WHEN EQUIPARTITION CORRUPTS RELATIONAL LEARNING: CAPACITY–GEOMETRY MISMATCH UNDER LONG TAILS

Anonymous authors
Paper under double-blind review

ABSTRACT

Balanced optimal-transport assignments are widely used in representation learning to prevent collapse by enforcing equal mass across latent components. Under long-tailed data, however, this constraint can conflict with the geometry supported by the evolving representation. We identify this capacity–geometry mismatch as a source of systematic distortion in relational learning: mismatched component capacities displace samples across representation-supported basins, perturbing transport assignments and, in turn, contaminating the relational teacher and distillation signal. We theoretically characterize this mechanism and establish a finite-batch connection between full-bank transport relations and the minibatch teacher used during optimization. Guided by this analysis, we introduce PACI, which estimates representation-supported pseudo-basin capacities and applies them consistently to both transport assignment and relation construction. PACI consistently outperforms a broad range of state-of-the-art deep clustering methods across long-tailed benchmarks. Controlled capacity perturbations trace the predicted progression from capacity mismatch to assignment displacement and teacher contamination, while intervention experiments isolate assignment displacement as the dominant pathway to degraded representation quality, distinct from relation-normalization effects. These findings generalize across training regimes and long-tailed settings, supporting geometry-matched capacity as a principled design rule for long-tailed relational representation learning. Code is available at the anonymous repository: `https://anonymous.4open.science/r/PACI-44C6/`.

1 INTRODUCTION

Deep clustering learns representations that organize unlabeled examples into semantic groups (Xie et al., 2016; Caron et al., 2018). Equipartitioned optimal transport is a widely used tool for learning such representations: assigning equal mass to latent components prevents assignments from concentrating in only a few components (Asano et al., 2020; Caron et al., 2020). When relations are constructed from transport assignments, component capacity also determines which examples supervise one another. It therefore shapes both the partition of the data and the teacher used to learn the representation.

Under long-tailed data, equal capacities can conflict with the geometry already supported by the representation. Consider two well-separated basins containing 80% and 20% of the sample mass. Requiring both components to receive 50% forces at least 30% of the mass to be assigned outside its native basin, even with the features and prototypes held fixed. We call this conflict *capacity–geometry mismatch*. When relations connect samples through shared components, the resulting change in component composition propagates into the relational teacher (Figure 1). A constraint introduced to organize assignments can thus distort the supervision that subsequently shapes the features.

Imbalance-aware OT can infer nonuniform class marginals or relax balance constraints (Zheng et al., 2023; Zhang et al., 2024a). We ask how capacity should be tied to the geometry that generates transport relations, and how assignment and relation normalization affect learning. Mass conservation

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.

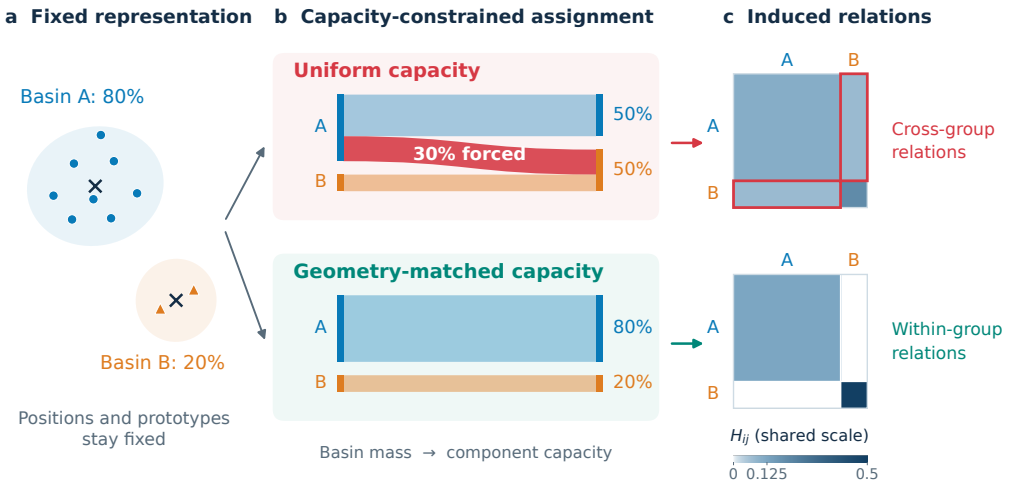

Figure 1: **Fixed geometry, different capacities, different relations.** Uniform capacity forces 30% of sample mass across basins (red ribbon), producing the highlighted cross-group relations. Geometry-matched capacity removes this compulsory reassignment. Ribbon widths show assignment mass; both relation matrices share one color scale.

gives a lower bound on assignment displacement. A forward–reverse interpretation of sample relations expresses teacher contamination through the semantic composition of transported components. A finite-batch result connects these full-bank relations to the normalized teacher used during training. The analysis identifies capacity matching as a way to remove forced reassignment at its source.

Guided by this analysis, we introduce PACI, a geometry-matched relational learning method. PACI estimates pseudo-basin capacities from an exponential-moving-average target representation and uses the same capacities in both transport assignment and the reverse transition from components to samples. Soft assignments produce a relational teacher that is distilled across augmented views; periodic refreshes update capacities and relations as the representation evolves. This construction turns the geometry-supported partition into supervision for further representation learning.

Across the five benchmark settings, PACI improves class-averaged accuracy over BYOL by 4.31–12.51 points, including CIFAR-100 and Tiny-ImageNet at imbalance factor 100. Controlled capacity perturbations trace the predicted progression from mismatch to assignment displacement and teacher contamination. Matched training interventions locate the principal capacity benefit in the assignment stage, while capacity–basin shuffles show why the association between each basin and its capacity matters.

Our contributions are threefold:

- We identify capacity–geometry mismatch as a source of distortion in relational learning and characterize its effects on assignments, component composition, and finite-batch supervision.
- We introduce PACI, which couples evolving pseudo-basin capacity estimation with a consistent transport-based relation teacher and cross-view distillation.
- We demonstrate PACI's effectiveness on long-tailed clustering benchmarks and use controlled interventions to identify assignment capacity as the dominant source of the capacity-matching benefit.

2 Related Work

Deep clustering under long tails. Deep clustering derives supervision from the evolving representation: DEC refines frequency-corrected soft targets, while DeepCluster alternates clustering and representation updates (Xie et al., 2016; Caron et al., 2018). Neighborhood consistency and joint instance–cluster discrimination provide complementary signals (Van Gansbeke et al., 2020; Li

et al., 2021). SDCLR uses a pruned competitor to emphasize easily forgotten examples (Jiang et al., 2021); LFSS uses across-epoch representation stability for instance- and cluster-level supervision (Li et al., 2025). ConMix mixes representations before contrastive learning to rebalance their classwise contributions under long tails (Li & Jia, 2025).

Transport capacities under imbalance. SeLa and SwAV use equipartitioned transport assignments to prevent collapse (Asano et al., 2020; Caron et al., 2020). Equipartition is one choice of an exact marginal: unbalanced OT relaxes marginal constraints through divergence penalties (Chizat et al., 2018), while partial OT transports a prescribed portion of the available mass (Caffarelli & McCann, 2010). P^2OT combines a KL penalty toward a uniform reference with progressively increasing selected mass; SP^2OT adds an initial-feature neighborhood graph to regularize pseudo-label inference (Zhang et al., 2024a;b). Both infer nonuniform component masses. DB-OT instead bounds target masses within a prescribed range (Shi et al., 2024). PROTOCOL uses progressive partial-transport labels from multi-view predictions for contrastive logit adjustment and class-level rebalancing (Xue et al., 2025).

Estimating imbalance from predictions and geometry. Imbalance estimates can represent class priors or the occupancy of learned clusters. SoLar updates a semantic class prior from predictions restricted to annotated candidate labels, using it for transport assignment and class-wise sample selection in partial-label learning (Wang et al., 2022). SAOT jointly infers class marginals and transport pseudo-labels from prediction costs; hard labels then guide class-wise contrastive learning (Zheng et al., 2023). CAOT additionally regularizes assignment with learned semantic similarities and uses the inferred hard labels to train clustering predictions and an attention network (Yao et al., 2026). MiniClustering recomputes mini-cluster occupancy counts from current predictions to estimate classwise self-training weights, and aligns similarities between target-cluster and mini-cluster predictions (Li et al., 2026).

Constructing relational teachers. BYOL and DINO train across views using momentum targets for representations and output distributions, respectively (Grill et al., 2020; Caron et al., 2021). Relational distillation transfers inter-example structure (Park et al., 2019); ReSSL learns similarity distributions from a momentum teacher (Zheng et al., 2021). BootSC uses separate OT problems to infer sample-affinity and cluster-assignment targets, then predicts both across views (Guo et al., 2026). AnchorGAE composes sample–anchor affinities through inverse anchor degrees for factorized graph convolution; these degrees are induced by the affinities (Zhang et al., 2023). This construction distinguishes a relation composed through components from an affinity target inferred directly between samples.

PACI pairs each basin proportion in the current representation with the prototype defining its transport cost. These proportions constrain the coupling used to construct the sample–sample teacher, so capacity estimation determines both component occupancy and relational supervision. Our theory characterizes how capacity mismatch changes assignments and teachers, including finite-batch supervision. Interventions test these responses and distinguish assignment and normalization effects on teacher contamination and training quality.

3 Capacity–Geometry Mismatch in Relational Learning

We follow the effect of capacity through three objects: sample–component assignments, the sample relations induced by those assignments, and the minibatch teacher used for distillation. The representation is held fixed throughout the analysis, so changes can be attributed to the capacity constraint.

3.1 Geometry, capacity, and relational supervision

For unlabeled samples $\{x_i\}_{i=1}^N$ and K nonempty representation-supported basins, we use entropy-regularized transport (Cuturi, 2013). Prototypes define a geometric partition with basin identities $b_i \in \{1, \dots, K\}$, empirical masses $\widehat{q}_k = N^{-1} \sum_i \mathbf{1}[b_i = k]$, and costs C_{ik}. The basin of a sample describes its location in the current representation; a transport component is a destination to which its mass can be assigned. With $a = N^{-1}\mathbf{1}_N$, capacities $q_k > 0$ summing to one specify the transport

marginal ($\varepsilon > 0$):

$$\Pi(q) = \arg\min_{\Pi \in \mathcal{U}(a,q)} \langle \Pi, C \rangle + \varepsilon \sum_{ik} \Pi_{ik}(\log \Pi_{ik} - 1), \qquad \mathcal{U}(a,q) = \{\Pi \geq 0 : \Pi \mathbf{1} = a, \Pi^\top \mathbf{1} = q\}. \tag{1}$$

The capacity–geometry mismatch is $\mathrm{TV}(q, \widehat{q}) = \frac{1}{2}\|q - \widehat{q}\|_1$. Equipartition sets $q = u = \mathbf{1}_K/K$. Geometry matching instead uses the masses of the basins that produced the transport cost.

Transport defines a two-step relation between samples. The forward probability $p(k \mid i) = N\Pi_{ik}$ assigns sample i to a component, and the reverse probability $p(j \mid k) = \Pi_{jk}/q_k$ returns to a sample. Their composition has the two-step bipartite-graph form (Zhang et al., 2023):

$$H_{ij} = \sum_k p(k \mid i) p(j \mid k), \qquad H = N \Pi D_q^{-1} \Pi^\top, \quad D_q = \mathrm{diag}(q). \tag{2}$$

Thus H is symmetric and row-stochastic. Changing capacity affects both the component assignments and the relations carried through them.

3.2 Capacity mismatch forces assignment displacement

For couplings Π, Π_0, write $D_\Pi(\Pi, \Pi_0) = \frac{1}{2}\|\Pi - \Pi_0\|_1$ and define cross-basin mass as $E_{\mathrm{basin}}(\Pi) = \sum_i \sum_{k \neq b_i} \Pi_{ik}$.

Proposition 1 (Capacity-induced displacement)**.** *For any $\Pi_0 \in \mathcal{U}(a, \widehat{q})$ and $\Pi \in \mathcal{U}(a, q)$,*

$$D_\Pi(\Pi, \Pi_0) \geq \mathrm{TV}(q, \widehat{q}), \qquad E_{\mathrm{basin}}(\Pi) \geq \mathrm{TV}(q, \widehat{q}). \tag{3}$$

The cross-basin bound is tight over the transport polytope.

Each basin can retain at most $\min(q_k, \widehat{q}_k)$ of its native mass; the remaining surplus must enter other components. This constraint holds for every feasible coupling, independently of the transport solver. Matching q to $\widehat{q}$ removes the compulsory capacity-induced floor.

3.3 Transported composition determines teacher contamination

Displacement changes which samples share components. To determine the resulting supervision, we now measure the semantic composition of each destination. Let y_i be a semantic label and define $A_{ak} = \sum_{i: y_i = a} \Pi_{ik}$ and $p(a \mid k) = A_{ak}/q_k$. Basin identities describe geometry, while y_i describes semantic identity; labels enter only this analysis and offline evaluation.

Proposition 2 (Semantic composition of the relation)**.** *The wrong-semantic mass of the raw relation in Equation* (2) *is*

$$\mathcal{C}_H = \frac{1}{N} \sum_i \sum_{j: y_j \neq y_i} H_{ij} = \sum_k q_k \left(1 - \sum_a p(a \mid k)^2\right). \tag{4}$$

For a designated native semantic class in each component, let r_k be its transported mass and $x_k = q_k - r_k$ the remaining mass. Then

$$\mathcal{C}_H \geq \sum_k \frac{2 r_k x_k}{q_k}. \tag{5}$$

Equation (4) identifies the teacher's wrong-semantic mass with the capacity-weighted impurity of transported components. When mass from other semantic classes mixes with retained native semantic content, Equation (5) quantifies the resulting cross-semantic supervision. The displacement bound locates the forced reassignment; the composition identity determines its semantic consequence. Section 5.3 tests their connection in learned representations.

3.4 From full-bank relations to minibatch supervision

The implemented teacher removes self-relations and applies symmetric degree normalization:

$$H^0 = H - \mathrm{diag}(H), \quad d = H^0 \mathbf{1}, \quad R = \tfrac{1}{2}(D_d^{-1} H^0 + H^0 D_d^{-1}), \quad W_{ij} = \frac{R_{ij}}{\sum_{\ell \neq i} R_{i\ell}}. \tag{6}$$

For a batch $\mathcal{B}$ containing anchor i, its target is $T^{\mathcal{B}}_{ij} = R_{ij}/\sum_{\ell\in\mathcal{B}\setminus\{i\}} R_{i\ell}$ for $j \neq i$, and zero on the diagonal. Thus training samples and renormalizes a row of W. Write $C^{\mathcal{B}}_i = \sum_{j\in\mathcal{B}\setminus\{i\}} T^{\mathcal{B}}_{ij}\mathbf{1}[y_j \neq y_i]$ for its contamination.

The minibatch teacher differs from the full-bank relation through two operations: relation preprocessing and finite-batch sampling.

Theorem 1 (Finite-batch relation connection)**.** *For each anchor, sample $m = B - 1$ neighbors uniformly without replacement from $M = N - 1 > 1$, with $1 \leq m \leq M$. Let W have positive off-diagonal entries and zero diagonal, and set $\kappa_i = M\sum_{j\neq i} W^2_{ij}$, $\delta_i = \mathrm{TV}(W_i, H_i)$, and $f_m = (M-m)/(m(M-1))$. Then*

$$\left|\mathbb{E}_{\mathcal{B}}\frac{1}{N}\sum_i C^{\mathcal{B}}_i - \mathcal{C}_H\right| \leq \frac{1}{N}\sum_i\left[\delta_i + \min\left\{1,\, 2\sqrt{f_m\kappa_i} + 4f_m(\kappa_i - 1)\right\}\right]. \tag{7}$$

Here δ_i measures the preprocessing change, while κ_i measures how strongly a relation row concentrates on a few neighbors; the sampling term vanishes for the full bank. This connects the composition analysis to the normalized training target, including its random row denominator. Appendix A.3 gives the per-anchor bound and proof.

For the distillation loss, write $P_i = \mathrm{softmax}(s_i/\tau)$ and $\ell_i(T) = -\sum_j T_{ij}\log P_{ij}$. Then

$$\nabla_{s_i}\ell_i(T) - \nabla_{s_i}\ell_i(T') = -\tau^{-1}(T_i - T'_i). \tag{8}$$

Capacity-induced changes to the teacher therefore change the distillation gradient directly. The resulting design principle is to estimate capacity from the transport geometry and use that marginal in both directions of the sample–component relation.

4 PACI: Geometry-Matched Relational Learning

PACI alternates periodic teacher refreshes with minibatch representation updates (Figure 2). Following BYOL (Grill et al., 2020), an online encoder f_θ, projector g_θ, and predictor h_θ learn from two augmented views; an EMA target encoder–projector supplies instance targets and the feature bank used to construct relations.

4.1 Estimating capacities from the transport geometry

We estimate the mass of each geometric basin supported by the current target representation. Pairing these capacities with the same prototypes that define the transport costs aligns the assignment constraints with the geometry used to construct the teacher.

At each refresh, we extract deterministic target features $u_i = \mathrm{norm}(g_\xi(f_\xi(x_i)))$ and obtain centered, PCA-processed, normalized features $\widetilde{u}_i$. Spherical K-means gives basin assignments b_i and unit prototypes c_k, defining

$$\widehat{q}_k = \frac{1}{N}\sum_i \mathbf{1}[b_i = k], \qquad C_{ik} = 1 - \widetilde{u}_i^\top c_k. \tag{9}$$

The basin partition specifies each component's total capacity; soft transport satisfies these capacities while allowing each sample to associate with multiple components, yielding sample relations for distillation. Hard basin identities are not used as classification targets.

4.2 Constructing a capacity-matched relation teacher

We solve Equation (1) with $q = \widehat{q}$ and store the resulting teacher through the low-rank factor

$$F = \sqrt{N}\Pi(\widehat{q})D_{\widehat{q}}^{-1/2}, \qquad H = FF^\top. \tag{10}$$

Raw relation entries are inner products $F_iF_j^\top$, and full-bank products use $F(F^\top v)$. These products compute the degrees in Equation (6) without materializing H. For each minibatch, we evaluate its pairwise entries, remove self-relations, apply the stored degrees, and normalize each row to obtain $T^{\mathcal{B}}$. The relation bank requires $O(NK)$ space (Appendix B).

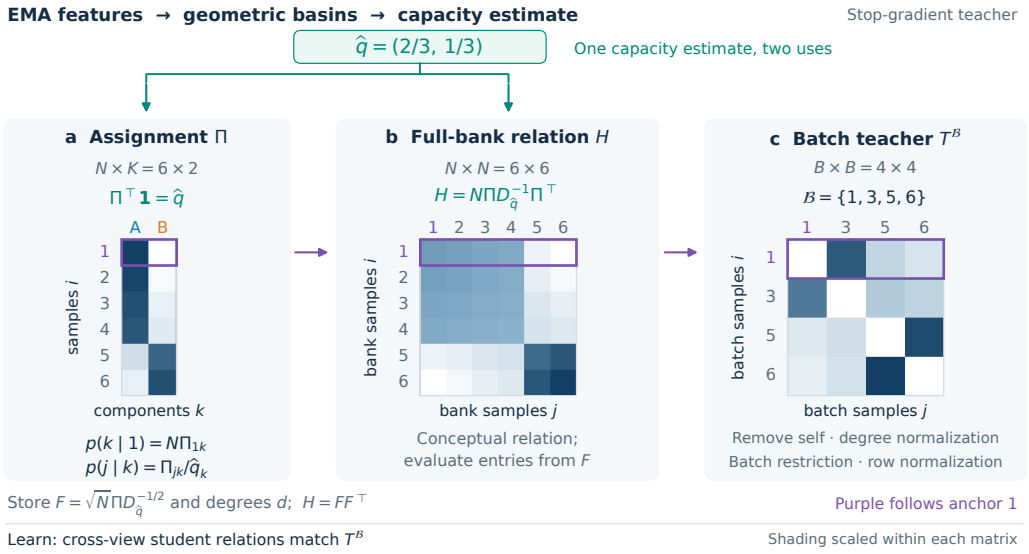

Figure 2: **From a sample–component assignment to the teacher of one anchor.** The schematic uses $N = 6$, $K = 2$, and $B = 4$; shading is scaled within each object. Purple tracks sample 1 across assignment rows, conceptual full-bank relations, and the normalized minibatch teacher. The shared capacity $\widehat{q}$ sets both the transport marginal and reverse conditional. Only the $N \times K$ factor F and degrees are stored; full-bank relations are evaluated through factor products. Periodic refresh updates this stop-gradient teacher.

4.3 Cross-view relational distillation

Write $\mathrm{norm}(\cdot)$ for ℓ_2 normalization and $\mathrm{sg}[\cdot]$ for stop-gradient. For view v, let $z_i^{(v)} = \mathrm{norm}(g_\theta(f_\theta(x_i^{(v)})))$, $p_i^{(v)} = \mathrm{norm}(h_\theta(g_\theta(f_\theta(x_i^{(v)}))))$, and $t_i^{(v)} = \mathrm{sg}[\mathrm{norm}(g_\xi(f_\xi(x_i^{(v)})))]$. The student predicts cross-view relations,

$$S_{ij}^{1\to 2} = \frac{\exp(\langle z_i^{(1)}, t_j^{(2)}\rangle/\tau)}{\sum_{\ell\in\mathcal{B}\setminus\{i\}} \exp(\langle z_i^{(1)}, t_\ell^{(2)}\rangle/\tau)}, \quad j \neq i, \tag{11}$$

and analogously for $S^{2\to 1}$. The symmetric relation loss is

$$\mathcal{L}_{\mathrm{rel}} = -\frac{1}{2B\log(B-1)} \sum_{i\in\mathcal{B}} \sum_{j\in\mathcal{B}\setminus\{i\}} T_{ij}^{\mathcal{B}} \left(\log S_{ij}^{1\to 2} + \log S_{ij}^{2\to 1}\right). \tag{12}$$

We combine it with the instance-level bootstrap objective (Grill et al., 2020):

$$\mathcal{L} = \mathcal{L}_{\mathrm{BYOL}} + \beta(e)\mathcal{L}_{\mathrm{rel}}, \quad \mathcal{L}_{\mathrm{BYOL}} = \frac{1}{2B}\sum_i \left[2 - 2\langle p_i^{(1)}, t_i^{(2)}\rangle + 2 - 2\langle p_i^{(2)}, t_i^{(1)}\rangle\right]. \tag{13}$$

The instance objective learns the initial representation used to estimate basins. Relation learning then activates with a gradual ramp in $\beta(e)$. The target network and relation bank are stop-gradient objects; the student updates its representation through both objectives. EMA updates and periodic bank refreshes incorporate the learned geometry into the next capacity estimate. Training schedules and construction parameters are given in Section 5 and Appendix B.

5 Experiments

Benchmarks test overall competitiveness and the gain over the underlying representation learner. Fixed-geometry interventions examine teacher distortion and basin matching; controlled training then separates the capacity benefit from its assignment and normalization contributions.

Table 1: Class-averaged accuracy (%) on five long-tailed benchmarks. Subscripts give reported standard deviations. Bold marks the best reported mean. IF5 results and all four metrics are given in Appendix C.

Method	CIFAR-10 IF10	CIFAR-20 IF10	STL-10 IF10	CIFAR-100 IF100	Tiny-ImageNet IF100
SCAN (Van Gansbeke et al., 2020)	$51.24_{\pm 1.23}$	$32.12_{\pm 1.35}$	$43.14_{\pm 1.36}$	$11.98_{\pm 0.38}$	$12.25_{\pm 1.42}$
SDCLR (Jiang et al., 2021)	$47.40_{\pm 0.85}$	$34.20_{\pm 0.26}$	$38.13_{\pm 1.07}$	$19.50_{\pm 1.43}$	$15.20_{\pm 0.57}$
CC (Li et al., 2021)	$42.74_{\pm 0.59}$	$24.31_{\pm 0.14}$	$37.34_{\pm 1.23}$	$18.12_{\pm 0.14}$	$14.12_{\pm 0.96}$
IDFD (Tao et al., 2021)	$55.62_{\pm 0.86}$	$26.41_{\pm 1.04}$	$35.38_{\pm 0.47}$	$20.05_{\pm 1.07}$	$12.12_{\pm 1.04}$
CoNR (Yu et al., 2023)	$43.64_{\pm 0.31}$	$18.72_{\pm 1.46}$	$31.77_{\pm 0.47}$	$18.50_{\pm 0.57}$	$14.40_{\pm 1.23}$
ProPos (Huang et al., 2023)	$50.18_{\pm 0.86}$	$32.91_{\pm 0.96}$	$39.06_{\pm 0.62}$	$25.96_{\pm 0.09}$	$15.88_{\pm 1.35}$
DMICC (Li et al., 2023)	$38.76_{\pm 1.25}$	$22.63_{\pm 0.05}$	$41.11_{\pm 1.15}$	$18.90_{\pm 0.79}$	$14.80_{\pm 0.77}$
SeCu (Qian, 2023)	$58.12_{\pm 0.38}$	$29.86_{\pm 0.94}$	$45.47_{\pm 0.67}$	$25.08_{\pm 1.14}$	$16.44_{\pm 1.38}$
LFSS (Li et al., 2025)	$58.83_{\pm 0.71}$	$41.06_{\pm 0.88}$	$43.22_{\pm 1.23}$	$25.48_{\pm 0.96}$	$17.12_{\pm 0.57}$
ConMix (Li & Jia, 2025)	$56.08_{\pm 0.23}$	$38.13_{\pm 0.85}$	$46.94_{\pm 2.16}$	$24.77_{\pm 0.72}$	$13.15_{\pm 0.03}$
MiniClustering (Li et al., 2026)	$57.52_{\pm 1.97}$	$41.53_{\pm 1.82}$	$47.63_{\pm 2.60}$	$25.55_{\pm 0.90}$	$14.89_{\pm 0.62}$
SimCLR (Chen et al., 2020)	$56.50_{\pm 0.62}$	$40.70_{\pm 1.33}$	$47.20_{\pm 0.67}$	$25.23_{\pm 0.53}$	$13.98_{\pm 0.36}$
BYOL (Grill et al., 2020)	$54.01_{\pm 1.94}$	$37.00_{\pm 2.11}$	$38.78_{\pm 1.18}$	$25.29_{\pm 0.81}$	$14.07_{\pm 0.43}$
PACI	$\mathbf{64.36}_{\pm 2.29}$	$\mathbf{42.18}_{\pm 1.24}$	$\mathbf{51.29}_{\pm 2.03}$	$\mathbf{29.60}_{\pm 0.40}$	$\mathbf{18.72}_{\pm 0.33}$

5.1 Experimental setup

Datasets and evaluation. We evaluate CIFAR-10, CIFAR-20 (the 20 superclasses of CIFAR-100), STL-10 (Krizhevsky, 2009; Coates et al., 2011), CIFAR-100, and Tiny-ImageNet. The main table reports IF10 for the first three datasets and IF100 for the latter two; additional IF5 results appear in Appendix C. Exponentially imbalanced subsets have class counts $n_c = n_{\max}\mathrm{IF}^{-c/(K-1)}$, up to integer rounding. The learner knows K and uses unlabeled training examples. PACI and BYOL evaluation uses spherical K-means on EMA target-projector features and Hungarian alignment. Our primary metric, class-averaged accuracy (CAA), weights classes equally; ACC, NMI, and ARI characterize the complete partition. All reported metrics are percentages.

Training and controlled comparisons. All results use three random seeds, 4801–4803. Across all five datasets, PACI uses ResNet-18 (He et al., 2016) for 800 epochs with batch size 256. All five benchmarks use relation weight $\beta = 0.05$, temperature $\tau = 0.40$, and refresh every 25 epochs. Controlled training on CIFAR-10/20 and STL-10 uses a common epoch-200 checkpoint, $\beta = 0.20$, $\tau = 0.20$, and refresh every 50 epochs. Within each controlled comparison, data, optimizer state, loss, evaluator, and training schedule are matched. Both start relation learning at epoch 200 after BYOL warm-up, with a 100-epoch ramp. Appendix B provides implementation and reporting details.

5.2 Comparison with state-of-the-art methods

PACI exceeds ConMix and MiniClustering in CAA across all five settings in Table 1. The complete results in Appendix C also show metric-specific differences: MiniClustering has higher ACC on CIFAR-100, while other methods lead ACC, NMI, and ARI on Tiny-ImageNet.

Against its BYOL foundation, PACI improves CAA by 10.35, 5.18, and 12.51 points on CIFAR-10, CIFAR-20, and STL-10 at IF10, respectively. The corresponding gains are 4.31 and 4.65 points on CIFAR-100 and Tiny-ImageNet at IF100.

5.3 Does capacity mismatch distort the relation teacher?

We freeze features, basin identities, prototypes, and transport cost in each of 18 controlled units (three datasets, two IFs, three seeds). The intervention follows $q(\lambda) = (1-\lambda)\widehat{q} + \lambda u$, with $\lambda \in \{0, .25, .5, .75, 1\}$, using the path capacity consistently in assignment and relation construction. The diagnostics are coupling displacement, cross-basin mass, and sample-weighted wrong-semantic mass in the implemented minibatch teacher.

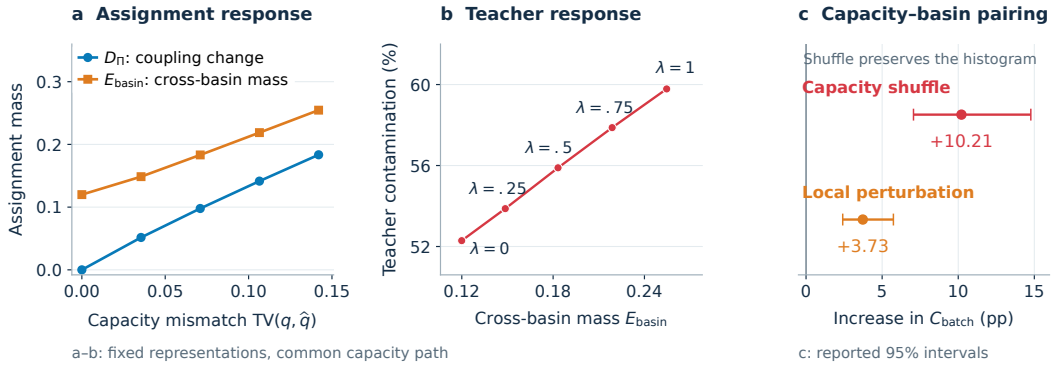

Figure 3: **Capacity mismatch and basin pairing shape the teacher.** (a,b) Joint responses along the capacity path on 18 fixed final representations. (c) Contamination increases relative to matched capacity, with 95% hierarchical-bootstrap intervals. Increases occur in 72/72 shuffle and 142/144 local-perturbation conditions. Shuffling preserves the capacity histogram.

All three diagnostics increase monotonically in all 18 final geometries. Uniform capacity increases coupling displacement by 0.183, cross-basin mass by 0.135, and contamination by 7.49 points. Their joint responses in Figure 3a–b are consistent with the transport and composition analysis (endpoint intervals in Appendix D).

At the common epoch-200 geometry, the endpoint contamination increase is 0.88 points and 16/18 paths are monotone; capacity mismatch has a larger teacher-level effect on final than on warm-up representations.

5.4 Does capacity need to match the basin identity?

Shuffling capacities among basins preserves their histogram but raises contamination by 10.21 points on average, with increases in all 72 conditions. Local perturbations raise it by 3.73 points on average (142/144 increases). These interventions on 18 final representations show that capacity–basin pairing matters beyond nonuniformity alone (Figure 3c; intervals in Table 11).

5.5 Which capacity choice drives the training benefit?

Capacity-matched versus balanced relations. Comparing PACI with BYOL evaluates the complete method. Matched versus balanced relations isolate capacity matching within the same relational learner: both branches share the warm-up checkpoint, training objective, and teacher-refresh rule, while using either basin-estimated capacity ($\widehat{q}$) or uniform capacity (u) in both transport assignment and relation construction.

Matched capacity improves final CAA by 3.516 points on average (95% CI $[1.218, 6.413]$), with positive means in all six dataset–imbalance settings and gains in 15/18 paired runs. ACC, NMI, and ARI also improve, by 5.200, 3.125, and 7.039 points, respectively.

Assignment versus relation normalization. To locate this training benefit, we next distinguish the assignment and normalization pathways by independently intervening on assignment capacity q_A and relation-normalization capacity q_R:

$$H(q_A, q_R) = N\Pi(q_A)D_{q_R}^{-1}\Pi(q_A)^\top. \tag{14}$$

Estimated (E) and uniform (U) capacities define EE (PACI), EU, UE, and UU (balanced). All four arms share the warm-up state and training procedure (Figure 4).

Replacing estimated assignment with uniform assignment lowers CAA by 3.632 points under estimated normalization and 3.168 points under uniform normalization. In contrast, changing normalization lowers CAA by 0.348 points with estimated assignments and improves it by 0.116 points with uniform assignments. Assignment thus dominates under both normalization choices, while normalization has an assignment-dependent effect.

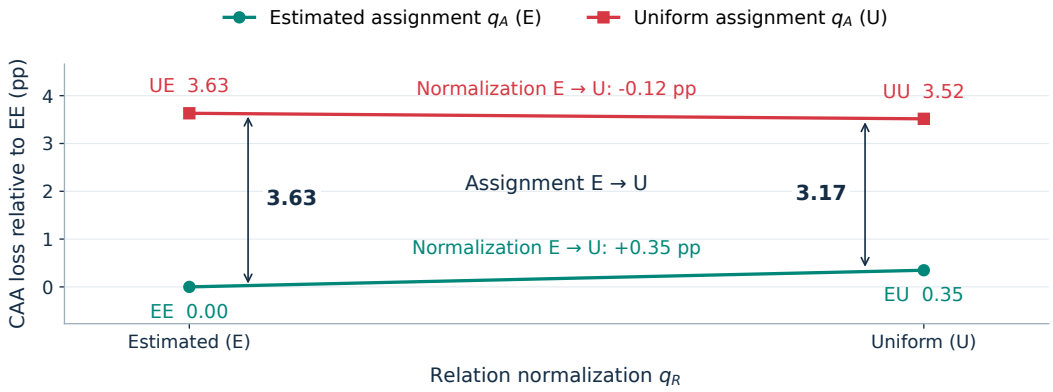

Figure 4: **Assignment changes dominate under either normalization.** Final CAA loss relative to EE across 18 paired units. E/U denote estimated/uniform capacity; arm codes list assignment first, normalization second. Vertical gaps and line slopes show the conditional assignment and normalization effects.

On final frozen geometries, a separate teacher-level intervention gives an average assignment effect of $+9.03$ contamination points (95% CI $[5.72, 13.24]$), versus -1.54 ($[-2.12, -1.02]$) for normalization. This larger assignment effect is consistent with Equation (14): assignment changes component composition, whereas normalization reweights existing components. Appendix D reports conditional effects and the secondary contamination–accuracy association.

5.6 Transfer across training configurations and long-tailed settings

The benchmark configuration changes relation weight, temperature, and refresh frequency across 18 units with the same three seeds. Displacement and cross-basin paths are monotone in all 18 units; contamination is monotone in 17/18. Nine CIFAR100-LT geometries ($K = 100$; IFs 10, 50, 100; seeds 4801–4803) all exhibit monotone contamination paths. Assignment dominance persists in both panels (Table 2), extending the teacher-level finding across configurations and cluster scales.

Table 2: Transfer of the capacity intervention. Endpoint changes compare uniform and estimated capacities; $|A|/|R|$ is the ratio of mean factorial effect magnitudes on contamination. CIFAR100 reports teacher diagnostics.

Panel	ΔD_{Π}	ΔE_{basin}	ΔC (pp)	$\lvert A\rvert/\lvert R\rvert$
Benchmark configuration (18 units)	0.201	0.123	+7.11	6.11
CIFAR100-LT (9 units)	0.193	0.073	+2.90	3.39

6 Conclusion

We identified capacity–geometry mismatch as a source of distortion in long-tailed relational learning. Prescribed component capacities determine which assignments are feasible, and these assignments carry semantic structure into the relation teacher. Our analysis characterizes the required transport displacement, the teacher's semantic composition, and its connection to finite-batch supervision. PACI estimates evolving pseudo-basin capacities and uses them consistently in transport assignment and relation construction. Benchmark results and matched training interventions show the performance benefit of geometry matching, with assignment accounting for the dominant capacity effect. These results support matching capacities to representation geometry when constructing relational teachers for long-tailed data.

Reproducibility Statement

Proofs of our theoretical results are provided in Appendix A. Experimental settings and implementation details are described in Section 5.1 and Appendix B. Code is available at the anonymous repository linked in the abstract.

AI Use Statement

We used large language models (LLMs) to assist with drafting sections from our existing drafts and outlines, and to improve grammar, clarity, coherence, and flow. LLMs also assisted with literature and reference checks, figure and table preparation, and the interpretation and presentation of reported experimental results. The authors take full responsibility for the paper's content, including all AI-assisted text, claims, and artifacts.

References

Yuki M. Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1911.05371v3.

Luis A. Caffarelli and Robert J. McCann. Free boundaries in optimal transport and Monge-Ampère obstacle problems. *Annals of Mathematics*, 171(2):673–730, 2010. doi: 10.4007/annals.2010.171.673. URL https://annals.math.princeton.edu/2010/171-2/p02.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision*, pp. 132–149, 2018. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Mathilde_Caron_Deep_Clustering_for_ECCV_2018_paper.html.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/70feb62b69f16e0238f741fab228fec2-Abstract.html.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper.html.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

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

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pp. 215–223. PMLR, 2011. URL https://proceedings.mlr.press/v15/coates11a.html.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar,

Bilal Piot, Koray Kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf.

Wengang Guo, Wei Ye, Chunchun Chen, Xin Sun, Christian Böhm, Claudia Plant, and Susanto Rahardja. Bootstrap deep spectral clustering with optimal transport. *IEEE Transactions on Multimedia*, 28:531–544, 2026. doi: 10.1109/TMM.2025.3623492.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Zhizhong Huang, Jie Chen, Junping Zhang, and Hongming Shan. Learning representation for clustering via prototype scattering and positive sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7509–7524, 2023. doi: 10.1109/TPAMI.2022.3216454. URL https://arxiv.org/abs/2111.11821.

Ziyu Jiang, Tianlong Chen, Bobak J Mortazavi, and Zhangyang Wang. Self-damaging contrastive learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4927–4939. PMLR, 2021. URL https://proceedings.mlr.press/v139/jiang21a.html.

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

Hongyu Li, Lefei Zhang, and Kehua Su. Dual mutual information constraints for discriminative clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8571–8579, 2023. doi: 10.1609/aaai.v37i7.26032. URL https://ojs.aaai.org/index.php/AAAI/article/view/26032.

Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8547–8555, 2021. doi: 10.1609/aaai.v35i10.17037. URL https://arxiv.org/abs/2009.09687.

Zhixin Li and Yuheng Jia. ConMix: Contrastive mixup at representation level for long-tailed deep clustering. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=3lH8WT0fhu.

Zhixin Li, Yuheng Jia, Hui Liu, and Junhui Hou. Learning from sample stability for deep clustering. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 34904–34919. PMLR, 2025. URL https://proceedings.mlr.press/v267/li25am.html.

Zhixin Li, Yuheng Jia, Guanliang Chen, Hui Liu, and Junhui Hou. Mini-cluster guided long-tailed deep clustering. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=3JlljaiQwR.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Park_Relational_Knowledge_Distillation_CVPR_2019_paper.html.

Qi Qian. Stable cluster discrimination for deep clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16645–16654, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Qian_Stable_Cluster_Discrimination_for_Deep_Clustering_ICCV_2023_paper.html.

Liangliang Shi, Zhaoqi Shen, and Junchi Yan. Double-bounded optimal transport for advanced clustering and classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14982–14990, 2024. doi: 10.1609/aaai.v38i13.29419. URL https://doi.org/10.1609/aaai.v38i13.29419.

Yaling Tao, Kentaro Takagi, and Kouta Nakata. Clustering-friendly representation learning via instance discrimination and feature decorrelation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=e12NDM7wkEY.

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. SCAN: Learning to classify images without labels. In *European Conference on Computer Vision*, pp. 268–285, 2020. URL https://arxiv.org/abs/2005.12320.

Haobo Wang, Mingxuan Xia, Yixuan Li, Yuren Mao, Lei Feng, Gang Chen, and Junbo Zhao. SoLar: Sinkhorn label refinery for imbalanced partial-label learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 8104–8117, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/357a0a771bf65ee07926d6af41b75030-Abstract-Conference.html.

Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pp. 478–487. PMLR, 2016. URL https://proceedings.mlr.press/v48/xieb16.html.

Xuqian Xue, Yiming Lei, Qi Cai, Hongming Shan, and Junping Zhang. PROTOCOL: Partial optimal transport-enhanced contrastive learning for imbalanced multi-view clustering. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70105–70119. PMLR, 2025. URL https://proceedings.mlr.press/v267/xue25c.html.

Zhihao Yao, Yuxuan Gu, Jixuan Yin, and Bo Li. Beyond looking up, try looking around: Harmonizing global structure and local consistency in optimal transport for short text clustering. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=0EANaZfd1l.

Chunlin Yu, Ye Shi, and Jingya Wang. Contextually affinitive neighborhood refinery for deep clustering. In *Advances in Neural Information Processing Systems*, volume 36, pp. 5778–5790, 2023. doi: 10.52202/075280-0253. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/123cfe7d8b7702ac97aaf4468fc05fa5-Abstract-Conference.html.

Chuyu Zhang, Hui Ren, and Xuming He. P^2OT: Progressive partial optimal transport for deep imbalanced clustering. In *International Conference on Learning Representations*, 2024a. URL https://arxiv.org/abs/2401.09266.

Chuyu Zhang, Hui Ren, and Xuming He. SP^2OT: Semantic-regularized progressive partial optimal transport for imbalanced clustering. *arXiv preprint arXiv:2404.03446*, 2024b. URL https://arxiv.org/abs/2404.03446v2. Revised June 2025.

Hongyuan Zhang, Jiankun Shi, Rui Zhang, and Xuelong Li. Non-graph data clustering via $\mathcal{O}(n)$ bipartite graph convolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (7):8729–8742, 2023. doi: 10.1109/TPAMI.2022.3231470.

Mingkai Zheng, Shan You, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. ReSSL: Relational self-supervised learning with weak augmentation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 2543–2555, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/14c4f36143b4b09cbc320d7c95a50ee7-Paper.pdf.

Xiaolin Zheng, Mengling Hu, Weiming Liu, Chaochao Chen, and Xinting Liao. Robust representation learning with reliable pseudo-labels generation via self-adaptive optimal transport for short text clustering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10493–10507. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.585. URL https://aclanthology.org/2023.acl-long.585/.

A Proofs and Further Analysis

A.1 Capacity-induced displacement

For Proposition 1, let $\Delta = \Pi - \Pi_0$. Marginalization contracts the entrywise ℓ_1 norm:

$$\|q - \widehat{q}\|_1 = \sum_k \left| \sum_i \Delta_{ik} \right| \le \sum_{ik} |\Delta_{ik}|. \tag{15}$$

Dividing by two gives the coupling bound. For the cross-basin bound, let $s_k = \sum_{i:b_i=k} \Pi_{ik}$. Source basin k contains mass $\widehat{q}_k$ and destination k receives q_k, so $s_k \le \min(\widehat{q}_k, q_k)$. Therefore

$$E_{\text{basin}} = 1 - \sum_k s_k \ge 1 - \sum_k \min(\widehat{q}_k, q_k) = \mathrm{TV}(q, \widehat{q}). \tag{16}$$

Retaining $\min(\widehat{q}_k, q_k)$ in each basin and routing surplus mass to deficit components attains the bound in the transport polytope. Entropy regularization determines the actual coupling within that polytope; Equation (3) specifies its compulsory floor.

Construction of Figure 1. The schematic fixes ten samples in two semantically aligned basins, with eight in A and two in B. For uniform capacity, each A sample has coupling row $(0.0625, 0.0375)$ and each B sample $(0, 0.1)$; for matched capacity these rows are $(0.1, 0)$ and $(0, 0.1)$. Both constructions have row mass $1/10$. Their column sums are $(0.5, 0.5)$ and $(0.8, 0.2)$, respectively. These illustrative, unregularized couplings attain the corresponding cross-basin floors. The matrices plot $H = 10\Pi D_q^{-1}\Pi^\top$ with a shared square-root color normalization over $[0, 0.5]$. Red outlines identify the A–B and B–A blocks.

A.2 Semantic composition of the teacher

To prove Proposition 2, first note that Equation (2) has $H\mathbf{1} = N\Pi D_q^{-1} q = N\Pi\mathbf{1} = \mathbf{1}$. Its average equal-semantic mass is

$$\frac{1}{N}\sum_i \sum_{j:y_j=y_i} H_{ij} = \sum_k \frac{1}{q_k} \sum_a \left(\sum_{i:y_i=a} \Pi_{ik} \right)^2 = \sum_k \frac{1}{q_k} \sum_a A_{ak}^2. \tag{17}$$

Subtracting this expression from one yields Equation (4). For a designated native class ℓ_k, set $r_k = A_{\ell_k k}$ and $x_k = \sum_{a \ne \ell_k} A_{ak}$. Since $\sum_{a \ne \ell_k} A_{ak}^2 \le x_k^2$,

$$q_k - \frac{1}{q_k}\sum_a A_{ak}^2 \ge q_k - \frac{r_k^2 + x_k^2}{q_k} = \frac{2 r_k x_k}{q_k}. \tag{18}$$

This proves Equation (5). In particular, if native-class retention satisfies $r_k \ge \alpha q_k$ and incompatible mass satisfies $\sum_k x_k \ge \eta E_{\text{basin}}$, then

$$\mathcal{C}_H \ge 2\alpha\eta E_{\text{basin}} \ge 2\alpha\eta\, \mathrm{TV}(q, \widehat{q}). \tag{19}$$

The native-class designation is fixed for the semantic analysis of each destination component.

A.3 Finite-batch relation connection

The per-anchor statement underlying Theorem 1 makes the sampling contribution explicit. Under its sampling rule, set $c_i = \sum_{j \ne i} W_{ij} \mathbf{1}[y_j \ne y_i]$ and $\kappa_i = M \sum_{j \ne i} W_{ij}^2$. Then

$$\mathbb{E}_{\mathcal{B}} |C_i^{\mathcal{B}} - c_i| \le e_i(m) := \min\{1,\, 2\sqrt{f_m \kappa_i} + 4 f_m(\kappa_i - 1)\}. \tag{20}$$

The proof controls the sampled numerator and the probability that its normalization denominator becomes small. Fix an anchor and abbreviate $w_j = W_{ij}$, $z_j = \mathbf{1}[y_j \ne y_i]$, and $c = \sum_j w_j z_j$. There are M neighbors, with $w_j > 0$ and $\sum_j w_j = 1$. Draw a uniform size-m subset S, where $1 \le m \le M$, and define

$$Z = \sum_{j \in S} w_j, \qquad U = \sum_{j \in S} w_j (z_j - c), \qquad \mu = m/M. \tag{21}$$

Then $C_i^{\mathcal{B}} - c = U/Z$, $\mathbb{E}Z = \mu$, and $\mathbb{E}U = 0$ because $\sum_j w_j(z_j - c) = 0$. The exact without-replacement variance formula is

$$\mathrm{Var}\left(\sum_{j\in S} v_j\right) = \frac{m(M-m)}{M(M-1)} \sum_j (v_j - \overline{v})^2. \tag{22}$$

Writing $\kappa = M\sum_j w_j^2$, $\nu = M\sum_j w_j^2(z_j - c)^2$, and $f_m = (M-m)/(m(M-1))$, we obtain

$$\frac{\mathrm{Var}(Z)}{\mu^2} = f_m(\kappa - 1), \qquad \frac{\mathrm{Var}(U)}{\mu^2} = f_m\nu \le f_m\kappa. \tag{23}$$

On $Z \ge \mu/2$, $|U/Z| \le 2|U|/\mu$. On its complement, $|C_i^{\mathcal{B}} - c| \le 1$. Cauchy–Schwarz and Chebyshev's inequality give

$$\mathbb{E}|C_i^{\mathcal{B}} - c| \le \frac{2\mathbb{E}|U|}{\mu} + \Pr(Z < \mu/2) \tag{24}$$

$$\le 2\sqrt{f_m\nu} + 4f_m(\kappa - 1) \tag{25}$$

$$\le 2\sqrt{f_m\kappa} + 4f_m(\kappa - 1). \tag{26}$$

Clipping the right side at one proves Equation (20); retaining ν gives the sharper semantic bound.

Extend W_i to all N entries with $W_{ii} = 0$. Both W_i and H_i are probability distributions, so for the wrong-semantic subset,

$$\left| c_i - \sum_{j:y_j \ne y_i} H_{ij} \right| \le \mathrm{TV}(W_i, H_i) = \delta_i. \tag{27}$$

Triangle inequality, averaging over anchors, and linearity of expectation prove Equation (7). Anchor batches need not be independent for this average bound.

Comparing two teachers. Let W, W' be two processed full-bank teachers, with average contaminations $\overline{c}, \overline{c}'$ and sampling bounds $\overline{e}, \overline{e}'$. Applying the theorem to each teacher yields

$$|\mathbb{E}(C_{\mathrm{batch}} - C'_{\mathrm{batch}}) - (\overline{c} - \overline{c}')| \le \overline{e} + \overline{e}'. \tag{28}$$

Consequently a processed full-bank contamination difference exceeding the combined sampling bound retains its sign in expectation. For raw relations, the corresponding processing discrepancies are included as in Equation (7).

Distillation gradients. For the row objective in Equation (8), differentiation gives $\partial\ell_i/\partial s_{ij} = \tau^{-1}(P_{ij} - T_{ij})$. Subtracting two teacher-induced gradients proves the identity. Each direction of the implemented loss in Equation (12) carries the additional factor $[2B\log(B-1)]^{-1}$, and the joint objective multiplies its contribution by $\beta(e)$.

A.4 Assignment and relation normalization

Write the kth coupling column as $\pi_k(q_A)$. The intervention operator in Equation (14) decomposes as

$$H(q_A, q_R) = \sum_k \frac{N}{q_{R,k}} \pi_k(q_A)\pi_k(q_A)^\top. \tag{29}$$

Holding q_A fixed preserves the component composition and changes its weight. For globally normalized raw relation mass, set $g_k = 1 - \sum_a p(a \mid k)^2$ and

$$\omega_k = \frac{q_{A,k}^2/q_{R,k}}{\sum_\ell q_{A,\ell}^2/q_{R,\ell}}. \tag{30}$$

The raw contamination is then $\sum_k \omega_k g_k$. Two choices r, r' of the normalization capacity q_R satisfy

$$|\widetilde{C}(r) - \widetilde{C}(r')| \le (\max_k g_k - \min_k g_k)\,\mathrm{TV}(\omega(r), \omega(r')). \tag{31}$$

The implemented interventions apply the same self-relation removal and degree/batch normalization to all four cells. The finite-batch theorem applies directly to their processed rows.

B Additional Method Details

Low-rank normalization. Using the factor F in Equation (10), let $h_i = \|F_i\|_2^2$. For any vector v, $H^0 v = F(F^\top v) - h \odot v$. Thus the degrees in Equation (6), $d = H^0\mathbf{1}$ and $R\mathbf{1} = \frac{1}{2}[\mathbf{1} + H^0(d^{-1})]$ can be computed with the factor alone. Each batch uses pairwise factor products and the stored degrees to construct R_{ij} before row normalization. Relation storage and full-bank matrix–vector products cost $O(NK)$; forming a batch relation costs $O(B^2K)$, apart from feature extraction, clustering, and transport.

Architecture and optimization. ResNet-18 is trained from scratch on all five datasets, using a 3×3 stride-one stem without max pooling on CIFAR-10/20/100 and the standard stem on STL-10 and Tiny-ImageNet. The projector and predictor have two layers and a hidden dimension of 4096. Output dimensions are 512 on CIFAR-10/20/100 and 256 on STL-10 and Tiny-ImageNet. SGD uses learning rate 0.05, momentum 0.9, weight decay 5×10^{-4}, and a $10\times$ predictor learning-rate multiplier. Learning rate increases linearly for 10 epochs and then follows cosine decay. Target parameters use EMA momentum 0.996. Training views use random resized crop with scale $[0.08, 1]$, horizontal flipping, color jitter, and random grayscale; relation refresh uses deterministic evaluation views.

Relation construction. All five datasets use the same relation construction: target-projector features are centered, reduced by PCA to at most 384 dimensions when necessary, and normalized before clustering. Spherical K-means uses 12 initializations and at most 100 iterations. Sinkhorn uses $\varepsilon = 0.12$, at most 2000 iterations, and marginal tolerance 10^{-6}. Evaluation uses spherical K-means with 10 initializations and at most 300 iterations. Section 5.1 specifies the benchmark and controlled configurations, seeds, and relation schedules.

Uncertainty and aggregation. All benchmark entries report means and standard deviations over three runs with seeds 4801–4803. Controlled training comparisons use the same three seeds and 10,000 hierarchical-bootstrap draws. Teacher-level summaries use 20,000 draws over dataset, IF, and seed, keeping repeated interventions associated with their representation unit.

Comparison sources. The result tables cite each comparison method's original paper. For the five main-table settings, we report means and standard deviations for all methods. Additional IF5 results come from the PACI/BYOL and ConMix/MiniClustering reproductions.

C Complete Benchmark Results

The following tables extend Table 1 to ACC, CAA, NMI, and ARI for all eight settings, including IF5. Metrics are percentages; subscripts give reported standard deviations. Only methods with available results are listed for each setting, and bold marks the best reported mean. Run counts and comparison sources are given in Appendix B.

Table 3: CIFAR-10 at IF5.

Method	ACC	CAA	NMI	ARI
ConMix (Li & Jia, 2025)	$62.86_{\pm 0.99}$	$64.33_{\pm 1.29}$	$56.76_{\pm 4.24}$	$45.62_{\pm 1.41}$
MiniClustering (Li et al., 2026)	$68.45_{\pm 1.83}$	$68.08_{\pm 0.41}$	$\mathbf{66.06}_{\pm 0.39}$	$54.35_{\pm 0.88}$
BYOL (Grill et al., 2020)	$64.54_{\pm 0.79}$	$68.58_{\pm 0.88}$	$59.52_{\pm 1.42}$	$45.87_{\pm 2.35}$
PACI	$\mathbf{73.24}_{\pm 1.51}$	$\mathbf{72.91}_{\pm 1.30}$	$65.86_{\pm 2.47}$	$\mathbf{61.55}_{\pm 1.92}$

Table 4: CIFAR-20 at IF5.

Method	ACC	CAA	NMI	ARI
ConMix (Li & Jia, 2025)	$42.55_{\pm 1.18}$	$39.90_{\pm 0.66}$	$44.59_{\pm 0.59}$	$27.21_{\pm 0.91}$
MiniClustering (Li et al., 2026)	$44.40_{\pm 1.50}$	$43.37_{\pm 1.40}$	$49.31_{\pm 1.27}$	$31.17_{\pm 0.51}$
BYOL (Grill et al., 2020)	$40.22_{\pm 0.64}$	$39.03_{\pm 1.33}$	$43.71_{\pm 1.33}$	$25.60_{\pm 0.82}$
PACI	$\mathbf{48.95}_{\pm 2.17}$	$\mathbf{46.62}_{\pm 0.87}$	$\mathbf{51.12}_{\pm 1.22}$	$\mathbf{34.66}_{\pm 1.83}$

Table 5: STL-10 at IF5.

Method	ACC	CAA	NMI	ARI
ConMix (Li & Jia, 2025)	$47.99_{\pm 1.02}$	$48.37_{\pm 0.64}$	$47.06_{\pm 0.47}$	$34.36_{\pm 1.30}$
MiniClustering (Li et al., 2026)	$51.77_{\pm 0.79}$	$50.20_{\pm 1.05}$	$\mathbf{51.50}_{\pm 1.76}$	$\mathbf{40.89}_{\pm 1.08}$
BYOL (Grill et al., 2020)	$47.89_{\pm 1.05}$	$47.23_{\pm 1.00}$	$50.07_{\pm 1.20}$	$34.33_{\pm 0.20}$
PACI	$\mathbf{52.54}_{\pm 0.51}$	$\mathbf{51.64}_{\pm 1.59}$	$50.90_{\pm 0.20}$	$40.79_{\pm 0.54}$

Table 6: CIFAR-10 at IF10.

Method	ACC	CAA	NMI	ARI
SCAN (Van Gansbeke et al., 2020)	$48.21_{\pm 0.50}$	$51.24_{\pm 1.23}$	$47.56_{\pm 0.33}$	$34.71_{\pm 0.08}$
SDCLR (Jiang et al., 2021)	$44.63_{\pm 1.40}$	$47.40_{\pm 0.85}$	$44.78_{\pm 1.28}$	$31.12_{\pm 1.47}$
CC (Li et al., 2021)	$36.86_{\pm 0.83}$	$42.74_{\pm 0.59}$	$26.32_{\pm 0.93}$	$19.81_{\pm 0.60}$
IDFD (Tao et al., 2021)	$53.82_{\pm 0.71}$	$55.62_{\pm 0.86}$	$54.54_{\pm 1.09}$	$39.35_{\pm 0.16}$
CoNR (Yu et al., 2023)	$43.43_{\pm 0.24}$	$43.64_{\pm 0.31}$	$44.01_{\pm 0.41}$	$31.75_{\pm 0.47}$
ProPos (Huang et al., 2023)	$48.33_{\pm 1.40}$	$50.18_{\pm 0.86}$	$48.98_{\pm 1.06}$	$35.33_{\pm 0.45}$
DMICC (Li et al., 2023)	$38.72_{\pm 1.11}$	$38.76_{\pm 1.25}$	$39.24_{\pm 0.97}$	$28.31_{\pm 1.36}$
SeCu (Qian, 2023)	$55.87_{\pm 0.09}$	$58.12_{\pm 0.38}$	$56.62_{\pm 0.35}$	$40.85_{\pm 1.45}$
LFSS (Li et al., 2025)	$58.52_{\pm 0.16}$	$58.83_{\pm 0.71}$	$59.31_{\pm 1.36}$	$42.79_{\pm 1.25}$
ConMix (Li & Jia, 2025)	$52.13_{\pm 0.91}$	$56.08_{\pm 0.23}$	$55.59_{\pm 1.06}$	$39.70_{\pm 1.52}$
MiniClustering (Li et al., 2026)	$59.30_{\pm 1.17}$	$57.52_{\pm 1.97}$	$55.66_{\pm 2.17}$	$41.73_{\pm 0.69}$
SimCLR (Chen et al., 2020)	$55.38_{\pm 0.68}$	$56.50_{\pm 0.62}$	$56.13_{\pm 1.09}$	$40.49_{\pm 0.70}$
BYOL (Grill et al., 2020)	$49.61_{\pm 1.34}$	$54.01_{\pm 1.94}$	$53.77_{\pm 0.83}$	$37.78_{\pm 1.46}$
PACI	$\mathbf{67.74}_{\pm 1.30}$	$\mathbf{64.36}_{\pm 2.29}$	$\mathbf{65.11}_{\pm 0.78}$	$\mathbf{59.54}_{\pm 4.88}$

Table 7: CIFAR-20 at IF10.

Method	ACC	CAA	NMI	ARI
SCAN (Van Gansbeke et al., 2020)	$32.54_{\pm 1.03}$	$32.12_{\pm 1.35}$	$35.78_{\pm 0.89}$	$23.75_{\pm 0.83}$
SDCLR (Jiang et al., 2021)	$37.10_{\pm 1.11}$	$34.20_{\pm 0.26}$	$42.67_{\pm 0.49}$	$27.95_{\pm 0.51}$
CC (Li et al., 2021)	$28.46_{\pm 1.49}$	$24.31_{\pm 0.14}$	$16.37_{\pm 1.13}$	$9.78_{\pm 0.78}$
IDFD (Tao et al., 2021)	$29.02_{\pm 0.93}$	$26.41_{\pm 1.04}$	$30.74_{\pm 1.16}$	$18.91_{\pm 0.29}$
CoNR (Yu et al., 2023)	$19.10_{\pm 0.34}$	$18.72_{\pm 1.46}$	$20.23_{\pm 0.01}$	$12.44_{\pm 1.40}$
ProPos (Huang et al., 2023)	$35.84_{\pm 0.46}$	$32.91_{\pm 0.96}$	$37.98_{\pm 1.26}$	$23.36_{\pm 0.45}$
DMICC (Li et al., 2023)	$23.36_{\pm 0.63}$	$22.63_{\pm 0.05}$	$24.75_{\pm 0.69}$	$15.22_{\pm 0.72}$
SeCu (Qian, 2023)	$32.75_{\pm 1.08}$	$29.86_{\pm 0.94}$	$34.70_{\pm 0.96}$	$21.34_{\pm 1.43}$
LFSS (Li et al., 2025)	$42.67_{\pm 0.14}$	$41.06_{\pm 0.88}$	$45.21_{\pm 0.11}$	$27.80_{\pm 1.12}$
ConMix (Li & Jia, 2025)	$40.45_{\pm 1.96}$	$38.13_{\pm 0.85}$	$42.76_{\pm 1.40}$	$27.03_{\pm 1.01}$
MiniClustering (Li et al., 2026)	$\mathbf{45.44}_{\pm 1.70}$	$41.53_{\pm 1.82}$	$46.38_{\pm 1.56}$	$28.28_{\pm 1.60}$
SimCLR (Chen et al., 2020)	$43.42_{\pm 1.19}$	$40.70_{\pm 1.33}$	$46.00_{\pm 0.53}$	$28.29_{\pm 1.23}$
BYOL (Grill et al., 2020)	$39.02_{\pm 1.19}$	$37.00_{\pm 2.11}$	$42.15_{\pm 0.90}$	$25.21_{\pm 1.22}$
PACI	$45.25_{\pm 2.43}$	$\mathbf{42.18}_{\pm 1.24}$	$\mathbf{49.11}_{\pm 1.22}$	$\mathbf{33.87}_{\pm 1.60}$

Table 8: STL-10 at IF10.

Method	ACC	CAA	NMI	ARI
SCAN (Van Gansbeke et al., 2020)	$41.97_{\pm 0.36}$	$43.14_{\pm 1.36}$	$47.81_{\pm 0.14}$	$34.73_{\pm 1.45}$
SDCLR (Jiang et al., 2021)	$36.78_{\pm 1.26}$	$38.13_{\pm 1.07}$	$37.75_{\pm 0.30}$	$27.41_{\pm 0.69}$
CC (Li et al., 2021)	$34.37_{\pm 0.50}$	$37.34_{\pm 1.23}$	$34.98_{\pm 0.10}$	$28.46_{\pm 1.03}$
IDFD (Tao et al., 2021)	$34.22_{\pm 0.10}$	$35.38_{\pm 0.47}$	$34.29_{\pm 0.78}$	$26.01_{\pm 0.82}$
CoNR (Yu et al., 2023)	$32.05_{\pm 0.90}$	$31.77_{\pm 0.47}$	$32.12_{\pm 0.92}$	$24.36_{\pm 0.03}$
ProPos (Huang et al., 2023)	$37.67_{\pm 1.42}$	$39.06_{\pm 0.62}$	$37.75_{\pm 0.01}$	$28.63_{\pm 0.28}$
DMICC (Li et al., 2023)	$41.32_{\pm 0.17}$	$41.11_{\pm 1.15}$	$41.41_{\pm 0.04}$	$31.41_{\pm 0.08}$
SeCu (Qian, 2023)	$48.92_{\pm 0.25}$	$45.47_{\pm 0.67}$	$\mathbf{49.01}_{\pm 0.10}$	$37.18_{\pm 0.65}$
LFSS (Li et al., 2025)	$41.72_{\pm 0.07}$	$43.22_{\pm 1.23}$	$41.80_{\pm 1.47}$	$31.71_{\pm 0.95}$
ConMix (Li & Jia, 2025)	$46.02_{\pm 1.04}$	$46.94_{\pm 2.16}$	$47.38_{\pm 0.59}$	$32.54_{\pm 1.68}$
MiniClustering (Li et al., 2026)	$47.07_{\pm 1.92}$	$47.63_{\pm 2.60}$	$46.06_{\pm 1.99}$	$35.53_{\pm 1.34}$
SimCLR (Chen et al., 2020)	$46.55_{\pm 0.96}$	$47.20_{\pm 0.67}$	$46.64_{\pm 1.36}$	$35.38_{\pm 1.20}$
BYOL (Grill et al., 2020)	$41.20_{\pm 0.57}$	$38.78_{\pm 1.18}$	$44.29_{\pm 0.78}$	$29.21_{\pm 1.53}$
PACI	$\mathbf{50.48}_{\pm 0.89}$	$\mathbf{51.29}_{\pm 2.03}$	$47.98_{\pm 1.37}$	$\mathbf{40.04}_{\pm 0.71}$

Table 9: CIFAR-100 at IF100.

Method	ACC	CAA	NMI	ARI
SCAN (Van Gansbeke et al., 2020)	$18.16_{\pm 0.79}$	$11.98_{\pm 0.38}$	$14.94_{\pm 0.36}$	$2.94_{\pm 1.49}$
SDCLR (Jiang et al., 2021)	$23.89_{\pm 0.73}$	$19.50_{\pm 1.43}$	$40.24_{\pm 0.05}$	$17.74_{\pm 1.40}$
CC (Li et al., 2021)	$24.84_{\pm 1.26}$	$18.12_{\pm 0.14}$	$45.74_{\pm 0.54}$	$17.63_{\pm 0.04}$
IDFD (Tao et al., 2021)	$28.10_{\pm 1.17}$	$20.05_{\pm 1.07}$	$44.98_{\pm 0.97}$	$20.74_{\pm 1.39}$
CoNR (Yu et al., 2023)	$22.66_{\pm 0.00}$	$18.50_{\pm 0.57}$	$38.17_{\pm 0.91}$	$16.83_{\pm 1.42}$
ProPos (Huang et al., 2023)	$32.29_{\pm 0.97}$	$25.96_{\pm 0.09}$	$53.26_{\pm 1.02}$	$24.43_{\pm 0.18}$
DMICC (Li et al., 2023)	$23.15_{\pm 0.33}$	$18.90_{\pm 0.79}$	$39.00_{\pm 0.38}$	$17.20_{\pm 0.78}$
SeCu (Qian, 2023)	$28.98_{\pm 1.21}$	$25.08_{\pm 1.14}$	$53.39_{\pm 1.34}$	$21.21_{\pm 1.36}$
LFSS (Li et al., 2025)	$31.21_{\pm 0.06}$	$25.48_{\pm 0.96}$	$53.55_{\pm 1.25}$	$23.59_{\pm 1.50}$
ConMix (Li & Jia, 2025)	$30.31_{\pm 0.46}$	$24.77_{\pm 0.72}$	$51.63_{\pm 0.26}$	$22.59_{\pm 0.34}$
MiniClustering (Li et al., 2026)	$\mathbf{37.59}_{\pm 0.36}$	$25.55_{\pm 0.90}$	$50.69_{\pm 0.93}$	$23.25_{\pm 0.99}$
SimCLR (Chen et al., 2020)	$30.28_{\pm 0.83}$	$25.23_{\pm 0.53}$	$52.06_{\pm 1.49}$	$22.95_{\pm 1.35}$
BYOL (Grill et al., 2020)	$28.67_{\pm 1.17}$	$25.29_{\pm 0.81}$	$51.63_{\pm 0.23}$	$20.66_{\pm 0.39}$
PACI	$33.86_{\pm 1.19}$	$\mathbf{29.60}_{\pm 0.40}$	$\mathbf{54.10}_{\pm 0.50}$	$\mathbf{26.20}_{\pm 0.78}$

Table 10: Tiny-ImageNet at IF100.

Method	ACC	CAA	NMI	ARI
SCAN (Van Gansbeke et al., 2020)	$14.88_{\pm 0.06}$	$12.25_{\pm 1.42}$	$37.41_{\pm 0.25}$	$7.72_{\pm 0.37}$
SDCLR (Jiang et al., 2021)	$17.51_{\pm 0.15}$	$15.20_{\pm 0.57}$	$41.27_{\pm 0.55}$	$11.67_{\pm 1.17}$
CC (Li et al., 2021)	$15.45_{\pm 1.44}$	$14.12_{\pm 0.96}$	$42.78_{\pm 0.89}$	$14.78_{\pm 0.92}$
IDFD (Tao et al., 2021)	$14.88_{\pm 0.17}$	$12.12_{\pm 1.04}$	$35.55_{\pm 0.59}$	$8.23_{\pm 1.46}$
CoNR (Yu et al., 2023)	$16.59_{\pm 0.14}$	$14.40_{\pm 1.23}$	$39.10_{\pm 0.90}$	$11.05_{\pm 0.33}$
ProPos (Huang et al., 2023)	$20.51_{\pm 0.53}$	$15.88_{\pm 1.35}$	$41.87_{\pm 0.89}$	$13.39_{\pm 1.32}$
DMICC (Li et al., 2023)	$17.05_{\pm 1.46}$	$14.80_{\pm 0.77}$	$40.19_{\pm 0.74}$	$11.36_{\pm 0.01}$
SeCu (Qian, 2023)	$18.18_{\pm 0.59}$	$16.44_{\pm 1.38}$	$\mathbf{44.60}_{\pm 1.08}$	$12.62_{\pm 0.41}$
LFSS (Li et al., 2025)	$\mathbf{22.52}_{\pm 1.49}$	$17.12_{\pm 0.57}$	$43.50_{\pm 1.03}$	$\mathbf{15.08}_{\pm 0.62}$
ConMix (Li & Jia, 2025)	$15.10_{\pm 0.10}$	$13.15_{\pm 0.03}$	$36.14_{\pm 0.25}$	$8.12_{\pm 0.22}$
MiniClustering (Li et al., 2026)	$17.15_{\pm 0.44}$	$14.89_{\pm 0.62}$	$40.43_{\pm 0.64}$	$13.27_{\pm 0.36}$
SimCLR (Chen et al., 2020)	$14.57_{\pm 1.04}$	$13.98_{\pm 0.36}$	$35.88_{\pm 0.31}$	$7.02_{\pm 1.39}$
BYOL (Grill et al., 2020)	$17.97_{\pm 0.27}$	$14.07_{\pm 0.43}$	$38.83_{\pm 0.33}$	$11.00_{\pm 0.10}$
PACI	$21.44_{\pm 0.71}$	$\mathbf{18.72}_{\pm 0.33}$	$40.30_{\pm 0.28}$	$13.46_{\pm 0.42}$

D Additional Intervention Results

D.1 Capacity perturbations and basin pairing

Table 11 gives the endpoint intervals and pairing diagnostics supporting Figure 3; its consistency counts refer to interventions on the same 18 final representations.

Table 11: Capacity-path endpoints and geometry-pairing interventions on 18 final representations. Changes are relative to geometry-matched capacity; contamination is in percentage points. Intervals are hierarchical-bootstrap 95% CIs.

Measurement	Mean change	95% CI	Consistency
Coupling displacement	$+0.183$	$[0.154, 0.211]$	18/18 paths
Cross-basin mass	$+0.135$	$[0.101, 0.161]$	18/18 paths
Teacher contamination	$+7.49$	$[4.68, 11.16]$	18/18 paths
Capacity–basin shuffle	$+10.21$	$[7.06, 14.77]$	72/72 conditions
Local capacity perturbation	$+3.73$	$[2.41, 5.74]$	142/144 conditions

D.2 Four-arm interventions

To supplement the conditional training effects in Figure 4, Table 12 pairs each arm's fork-time contamination change with its final CAA loss.

Matched training interventions. The four branches share the epoch-200 model, optimizer and scaler state, training subset, batch-generator rule with seed offset 700001, losses, schedule, and evaluator. EE and UU are the matched-capacity and balanced-capacity branches; UE and EU change the corresponding capacity independently. The controlled panel has 18 paired units and 72 arm–unit outcomes. After branching, representations and their refreshed geometries evolve within each arm. Labels are used for evaluation and offline semantic diagnostics; pseudo-basins and relations are constructed without labels.

Table 12: Four-arm training, relative to PACI (EE), averaged over 18 paired units. Positive CAA loss means lower final accuracy; contamination is measured at the common fork. Both changes are percentage points.

Assignment	Relation normalization	$\Delta C_{\mathrm{batch}}$	CAA loss
Estimated	Estimated (EE)	0	0
Estimated	Uniform (EU)	-0.023	$+0.348$
Uniform	Estimated (UE)	$+1.204$	$+3.632$
Uniform	Uniform (UU)	$+0.875$	$+3.516$

For contamination C, define $A_E = C(UE) - C(EE)$, $A_U = C(UU) - C(EU)$, $R_E = C(EU) - C(EE)$, and $R_U = C(UU) - C(UE)$. Average effects are $A_{\mathrm{avg}} = (A_E + A_U)/2$ and $R_{\mathrm{avg}} = (R_E + R_U)/2$. The interaction is $I = C(UU) - C(EU) - C(UE) + C(EE)$.

For final CAA loss, the analogous interaction is -0.464 points (Figure 4).

Teacher-level effects. On final controlled geometries, $A_E = 10.56$ and $A_U = 7.50$ contamination points; each is positive in 18/18 units. $R_E = -0.007$ points, with an interval crossing zero, and $R_U = -3.07$ points. The interaction is -3.06 points. Thus normalization interacts with assignment, while assignment produces the large positive contamination effect under both normalization choices.

Class-group results. Relative to balanced relations, matched capacity changes head-, medium-, and tail-group accuracy by $+7.199$, -1.018, and $+3.601$ points.

Contamination and final utility. Pooling the three nonreference arms, regression of final CAA loss on fork-time contamination change gives slope 2.736, with a 95% hierarchical-bootstrap interval of $[0.284, 6.158]$. Adding arm fixed effects gives slope 2.501 with interval $[-3.497, 6.499]$. These regressions measure association; the training interventions estimate the effects of capacity choices. A separate stability-filter intervention reduced contamination without improving final PACI accuracy.

D.3 Diagnostics across long-tailed settings

Table 13 expands the CIFAR100-LT summary in Table 2 by imbalance factor, using the same endpoint changes and teacher-level factorial effects defined above.

Table 13: CIFAR100-LT teacher diagnostics by imbalance factor, averaged over three seeds. Contamination and factorial effects are percentage points.

IF	ΔD_{Π}	ΔE_{basin}	ΔC	A_{avg}	R_{avg}
10	0.181	0.078	2.802	3.853	-1.052
50	0.206	0.080	3.208	4.522	-1.315
100	0.193	0.061	2.686	3.961	-1.275