# CoMove: Evolving Supervision for Class-Incremental Learning

**Anonymous authors**
Paper under double-blind review

## Abstract

Class-incremental learning must accommodate new classes while retaining knowledge supported by limited past data. As the classifier expands and replay becomes sparse, both the geometry of learning targets and the evidence supporting them change. We introduce CoMove, a method that adapts the geometry and allocation of supervision throughout learning. For progressive equiangular tight frame (ETF) classifiers, we characterize the displacement of inherited prototypes and derive a spherical transport that aligns teacher references with the expanded geometry while preserving rankings among old classes. Peer, prototype, and transported-teacher relations are integrated through a common objective with shared negative competition and source-specific responses. Class composition and replay support determine their contributions, transferring under-supported supervision to the teacher while preserving total query weight. A single allocation rule applies across datasets and replay budgets, without tuning source weights; inference uses the ETF directly. Experiments on CIFAR-10, CIFAR-100, and TinyImageNet demonstrate consistent improvements over competing methods across low-memory settings. CoMove achieves the highest mean final class-incremental accuracy in all six dataset–buffer settings, with gains over the best baseline in each setting averaging 4.69 percentage points, alongside substantially reduced forgetting. The same incremental objective also supports learning without exemplar replay. Our code is provided in the supplementary material.

## 1 Introduction

Class-incremental learning asks a model to recognize an expanding set of classes while retaining access to only a small fraction of past data. Labels, examples, and a previous model provide complementary supervision: class labels specify the required decision, examples describe within-class variation, and the previous model supplies information about earlier tasks. Replay and distillation exploit these sources in different ways (Rebuffi et al., 2017; Buzzega et al., 2020; Cha et al., 2021). As learning proceeds, however, both the geometry of their targets and the amount of data supporting them change.

Progressive equiangular tight frame (ETF) expansion moves the target directions of inherited classes, while the frozen teacher still represents them in the preceding geometry. At the same time, a fixed replay budget must support an increasing number of classes. Reweighting can compensate for class imbalance, but cannot restore missing examples or reconcile different target directions. Continual supervision must therefore account for both the movement of its references and the evidence available to support them.

We introduce CoMove, a class-incremental learning method that adapts both the references used for training and their contributions to the objective (Figure 1). We use a progressively expanding ETF to represent the seen classes (Wang et al., 2026). An exact characterization of its inherited vertices yields a common contraction and displacement. We extend this update to frozen teacher features through a spherical transport, aligning the inherited prototypes while preserving rankings among old labels.

The objective places peer, prototype, and transported-teacher references against the same negative competition while retaining their distinct positive scores. Class composition sets their base weights; replay support determines how much prototype and peer supervision is retained or transferred to the



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.

teacher. This reallocation conserves the total source weight. Even when replay offers little support, each query retains its full supervision budget.

On CIFAR-10, CIFAR-100, and TinyImageNet, CoMove achieves the highest mean final Class-IL accuracy in all six main comparisons at replay budgets 200 and 500. The gain over the best baseline in each setting averages 4.69 percentage points and reaches 6.36 points. On TinyImageNet with 200 exemplars, final forgetting is 7.40 points, compared with 31.06 for the lowest-forgetting baseline. All our scores come from the expanding ETF classifier; they include no post-hoc classifier fitting.

**Contributions.** **(i) Geometric characterization.** We derive the symmetric affine lift of inherited ETF vertices and establish exact prototype alignment and old-label ranking preservation for its spherical continuation to teacher features. **(ii) Evolving supervision.** We introduce *CoMove*, a learning objective coupling prototype, peer, and temporal relations through shared negative competition and a support-dependent transfer of supervision. **(iii) Empirical evaluation.** We establish gains across the main low-memory comparisons and evaluate the same incremental formulation without replay. Paired ablations distinguish the effects of supervision allocation, source responses, and reference transport on acquisition and retention.

## 2 Related Work

**Continual learning objectives.** Methods retain past knowledge through parameter or output regularization (Kirkpatrick et al., 2017; Zenke et al., 2017; Li & Hoiem, 2018), replay (Rebuffi et al., 2017; Riemer et al., 2019; Rolnick et al., 2019), or constraints on updates (Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2019). DER++ combines replay and logit matching (Buzzega et al., 2020), LODE separates discrimination among new classes from discrimination between old and new classes (Liang & Li, 2023), and IDER enforces idempotence across current and previous models (Liu et al., 2026).

**Contrastive supervision.** Supervised contrastive learning uses same-class positives (Khosla et al., 2020), with alignment and uniformity providing a geometric interpretation (Wang & Isola, 2020). $Co^2L$ preserves instance relations through distillation (Cha et al., 2021). CCLIS combines prototype–instance distillation with importance sampling and buffer selection to approximate past data distributions (Li et al., 2024). Replay construction (Boschini et al., 2023; Tong et al., 2025) and adaptive contrastive learning (Wen et al., 2024) address related aspects of limited memory. Our allocation operates on the available training set: it transfers weight between supervision sources while leaving buffer selection unchanged.

**Neural collapse and classifier geometry.** Neural collapse connects concentrated class features with simplex geometry (Papyan et al., 2020); subsequent analyses examine its dynamics and behavior under imbalance (Han et al., 2022; Fang et al., 2021; Xie et al., 2023). Fixed ETF classifiers and dot regression address class imbalance (Yang et al., 2022) and incremental learning (Yang et al., 2023a;b); adaptive fitting instead finds the ETF nearest to current features (Markou et al., 2024). NC also motivates contrastive objectives (Montmaur et al., 2024; Dang et al., 2025). SAMix interpolates fixed prototypes on the sphere to supervise mixed examples (Dang et al., 2026); our map instead updates teacher references after classifier expansion. ProNC expands a nested ETF as classes arrive and supports different alignment losses (Wang et al., 2026). We build on this construction, characterize its inherited-vertex displacement exactly, and use it within a different learning objective and source-allocation rule.

**Feature drift and inference.** Distillation can preserve outputs or intermediate representations (Hou et al., 2019; Douillard et al., 2020; Fini et al., 2022). LDC learns a map for old prototypes (Gomez-Villa et al., 2024); BiCyc learns bidirectional maps with cycle consistency to transport class statistics (Xu & Krawczyk, 2026). Our analytic transport instead follows ETF expansion and acts on teacher references during training. LCA aligns task classifiers with a consolidated pretrained backbone (Tran et al., 2026), addressing drift in the learned representation. FeCAM models class-specific feature covariances for inference (Goswami et al., 2023). Feature preservation and classifier performance need not coincide (Lanzillotta et al., 2026); we therefore distinguish native ETF results from post-hoc compositions.

# 3 PRELIMINARIES

**Class-incremental learning.** At stage $t$, the learner receives a dataset $\mathcal{D}_t$ with labels $\mathcal{Y}_t$ disjoint from earlier tasks, and a replay set $\mathcal{M}_{t-1}$ of at most $M$ examples (Masana et al., 2023). Training uses $\widetilde{\mathcal{D}}_t = \mathcal{D}_t \cup \mathcal{M}_{t-1}$. The seen labels are $\mathcal{C}_t = \bigcup_{s\le t} \mathcal{Y}_s$. For equal task sizes $q = |\mathcal{Y}_t| \ge 2$, incremental stages have $O = (t-1)q$ old and $K = tq = O + q$ seen classes. Labels follow arrival order: $c \le O$ is old and $g(c) = \lfloor (c-1)/q \rfloor$ denotes its zero-based task-block index. Class-IL predicts over all seen labels without task identity. Task-IL restricts the same scores to the test sample's true block.

**Representations and classifier.** With encoder $f_t$ and linear projector $\boldsymbol{W}_t$, define the student feature and, at incremental stages $t \ge 2$, the frozen teacher feature:

$$\boldsymbol{z}_i = \nu(\boldsymbol{W}_t f_t(x_i)), \qquad \boldsymbol{a}_i = \mathrm{sg}[\nu(\boldsymbol{W}_{t-1} f_{t-1}(x_i))], \qquad \nu(\boldsymbol{x}) = \boldsymbol{x}/\|\boldsymbol{x}\|. \tag{1}$$

Here $d$ is the projected feature dimension and $\mathbb{S}^{d-1} = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| = 1\}$. Unqualified norms are Euclidean; normalization applies to nonzero representations and sg stops gradients. We suppress the current stage in $\boldsymbol{z}_i$ and $\boldsymbol{a}_i$, and write $\boldsymbol{z}(x)$ and $\boldsymbol{a}(x)$ for their evaluations at input $x$. The teacher receives the same augmented input as the student, including new-class inputs. The classifier consists of unit simplex ETF vertices $\boldsymbol{U}_K = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_K]$ satisfying

$$\boldsymbol{u}_c^\top \boldsymbol{u}_\ell = -\frac{1}{K-1}\ (c \ne \ell), \qquad \sum_{c=1}^{K} \boldsymbol{u}_c = \boldsymbol{0}, \qquad \widehat{y}(x) = \operatorname*{arg\,max}_{c \le K} \boldsymbol{z}(x)^\top \boldsymbol{u}_c. \tag{2}$$

Prediction uses a fixed tie-breaking rule when several labels maximize the score. ETF targets are motivated by neural collapse (Papyan et al., 2020); for completeness, Appendix D.1.1 recalls NC1–NC4. We use a fixed initial frame and its nested expansion (Wang et al., 2026), with $d \ge K$ for this basis construction (an abstract simplex only requires $d \ge K-1$). At a task transition, write $\boldsymbol{v}_c$ for the previous vertex and $\boldsymbol{u}_c$ for its expanded location. Keeping the inherited basis columns does not keep these vertices fixed; the next section describes their displacement and the corresponding temporal reference.

# 4 COMOVE: EVOLVING SUPERVISION

CoMove combines same-class peers, current class prototypes, and a transported teacher. We construct the temporal reference in Section 4.1, the common objective in Section 4.2, and its support-dependent weights in Section 4.3.

## 4.1 TRANSPORTING TEMPORAL REFERENCES

**From a symmetric lift to the task endpoint.** A task introduces $q$ classes. We express its ETF expansion as a chain of single-class lifts, then compute their combined effect on inherited vertices.

**Definition 1** (Symmetric lift)**.** For an $m$-class simplex ETF, $m \ge 2$, and a unit vector $\boldsymbol{n}_m$ orthogonal to its span, define

$$\boldsymbol{v}_c^{(m+1)} = a_m \boldsymbol{v}_c^{(m)} - \boldsymbol{n}_m/m \quad (c \le m), \qquad \boldsymbol{v}_{m+1}^{(m+1)} = \boldsymbol{n}_m, \qquad a_m = \sqrt{1 - m^{-2}}. \tag{3}$$

The old–new inner product fixes the shift $-\boldsymbol{n}_m/m$; unit norm fixes the positive contraction $a_m$. Old–old products then equal $-1/m$, so the result is an $(m+1)$-class ETF. The lift is symmetric *among inherited labels*: permuting them commutes with their common update. It is the vertex-space form of nested basis expansion (Appendix D.1.2).

**Theorem 1** (Inherited-vertex state)**.** *Chaining the symmetric lifts from $2 \le O < K$ yields*

$$\boxed{\boldsymbol{u}_c = \alpha \boldsymbol{v}_c + \boldsymbol{b} \quad (c \le O)}, \qquad \alpha = \sqrt{\frac{K(O-1)}{O(K-1)}}, \qquad \boldsymbol{b} = \frac{1}{O}\sum_{c \le O} \boldsymbol{u}_c, \tag{4}$$

*where $\boldsymbol{b} \perp \mathrm{span}\{\boldsymbol{v}_c\}_{c \le O}$ and $\|\boldsymbol{b}\|^2 = 1 - \alpha^2$.*

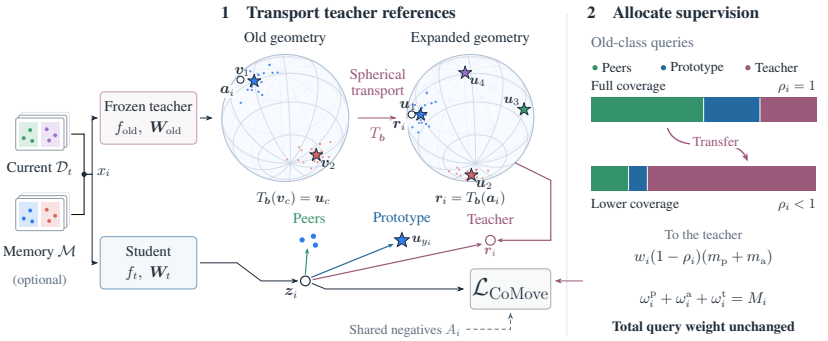


Figure 1: **CoMove: evolving supervision.** Spherical transport aligns teacher references with the expanded ETF geometry. Reduced replay coverage transfers peer and prototype weight to the teacher, preserving total query weight. Geometry and allocation are schematic.

The squared contractions telescope, and averaging over the centered old simplex gives the displacement (Section D.2). Computing the transition therefore requires only the class counts and the centroid of the inherited vertices in the expanded ETF. These determine $\alpha$ and $\boldsymbol{b}$ once per task, without fitting a map to features.

**From prototype displacement to a teacher reference.** Even an ideally collapsed teacher returns $\boldsymbol{v}_c$, whereas current alignment targets $\boldsymbol{u}_c$, with $\|\boldsymbol{u}_c - \boldsymbol{v}_c\|^2 = 2(1-\alpha) > 0$. Losses uniquely minimized at their reference therefore have different minimizers. Reweighting those losses cannot make their targets coincide (Proposition 3). The affine continuation $F(\boldsymbol{x}) = \alpha\boldsymbol{x} + \boldsymbol{b}$ matches prototypes exactly. Its output norm depends on the feature direction (Section D.3), motivating a spherical continuation that also preserves local angles.

**Definition 2** (Möbius continuation)**.** For $\|\boldsymbol{b}\| < 1$ and $\alpha = \sqrt{1 - \|\boldsymbol{b}\|^2}$, define on $\mathbb{S}^{d-1}$

$$\boxed{T_{\boldsymbol{b}}(\boldsymbol{x}) = \frac{\alpha\boldsymbol{x} + \boldsymbol{b} + \frac{\boldsymbol{b}^\top \boldsymbol{x}}{1+\alpha}\boldsymbol{b}}{1 + \boldsymbol{b}^\top \boldsymbol{x}}}, \qquad \boldsymbol{r}_i = \mathrm{sg}[T_{\boldsymbol{b}}(\boldsymbol{a}_i)]. \tag{5}$$

The denominator is positive on the unit sphere. The map has unit outputs and inverse $T_{-\boldsymbol{b}}$ (Section D.4). It is applied once between task endpoints, to teacher features on both old and new inputs. For new classes, the transported teacher features provide a representation reference. The frozen teacher has no classifier for these classes; classification uses the current ETF. Appendix C.6 illustrates the construction on a real two-task checkpoint.

**Theorem 2** (Prototype alignment and old-label rankings)**.** *Under Equation* (4)*,* $T_{\boldsymbol{b}}(\boldsymbol{v}_c) = \boldsymbol{u}_c$ *for all* $c \le O$*. For any unit* $\boldsymbol{x}$ *and old labels* $c, \ell$*,*

$$T_{\boldsymbol{b}}(\boldsymbol{x})^\top(\boldsymbol{u}_c - \boldsymbol{u}_\ell) = \frac{\alpha^2}{1 + \boldsymbol{b}^\top \boldsymbol{x}} \boldsymbol{x}^\top(\boldsymbol{v}_c - \boldsymbol{v}_\ell). \tag{6}$$

For a frozen teacher feature, this gives $\arg\max_{c \le O} \boldsymbol{a}_i^\top \boldsymbol{v}_c = \arg\max_{c \le O} \boldsymbol{r}_i^\top \boldsymbol{u}_c$. Thus transport preserves its old-class prediction, including ties (Section D.5). The training objective still combines these references with current examples and prototypes.

### 4.2 A Common Learning Objective

Peers supply current examples, prototypes specify class directions, and the transported teacher supplies a temporal reference. We keep their positive scores separate but place them against the same negative partition. Each source can then have its own attraction response without changing the competing classes. For query $i$, the peer, prototype, and teacher scores use a fixed scale $\gamma > 0$:

$$s_{ij}^{\mathrm{p}} = \gamma \boldsymbol{z}_i^\top \boldsymbol{z}_j, \qquad s_i^{\mathrm{a}} = \gamma \boldsymbol{z}_i^\top \boldsymbol{u}_{y_i}, \qquad s_i^{\mathrm{t}} = \gamma \boldsymbol{z}_i^\top \boldsymbol{r}_i. \tag{7}$$

**Shared negative competition.** For a minibatch indexed by $i, j \in \{1, \dots, B\}$, define the same-class set $\mathcal{B}_i = \{j : y_j = y_i\}$, positive peers $\mathcal{P}_i = \mathcal{B}_i \setminus \{i\}$, and empirical negatives $\mathcal{N}_i = \{j : y_j \neq y_i,\ g(y_j) = g(y_i)\}$. Every seen prototype competes with the true class; empirical negatives come from different classes within the query's task block:

$$P_i = \sum_{c \in \mathcal{C}_t \setminus \{y_i\}} e^{\gamma \boldsymbol{z}_i^\top \boldsymbol{u}_c}, \qquad E_i = \sum_{j \in \mathcal{N}_i} e^{\gamma \boldsymbol{z}_i^\top \boldsymbol{z}_j}. \tag{8}$$

Empirical negatives distinguish classes within each task block, while prototype negatives provide competition across all seen classes, including those absent from the batch. Set $\xi_i = E_i / P_i$. The unpooled partition is $\log(P_i + E_i) = \log P_i + \log(1 + \xi_i)$. Pooling averages $\xi_j$ across same-class queries before applying the logarithm:

$$A_i = \log P_i + \mu_{y_i}, \qquad \mu_{y_i} = \log \left( 1 + \frac{1}{|\mathcal{B}_i|} \sum_{j \in \mathcal{B}_i} \xi_j \right). \tag{9}$$

Only $\mu_c$ is shared; $P_i$ remains query-specific. For $|\mathcal{B}_i| = 1$ this recovers the unpooled partition, and without empirical negatives the correction vanishes. Gradients pass through the mean, coupling each query to its same-class neighbors' ratios (Sections D.7 and D.8).

**Source-dependent loss responses.** The function $\psi_i$ evaluates a positive score $s$ against query $i$'s shared negative partition $A_i$:

$$\psi_i(s; \kappa, \lambda) = \log \left[ \kappa e^{-\lambda(s - s_i^0)} + e^{A_i - s} \right], \qquad s_i^0 = \gamma \boldsymbol{a}_i^\top \boldsymbol{r}_i. \tag{10}$$

The null mass $\kappa \geq 0$ controls saturation, while the rate $\lambda \geq 0$ controls the score dependence of the null branch. The case $(1, 0)$ is the logistic margin loss; $\kappa = 0$ gives the affine margin $A_i - s$. The frozen score $s_i^0$ compares the original and transported teacher. For incremental tasks, the objective is $\mathcal{L}_{\text{CoMove}} = B^{-1} \sum_{i=1}^{B} \ell_i$. For $|\mathcal{P}_i| > 0$, the per-query objective is

$$\boxed{\ell_i = \underbrace{\frac{\omega_i^{\text{p}}}{|\mathcal{P}_i|} \sum_{j \in \mathcal{P}_i} \psi_i(s_{ij}^{\text{p}}; 1, 0)}_{\text{peers}} + \underbrace{\omega_i^{\text{a}} \psi_i(s_i^{\text{a}}; \kappa_i, 0)}_{\text{prototype}} + \underbrace{\omega_i^{\text{t}} \psi_i(s_i^{\text{t}}; \kappa_i, \lambda_i)}_{\text{teacher}}}. \tag{11}$$

Queries without positive peers use a prototype fallback detailed in Appendix A.2. The weights in Equation (13) allocate supervision across these terms; $\kappa_i$ and $\lambda_i$ determine their responses.

**Computed source responses.** We compute the prototype and teacher responses from an entropy deficit; actual peers retain $(\kappa, \lambda) = (1, 0)$. Let $\delta_i = \text{sg}([1 - h_i(\boldsymbol{z}_i) - h_i(\boldsymbol{r}_i)]_+)$. The response parameters are

$$(\kappa_i, \lambda_i) = \left( 1 - \mathbf{1}_{\{y_i > O\}} \delta_i,\ \mathbf{1}_{\{y_i \leq O\}} \frac{O}{q} \delta_i \right). \tag{12}$$

Here $h_i$ is normalized entropy of prototype scores over the wrong classes in the query's task block, clipped to $[0, 1]$; for binary tasks, $h_i = 0$. These values are recomputed but detached in backward. The policy controls the shape of the loss response as well as its weight: at rate one, positive-score attraction is constant while negative demand can still vary (Section D.9). Equation (22) defines $h_i$ explicitly; Table 6 lists the responses by source and query type.

### 4.3 Support-Dependent Allocation

The responses above determine how each reference acts on a query. We now set how much weight each source receives, first from class composition and then from the number of available old-class examples.

**Class-count budget.** We assign the new-class share $q/K$ to peers and the inherited share $O/K$ to prototype and teacher supervision. Treating these two inherited sources symmetrically gives $m_{\text{p}} = q/K$ and $m_{\text{a}} = m_{\text{t}} = O/(2K)$ (Figure 2). Appendix A.3 derives this equal split from invariance under exchanging the sources and conservation of the inherited budget.

**From class frequency to support.** This base budget does not account for the examples supporting each class. Let $n_c$ count class-$c$ examples in the merged training dataset and $N = \sum_c n_c$. The frequency weight $w_i = N/(Kn_{y_i})$ balances class contributions. Reweighting compensates for imbalance, but does not increase the number of available examples. We use a count-based replay coverage rule, $\rho_i = \min\{1, (n_{y_i} - 1)/(K - 1)\}$ for old classes, and $\rho_i = 1$ for new classes. Thus $n_{y_i} = 1$ gives zero coverage for an old class, while $K$ observations give full coverage.

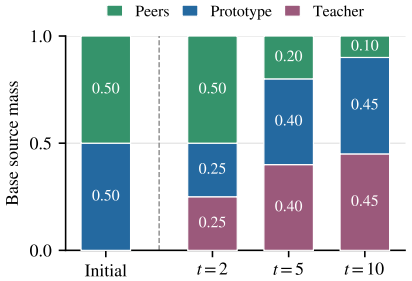


Figure 2: **Base source allocation.** Equal-sized tasks, before coverage transfer.

**Coverage transfer.** Starting from the base masses in Figure 2, we retain a fraction $\rho_i$ of the frequency-weighted prototype and peer supervision and transfer the remainder to the teacher. The resulting per-query weights are

$$\omega_i^{\mathrm{a}} = w_i \rho_i m_{\mathrm{a}}, \quad \omega_i^{\mathrm{p}} = w_i \rho_i m_{\mathrm{p}}, \quad \omega_i^{\mathrm{t}} = m_{\mathrm{t}} + w_i(1 - \rho_i)(m_{\mathrm{a}} + m_{\mathrm{p}}). \tag{13}$$

The peer term averages its aggregate weight over the available positives, or assigns it to the prototype fallback. The three source weights sum to $M_i = w_i(m_{\mathrm{a}} + m_{\mathrm{p}}) + m_{\mathrm{t}}$, independent of $\rho_i$. **Sparse support reallocates supervision at constant total query weight.** With $K = 20$ and three old-class examples, $\rho_i = 2/19$: this fraction of the frequency-weighted prototype and peer budget remains, and the rest goes to the teacher. Setting $\rho_i = 1$ gives a no-transfer control at the same total weight. Appendix A.4 gives the developed coverage form and exact correction; Section D.6 establishes the equivalence to Equation (11).

With $M = 0$, only current-class queries are observed, so $\rho_i = 1$ and coverage transfer is inactive. The transported teacher still supervises current inputs through its base weight $m_{\mathrm{t}}$, allowing the incremental objective to operate without exemplar replay.

# 5 Experiments

## 5.1 Experimental Setup

**CL benchmarks and evaluation metrics.** Following standard continual-learning protocols (Yang et al., 2023b; Wang et al., 2026), we evaluate on Seq-CIFAR-10/100 (Krizhevsky, 2009) and Seq-TinyImageNet (Le & Yang, 2015), partitioned into five, ten, and ten tasks, respectively. Each task introduces two, ten, or twenty classes. We report final average accuracy (FAA) and final forgetting (FF) under Class-IL and Task-IL, with replay budgets of 200 and 500 examples. We also evaluate the exemplar-free setting ($M = 0$). Appendices B.1 and B.2 give dataset and metric definitions.

**Baselines.** We compare with replay-based methods ER (Riemer et al., 2019), iCaRL (Rebuffi et al., 2017), DER/DER++ (Buzzega et al., 2020), LODE (Liang & Li, 2023), STAR (Eskandar et al., 2025), and CSReL (Tong et al., 2025); contrastive methods Co$^2$L (Cha et al., 2021), CILA (Wen et al., 2024), and FC-NCCL (Dang et al., 2025); and neural-collapse methods NCT (Yang et al., 2023b) and ProNC (Wang et al., 2026). SAMix (Dang et al., 2026) appears in the separate contrastive comparison. Exemplar-free baselines include LwF (Li & Hoiem, 2018), EFC (Magistri et al., 2024), and EFC++ (Magistri et al., 2025). Baseline results are taken from published reports or obtained with the implementations detailed in Appendix B.3.

**Implementation details.** We use ResNet-18 (He et al., 2016) with a 512-dimensional linear projector, batch size 32, and 50/100 epochs per task on CIFAR/TinyImageNet. **The same source-allocation formulas apply across datasets and memory budgets**, without configuration-specific tuning of source weights. We report means and standard deviations. Appendix B.1 gives the hyperparameters; Appendices B.2 and B.4 specify repeated runs and paired controls.

Table 1: **Main comparison at replay budgets 200 and 500.** FAA (↑) and FF (↓); bold and underline mark the best and second-best distinct means.

| Method | Class-IL | | | | Task-IL | | | |
|---|---|---|---|---|---|---|---|---|
| | $M = 200$ | | $M = 500$ | | $M = 200$ | | $M = 500$ | |
| | FAA↑ | FF↓ | FAA↑ | FF↓ | FAA↑ | FF↓ | FAA↑ | FF↓ |
| **(a) Seq-CIFAR-10** | | | | | | | | |
| ER (Riemer et al., 2019) | 44.79±1.86 | 59.30±2.48 | 57.74±2.48 | 43.22±2.10 | 91.19±0.94 | 6.07±1.09 | 93.61±0.27 | 3.50±0.53 |
| iCaRL (Rebuffi et al., 2017) | 49.02±3.20 | 23.52±1.27 | 47.55±3.95 | 28.20±2.41 | 88.99±2.13 | 25.34±1.64 | 88.22±2.62 | 22.61±3.97 |
| DER (Buzzega et al., 2020) | 61.93±1.79 | 35.79±2.59 | 70.51±1.67 | 24.02±1.63 | 91.40±0.92 | 6.08±0.70 | 93.40±0.21 | 3.72±0.55 |
| DER++ (Buzzega et al., 2020) | 64.88±1.17 | 32.59±2.32 | 72.70±1.36 | 22.38±4.41 | 91.92±0.60 | 5.16±0.21 | 93.88±0.50 | 4.66±1.15 |
| LODE (Liang & Li, 2023) | 68.01 | 24.63 | 75.91 | 18.18 | 93.11 | 4.75 | 94.19 | 3.94 |
| Co$^2$L (Cha et al., 2021) | 51.27±1.86 | 30.17±8.57 | 61.78±4.22 | 17.79±3.36 | 84.69±1.52 | 2.91±5.25 | 89.51±2.45 | 2.65±1.00 |
| CILA (Wen et al., 2024) | 59.68±0.65 | 37.52±5.84 | 67.82±0.33 | 18.22±1.32 | 91.36±0.08 | 5.49±0.74 | 93.29±0.24 | 0.65±0.13 |
| FC-NCCL (Dang et al., 2025) | 52.20±1.56 | 33.74±1.65 | 52.20±1.56 | 27.88±2.75 | 85.94±0.22 | 4.90±0.15 | 85.94±0.22 | 3.15±0.42 |
| STAR (Eskandar et al., 2025) | 62.10±2.21 | **21.78**±3.16 | 69.15±3.53 | 18.59±2.91 | 93.54±1.48 | 6.23±2.06 | 95.36±0.37 | 2.38±0.34 |
| CSReL (Tong et al., 2025) | 37.46±1.57 | 26.34±2.18 | 40.82±4.09 | 29.18±6.91 | 69.22±3.03 | 17.16±2.91 | 73.30±5.92 | 16.89±3.19 |
| NCT (Yang et al., 2023b) | 51.59±0.41 | 22.48±19.50 | 60.93±0.94 | **13.82**±3.64 | 80.63±0.46 | 1.41±1.09 | 81.27±0.24 | 3.84±0.35 |
| ProNC (Wang et al., 2026) | 65.58±0.15 | 32.75±4.71 | 73.95±0.68 | 26.75±1.29 | 96.86±0.10 | 0.65±0.08 | 96.95±0.14 | **0.24**±0.14 |
| **Ours** | **73.55**±0.92 | 24.12±0.58 | **78.00**±0.28 | 18.29±0.24 | **96.96**±0.18 | **0.60**±0.18 | **97.15**±0.03 | 0.38±0.08 |
| **(b) Seq-CIFAR-100** | | | | | | | | |
| ER (Riemer et al., 2019) | 21.78±0.48 | 75.06±0.63 | 22.35±0.61 | 73.08±0.78 | 60.19±1.01 | 27.38±1.46 | 73.98±1.52 | 16.23±1.06 |
| iCaRL (Rebuffi et al., 2017) | 28.00±0.91 | 47.20±1.23 | 33.25±1.25 | 40.99±1.02 | 51.43±1.47 | 36.20±1.85 | 58.16±1.76 | 27.90±1.37 |
| DER (Buzzega et al., 2020) | 31.23±1.38 | 62.72±2.69 | 41.36±1.76 | 49.07±2.54 | 63.09±1.09 | 25.98±1.55 | 71.73±0.74 | 25.98±1.55 |
| DER++ (Buzzega et al., 2020) | 28.13±0.51 | 60.99±1.52 | 38.20±1.00 | 49.18±2.19 | 66.80±0.41 | 23.91±0.55 | 74.77±0.31 | 15.75±0.48 |
| LODE (Liang & Li, 2023) | 26.65 | 44.29 | 40.01 | 32.58 | 71.23 | 18.75 | 80.06 | 8.96 |
| Co$^2$L (Cha et al., 2021) | 18.09±0.49 | 64.14±0.69 | 26.64±1.42 | 48.60±1.34 | 49.19±0.91 | 36.81±0.63 | 62.32±0.19 | 23.47±0.27 |
| CILA (Wen et al., 2024) | 19.49±0.53 | 64.01±0.18 | 31.27±0.17 | 45.67±1.02 | 53.93±1.02 | 33.07±0.96 | 68.29±0.46 | 16.98±0.56 |
| FC-NCCL (Dang et al., 2025) | 15.81±0.48 | 62.51±0.40 | 22.29±0.18 | 46.09±0.58 | 43.91±0.76 | 39.79±0.98 | 56.43±0.29 | 24.29±0.50 |
| STAR (Eskandar et al., 2025) | 18.29±2.58 | 68.40±8.52 | 28.45±1.70 | 53.11±1.57 | 58.53±13.06 | 27.84±8.01 | 74.06±1.63 | 15.22±0.96 |
| CSReL (Tong et al., 2025) | 29.06±1.00 | 58.23±1.54 | 39.22±1.70 | 47.03±2.57 | 66.99±0.35 | 23.20±1.74 | 75.09±0.78 | 15.55±0.76 |
| NCT (Yang et al., 2023b) | 26.38±0.57 | 27.40±1.65 | 33.84±0.38 | 24.37±5.42 | 75.75±0.17 | 4.79±0.07 | 76.06±0.52 | 3.97±0.75 |
| ProNC (Wang et al., 2026) | 42.99±0.85 | 36.07±0.51 | 48.94±0.44 | 24.32±0.24 | 85.63±0.73 | 4.40±0.82 | **86.38**±0.43 | 4.35±0.17 |
| **Ours** | **49.02**±0.51 | **19.69**±0.60 | **50.92**±0.54 | **13.94**±0.47 | **85.83**±0.38 | **1.85**±0.23 | 85.86±0.59 | **1.84**±0.11 |
| **(c) Seq-TinyImageNet** | | | | | | | | |
| ER (Riemer et al., 2019) | 8.49±0.16 | 76.53±0.51 | 9.99±0.29 | 75.21±0.54 | 38.17±2.00 | 40.47±1.54 | 48.64±0.46 | 30.73±0.62 |
| iCaRL (Rebuffi et al., 2017) | 7.53±0.79 | 31.06±1.91 | 9.38±1.53 | 37.30±1.42 | 28.19±1.47 | 42.47±2.47 | 31.55±3.27 | 39.44±0.84 |
| DER (Buzzega et al., 2020) | 11.87±0.78 | 64.83±1.48 | 17.75±1.14 | 59.95±2.31 | 40.22±0.67 | 40.43±1.05 | 51.78±0.88 | 28.21±0.97 |
| DER++ (Buzzega et al., 2020) | 11.34±1.17 | 73.47±1.23 | 19.38±1.41 | 58.75±1.93 | 43.06±1.16 | 39.02±0.43 | 51.91±0.68 | 25.47±1.03 |
| LODE (Liang & Li, 2023) | 15.13 | 64.00 | 20.50 | 56.51 | 51.42 | 29.66 | 61.49 | 18.61 |
| Co$^2$L (Cha et al., 2021) | 12.95±0.06 | 62.04±0.28 | 18.71±0.84 | 49.64±1.14 | 37.07±1.62 | 40.75±3.55 | 50.74±1.24 | 13.80±2.10 |
| CILA (Wen et al., 2024) | 12.98±0.10 | 63.11±0.61 | 18.09±0.49 | 64.14±0.69 | 37.32±1.87 | 41.40±0.51 | 49.19±0.91 | 27.83±1.32 |
| FC-NCCL (Dang et al., 2025) | 10.57±1.66 | 59.68±2.02 | 11.52±0.01 | 44.96±0.09 | 32.78±3.54 | 45.10±1.57 | 36.32±0.05 | 39.00±0.17 |
| STAR (Eskandar et al., 2025) | 11.55±3.23 | 67.43±1.48 | 15.19±2.61 | 63.42±3.19 | 41.70±0.95 | 34.78±2.08 | 55.06±2.07 | 15.19±2.61 |
| CSReL (Tong et al., 2025) | 18.14±3.10 | 49.77±2.27 | 22.13±0.35 | 50.01±1.87 | 45.04±5.86 | 34.12±2.18 | 51.94±0.22 | 28.20±0.36 |
| NCT (Yang et al., 2023b) | 10.95±1.45 | 49.33±4.47 | 18.24±0.62 | 50.90±1.16 | 52.71±4.12 | 15.88±0.95 | 62.30±0.41 | 6.84±0.87 |
| ProNC (Wang et al., 2026) | 27.44±1.00 | 42.81±0.56 | 29.06±0.32 | 38.58±0.39 | 69.09±0.65 | 9.42±0.58 | 69.77±0.89 | 9.52±0.96 |
| **Ours** | **33.80**±0.12 | **7.40**±0.25 | **35.22**±0.35 | **5.47**±0.19 | **72.12**±0.06 | **1.64**±0.28 | **72.41**±0.45 | **1.30**±0.47 |

### 5.2 Main Results

**Consistent gains across all six settings.** For a fair comparison, Table 1 follows ProNC (Wang et al., 2026): methods share a ResNet-18 backbone, batch size, and number of representation-training epochs, while retaining their own heads and objectives. CoMove exceeds every baseline in Class-IL FAA at both budgets: **4.69 points on average**, up to 6.36 points. With 200 examples, the gains over the best baseline are 5.54, 6.03, and 6.36 points on CIFAR-10, CIFAR-100, and TinyImageNet. Task-IL FAA is also highest in five of six settings.

**Stronger accuracy with less replay.** On TinyImageNet, 200 examples suffice to exceed the best baseline using 500 (33.80% versus 29.06%): 4.74 points higher with **60% fewer replay examples**. CoMove also achieves the *lowest Class-IL FF* on CIFAR-100 and TinyImageNet at both budgets. TinyImageNet FF is 7.40/5.47, versus the next-lowest 31.06/37.30 at buffers 200/500.

**Comparison with contrastive methods.** Table 2 complements this comparison with Co$^2$L, FC-NCCL, and SAMix under their published protocols, including longer initial training and separate classifier fitting. Here, zero buffer applies to representation learning; classifier fitting still uses 200 auxiliary examples (Dang et al., 2026). CoMove stores none when $M = 0$. SAMix combines sphere-adaptive mixup with hardness-softness distillation. With *neither mixed samples nor post-hoc classifier training*, CoMove exceeds SAMix in Class-IL FAA in **four of six replay settings**: all three datasets at buffer 200 and TinyImageNet at buffer 500. The TinyImageNet margins are 15.41/12.27 points. SAMix remains higher on CIFAR at buffer 500. Batch sizes and the CIFAR-100 task split also differ (Appendix B).

Table 2: **Published contrastive comparisons.** FAA (↑) and FF (↓), with available deviations. Bold and underline mark the best and second-best distinct means. *Source protocols differ in task splits, training, or classifier fitting (Section 5.2); numerical ranks are not matched-protocol comparisons.

| Method | CIFAR-10 | | | | CIFAR-100 | | | | TinyImageNet | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Class-IL | | Task-IL | | Class-IL | | Task-IL | | Class-IL | | Task-IL | |
| | FAA↑ | FF↓ | FAA↑ | FF↓ | FAA↑ | FF↓ | FAA↑ | FF↓ | FAA↑ | FF↓ | FAA↑ | FF↓ |
| **Replay buffer: 0** | | | | | | | | | | | | |
| Co$^2$L* | 58.89±2.61 | 35.81±1.08 | 86.65±1.05 | 14.33±0.87 | 26.89±0.78 | 66.51±0.28 | 51.91±0.63 | 39.63±0.62 | 13.43±0.57 | 62.80±0.77 | 40.21±0.68 | 39.54±1.08 |
| FC-NCCL* | 69.26±0.32 | 23.85±0.30 | 94.41±0.43 | 4.72±0.28 | 32.57±0.55 | 52.03±0.63 | 57.87±0.62 | 36.20±0.48 | 14.54±0.52 | 53.97±0.63 | 43.81±0.47 | 37.57±0.88 |
| SAMix* | **70.32**±0.43 | **21.47**±0.68 | 95.89±0.74 | 2.01±0.29 | **44.41**±0.35 | **42.53**±0.97 | 70.33±0.48 | 20.69±0.75 | 17.19±0.48 | 51.61±0.55 | 45.12±0.73 | 36.53±0.62 |
| ProNC | 22.03±1.44 | 93.98±1.86 | 96.31±0.09 | 1.23±0.15 | 24.38±0.21 | 64.27±0.82 | 83.58±0.34 | 6.48±0.32 | 24.09±0.41 | 47.31±0.33 | 67.65±0.29 | 10.60±0.42 |
| **Ours** | 32.41±1.52 | 79.28±1.82 | **96.73**±0.11 | **0.79**±0.07 | 36.59±0.45 | 43.63±0.54 | **85.74**±0.48 | **1.67**±0.39 | **35.18**±0.41 | **11.44**±0.54 | **74.00**±0.64 | **1.02**±0.24 |
| **Replay buffer: 200** | | | | | | | | | | | | |
| Co$^2$L* | 65.57±1.37 | 36.35±1.16 | 93.43±0.78 | 6.71±0.35 | 27.38±0.85 | 67.82±0.41 | 53.94±0.76 | 38.22±0.34 | 13.88±0.40 | 73.25±0.21 | 42.37±0.74 | 47.11±1.04 |
| FC-NCCL* | 72.63±0.78 | 25.24±0.69 | 95.31±0.32 | 4.28±0.32 | 34.04±0.42 | 52.40±0.83 | 59.46±0.65 | 33.66±0.24 | 15.52±0.53 | 52.07±0.46 | 44.59±0.72 | 33.76±0.58 |
| SAMix* | 73.25±0.54 | **20.54**±0.44 | 96.68±0.32 | 1.27±0.21 | 46.95±0.71 | 39.24±1.26 | 73.48±0.29 | 16.51±0.53 | 18.39±0.38 | 51.66±0.75 | 49.51±0.83 | 30.89±0.53 |
| ProNC | 65.58±0.15 | 32.75±4.71 | 96.86±0.10 | 0.65±0.08 | 42.99±0.85 | 36.07±0.51 | 85.63±0.73 | 4.40±0.82 | 27.44±1.00 | 42.81±0.56 | 69.09±0.65 | 9.42±0.58 |
| **Ours** | **73.55**±0.92 | 24.12±0.58 | **96.96**±0.18 | **0.60**±0.18 | **49.02**±0.51 | **19.69**±0.60 | **85.83**±0.38 | **1.85**±0.23 | **33.80**±0.12 | **7.40**±0.25 | **72.12**±0.06 | **1.64**±0.28 |
| **Replay buffer: 500** | | | | | | | | | | | | |
| Co$^2$L* | 74.26±0.77 | 25.33±0.99 | 95.90±0.26 | 3.41±0.80 | 37.02±0.76 | 51.23±0.65 | 62.44±0.36 | 26.30±0.57 | 20.12±0.42 | 65.15±0.26 | 53.04±0.69 | 39.22±0.69 |
| FC-NCCL* | 75.51±0.52 | 22.59±1.02 | 96.14±0.25 | 3.21±0.25 | 40.25±0.58 | 41.66±0.78 | 65.85±0.44 | 24.84±0.91 | 20.31±0.34 | 46.08±0.56 | 53.46±0.59 | 26.45±0.79 |
| SAMix* | **78.62**±0.38 | **15.25**±0.53 | 96.98±0.57 | 1.16±0.12 | **55.85**±0.61 | 29.94±1.06 | 79.02±0.78 | 9.98±0.91 | 22.95±0.47 | 43.53±0.59 | 57.13±0.32 | 22.87±0.61 |
| ProNC | 73.95±0.68 | 26.75±1.29 | 96.95±0.14 | **0.24**±0.14 | 48.94±0.44 | 24.32±0.24 | **86.38**±0.43 | 4.35±0.17 | 29.06±0.32 | 38.58±0.39 | 69.77±0.89 | 9.52±0.96 |
| **Ours** | 78.00±0.28 | 18.29±0.24 | **97.15**±0.03 | 0.38±0.08 | 50.92±0.54 | **13.94**±0.47 | 85.86±0.59 | **1.84**±0.11 | **35.22**±0.35 | **5.47**±0.19 | **72.41**±0.45 | **1.30**±0.47 |

**Gains persist without replay.** Without storing past examples, CoMove improves over ProNC by 10.38, 12.21, and 11.09 Class-IL points across the three datasets: mean accuracies of 32.41%, 36.59%, and 35.18% (Table 2). TinyImageNet FF falls from 47.31 to 11.44. Table 3 extends the comparison to additional exemplar-free methods. On TinyImageNet, the 200-example buffer lowers FAA despite reducing forgetting; matched-initialization controls show better retention of early classes but lower accuracy on later ones (Appendix C.2). The teacher supplies temporal references on current inputs; the large Class-IL/Task-IL gap on CIFAR-10 still indicates difficult cross-task discrimination.

Table 3: **Exemplar-free comparison (buffer 0).** Class-IL/Task-IL FAA (↑). CIFAR-10 contrastive adaptations: Appendix B.3.

| Method | CIFAR-10 | | CIFAR-100 | | TinyImageNet | |
|---|---|---|---|---|---|---|
| | Class-IL | Task-IL | Class-IL | Task-IL | Class-IL | Task-IL |
| LwF | 19.29 | 91.81 | 8.46 | 72.99 | 7.25 | 64.21 |
| Co$^2$L | 19.68 | 54.75 | 26.06 | 51.91 | 13.43 | 38.98 |
| FC-NCCL | 19.32 | 79.88 | 30.48 | 56.69 | 14.04 | 42.59 |
| EFC | 19.84 | 96.10 | 18.27 | 84.21 | 17.37 | 73.29 |
| EFC++ | 20.01 | 96.29 | 18.65 | 84.12 | 17.61 | 73.60 |
| ProNC | 22.03±1.44 | 96.31±0.09 | 24.38±0.21 | 83.58±0.34 | 24.09±0.41 | 67.65±0.29 |
| Ours | **32.41**±1.52 | **96.73**±0.11 | **36.59**±0.45 | **85.74**±0.48 | **35.18**±0.41 | **74.00**±0.64 |

**Beyond ETF inference.** We also evaluate the learned representations with LDC (Gomez-Villa et al., 2024). CoMove retains the highest Class-IL accuracy among the evaluated methods across all nine dataset–buffer settings (Appendix C.7).

### 5.3 Ablation Studies

Each ablation shares its control's initial-task checkpoint, augmented inputs, and batch order. Table 4 reports Class-IL FAA; Appendix C.1 gives complete FAA/FF results.

Table 4: **Component ablations.** Class-IL FAA (%). —: identical.

| Variant | CIFAR-10 | | CIFAR-100 | | TinyImageNet | |
|---|---|---|---|---|---|---|
| | 200 | 500 | 200 | 500 | 200 | 500 |
| CoMove (full) | 72.36 | 78.30 | 47.74 | 50.46 | 34.80 | 35.36 |
| w/o pooling | 72.22 | 77.72 | 48.18 | 50.47 | 35.40 | 35.55 |
| w/o coverage transfer | — | — | 43.01 | 47.15 | 32.74 | 34.17 |
| w/o source responses | 67.28 | 74.86 | 48.01 | 50.15 | 34.54 | 35.70 |
| Own-ratio only | 67.92 | 74.35 | 47.09 | 48.91 | 35.15 | 35.35 |
| Raw teacher | 70.66 | 77.36 | 47.70 | 50.34 | 34.40 | 36.06 |

**Coverage transfer improves acquisition.** To isolate reallocation, we set $\rho_i = 1$ while preserving the total query weight. This reduces Class-IL accuracy by 4.73/3.62 points on CIFAR-100 and 2.06/1.19 on TinyImageNet. The CIFAR-100/500 coverage comparison uses its paired full-method control (50.77%; Appendix C.1). On CIFAR-100/200, final new-block accuracy falls from 41.5% to 21.3%, while old-block accuracy falls from 48.43% to 45.42%. In this paired run, *transfer benefits acquisition as well as retention*. CIFAR-10 has full coverage throughout these runs, making this intervention identical to the full method.

**Source responses favor retention.** Logistic responses preserve the allocation but cost 5.08/3.44 CIFAR-10 points. At buffer 200, new-block accuracy rises from 87.85% to 93.45%, while old-block accuracy falls from 68.49% to 60.74%: *acquisition improves at a larger retention cost*. Figure 3 traces these effects. The entropy deficit is one in binary tasks and mostly inactive on CIFAR-100.

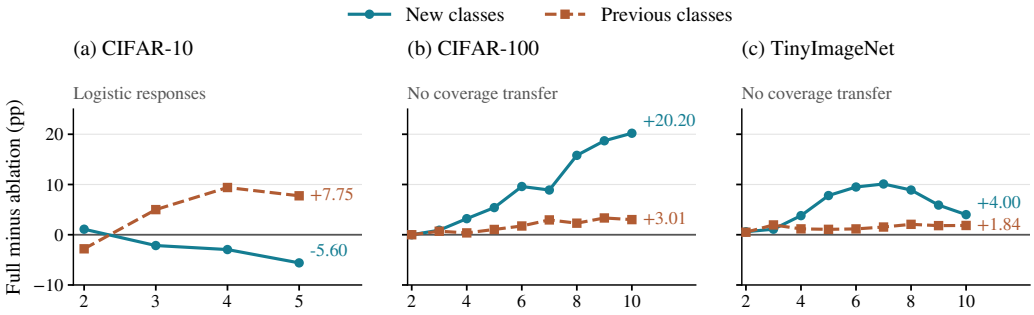


Figure 3: **Acquisition and retention respond to different components.** Full-minus-ablation Class-IL accuracy on newly learned and previous classes, with a common vertical scale; paired runs, buffer 200. **(a)** Source responses trade acquisition for retention on CIFAR-10. **(b,c)** Coverage transfer benefits acquisition on CIFAR-100 and TinyImageNet. Previous-class accuracy averages all preceding blocks, whose population grows with each task. Absolute trajectories are in Figure 5.

**Transport and shared competition.** Transport gains 1.70/0.94 Class-IL points on CIFAR-10 and 0.04/0.12 on CIFAR-100; on TinyImageNet, it helps at buffer 200 but hurts at 500. Neither Möbius nor normalized affine transport consistently outperforms the other (Appendix C.5). Removing cross-query gradients while preserving pooled loss values reduces CIFAR accuracy; removing pooling entirely has a smaller effect. Appendix C.1 reports paired acquisition/retention changes for all interventions.

**Evolution of supervision.** Replay support decreases as classes accumulate: coverage transfer starts at T2 on TinyImageNet but remains inactive on CIFAR-10 (Figure 6). Table 5 reports the teacher's share of source weight over final-task training, separately for old and new queries. On TinyImageNet, old-class queries with $n_{y_i} = 1$ have $\rho_i = 0$ and allocate all weight to the teacher; new queries ($\rho_i = 1$) retain peer and prototype supervision under the same allocation rule. Even at full coverage, the base temporal term can dominate the total source weight. Appendix C.3 distinguishes these weights from each source's attraction and competition.

Table 5: **Teacher allocation (%).** Final task, buffer 200.

| Dataset | Old | New |
|---|---|---|
| CIFAR-10 | 1.61 | 76.57 |
| CIFAR-100 | 99.02 | 88.76 |
| TinyImageNet | 100.00 | 88.93 |

## 6 Conclusion

Progressive ETF expansion creates a geometric mismatch between prototype targets and teacher references. CoMove addresses this mismatch through spherical transport and jointly allocates peer, prototype, and temporal supervision according to available support. It achieves the highest Class-IL accuracy in all six main low-replay comparisons and improves performance without replay. These results support coordinating the geometry of learning targets with the allocation of supervision as classes accumulate.

**Limitations and future work.** Our guarantees assume exact symmetric ETF expansion, and the basis construction requires $d \geq K$. Class-dependent expansion and approximate ETF targets in lower dimensions would require new transport guarantees. Unequal task sizes remain untested; adapting the allocation to actual task sizes is a natural extension. Count-based coverage also leaves feature diversity unmodeled. We hope this work inspires further research on evolving supervision, where learning targets and their relative influence adapt jointly to changes in class geometry and data availability.

## Ethics Statement

This work studies class-incremental learning on established image-classification benchmarks. It involves no new data collection or human-subject experiments.

## Reproducibility Statement

Section 5.1 summarizes the implementation, and Appendix A gives the complete training procedure. Hyperparameters, evaluation protocols, baseline implementations and paired controls are described in Appendices B.1–B.4. Appendix D contains the mathematical proofs. Additionally, the source code for our implementation accompanies this paper as supplementary material.

## AI Use Statement

Generative AI tools were used to assist with manuscript drafting and language polishing, as well as figure preparation; the authors reviewed all AI-assisted content and take full responsibility for the final text, figures, results, and claims.

## References

Matteo Boschini, Lorenzo Bonicelli, Pietro Buzzega, Angelo Porrello, and Simone Calderara. Class-incremental continual learning into the extended der-verse. *IEEE TPAMI*, 45(5):5497–5512, 2023. doi: 10.1109/TPAMI.2022.3206549.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. In *NeurIPS*, 2020. URL https://arxiv.org/abs/2004.07211.

Hyuntak Cha, Jaeho Lee, and Jinwoo Shin. Co2L: Contrastive continual learning. In *ICCV*, pp. 9496–9505, 2021. doi: 10.1109/ICCV48922.2021.00938. URL https://doi.org/10.1109/ICCV48922.2021.00938.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-GEM. In *ICLR*, 2019. URL https://openreview.net/forum?id=Hkf2_sC5FX.

Trung-Anh Dang, Vincent Nguyen, Ngoc-Son Vu, and Christel Vrain. Memory-efficient continual learning with neural collapse contrastive. In *WACV*, 2025.

Trung-Anh Dang, Vincent Nguyen, Ngoc-Son Vu, and Christel Vrain. Samix: Calibrated and accurate continual learning via sphere-adaptive mixup and neural collapse. *Mach. Learn.*, 115: 168, 2026. doi: 10.1007/s10994-026-07049-3. URL https://link.springer.com/article/10.1007/s10994-026-07049-3.

Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. PODNet: Pooled outputs distillation for small-tasks incremental learning. In *ECCV*, pp. 86–102. Springer, 2020.

Masih Eskandar, Tooba Imtiaz, Davin Hill, Zifeng Wang, and Jennifer Dy. STAR: Stability-inducing weight perturbation for continual learning. In *ICLR*, 2025. URL https://openreview.net/forum?id=6N5OM5Duuj.

Cong Fang, Hangfeng He, Qi Long, and Weijie J. Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *PNAS*, 118(43):e2103091118, 2021.

Enrico Fini, Victor G. Turrisi da Costa, Xavier Alameda-Pineda, Elisa Ricci, Karteek Alahari, and Julien Mairal. Self-supervised models are continual learners. In *CVPR*, pp. 9621–9630, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Fini_Self-Supervised_Models_Are_Continual_Learners_CVPR_2022_paper.html.

Alex Gomez-Villa, Dipam Goswami, Kai Wang, Andrew D. Bagdanov, Bartlomiej Twardowski, and Joost van de Weijer. Exemplar-free continual representation learning via learnable drift compensation. In *ECCV*, pp. 473–490. Springer-Verlag, 2024. doi: 10.1007/978-3-031-72667-5_27.

Dipam Goswami, Yuyang Liu, Bartlomiej Twardowski, and Joost van de Weijer. Fecam: Exploiting the heterogeneity of class distributions in exemplar-free continual learning. In *NeurIPS*, 2023.

X. Y. Han, Vardan Papyan, and David L. Donoho. Neural collapse under MSE loss: Proximity to and dynamics on the central path. In *ICLR*, 2022. URL https://openreview.net/forum?id=w1UbdvWH_R3.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *CVPR*, pp. 831–839, 2019.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020. URL https://arxiv.org/abs/2004.11362.

Gyuhak Kim, Changnan Xiao, Tatsuya Konishi, Zixuan Ke, and Bing Liu. A theoretical study on solving continual learning. In *NeurIPS*, 2022.

Gyuhak Kim, Changnan Xiao, Tatsuya Konishi, and Bing Liu. Learnability and algorithm for continual learning. In *ICML*, volume 202, pp. 16877–16896. PMLR, 2023. URL https://proceedings.mlr.press/v202/kim23x.html.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *PNAS*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Guannan Lai, Da-Wei Zhou, Xin Yang, and Han-Jia Ye. The lie of the average: How class incremental learning evaluation deceives you? In *ICLR*, 2026. URL https://openreview.net/forum?id=19LHXi9uLw.

Giulia Lanzillotta, Damiano Meier, and Thomas Hofmann. Heads collapse, features stay: Why replay needs big buffers. In *ICLR*, 2026. URL https://openreview.net/forum?id=IdW0d0mRnG.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. Technical report, Stanford University, 2015. URL https://cs231n.stanford.edu/reports/2015/pdfs/yle_project.pdf.

Jiyong Li, Dilshod Azizov, Yang Li, and Shangsong Liang. Contrastive continual learning with importance sampling and prototype-instance relation distillation. In *AAAI*, volume 38, pp. 13554–13562, 2024. doi: 10.1609/aaai.v38i12.29259.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE TPAMI*, 40(12):2935–2947, 2018. doi: 10.1109/TPAMI.2017.2773081.

Yan-Shuo Liang and Wu-Jun Li. Loss decoupling for task-agnostic continual learning. In *NeurIPS*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/249f73e01f0a2bb6c8d971b565f159a7-Abstract.html.

Haowei Lin, Yijia Shao, Weinan Qian, Ningxin Pan, Yiduo Guo, and Bing Liu. Class incremental learning via likelihood ratio based task prediction. In *ICLR*, 2024. URL https://openreview.net/forum?id=8QfK9Dq4q0.

Zhanwang Liu, Yuting Li, Haoyuan Gao, Yexin Li, Linghe Kong, Lichao Sun, and Weiran Huang. IDER: IDempotent experience replay for reliable continual learning. In *ICLR*, 2026. URL https://arxiv.org/abs/2603.00624v2.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *NeurIPS*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/f87522788a2be2d171666752f97ddebb-Paper.pdf.

Simone Magistri, Tomaso Trinci, Albin Soutif-Cormerais, Joost van de Weijer, and Andrew D. Bagdanov. Elastic feature consolidation for cold start exemplar-free incremental learning. In *ICLR*, 2024. URL https://openreview.net/forum?id=7D9X2cFnt1.

Simone Magistri, Tomaso Trinci, Albin Soutif-Cormerais, Joost van de Weijer, and Andrew D. Bagdanov. EFC++: Elastic feature consolidation with prototype re-balancing for cold start exemplar-free incremental learning, 2025. URL https://arxiv.org/abs/2503.10439v3.

Evan Markou, Thalaiyasingam Ajanthan, and Stephen Gould. Guiding neural collapse: Optimising towards the nearest simplex equiangular tight frame. In *NeurIPS*, 2024. URL https://arxiv.org/abs/2411.01248.

Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D. Bagdanov, and Joost Van De Weijer. Class-incremental learning: Survey and performance evaluation on image classification. *IEEE TPAMI*, 45(5):5513–5533, 2023. doi: 10.1109/TPAMI.2022.3213473.

Antoine Montmaur, Nicolas Larue, and Ngoc-Son Vu. Neural collapse inspired contrastive continual learning. In *BMVC*. BMVA, 2024. URL https://papers.bmvc2024.org/0579.pdf.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *PNAS*, 117(40):24652–24663, 2020. doi: 10.1073/pnas.2015509117.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, 2017.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *ICLR*, 2019. URL https://openreview.net/forum?id=B1gTShAct7.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *NeurIPS*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/fa7cdfad1a5aaf8370ebeda47a1ff1c3-Paper.pdf.

Ruilin Tong, Yuhang Liu, Javen Qinfeng Shi, and Dong Gong. Coreset selection via reducible loss in continual learning. In *ICLR*, pp. 13964–13999, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/24a8d40f6656e542f3fd43bac678e71b-Paper-Conference.pdf.

Tung Tran, Danilo Vasconcellos Vargas, and Khoat Than. LCA: Local classifier alignment for continual learning. In *ICLR*, 2026. URL https://openreview.net/forum?id=3uINmRldVW.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, volume 119, pp. 9929–9939, 2020. URL https://proceedings.mlr.press/v119/wang20k.html.

Zheng Wang, Wanhao Yu, Li Yang, and Sen Lin. Rethinking continual learning with progressive neural collapse. In *ICLR*, 2026. URL https://openreview.net/forum?id=E3bBZ02Qcc.

Yichen Wen, Zhiquan Tan, Kaipeng Zheng, Chuanlong Xie, and Weiran Huang. Provable contrastive continual learning. In *ICML*, volume 235, pp. 52819–52838. PMLR, 2024. URL https://proceedings.mlr.press/v235/wen24f.html.

Liang Xie, Yibo Yang, Deng Cai, and Xiaofei He. Neural collapse inspired attraction–repulsion-balanced loss for imbalanced learning. *Neurocomputing*, 527:60–70, 2023.

Hongye Xu and Bartosz Krawczyk. Two-way is better than one: Bidirectional alignment with cycle consistency for exemplar-free class-incremental learning. In *ICLR*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/c1e16692b361782abafe936bd7ac2d75-Abstract-Conference.html.

Yibo Yang, Shixiang Chen, Xiangtai Li, Liang Xie, Zhouchen Lin, and Dacheng Tao. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? In *NeurIPS*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/f7f5f501282771c96bb3fedcc96bedfe-Abstract.html.

Yibo Yang, Haobo Yuan, Xiangtai Li, Zhouchen Lin, Philip Torr, and Dacheng Tao. Neural collapse inspired feature-classifier alignment for few-shot class-incremental learning. In *ICLR*, 2023a. URL https://openreview.net/forum?id=y5W8tpojhtJ.

Yibo Yang, Haobo Yuan, Xiangtai Li, Jianlong Wu, Lefei Zhang, Zhouchen Lin, Philip Torr, Dacheng Tao, and Bernard Ghanem. Neural collapse terminus: A unified solution for class incremental learning and its variants, 2023b. URL https://arxiv.org/abs/2308.01746.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *ICML*, pp. 3987–3995. PMLR, 2017. URL https://proceedings.mlr.press/v70/zenke17a.html.

# SUPPLEMENTARY MATERIAL

The appendix separates implementation, experimental protocols, additional results, and proofs. All section titles and page numbers below are clickable.

**A Method and Implementation Details . . . 15**
A.1 Training Procedure and Notation . . . 15
A.2 Relation Responses and Boundary Cases . . . 16
A.3 Source Allocation . . . 17
A.4 Equivalent Objectives and Score Forces . . . 18
**B Experimental Setup . . . 19**
B.1 Benchmarks and Training Configuration . . . 19
B.2 Evaluation Metrics and Repeated Runs . . . 20
B.3 Baseline Sources and Protocol Differences . . . 20
B.4 Paired Controls and Reproducibility . . . 21
**C Additional Results and Analysis . . . 22**
C.1 Complete Component Ablations . . . 22
C.2 Acquisition and Retention Trajectories . . . 22
C.3 Measured Objective Dynamics . . . 23
C.4 Pooling: Values and Gradient Paths . . . 26
C.5 Transport Alternatives . . . 27
C.6 Geometric Illustrations . . . 27
C.7 Post-Hoc Classifier Evaluation . . . 28
**D Proofs . . . 30**
D.1 Technical Preliminaries . . . 30
D.2 Proof of Theorem 1 . . . 32
D.3 The affine feature squared-norm defect . . . 32
D.4 Spherical well-posedness, inverse and conformality . . . 32
D.5 Proof of Theorem 2 . . . 33
D.6 Equivalent potential and supervision allocation . . . 34
D.7 Negative-demand redistribution . . . 34
D.8 Entropic aggregation and the binary affine subloss . . . 35
D.9 Temporal score attraction . . . 35
D.10 Endpoint drift and chordal optimality . . . 35

# A METHOD AND IMPLEMENTATION DETAILS

The following procedure specifies the forward computation, gradient paths and boundary cases of CoMove; proofs are in Appendix D.

## A.1 TRAINING PROCEDURE AND NOTATION

**Initialization.** Fit the encoder and projector against the initial ETF using equal prototype and peer masses, logistic responses, and the unpooled partition. There is no teacher. Equation (28) defines this initial stage; the incremental class-count allocation applies from task two onward.

**An incremental stage.** Given the preceding encoder and projector, its ETF, the current-task data, and the retained replay examples, the computation is:

1. Freeze a copy of the preceding encoder and projector in evaluation mode. Expand the ETF from $O$ to $K = O + q$ classes and compute $\alpha$ and $\boldsymbol{b}$ from Theorem 1.
2. Merge replay with current-task data once. Compute class counts $n_c$, total size $N$, and the source masses in Equation (25).
3. For each augmented minibatch in each training epoch, compute normalized student features $\boldsymbol{z}_i$ and frozen teacher features $\boldsymbol{a}_i$ on the *same augmented inputs*. Transport the latter to $\boldsymbol{r}_i$ using Equation (5), including for new-class inputs.
4. Construct positive peers and within-task empirical negatives (Equation (16)); compute the shared negative partition (Equation (9)). Compute the detached responses in Equation (24) and the coverage-dependent weights in Equation (13).
5. Evaluate Equation (11), using the prototype fallback when a query has no positive peer. Average over queries and update only the student encoder and projector. Repeat steps 3–5 over the stage.
6. Evaluate the student with the current ETF. Neither the teacher nor a newly fitted classifier is required for native prediction.

Teacher parameters, prototypes, logit scale, dataset counts and response policies are fixed in backward; student peer keys and the classwise pooled partition are not detached. Training schedules and evaluation protocols are specified in Appendix B.

**Notation used in the implementation.**

The source index $s$ ranges over $\{\mathrm{p}, \mathrm{a}, \mathrm{t}\}$, denoting peers, the prototype anchor, and the teacher.

| Quantity | Meaning |
|---|---|
| $O, q, K$ | Inherited, arriving, and total class counts; $K = O + q$. |
| $t, T, B, d$ | Current stage, total stages, minibatch size, and feature dimension. |
| $M, \mathcal{M}_{t-1}$ | Replay capacity and retained examples from previous tasks. |
| $\boldsymbol{v}_c, \boldsymbol{u}_c$ | Inherited prototype before and after expansion. |
| $\alpha, \boldsymbol{b}$ | Common contraction and displacement of inherited prototypes. |
| $\boldsymbol{z}_i, \boldsymbol{a}_i, \boldsymbol{r}_i$ | Student feature, frozen teacher feature, transported reference. |
| $\mathcal{P}_i, \mathcal{N}_i$ | Same-class peers and different-class empirical negatives within the task block. |
| $\mathcal{B}_i, \xi_i, \mu_c$ | Same-class batch indices, empirical-to-prototype mass ratio, and class-shared log correction. |
| $P_i, E_i, A_i$ | Prototype mass, empirical mass, and shared log-partition. |
| $n_c, N, w_i$ | Merged-dataset class counts, total count, frequency weight. |
| $\rho_i, \omega_i^s$ | Count-based coverage and effective weight of source $s$. |
| $\delta_i, \kappa_i, \lambda_i$ | Detached entropy deficit, null mass, and temporal response rate. |
| $\mathcal{R}_i, \boldsymbol{d}_{ie}, \omega_{ie}$ | Tagged relations, reference direction, and weight of relation $e$. |
| $M_i, p_{ie}$ | Total query weight and normalized relation weight $\omega_{ie}/M_i$. |
| $p_{ic}(\boldsymbol{x}), p_{\mathrm{N},ie}$ | Wrong-class softmax probability and negative-branch responsibility. |
| $\psi_i, \ell_i^{\mathrm{a}}, \ell_i^{\mathrm{p}}, \ell_i^{\mathrm{t}}$ | Relation potential and prototype, peer, and teacher source losses. |
| $\ell_i, \mathcal{L}_{\mathrm{CoMove}}$ | Per-query objective and its minibatch average. |

### A.2 Relation Responses and Boundary Cases

This section specifies the relation policies needed to reproduce the objective. It completes Sections 4.2 and 4.3. The algebraic correspondence is established in Section D.6.

**Teacher evaluation and data access.** Replay is merged into the current training dataset once per stage. The frozen encoder and projector remain in evaluation mode and receive the same augmented input as the student, including new-class inputs. Their output supplies a temporal reference, without assigning a new class label. Prototypes and the logit scale are fixed during incremental training.

**Scores and negative sets.** For a minibatch of size $B$, use the positive scores

$$s_{ij}^{\mathrm{p}} = \gamma \boldsymbol{z}_i^\top \boldsymbol{z}_j, \qquad s_i^{\mathrm{a}} = \gamma \boldsymbol{z}_i^\top \boldsymbol{u}_{y_i}, \qquad s_i^{\mathrm{t}} = \gamma \boldsymbol{z}_i^\top \boldsymbol{r}_i. \tag{14}$$

and the same-class, positive-peer, and empirical-negative sets

$$\mathcal{B}_i = \{j \in \{1, \dots, B\} : y_j = y_i\}, \qquad \mathcal{P}_i = \mathcal{B}_i \setminus \{i\}, \tag{15}$$

$$\mathcal{N}_i = \{j \in \{1, \dots, B\} : y_j \neq y_i,\ g(y_j) = g(y_i)\}. \tag{16}$$

where $g$ is the task-block map from Section 3. The negative masses are

$$P_i = \sum_{c \in \mathcal{C}_t \setminus \{y_i\}} \exp(\gamma \boldsymbol{z}_i^\top \boldsymbol{u}_c), \quad E_i = \sum_{j \in \mathcal{N}_i} \exp(\gamma \boldsymbol{z}_i^\top \boldsymbol{z}_j). \tag{17}$$

and the unpooled decomposition is

$$\log(P_i + E_i) = \log P_i + \log(1 + \xi_i), \qquad \xi_i = \frac{E_i}{P_i}. \tag{18}$$

**Relation losses.** The potential in Equation (10) gives

$$\ell_i^{\mathrm{a}} = \psi_i(s_i^{\mathrm{a}}; \kappa_i, 0), \qquad \ell_i^{\mathrm{t}} = \psi_i(s_i^{\mathrm{t}}; \kappa_i, \lambda_i), \tag{19}$$

$$\ell_i^{\mathrm{p}} = \begin{cases} \dfrac{1}{|\mathcal{P}_i|} \displaystyle\sum_{j \in \mathcal{P}_i} \psi_i(s_{ij}^{\mathrm{p}}; 1, 0), & |\mathcal{P}_i| > 0, \\ \ell_i^{\mathrm{a}}, & |\mathcal{P}_i| = 0. \end{cases} \tag{20}$$

Actual peers use $(\kappa, \lambda) = (1, 0)$. The no-peer fallback uses the anchor score and anchor policy, including its null mass. Peer-key features and the shared partition remain differentiable.

**Relation policies.** Table 6 expands the source parameters in Equation (12) by query type. Missing peers retain their source allocation but use the prototype direction and its null mass; they do not create a logistic peer relation.

Table 6: **Relation policies.** Parameters of $\psi_i$ for incremental tasks. Actual peers use the logistic response; the no-peer fallback uses the prototype response.

| Query | Source | Reference | Null mass | Rate |
|---|---|---|---|---|
| New | Peer | $\boldsymbol{z}_j$ | 1 | 0 |
| New | Prototype | $\boldsymbol{u}_{y_i}$ | $1 - \delta_i$ | 0 |
| New | Teacher | $\boldsymbol{r}_i$ | $1 - \delta_i$ | 0 |
| Old | Peer | $\boldsymbol{z}_j$ | 1 | 0 |
| Old | Prototype | $\boldsymbol{u}_{y_i}$ | 1 | 0 |
| Old | Teacher | $\boldsymbol{r}_i$ | 1 | $\delta_i O/q$ |

**Detached entropy rule.** For $\mathcal{W}_i = \{c \in \mathcal{C}_t \setminus \{y_i\} : g(c) = g(y_i)\}$, define, for $c \in \mathcal{W}_i$ and $\boldsymbol{x} \in \mathbb{S}^{d-1}$,

$$p_{ic}(\boldsymbol{x}) = \frac{\exp(\gamma \boldsymbol{x}^\top \boldsymbol{u}_c)}{\sum_{\ell \in \mathcal{W}_i} \exp(\gamma \boldsymbol{x}^\top \boldsymbol{u}_\ell)}. \tag{21}$$

The normalized, clipped entropy is

$$h_i(\boldsymbol{x}) = \begin{cases} \mathrm{clip}_{[0,1]}\left(-\dfrac{\sum_{c\in\mathcal{W}_i} p_{ic}(\boldsymbol{x})\log p_{ic}(\boldsymbol{x})}{\log(q-1)}\right), & q>2, \\ 0, & q=2. \end{cases} \tag{22}$$

The resulting policy parameters are

$$\delta_i = \mathrm{sg}([1 - h_i(\boldsymbol{z}_i) - h_i(\boldsymbol{r}_i)]_+), \tag{23}$$

$$\kappa_i = 1 - \mathbf{1}_{\{y_i>O\}}\delta_i, \qquad \lambda_i = \mathbf{1}_{\{y_i\le O\}}\frac{O}{q}\delta_i. \tag{24}$$

Here $[u]_+ = \max\{u, 0\}$ and $\mathbf{1}_{\mathcal{E}}$ is the indicator of event $\mathcal{E}$. These responses are recomputed at each forward pass and held constant in backward through sg in Equation (23). Geometry, teacher, scale and counts are detached too; the shared negative margin remains differentiable. We chose this entropy-based rule for the loss response; the ETF expansion determines the transport geometry. The geometric transport specifies the reference direction; the response determines its positive-score attraction and negative demand (Section D.9).

**Source masses and coverage.** The base source masses are

$$m_{\mathrm{p}} = \frac{q}{K}, \qquad m_{\mathrm{a}} = m_{\mathrm{t}} = \frac{O}{2K}, \qquad m_{\mathrm{a}} + m_{\mathrm{p}} + m_{\mathrm{t}} = 1. \tag{25}$$

For merged-dataset counts $n_c$ and total size $N = \sum_c n_c$, the class weight and coverage are

$$w_i = \frac{N}{Kn_{y_i}}, \qquad \rho_i = \begin{cases} \min\{1, (n_{y_i}-1)/(K-1)\}, & y_i \le O, \\ 1, & y_i > O. \end{cases} \tag{26}$$

Only represented classes occur as queries, so $n_{y_i} \ge 1$. The counts come from the merged training dataset; minibatch peer counts enter the peer loss separately. Coverage uses dataset counts as a proxy for support, with the chosen threshold $K-1$. It measures neither feature diversity nor teacher reliability; those properties require separate measurements. At zero coverage, prototype and peer positive weights vanish and their mass moves to the teacher; the negative competition remains active. The total query weight is

$$M_i = w_i(m_{\mathrm{a}} + m_{\mathrm{p}}) + m_{\mathrm{t}}, \tag{27}$$

and is unchanged by coverage transfer.

**First task.** CoMove starts from a fixed initial ETF; unlike ProNC, it does not fit the initial frame to learned first-task class means. During the initial task, $K = q$ and no teacher or pooling is used. With the unpooled partition $A_i^{(1)} = \log(P_i + E_i)$, define $\psi_i^{(1)}(s) = \log(1 + e^{A_i^{(1)} - s})$. Set $\ell_i^{\mathrm{a},(1)} = \psi_i^{(1)}(s_i^{\mathrm{a}})$ and define $\ell_i^{\mathrm{p},(1)}$ by averaging $\psi_i^{(1)}(s_{ij}^{\mathrm{p}})$ over positive peers, with the same anchor fallback. The initial-task objective is

$$\mathcal{L}_{\mathrm{birth}} = \frac{1}{B}\sum_i \frac{N}{qn_{y_i}}\left(\tfrac{1}{2}\ell_i^{\mathrm{a},(1)} + \tfrac{1}{2}\ell_i^{\mathrm{p},(1)}\right). \tag{28}$$

### A.3 Source Allocation

The base allocation assigns class fractions to the three supervision sources. The symmetry argument below determines an equal split under an exchangeability assumption on the inherited sources. We adopt this fixed rule across settings. The argument leaves the statistical reliability of the two sources and the optimality of their allocation unresolved.

**Incremental budget.** Let $O$ denote the number of inherited classes, let $q$ denote the number of new classes, and let $K = O + q$ denote the number of classes observed after the current expansion. We first allocate mass according to the composition of the incremental problem:

$$m_{\mathrm{inherited}} = \frac{O}{K}, \qquad m_{\mathrm{new}} = \frac{q}{K}, \qquad m_{\mathrm{inherited}} + m_{\mathrm{new}} = 1.$$

The new-class share is assigned to same-class peer supervision, so the two quantities are identical:

$$\boxed{m_{\mathrm{p}} = m_{\mathrm{new}} = \frac{q}{K}}.$$

The inherited share must represent two distinct pieces of information: current classifier alignment through the anchor and temporal continuity through the transported teacher. We split this inherited budget symmetrically:

$$m_{\mathrm{a}} = m_{\mathrm{t}} = \frac{1}{2} m_{\mathrm{inherited}} = \frac{O}{2K}.$$

Consequently,

$$\boxed{m_{\mathrm{a}} = \frac{O}{2K}, \qquad m_{\mathrm{p}} = \frac{q}{K}, \qquad m_{\mathrm{t}} = \frac{O}{2K}} \qquad \text{and} \qquad m_{\mathrm{a}} + m_{\mathrm{p}} + m_{\mathrm{t}} = 1. \tag{29}$$

**Why one half.** The inherited budget is shared by the anchor and teacher, giving

$$m_{\mathrm{a}} + m_{\mathrm{t}} = m_{\mathrm{inherited}} = \frac{O}{K}.$$

More generally, this conservation equation can be parameterized by a single fraction $\theta \in [0, 1]$:

$$m_{\mathrm{a}} = \theta m_{\mathrm{inherited}} = \theta \frac{O}{K}, \qquad m_{\mathrm{t}} = (1-\theta) m_{\mathrm{inherited}} = (1-\theta) \frac{O}{K}.$$

When no prior reason privileges the anchor over the teacher, invariance under exchanging these two sources imposes $\theta = 1 - \theta$, hence $\theta = 1/2$. We fix this symmetric split throughout the experiments.

**Coverage interpretation.** Replay coverage redistributes the semantic budget: supported mass remains associated with semantic sources, while unsupported mass is assigned to the temporal reference. The base temporal share remains present in every incremental transition.

### A.4 Equivalent Objectives and Score Forces

The following identities express coverage transfer and distinguish source weights from their contributions to score attraction and competition.

**Source-grouped and transfer forms.** Using the source losses defined in Appendix A.2, Equation (11) becomes

$$\begin{aligned} \ell_i &= \omega_i^{\mathrm{a}} \ell_i^{\mathrm{a}} + \omega_i^{\mathrm{p}} \ell_i^{\mathrm{p}} + \omega_i^{\mathrm{t}} \ell_i^{\mathrm{t}} \\ &= \underbrace{w_i \rho_i (m_{\mathrm{a}} \ell_i^{\mathrm{a}} + m_{\mathrm{p}} \ell_i^{\mathrm{p}})}_{\text{covered supervision}} + \underbrace{m_{\mathrm{t}} \ell_i^{\mathrm{t}}}_{\text{base distillation}} \\ &\quad + \underbrace{w_i (1-\rho_i)(m_{\mathrm{a}} + m_{\mathrm{p}}) \ell_i^{\mathrm{t}}}_{\text{coverage transfer}}. \end{aligned} \tag{30}$$

The temporal coefficient comprises a base allocation and the transferred mass, both multiplying the same relation loss. The relation responses and peer fallback are those of Appendix A.2.

**Coverage correction.** Evaluate Equation (11) at $\rho_i = 1$, holding all other quantities fixed. The change from this no-transfer objective is

$$\begin{aligned} \ell_i - \ell_i|_{\rho_i = 1} = w_i (1 - \rho_i) \big[ &(m_{\mathrm{a}} + m_{\mathrm{p}}) \ell_i^{\mathrm{t}} \\ &- m_{\mathrm{a}} \ell_i^{\mathrm{a}} - m_{\mathrm{p}} \ell_i^{\mathrm{p}} \big]. \end{aligned} \tag{31}$$

This identifies a no-transfer control that preserves total query weight. The correction can have either sign; conservation of weights does not imply conservation of gradient norms.

**Tagged relations.** For query $i$, let $\mathcal{R}_i$ contain the prototype and teacher relations, together with one tagged relation $(\mathrm{p}, j)$ for each $j \in \mathcal{P}_i$, or one prototype fallback when $\mathcal{P}_i$ is empty. Their directions are respectively $\boldsymbol{u}_{y_i}$, $\boldsymbol{r}_i$, $\boldsymbol{z}_j$, and $\boldsymbol{u}_{y_i}$. Each relation $e$ has a unit direction $\boldsymbol{d}_{ie}$, weight $\omega_{ie}$, and the null mass and rate of its source. The aggregate source weights are

$$\omega_i^{\mathrm{a}} = w_i \rho_i m_{\mathrm{a}}, \qquad \omega_i^{\mathrm{p}} = w_i \rho_i m_{\mathrm{p}},$$
$$\omega_i^{\mathrm{t}} = m_{\mathrm{t}} + w_i(1-\rho_i)(m_{\mathrm{a}} + m_{\mathrm{p}}).$$

Each actual peer receives $\omega_i^{\mathrm{p}}/|\mathcal{P}_i|$. If there are no peers, one fallback relation receives $\omega_i^{\mathrm{p}}$, the anchor direction and its policies. With $s_{ie} = \gamma \boldsymbol{z}_i^\top \boldsymbol{d}_{ie}$, this gives

$$\ell_i = \sum_{e \in \mathcal{R}_i} \omega_{ie} \psi_i(s_{ie}; \kappa_{ie}, \lambda_{ie}), \qquad M_i = \sum_{e \in \mathcal{R}_i} \omega_{ie}. \tag{32}$$

The normalized weights $p_{ie} = \omega_{ie}/M_i$ define a distribution over relations. The objective averages the relation-wise potentials under this distribution and scales the result by $M_i$. Each potential applies its own logarithm before this average; averaging the positive scores first would give a different loss.

For each relation, define the odds

$$\zeta_{ie} = \kappa_{ie} \exp[s_{ie} - A_i - \lambda_{ie}(s_{ie} - s_i^0)].$$

This quantity also determines the score forces below.

**Score forces.** Treat the partition and positive scores as independent coordinates, holding the policy parameters and frozen score $s_i^0$ fixed. Write

$$p_{\mathrm{N},ie} = (1 + \zeta_{ie})^{-1}.$$

Differentiating one relation gives

$$\begin{aligned} \partial_{A_i} \psi_i(s_{ie}; \kappa_{ie}, \lambda_{ie}) &= p_{\mathrm{N},ie}, \\ -\partial_{s_{ie}} \psi_i(s_{ie}; \kappa_{ie}, \lambda_{ie}) &= p_{\mathrm{N},ie} + \lambda_{ie}(1 - p_{\mathrm{N},ie}). \end{aligned} \tag{33}$$

The weighted negative demand is $\omega_{ie} p_{\mathrm{N},ie}$ and the weighted positive score force is $\omega_{ie}[p_{\mathrm{N},ie} + \lambda_{ie}(1 - p_{\mathrm{N},ie})]$. For rate one the positive score force before weighting is exactly one, whereas the negative responsibility can be smaller. These scalar score derivatives enter the full feature gradients through the normalization, peer-key and shared-partition paths. The plotted force fractions therefore describe score responses; they do not measure parameter gradient norms.

# B Experimental Setup

We specify the training and evaluation protocols used in Section 5.

## B.1 Benchmarks and Training Configuration

**Training configuration.** We follow the experimental protocol of ProNC (Wang et al., 2026), using the same ResNet-18 backbone, batch size of 32, and training epochs. CIFAR-10 comprises five two-class tasks; CIFAR-100 ten ten-class tasks; TinyImageNet ten twenty-class tasks. CIFAR models train for 50 epochs per task, with learning rates 0.01 (CIFAR-10) and 0.03 (CIFAR-100), decayed at epochs 35 and 45. TinyImageNet uses 100 epochs, learning rate 0.03, and milestones 45, 75, and 95. These epoch counts also apply to the paired component ablations. The learning rates, zero momentum, and zero weight decay match ProNC's reported settings. We use SGD and reset the learning rate at each task, multiplying it by 0.1 at each milestone. The same source-weight formulas apply across the six replay settings and the zero-buffer evaluation, with no mixing coefficient selected separately for each configuration.

**Data augmentation and replay.** We use the dataset transforms and class-balanced random replay policy from the ProNC implementation (Wang et al., 2026). Training images undergo random cropping with four-pixel padding, horizontal flipping, and dataset normalization. Crop sizes are $32 \times 32$ for CIFAR and $64 \times 64$ for TinyImageNet. After each task with $K$ seen classes, memory retains $\lfloor M/K \rfloor$ examples per class: old class memories are truncated to this quota, and new examples are selected uniformly without replacement. Unused capacity is left unfilled. At the next task, memory is merged with the current data and shuffled into minibatches of 32 examples.

**Optimization and inference.** CoMove uses a trainable, bias-free $512 \times 512$ linear projector initialized to the identity, followed by $\ell_2$ normalization. The logit scale is fixed at $\gamma \approx 6.5$ throughout training on all datasets and memory budgets. We obtain this value by initializing the scale at $1/0.07$ and optimizing it on the first CIFAR-100 task. It controls the sharpness of the similarity scores, akin to an inverse temperature in contrastive learning. At each incremental task, we freeze the preceding encoder and projector, expand the ETF, and compute its displacement and merged-data counts. Each augmented minibatch supplies student features and transported teacher references. Gradients pass through student queries, peer keys, and the shared negative partition; the teacher, prototypes, scale, counts, and entropy responses are detached. Prediction uses only the student and current ETF (Equation (2)), without a fitted post-hoc classifier.

### B.2 Evaluation Metrics and Repeated Runs

**Repeated runs.** Tables 1, 2, and 3 report CoMove means and sample standard deviations over three runs. These include both resumed and from-scratch trajectories; the deviations therefore include differences in initialization and execution settings. For published baselines, uncertainties follow the source publications; for repeated local reproductions, we compute sample standard deviations. Seed values and per-run configurations are supplied with the code.

**Accuracy and forgetting.** Let $T$ be the total number of tasks and $a_{t,k}$ the accuracy on task $k$ after learning stage $t$. For either Class-IL or Task-IL, we use

$$\mathrm{FAA} = \frac{1}{T}\sum_{k=1}^{T} a_{T,k}, \qquad \mathrm{FF} = \frac{1}{T-1}\sum_{k=1}^{T-1}\left(\max_{k\le t<T} a_{t,k} - a_{T,k}\right). \tag{34}$$

The maximum includes the initial-task evaluation and excludes the final stage. Class-IL selects among all seen labels. Task-IL restricts the same scores to the true task block. Route accuracy is the proportion of samples whose globally predicted class belongs to their true block. In the acquisition/retention figure, new-block accuracy is evaluated on the block just learned; old-block accuracy averages the previous blocks, whose population grows with each transition.

**Initialization.** The initial ETF is constructed from a random orthonormal basis and held fixed during training. ProNC instead fits its initial frame to the learned features (Wang et al., 2026). The initial task uses equally weighted peer and fixed-ETF prototype supervision, without a teacher or pooling. Both sources use logistic responses; Appendix A.2 gives the complete objective. The incremental allocation applies from the second task onward.

### B.3 Baseline Sources and Protocol Differences

**Main comparison and headline gains.** The main comparison reports native-classifier results from the reference literature suite and our three-seed means. For each of its six dataset–budget cells, we subtract the largest compared baseline Class-IL FAA from our mean. The six differences average 4.69 points; the largest is 6.36. Each difference uses the best baseline in its dataset–buffer cell. In particular, the SAMix CIFAR-100 task split and its post-hoc classifier fitting differ from ours. The main claim concerns the complete evaluated system: its fixed-frame initialization also differs from ProNC's initial-frame fitting.

**Baseline sources.** The main literature comparison uses the results reported by ProNC (Wang et al., 2026). Additional Co$^2$L, FC-NCCL, and SAMix entries follow their stated source protocols. The local ProNC runs use its Mammoth implementation; EFC, EFC++, and LwF use Mammoth adaptations. At zero buffer, ProNC uses local three-run means on all three datasets.

**CIFAR-10 contrastive adaptations.** The Co$^2$L (Cha et al., 2021) and FC-NCCL (Dang et al., 2025) CIFAR-10 entries in Table 3 use five two-class tasks, zero replay, batch size 32, and 50 epochs per task. We retain their representation objectives and train a persistent linear classifier online on detached current-task features, using cross-entropy over seen classes. The classifier is carried across tasks. Each method uses two independently augmented views with our crop, flip, and normalization pipeline. These single-run adaptations use no additional post-hoc classifier training or stored past examples. Their original evaluation protocols are represented separately in Table 2.

**SAMix comparison.** The SAMix row corresponds to TA-NCCL+SAMix in the source paper's Table 1 (Dang et al., 2026). It adds mixed inputs and spherical prototype targets to representation learning with hardness-softness distillation, then fits a linear classifier for 100 epochs with the encoder frozen. The reported schedule uses batch size 512 and 500 initial epochs, followed by 100 epochs per CIFAR task and 50 per TinyImageNet task; our schedule uses batch size 32 and 50/100 epochs per task on CIFAR/TinyImageNet. The comparison consequently spans different representation-learning and classifier-fitting protocols; training time is not compared.

LODE values are taken from ProNC's main comparison, without standard deviations because none are reported for that row. BFP+ID is the fixed BFP composition of IDER (Liu et al., 2026), using arXiv v2 Tables 1 and 7, averaged over five runs. We leave BFP+ID out of Table 1 because its backbone and replay policy differ from ours. For context, its published Class-IL FAA is 71.99, 76.65, 48.53 and 30.62 on C10/200, C10/500, C100/500 and Tiny/500; unreported budgets and Task-IL metrics are unavailable. Its splits and epoch counts match ours, but its replay policy, backbone modification and schedules differ. The quoted results come from their published experiments; we did not run a paired reproduction.

**Evaluation scope.** Task order can substantially affect CIL performance, and averages over a few sequences need not capture its full distribution (Lai et al., 2026). The acquisition/retention analyses resolve task-wise behavior along the evaluated sequences. Native and post-hoc classifiers are evaluated separately, since useful representations may survive classifier-level forgetting (Lanzillotta et al., 2026). Task identification and within-task discrimination also impose distinct requirements (Kim et al., 2022; 2023; Lin et al., 2024).

### B.4 Paired Controls and Reproducibility

Component ablations use one matched run per condition, starting from the same seeded initial-task checkpoint and using the same training schedule and augmentation policy. They isolate incremental interventions after the same initialization; Appendix C.1 defines the controls.

# C Additional Results and Analysis

We report complete ablations, supervision dynamics, geometric diagnostics and post-hoc classifier evaluations.

## C.1 Complete Component Ablations

Tables 7 and 8 give the complete component comparisons from Section 5.3.

Table 7: **Complete component ablations.** Final Class-IL/Task-IL accuracy (%). The CIFAR-100/500 coverage comparison uses the full control in the last row; other comparisons use the first row. Identical denotes an unchanged objective.

| | C10/200 | C10/500 | C100/200 | C100/500 |
|---|---|---|---|---|
| Full method | 72.36/97.04 | 78.30/96.99 | 47.74/85.39 | 50.46/85.94 |
| No pooling | 72.22/96.96 | 77.72/96.99 | 48.18/85.30 | 50.47/86.16 |
| No coverage transfer | identical | identical | 43.01/83.35 | 47.15/84.01 |
| Logistic responses | 67.28/95.43 | 74.86/95.81 | 48.01/85.31 | 50.15/85.56 |
| Raw teacher | 70.66/96.69 | 77.36/97.00 | 47.70/85.59 | 50.34/85.94 |
| Normalized affine | 72.91/97.01 | 78.55/96.93 | 47.79/85.71 | 50.08/86.00 |
| Own-ratio pooling gradient | 67.92/95.85 | 74.35/96.42 | 47.09/85.42 | 48.91/85.88 |
| Full control for coverage | — | — | 47.74/85.39 | 50.77/85.74 |

Table 8 gives the corresponding complete TinyImageNet ablation results for both replay budgets.

Table 8: **TinyImageNet component ablations.** Final FAA/FF (Class-IL / Task-IL, %) for paired runs initialized from the same seeded checkpoint within each buffer setting.

| Variant | Buffer 200 | Buffer 500 |
|---|---|---|
| Full objective | 34.80/72.95 (8.17/1.62) | 35.36/73.29 (6.01/1.73) |
| No pooling | 35.40/73.28 (7.34/1.64) | 35.55/73.35 (6.36/1.70) |
| No coverage transfer | 32.74/70.48 (9.17/3.34) | 34.17/70.24 (10.10/3.88) |
| Logistic responses | 34.54/72.96 (7.77/1.32) | 35.70/73.27 (6.02/1.48) |
| Raw teacher | 34.40/72.49 (8.63/2.19) | 36.06/73.55 (6.71/1.48) |
| Normalized affine | 34.40/73.24 (8.09/1.31) | 36.07/73.33 (5.68/1.73) |
| Own-ratio gradients only | 35.15/72.88 (8.20/1.81) | 35.35/73.56 (6.68/1.67) |

**Paired protocol.** All interventions start from the same seeded initial-task checkpoint, trained with equal anchor and peer source masses and fixed logit scale $\gamma \approx 6.5$. Each continuation trains for 50 epochs per task on CIFAR and 100 on TinyImageNet, with batch size 32. Within each comparison, the batch order and augmented inputs are matched.

The CIFAR-100/500 coverage experiment has its own matched full-method control (50.77/85.74); the other interventions use 50.46/85.94. Both controls are reported in Table 7.

**Interventions.** No pooling replaces the shared partition by $\log(P_i + E_i)$. No coverage transfer sets $\rho_i = 1$ without changing total query weight. Logistic responses set $\delta_i = 0$, hence unit null masses and zero temporal rates, while retaining the sources and partition. The raw-teacher and normalized-affine controls change the teacher target; both recompute the reference score and entropy policy from that actual target. They thus test transport together with its induced policies.

## C.2 Acquisition and Retention Trajectories

**Replay on TinyImageNet.** We compare three pairs of continuations with buffers of zero and 200, each initialized from the same first-task checkpoint. The zero-buffer runs finish 1.44 points higher in mean Class-IL FAA, while replay reduces mean forgetting from 11.29 to 7.40 points. Replay leads in mean accuracy through task 7; the zero-buffer runs overtake it at task 8. At the final evaluation, replay improves accuracy on each of the first four task blocks but reduces it on each later block. This comparison measures the effect of the configured memory budget, including its induced sampling and loss weights.

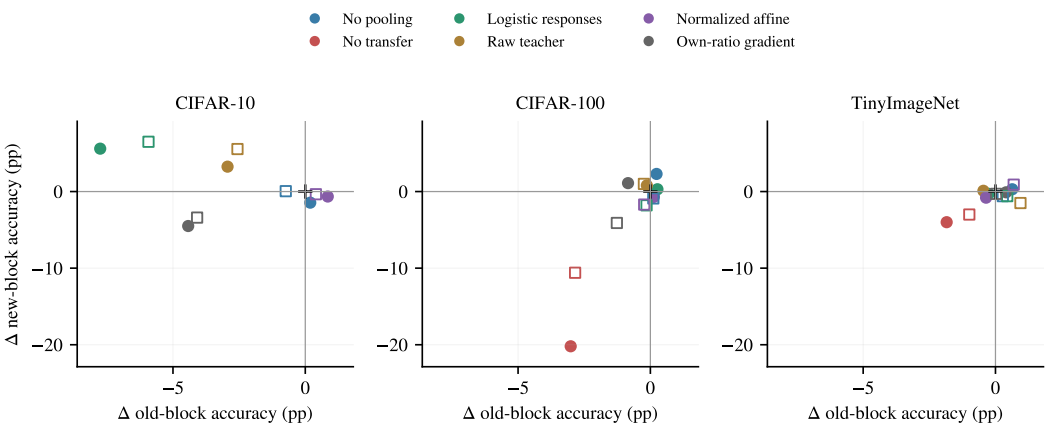


Figure 4: **Paired acquisition and retention effects.** Changes relative to the full control in final old-block and new-block accuracy. Filled circles: buffer 200; open squares: buffer 500. The cross marks the full method. CIFAR-10 without transfer coincides with the origin.

**Acquisition–retention trade-offs.** Figure 4 reports endpoint changes for all interventions and budgets. Each point subtracts its matched full-method control, including the separate C100/500 coverage control. If $T$ is the number of tasks, the overall Class-IL change satisfies $\Delta\mathrm{FAA} = ((T-1)\Delta\mathrm{old} + \Delta\mathrm{new})/T$. Thus the plot separates the contributions hidden by the average score.

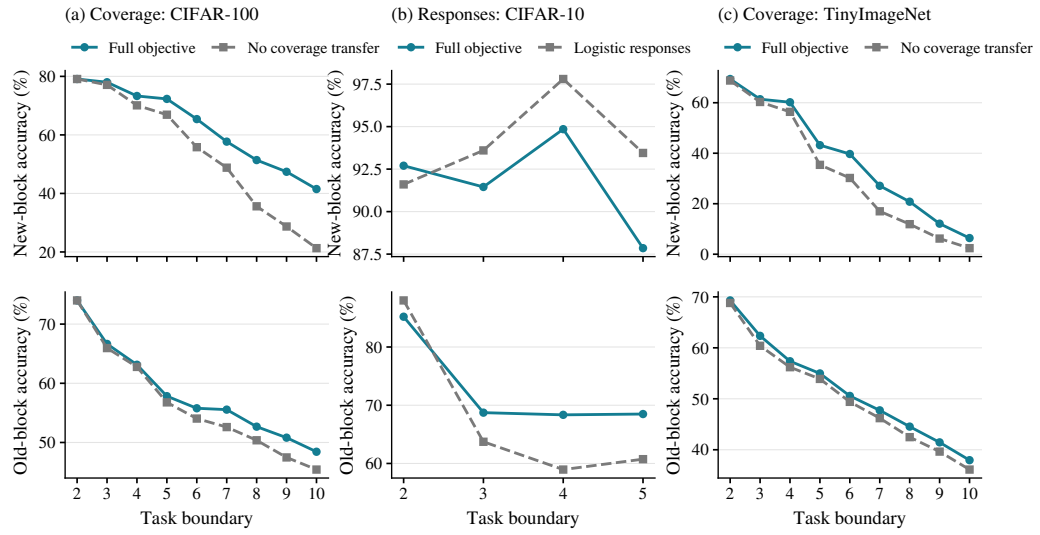


Figure 5: **Absolute acquisition and retention trajectories.** Class-IL accuracy on the newly learned block (top) and the mean over all previous blocks (bottom), for the paired runs underlying Figure 3. Buffer 200; the previous-block population grows with each task. Panel-specific vertical scales show the absolute trajectories; the main figure uses a common scale for differences.

### C.3 Measured Objective Dynamics

These measurements complement the interventions in Section 5.3. Table 9 gives the complete source allocation for old and new queries during final-task training.

**Coverage and transferred weight.** At the first transition, $\|\boldsymbol{b}\|^2 = 1/(2q-1)$, giving prototype displacements of $35.3°$, $13.3°$, and $9.2°$ on CIFAR-10, CIFAR-100 and TinyImageNet, respectively. Figure 6 follows the full objective on all six CIFAR and TinyImageNet configurations. For old

Table 9: **Source allocation (%).** Final-task training; buffer 200.

| Source | C10 | C100 | Tiny |
|---|---|---|---|
| *Old queries* | | | |
| Peers | 32.80 | 0.18 | 0.00 |
| Prototype | 65.59 | 0.80 | 0.00 |
| Teacher | 1.61 | 99.02 | 100.00 |
| *New queries* | | | |
| Peers | 7.81 | 2.04 | 2.01 |
| Prototype | 15.62 | 9.20 | 9.05 |
| Teacher | 76.57 | 88.76 | 88.93 |

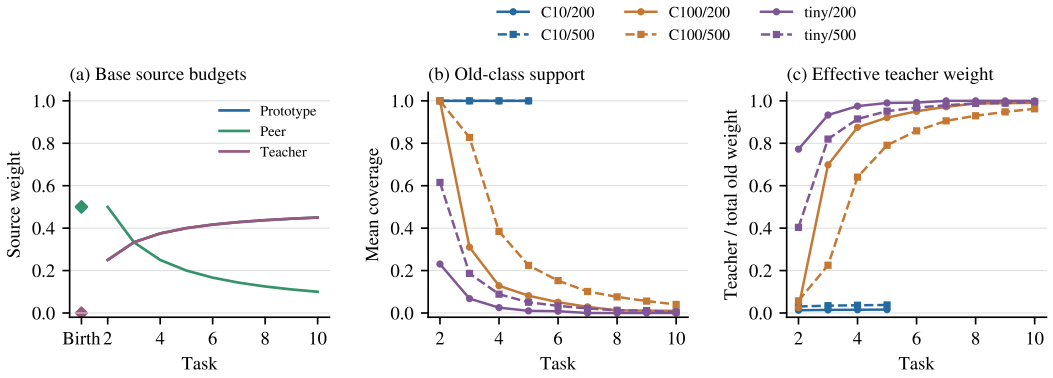


Figure 6: **Evolution of CoMove supervision.** (a) Base prototype, peer, and teacher masses; diamonds denote initialization. (b) Measured old-query coverage. (c) Effective teacher weight divided by total old-query weight, including both its base share and the transferred supervision. Solid/dashed curves use buffers 200/500. Measurements aggregate all training batches of the full method.

queries we report mean coverage and effective teacher weight as a fraction of total source weight. This fraction is $\overline{\omega^{\mathrm{t}}}/\overline{M}$, where bars denote averages over old-query occurrences during training of the task. The teacher weight includes both its base allocation and the transferred mass. CIFAR-10 has unit coverage and zero transfer at both budgets. On CIFAR-100 the reduction of support progressively reallocates supervision, consistent with the delayed divergence of the no-transfer runs.

**Allocated weight versus score forces.** Figures 7 and 8 distinguish the effective source weights $\omega_i^s$, $s \in \{\mathrm{p}, \mathrm{a}, \mathrm{t}\}$, from the weighted positive and negative coefficients in Equation (33). For each family, we sum its coefficients over all old-query occurrences in every batch and epoch of a task, and divide the teacher sum by the sum over all three sources. Aggregation therefore precedes normalization. The weights include class reweighting and coverage transfer, unlike the base masses in Figure 2. The peer term includes its prototype fallback.

At the final CIFAR-10/200 task, the teacher accounts for 1.61% of allocated weight but 56.86% of positive attraction and 0.15% of negative demand. At buffer 500 these shares are 3.79%, 71.56%, and 0.34%. Thus the temporal response separates attraction from competition even when the teacher has little allocated weight. On CIFAR-100/200 the corresponding shares are 99.02%, 96.58%, and 96.58%; TinyImageNet/200 reaches 100% in all three quantities. In these late-task regimes, coverage transfer dominates the allocation.

The lower panels show the corresponding paired accuracy differences. The initial-task test set is fixed, unlike the growing collection of previous blocks in the main trajectory figure. On CIFAR-10/200, the full method improves final initial-block accuracy by 7.95 points over logistic responses, while new-block accuracy is 5.60 points lower. On CIFAR-100/200, transfer improves the same two quantities by 8.60 and 20.20 points.

Numerical checks on fixed feature batches reproduce the analytic prototype and teacher score coefficients within $10^{-5}$ absolute and $10^{-4}$ relative tolerance. At a common state, setting $\rho_i = 1$ preserves total query weight, and logistic responses preserve all three source weights. The diagnostics cover incremental tasks, when a teacher reference is available.

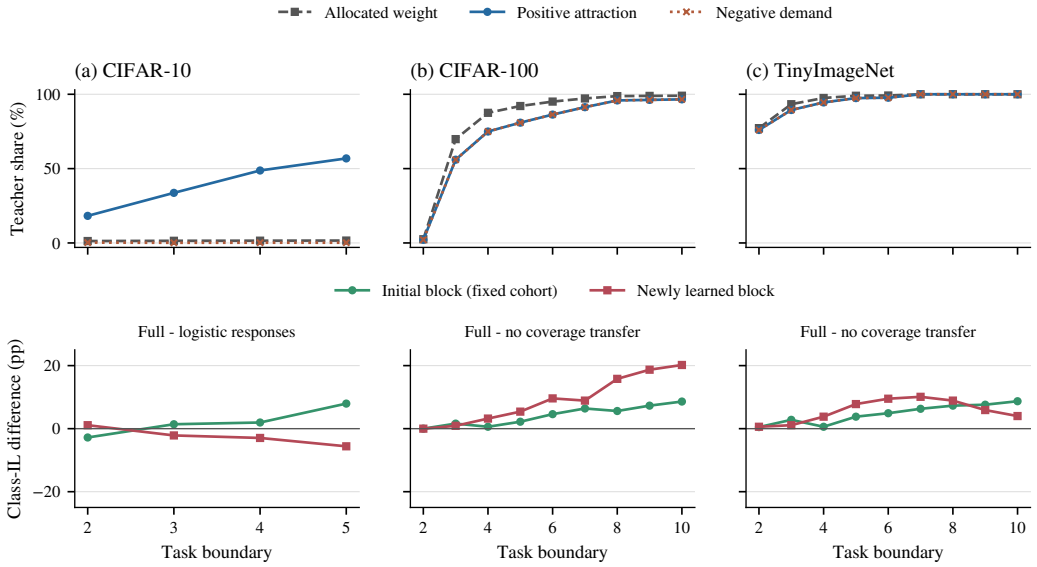


Figure 7: **Allocated weight and score forces at buffer 200.** Top: teacher shares of allocated weight, positive attraction, and negative demand, aggregated over all old-query occurrences during each task. Bottom: paired Class-IL differences, full method minus logistic responses on CIFAR-10, or minus no coverage transfer on CIFAR-100 and TinyImageNet. The initial block is a fixed test set; the newly learned block changes with the task. Forces are scalar score derivatives, computed before backpropagation through normalized features and model parameters.

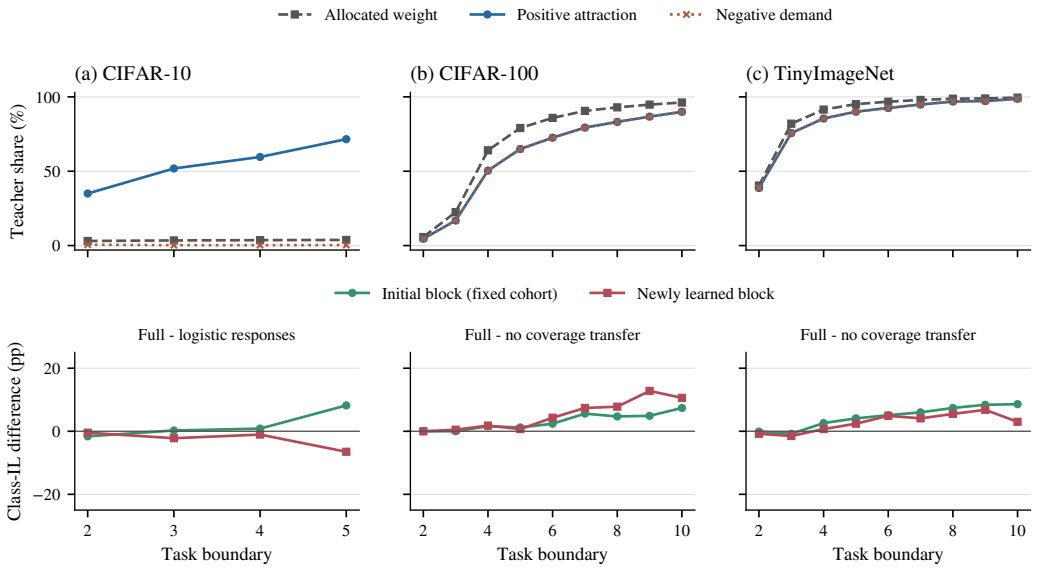


Figure 8: **Allocated weight and score forces at buffer 500.** The definitions and interventions match Figure 7. Teacher shares aggregate all old-query occurrences; positive and negative shares coincide when the temporal rate is zero. Bottom: paired full-minus-ablation Class-IL differences on the fixed initial block and the newly learned block. The CIFAR-100 no-transfer comparison uses its separately paired full control (FAA 50.77); the other interventions use the full control at 50.46.

**Policy activity and signed coefficients.** Figure 9 distinguishes two measured quantities. The entropy deficit is exactly one for binary CIFAR-10 tasks by definition, whereas its old-query mean is close to zero on CIFAR-100. The signed-channel diagnostic counts new queries whose coefficient of $\log P_i$ is below $-10^{-8}$, including the batch-average factor. It extends the common-batch example in Figure 10 across the whole training sequence. The frequency is highest at T2 and decreases substantially thereafter.

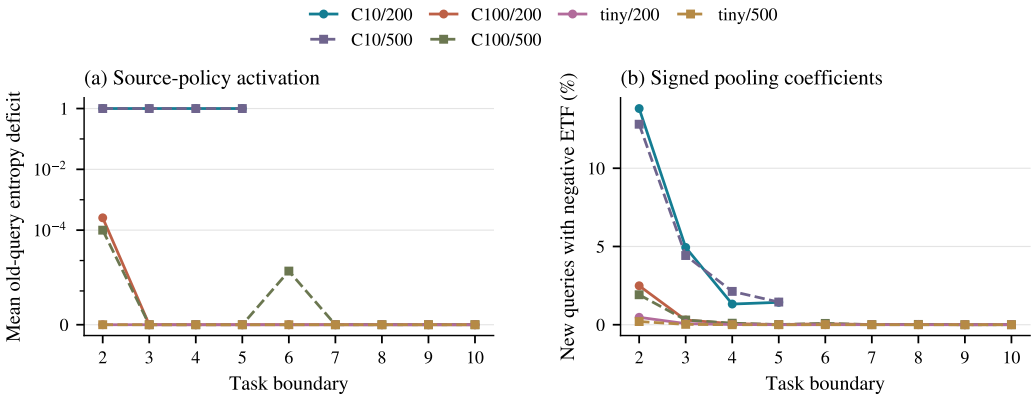


Figure 9: **Source policies and pooling operate in different regimes.** Left: mean entropy deficit on old queries (symmetric-log axis, linear below $10^{-6}$). Right: fraction of new queries with materially negative ETF channel coefficient. All epochs and batches are included, weighted by valid query counts.

### C.4 Pooling: Values and Gradient Paths

**Sharing values versus sharing gradients.** The own-ratio control preserves the pooled forward value but retains only the derivative through the query's own log-ratio inside the shared correction; derivatives through other same-class queries' ratios are stopped. Other query/key paths and positive peers remain differentiable. This costs 4.44, 3.95, 0.65, and 1.55 Class-IL points in the table's column order. Removing pooling entirely costs 0.14/0.58 points on CIFAR-10 but gains 0.44/0.01 on CIFAR-100. The own-ratio control is not gradient-norm matched.

Figure 10 evaluates all three objectives on the same frozen minibatch from the second CIFAR-10/200 task. We show the most frequent new class, with ties broken by ascending label, and retain batch order. The batch is frozen. The full and own-ratio objectives have identical forward values.

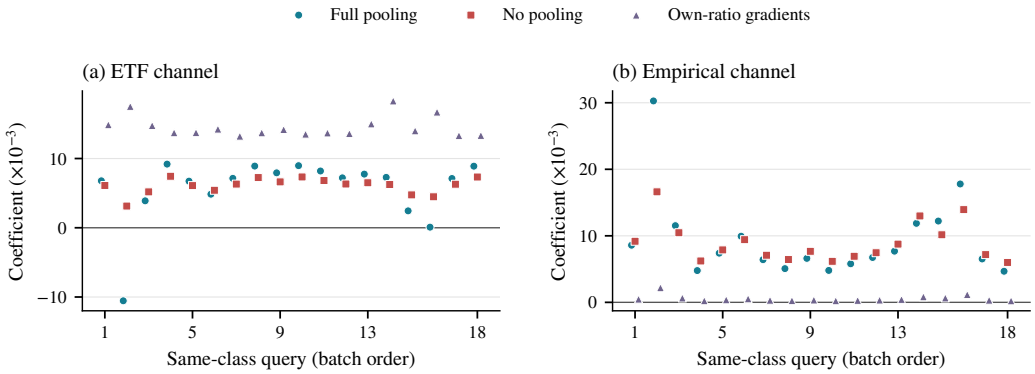


Figure 10: **Pooling redistributes negative demand on a common state.** Signed coefficients of $\log P_i$ and $\log E_i$, including the batch-average factor. Each position is one query, evaluated under three objectives on the same frozen batch. Full pooling can assign a negative ETF coefficient. These scalar partition derivatives do not include the full gradient through the features and model parameters.

### C.5 Transport Alternatives

The normalized-affine control uses $\nu(\alpha \boldsymbol{a}_i + \boldsymbol{b})$, without a fitted residual rotation. Like Möbius, it aligns inherited prototypes and preserves old-label rankings, but it does not in general preserve local angles. Its Class-IL changes relative to the full method are $+0.55, +0.25, +0.05, -0.38, -0.40, +0.71$ on CIFAR-10/200, CIFAR-10/500, CIFAR-100/200, CIFAR-100/500, TinyImageNet/200 and TinyImageNet/500, respectively. Tables 7 and 8 report the absolute scores. These paired comparisons do not establish a uniform accuracy advantage for either transport. Möbius provides a conformal spherical continuation; the results do not establish an accuracy benefit attributable to conformality alone.

The transport is fixed across datasets and replay budgets.

### C.6 Geometric Illustrations

Figure 11 connects the affine prototype update to spherical transport in a low-dimensional CIFAR-10/200 experiment with $d = 3$. The transition adds two classes to a two-class ETF, so $\alpha = \sqrt{2/3}$ and $\|\boldsymbol{b}\| = 1/\sqrt{3}$. This illustration uses symmetric vertex lifts in the minimal dimension $d = K - 1$; the nested orthonormal-basis implementation used in the main experiments requires $d \geq K$. We use the same 120 task-1 buffer features in all panels, selecting the first 60 stored examples of each class in buffer order. The second panel updates only the classifier; the third transports the frozen features with Equation (5). A common rigid rotation is used throughout, without PCA. The experiment covers the first two tasks supported by the four-class ETF.

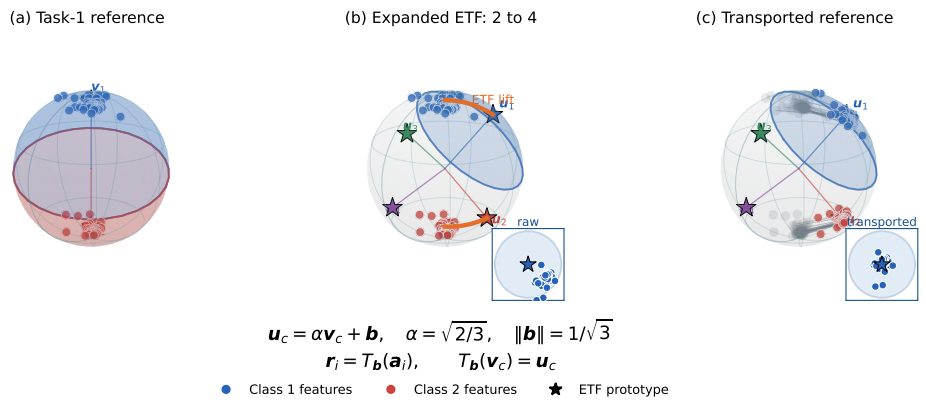


Figure 11: **From ETF expansion to transported references.** CIFAR-10/200, $d = 3$, $2 \to 4$ classes. (a) Frozen task-1 features and prototypes. (b) The expanded ETF with those features unchanged. (c) The same features after Möbius transport. Caps are centered at ETF vertices with half the nearest-prototype angular separation; only the blue cap is shown in (b,c). The insets show class 1 under the expanded ETF.

**Synthetic reference construction.** Figure 12 complements the checkpoint visualization with a single synthetic $2 \to 3$ expansion. It separates exact prototype alignment, the affine squared-norm defect, and preservation of old-only rankings from the decision against new classes. The corresponding identities are derived in Appendix D.1.2 and proved in Appendix D; neither visualization is a global-accuracy guarantee.

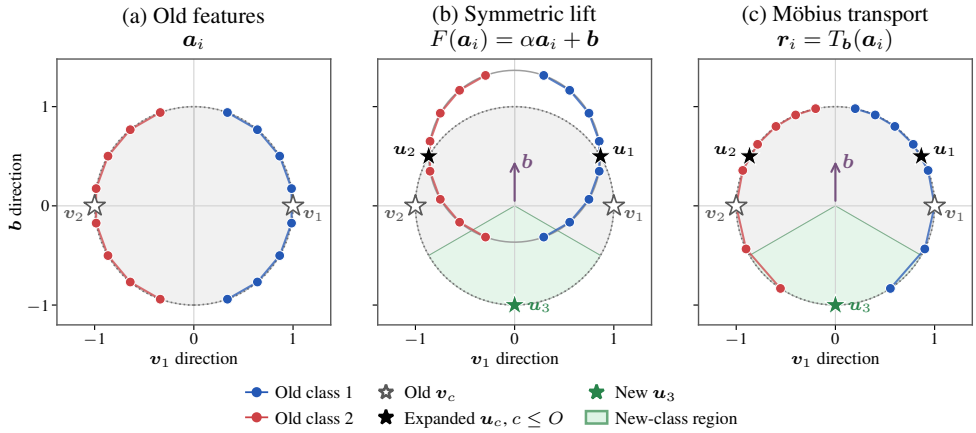


Figure 12: **From symmetric lift to spherical feature transport.** Synthetic $2 \to 3$ extension in a two-dimensional slice of $\mathbb{R}^3$, with $\alpha = \sqrt{3}/2$ and $\|b\| = 1/2$. The same old-class features are shown (a) before expansion, (b) under the affine continuation, which matches inherited vertices but leaves the sphere, and (c) after Möbius transport. Green marks the region won by the new class: preserved old-class rankings do not imply preserved global predictions. The illustration isolates a single geometric expansion step; no model is trained through a task.

### C.7 Post-Hoc Classifier Evaluation

We assess whether CoMove's advantage extends to a different inference rule, separating the quality of its learned representations from their alignment with the prescribed ETF classifier. We compare LwF (Li & Hoiem, 2018), ProNC (Wang et al., 2026), and CoMove with LDC (Gomez-Villa

et al., 2024). For ProNC and CoMove, this evaluates the effect of replacing the ETF classifier with compensated class means while keeping the encoder frozen.

Table 10: **Post-hoc classifier evaluation with LDC.** FAA (↑) and FF (↓). Mean ± sample standard deviation. Dashes denote unavailable results.

**(a) Seq-CIFAR-10**

| Method | Class-IL | | | | | | Task-IL | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $M=0$ | | $M=200$ | | $M=500$ | | $M=0$ | | $M=200$ | | $M=500$ | |
| | FAA↑ | FF↓ | FAA↑ | FF↓ | FAA↑ | FF↓ | FAA↑ | FF↓ | FAA↑ | FF↓ | FAA↑ | FF↓ |
| LwF + LDC | 46.72 | **34.60** | – | – | – | – | 88.53 | 7.26 | – | – | – | – |
| ProNC + LDC | 45.60±1.78 | 42.97±0.94 | 56.21±1.15 | **30.53**±1.51 | 57.68±3.06 | **28.26**±1.78 | **94.21**±0.96 | **3.85**±1.23 | **95.69**±0.54 | **1.72**±0.75 | 94.84±1.70 | **1.58**±0.11 |
| Ours + LDC | **55.56**±0.11 | 43.26±0.79 | **61.11**±0.37 | 35.92±0.49 | **63.50**±0.88 | 32.80±1.37 | 92.52±0.46 | 5.98±0.50 | 95.06±0.54 | 2.85±0.59 | **95.59**±0.24 | 2.15±0.23 |

**(b) Seq-CIFAR-100**

| Method | Class-IL | | | | | | Task-IL | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $M=0$ | | $M=200$ | | $M=500$ | | $M=0$ | | $M=200$ | | $M=500$ | |
| | FAA↑ | FF↓ | FAA↑ | FF↓ | FAA↑ | FF↓ | FAA↑ | FF↓ | FAA↑ | FF↓ | FAA↑ | FF↓ |
| LwF + LDC | 24.98 | 38.96 | – | – | – | – | 66.65 | 9.41 | – | – | – | – |
| ProNC + LDC | 33.51±0.41 | 44.76±0.35 | 38.20±0.60 | 38.09±0.85 | 40.87±0.33 | 34.19±0.54 | 82.30±0.14 | 7.82±0.18 | 84.40±0.20 | 5.43±0.12 | 85.00±0.13 | 4.61±0.19 |
| Ours + LDC | **45.60**±0.26 | **26.50**±0.54 | **48.18**±0.63 | **23.08**±0.60 | **49.19**±0.22 | **21.93**±0.32 | **84.51**±0.20 | **2.44**±0.13 | **85.09**±0.44 | **2.16**±0.23 | **85.44**±0.45 | **1.83**±0.27 |

**(c) Seq-TinyImageNet**

| Method | Class-IL | | | | | | Task-IL | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $M=0$ | | $M=200$ | | $M=500$ | | $M=0$ | | $M=200$ | | $M=500$ | |
| | FAA↑ | FF↓ | FAA↑ | FF↓ | FAA↑ | FF↓ | FAA↑ | FF↓ | FAA↑ | FF↓ | FAA↑ | FF↓ |
| LwF + LDC | 21.59 | 38.14 | – | – | – | – | 57.51 | 11.88 | – | – | – | – |
| ProNC + LDC | 27.53±0.35 | 34.96±0.15 | 28.02±0.35 | 34.09±0.61 | 28.22±0.49 | 33.21±0.55 | 67.77±0.19 | 10.67±0.23 | 68.06±0.28 | 10.35±0.21 | 68.17±0.45 | 10.00±0.47 |
| Ours + LDC | **37.94**±0.26 | **16.56**±0.09 | **38.42**±0.09 | **14.85**±0.41 | **39.10**±0.46 | **13.77**±0.67 | **72.82**±0.31 | **0.66**±0.31 | **71.59**±0.11 | **1.10**±0.13 | **72.20**±0.19 | **0.86**±0.17 |

For ProNC and CoMove, we fit a bias-free ridge map on current-task paired features, transport old class means, and use nearest compensated means instead of the ETF, without updating the encoder. LwF uses the supervised LDC recipe (Gomez-Villa et al., 2024). Table 10 compares the resulting classifiers.

**The representation advantage persists under LDC.** CoMove achieves the highest Class-IL FAA in all nine dataset–buffer settings. Without replay, its margins over ProNC + LDC are 9.96, 12.09, and 10.41 points on CIFAR-10, CIFAR-100, and TinyImageNet, respectively. The advantage therefore extends to evaluation with compensated class means.

**Classifier replacement changes cross-task discrimination.** Without replay, LDC raises ProNC Class-IL FAA on CIFAR-100 from 24.38% to 33.51%, while Task-IL FAA falls from 83.58% to 82.30%. This contrast points to improved discrimination across task blocks, despite slightly weaker discrimination within each block. The benefit depends on the dataset and memory budget: with replay, native ETF inference gives higher Class-IL FAA on CIFAR for both ProNC and CoMove, whereas LDC improves their TinyImageNet results. On the evaluated ProNC encoders, replacing ETF inference with LDC raises TinyImageNet FAA from 25.77% to 28.02% at buffer 200 and from 26.62% to 28.22% at buffer 500.

**Higher final accuracy can coexist with greater forgetting.** For CoMove on TinyImageNet, LDC raises Class-IL FAA from 33.80% to 38.42% at buffer 200 and from 35.22% to 39.10% at buffer 500. Class-IL FF also rises, from 7.40 to 14.85 and from 5.47 to 13.77, respectively. The two metrics capture different aspects of the trajectory: FAA measures final performance, while FF measures the decline from each task's previous peak. LDC offers a useful alternative inference rule, with a different accuracy–forgetting trade-off from the native ETF.

# D PROOFS

We first recall the required geometric preliminaries, then prove the reference-transport results of Section 4 and the objective identities used in Sections 4.2 and 4.3.

## D.1 TECHNICAL PRELIMINARIES

### D.1.1 NEURAL-COLLAPSE BACKGROUND

For completeness, we recall NC1–NC4 (Papyan et al., 2020). Consider balanced training data with $n$ examples per class and features $\boldsymbol{g}_{c,i} \in \mathbb{R}^d$. Write

$$\boldsymbol{\mu}_c = \frac{1}{n}\sum_i \boldsymbol{g}_{c,i}, \qquad \boldsymbol{\mu}_{\mathrm{G}} = \frac{1}{K}\sum_c \boldsymbol{\mu}_c, \qquad \boldsymbol{m}_c = \boldsymbol{\mu}_c - \boldsymbol{\mu}_{\mathrm{G}}.$$

These are empirical means, distinct from the prescribed ETF vertices.

**NC1: within-class variability collapse.** Within-class variation becomes negligible relative to between-class scale:

$$\frac{\mathrm{tr}(\boldsymbol{\Sigma}_{\mathrm{W}})}{\mathrm{tr}(\boldsymbol{\Sigma}_{\mathrm{B}})} \longrightarrow 0, \quad \boldsymbol{\Sigma}_{\mathrm{W}} = \frac{1}{Kn}\sum_{c,i}(\boldsymbol{g}_{c,i} - \boldsymbol{\mu}_c)(\boldsymbol{g}_{c,i} - \boldsymbol{\mu}_c)^\top, \quad \boldsymbol{\Sigma}_{\mathrm{B}} = \frac{1}{K}\sum_c \boldsymbol{m}_c\boldsymbol{m}_c^\top.$$

**NC2: simplex ETF geometry.** The centered class means approach equal lengths and equiangular directions:

$$\frac{\|\boldsymbol{m}_c\|}{\|\boldsymbol{m}_\ell\|} \longrightarrow 1, \qquad \frac{\boldsymbol{m}_c^\top \boldsymbol{m}_\ell}{\|\boldsymbol{m}_c\|\|\boldsymbol{m}_\ell\|} \longrightarrow -\frac{1}{K-1} \quad (c \neq \ell).$$

**NC3: self-duality.** The learned classifier weights approach the centered class means up to a common scale. For matrices $\boldsymbol{V} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K]$ and $\boldsymbol{M} = [\boldsymbol{m}_1, \ldots, \boldsymbol{m}_K]$,

$$\left\| \frac{\boldsymbol{V}}{\|\boldsymbol{V}\|_{\mathrm{F}}} - \frac{\boldsymbol{M}}{\|\boldsymbol{M}\|_{\mathrm{F}}} \right\|_{\mathrm{F}} \longrightarrow 0.$$

**NC4: nearest-class-center decisions.** Classifier decisions approach the nearest-class-center rule $\arg\min_c \|\boldsymbol{g} - \boldsymbol{\mu}_c\|^2$. For equal-length centered means, this rule is equivalent to maximizing $(\boldsymbol{g} - \boldsymbol{\mu}_{\mathrm{G}})^\top \boldsymbol{m}_c / \|\boldsymbol{m}_c\|$.

These properties describe an asymptotic training regime. Our construction fixes the classifier geometry; feature concentration, feature–vertex alignment, and held-out generalization depend on the learned representations. We make no neural-collapse assumption about the finite continual-learning features used in the experiments.

### D.1.2 EXPANSION GEOMETRY AND TARGET COMPATIBILITY

Nested basis expansion yields the symmetric lift in Section 4.1. The identities below give its accumulated displacement and the resulting mismatch between targets.

**Nested ETF construction.** Let $\boldsymbol{Q}_K = [\boldsymbol{e}_1, \ldots, \boldsymbol{e}_K]$ have orthonormal columns, and let $\boldsymbol{H}_K = \boldsymbol{I}_K - K^{-1}\mathbf{1}_K\mathbf{1}_K^\top$. Here $\boldsymbol{I}_K$ is the identity matrix and $\mathbf{1}_K$ the all-ones vector in $\mathbb{R}^K$. The progressive ETF is

$$\boldsymbol{U}_K = \sqrt{\frac{K}{K-1}}\,\boldsymbol{Q}_K\boldsymbol{H}_K. \tag{35}$$

The basis update keeps inherited columns:

$$\boldsymbol{Q}_K = [\boldsymbol{Q}_O\ \boldsymbol{Q}_{\mathrm{new}}], \qquad \boldsymbol{Q}_O^\top\boldsymbol{Q}_{\mathrm{new}} = \mathbf{0}_{O\times q}, \qquad \boldsymbol{Q}_{\mathrm{new}}^\top\boldsymbol{Q}_{\mathrm{new}} = \boldsymbol{I}_q. \tag{36}$$

This is the nested expansion of ProNC (Wang et al., 2026). It requires $d \geq K$; the rank of the resulting simplex is only $K-1$. The symmetric lift in Equation (3) uses $\boldsymbol{n}_m = \sqrt{m/(m+1)}(\boldsymbol{e}_{m+1} - m^{-1}\sum_{j\leq m}\boldsymbol{e}_j)$. It is unit and orthogonal to the centered old simplex. Thus preserving basis columns does not preserve the centered, rescaled ETF vertices.

**Dimension and exactness of a single lift.** An $m$-class simplex spans $m-1$ dimensions, so an orthogonal lift axis exists for $d \geq m$. Its old–new inner products must equal $-1/m$, fixing the common shift to $-\boldsymbol{n}_m/m$. Unit norm then gives $a_m^2 + m^{-2} = 1$; choosing $a_m > 0$ yields $a_m = \sqrt{1-m^{-2}}$. The inherited pairwise products become $-a_m^2/(m-1) + m^{-2} = -1/m$, establishing exact ETF geometry. Each inherited vertex moves through angle $\arccos a_m$.

**Chained state.** At intermediate size $m$, write $\boldsymbol{v}_c^{(m)} = \alpha_m \boldsymbol{v}_c + \boldsymbol{b}_m$ for $c \leq O$. With $(\alpha_O, \boldsymbol{b}_O) = (1, \boldsymbol{0})$,

$$\alpha_{m+1} = a_m \alpha_m, \qquad \boldsymbol{b}_{m+1} = a_m \boldsymbol{b}_m - \boldsymbol{n}_m/m. \tag{37}$$

Each new lift contracts the accumulated shift before adding its own. Consequently,

$$\boldsymbol{b} = -\sum_{m=O}^{K-1} \frac{\boldsymbol{n}_m}{m} \prod_{\ell=m+1}^{K-1} a_\ell = \frac{1}{O} \sum_{c \leq O} \boldsymbol{u}_c. \tag{38}$$

Empty products equal one. The common endpoint displacement satisfies

$$\|\boldsymbol{b}\|^2 = 1 - \alpha^2 = \frac{K-O}{O(K-1)}, \qquad \phi_{\mathrm{sym}}(O \to K) = \arccos \alpha. \tag{39}$$

The full derivation is Section D.2. The minimum total squared chordal displacement over exact expanded ETF orientations is established in Section D.10.

**Alignment and distillation targets.** For unit directions $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^{d-1}$, neural-collapse-inspired alignment may use dot regression (Yang et al., 2022; 2023b),

$$\varphi(\boldsymbol{x}, \boldsymbol{y}) = \tfrac{1}{2}(\boldsymbol{x}^\top \boldsymbol{y} - 1)^2. \tag{40}$$

Applying it to classifier and previous-feature references gives

$$\mathcal{L}_{\mathrm{align}} = B^{-1} \sum_i \varphi(\boldsymbol{z}_i, \boldsymbol{u}_{y_i}), \qquad \mathcal{L}_{\mathrm{distill}} = B^{-1} \sum_i \varphi(\boldsymbol{z}_i, \boldsymbol{a}_i). \tag{41}$$

These standard alignment objectives illustrate the target mismatch that motivates the transported reference in CoMove. CoMove uses the relation potential of Equation (10) for training.

**Proposition 3** (Expansion-induced target incompatibility)**.** *Under Theorem 1, an inherited class $c \leq O$ satisfies*

$$\boldsymbol{v}_c^\top \boldsymbol{u}_c = \alpha < 1, \qquad \|\boldsymbol{u}_c - \boldsymbol{v}_c\|^2 = 2(1-\alpha) > 0. \tag{42}$$

*For alignment losses $\varphi_{\mathrm{a}}$ and $\varphi_{\mathrm{t}}$ uniquely minimized at their unit reference,*

$$\arg\min_{\boldsymbol{z} \in \mathbb{S}^{d-1}} \varphi_{\mathrm{a}}(\boldsymbol{z}, \boldsymbol{u}_c) = \{\boldsymbol{u}_c\} \neq \{\boldsymbol{v}_c\} = \arg\min_{\boldsymbol{z} \in \mathbb{S}^{d-1}} \varphi_{\mathrm{t}}(\boldsymbol{z}, \boldsymbol{v}_c). \tag{43}$$

Orthogonality of $\boldsymbol{b}$ gives the inner product in Equation (42), and unit norms give its distance identity. Distinct references imply the distinct singleton minimizers. For dot regression, $\varphi(\boldsymbol{u}_c, \boldsymbol{v}_c) = \frac{1}{2}(1-\alpha)^2 > 0$. Squared Euclidean and cosine distance satisfy the same unique-minimum condition on the unit sphere.

**Residuals and decision scope.** Figure 12 illustrates the distinction between the affine vertex update and its spherical continuation on synthetic features. For a unit feature $\boldsymbol{x} = \boldsymbol{v}_c + \boldsymbol{\varepsilon}$, the unnormalized affine continuation satisfies

$$F(\boldsymbol{x}) = \boldsymbol{u}_c + \alpha \boldsymbol{\varepsilon}, \qquad \|F(\boldsymbol{x})\|^2 = 1 + 2\alpha \boldsymbol{b}^\top \boldsymbol{\varepsilon}. \tag{44}$$

Its unit-norm outputs correspond exactly to the equator containing inherited vertices when the shift is nonzero (Section D.3). The spherical transport gives

$$\arg\max_{c \leq O} \boldsymbol{u}_c^\top T_{\boldsymbol{b}}(\boldsymbol{x}) = \arg\max_{c \leq O} \boldsymbol{v}_c^\top \boldsymbol{x}, \tag{45}$$

including ties. Raw features also preserve old-only rankings under this specific expansion, since $\boldsymbol{x}^\top(\boldsymbol{u}_c - \boldsymbol{u}_\ell) = \alpha \boldsymbol{x}^\top(\boldsymbol{v}_c - \boldsymbol{v}_\ell)$. The transport additionally aligns the inherited prototypes exactly. Neither old-only ranking identity guarantees the winning label against new classes: the green region in Figure 12 makes that distinction explicit. Both identities hold for fixed references; they give no guarantee about retention after student optimization. We evaluate retention experimentally.

### D.2 Proof of Theorem 1

**Claim.** Theorem 1, including the accumulated shift and its norm.

Write the columns of the nested basis as $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_K$ and $\bar{\boldsymbol{e}}_m = m^{-1} \sum_{j \le m} \boldsymbol{e}_j$. From Equation (35), the old and new prototypes satisfy

$$\begin{aligned} \boldsymbol{v}_c &= \sqrt{\frac{O}{O-1}}(\boldsymbol{e}_c - \bar{\boldsymbol{e}}_O), \\ \boldsymbol{u}_c &= \sqrt{\frac{K}{K-1}}(\boldsymbol{e}_c - \bar{\boldsymbol{e}}_K) = \alpha \boldsymbol{v}_c + \boldsymbol{b}, \\ \boldsymbol{b} &= \sqrt{\frac{K}{K-1}}(\bar{\boldsymbol{e}}_O - \bar{\boldsymbol{e}}_K). \end{aligned}$$

Matching the coefficient of $\boldsymbol{e}_c$ yields the stated $\alpha$. Every old centered direction is orthogonal to the old column mean and to all new columns. Therefore $\boldsymbol{b}$ is orthogonal to the old simplex span. Since the old prototypes sum to the zero vector, averaging their new positions gives $\boldsymbol{b} = O^{-1} \sum_{c \le O} \boldsymbol{u}_c$. Orthonormality of the basis also gives

$$\|\boldsymbol{b}\|^2 = \frac{K}{K-1}\left(\frac{1}{O} - \frac{1}{K}\right) = \frac{K-O}{O(K-1)} = 1 - \alpha^2. \tag{46}$$

**Connection to the one-class chain.** For an intermediate size $m$, set $\boldsymbol{n}_m = \sqrt{m/(m+1)}(\boldsymbol{e}_{m+1} - \bar{\boldsymbol{e}}_m)$. This is a unit vector orthogonal to the current centered simplex span. Substitution into the centered-basis formula yields

$$\boldsymbol{v}_c^{(m+1)} = a_m \boldsymbol{v}_c^{(m)} - \frac{\boldsymbol{n}_m}{m}, \qquad a_m = \sqrt{1 - m^{-2}},$$

with endpoints $\boldsymbol{v}_c^{(O)} = \boldsymbol{v}_c$ and $\boldsymbol{v}_c^{(K)} = \boldsymbol{u}_c$. Repeated substitution, or induction on $m$, gives

$$\alpha = \prod_{m=O}^{K-1} a_m, \qquad \boldsymbol{b} = -\sum_{m=O}^{K-1} \frac{\boldsymbol{n}_m}{m} \prod_{\ell=m+1}^{K-1} a_\ell. \tag{47}$$

Indeed, each further step multiplies the existing coefficients by $a_m$ and adds its new displacement. The squared contraction telescopes:

$$\alpha^2 = \prod_{m=O}^{K-1} \frac{(m-1)(m+1)}{m^2} = \frac{O-1}{K-1}\frac{K}{O}.$$

Thus the direct formula gives the endpoint of the nested-basis construction. □

### D.3 The affine feature squared-norm defect

**Claim.** Equation (44) and the scope of affine spherical exactness.

Take any unit direction $\boldsymbol{x} = \boldsymbol{v}_c + \boldsymbol{\varepsilon}$. By Theorem 1, $F(\boldsymbol{x}) = \alpha \boldsymbol{x} + \boldsymbol{b} = \boldsymbol{u}_c + \alpha \boldsymbol{\varepsilon}$. Expanding its norm and using Equation (46) gives

$$\begin{aligned} \|F(\boldsymbol{x})\|^2 &= \alpha^2 \|\boldsymbol{x}\|^2 + \|\boldsymbol{b}\|^2 + 2\alpha \boldsymbol{b}^\top \boldsymbol{x} \\ &= 1 + 2\alpha \boldsymbol{b}^\top \boldsymbol{\varepsilon}, \end{aligned}$$

because $\boldsymbol{b}^\top \boldsymbol{v}_c = 0$. Thus, for a nonzero shift, $F$ is unit-norm exactly when $\boldsymbol{b}^\top \boldsymbol{x} = 0$. The preserved set is the equator orthogonal to $\boldsymbol{b}$ and contains the old prototype sphere. For example, $\boldsymbol{x} = \boldsymbol{b}/\|\boldsymbol{b}\|$ has squared-norm defect $\|F(\boldsymbol{x})\|^2 - 1 = 2\alpha\|\boldsymbol{b}\| > 0$. At zero shift $F$ is the identity. □

### D.4 Spherical well-posedness, inverse and conformality

**Claim.** The map in Equation (5) is spherical, globally defined, invertible, and obeys the distance identity below.

**Axial construction.** For $\boldsymbol{b} \neq \mathbf{0}$, let $\tau = \|\boldsymbol{b}\|$, $\boldsymbol{e} = \boldsymbol{b}/\tau$, and decompose a unit direction as $\boldsymbol{x} = s\boldsymbol{e}+\boldsymbol{w}$ with $\boldsymbol{w} \perp \boldsymbol{e}$. On the equator, $s = 0$, the desired image is $\tau\boldsymbol{e}+\alpha\boldsymbol{w}$. To preserve local angles as well, we use an axial Möbius transformation. Its fractional-linear axial map is $s \mapsto (s+\tau)/(1+\tau s)$, which sends 0 to $\tau$ and fixes the poles $s = \pm 1$. Keeping the transverse direction unchanged, unit norm then determines its positive scale:

$$T_{\boldsymbol{b}}(s\boldsymbol{e} + \boldsymbol{w}) = \frac{s+\tau}{1+\tau s}\boldsymbol{e} + \frac{\alpha}{1+\tau s}\boldsymbol{w}, \qquad \alpha = \sqrt{1-\tau^2}. \tag{48}$$

Indeed, $(s+\tau)^2 + (1-\tau^2)(1-s^2) = (1+\tau s)^2$. Rewriting Equation (48) without axial coordinates gives Equation (5). The zero-displacement case extends as the identity.

**Well-posedness and geometric properties.** For $\boldsymbol{b} \neq \mathbf{0}$, let $\tau = \|\boldsymbol{b}\| < 1$, $\boldsymbol{e} = \boldsymbol{b}/\tau$, and decompose a unit input as $\boldsymbol{x} = s\boldsymbol{e} + \boldsymbol{w}$, with $\boldsymbol{w} \perp \boldsymbol{e}$ and $s^2 + \|\boldsymbol{w}\|^2 = 1$. The map becomes

$$T_{\boldsymbol{b}}(\boldsymbol{x}) = \frac{s+\tau}{1+\tau s}\,\boldsymbol{e} + \frac{\alpha}{1+\tau s}\,\boldsymbol{w}.$$

The denominator is at least $1 - \tau > 0$. Since $(s+\tau)^2 + (1-\tau^2)(1-s^2) = (1+\tau s)^2$, the output has unit norm. Replacing $\tau$ by $-\tau$ in the same axial coordinates recovers both $s$ and $\boldsymbol{w}$, establishing $T_{-\boldsymbol{b}} \circ T_{\boldsymbol{b}} = \mathrm{id}$. At $\boldsymbol{b} = \mathbf{0}$ all statements reduce to the identity map.

The squared chordal-distance identity is

$$\|T_{\boldsymbol{b}}(\boldsymbol{x}) - T_{\boldsymbol{b}}(\boldsymbol{y})\|^2 = \frac{\alpha^2\|\boldsymbol{x}-\boldsymbol{y}\|^2}{(1+\boldsymbol{b}^\top\boldsymbol{x})(1+\boldsymbol{b}^\top\boldsymbol{y})}. \tag{49}$$

Write $\boldsymbol{y} = r\boldsymbol{e} + \boldsymbol{v}$ in the same axial coordinates. The numerator of one minus the transported inner product is

$$\begin{aligned}(1+\tau s)(1+\tau r) - (s+\tau)(r+\tau) - (1-\tau^2)\boldsymbol{w}^\top\boldsymbol{v} \\ = (1-\tau^2)(1 - sr - \boldsymbol{w}^\top\boldsymbol{v}) = \alpha^2(1 - \boldsymbol{x}^\top\boldsymbol{y}).\end{aligned}$$

Dividing by the two denominators and using $\|\boldsymbol{x}-\boldsymbol{y}\|^2 = 2(1-\boldsymbol{x}^\top\boldsymbol{y})$ proves Equation (49). For a tangent perturbation, divide this identity by the squared perturbation length and pass to the limit. Every tangent direction is scaled by $\alpha/(1+\boldsymbol{b}^\top\boldsymbol{x})$. Polarization shows that tangent angles are preserved. □

### D.5 Proof of Theorem 2

**Claim.** Theorem 2.

For an old prototype, $\boldsymbol{b}^\top\boldsymbol{v}_c = 0$ by Theorem 1. Substitution into Equation (5) gives $T_{\boldsymbol{b}}(\boldsymbol{v}_c) = \alpha\boldsymbol{v}_c + \boldsymbol{b} = \boldsymbol{u}_c$. For two old labels, their expanded difference is

$$\boldsymbol{u}_c - \boldsymbol{u}_\ell = \alpha(\boldsymbol{v}_c - \boldsymbol{v}_\ell), \qquad \boldsymbol{b}^\top(\boldsymbol{v}_c - \boldsymbol{v}_\ell) = 0.$$

Taking the inner product with Equation (5) removes all numerator terms parallel to $\boldsymbol{b}$, leaving exactly Equation (6). The multiplier is strictly positive by $\alpha > 0$ and $1 + \boldsymbol{b}^\top\boldsymbol{x} \geq 1 - \|\boldsymbol{b}\| > 0$. Every sign and zero of an old-class score difference is therefore preserved. □

**Decision-level consequence of Theorem 2.** **Claim.** Theorem 2.

The nonzero-output assumption makes the frozen representation $\boldsymbol{a}(x_{\mathrm{old}})$ in Equation (1) well defined and unit norm. Apply Equation (6) with that representation. Every old-class logit gap is multiplied by the same positive factor $\alpha^2/(1 + \boldsymbol{b}^\top\boldsymbol{a}(x_{\mathrm{old}}))$. Hence an old label maximizes the old logits if and only if it maximizes the transported logits restricted to the same labels, proving Equation (45). Equality of maximizer sets also covers ties. A fixed common tie-breaking rule then gives identical predictions, and therefore identical accuracy on any fixed old-labelled evaluation set. □

**Scope of the guarantee.** The argument applies to any common subset of old labels, including a previous task block. It does not compare with new labels, or replace the frozen feature by the optimized student. It preserves incorrect old predictions too. Old-restricted softmax probabilities generally change: after discarding a common logit offset, transport replaces the scale $\gamma$ by $\gamma\alpha^2/(1+\boldsymbol{b}^\top\boldsymbol{a}(x_{\mathrm{old}}))$. Thus exact decisions do not imply probability preservation. These limitations remain even when the prototype relation is exact.

### D.6 Equivalent potential and supervision allocation

**Claim.** Equation (11) equals the evaluated source-weighted objective under the same detach and fallback policies.

The implementation evaluates the potential $\psi_i$ in Equation (10); its relation policies and peer averaging give Equations (19) and (20).

For an equivalent normalized-allocation form, define $\beta = m_{\mathrm{t}} = m_{\mathrm{a}}$, so that $m_{\mathrm{p}} = 1 - 2\beta$, and set

$$\ell_i^{\mathrm{sup}} = \frac{\beta \ell_i^{\mathrm{a}} + (1-2\beta)\ell_i^{\mathrm{p}}}{1-\beta}, \qquad \pi_i = \frac{w_i(1-\beta)\rho_i}{M_i}.$$

Since $M_i = w_i(1-\beta) + \beta$,

$$\begin{aligned} M_i[\pi_i \ell_i^{\mathrm{sup}} + (1-\pi_i)\ell_i^{\mathrm{t}}] = {} & w_i\beta\rho_i\ell_i^{\mathrm{a}} \\ & + w_i(1-2\beta)\rho_i\ell_i^{\mathrm{p}} \\ & + [\beta + w_i(1-\beta)(1-\rho_i)]\ell_i^{\mathrm{t}}. \end{aligned} \tag{50}$$

Grouping the relations in Equation (11) gives these weights and the coverage form in Equation (30). Their sum is $M_i$. Subtracting the expression at $\rho_i = 1$ proves Equation (31). Positive dataset counts give $M_i > 0$ and $0 \le \pi_i \le \rho_i \le 1$.

Factoring out $e^{-\lambda(s-s_i^0)}$ gives the equivalent margin form:

$$\psi_i(s;\kappa,\lambda) = \log\left(\kappa + e^{A_i - s + \lambda(s-s_i^0)}\right) - \lambda(s - s_i^0).$$

This identity includes $\kappa = 0$. Each rewrite has the same gradients when the same detached policies are held fixed. □

### D.7 Negative-demand redistribution

**Claim.** The shared margin induces the following exact partition differential, holding positive scores and detached policies fixed.

For a class $c$ represented in the batch, write $\mathcal{I}_c = \{i \in \{1,\dots,B\} : y_i = c\}$, so the query-indexed set in Equation (16) satisfies $\mathcal{B}_i = \mathcal{I}_{y_i}$. For $j \in \mathcal{I}_c$, define

$$G_i = \frac{\partial \mathcal{L}_{\mathrm{CoMove}}}{\partial A_i}, \qquad S_c = \sum_{i \in \mathcal{I}_c} G_i, \qquad \eta_j = \frac{\xi_j}{\sum_{i \in \mathcal{I}_c}(1+\xi_i)}.$$

Differentiating Equation (9) and using $\mathrm{d}\xi_j = \xi_j(\mathrm{d}\log E_j - \mathrm{d}\log P_j)$ gives

$$\mathrm{d}\mu_c = \sum_{j \in \mathcal{I}_c} \eta_j(\mathrm{d}\log E_j - \mathrm{d}\log P_j).$$

Write $\mathrm{d}_{\mathrm{neg}}\mathcal{L}_{\mathrm{CoMove}} := \sum_i G_i \mathrm{d}A_i$ for the differential through the negative partitions at fixed positive scores and policies. Substituting $\mathrm{d}A_i = \mathrm{d}\log P_i + \mathrm{d}\mu_{y_i}$ and collecting terms over represented classes gives

$$\mathrm{d}_{\mathrm{neg}}\mathcal{L}_{\mathrm{CoMove}} = \sum_c \sum_{j \in \mathcal{I}_c} [(G_j - \eta_j S_c)\mathrm{d}\log P_j + \eta_j S_c \mathrm{d}\log E_j]. \tag{51}$$

When the empirical support is empty, its mass is identically zero and its contribution is omitted; we do not differentiate $\log 0$. Source weights, peer averages and the batch factor are included in $G_i$. □

**Signed coefficients and full feature gradients.** The coefficient $G_j - \eta_j S_c$ can be negative; positivity of the forward partition does not forbid this. The two channel coefficients sum to $G_j$, but neither gradient norms nor per-query empirical pressure are conserved. To obtain feature gradients, compose this differential with every query and key path and the normalization Jacobian. The formula is not a convex average of unpooled feature gradients. A class with $|\mathcal{I}_c| = 1$ recovers the unpooled value and derivative, whereas equal ratios at a point need not imply equal gradients when query demands differ.

### D.8 Entropic aggregation and the binary affine subloss

**Claim.** The shared margin has an exact KL decomposition, but its interpretation as a loss difference is restricted to the identified affine terms.

Fix a represented class $c$ and let $\mathcal{I}_c = \{j \in \{1, \dots, B\} : y_j = c\}$. All sums below run over this set. Let $\chi_j = \log(1 + \xi_j)$, $Z = \sum_j e^{\chi_j}$, and $\varpi_j = e^{\chi_j}/Z$. Write $\boldsymbol{\varpi}_c = (\varpi_j)_{j \in \mathcal{I}_c}$ for this probability vector and $\mathsf{U}_c$ for the uniform law on the same set. The shared margin is $\mu_c = \log Z - \log |\mathcal{I}_c|$. Direct expansion gives

$$D_{\mathrm{KL}}(\mathsf{U}_c \| \boldsymbol{\varpi}_c) = \frac{1}{|\mathcal{I}_c|} \sum_j [-\log |\mathcal{I}_c| - \chi_j + \log Z] = \mu_c - \frac{1}{|\mathcal{I}_c|} \sum_j \chi_j.$$

For any probability vector $\boldsymbol{p} = (p_j)_{j \in \mathcal{I}_c}$, the same expansion gives

$$\sum_j p_j \chi_j - D_{\mathrm{KL}}(\boldsymbol{p} \| \mathsf{U}_c) = \mu_c - D_{\mathrm{KL}}(\boldsymbol{p} \| \boldsymbol{\varpi}_c) \le \mu_c,$$

with equality at $\boldsymbol{p} = \boldsymbol{\varpi}_c$. This proves both standard logsumexp identities.

Every query in $\mathcal{I}_c$ has weight $w_i = N/(Kn_c)$, where $n_c$ counts class-$c$ examples in the merged dataset. For $q = 2$, the entropy deficit is one. On new classes, anchor and temporal losses are affine in the partition, with total coefficient $m_{\mathrm{t}}(N/(Kn_c) + 1)$ per query. Replacing each local $\chi_j$ by $\mu_c$ therefore changes their class-summed contribution to the batch-averaged objective by exactly

$$\frac{|\mathcal{I}_c| m_{\mathrm{t}}}{B} \left( \frac{N}{Kn_c} + 1 \right) D_{\mathrm{KL}}(\mathsf{U}_c \| \boldsymbol{\varpi}_c). \tag{52}$$

Peer softplus terms and old temporal terms are not covered by this statement. For $|\mathcal{I}_c| = 1$, the KL is zero and the fallback also coincides. No unbiasedness or variance-reduction guarantee follows. □

### D.9 Temporal score attraction

**Claim.** For an old temporal edge, the negative responsibility and positive attraction need not have the same magnitude.

Fix query $i$ and the policy coefficients. For an old temporal relation, $\kappa = 1$; write $\lambda = \lambda_i$. Let $Q = e^{-\lambda(s - s_i^0)}$, $V = e^{A_i - s}$, and $p_{\mathrm{N}} = V/(Q + V)$. Differentiating $\psi_i(s; 1, \lambda) = \log(Q + V)$ yields

$$\partial_{A_i} \psi_i = p_{\mathrm{N}}, \qquad -\partial_s \psi_i = \lambda(1 - p_{\mathrm{N}}) + p_{\mathrm{N}}.$$

Since $p_{\mathrm{N}} \in [0, 1]$, the positive-score attraction lies between $\min(1, \lambda)$ and $\max(1, \lambda)$; it equals one when $\lambda = 1$. This partial derivative holds the partition fixed; the full feature gradient also differentiates normalization and shared partitions. □

### D.10 Endpoint drift and chordal optimality

**Claim.** Equations (39) and (53).

For any exact $K$-class ETF with a fixed correspondence for the first $O$ old labels, the total squared chordal displacement satisfies

$$D_{\mathrm{ch}} := \sum_{c=1}^{O} \|\widetilde{\boldsymbol{u}}_c - \boldsymbol{v}_c\|^2 \ge 2O(1 - \alpha). \tag{53}$$

The symmetric construction attains equality for this squared chordal objective. The bound says nothing about the sum of angular displacements, and the teacher-transport construction does not depend on chordal optimality.

For the symmetric lift, Theorem 1 gives $\boldsymbol{u}_c = \alpha \boldsymbol{v}_c + \boldsymbol{b}$ with $\boldsymbol{b} \perp \boldsymbol{v}_c$. Hence $\boldsymbol{v}_c^\top \boldsymbol{u}_c = \alpha$, so the endpoint angle is $\arccos \alpha$ and the squared chordal displacement is $2(1 - \alpha)$. Their totals over old labels are respectively $O \arccos \alpha$ and $2O(1 - \alpha)$. The one-class specialization is $\arccos \sqrt{1 - 1/m^2}$. These are displacements between endpoints; summing intermediate angles generally measures a different quantity.

**A lower bound over all expanded ETF orientations.** Consider any exact $K$-class ETF with inherited matched vertices $\widetilde{\boldsymbol{u}}_1, \ldots, \widetilde{\boldsymbol{u}}_O$. Let $\bar{\boldsymbol{u}} = O^{-1} \sum_{c \le O} \widetilde{\boldsymbol{u}}_c$. Its Gram matrix implies

$$\|\bar{\boldsymbol{u}}\|^2 = \frac{K - O}{O(K-1)}, \qquad \|\widetilde{\boldsymbol{u}}_c - \bar{\boldsymbol{u}}\|^2 = \frac{K(O-1)}{O(K-1)} = \alpha^2.$$

Set $\boldsymbol{w}_c = (\widetilde{\boldsymbol{u}}_c - \bar{\boldsymbol{u}})/\alpha$, which has unit norm. Using $\sum_c \boldsymbol{v}_c = \mathbf{0}$ and the unit-vector bound gives

$$\sum_{c \le O} \widetilde{\boldsymbol{u}}_c^\top \boldsymbol{v}_c = \alpha \sum_{c \le O} \boldsymbol{w}_c^\top \boldsymbol{v}_c \le O\alpha.$$

Expanding the squared distances proves Equation (53). Equality is attained by the symmetric construction, where $\boldsymbol{w}_c = \boldsymbol{v}_c$. Thus the positive total movement is unavoidable, but its distribution across individual classes need not be symmetric for other orientations. No minimizer of the angular sum follows from this argument. □