FROM NEURAL COLLAPSE GEOMETRY TO STRUCTURAL MEMORY: PRESERVING DECISION STRUCTURE IN CONTINUAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural collapse (NC) reveals a structured geometry that links sample representations, class prototypes, and classification decisions, providing a unified perspective on how class knowledge can be organized across learning stages. Existing NC-based continual learning methods adopt fixed or progressively expanded simplex equiangular tight frames (ETFs) as learning targets. However, specifying such a geometric target does not by itself ensure that the induced discriminative structure persists across stages; in particular, globally expanding an ETF shifts prototypes that have already been established for old classes. We propose **Structure-Preserving Continual Learning (SPCL)**, which treats the NC-induced decision structure as structural memory—the class reference, the coordinate response, and the decision margin of the established classes—and preserves it across stages. SPCL freezes established prototypes and embeds each new task's simplex block in the orthogonal complement of the old span; distills each sample's response distribution over the frozen prototypes; and penalizes old-class margin decreases on replayed samples in the post-plasticity window, with mixed-batch feature distillation stabilizing the underlying representation. These mechanisms preserve the class references and sample-level relations anchored to them, letting structural memory persist as the label space expands. Experiments on Seq-CIFAR-100 with a total replay buffer capacity of 500 samples show that SPCL improves final average accuracy while reducing average forgetting by approximately **70%** relative to ProNC under matched protocols.

1 INTRODUCTION

Continual learning (CL) studies how a model can learn from a sequence of tasks without losing knowledge acquired from earlier ones; the core difficulty is catastrophic forgetting (McCloskey & Cohen, 1989; Kirkpatrick et al., 2017). In class-incremental learning (CIL), this challenge is particularly severe: at inference time, the model receives no task identity and must discriminate among all classes observed so far. Replay and distillation mitigate forgetting by reusing buffered examples or constraining the current model to match a previous one (Rebuffi et al., 2017; Rolnick et al., 2019; Buzzega et al., 2020; Li & Hoiem, 2017). Yet these mechanisms mainly regularize data, features, logits, or parameters. They do not directly answer a more structural question: as the label space expands, what is the decision structure that old knowledge should preserve?

Neural collapse (NC) provides a promising geometric lens for this question. In well-trained classifiers, within-class features collapse toward class means, class means align with classifier weights, and the resulting class prototypes form a simplex equiangular tight frame (ETF), under which prediction reduces to a nearest-prototype rule (Papyan et al., 2020; Zhu et al., 2021). This geometry turns class knowledge into an explicit object—a set of class references and the sample positions relative to them. Recent NC-based continual learning methods therefore use ETF vertices as classification targets (fig. 1a). Fixed-ETF methods pre-assign a global ETF for all classes (Yang et al., 2023a;b), while Progressive Neural Collapse (ProNC) expands the ETF target as new classes arrive (Wang et al., 2026).

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

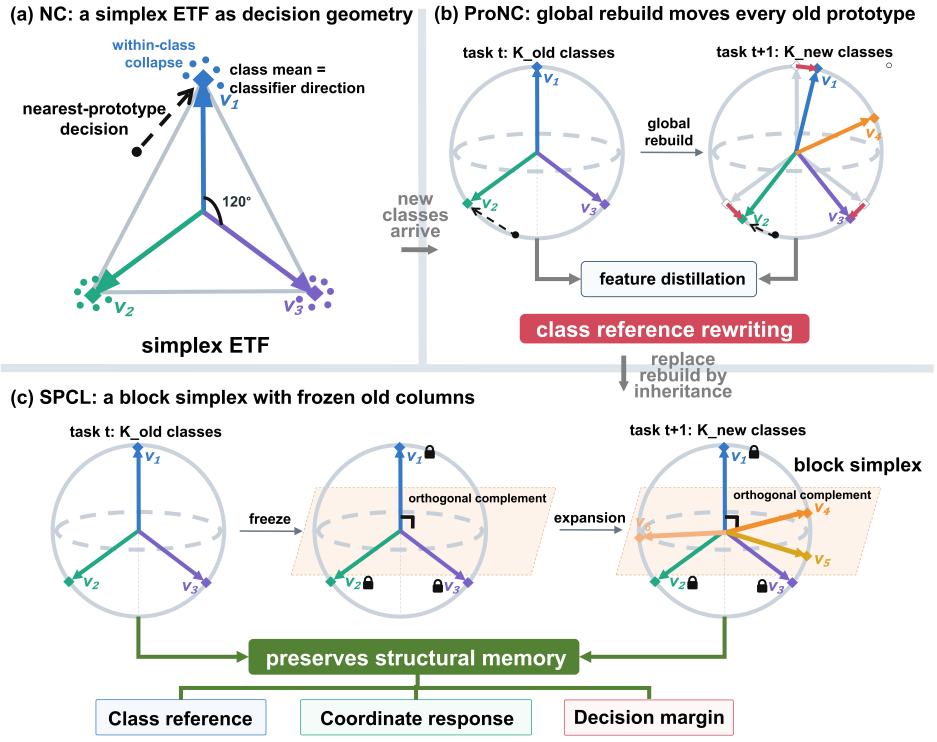


Figure 1: **From neural-collapse geometry to structural memory.** **(a)** Under neural collapse, within-class features concentrate at their class means and the class prototypes form a simplex ETF on the unit sphere; prediction reduces to a nearest-prototype query. **(b)** ProNC expands by rebuilding the global simplex ETF at every task boundary: the required pairwise inner product changes from $-1/(K_{\text{old}}-1)$ to $-1/(K_{\text{new}}-1)$, so every established prototype moves and the class reference is rewritten. **(c)** SPCL freezes the old columns and places each new task's simplex block in the orthogonal complement; the structural memory is preserved under this frozen reference.

However, a geometric target that is optimal for a static classifier is not automatically a memory structure that persists over time. Fixed global ETFs keep a stable reference but require the final label space in advance; progressive ETFs adapt to the classes seen so far, but at a hidden cost: expanding a global simplex ETF changes the pairwise equiangular value from $-1/(K_{\text{old}}-1)$ to $-1/(K_{\text{new}}-1)$, so no expansion can keep all old prototypes unchanged (fig. 1b)—even with the feature $\hat{h}(x)$ perfectly preserved, moving the reference rewrites the old scores $(E^{\text{old}})^{\top}\hat{h}(x)$ and may alter the induced decisions.

This reveals a *static–temporal conflict*: global equiangularity favors rebuilding the prototype set for the current class count, whereas continual learning requires established references to remain inheritable across stages(proposition 4). The object to preserve is therefore not an ETF target at a single time step but the old decision structure it induces. For a cosine nearest-prototype classifier, this structure has three components: the class reference E^{old} (the old-class columns of the prototype matrix) defines the coordinate system of old-class evidence; the coordinate response $q(x)$ records how a sample responds to E^{old}; and the decision margin $m(x)$ protects the old-class prediction against its strongest old-class competitor. We call this coupled object *structural memory*: the NC-induced decision structure that should persist as new classes are incorporated.

We propose Structure-Preserving Continual Learning (SPCL), which preserves structural memory through these three components (fig. 1c). The class reference is frozen by construction, with each incoming task's simplex block placed in the orthogonal complement of the old span. Under this frozen reference, sample responses are distilled and old-class margins guarded by a gated hinge that activates only in the final third of each task's epochs, on a representation stabilized by mixed-batch feature distillation.

Our contributions are as follows:

1. **Static–temporal conflict in NC-based expansion.** Global ETF expansion improves adaptivity to the current class set, but exact global equiangularity necessarily moves old prototypes as the number of classes changes (proposition 4). Whether this cumulative reference rewriting burdens continual learning is left implicit by existing NC-based methods.
2. **Structural-memory formulation of old knowledge.** Rather than treating NC only as a static endpoint geometry, we organize the NC-induced decision structure into class references, coordinate responses, and decision margins, corresponding to the direct dependencies of the old classification function.
3. **SPCL: structure preservation across stages.** SPCL freezes old references, places new ETF blocks in the orthogonal complement, distills old-coordinate responses, and protects old margins with a hinge applied only in the post-plasticity window of each task, while retaining mixed-batch feature distillation as the representational substrate.
4. **Matched-protocol evidence and structural attribution.** On Seq-CIFAR-100 with a 500-sample replay buffer, SPCL improves final average accuracy while reducing average forgetting by approximately 70% (7.37 vs. 24.32) relative to the results reported in the ProNC paper. Structural diagnostics verify that prototype drift is zero by construction, and reference expansion produces no old-class prediction flip on the 45,000 audited inputs; by elimination, the residual forgetting concentrates on the feature side within the audit scope.

2 Related work

Replay and per-sample preservation. Experience replay anchors old distributions with buffered exemplars, selected by herding and classified by nearest class mean in iCaRL (Rebuffi et al., 2017; Rolnick et al., 2019); distillation-based variants constrain what the current model reproduces on the mixed batch—past responses (LwF, Li & Hoiem, 2017), buffered logits (DER++, Buzzega et al., 2020), refined logit distillation (LODE, Liang & Li, 2023), or redesigned rehearsal (CSReL, Tong et al., 2025). These mechanisms regulate sample-level quantities—features, logits, outputs—and leave implicit the class references, sample-to-prototype relations, and inter-class margins that generate the predictions. SPCL retains mixed-batch feature distillation as its substrate and preserves these structural objects on top of it (section 3.2); response-space alternatives such as asymmetric CE (Boschini et al., 2022) are examined in section D.

Neural collapse and ETF classifiers. Neural collapse shows that near the end of training, features collapse onto class means, the means approach a simplex ETF, and classifier weights align with them, so the decision reduces to a nearest-prototype rule (Papyan et al., 2020; Zhu et al., 2021). Fixed-global-ETF methods (Yang et al., 2023a;b) pre-assign one frame for all classes: a stable reference, but one requiring the final label space in advance. ProNC (Wang et al., 2026) instead rebuilds the frame for the classes seen so far; since a global simplex ETF has pairwise inner product $-1/(K-1)$, each rebuild moves the old prototypes (the static–temporal conflict of section 1). Training-side improvements refine augmentation and stability within this rebuild paradigm (Dang et al., 2026; Eskandar et al., 2025). Our construction shares ProNC's initialization and per-task expansion, but freezes the old columns and embeds each new block orthogonally; the resulting block simplex trades global equiangularity for temporal consistency of the reference (section 3.3).

Structural attribution of forgetting. Parameter regularization limits how far the weights may move (Kirkpatrick et al., 2017; Lopez-Paz & Ranzato, 2017); output regularization preserves responses (Li & Hoiem, 2017); relational distillation preserves relations among samples (Park et al., 2019). Each constrains a proxy object rather than the class-reference structure through which old predictions are generated, and how reference motion and feature change each contribute to rewriting old predictions has received little direct measurement. SPCL makes this question operational: preservation is organized around E^{old}, q, and m (section 3.2), the diagnostics separate reference-side from feature-side forgetting (section 4.4), and section D examines why these direct dependencies are the right targets.

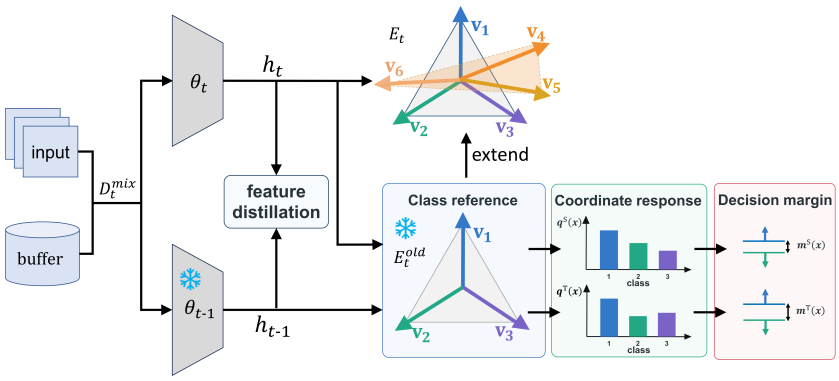


Figure 2: **The SPCL framework.** At each task boundary, the the previous reference is frozen as the old block ($E_t^{\text{old}} = E_{t-1}$) and the incoming classes receive a simplex block E_t^{new} in the orthogonal complement. Within a task, a frozen teacher and the student share this reference: the structural memory is preserved by response distillation ($q^{\text{T}} \approx q^{\text{S}}$) and the time-gated margin hinge (active in the final third of each task), on a representation anchored by vertex alignment and point-to-point feature distillation over the mixed batch; all terms are combined in the total objective (eq. (11)).

3 Structure-Preserving Continual Learning

SPCL preserves the NC-induced decision structure as *structural memory* through a reference construction and complementary preservation losses (fig. 2): the class reference is frozen by construction, with new-class vertices placed in the orthogonal complement (fig. 1c); the coordinate response is distilled under the frozen reference, with a time-gated margin hinge; and a mixed-batch feature-distillation term stabilizes the feature space.

3.1 Problem Setting

A class-incremental learner receives T tasks in sequence; task t introduces c_t new classes and K_t classes have been seen by its end. Write $\mathcal{C}_{\text{old}} = \mathcal{C}_{\leq t-1}$ for the classes placed before task t; no task identity is given, and at test time the model must predict among all classes seen so far. A replay buffer $\mathcal{M}$ of B herding-selected exemplars is maintained (section B); every training step combines a current-task mini-batch with a replayed mini-batch (the *mixed batch* $\mathcal{D}_t^{\text{mix}}$, on which all losses are computed).

The backbone network maps an input x to a feature vector, which we ℓ_2-normalize to $\hat{h}(x) \in \mathbb{R}^d$. Each class j has a unit-norm prototype vector v_j; we collect them into the prototype matrix $E_t = [v_1, \ldots, v_{K_t}] \in \mathbb{R}^{d \times K_t}$, serving as the class reference for all predictions. Classification is nearest-prototype under cosine similarity: the score of class j is $s_{j,t}(x) = \langle \hat{h}_t(x), v_j \rangle$, and the predicted label is $\hat{y}(x) = \arg\max_j s_{j,t}(x)$.

The network trained on task t is the *student* θ_t; the copy frozen at the end of task $t-1$ is the *teacher* θ_{t-1}^*, whose features $\hat{h}_{t-1}$ serve as gradient-free distillation targets. The direction-regression loss between two unit vectors, used by the alignment and feature-distillation terms below, is

$$\text{DR}(a, b) = \tfrac{1}{2}\big(a^\top b - 1\big)^2, \qquad \|a\|_2 = \|b\|_2 = 1. \tag{1}$$

3.2 Structural Memory

We organize the decision structure that the established classes contribute at task t into three components and formalize them as *structural memory*:

$$\mathcal{S}_t^{\text{old}} = \big(E_t^{\text{old}},\ q_t,\ m_t\big), \tag{2}$$

where E_t^{old} is the class reference—the columns of E_t indexed by $\mathcal{C}_{\text{old}}$ (superscripts mark the class scope, subscripts the task)—q_t the coordinate response of samples under that reference, and m_t the

decision margin. The subscript t marks the network state at which the latter two are evaluated: both are computed with the current features $\hat{h}_t$. Specifically,

$$q_t(x) = \sigma\big(\tau^{-1}(E_t^{\text{old}})^\top \hat{h}_t(x)\big), \qquad m_t(x) = s_{y,t}(x) - \max_{k \in \mathcal{C}_{\text{old}} \setminus \{y\}} s_{k,t}(x), \tag{3}$$

where σ is the softmax and τ is a fixed temperature.

These quantities form a one-way dependency: E_t^{old} and $\hat{h}$ jointly determine q_t, and m_t is a function of q_t—the gap to the strongest old-class competitor, which a prediction flip crosses at zero. Were q preserved exactly, m would follow automatically; in practice, distribution matching is imperfect, and residual errors can flip confusable samples—exactly the case the gated margin hinge targets.

Preservation acts across the task boundary: the target is $\mathcal{S}_t^{\text{old}} \approx \mathcal{S}_{t-1}$, where $\mathcal{S}_{t-1} = (E_{t-1},\, q_{t-1},\, m_{t-1})$ is the same structure evaluated at the end of task $t-1$, whose full label space is exactly $\mathcal{C}_{\text{old}}$. The class reference holds exactly—no-retroactive placement yields $E_t^{\text{old}} = E_{t-1}$ (proposition 1), placing both sides over the same columns—so response distillation and the gated margin hinge target $q_t \approx q_{t-1}$ and $m_t \approx m_{t-1}$ under this shared reference (section 3.4).

3.3 Class reference: no-retroactive construction and alignment

Any global ETF rebuild moves old prototypes (section 1), and no differentiable loss can prevent a discrete geometric operation; the class reference must therefore be preserved by construction. SPCL does so through *no-retroactive placement*, which freezes old columns and places new-class vertices in their orthogonal complement, and *vertex alignment*, which trains features to use the frozen reference.

No-retroactive placement. After task 0, the initial reference E_0 is obtained from normalized class means via a QR decomposition of the centered projector (section B). At every subsequent task boundary, the old columns are frozen element-wise, and the new-class vertices form a simplex ETF in the orthogonal complement of the span of the old prototypes:

$$E_t^{\text{new}} = B_\perp V_{c_t}, \tag{4}$$

where $B_\perp \in \mathbb{R}^{d \times (d-r)}$ is an orthonormal basis of the complement, $r = \operatorname{rank}(E_t^{\text{old}})$, and V_{c_t} is a simplex block (unit-norm columns, pairwise inner products $-1/(c_t-1)$), possible iff $d - r \geq c_t - 1$. The expanded reference is

$$E_t = [E_t^{\text{old}} \mid E_t^{\text{new}}]. \tag{5}$$

The construction guarantees the following properties (proofs in section A):

Proposition 1 (Zero retroactive drift)**.** *$E_t^{\text{old}} = E_{t-1}$, and this identity persists: every old column remains element-wise unchanged in all subsequent tasks.*

Proposition 2 (Prototype-block orthogonality)**.** *$(E_t^{\text{old}})^\top E_t^{\text{new}} = 0$: a feature in $\operatorname{span}(E_t^{\text{old}})$ scores zero on every new-class prototype; in general, the new-class score depends only on the complement residual.*

Features outside the old span can still receive nonzero new-class scores; the result is a **block simplex**—global equiangularity is traded for temporal consistency. Orthogonality alone does not guarantee prediction stability: an old prediction survives expansion exactly when $s_{y,t}(x) \geq \max_{j \in \mathcal{C}_{\text{new}}} s_{j,t}(x)$ on top of a nonnegative old-class margin. We note that $\mathcal{S}_t^{\text{old}}$ constrains the old-class *internal* structure (q over $\mathcal{C}_{\text{old}}$; m against old-class competitors)—the old–new margin above is not directly constrained by any loss; section 4.4 audits it empirically and observes zero flips on the audited inputs.

Proposition 3 (Capacity-limited extension)**.** *The block embeds completely iff $d - \operatorname{rank}(E_t^{\text{old}}) \geq c_t - 1$; beyond capacity, SPCL places new vertices in the remaining space with old columns still frozen (section B).*

Vertex alignment. The construction keeps the reference stationary, but classification also requires each sample's feature to lie near its class prototype; alignment maintains this correspondence under the frozen reference:

$$\mathcal{L}_{\text{align}} = \mathbb{E}_{(x,y) \in \mathcal{D}_t^{\text{mix}}} \big[\mathrm{DR}\big(\hat{h}_t(x),\, v_{y(x)}\big)\big], \tag{6}$$

where $v_{y(x)}$ is the sample's own prototype. The form follows ProNC; the difference is that targets, once placed, never change: replayed features are kept at their frozen vertices, new features are placed at theirs, and E_t is a constant store—within-task training updates only the feature extractor θ_t.

3.4 Preservation under the frozen reference: response distillation and the gated margin hinge

Freezing the reference eliminates reference-side change, but features may still drift within the frozen reference, altering each sample's position relative to the old-class references. Two complementary losses constrain this drift at different granularity: response distillation preserves the full response distribution, and the gated margin hinge prioritizes the decision-critical gap.

Response distillation. With teacher and student sharing the same frozen reference (proposition 1), their old-class response distributions are

$$q^{\mathrm{T}}(x) = \mathrm{sg}\big[\sigma\big(\tau^{-1}(E_t^{\mathrm{old}})^{\top}\hat{h}_{t-1}(x)\big)\big], \qquad q^{\mathrm{S}}(x) = \sigma\big(\tau^{-1}(E_t^{\mathrm{old}})^{\top}\hat{h}_t(x)\big), \tag{7}$$

where $\mathrm{sg}[\cdot]$ denotes stop-gradient and $\tau = 1/4$ (i.e., a sharpening factor of 4). Response distillation is

$$\mathcal{L}_{\mathrm{resp}} = -\,\mathbb{E}_{x\in\mathcal{D}_t^{\mathrm{mix}}} \sum_{j\in\mathcal{C}_{\mathrm{old}}} q_j^{\mathrm{T}}(x)\,\log q_j^{\mathrm{S}}(x), \tag{8}$$

which equals $\mathrm{KL}(q^{\mathrm{T}}\|q^{\mathrm{S}})$ up to a student-independent constant. The loss applies to all mixed-batch samples over the old columns—preserving the old-class response ordering of new samples and the responses of replayed samples together; restricting it to replayed samples conflicts with new-class learning (section D).

Gated margin hinge. The margin m is functionally determined by q (section 3.2); the hinge exists because aggregate distribution matching can still miss the sample–competitor pair closest to a flip. For a replayed old-class sample (x, y), the teacher and student margins $m^{\mathrm{T}}(x)$ and $m^{\mathrm{S}}(x)$ are obtained from eq. (3) by substituting $\hat{h}_{t-1}$ and $\hat{h}_t$; the constraint is

$$\mathcal{L}_{\mathrm{margin}} = \mathbb{E}_{(x,y)\in\mathcal{M}_t^{\mathrm{old}}}\Big[\mathbf{1}\big[e \geq \lceil \tfrac{2}{3}\mathcal{E}\rceil\big]\big[m^{\mathrm{T}}(x) - m^{\mathrm{S}}(x)\big]_+\Big], \tag{9}$$

where $\mathcal{M}_t^{\mathrm{old}}$ is the subset of replayed old-class samples in the current mixed mini-batch, e the current epoch, and $\mathcal{E}$ the per-task epoch budget: margin decreases are penalized only in the final third of each task. The gate follows the activation profile measured in section 4.4—the below-teacher transient of the plasticity phase self-repairs, and only the non-self-repaired residual needs carrying through the boundary; applying the hinge throughout the task instead fights the plasticity phase and yields no measurable gain (section 4.3).

3.5 Representational substrate and the overall objective

The three structural quantities are ultimately computed on the feature representation. To stabilize this shared input, we retain mixed-batch point-to-point feature distillation as the representational substrate:

$$\mathcal{L}_{\mathrm{p2p}} = \mathbb{E}_{x\in\mathcal{D}_t^{\mathrm{mix}}}\big[\mathrm{DR}\big(\hat{h}_t(x),\, \hat{h}_{t-1}(x)\big)\big]. \tag{10}$$

This constrains the numerical position of features—a different object from the structural constraints, which target discriminative relations under the frozen reference.

The task-learning term is the cross-entropy on the mixed batch over all placed prototypes, $\mathcal{L}_{\mathrm{CE}} = \mathbb{E}[-\log \mathrm{softmax}(s_{y,t})]$. Combining all terms, the overall objective for task $t \geq 1$ is

$$\mathcal{L} = \underbrace{\mathcal{L}_{\mathrm{CE}}}_{\text{task learning}} + \underbrace{\lambda_a\,\mathcal{L}_{\mathrm{align}} + \lambda_c\,\mathcal{L}_{\mathrm{resp}} + \lambda_m\,\mathcal{L}_{\mathrm{margin}}}_{\text{structural-memory preservation}} + \underbrace{\lambda_p\,\mathcal{L}_{\mathrm{p2p}}}_{\text{representational substrate}}, \tag{11}$$

with weights λ_a=18, λ_c=10, λ_m=1, λ_p=260 (all defaults fixed a priori; see section B). The reference is stored, not learned: alignment trains representations to *use* it, and the response and margin losses preserve relations measured *within* it.

Table 1: **Main results on Seq-CIFAR-100 and Seq-TinyImageNet.** SPCL results average 3 seeds ($\pm$std); †: quoted from the matched-protocol table of the ProNC paper. Best per column in bold, second best in italic.

	Seq-CIFAR-100				Seq-TinyImageNet			
	$B = 200$		$B = 500$		$B = 200$		$B = 500$	
Method	FAA↑	FF↓	FAA↑	FF↓	FAA↑	FF↓	FAA↑	FF↓
iCaRL (Rebuffi et al., 2017)†	28.00	47.20	33.25	40.99	7.53	*31.06*	9.38	*37.30*
DER++ (Buzzega et al., 2020)†	28.13	60.99	38.20	49.18	11.34	73.47	19.38	58.75
LODE (Liang & Li, 2023)†	26.65	44.29	40.01	32.58	15.13	64.00	20.50	56.51
CSReL (Tong et al., 2025)†	29.06	58.23	39.22	47.03	18.14	49.77	22.13	50.01
NCT (Yang et al., 2023b)†	26.38	*27.40*	33.77	27.85	10.95	49.33	18.24	50.90
STAR (Eskandar et al., 2025)†	38.15	42.17	47.56	28.68	13.64	68.51	21.31	57.20
ProNC (Wang et al., 2026)†	*42.99*	36.07	*48.94*	*24.32*	*27.44*	42.81	*29.06*	38.58
SPCL (ours)	**45.97** $\pm$0.6	**13.65** $\pm$0.4	**49.98** $\pm$1.0	**7.37** $\pm$0.4	**28.47** $\pm$0.2	**30.26** $\pm$0.6	**29.33** $\pm$0.6	**28.04** $\pm$0.3

Inference. $\hat{y}(x) = \arg\max_j \langle \hat{h}(x), v_j \rangle$: evaluation and learning share one never-moving reference; no new trainable parameters are introduced.

4 Experiments

4.1 Experimental setup

Datasets and baselines. We evaluate on Seq-CIFAR-100 (Krizhevsky, 2009) (10 tasks $\times$ 10 classes) and Seq-TinyImageNet (Le & Yang, 2015) (10 tasks $\times$ 20 classes), following the standard class-incremental protocol of Rebuffi et al. (2017). We compare against replay-based methods (iCaRL, DER++, LODE, CSReL, STAR) and NC-based methods (NCT, ProNC). †-marked values are quoted from the matched-protocol table of the ProNC paper (same backbone, same task splits, same budgets); protocol deviations of related methods are examined in section C. Our same-protocol reproduction of ProNC is consistent (section B).

Implementation. We train a ResNet-18 backbone (d=512) from scratch for 50 epochs per task using SGD (learning rate 0.03, momentum 0, weight decay 0, 10$\times$ decay at epochs 35/45); Seq-TinyImageNet uses 100 epochs, decay at 45/75/95, and λ_a=20. The batch size is 32; the replay buffer uses herding with capacity $B \in \{200, 500\}$. All SPCL results average 3 random seeds (initialization and shuffling vary; class order fixed).

Metrics. We report the final average accuracy (FAA) and the mean forgetting (FF), both defined as in Wang et al. (2026). FAA is the average test accuracy over all tasks after training completes; FF is the average drop from each task's peak accuracy to its final accuracy.

4.2 Main results

Table 1 shows the comparison under matched protocols and budgets. Under these protocols, SPCL attains the highest FAA and the lowest FF in every column of both datasets and both buffer sizes; our same-code-base reproduction of ProNC confirms the quoted values and serves as a protocol check (section B).

The pattern behind these numbers is informative: the retention gain far exceeds the accuracy gain. On Seq-CIFAR-100, FF falls by 62–70% relative to ProNC (e.g., 7.37 vs. 24.32 at B=500, a 69.7% reduction) while FAA improves by only 1–3 points; on Seq-TinyImageNet, FF is 10–13 points lower at both budgets. This asymmetry indicates where the improvement comes from: the frozen reference and response distillation keep old predictions from being rewritten—what the method removes is forgetting, not inaccuracy—and it grows as the buffer shrinks, where the structural constraints contribute most.

Among the NC-based designs, this resolves a long-standing trade-off. NCT pre-assigns the reference and concedes 15–17 FAA points to ProNC on Seq-CIFAR-100; ProNC rebuilds the reference and

Table 2: **Per-component removal ablation on Seq-CIFAR-100.** Best per column in bold. All entries average 3 seeds; the SPCL row repeats the 3-seed means of table 1 for reference, and *gated* marks the time-gated hinge of section 3.4.

			$B = 200$		$B = 500$	
Configuration	resp.	marg.	FAA	FF	FAA	FF
w/o response & margin	–	–	**46.63**	18.52	**50.98**	11.02
w/o response	–	✓	46.32	18.06	50.87	9.91
w/o margin	✓	–	45.99	14.77	50.02	7.96
full, ungated hinge	✓	✓	45.48	14.92	50.07	7.96
SPCL (full)	✓	gated	45.97	**13.65**	49.98	**7.37**

concedes 10–22 FF points to SPCL. By expanding the reference without rebuilding it, SPCL combines the stability of the former with the adaptivity of the latter, with retention gains concentrated on early tasks and late-task accuracy comparable to the rebuild baseline (section E).

4.3 Ablation

Table 2 removes one preservation component at a time from the full objective; all variants retain the class-reference construction and the substrate, and every entry is a 3-seed mean.

Response distillation carries retention. Removing it degrades FF by 2.5–4.4 points at the two buffers, and 3.6–4.9 points together with the hinge, while buying back only 0.4–1.0 FAA points. The exchange rate is itself the finding: several points of retention are purchased for well under one point of plasticity, and the removal variants are indeed slightly *more* accurate yet far more forgetful—the constraint blocks the rewriting of old decisions, not the learning of new ones.

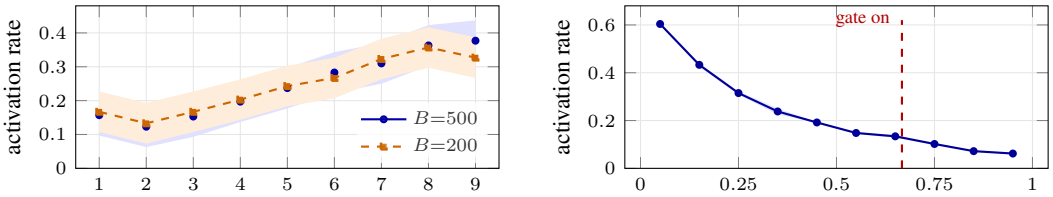
The gated hinge adds a separable gain. Against same-batch controls at $B{=}200$, the gated hinge lowers FF below both the no-hinge (15.27) and the ungated-hinge (15.54, per-seed signs flip) states in every seed, with FAA above the ungated baseline. The ungated hinge fails as a scheduling artifact—it fights the plasticity phase, whose transient self-repairs (section 4.4)—and a risk-gated variant that reweights but does not time the constraint fails identically (FF 16.52): the interference is temporal, not distributional.

Alignment and the substrate are necessary. Setting $\lambda_a{=}0$ collapses accuracy to 9–10%—cross-entropy still couples features to the reference but cannot pull new-class features to their placed vertices—and zeroing the substrate collapses FAA to 25.91: no structural loss replaces per-sample feature anchoring.

Overhead and parameter choice. SPCL adds no trainable parameters; its cost is one gradient-free teacher forward per step and one orthonormalization per boundary, and all weights were fixed before the final multi-seed batch. Both the substrate weight and the gating fraction admit broad operating ranges, supported by the dose–response plateau and the activation profile respectively (section B); beyond capacity ($\sum(c_r{-}1) > d$), the degraded placement stays within one FAA point of ProNC.

4.4 Analysis: structural diagnostics

The construction in section 3.3 provides two guarantees: zero prototype drift (proposition 1) and block orthogonality (proposition 2). The first three diagnostics verify each, measure what orthogonality alone cannot determine—the flips that reference expansion actually induces—and then attribute the remaining forgetting within the audit scope; the last profiles the within-task transient that the gated hinge polices.



(a) task index **(b)** within-task progress

Figure 3: **Margin-hinge activation dynamics. (a)** Fraction of replayed samples below the teacher margin, per task (B=500/200, 3 seeds; bands: seed range). **(b)** Within-task course (ungated, B=200): the transient self-repairs; the gated hinge activates at progress $2/3$ (dashed).

Reference drift. SPCL's prototype drift is exactly zero; ProNC's rebuild moves every old prototype at each boundary, matching the closed form

$$\delta(t) = \sqrt{2\Big(1 - \sqrt{\tfrac{(K_{\text{old}}-1)\,K_{\text{new}}}{K_{\text{old}}\,(K_{\text{new}}-1)}}\Big)}, \tag{12}$$

with K_{old} and K_{new} the class counts before and after the boundary: the per-boundary displacement shrinks as classes accumulate, reaching about 0.75 unit radii by the end of training, while SPCL stays exactly at zero (fig. 5a).

The reference-expansion channel. Orthogonality does not by itself determine old predictions (section 3.3), so this channel is measured directly through *flips*: comparing the teacher's predictions under old versus all prototypes at pre-overwrite checkpoints (five audits per boundary, nine boundaries) shows that reference expansion flips **no old-class prediction** on any of the 45,000 audited inputs—consistent with proposition 2, though not implied by it alone.

The feature channel. With both structural channels closed, replacing only the feature side (student vs. teacher features under the same all-prototype reference) concentrates all observed flips within the audit scope—2151 gross, 87% self-repaired within the task, 280 net, with buffer labels agreeing at task end: the residual forgetting is consistent with feature-side change on non-replayed inputs, the inherent boundary of the replay paradigm.

Where the hinge acts. The final diagnostic profiles the transient the gate polices: the below-teacher fraction grows with the task index (0.157 $\rightarrow$ 0.377; fig. 3a), and its within-task course peaks at the switch (0.61), decays monotonically, and leaves a 0.09 residual by the final third (fig. 3b) —precisely the gated hinge's window (median margin ratio 1.4).

4.5 Discussion

Prototype drift is eliminated by construction and expansion changes are suppressed on audited inputs; feature-side forgetting remains partially controlled. (i) Diagnostics cover audited inputs only. (ii) One backbone, two datasets; a permuted class order is consistent (section E); beyond capacity, only old-column immutability holds. (iii) Three seeds only: the ungated-hinge effect is unresolvable at this scale, and the gated-hinge gain rests on the same seeds. (iv) †-marked baselines are quoted; strict comparisons are same-code-base only.

5 Conclusion

NC-based continual learners rebuild the global ETF at every task boundary, rewriting past training targets. We propose Structure-Preserving Continual Learning (SPCL), which organizes the old decision structure as *structural memory*—class reference, coordinate response, decision margin—and preserves it by no-retroactive placement, response distillation under the frozen reference, a gated margin hinge, and a feature-distillation substrate. On both benchmarks, SPCL attains the highest accuracy and lowest forgetting, reducing mean forgetting by up to 70% over the strongest baseline with zero prototype drift. We hope this work opens directions for treating neural-collapse geometry as an inheritable memory and controlling feature-side forgetting beyond the reach of replay.

REPRODUCIBILITY STATEMENT

All experiments ran on the official ProNC code base under fully matched protocols. Complete training logs, metric scripts (FAA/FF recomputable to the last digit from logs), and geometry-verification scripts are archived with the code (reproduction details in section B). Every number in this paper is traceable to an archived log file.

ETHICS STATEMENT

This work uses only publicly available benchmarks (Seq-CIFAR-100 and Seq-TinyImageNet) and involves no human subjects or sensitive data. All experiments are reproducible from the archived code and training logs released with the paper.

AI ASSISTANCE DISCLOSURE

In this work, large language models (AI coding assistants) were used under the authors' direction and supervision to: (i) assist with retrieving and surveying related work; (ii) implement the experimental code for the methods and diagnostics studied here and assist with running experiments; (iii) assist with analyzing experiment logs and preparing statistical comparisons; and (iv) assist with drafting and polishing the manuscript, including Chinese–English translation. All research decisions—choice of methods, experimental design, interpretation and acceptance of results—were made by the authors; the authors verified every reported number against archived training logs and take full responsibility for the content of the paper.

REFERENCES

Matteo Boschini, Lorenzo Bonicelli, Pietro Buzzega, Angelo Porrello, and Simone Calderara. Class-incremental continual learning into the extended DER-verse. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5), 2022.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Trung-Anh Dang, Vincent Nguyen, Ngoc-Son Vu, and Christel Vrain. SAMix: Calibrated and accurate continual learning via sphere-adaptive mixup and neural collapse. *Machine Learning*, 2026. arXiv:2510.15751.

Masih Eskandar, Tooba Imtiaz, Davin Hill, Zifeng Wang, and Jennifer Dy. STAR: Stability-inducing weight perturbation for continual learning. In *International Conference on Learning Representations (ICLR)*, 2025.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13), 2017.

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

Ya Le and Xuan Yang. Tiny ImageNet visual recognition challenge. Technical report, Stanford University, 2015.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12), 2017.

Yan-Shuo Liang and Wu-Jun Li. Loss decoupling for task-agnostic continual learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continuum learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165, 1989.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40), 2020.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Sylvestre-Alvis Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Greg Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Ruilin Tong, Yuhang Liu, Javen Qinfeng Shi, and Dong Gong. Coreset selection via reducible loss in continual learning. In *International Conference on Learning Representations*, 2025.

Zheng Wang, Wanhao Yu, Li Yang, and Sen Lin. Rethinking continual learning with progressive neural collapse. In *International Conference on Learning Representations (ICLR)*, 2026.

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 *International Conference on Learning Representations (ICLR)*, 2023a.

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. *arXiv preprint arXiv:2308.01746*, 2023b.

Zhihui Zhu, Tianyu Ding, Jinqing Zhou, Xinyu Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

A PROOFS

Proposition 4 (Static–temporal conflict)**.** *Let E_{t-1} be the K_{old}-class simplex ETF at the end of task $t-1$ (so $K_{\text{old}} = K_{t-1}$, and likewise $K_{\text{new}} = K_t$). If the prototype set E_t at the end of task t forms a global $K_{\text{new}} > K_{\text{old}}$-class simplex ETF, then E_t contains no subset equal to E_{t-1} column by column—any expansion that preserves global equiangularity must move the old prototypes.*

Proof. Suppose such a subset existed; its pairwise inner products would still be $-1/(K_{\text{old}} - 1)$; but a global K_{new}-class simplex ETF requires every pairwise inner product of distinct classes to equal $-1/(K_{\text{new}} - 1)$. Since $K_{\text{new}} > K_{\text{old}} \geq 2$, the two target values differ—a contradiction. □

Scope of proposition 4. The claim is specific to strict unit-norm global simplex ETFs under column-preserving expansion; classifier variants with biases, non-unit norms, or learned scales are outside its scope, and the proposition makes no claim about how such variants expand.

Proof of proposition 1. By construction, no-retroactive placement copies every column of the previous frame E_{t-1} verbatim as the old block E_t^{old} at each task boundary; no loss term or post-processing writes these columns. Hence $E_t^{\text{old}} = E_{t-1}$, every v_k ($k \leq K_{\text{old}}$) is element-wise identical for all $t' \geq t$, and the score vector $(E_t^{\text{old}})^\top \hat{h}(x)$ is evaluated against a time-invariant reference. □

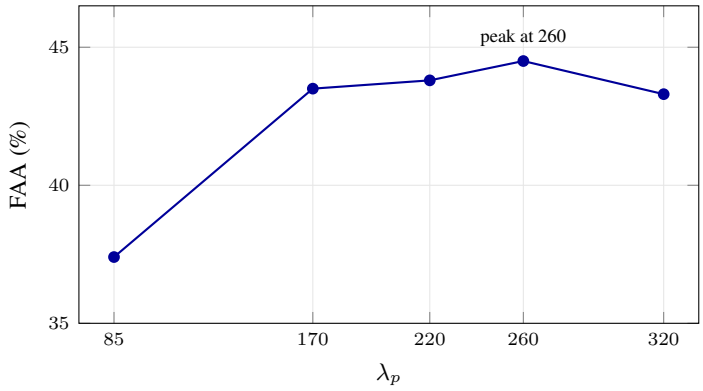


Figure 4: **Dose–response of the substrate weight** λ_p (B=200). The curve is inverted-U with a peak at λ_p=260; the plateau between 170 and 320 (within 1.2 points) indicates a broad operating range. Zero-drift placement is the prerequisite for this budget: under ProNC's drifting reference the same reweighting is harmful and the optimum returns to 170.

Proof of proposition 2. $B_\perp$ is an orthonormal basis of the orthogonal complement of $\mathrm{span}(E_t^{\mathrm{old}})$, so $(E_t^{\mathrm{old}})^\top B_\perp = 0$ and therefore $(E_t^{\mathrm{old}})^\top E_t^{\mathrm{new}} = (E_t^{\mathrm{old}})^\top B_\perp V_{c_t} = 0$. If $\hat{h}(x) \in \mathrm{span}(E_t^{\mathrm{old}})$, then $(E_t^{\mathrm{new}})^\top \hat{h}(x) = V_{c_t}^\top B_\perp^\top \hat{h}(x) = 0$: all new-class scores vanish and the arg-max over all prototypes equals the arg-max over the old prototypes. In general the new-class score depends only on the residual $B_\perp^\top \hat{h}(x)$; section 4.4 measures its induced flips at zero. □

Proof of proposition 3. The Gram matrix of a K-vertex simplex ETF is $G = \frac{KI - \mathbf{1}\mathbf{1}^\top}{K-1}$, whose null space contains $\mathbf{1}$, so $\mathrm{rank}(G) = K-1$ and the K prototypes span a $(K-1)$-dimensional subspace. A c-vertex simplex block likewise spans $c-1$ dimensions, and by proposition 2 its span is orthogonal to $\mathrm{span}(E_t^{\mathrm{old}})$, so after t tasks the reference E_t has rank $\sum_{r=0}^{t}(c_r - 1) = K_t - (t+1)$ (each block contributes $c_r - 1$ effective dimensions, its Gram having the all-ones null direction). Exact block placement requires $d \geq \sum_{r \leq t}(c_r - 1)$; beyond capacity no assignment can keep all cross-block inner products at zero, and the implementation degrades to placing new simplices in the available space while old columns stay frozen—the feasibility probe of section 4.3. □

B Reproduction Details

Our full configuration uses herding selection with the no-retroactive placement, vertex alignment, response distillation, the time-gated margin hinge (final third of each task's epochs), and mixed-batch feature distillation, with the weights of eq. (11); launch commands and training logs are archived with the code. The w/o response & margin variant drops the last two structural losses; the substrate-only variant additionally drops vertex alignment. When the rank budget is exhausted, the degraded placement of proposition 3 spreads new vertices uniformly over the remaining space while old columns stay frozen; in a reduced-width stress test (d=64, so that $\sum(c_r - 1) > d$), this regime leaves SPCL within one FAA point of ProNC (20.67 vs. 21.75, single run). Task 0 is trained with plain cross-entropy; the initial ETF is obtained from class means via QR of the centered projector at the first task boundary. Optimizer, schedule, and seeds are as in section 4.1.

Parameter choice. λ_p=260 sits at the peak of the dose–response plateau in fig. 4 (FAA within 1.2 points over $\lambda_p \in [170, 320]$); a frozen reference itself releases this budget—under ProNC's drifting reference the optimum returns to 170. The gating fraction $2/3$ is read off the activation profile: the below-teacher transient has largely self-repaired by the final third (fig. 3b), so any gate within that window isolates the same non-self-repaired residual.

ProNC reproduction. Our same-protocol reproduction of ProNC on Seq-CIFAR-100 (B=200, archived log) gives FAA 42.73 / FF 37.02, against the quoted 42.99 / 36.07—consistent within seed noise, which is the protocol check referenced in section 4.1.

Memory accounting. Beyond the replay buffer (B exemplars, herding re-selected at each task boundary—two or five per class at $B{=}200/500$), SPCL stores one frozen teacher copy of the backbone and the constant prototype matrix $E_t \in \mathbb{R}^{d\times K_t}$ (512×100 floats, negligible); class means and the boundary QR decomposition are computed transiently and not stored. The teacher adds one gradient-free forward per step.

C Protocol alignment with recent methods

SAMix (Dang et al., 2026) reports FAA 55.85 / FF 29.94 at $B{=}500$, but its protocol includes 500-epoch training on task one, an independently trained linear evaluation classifier, and a projection head. STAR (Eskandar et al., 2025) reports its strongest results under a 5×20 split; the 10×10 numbers quoted in table 1 (47.56/28.68 at $B{=}500$) are the matched-protocol reproduction reported in the ProNC paper. Among matched-protocol methods, our FF is 1.4–$3.3\times$ lower than ProNC, the strongest baseline, in every setting; direct comparisons in the main table are restricted to same-protocol, same-budget numbers.

D Why the direct dependencies: a falsification argument

The structural-memory formulation rests on a design principle: preservation constraints must sit on the *direct dependencies* of the old classification function (E, q, m, and their substrate $\hat{h}$), not on statistics derived from them. This appendix substantiates that principle by examining the alternative—constraining objects that lie off the dependency chain—and showing that each such constraint fails to protect the decision structure. All comparisons use the matched protocol and paired seeds of the main experiments.

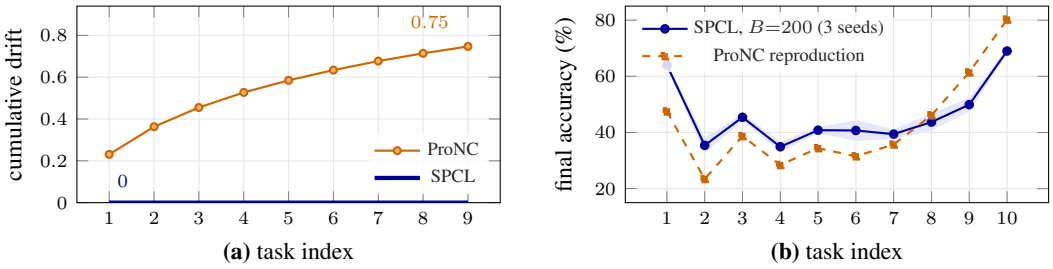
Derived statistics do not imply the function. Sample-relation statistics (Gram constraints, relational knowledge distillation (Park et al., 2019)) and distillation-replacement forms (an angle-matrix objective in place of point-to-point anchoring) preserve only many-to-one maps of the feature configuration: identical statistics are compatible with different decision boundaries, so preserving them leaves the prediction unprotected. Empirically, the former ranged from harmful (-6.8 FAA) to neutral, and the latter collapsed outright (FAA 14.0).

Gauge freedoms absorb constraints without preserving behavior. Compensating "nuisance" degrees of freedom—Procrustes and bounded rotations, scale–offset compensations, or rigid (R, s, b) corrections of the coordinate frame, whether applied on the training or the inference side—can re-express the old function exactly, so constraining them changes nothing the classifier computes. These forms were neutral to within the seed band (one variant collapsed to FAA 8.4), confirming that the constraint was absorbed by the gauge rather than by the structure it meant to protect.

Restricting or reweighting the anchoring loses information. Recomputing per-sample distillation weights from class-level stress, or restricting distillation to old samples with an all-class softmax, discards part of the per-sample information: the former was exactly neutral, the latter conflicted with new-class learning (-4 FAA points). Anchoring strongly augmented features to vertices was strongly harmful (-11 to -14), indicating that the anchor must sit on the same view of the data on which decisions are made.

Masking rather than aligning. Asymmetric cross-entropy (Boschini et al., 2022) masks old-class logits on new samples instead of aligning features to references; it carries a real accuracy cost (paired -0.78, consistent in $3/3$ seeds) and was therefore not included in SPCL.

Together these observations close the design argument: objects off the dependency chain either fail to determine the old predictions (derived statistics, gauge freedoms) or discard information that the predictions depend on (reweighted or restricted anchoring). Effective preservation must therefore anchor the direct dependencies themselves—the class reference, the coordinate response, the decision margin, and the substrate that carries them—which is precisely the construction of the main text.



(a) task index **(b)** task index

Figure 5: **Structural diagnostics at B=200.** **(a)** Cumulative drift of the task-0 prototypes, $\sum_{s\le t}\delta(s)$: ProNC's global rebuild versus SPCL's no-retroactive placement (mean of 3 runs). **(b)** Final per-task accuracy for SPCL (three seeds of the w/o-response variant, whose FAA is 0.8 points above the full method, table 2) versus the same-protocol ProNC reproduction; the band spans the seed range.

E Diagnostic curves and task-level analysis

Old–new balance under the block simplex. Trading global equiangularity for a frozen reference could in principle bias the score scales of early (old-block) and late (new-block) classes differently. Figure 5b shows the opposite pattern: the old–new final-accuracy gap is $+4.4$ points for SPCL (w/o-response variant) versus $+16.5$ for the global-rebuild baseline—the recency imbalance characteristic of replay is *reduced*, not introduced, by the frozen reference, and late-task accuracy remains comparable to the baseline (48.5 vs. 51.0) while early-task accuracy improves by about ten points.

Class-order robustness. Because each task's simplex block is placed in arrival order, a different class order could in principle reshape the block structure. Running the method under a permuted class order (two seeds) gives FAA 46.18/46.35 and FF 13.08/12.21, against 45.97/13.65 under the fixed order (three seeds, table 2): the construction is not order-sensitive, and the frozen reference retains its advantage under reordering.