.6_{\pm 0.5}$	$\mathbf{55.4}_{\pm 0.3}$
FedTaxo-L	$79.7_{\pm 0.1}$	$73.5_{\pm 0.5}$	$93.8_{\pm 0.2}$	$89.2_{\pm 0.3}$	$85.5_{\pm 0.2}$	$76.1_{\pm 0.3}$	$57.7_{\pm 0.2}$	$54.4_{\pm 0.2}$
oracle (ref.)	$81.3_{\pm 0.2}$	$76.9_{\pm 0.4}$	$93.9_{\pm 0.2}$	$91.7_{\pm 0.4}$	$86.2_{\pm 0.1}$	$81.4_{\pm 0.3}$	$59.6_{\pm 0.2}$	$56.6_{\pm 0.2}$

The left half of Table 21 reports Avg, PD and novel-class accuracy on CIFAR-100 for our from-scratch UOPP core with only the front-end varied; bold marks the best non-oracle per column and an asterisk a significant win of FedTaxo over every other front-end ($p < 0.05$), over 5 seeds. The right half summarizes robustness across the ten loop settings: wins, mean rank, and mean and worst-case regret, the Avg gap to the best front-end.

Table 21: A trained prompt-based method (left) and robustness over ten settings (right).

Front-end	5-shot Avg↑	5-shot PD↓	5-shot novel↑	1-shot Avg↑	1-shot novel↑
independent (no fed.)	85.5±0.1	5.0±0.5	72.8±0.8	79.6±0.1	34.6±0.5
string	85.5±0.2	6.5±0.6	68.7±1.0	82.5±0.5	48.3±1.9
text-NN (CLIP)	85.5±0.2	6.7±0.6	68.3±1.0	82.6±0.4	49.5±1.8
text-strong (BGE)	85.5±0.2	6.6±0.6	68.7±1.0	82.5±0.5	48.3±1.9
visual	86.5±0.1	3.4±0.5	79.9±0.8	81.4±0.3	45.3±1.0
discovery (GCD)	86.7±0.2	3.0±0.7	81.3±0.5	82.1±0.3	48.8±1.3
FedTaxo-C (greedy)	86.6±0.2	3.4±0.5	80.3±0.8	82.0±0.4	49.9±2.1
FedTaxo-B (greedy)	86.2±0.2	4.5±1.0	76.5±2.5	79.5±0.7	26.9±5.9
FedTaxo-A (greedy)	85.4±0.4	5.9±0.9	71.2±2.4	76.8±1.1	7.6±6.3
FedTaxo	**86.8**±0.2*	**2.8**±0.4	**82.1**±1.0	**83.7**±0.4*	**58.8**±1.7
FedTaxo-L	86.7±0.2	3.0±0.4	81.6±0.9	83.2±0.3	56.2±2.1
oracle (ref.)	86.8±0.2	2.8±0.4	82.2±1.1	84.2±0.3	63.0±1.1

Front-end	wins	mean rank↓	mean regret↓	worst regret↓
FedTaxo	7	2.1	0.12	1.01
FedTaxo-L	3	2.5	0.20	0.59
FedTaxo-C (greedy)	0	5.2	0.88	2.22
text-NN (CLIP)	0	5.3	0.87	1.56
text-strong (BGE)	0	5.8	0.86	1.73
string	0	6.0	1.03	2.45
discovery (GCD)	0	6.3	1.77	4.99
visual	0	6.8	1.57	3.75
FedTaxo-B (greedy)	0	7.3	2.24	7.92
FedTaxo-A (greedy)	0	8.7	2.66	9.13
independent (no fed.)	0	10.0	2.96	6.02

Table 22 runs the UOPP repository's Fed-L2P on CIFAR-100 under its own protocol (20 clients, 5 shots) with only the front-end's label maps injected, and reports Avg, PD and last-session accuracy A_8 in each client's own label space over 5 seeds. Its FedAvg-trained linear head forgets base classes faster the more clients share a row, so not federating novel classes (69.5) beats every merging front-end, and the ordering among them tracks how much each merges rather than how correctly: FedTaxo recovers almost exactly the oracle's concept count (40–42 ids for 40 classes) and scores almost exactly the oracle's Avg (55.3 against 55.4), while the least-merging front-end, greedy cosine at 59–67 ids, scores highest (58.0). We therefore read this experiment as evidence that a published classifier can foreclose the gain alignment provides, and draw no comparison among front-ends from it.

Table 22: Fed-L2P from the UOPP repository under taxonomy mismatch.

Fed-L2P	indep.	text-NN	visual	discovery	FedTaxo-C	FedTaxo-B	**FedTaxo**	**FedTaxo-L**	oracle
Avg↑	$69.5_{\pm 0.9}$	$52.2_{\pm 1.5}$	$57.5_{\pm 1.8}$	$56.1_{\pm 1.2}$	$\mathbf{58.0}_{\pm 1.7}$	$55.5_{\pm 0.8}$	$55.3_{\pm 1.9}$	$55.9_{\pm 1.3}$	$55.4_{\pm 2.0}$
PD↓	$25.2_{\pm 0.8}$	$42.8_{\pm 1.0}$	$37.9_{\pm 1.3}$	$38.6_{\pm 1.5}$	$\mathbf{36.8}_{\pm 1.4}$	$39.5_{\pm 0.8}$	$39.1_{\pm 1.4}$	$38.5_{\pm 0.9}$	$38.8_{\pm 1.9}$
A_8 ↑	$62.1_{\pm 0.7}$	$44.5_{\pm 0.8}$	$49.4_{\pm 1.2}$	$48.7_{\pm 1.3}$	$\mathbf{50.5}_{\pm 1.3}$	$47.9_{\pm 0.7}$	$48.2_{\pm 1.4}$	$48.9_{\pm 0.9}$	$48.6_{\pm 1.8}$

In Table 23 each client partitions the novel classes into its own groups of one to three classes, so client A may hold $\{A, B\}$ where client B holds $\{B, C\}$ and no single global partition is consistent with both; it reports Avg and final novel-class accuracy at five shots over 5 seeds, and Appendix D.2 discusses the result. Text alignment is the lower of the two consistently negative rules on every dataset except *mini*ImageNet.

Table 23: Cross-cutting client taxonomies at five shots (%).

	CIFAR-100		*mini*ImageNet		CUB-200		FGVC-Aircraft	
Front-end	Avg	novel	Avg	novel	Avg	novel	Avg	novel
independent (no fed.)	79.2	47.8	92.8	76.4	83.4	52.0	57.8	26.8
text-NN (CLIP)	77.7	37.4	92.2	69.8	80.4	36.7	56.9	19.8
visual	79.6	49.0	92.5	71.1	83.9	53.0	58.1	29.6
discovery (GCD)	79.9	48.9	92.4	68.8	84.4	60.9	57.9	31.4
FedTaxo-C	79.4	49.4	92.7	75.5	83.6	52.4	58.0	28.5
FedTaxo	79.2	47.4	92.7	75.0	83.6	52.9	58.0	28.7
FedTaxo-L	78.8	45.6	92.8	75.1	84.1	54.2	57.8	26.9
ground-truth overlap merging	78.6	39.2	91.0	56.0	81.8	38.0	57.6	23.1

D.2 Cross-cutting taxonomies in detail

Outside the formulation. Every result above assumes client taxonomies refine a common structure. Where each client instead partitions the novel classes its own way, so that one holds $\{A, B\}$ where another holds $\{B, C\}$ and no single partition satisfies both, *no method helps consistently and which one helps depends on the dataset*: the best front-end gains 1.6, 9.0 and 4.6 novel points over not federating on CIFAR-100, CUB-200 and Aircraft while every front-end loses on *mini*ImageNet, and

discovery alone swings from $+9.0$ to -7.6. Consistently negative are text alignment and *ground-truth overlap merging*, which joins two groups whenever they truly share a class: both fall below not federating on all four datasets, overlap merging because overlap is not transitive while merging is. Neither is uniformly the worse of the two, and text alignment is in fact lower on three, at 37.4 against 39.2 novel points on CIFAR-100, 36.7 against 38.0 on CUB-200 and 19.8 against 23.1 on Aircraft. That indicts the operator, not the use of true correspondence: preserving the clients' own groupings needs overlapping sets over a fine partition, which we regard as the next formulation (Appendix D, Table 23).

Table 24 repeats Table 19 with a coarse label receiving K shots in total, split over its members, rather than K shots of each member, so that coarse and fine labels carry the same budget. Edge recall falls from 53 to 44 on CIFAR-100 and 47 to 42 on CUB-200 and precision from 83 to 74 and 84 to 80, so the structure is recovered less completely, while coarse-client accuracy is unchanged.

Table 24: The hierarchy regime at equal shot budget (%).

	CIFAR-100					*mini*ImageNet					CUB-200					FGVC-Aircraft				
Front-end	Avg	coarse	fine	edge R/P	merged	Avg	coarse	fine	edge R/P	merged	Avg	coarse	fine	edge R/P	merged	Avg	coarse	fine	edge R/P	merged
independent	79.2	$56.6_{\pm 4.0}$	$54.9_{\pm 1.8}$	0/–	0	93.0	$83.3_{\pm 0.9}$	$83.1_{\pm 0.7}$	0/–	0	84.7	$71.0_{\pm 1.3}$	$68.1_{\pm 1.1}$	0/–	0	57.8	$37.2_{\pm 3.2}$	$36.7_{\pm 1.6}$	0/–	0
text-NN (CLIP)	80.0	$61.1_{\pm 2.5}$	$55.8_{\pm 1.7}$	0/–	37	93.3	$84.9_{\pm 1.1}$	$84.3_{\pm 1.3}$	0/–	8	85.2	$73.4_{\pm 1.2}$	$71.1_{\pm 0.7}$	0/–	14	58.9	$46.4_{\pm 3.6}$	$44.1_{\pm 2.4}$	0/–	41
visual	80.4	$63.5_{\pm 2.9}$	$62.6_{\pm 0.9}$	0/–	52	93.6	$87.6_{\pm 0.8}$	$87.7_{\pm 0.9}$	0/–	75	85.6	$76.6_{\pm 1.1}$	$74.0_{\pm 0.9}$	0/–	58	58.9	$49.0_{\pm 4.1}$	$46.7_{\pm 2.6}$	0/–	70
discovery (GCD)	80.6	$56.8_{\pm 2.5}$	$67.6_{\pm 1.2}$	0/–	87	93.9	$87.9_{\pm 0.8}$	$90.6_{\pm 0.4}$	0/–	99	85.7	$69.3_{\pm 2.5}$	$77.3_{\pm 1.3}$	0/–	86	58.7	$47.4_{\pm 3.3}$	$44.9_{\pm 4.2}$	0/–	91
FedTaxo-C	80.3	$60.1_{\pm 3.3}$	$62.6_{\pm 0.4}$	1/58	32	93.4	$85.7_{\pm 0.5}$	$86.9_{\pm 0.7}$	0/0	16	85.4	$73.9_{\pm 1.0}$	$72.7_{\pm 0.7}$	0/7	14	58.9	$46.0_{\pm 3.5}$	$46.0_{\pm 2.4}$	0/21	40
FedTaxo-B	80.1	$60.8_{\pm 1.7}$	$61.3_{\pm 1.0}$	23/71	53	93.1	$81.0_{\pm 4.9}$	$82.8_{\pm 3.1}$	31/69	59	85.3	$73.4_{\pm 1.1}$	$71.8_{\pm 1.1}$	34/72	33	58.3	$44.5_{\pm 3.6}$	$41.8_{\pm 2.4}$	9/23	62
FedTaxo	80.2	$62.3_{\pm 2.0}$	$64.2_{\pm 1.5}$	5/26	64	93.8	$87.2_{\pm 2.4}$	$89.4_{\pm 1.1}$	3/0	89	85.4	$73.8_{\pm 1.0}$	$72.9_{\pm 0.6}$	0/–	14	59.0	$48.9_{\pm 2.8}$	$47.7_{\pm 2.6}$	0/2	85
FedTaxo-L	80.8	$67.1_{\pm 2.7}$	$66.8_{\pm 0.9}$	44/74	32	93.8	$89.6_{\pm 0.4}$	$89.1_{\pm 0.5}$	12/96	75	86.1	$81.7_{\pm 1.8}$	$76.3_{\pm 1.0}$	42/80	47	59.0	$46.9_{\pm 3.7}$	$46.8_{\pm 2.6}$	2/56	43
oracle (ref.)	80.9	$73.9_{\pm 2.1}$	$67.3_{\pm 0.8}$	100/100	0	93.8	$91.6_{\pm 0.5}$	$88.5_{\pm 1.1}$	100/100	0	86.0	$84.1_{\pm 1.3}$	$75.4_{\pm 0.5}$	100/100	0	59.2	$54.1_{\pm 3.6}$	$46.8_{\pm 1.8}$	100/100	0

Table 25 separates forgetting from the UOPP summary PD, base-session accuracy minus final accuracy, which mixes forgetting with the growing difficulty of a larger label space; *base forget.* is the drop in accuracy on the base classes themselves. Base accuracy moves by at most a few points for every front-end, so PD is dominated by novel-class acquisition rather than by forgetting.

Table 25: PD and base-class forgetting (%).

	CIFAR-100		*mini*ImageNet		CUB-200		FGVC-Aircraft	
Front-end	PD	base forget.	PD	base forget.	PD	base forget.	PD	base forget.
5-shot								
independent (no fed.)	8.7	1.1	3.7	0.5	8.8	1.4	8.2	1.4
text-NN (CLIP)	8.2	1.9	3.4	0.9	7.1	2.3	6.9	2.5
text-strong (BGE)	8.3	1.9	3.0	1.0	6.5	2.3	7.1	2.4
visual	6.3	1.6	2.2	0.9	6.1	2.4	6.2	2.6
discovery (GCD)	5.7	1.7	1.9	1.0	5.8	2.3	6.9	2.8
FedTaxo-C (greedy)	6.1	1.6	2.1	0.9	6.4	2.1	5.4	2.6
FedTaxo-B (greedy)	6.1	1.8	2.7	1.0	6.3	2.4	6.3	2.6
FedTaxo-A (greedy)	8.7	2.0	6.5	1.2	8.0	2.5	6.1	2.6
FedTaxo	4.8	1.8	1.9	1.0	7.6	2.6	5.2	2.7
FedTaxo-L	5.5	2.0	1.8	1.0	5.6	2.5	5.2	2.6
oracle (ref.)	4.4	1.7	1.8	1.0	5.3	2.3	4.7	2.6
1-shot								
independent (no fed.)	20.8	0.1	16.1	0.2	25.4	0.2	13.8	0.3
text-NN (CLIP)	15.2	0.6	9.8	0.4	17.3	0.8	11.2	0.7
text-strong (BGE)	15.5	0.5	10.0	0.4	16.5	0.8	11.3	0.7
visual	19.2	0.5	13.1	0.3	22.1	1.1	12.6	1.0
discovery (GCD)	22.4	1.1	17.1	0.6	25.4	1.1	13.5	1.0
FedTaxo-C (greedy)	15.7	0.5	10.0	0.4	18.9	0.7	11.2	0.8
FedTaxo-B (greedy)	18.9	1.0	21.5	0.8	21.0	1.1	13.6	0.9
FedTaxo-A (greedy)	17.3	0.8	24.5	0.9	21.1	1.0	11.3	0.7
FedTaxo	14.5	0.7	6.9	0.4	15.6	1.0	10.6	0.8
FedTaxo-L	15.3	0.7	7.9	0.4	15.4	1.0	11.2	0.7
oracle (ref.)	12.4	0.4	6.1	0.4	13.5	0.8	10.2	0.7

In Table 26 the simulator sees the first half of every class's synonym forms and deployment uses the second half; classes with a single form keep it, and 63, 80 and 115 of the classes have two or more forms on the three datasets. The default FedTaxo, whose score is not learned, is affected only through its dry-run threshold, while FedTaxo-L retrains its matcher on the first half alone. Mean±std over 5 seeds.

Table 26: Held-out paraphrases (Avg, %).

Front-end	CIFAR-100	*mini*ImageNet	CUB-200
independent	$78.9_{\pm 0.1}$	$93.0_{\pm 0.1}$	$84.3_{\pm 0.2}$
text-NN (CLIP)	$78.5_{\pm 0.1}$	$93.3_{\pm 0.2}$	$84.4_{\pm 0.2}$
text-strong (BGE)	$78.5_{\pm 0.1}$	$93.3_{\pm 0.2}$	$84.8_{\pm 0.2}$
FedTaxo-C	$80.1_{\pm 0.4}$	$93.7_{\pm 0.1}$	$85.6_{\pm 0.2}$
FedTaxo	$81.1_{\pm 0.2}$	$93.9_{\pm 0.2}$	$85.7_{\pm 0.2}$
FedTaxo-L	$81.1_{\pm 0.2}$	$93.8_{\pm 0.2}$	$86.0_{\pm 0.1}$
oracle (ref.)	$81.3_{\pm 0.2}$	$93.9_{\pm 0.2}$	$86.2_{\pm 0.1}$

Table 27 lets every client observe every new class at one shot, so each concept can pool five single shots and the oracle's gain over not federating is largest; FedTaxo and FedTaxo-L are listed separately, since they are distinct configurations, over 5 seeds.

Table 27: One shot with every client observing every new class (Avg, %).

Front-end	CIFAR-100	*mini*ImageNet	CUB-200	FGVC-Aircraft
independent	$67.1_{\pm 0.2}$	$81.6_{\pm 0.3}$	$68.8_{\pm 0.3}$	$50.7_{\pm 0.1}$
text-NN (CLIP)	$71.9_{\pm 0.2}$	$89.0_{\pm 0.2}$	$76.7_{\pm 0.2}$	$53.6_{\pm 0.1}$
discovery (GCD)	$64.5_{\pm 0.1}$	$82.3_{\pm 2.2}$	$65.3_{\pm 2.4}$	$50.7_{\pm 1.3}$
FedTaxo-C	$71.3_{\pm 0.2}$	$88.1_{\pm 0.8}$	$75.3_{\pm 0.3}$	$53.8_{\pm 0.3}$
FedTaxo	$75.4_{\pm 0.5}$	$91.9_{\pm 0.3}$	$80.5_{\pm 0.5}$	$54.9_{\pm 0.2}$
FedTaxo-L	$74.1_{\pm 0.5}$	$91.2_{\pm 0.4}$	$80.2_{\pm 0.3}$	$54.4_{\pm 0.2}$
oracle (ref.)	$76.5_{\pm 0.4}$	$91.9_{\pm 0.3}$	$81.4_{\pm 0.3}$	$55.5_{\pm 0.2}$

Table 28 reports Avg at five shots when a client holds a second, duplicate label for a new class with probability p, which violates the cannot-link assumption, and when every shot embedding is perturbed by Gaussian noise of norm σ times the embedding norm before it leaves the client. FedTaxo and FedTaxo-L are reported in separate columns, and the *none* row reproduces Table 2; 5 seeds.

Table 28: Duplicate labels and embedding noise (Avg, %).

	CIFAR-100				CUB-200			
Perturbation	FedTaxo-C	**FedTaxo**	FedTaxo-L	oracle	FedTaxo-C	**FedTaxo**	FedTaxo-L	oracle
none	80.4	81.1	80.9	81.3	85.5	85.1	86.1	86.2
duplicate labels, p=0.2	79.1	79.0	79.8	81.5	83.7	82.9	84.0	86.4
duplicate labels, p=0.4	78.2	77.8	78.3	81.8	81.8	81.2	82.1	86.4
embedding noise σ=0.25	79.5	80.9	80.0	81.1	84.9	85.1	86.0	86.1
embedding noise σ=0.5	77.7	79.9	78.2	80.0	83.3	84.3	84.9	85.0
embedding noise σ=1	73.4	74.4	72.7	74.4	75.9	78.6	78.8	79.3

Table 29 gives the capsules arriving per session and the single-threaded wall-clock seconds per session for FedTaxo, for FedTaxo-L, whose pair evaluations over the eight sessions are also listed, and for the greedy FedTaxo-C. Re-clustering with the cosine score costs a fraction of the learned matcher, which must evaluate a gradient-boosted model on every pair.

Table 29: Cost of re-clustering on CIFAR-100 at five shots.

Clients M	capsules per session	**FedTaxo** (cosine) s/session	FedTaxo-L (learned) s/session	FedTaxo-L pair evaluations	FedTaxo-C (greedy) s/session
5	15	0.4	0.9	50832	0.00
10	30	0.6	3.3	188978	0.01
20	60	1.0	12.4	739402	0.02

Table 30 reports Avg and final novel accuracy of FedTaxo at five shots over 5 seeds when a client holds a duplicate label for a new class with probability *dup. rate*, so that the within-client cannot-link is wrong for that pair. The *hard* rule forbids merging two labels of one client and is used everywhere else in the paper; the *soft* rule permits it above a raised threshold $\tau+\sigma$, with σ the standard deviation of the session's pair scores; *none* drops the constraint for every front-end. Bold marks the best Avg of the three rules. The soft rule is the best in five of six cells and always better than no constraint, so a partly wrong constraint is still worth keeping; it recovers part of the damage, not all of it: 36 and 38% of the clean-to-corrupted gap on CIFAR-100 at p=0.2 and 0.4, and 59 and 53% on CUB-200.

Table 30: Hard, soft and no within-client constraint (%).

		hard (main results)		soft ($\tau+\sigma$)		none	
Dataset	dup. rate	Avg	novel	Avg	novel	Avg	novel
CIFAR-100	0.0	**81.14**	71.2	80.45	66.3	80.18	63.9
	0.2	79.00	59.0	**79.77**	62.1	79.53	60.0
	0.4	77.81	52.4	**79.07**	58.6	78.74	56.6
reference		independent 77.38			oracle 81.77		
CUB-200	0.0	85.09	73.1	**85.62**	75.7	85.41	74.5
	0.2	82.94	63.9	**84.21**	70.0	84.14	69.3
	0.4	81.15	56.0	**83.23**	65.3	83.14	64.6
reference		independent 81.57			oracle 86.41		

Table 31 re-extracts every image feature with CLIP ViT-B/32 (512-d) instead of ViT-L/14 (768-d) and reports Avg and final novel accuracy over 5 seeds; the label-text embeddings are unchanged, since the pipeline never combines text with image features. The last two rows give the Avg that FedTaxo loses to the smaller encoder and the fraction of the oracle's Avg gain over not federating that it keeps: absolute accuracy falls by 6 to 20 points while that fraction moves by at most 6. Five-shot CUB-200 is the exception under both encoders, and the weaker one sharpens it.

Table 31: A weaker visual encoder, CLIP ViT-B/32 (Avg/novel, %).

Front-end	CIFAR-100 5-shot	CUB-200 5-shot	miniImageNet 5-shot	Aircraft 5-shot	CIFAR-100 1-shot	CUB-200 1-shot
independent (no fed.)	66.9/44.5	69.3/46.2	86.6/71.6	38.4/22.6	61.9/9.6	61.7/8.3
string	67.4/46.8	70.8/55.6	87.4/76.4	39.1/27.0	64.4/25.3	64.9/24.3
text-NN (CLIP)	67.5/46.9	70.9/56.0	87.4/76.8	39.0/26.6	64.6/26.8	65.1/25.0
visual	67.8/52.0	71.2/56.9	87.5/78.7	39.0/27.5	62.5/15.0	63.0/16.4
discovery (GCD)	68.3/54.3	71.7/59.7	88.0/82.0	39.0/26.1	62.2/13.0	62.1/12.6
FedTaxo-C	68.2/54.9	71.2/57.0	87.7/79.8	39.3/30.5	64.5/27.1	64.6/23.3
FedTaxo	68.6/57.3	70.6/55.1	87.9/81.7	39.5/31.8	64.5/26.7	65.7/27.7
FedTaxo-L	68.3/55.6	71.9/61.5	88.0/81.7	39.5/31.6	64.1/25.1	65.8/28.6
oracle (ref.)	68.8/59.4	72.0/61.9	88.0/81.9	39.6/32.4	65.4/33.7	66.5/31.8
FedTaxo: ViT-L/14 − ViT-B/32	12.5	14.5	6.0	19.8	11.6	14.7
FedTaxo share of the oracle gain	91%	47%	95%	92%	75%	83%

Table 32 describes the coarse level the coarse-labeling clients use: the number of fine classes and coarse groups, the mean number of fine classes per group, the fraction of groups holding a single class, the fraction of fine classes whose name equals their parent's, and the fraction of parent names WordNet knows. A single-class group makes sub and same-as designate the same images. The last two rows are the non-degenerate levels used in Table 33.

Table 32: Structure of each benchmark's coarse level.

Coarse level	classes/groups	members per group	singleton groups	child name = parent	parent in WordNet
CIFAR-100	100/20	5.0	0%	0%	100%
CUB-200	200/70	2.9	47%	5%	98%
miniImageNet	100/77	1.3	79%	0%	100%
Aircraft	100/70	1.4	77%	51%	12%
miniImageNet, WordNet 4 levels up	100/46	2.2	61%	0%	100%
Aircraft, manufacturer	100/30	3.3	43%	0%	39%

Table 33 gives the sub-edge recall and precision of FedTaxo-L at five shots over 5 seeds. *default* is the frozen recipe of Table 19; + *single-member units* lets the training simulator draw coarse capsules that pool a single fine class, the case the deployed coarse clients mostly present on the two nearly degenerate hierarchies; *coarser level* relabels the coarse clients at the non-degenerate level of Table 32 (*mini*ImageNet and Aircraft only). The two interventions are not additive and the level

carries the result: on *mini*ImageNet recall is 94 at the coarser level alone and 91 with both, against 38 from the simulator fix alone and 3 by default, while Aircraft resists every configuration. A dash marks a configuration that does not exist, since CIFAR-100 and CUB-200 have no degenerate level to replace, or a precision left undefined because no sub-edge was produced (Aircraft at the coarser level). Avg changes by at most 0.5 across all configurations, so what moves is the structure, not the accuracy.

Table 33: Sub-edge recall and precision under the two diagnosed causes (%).

	sub-edge recall				sub-edge precision			
Dataset	default	+ single-member units	coarser level	coarser level + units	default	+ single-member units	coarser level	coarser level + units
CIFAR-100	53	71	–	–	83	76	–	–
CUB-200	47	34	–	–	84	73	–	–
miniImageNet	3	38	94	91	50	99	99	99
Aircraft	2	26	0	0	56	60	–	–

Table 34 retains the shared canonical base initialization: every client re-declares half of the base classes under its own private label, and base nodes are candidates for those capsules. *base-align* is the fraction of re-declared base capsules that reach the base node of their own class. The graph, the canonical base-to-label mapping used at evaluation and the flag that marks a capsule as a returning base class all remain in place, so this table measures matching privately named reappearances into an already aligned graph; Table 35 removes the shared base initialization itself. Five shots, 5 seeds.

Table 34: Base classes that reappear under private names (%).

	CIFAR-100		CUB-200		miniImageNet		Aircraft	
Front-end	base-align↑	Avg↑	base-align↑	Avg↑	base-align↑	Avg↑	base-align↑	Avg↑
independent (no fed.)	0	78.8	0	84.3	0	93.0	0	57.6
string	62	79.7	60	85.2	43	93.5	97	58.8
text-NN (CLIP)	82	79.9	65	85.2	55	93.6	95	58.5
visual	99	80.3	100	85.9	100	93.7	97	58.8
FedTaxo-C	62	80.3	60	85.6	57	93.7	97	59.2
FedTaxo	99	81.0	90	85.2	96	93.9	99	59.4
FedTaxo-L	96	80.9	100	86.1	99	94.0	97	59.4
oracle (ref.)	100	81.2	100	86.3	100	94.0	100	59.5

Table 35 removes the shared canonical base initialization: every client holds every base class under its own private label, the front-end aligns the base session with the rule it applies to novel classes, and base classes are scored in each client's own label space, so no front-end receives a base correspondence it did not build. *merged* is the fraction of cross-client pairs of one base class that end in the same node; the last two rows give what FedTaxo loses relative to the aligned protocol of Table 2 and the size of the base graph it builds. The regime removes the shared canonical labeling of the base session, not labeled data at the server: the merge threshold is still selected by the dry run on canonically labeled base classes, as everywhere else in the paper. The greedy rules fail in both directions at once: on CIFAR-100 the cosine rule merges 62% of the true pairs yet still leaves 87 nodes, a fifth of which mix classes, and the visual rule leaves 65 nodes that are only 51% pure. Five shots, 5 seeds.

Table 35: An unaligned base session (%).

	CIFAR-100		CUB-200		miniImageNet		Aircraft	
Front-end	merged↑	Avg↑	merged↑	Avg↑	merged↑	Avg↑	merged↑	Avg↑
independent (no fed.)	0	78.0	0	84.2	0	93.0	0	56.3
string	45	64.1	46	71.5	45	87.4	78	52.7
text-NN (CLIP)	56	64.0	49	70.6	55	87.2	72	54.2
visual	60	66.1	72	75.8	87	87.7	34	41.6
discovery (GCD)	100	80.7	100	86.0	100	94.0	100	58.2
FedTaxo-C	62	74.0	59	79.0	77	89.4	75	54.6
FedTaxo	100	81.2	97	84.8	100	94.0	100	58.9
FedTaxo-L	95	80.6	97	85.9	100	94.0	98	58.9
oracle (ref.)	100	81.3	100	86.2	100	94.0	100	59.0
FedTaxo: unaligned − aligned base	+0.04		-0.25		+0.03		-0.36	
base nodes, FedTaxo / true number of classes	60/60		103/100		60/60		60/60	

Table 36 scores the pairs of the alignment protocol of Appendix D.1 (40 base classes, five clients, 30 novel concepts each) by FedTaxo-L's $P(\text{same})$ and, on the same pairs, by the cosine fusion of FedTaxo-C; means over 5 seeds. As a ranker of single pairs the learned matcher is better under synonyms and mixed under homonyms, where the 40 base classes offer few collision groups to learn from, which is one reason it does not improve on the cosine in the loop.

Table 36: Pairwise alignment AP of the learned matcher.

Dataset	Score	SYN	HOM	SYNHOM	SYNHOM K=1/2/3
CIFAR-100	FedTaxo-C	0.958	0.918	0.820	0.587 / 0.670 / 0.749
	FedTaxo-L	0.964	0.791	0.721	0.472 / 0.555 / 0.657
FGVC-Aircraft	FedTaxo-C	0.919	0.929	0.875	0.653 / 0.729 / 0.832
	FedTaxo-L	0.996	0.926	0.899	0.541 / 0.762 / 0.836
CUB-200	FedTaxo-C	0.832	0.987	0.770	0.552 / 0.730 / 0.711
	FedTaxo-L	0.831	0.670	0.788	0.294 / 0.465 / 0.551
*mini*ImageNet	FedTaxo-C	0.983	0.982	0.911	0.792 / 0.864 / 0.882
	FedTaxo-L	0.984	0.964	0.957	0.743 / 0.850 / 0.869

Table 37 gives accuracy on the novel classes at the last session, at five and one shots under SYNHOM over 5 seeds, the quantity federation should raise.

Table 37: Final novel-class accuracy (%).

	CIFAR-100		*mini*ImageNet		CUB-200		FGVC-Aircraft	
Front-end	K=5	K=1	K=5	K=1	K=5	K=1	K=5	K=1
independent (no fed.)	$55.5_{\pm 0.6}$	$11.0_{\pm 1.0}$	$83.4_{\pm 0.5}$	$39.2_{\pm 1.9}$	$67.8_{\pm 0.4}$	$21.7_{\pm 1.9}$	$36.5_{\pm 0.3}$	$14.3_{\pm 1.4}$
text-NN (CLIP)	$59.7_{\pm 1.3}$	$32.5_{\pm 1.8}$	$85.1_{\pm 0.7}$	$60.8_{\pm 2.4}$	$74.1_{\pm 0.4}$	$44.3_{\pm 2.4}$	$43.7_{\pm 2.2}$	$24.5_{\pm 1.8}$
discovery (GCD)	$68.5_{\pm 1.7}$	$7.8_{\pm 2.1}$	$91.1_{\pm 0.6}$	$36.8_{\pm 17.0}$	$77.6_{\pm 0.9}$	$23.2_{\pm 1.5}$	$44.6_{\pm 3.0}$	$17.0_{\pm 2.1}$
FedTaxo-C (greedy)	$66.2_{\pm 1.1}$	$30.0_{\pm 2.4}$	$89.6_{\pm 0.5}$	$60.3_{\pm 1.8}$	$75.5_{\pm 0.5}$	$40.0_{\pm 2.5}$	$49.3_{\pm 1.5}$	$24.7_{\pm 1.5}$
FedTaxo	$71.2_{\pm 1.4}$	$35.1_{\pm 1.4}$	$90.6_{\pm 0.6}$	$71.5_{\pm 0.9}$	$73.1_{\pm 1.2}$	$49.2_{\pm 2.8}$	$50.4_{\pm 1.5}$	$26.9_{\pm 1.8}$
FedTaxo-L	$69.5_{\pm 1.4}$	$32.1_{\pm 2.0}$	$90.8_{\pm 0.4}$	$67.8_{\pm 1.6}$	$78.4_{\pm 0.6}$	$49.6_{\pm 3.0}$	$50.2_{\pm 1.7}$	$24.4_{\pm 2.2}$
oracle (ref.)	$72.1_{\pm 1.7}$	$42.0_{\pm 1.9}$	$90.9_{\pm 0.4}$	$74.1_{\pm 1.0}$	$78.8_{\pm 0.8}$	$54.5_{\pm 3.4}$	$51.9_{\pm 1.9}$	$27.8_{\pm 1.3}$

E THE HAND-DESIGNED SCORES

FedTaxo-C, FedTaxo-B, and FedTaxo-A score a capsule z against a node v by

$$s(z, v) = w_v \, \tilde{\rho}(z, v) \, \bar{D}_{\text{vis}}(z, v) + w_s \, \bar{D}_{\text{sem}}(z, v), \tag{17}$$

with $\bar{D}_{\text{sem}}$ the label-embedding cosine and $\bar{D}_{\text{vis}}$ the prototype cosine (FedTaxo-C, w_v=0.6, $\tilde{\rho}$=1) or the negative Bhattacharyya distance of the calibrated Gaussians weighted by a reliability $\tilde{\rho} = \max(\rho_z \rho_v, \rho_{\min})$, $\rho = \text{tr}\Sigma_{\text{base}}/(\text{tr}\Sigma_{\text{base}} + \text{tr}\Sigma_k/K)$, so that a low-shot capsule defers to text (FedTaxo-B, with split/merge repair); FedTaxo-A selects w_v and $\rho_{\min}$ by the dry run and assigns a session's capsules jointly by partial optimal transport (Appendix F). They commit each capsule greedily to its best node if the score clears τ and type the match by a spread ratio γ=2 or the lexical cue.

F ORDER-INVARIANT ASSIGNMENT VIA OPTIMAL TRANSPORT

The greedy rule merges one capsule at a time, so its outcome can depend on the arrival order within a session. FedTaxo-A therefore assigns a session's capsules jointly: capsules that arrive in the same session are first grouped with each other by complete linkage at the same threshold, then the groups are matched against the graph snapshot from the start of the session by solving an entropic partial-optimal-transport plan (Cuturi, 2013), with a reject column (costed at the decision threshold) standing in for "spawn a new node." Every group's arg max assignment is committed at once, so two capsules of the same session no longer see a graph that changed because of each other's earlier arrival. Table 38 isolates this step on FedTaxo-A without repair; Table 16 gives its greedy, joint, and repaired variants.

Table 38 compares FedTaxo-A without repair when a session's capsules are assigned one at a time (greedy) or jointly through an entropic partial optimal-transport plan (OT), at five and one shots, with p from a one-sided paired t-test for OT $>$ greedy over 5 seeds; bold marks the higher Avg. Under real homonyms joint assignment never helps: it is 0.5–0.8 points below greedy at five shots on CIFAR-100, *mini*ImageNet and CUB-200, ties on Aircraft, and is within 1.3 points at one shot.

Table 38: Joint optimal-transport against greedy assignment (Avg, %).

	5-shot		1-shot	
Dataset	greedy→OT (Avg)	p	greedy→OT (Avg)	p
CIFAR-100	**79.7**→79.2	0.99	**75.2**→75.1	0.92
*mini*ImageNet	**93.1**→92.3	0.98	**84.4**→83.1	0.95
CUB-200	**85.3**→84.7	1.00	**77.5**→77.2	0.98
FGVC-Aircraft	58.2→58.2	0.54	55.9→55.9	–

Under the frozen protocol the order invariance that joint assignment buys does not pay: the complete-linkage groups it matches are costed by their pooled statistics, which average away the per-capsule evidence the greedy rule acts on, and every significant difference in Table 38 favors greedy. Joint assignment is therefore kept only in FedTaxo-A; FedTaxo-C uses the greedy rule and FedTaxo re-clusters.

G PROOFS

Throughout, $\phi(x) = \mu_c + \xi$ denotes the embedding of a sample of concept c with independent coordinates and σ-sub-Gaussian noise ξ, and $p_c = \frac{1}{K}\sum_{i=1}^{K} \phi(x_i)$ is the K-shot prototype, so $\mathbb{E}[p_c] = \mu_c$ and $p_c - \mu_c$ is $(\sigma/\sqrt{K})$-sub-Gaussian per coordinate.

G.1 PROOF OF THEOREM 1

Theorem (Theorem 1, formal)**.** *Let p_a, p_b be independent K-shot prototypes with $p_c \sim \mathcal{N}(\mu_c, \sigma^2 I_d/K)$, write $s^2 = 2\sigma^2/K$ and $t = \log(1/\delta)$, fix a target separation $\Delta_0 > 0$, and consider the test*

$$\widehat{r}(p_a, p_b) = \textit{same} \iff \|p_a - p_b\|_2^2 < \tau^2, \qquad \tau^2 = s^2 d + \tfrac{1}{2}\Delta_0^2,$$

whose threshold depends only on d, s and Δ_0. Suppose $8s^2(\sqrt{dt} + t) \le \Delta_0^2 \le s^2 d$. Then $\Pr[\widehat{r} = \textit{different}] \le \delta$ under $H_0 : \mu_a = \mu_b$, and $\Pr[\widehat{r} = \textit{same}] \le \delta$ under any alternative $H_1 : \|\mu_a - \mu_b\|_2 \ge \Delta_0$. Conversely, if $\Delta_0 = o(\sigma/\sqrt{K})$ then every test errs with probability at least $\frac{1}{2}(1 - o(1))$, which exceeds $1/4$.

Upper bound. Write $X = p_a - p_b$, so $X \sim \mathcal{N}(\mu_a - \mu_b, s^2 I_d)$ and $\|X\|_2^2 = s^2 \chi_d^2(\lambda)$ with noncentrality $\lambda = \|\mu_a - \mu_b\|_2^2/s^2$. We use the chi-square tails of Laurent & Massart (2000): $\Pr[\chi_d^2 \ge d + 2\sqrt{dt} + 2t] \le e^{-t}$ and, in the noncentral case, $\Pr[\chi_d^2(\lambda) \le d + \lambda - 2\sqrt{(d + 2\lambda)t}] \le e^{-t}$.

Under H_0 ($\lambda = 0$). An error needs $s^2\chi_d^2 \ge \tau^2$, that is $\chi_d^2 - d \ge \Delta_0^2/(2s^2)$. Since $\Delta_0^2 \ge 8s^2(\sqrt{dt} + t) \ge 2s^2(2\sqrt{dt} + 2t)$, this deviation is at least the Laurent-Massart bound, so the error probability is at most δ.

Under H_1. Write $\lambda_0 = \Delta_0^2/s^2$, so $\lambda \ge \lambda_0$. An error needs $s^2\chi_d^2(\lambda) \le \tau^2 = s^2(d + \lambda_0/2)$. A noncentral chi-square is stochastically increasing in its noncentrality, so $\Pr[\chi_d^2(\lambda) \le d + \lambda_0/2] \le \Pr[\chi_d^2(\lambda_0) \le d + \lambda_0/2]$ and it suffices to treat $\lambda = \lambda_0$, where the error needs $\chi_d^2(\lambda_0) - (d + \lambda_0) \le -\lambda_0/2$. By the noncentral tail it is enough that $\lambda_0/2 \ge 2\sqrt{(d + 2\lambda_0)t}$. Under $\lambda_0 \le d$ we have $d + 2\lambda_0 \le 3d$, so $\lambda_0^2 \ge 48dt$ suffices, and $\Delta_0^2 \ge 8s^2\sqrt{dt}$ gives $\lambda_0 \ge 8\sqrt{dt}$, hence $\lambda_0^2 \ge 64dt \ge 48dt$. The error probability is again at most δ. Since $8s^2(\sqrt{dt} + t) \le 16s^2\sqrt{dt}$ for $d \ge t$, the hypothesis is implied by $\Delta_0 \ge 4\sigma(2dt)^{1/4}/\sqrt{K}$, the rate quoted in (5). □

Why the centering cannot be removed. For the uncentered rule $\|p_a - p_b\| < \Delta_{ab}/2$ of the earlier statement, a same-concept pair satisfies $\|X\|^2 = s^2\chi_d^2$ with mean $s^2 d$, so the rule is correct only if

$\chi^2_d < \Delta^2_{ab}/(4s^2)$, which for $\Delta_{ab} = \Theta(\sigma\sqrt{t/K})$ and $d \gg t$ has probability $\Pr[\chi^2_d < t/2] = e^{-\Theta(d)}$. Controlling the scalar projection $\langle u, X\rangle$ with $u = (\mu_a - \mu_b)/\Delta_{ab}$, as a dimension-free argument does, does not control $\|X\|$, because the $d-1$ noise directions orthogonal to u contribute $s^2(d-1)$ to the squared norm and nothing to the projection. A dimension-free guarantee is therefore available for the projection statistic, which requires knowing $\mu_a - \mu_b$, but not for the norm rule the front-ends actually use.

Lower bound. If $\Delta_0 = o(\sigma/\sqrt{K})$, the prototype laws $\mathcal{N}(\mu_a, \sigma^2 I_d/K)$ and $\mathcal{N}(\mu_b, \sigma^2 I_d/K)$ have total variation $\mathrm{TV} \le \Delta_0\sqrt{K}/(2\sigma) \to 0$ by Pinsker's inequality, so the sum of the two error probabilities of any test is at least $1 - \mathrm{TV} = 1 - o(1)$ and the worse of the two is at least $\frac{1}{2}(1 - o(1))$, which exceeds $1/4$ for Δ_0 small enough. This half is dimension-free and therefore loose. For this problem the known minimax sample complexity is $\Theta(\sqrt{d}/\varepsilon^2)$ (Acharya et al., 2020, p. 2); with one observation of per-coordinate variance $s^2 = 2\sigma^2/K$ it corresponds to a separation of order $s\, d^{1/4} \asymp \sigma d^{1/4}/\sqrt{K}$, which the upper bound above matches at constant δ. No gap in the rate remains; only our own converse falls short of it. ■

Theorem 2 (Homonym rejection requires the visual channel)**.** *Let $a \neq b$ be a homonym pair, so $e_a = e_b$ and $D_{\mathrm{sem}}(a,b) = 0$. Let a* monotone text-only matcher *be any rule of the form* same $\iff$ *$h(D_{\mathrm{sem}}) \ge \theta$ with h non-increasing. Any such matcher that merges some true synonym pair with semantic gap $\gamma > 0$ also merges (a,b): no threshold separates them. The reliability-fused score of Eq. 17 ranks (a,b) below that synonym pair, so that* some *threshold accepts the synonym and rejects (a,b), iff*

$$w_v \tilde{\rho}_{ab}\, D^{\mathrm{Bhat}}_{\mathrm{vis}}(a,b) \;>\; w_v \tilde{\rho}_{\mathrm{syn}}\, D^{\mathrm{Bhat}}_{\mathrm{vis}}(\mathrm{syn}) + w_s\, \gamma,$$

where $D^{\mathrm{Bhat}}_{\mathrm{vis}}(\mathrm{syn})$ is the synonym pair's own visual distance, which is small but not zero. Separability is all this gives: whether the threshold selected by the dry run falls in that interval is an empirical question, which Sec. 4.1 answers by measurement rather than by this ranking. Here $\tilde{\rho} = \max(\rho_z\rho_v, \rho_{\min})$ with $\rho = \mathrm{tr}\Sigma_{\mathrm{base}}/(\mathrm{tr}\Sigma_{\mathrm{base}} + \mathrm{tr}\Sigma_k/K)$ is a product of two per-capsule terms, so it admits no bound below beyond $\rho_{\min}$; we report its measured range instead.

G.2 Proof of Theorem 2

By assumption $e_a = e_b$, so $D_{\mathrm{sem}}(a,b) = 1 - \langle e_a, e_b\rangle = 0$ while a true synonym pair whose strings differ has $D_{\mathrm{sem}} = \gamma > 0$. For a monotone rule, $h(0) \ge h(\gamma) \ge \theta$, so admitting the synonym pair admits (a,b): a false merge. Monotonicity is needed: a text-only rule that is non-monotone in D_{sem} (for instance one that rejects exact string matches outright) is not covered, though it would then fail on genuine exact matches. For the fused score, write $d_{\mathrm{syn}} = D^{\mathrm{Bhat}}_{\mathrm{vis}}(\mathrm{syn})$ and $d_{ab} = D^{\mathrm{Bhat}}_{\mathrm{vis}}(a,b)$. The synonym pair scores $s_{\mathrm{syn}} = -w_v\tilde{\rho}_{\mathrm{syn}} d_{\mathrm{syn}} - w_s\gamma$ and the homonym pair, whose semantic term vanishes, $s_{ab} = -w_v\tilde{\rho}_{ab} d_{ab}$; hence $s_{ab} < s_{\mathrm{syn}}$ iff $w_v\tilde{\rho}_{ab} d_{ab} > w_v\tilde{\rho}_{\mathrm{syn}} d_{\mathrm{syn}} + w_s\gamma$. Setting $d_{\mathrm{syn}} = 0$, as an earlier version did, drops a non-negative term from the right-hand side, which leaves a necessary but generally insufficient condition that was presented as an equivalence. Two further limits are worth stating plainly. The inequality orders the two pairs; it does not say that the selected threshold θ lies between them, so it establishes that the fused score *can* reject the homonym while keeping the synonym, not that it does. And the ordering is pairwise: with several synonym pairs the relevant comparison is against the lowest-scoring one that must still be merged. We make no claim about the range of $\tilde{\rho}$. It is a product of two per-capsule reliabilities, each of which distribution calibration keeps below 1 but neither of which is bounded below by a constant, so the earlier assertion that $\tilde{\rho} \ge 1/2$ was wrong on two counts: the product of two terms near $1/2$ is near $1/4$, and covariance blending does not pin either term. Measured over novel capsules on CIFAR-100 and CUB-200, $\rho_z\rho_v$ averages 0.81 and 0.85 at five shots and 0.59 and 0.66 at one, with minima of 0.75 and 0.52. The observed one-shot failure of FedTaxo-B is consistent with either a shrinking $\tilde{\rho}$ or prototype noise (5); we do not claim to distinguish them. ■

Theorem 3 (Impurity accounting for the split pass)**.** *Let a node hold capsules of c distinct concepts and let a split send them to halves A and B holding c_A and c_B distinct concepts, with j concepts having capsules on both sides, so $c_A + c_B = c + j$. Writing $I = \sum_n(|\{\text{concepts in } n\}| - 1)$ for total impurity,*

$$I_{\textit{after}} - I_{\textit{before}} = j - 1.$$

Hence a split strictly reduces total impurity iff its cut separates concepts cleanly ($j = 0$), leaves it unchanged iff exactly one concept straddles the cut, and increases it otherwise. The number of

impure nodes is not monotone either: a node holding four concepts split into two holding two each raises that count from one to two.

G.3 Proof of Theorem 3

Every capsule goes to exactly one side, so a concept present in the node appears in A, in B, or in both, giving $c_A + c_B = c + j$ with j the number appearing in both. The node contributed $c - 1$ to I before the split and $(c_A - 1) + (c_B - 1) = c + j - 2$ after, so the difference is $j - 1$; the three cases and the counterexample follow by substitution. ■

What we do not prove. It is tempting to add that a pure node is left intact and that a node pooling two well-separated concepts is cut cleanly, each with high probability under (5). We state neither. The two halves produced by 2-means are not independent samples: their membership is chosen using the same prototypes that are then compared, so a concentration bound for a *fixed* pair of sample means does not apply to the selected pair, and a bound for one pair does not control a node with an arbitrary number of capsules. Whether repair helps is therefore an empirical question in this paper, answered by the ablations of Table 9 rather than by a guarantee.

G.4 Constrained re-clustering under a separating matcher

Proposition 1 (Exact recovery and order invariance)**.** *Let a session hold units $u_1, \dots, u_n$ (capsules, or the must-linked capsules of one client label), each of one latent concept, and let the symmetric matching score s be τ-separating on them: $s(u_i, u_j) \geq \tau$ whenever u_i, u_j share a concept and $< \tau$ otherwise. This covers either score in Eqs.* (3) *and* (11)*. Then average-linkage agglomeration on s that stops when the best merge falls below τ returns exactly the concept partition; the result does not depend on the order in which units arrived or on any earlier graph; and cannot-links between units of one client (which are of different concepts by assumption) leave the result unchanged.*

Proof. Induct on merges, with the invariant that every current cluster is pure. Two pure clusters of one concept have average linkage $\geq \tau$ (an average of within-concept pair scores, each $\geq \tau$); two pure clusters of different concepts have average linkage $< \tau$ (an average of cross-concept scores, each $< \tau$). Hence the arg-max merge is a within-concept merge whenever one exists, purity is preserved, and the procedure stops exactly when no within-concept pair of clusters remains, i.e. at the concept partition. The merge sequence is a function of the score matrix alone, and FedTaxo dissolves every non-base node before re-clustering, which gives order invariance and independence of earlier graphs. A cannot-link removes a candidate merge between units of one client; by assumption these are cross-concept merges, which are never taken, so the output is unchanged. □

Theorem 3 gives no such guarantee for the split pass: a split lowers total impurity only when its cut separates concepts cleanly, and raises it when two or more concepts straddle the cut. Re-clustering, by contrast, reaches zero impurity in one pass whenever the score is τ-separating on the session's units, and when it errs the error is confined to that session and revisited at the next.