# WHAT CONTINUAL LEARNING PRESERVES: AN INTERVENTIONAL STUDY OF VISION MODELS

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

## ABSTRACT

Continual learning requires retaining prior knowledge while acquiring new information. Beyond accuracy, prior work examines what continual learners preserve through representations, explanations, learned features, and important components. We extend this analysis to causal organization, using activation patching to measure component influences on final representations (causal profiles) and one another (effective networks). In controlled visual learning and continual self-supervised learning, preservation of this organization tracks retained performance across methods, preservation strengths and preservation targets. Two simple regularizers that preserve causal profiles or effective networks improve retention over fine-tuning. Preserving shuffled or reversed profiles can instead impair performance even when representations remain similar to their learned state. Controls further show that larger representation changes need not accompany larger accuracy losses. Decomposing performance changes across current, past, and future tasks reveals different learning–retention trade-offs. Strong weight regularization retains knowledge mainly by limiting further learning, whereas representation and causal-profile preservation allow continued acquisition. On Split-CIFAR-100, continued learning also improves previously learned tasks, allowing causal-profile preservation to outperform freezing and transferred weight-regularization baselines. On Split-ImageNet-R, causal-profile and effective-network preservation reduce forgetting relative to fine-tuning. These findings establish an intervention-based analysis framework and identify causal organization as a useful preservation target for retaining knowledge without unnecessarily restricting new learning.

## 1 INTRODUCTION

A continually learning network must acquire new knowledge without losing what it already knows (Wang et al., 2024; Zhou et al., 2024b; Shi et al., 2025). Existing methods constrain changes in parameters (Kirkpatrick et al., 2017; Liang & Li, 2024), outputs or features (Li & Hoiem, 2018; Fini et al., 2022; Asadi et al., 2023), or performance on stored examples (Rebuffi et al., 2017; Rolnick et al., 2019; Ibrahim et al., 2024). Accuracy alone does not reveal which aspects of the model these methods preserve. Researchers have examined this question through representation probing (Ramasesh et al., 2021; Davari et al., 2022), explanation drift (Cossu et al., 2024b), activity-weighted graphs (Carboni et al., 2025), learned features (Masip et al., 2026), and component importance (Vahedifar et al., 2026). Explanation-preserving objectives further show that maintaining the evidence underlying predictions can reduce forgetting (Dhar et al., 2019; Ebrahimi et al., 2021). Conversely, high accuracy can conceal limited acquisition of new representations (Kim & Han, 2023; Hess et al., 2024).

These approaches characterize different aspects of a model's internal computation: probing asks what information a simple readout can recover from representations, attribution methods identify the input evidence behind a prediction, and dictionary-learning methods such as crosscoders decompose activations into interpretable features that can be compared across training. Our approach is instead inspired by the brain, where learning reorganizes how regions interact: the modular structure of brain networks reconfigures during learning (Bassett et al., 2011), sensorimotor systems become more autonomous as a skill is mastered (Bassett et al., 2015), and learning shifts the balance between top-down and bottom-up influence in visual cortex (Makino & Komiyama, 2015). We therefore ask whether continual learning preserves the causal influence of each component on the model's output



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.

and the effective connections between components in successive layers, whether preserving them directly helps, and whether their exact correspondence matters.

We measure both with activation patching (Vig et al., 2020; Geiger et al., 2021; Meng et al., 2022; Zhang & Nanda, 2024): we replace the activation of one attention head or convolutional channel, fully or partially, with the activation of another input and record what changes. The change in the model's final representation measures that component's causal influence on the output, and these influences form the model's *causal profile*. The changes in the components of the next layer measure the component's effective connections to that layer, and these connections form the *effective network*. We track both of them, alongside parameter and representation changes, during continual learning, and separately test objectives that preserve them with correct targets or with shuffled or reversed targets. Following studies of representation plasticity and knowledge accumulation (Kim & Han, 2023; Hess et al., 2024), we separately account for performance changes on current, past, and future tasks.

We combine exposure-matched learning of synthetic objects with continual self-supervised learning on Split-CIFAR-100 and Split-ImageNet-R. The study makes three contributions.

- **Intervention-based measurements of preservation.** Cross-view correspondence learning strengthens causal-profile and effective-network identity specificity. Preservation of these measurements also tracks retained function among methods that do not explicitly optimize them (Section 5.1).
- **Directly preserving causal structure improves continual learning.** Objectives that preserve the causal profile or the effective network improve final accuracy relative to fine-tuning, whereas preserving shuffled or reversed targets impairs it. Which structure is preserved matters more than how much representations change: a target change that altered representations more than a reversed profile target left accuracy unchanged (Section 5.2).
- **Decomposing learning and forgetting.** Across both benchmarks, gains over fine-tuning arise mainly from more favorable changes on past and future tasks, rather than larger current-task gains. At the selected strengths, weight regularization sharply limits learning after the initial task, whereas representation and causal-profile preservation retain much of fine-tuning's current-task learning while reducing losses on other tasks (Section 5.3).

Our two direct regularizers preserve stored causal profiles or effective networks. They serve as experimental probes, rather than a claim of uniformly superior continual-learning performance.

## 2 Related work

**Diagnosing forgetting and acquisition.** Accuracy alone does not show what a continually trained model loses or gains, so previous analyses have looked inside the model. Forgetting concentrates in deeper layers (Ramasesh et al., 2021), and linear probes show that representations can retain information the classifier no longer uses (Davari et al., 2022). Conversely, strong class-incremental accuracy need not mean that the feature extractor keeps learning (Kim & Han, 2023), and excluding tasks from training separates knowledge accumulation from feature forgetting (Hess et al., 2024). Backward and forward transfer summarize effects on earlier and later tasks (Lopez-Paz & Ranzato, 2017); our decomposition divides the final gain over the frozen encoder exactly into current-, past-, and future-task contributions.

**Explanation preservation and drift.** Learning without Memorizing preserves attention maps (Dhar et al., 2019), while Remembering for the Right Reasons maintains explanations for replayed examples (Ebrahimi et al., 2021). CLEX measures explanation drift across continual-learning strategies (Cossu et al., 2024b), and ICICLE regularizes changes in prototype-based interpretations (Rymarczyk et al., 2023). These studies examine the evidence underlying predictions; we instead measure internal components' responses to activation patching.

**Graphs, features, and component interventions.** Carboni et al. (2025) study architectural graphs weighted by unit activity; our effective networks instead measure how receiver components respond when a sender is patched. Using crosscoder features, Masip et al. (2026) attribute forgetting to

feature degradation and readout misalignment. We measure the model's own heads and channels, which requires no learned feature dictionary but does not reveal which semantic features they encode. Shapley neuron valuation uses activation interventions to identify units worth protecting (Vahedifar et al., 2026), and circuit analyses compare preservation under reinforcement learning and supervised fine-tuning (Rojas Nunez et al., 2026). We measure each component's effect on both the final representation and the next processing stage, and use each as a preservation target.

**Continual learning with pretrained and self-supervised models.** Frozen pretrained features provide strong class-incremental baselines (Janson et al., 2022; Zhou et al., 2024a; McDonnell et al., 2023). In this regime, L2 regularization toward previous weights can outperform EWC and distillation (Smith et al., 2023). Continual self-supervised methods include predictor distillation (Fini et al., 2022), functional regularization (Gomez-Villa et al., 2022), and mixup (Madaan et al., 2022). Comparisons with supervised training report reduced forgetting under self-supervision (Madaan et al., 2022; Cossu et al., 2024a). We adopt this setting and compare ER, L2, EWC, and CaSSLe with preservation of representations and intervention-defined organization.

## 3 Measuring and Preserving Network Organization

### 3.1 Two Intervention-Defined Levels of Organization

For a base image $x_b$ and source image $x_s$, activation patching replaces component $k$'s activation with $(1-\alpha)h_k(x_b) + \alpha h_k(x_s)$. The remaining computation proceeds normally; $\alpha = 1$ denotes full replacement. Let $\widehat{z}$ denote the normalized final representation. The *causal profile* of the pair is defined as the change in this representation when each component $k$ is patched,

$$p_k(x_b; x_s) = \left\| \widehat{z}^{(k,x_s)}(x_b) - \widehat{z}(x_b) \right\|_2, \tag{1}$$

and the *effective network* is defined as the change in each receiver component $j$ of the next processing stage when sender component $k$ is patched (Wang et al., 2023), $e_{kj}(x_b; x_s) = \|h_j^{(k,x_s)}(x_b) - h_j(x_b)\|_2$. Components are attention heads in DINOv2-B and convolutional channels in CORnet-S (Appendix A). When a model is compared with its earlier state, each component is compared with the component at the same position, that is, the same layer and index. A profile describes both the pattern and the size of component effects, and a network row describes one sender's pattern of effects on the next stage. They describe responses to the specified patches, not anatomical connections.

### 3.2 Measuring What Is Preserved

We quantify preservation relative to a reference state using four measures, on held-out images that no method uses for preservation or replay. *Weight distance* is parameter displacement weighted by a diagonal Fisher estimate (Kirkpatrick et al., 2017; Mirzadeh et al., 2020), computed from squared loss gradients on first-task images, using the pretrained model on the public benchmarks and the model after initial-set learning in the controlled-exposure setting. For prediction, *weight preservation* is $-\log_2$ of displacement relative to fine-tuning in the same stimulus set or class order. *Representation similarity* uses cosine similarity and centered kernel alignment (CKA) (Kornblith et al., 2019): cosine compares image features, whereas CKA compares representational geometry. *Causal-profile similarity* is cosine similarity between mean-centered profiles. *Effective-network similarity* is the cosine similarity between each sender's effects on its receivers, after removing the effect pattern shared by all senders. On the continual benchmarks, the reference is the post-task-1 model, so these measurements quantify preservation of that state.

### 3.3 Preservation Objectives

Each regularized method minimizes $\mathcal{L}_{\text{SSL}} + \lambda\mathcal{R}$, with $\mathcal{R}$ constraining a specified property relative to a stored reference (Table S1). EWC (Kirkpatrick et al., 2017) and L2 penalize parameter displacement from the latest task boundary, using Fisher or uniform weights. Representation preservation penalizes changes in stored images' representations (Rebuffi et al., 2017; Hou et al., 2019; Douillard et al., 2020). Causal-profile preservation penalizes differences between current and stored standardized profiles; effective-network preservation penalizes changes in each sender's pattern of effects.

We refer to the latter two as *intervention-based regularizers*. Image-based targets are computed when their images or pairs enter memory and remain fixed thereafter. Memory-based methods store the same images, reaching 200 only after the final task (Table S1).

The objective scales differ by orders of magnitude, making their coefficients difficult to compare directly. We therefore control relative gradient strength (Chen et al., 2018), setting

$$\lambda_t = \gamma \, \|\nabla_\theta \mathcal{L}_{\mathrm{SSL}}\|_2 \, / \, \|\nabla_\theta \mathcal{R}\|_2, \tag{2}$$

at each update with a nonzero preservation gradient. $\lambda_t$ is recomputed at every update but not differentiated, is capped at $10^6$, and stays unchanged when the preservation gradient is zero, as at the first update after a task boundary. We call $\gamma$ the preservation *strength* and select it separately for each method using validation data. Profile gradients include differentiation through standardization.

### 3.4 Decomposing accuracy changes across tasks

Let $A_{i,j}$ denote test accuracy on task $j$ after training through task $i$ (Lopez-Paz & Ranzato, 2017); row zero represents the frozen pretrained encoder. We fit a linear probe over all classes after each stage, so accuracies are defined for past, current, and future tasks. Final mean accuracy is $\bar{A}_T = \frac{1}{T}\sum_j A_{T,j}$. The final gain over the frozen encoder is the sum of the accuracy changes made at each stage. At stage $i$, we partition the change from row $i-1$ to row $i$ into current-task ($j = i$), past-task ($j < i$), and future-task ($j > i$) contributions:

$$\bar{A}_T - \bar{A}_0 = \underbrace{\frac{1}{T}\sum_{i=1}^{T}\left(A_{i,i} - A_{i-1,i}\right)}_{\text{current task}} + \underbrace{\frac{1}{T}\sum_{i=1}^{T}\sum_{j<i}\left(A_{i,j} - A_{i-1,j}\right)}_{\text{past tasks}} + \underbrace{\frac{1}{T}\sum_{i=1}^{T}\sum_{j>i}\left(A_{i,j} - A_{i-1,j}\right)}_{\text{future tasks}}. \tag{3}$$

The past- and future-task terms equal $(T-1)/T$ times backward and forward transfer (Lopez-Paz & Ranzato, 2017), with the frozen encoder as the forward-transfer baseline. We report accuracy and absolute accuracy differences in %; differences are not relative percentage changes. Every method follows the same first-stage protocol; preservation methods primarily diverge from stage two onward.

## 4 Experimental setup

**Controlled-exposure setting.** Pretrained models have already learned natural images, so we measure the effects of new experience with novel synthetic objects. CORnet-S (Kubilius et al., 2019) and DINOv2-B (Oquab et al., 2024) learn three-dimensional Zeroverse objects (Xie et al., 2024) without class labels, under matched exposure, across twelve independently generated stimulus sets. Each model learns an initial object set, then continues on a new set; retention is evaluated on the initial objects. Preservation objectives use stored anchors but no initial-set task loss; ER provides an explicit replay comparison. Appendix A describes the complete protocol.

**Continual self-supervised learning.** Pretrained DINOv2-B continues a DINO/iBOT self-supervised objective (Caron et al., 2021; Zhou et al., 2022; Oquab et al., 2024) across five Split-CIFAR-100 tasks (Krizhevsky, 2009) or ten Split-ImageNet-R tasks (Hendrycks et al., 2021; Wang et al., 2022). Labels do not enter this loss. After each task, we fit an all-class linear probe on frozen features (Kolesnikov et al., 2019; Fini et al., 2022). We compare fine-tuning without preservation, experience replay (ER) (Rolnick et al., 2019), CaSSLe (Fini et al., 2022), EWC (Kirkpatrick et al., 2017), L2 (Smith et al., 2023), representation preservation (Rebuffi et al., 2017; Hou et al., 2019; Douillard et al., 2020), and causal-profile and effective-network preservation. Each class order defines a replication; comparisons use orders completed by all relevant methods. We select strengths on validation data from six Split-ImageNet-R development orders and transfer them unchanged to Split-CIFAR-100. Twelve additional ImageNet-R orders provide confirmation without further selection. Appendix B specifies the implementation and label use for task construction, memory, and evaluation. Comparisons are paired by class order or stimulus set and use exact two-sided sign-flip tests. Figures report means with 95% intervals or standard errors, as specified (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

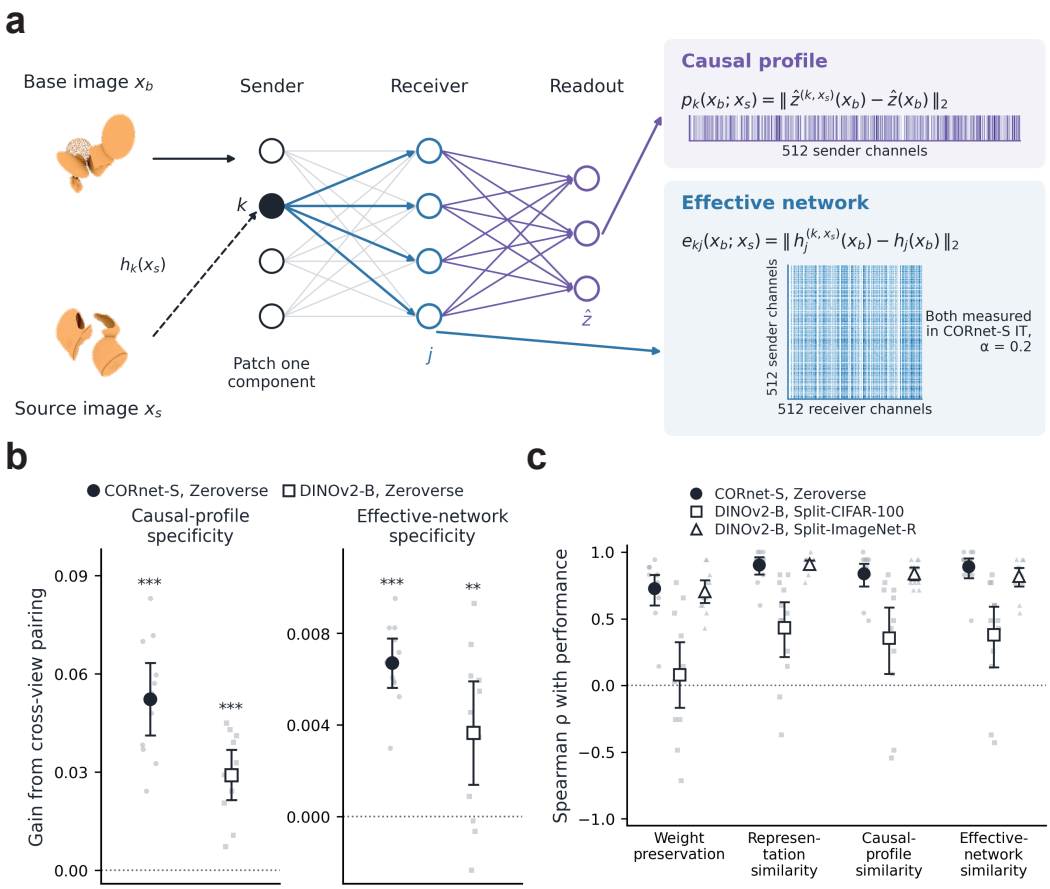



Figure 1: **Intervention-defined organization is shaped by learning and associated with retained performance.** **(a)** Activation patching measures effects on the final representation (causal profile) and the next processing stage (effective network). Examples use the displayed images and CORnet-S IT: 512 first-step sender channels, patched with $\alpha = 0.2$, and 512 second-step receiver channels. **(b)** Effect of cross-view learning on causal-profile and effective-network identity specificity, that is, how much more an object's profiles or networks resemble those of its other views than those of a matched lure object, in the controlled-exposure setting: gain from cross-view over within-view pairing, twelve stimulus sets per model (Appendix A); gray dots denote stimulus sets, black symbols means with 95% $t$ intervals, and marks exact sign-flip tests against zero. **(c)** Within-set or within-order Spearman correlations between each measure and retained function, including final accuracy on twelve orders per public benchmark; CORnet-S: six methods at validation-selected strengths across twelve stimulus sets, with the causal profile of Eq. 1. Symbols show means with 95% intervals; faint points show individual sets or orders. Intervention-based regularizers are excluded. Weight preservation is the negative log displacement ratio relative to fine-tuning; larger values indicate greater preservation for every measure.

## 5 Results

### 5.1 Intervention-derived measurements track retained function

We first tested whether learning alters the measured organization in the controlled-exposure setting. Cross-view learning increased causal-profile identity specificity in all 12 stimulus sets for both models, and effective-network identity specificity in most sets (Figure 1b), alongside identification accuracy (Figure S1).

We next examined whether preservation predicts performance among methods that do not explicitly optimize causal profiles or effective networks. This analysis included fine-tuning, ER, CaSSLe,

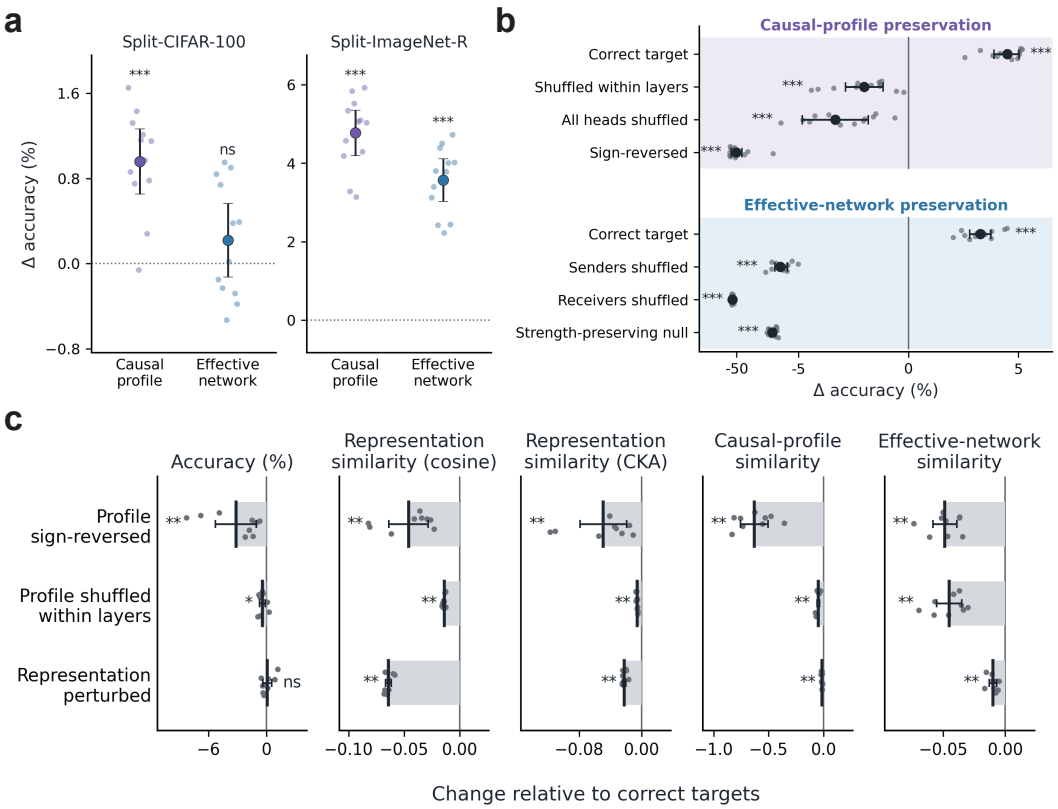



Figure 2: **Preservation-target correspondence affects performance.** DINOv2-B; dots denote class orders, and summary marks denote means with 95% intervals. Accuracy differences are in %. **(a)** Selected-strength preservation relative to fine-tuning: twelve Split-CIFAR-100 orders and twelve Split-ImageNet-R confirmation orders. **(b)** Standalone profile or network preservation with correct or rearranged targets, relative to fine-tuning: Split-ImageNet-R orders 1–12, symmetric logarithmic axis. Strengths as in (a). The strength-preserving null replaces each transition of the network target with the outer product of its sender and receiver totals divided by the total weight, keeping every sender's and receiver's total effect but no specific connection. **(c)** Joint representation and profile preservation ($\gamma = 0.5$ each), with one target perturbed, relative to both targets being correct. Split-ImageNet-R orders 4–12 were not used to select these perturbations. Effective-network similarity is measured but not directly constrained. Figure S2 reports all standalone target rearrangements.

EWC, L2 and representation preservation. In the controlled-exposure setting, with each method at a strength selected by mean familiar- and new-set accuracy on six stimulus sets (Eq. 2; Appendix A), causal-profile and effective-network similarity tracked retention within stimulus sets ($\rho = 0.84$ and 0.89; Figure 1c), as did representation similarity (0.90), whereas weight preservation showed a weaker association (0.73). On Split-CIFAR-100, representation, profile, and network similarity correlated with final accuracy (0.36–0.43), more than weight preservation (0.08). All four measures were strongly associated with accuracy on Split-ImageNet-R (0.70–0.91). Across methods, causal-profile and effective-network similarity thus tracked retained function in every setting, more closely than weight preservation.

### 5.2 Preservation objectives and their target correspondence affect performance

Associations across methods cannot isolate the contribution of a preservation target. We therefore first added intervention-based regularizers with measured targets, then altered those targets while holding the objective, strength, stored images, and training stream fixed.

**Correct targets can improve accuracy relative to fine-tuning.** On Split-CIFAR-100, selected-strength causal-profile preservation improved accuracy by $0.96\%$ over fine-tuning ($p < 0.001$). Effective-network preservation yielded a smaller, non-significant gain of $0.22\%$ (Figure 2a). On the Split-ImageNet-R confirmation orders, their gains were $4.8$ and $3.6\%$, respectively.

**Rearranging target correspondence can impair performance.** Correct-target profile and network preservation improved accuracy relative to fine-tuning by $4.5$ and $3.3\%$, whereas rearranged targets produced lower mean accuracy than fine-tuning (Figure 2b). Profile shuffling cost 2.0–$3.3\%$ and profile reversal cost $51\%$ relative to fine-tuning. Permuting network sender rows cost $9.8\%$; permuting receiver columns cost $59\%$. A strength-preserving null that keeps every sender's and receiver's total effect but no specific connection cost $13.3\%$, so aggregate sender and receiver strengths alone do not reproduce the benefit of the measured network. These experiments use the strengths selected for Figure 2a. Receiver-column permutation changed the target less than shuffling all entries, yet impaired performance as much (Appendix C; Figure S2b). Thus, target-change magnitude alone did not order the functional costs. Among tested targets, only those preserving the model's own measured organization yielded mean gains over fine-tuning.

**Causal-profile similarity tracks accuracy across preservation targets.** These experiments preserved either the correct target or a rearranged one, for example shuffled or sign-reversed (Figure 2b). Ranking all targets by accuracy shows how each preserved measure changes with it (Figure S3; $\rho$ denotes the Spearman correlation between the mean accuracy and the mean measure of each target). For profile targets, causal-profile similarity decreased in the same order as accuracy ($\rho = 1.00$), whereas representation and effective-network similarity did not ($0.60$ each), and weight preservation was negatively associated with accuracy ($-0.80$). For network targets, representation, causal-profile and effective-network similarity all decreased in the same order as accuracy ($\rho = 1.00$), whereas weight preservation was only weakly associated with it ($0.32$). Causal-profile and effective-network similarity thus tracked accuracy across preservation targets, causal-profile similarity most closely.

**Perturbing targets under joint preservation.** We next perturbed one target while jointly preserving representations and causal profiles, with stored images and training held fixed (Figure 2c). On orders excluded from perturbation selection, profile reversal reduced accuracy by $3.2\%$ ($p = 0.004$). It reduced profile similarity by $0.63$, representation cosine by $0.046$, CKA by $0.049$, and network similarity by $0.049$. Within-layer profile shuffling changed network similarity almost as much ($-0.045$), but cost only $0.43\%$ ($p = 0.02$). Network-fidelity loss alone therefore did not account for the difference in functional cost in these tests.

For comparison, we retained the correct profile target and permuted a fixed fraction of each representation target's coordinates. We calibrated this fraction on separate orders using representation similarity, without inspecting accuracy. The resulting cosine change exceeded that under profile reversal ($-0.064$), although CKA changed less ($-0.022$) and profile and network changes were small. Accuracy showed no detectable change ($+0.09\%$, 95% interval $[-0.36, +0.55]$). These perturbations reveal different relationships between coordinate drift, representational geometry, intervention-defined organization, and retained performance. In particular, representation metrics alone did not account for their functional effects.

### 5.3 Preservation objectives differ in how they balance retention and continued acquisition

Preservation can retain an early model state or permit further acquisition, with different consequences for other tasks. We compared selected-strength methods using accuracy over tasks learned so far (Chaudhry et al., 2019) (Figure 3a). We then decomposed each method's final gain over freezing into current-, past-, and future-task contributions (Figure 3b).

**Restricted change versus continued acquisition.** On Split-CIFAR-100, accuracy after the first task, which all methods learn with the same protocol, was 89.7–$89.9\%$ (Figure 3a). As tasks were added, representation and causal-profile preservation stayed near this level ($89.5\%$ at the end), whereas fine-tuning and CaSSLe declined to $88.5\%$ and $88.1\%$. EWC and L2 finished near the frozen encoder's accuracy on the same tasks ($88.7\%$ and $88.6\%$). The first stage, which all methods share, contributed $0.19$–$0.24\%$ on the current task and $-0.29$ to $-0.24\%$ on future tasks. EWC

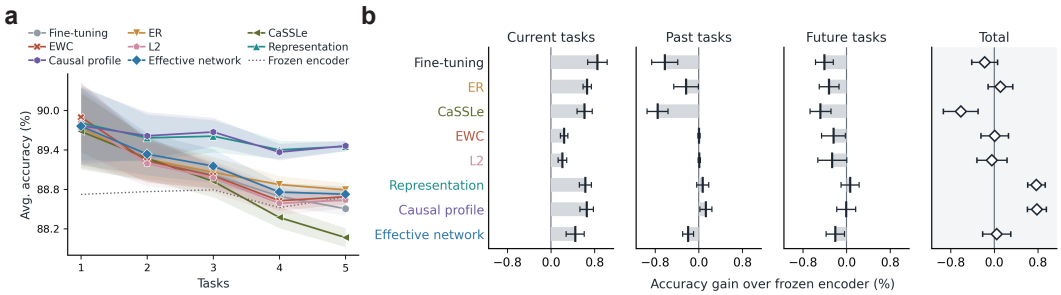



Figure 3: **Preservation objectives balance retention and acquisition differently.** DINOv2-B on Split-CIFAR-100, with selected strengths and twelve matched class orders. **(a)** Mean accuracy over tasks learned so far, evaluated by a linear probe fitted after each stage; shading shows s.e.m. The dotted line evaluates the frozen pretrained encoder on the same tasks. **(b)** Current-, past-, and future-task contributions to the final gain over freezing (Eq. 3); diamonds show their sum. All tasks use the same all-class probe, and all methods share the first-stage protocol. Means with 95% intervals; contributions are accuracy changes in %, not learning rates. Strength sweeps, each method's path through preserved measures and accuracy, and additional dynamics appear in Figures S4, S5 and S6.

and L2's final contributions differed from these first-stage values by at most $0.01\%$, indicating little subsequent change.

Fine-tuning produced the largest current-task contribution ($0.85\%$), but lost $0.62\%$ on past tasks and $0.40\%$ on future tasks. It therefore finished $0.18\%$ below freezing; CaSSLe's net loss was larger ($-0.62\%$). Representation and causal-profile preservation acquired less on current tasks ($0.63$ and $0.65\%$), but improved past tasks ($0.07$ and $0.13\%$). They also offset the initial future-task loss, yielding final future-task contributions of $0.07$ and $-0.01\%$. Both therefore finished approximately $0.8\%$ above the frozen encoder. Despite nearly identical final accuracy, these Split-CIFAR-100 results suggest slightly more favorable past-task changes under causal-profile preservation and future-task changes under representation preservation, indicating different emphases of the two methods, although these differences were not statistically significant.

**Causal-profile preservation exceeds freezing.** Continued acquisition supported the leading methods on Split-CIFAR-100 (Table 1). Causal-profile preservation exceeded the frozen pretrained encoder by $0.78\%$ and its own post-task-1 model by $0.86\%$. It also exceeded transferred EWC and L2 by $0.78$ and $0.83\%$, respectively (all four contrasts $p < 0.001$). Its advantage over ER was $0.67\%$ ($p = 0.001$). Representation preservation, the strongest baseline, exceeded freezing by $0.77\%$. Under matched image memories and validation-selected strengths, causal-profile preservation achieved nearly identical mean accuracy ($89.46$ vs. $89.45\%$). Across tested strengths, representation and profile preservation consistently outperformed fine-tuning, whereas EWC and L2 approached frozen-encoder performance (Figure S4, left).

**Split-ImageNet-R.** On the reserved confirmation orders, every method outperformed fine-tuning on average, and weight regularization gained most (6.0–6.2%; Table 1; Figure S4, right). Across all methods and tested strengths on the development orders, accuracy tracked representation similarity ($\rho = 0.90$) and causal-profile similarity ($0.85$) more closely than effective-network similarity ($0.67$) or weight preservation ($0.44$; Figure S5). Causal-profile similarity thus also tracked accuracy as preservation strength varied. The frozen encoder also exceeded fine-tuning, by $5.1\%$, and outperformed several preservation methods. Taskwise accounting shows why (Figure S6b). EWC and L2 again largely retained the post-task-1 state. Representation and profile preservation had larger current-task contributions ($0.98$ and $0.86\%$) than EWC and L2 ($0.37$ and $0.39\%$). However, their past-task contributions were negative ($-0.98$ and $-1.02\%$), versus approximately zero under weight regularization. Future-task contributions were also smaller ($-0.33$ and $-0.15\%$, versus $0.61$ and $0.77\%$). Thus, the benefits of further acquisition were outweighed by changes to other tasks, favoring retention of the earlier state.

Table 1: **Performance at validation-selected strengths.** Strengths $\gamma^*$ are selected on six Split-ImageNet-R development orders and transferred unchanged to Split-CIFAR-100 (Table S2). Accuracy is the final linear-probe result (%, mean ± s.e.m.) across twelve CIFAR orders or twelve reserved ImageNet-R confirmation orders. $\Delta$ denotes the paired gain over fine-tuning in %, followed by improved orders / total orders.

| | | Split-CIFAR-100 | | Split-ImageNet-R, confirmation orders | |
|---|---|---|---|---|---|
| Method | $\gamma^*$ | Accuracy | $\Delta$ | Accuracy | $\Delta$ |
| Frozen pretrained encoder | – | $88.68 \pm 0.00$ | +0.18 (9/12) | $70.73 \pm 0.05$ | +5.08 (12/12) |
| Fine-tuning | – | $88.50 \pm 0.11$ | | $65.65 \pm 0.24$ | |
| ER | 200 img | $88.79 \pm 0.10$ | +0.29 (10/12) | $67.63 \pm 0.22$ | +1.98 (11/12) |
| CaSSLe | 0.05 | $88.06 \pm 0.14$ | −0.44 (1/12) | $67.04 \pm 0.35$ | +1.38 (12/12) |
| EWC | 32 | $88.69 \pm 0.11$ | +0.18 (9/12) | $71.66 \pm 0.19$ | +6.01 (12/12) |
| L2 | 16 | $88.64 \pm 0.13$ | +0.13 (8/12) | $71.86 \pm 0.18$ | +6.21 (12/12) |
| Representation | 0.2 | $89.45 \pm 0.07$ | +0.95 (11/12) | $70.40 \pm 0.14$ | +4.75 (12/12) |
| Causal profile | 1 | $89.46 \pm 0.08$ | +0.96 (11/12) | $70.42 \pm 0.17$ | +4.77 (12/12) |
| Effective network | 4 | $88.72 \pm 0.12$ | +0.22 (7/12) | $69.22 \pm 0.21$ | +3.56 (12/12) |

## 6 Discussion and conclusion

This study extends analyses of preserved features and explanations to the intervention-defined organization of network components. Causal profiles and effective networks track retained function across methods, preservation strengths and preservation targets, and changing these targets can alter subsequent performance. Taskwise accounting further distinguishes retention through restricted adaptation from retention accompanied by new acquisition. The value of that acquisition depends on its effects on other tasks. These results support causal organization as both an analysis tool and a candidate preservation target, complementary to parameter and representation measurements.

A method may also perform well primarily by restricting further adaptation. This imposed restriction differs from the loss of plasticity studied by Dohare et al. (2024). Our results motivate preservation objectives that protect prior knowledge while retaining acquisition when its cross-task benefits outweigh its costs.

**Limitations.** We study two architectures, with public-benchmark evidence limited to one pretrained encoder, one benchmark family, and modest benefits from continued training. Wrong-target experiments establish effects of training targets, not a causal mediation pathway through a specific internal property. The calibrated control changes feature coordinates more than profile reversal but changes geometry less; representational geometry may therefore mediate some effects. Our decomposition accounts for performance changes without identifying their internal causes. Finally, full activation patching can move activations outside their natural range (Zhang & Nanda, 2024; Grant et al., 2026), although partial patching yielded nearly the same profiles (Appendix D).

## Reproducibility statement

Synthetic stimuli use public generation resources; CIFAR-100 uses its standard split, and Appendix B specifies the stratified ImageNet-R split. We will publicly release code, per-run results, figure notebooks, and figure source data upon acceptance. Preservation strengths were selected using validation data from pre-specified development orders. Confirmation orders and the representation-perturbation calibration rule were fixed before their results were examined.

## Ethics statement

This study does not collect or analyze human or animal experimental data.

## AI USE STATEMENT

We used large language models for polishing writing, literature discovery, identification of open-source repositories, and code checking and revision. The authors designed the research, reviewed the resulting content, and take full responsibility for the work.

## REFERENCES

Nader Asadi, MohammadReza Davari, Sudhir Mudur, Rahaf Aljundi, and Eugene Belilovsky. Prototype-sample relation distillation: Towards replay-free continual learning. In *International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, pp. 1093–1106, 2023.

Danielle S. Bassett, Nicholas F. Wymbs, Mason A. Porter, Peter J. Mucha, Jean M. Carlson, and Scott T. Grafton. Dynamic reconfiguration of human brain networks during learning. *Proceedings of the National Academy of Sciences*, 108(18):7641–7646, 2011.

Danielle S. Bassett, Muzhi Yang, Nicholas F. Wymbs, and Scott T. Grafton. Learning-induced autonomy of sensorimotor systems. *Nature Neuroscience*, 18:744–751, 2015. doi: 10.1038/nn.3993.

Lucrezia Carboni, Dwight Nwaigwe, Marion Mainsant, Raphael Bayle, Marina Reyboz, Martial Mermillod, Michel Dojat, and Sophie Achard. Exploring continual learning strategies in artificial neural networks through graph-based analysis of connectivity: Insights from a brain-inspired perspective. *Neural Networks*, 185:107125, 2025.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, 2021.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K. Dokania, Philip H. S. Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 794–803, 2018.

Andrea Cossu, Antonio Carta, Lucia Passaro, Vincenzo Lomonaco, Tinne Tuytelaars, and Davide Bacciu. Continual pre-training mitigates forgetting in language and vision. *Neural Networks*, 179:106492, 2024a.

Andrea Cossu, Francesco Spinnato, Riccardo Guidotti, and Davide Bacciu. Drifting explanations in continual learning. *Neurocomputing*, 597:127960, 2024b.

MohammadReza Davari, Nader Asadi, Sudhir Mudur, Rahaf Aljundi, and Eugene Belilovsky. Probing representation forgetting in supervised and unsupervised continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16712–16721, 2022.

Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. Learning without memorizing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Shibhansh Dohare, J. Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A. Rupam Mahmood, and Richard S. Sutton. Loss of plasticity in deep continual learning. *Nature*, 632:768–774, 2024.

Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. PODNet: Pooled outputs distillation for small-tasks incremental learning. In *European Conference on Computer Vision (ECCV)*, 2020.

Sayna Ebrahimi, Suzanne Petryk, Akash Gokul, William Gan, Joseph E. Gonzalez, Marcus Rohrbach, and Trevor Darrell. Remembering for the right reasons: Explanations reduce catastrophic forgetting. In *International Conference on Learning Representations (ICLR)*, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, 2021.

Alex Gomez-Villa, Bartlomiej Twardowski, Lu Yu, Andrew D. Bagdanov, and Joost van de Weijer. Continually learning self-supervised representations with projected functional regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3867–3877, 2022.

Satchel Grant, Simon Jerome Han, Alexa R. Tartaglini, and Christopher Potts. Addressing divergent representations from causal interventions on neural networks. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2511.04638.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Timm Hess, Eli Verwimp, Gido M. van de Ven, and Tinne Tuytelaars. Knowledge accumulation in continually learned representations and the issue of feature forgetting. *Transactions on Machine Learning Research*, 2024.

Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 831–839, 2019.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L. Richter, Quentin Anthony, Timothée Lesort, Eugene Belilovsky, and Irina Rish. Simple and scalable strategies to continually pre-train large language models. *Transactions on Machine Learning Research*, 2024.

Paul Janson, Wenxuan Zhang, Rahaf Aljundi, and Mohamed Elhoseiny. A simple baseline that questions the use of pretrained-models in continual learning. In *NeurIPS Workshop on Distribution Shifts*, 2022. arXiv:2210.04428.

Dongwan Kim and Bohyung Han. On the stability-plasticity dilemma of class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20196–20204, 2023.

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. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1920–1929, 2019.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, 2019.

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

Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Aran Nayebi, Daniel Bear, Daniel L. K. Yamins, and James J. DiCarlo. Brain-like object recognition with high-performing shallow recurrent ANNs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018.

Yan-Shuo Liang and Wu-Jun Li. InfLoRA: Interference-free low-rank adaptation for continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23638–23647, 2024.

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

Divyam Madaan, Jaehong Yoon, Yuanchun Li, Yunxin Liu, and Sung Ju Hwang. Representational continuity for unsupervised continual learning. In *International Conference on Learning Representations*, 2022.

Hiroshi Makino and Takaki Komiyama. Learning enhances the relative impact of top-down processing in the visual cortex. *Nature Neuroscience*, 18:1116–1122, 2015. doi: 10.1038/nn.4061.

Sergi Masip, Gido M. van de Ven, Javier Ferrando, and Tinne Tuytelaars. Putting a face to forgetting: Continual learning meets mechanistic interpretability. In *Conference on Lifelong Learning Agents (CoLLAs)*, 2026.

Mark D. McDonnell, Dong Gong, Amin Parvaneh, Ehsan Abbasnejad, and Anton van den Hengel. RanPAC: Random projections and pre-trained models for continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 2022.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Razvan Pascanu, and Hassan Ghasemzadeh. Understanding the role of training regimes in continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Vinay V. Ramasesh, Ethan Dyer, and Maithra Raghu. Anatomy of catastrophic forgetting: Hidden representations and task semantics. In *International Conference on Learning Representations (ICLR)*, 2021.

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

Jeanmely Rojas Nunez, Viraj Sawant, Nathan Allen, Nomgondalai Amgalanbaatar, Yannis Zongo, Vasu Sharma, and Maheep Chaudhary. Mechanistic origins of catastrophic forgetting: why RL preserves circuits better than SFT? In *ICML Workshop on Continual Adaptation at Scale (CATS)*, 2026. arXiv:2605.28860.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Greg Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Dawid Rymarczyk, Joost van de Weijer, Bartosz Zieliński, and Bartłomiej Twardowski. ICICLE: Interpretable class incremental continual learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1887–1898, 2023.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, 58(5):120, 2025. doi: 10.1145/3735633.

James Seale Smith, Junjiao Tian, Shaunak Halbe, Yen-Chang Hsu, and Zsolt Kira. A closer look at rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CLVision)*, 2023.

Mohammad Ali Vahedifar, Abhisek Ray, and Qi Zhang. Shapley neuron values for continual learning: Which neurons matter most? In *International Conference on Machine Learning (ICML)*, 2026.

Giuseppe Vecchio and Valentin Deschaintre. MatSynth: A modern PBR materials dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, 2020.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5362–5383, 2024. doi: 10.1109/TPAMI.2024.3367329.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Desai Xie, Sai Bi, Zhixin Shu, Kai Zhang, Zexiang Xu, Yi Zhou, Sören Pirk, Arie Kaufman, Xin Sun, and Hao Tan. LRM-Zero: Training large reconstruction models with synthesized data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024.

Da-Wei Zhou, Hai-Long Sun, Han-Jia Ye, and De-Chuan Zhan. Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need. *International Journal of Computer Vision*, 2024a. arXiv:2303.07338.

Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Class-incremental learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12), 2024b. doi: 10.1109/TPAMI.2024.3429383.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations (ICLR)*, 2022.

## A CONTROLLED-EXPOSURE SETTING

**Stimuli.** We generated objects with the public Zeroverse procedure (Xie et al., 2024) and rendered them using one public MatSynth material (Vecchio & Deschaintre, 2024), deliberately restricting the original distribution. Each of twelve stimulus sets contained 32 target objects and 32 lures, rendered from 32 camera positions. Each lure had independently generated geometry and shared its target's camera positions. Views 0–7 were used for learning and views 8–15 for retrieval candidates. Views 16–23 and 24–31 formed two held-out query families. Targets were divided into four sets of eight identities. Sets A and B were learned; sets C and G supplied source images for activation patching. Held-out query images served as base images.

**Learning from experience.** Two copies of a pretrained model started from a common state and learned set A or set B. Each identity was therefore evaluated in a directly trained copy and a copy without direct experience of that identity. Cross-view positive pairs contained different views of an identity; same-view positive pairs contained two augmentations of one view. Both conditions received identical inputs and differed only in positive-pair assignment. Training used a symmetric contrastive loss (temperature 0.1), with additional self-distillation and masked-patch terms for DINOv2-B. Each copy received 480 updates. We measured the effect of experience on a quantity $Y$ using the reciprocal difference-in-differences $\tau(Y) = \frac{1}{16}\sum_i \left[(Y_{i,\text{learned}} - Y_{i,\text{holdout}})\right]_{\text{cross}} - \frac{1}{16}\sum_i \left[(Y_{i,\text{learned}} - Y_{i,\text{holdout}})\right]_{\text{same}}$, which compares the same identities across copies and removes their shared starting point. Cross-view pairing can be learned only by linking different views of an object, that is, its three-dimensional identity, whereas same-view pairing receives the same images and can be learned from two-dimensional image features; their difference therefore isolates learning of objects across views.

**Intervention-defined organization.** DINOv2-B components are its 144 attention heads, patched before the attention output projection. The causal-profile assay uses full replacement ($\alpha = 1$); the effective-network assay uses $\alpha = 0.2$ between adjacent blocks. Full replacement, the standard form of activation patching, measures each component's total contribution to the final representation, and partial patching yields nearly the same profiles (Appendix D). A smaller patch for the network measures each receiver's sensitivity to a moderate change in one sender while keeping the next stage closer to its natural activity. CORnet-S profiles contain all 2,368 channels of its nine processing steps: V1, two V2 steps, four V4 steps and two IT steps. Each component is patched fully ($\alpha = 1$) with a same-view source image from set C, and profiles of both models use Eq. 1. The retrieval margin is the cosine similarity to the correct candidate minus that to the paired lure.

CORnet-S effective networks are measured between individual channels of adjacent processing steps. Each sender channel's activation moves 20% toward its source-image activation ($\alpha = 0.2$), with the perturbation propagated through the model's computation. The receiver effect is the root-mean-square change in its spatial activation map. Identity specificity subtracts similarity to the paired lure from same-identity similarity. For profiles, similarity is the Pearson correlation between view-specific profiles within a held-out query family, averaged across view pairs and then across the two families; for effective networks, it compares the same identity across the two query families, using sources from C and G. Cross-view pairing improved identification and organization measures for directly learned identities (Figures 1b and S1).

**Continual learning and methods.** For the association analysis, a model first learned set A and stored its state and targets on A anchor images. It then continued learning set B for 480 updates. Methods included fine-tuning, experience replay of A images, CaSSLe, EWC, L2, and representation, causal-profile and effective-network preservation. For Figure 1c, the conventional methods used the controller of Eq. 2 at strengths selected by mean familiar- and new-set accuracy on six stimulus sets (EWC $\gamma = 4$, L2 1, CaSSLe 2, representation 4). Preservation methods used stored anchors and targets without the A training loss. The network objective preserved the centered, unit-length pattern of each sender channel's effects on IT, referenced to the post-A state. The representation objective preserved normalized anchor representations. The causal-profile objective preserved each anchor pair's standardized representation-effect profile, with gradients propagated through standardization. Retention was the familiar-object retrieval-margin change from the post-A state, with candidate prototypes fixed at post-A. Acquisition was the corresponding change for B objects. With coefficients fixed for the whole run, and representation preservation calibrated so that its gradient was a quarter of the task gradient at a common perturbation of the post-A model, fine-tuning on B lowered the familiar-set margin by 0.149; effective-network and representation preservation reduced this loss by 0.117 and 0.113, respectively, with benefits in every stimulus set. Causal-profile preservation, with the controller of Eq. 2 at $\gamma = 0.2$, reduced it by 0.095, also in every stimulus set.

**Preservation levels and inference.** Parameter preservation was assessed from post-A displacement, weighted by a diagonal Fisher estimate at the post-A state. Representation preservation used the cosine similarity of held-out familiar-query representations to their post-A values. Profile preservation used the centered cosine between Eq. 1 profiles of held-out familiar images, with C-set source images, and their post-A values. Effective-network preservation used the centered row cosine of the IT channel network on the same images and its post-A state. Quantities were averaged within each

stimulus set before inference. We used exact sign-flip tests across twelve stimulus sets, bootstrap intervals over stimulus sets for the associations, and $t$ intervals for the effects of cross-view learning (Figures 1b and S1).

## B CONTINUAL SELF-SUPERVISED LEARNING: IMPLEMENTATION

**Data and training.** ImageNet-R's 200 classes were divided into ten tasks of twenty classes, with training, validation and test portions of 64%, 16% and 20%. CIFAR-100 was divided into five tasks using its standard split. Each class order defined one replication and was shared across methods. DINOv2-B/14 continued self-supervised training with freshly initialized DINO and iBOT heads. Each task received 200 updates with batch size 8, two global crops and four local crops. AdamW used backbone and head learning rates of $2 \times 10^{-6}$ and $10^{-4}$, respectively. After each task, we fitted an all-class linear probe using up to 500 training images per class. Probes were evaluated on the test split and, for strength selection, the validation split.

**Memory.** Memory-based methods store the same images, with method-specific stored state and sampling (Table S1). Following the per-task allocation of GEM (Lopez-Paz & Ranzato, 2017), $200/(2T)$ disjoint image pairs are stored per completed task and never replaced. Intervention-based regularizers use each pair in both directions, exchanging base and source roles; ER and representation preservation use its two images. Memory therefore reaches 200 images only after the final task. Labels determine tasks, class-balanced memories and cross-class pairs, but do not enter the self-supervised loss.

Table S1: **Method specifications and resource accounting for continual self-supervised learning.** Methods share the optimizer, eight current images per update, 200 updates per task and the probe protocol. The gradient-pressure controller sets preservation strengths (Eq. 2); ER uses a fixed weight for its stored-image self-supervised loss. Here, $t$ denotes the number of completed tasks. Intervention-based regularizers store two images per pair and a separate target for each assignment of base and source roles.

| Method | Preservation term | Target and its update | Stored images while training task $t+1$ | Additional state | Sampled per update |
|---|---|---|---|---|---|
| Fine-tuning | – | – | 0 | – | 8 current images |
| ER | Self-supervised loss on stored images (weight 0.5) | Images only | $20t$ (ImageNet-R), $40t$ (CIFAR-100) | – | + 8 stored images |
| CaSSLe | Predictor distillation to the previous encoder on current images | Previous encoder, replaced at every task boundary | 0 | Encoder copy, predictor | Current images |
| EWC | Fisher-weighted distance to reference weights | Weights at the latest boundary; Fisher summed over tasks (8 images each) | 0 | Reference weights, Fisher diagonal | All parameters |
| L2 | Uniform distance to reference weights | Weights at the latest boundary | 0 | Reference weights | All parameters |
| Representation | Distance to stored representations | Computed when the image is stored; never updated | $20t$ (ImageNet-R), $40t$ (CIFAR-100) | One 768-d target per image | 4 stored images |
| Causal profile | Difference of standardized 144-head profiles | Computed when the pair is stored; never updated | $20t$ (ImageNet-R), $40t$ (CIFAR-100) | One 144-value profile per pair direction | 1 stored pair, all 144 heads |
| Effective network | Row-wise cosine distance | Computed when the pair is stored; never updated | $20t$ (ImageNet-R), $40t$ (CIFAR-100) | One $132 \times 12$ matrix per pair direction | 1 stored pair, 4 of 132 rows |

**Objectives and probe.** Representations are the last layer's L2-normalized class tokens. ER adds a self-supervised loss on eight stored images with weight 0.5. CaSSLe applies a two-layer predictor to each current representation and minimizes $2 - 2\cos$ against the previous encoder's representation of the same image. EWC minimizes $\frac{1}{2}\sum_i F_i(\theta_i - \theta_i^*)^2$, where $\theta^*$ denotes weights at the latest task boundary. The Fisher estimate $F$ sums squared self-supervised-loss gradients over completed tasks. Each task contributes gradients from eight images evaluated using its post-task model. L2 assigns uniform weights to parameter differences. Representation preservation minimizes the mean squared difference between current and stored representations of four stored images.

Causal-profile preservation minimizes the mean squared difference between current and stored standardized profiles for one stored pair. Profiles are mean-centered and scaled to unit root-mean-square across the 144 heads, with gradients propagated exactly through standardization. Effective-network preservation samples four rows from one stored pair's matrix. Each row contains the responses

of the next block's 12 receiver heads to one sender head. The objective minimizes one minus the cosine similarity between current and stored rows. These rows are uncentered, whereas the effective-network similarity measure uses centering. At each update, the controller in Eq. 2 scales each preservation term, and the combined gradient is clipped to norm 3.

After each task, an all-class linear classifier is trained on representations from up to 500 training images per class. Training uses 500 full-batch stochastic-gradient-descent steps, with learning rate 1, momentum 0.9 and weight decay $10^{-4}$. The classifier is then evaluated on the test split.

**Preserving a subset of heads.** We ranked heads by influence at the reference state. At the selected strength ($\gamma = 1$), preserving the most influential half or quarter did not outperform preserving all heads. Accuracy differences were $-0.15$ and $-1.05\%$, respectively, across six class orders. These subsets also did not differ from equally sized, layer-matched random subsets ($+0.05$ and $-0.14\%$, respectively).

**Selection.** Strengths were selected by validation accuracy on class orders 1–6 (Table S2; Figure S4). Class orders 7–18 were reserved for confirmation. Held-out measurements use task-1 images stored by no method and compare the final model with its post-task-1 state. In the Split-ImageNet-R strength sweep, stronger profile constraints allowed feature coordinates to drift while retaining representational geometry. Between $\gamma = 1$ and 16, the representation-cosine gain over fine-tuning fell from $0.034$ to $0.005$, while the CKA gain remained near $0.036$.

**Paired contrasts and learning dynamics.** Table S3 reports paired contrasts supporting the main claims. Figure S6 tracks accuracy on the complete task set, holding the evaluated tasks fixed across stages. It also shows the accuracy decomposition for Split-ImageNet-R.

Table S2: **Strength selection on six Split-ImageNet-R development class orders.** The selected strength $\gamma^*$ maximizes mean validation accuracy (%) after the final task. Strengths transfer unchanged to Split-CIFAR-100 and the twelve confirmation class orders.

| Method | Tested $\gamma$ | $\gamma^*$ | Validation accuracy at $\gamma^*$ |
|---|---|---|---|
| Fine-tuning | – | – | 65.80 |
| ER | 200 img | 200 img | 67.19 |
| CaSSLe | 0.02, 0.05, 0.1, 0.2 | 0.05 | 66.79 |
| EWC | 0.05, 0.2, 0.4, 1, 2, 4, 8, 16, 32 | 32 | 71.16 |
| L2 | 0.5, 1, 2, 4, 8, 16 | 16 | 71.37 |
| Representation | 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 4 | 0.2 | 70.44 |
| Causal profile | 0.05, 0.1, 0.2, 0.5, 1, 2, 4, 8, 16 | 1 | 70.11 |
| Effective network | 0.05, 0.1, 0.2, 0.5, 1, 2, 4, 8, 16 | 4 | 68.73 |

## C WRONG-TARGET EXPERIMENTS

**Screen.** We screened seven combinations on three development class orders. Five combined representation preservation ($\gamma = 0.5$) with profile targets shuffled across all components, within layers, across layers, sign-reversed, or shuffled at four times the strength. Two combined a shuffled profile target with stronger representation preservation or L2 regularization. Stronger anchors almost completely suppressed the wrong target's effects. Sign-reversed and within-layer shuffled targets were carried forward to class orders 4–12 (Figure 2c).

**How far each rearrangement moves the target.** We recomputed the pretrained model's network and profile targets on held-out image pairs and applied each rearrangement. Different sender heads evoked similar receiver patterns, with a mean cosine of $0.96$ between rows of a transition. Mean row cosines between rearranged and correct network targets were $0.98$ for the strength-preserving null and $0.96$ for permuted sender rows. Permuting receiver columns and shuffling within rows both yielded $0.92$; transposition and shuffling all entries yielded $0.87$ and $0.86$, respectively. Profile cosines with the correct target were $0.12$ after shuffling all heads, $0.45$ after shuffling within layers and $0.12$ after shuffling whole layers. Figure S2 compares the other rearranged standalone

Table S3: **Essential paired contrasts.** Final accuracy differences are measured in % and paired by class order. Columns report the mean, 95% $t$ interval, number of class orders favoring the first method, and exact sign-flip $p$. Twelve class orders allow $p \geq 0.0005$.

| Contrast (accuracy, %) | $\Delta$ | 95% interval | Improved | $p$ |
|---|---|---|---|---|
| *Split-CIFAR-100, twelve class orders* | | | | |
| Causal-profile preservation − frozen pretrained encoder | +0.78 | [+0.61, +0.95] | 12/12 | 0.0005 |
| Causal-profile preservation − its model after task 1 | +0.86 | [+0.73, +0.99] | 12/12 | 0.0005 |
| Causal-profile preservation − EWC | +0.78 | [+0.61, +0.94] | 12/12 | 0.0005 |
| Causal-profile preservation − L2 | +0.83 | [+0.67, +0.98] | 12/12 | 0.0005 |
| Causal-profile preservation − ER | +0.67 | [+0.35, +0.99] | 11/12 | 0.0010 |
| Representation − causal-profile preservation | −0.01 | [−0.14, +0.12] | 6/12 | 0.878 |
| Representation preservation − frozen pretrained encoder | +0.77 | [+0.61, +0.93] | 12/12 | 0.0005 |
| Representation preservation − its model after task 1 | +0.80 | [+0.60, +1.00] | 12/12 | 0.0005 |
| Effective-network preservation − fine-tuning | +0.22 | [−0.12, +0.57] | 7/12 | 0.185 |
| *Split-ImageNet-R, twelve confirmation orders* | | | | |
| Causal-profile preservation − fine-tuning | +4.77 | [+4.19, +5.35] | 12/12 | 0.0005 |
| Effective-network preservation − fine-tuning | +3.56 | [+3.02, +4.11] | 12/12 | 0.0005 |
| ER − fine-tuning | +1.98 | [+1.40, +2.56] | 11/12 | 0.0010 |
| L2 − EWC | +0.19 | [+0.07, +0.32] | 11/12 | 0.002 |
| EWC − representation preservation | +1.26 | [+0.83, +1.68] | 12/12 | 0.0005 |
| Representation − causal-profile preservation | −0.02 | [−0.40, +0.37] | 7/12 | 0.930 |
| Causal-profile − effective-network preservation | +1.20 | [+1.02, +1.39] | 12/12 | 0.0005 |

targets with their correct-target controls, and Figure 2b shows the strength-preserving null. Each rearrangement produced lower final accuracy than its correct target in every class order.

**Calibrated representation-perturbation control.** This control retains the correct profile target and permutes a fraction $f$ of each stored representation target's coordinates. We tested $f \in \{0.05, 0.1, 0.2, 0.4, 1\}$ on three development class orders. A prespecified rule selected the fraction whose representation cosine and CKA were closest to those of the sign-reversed arm, without examining accuracy. The selected fraction, $f = 0.4$, was then evaluated on class orders 4–12. On screening orders 1–3, final-versus-post-task-1 representation cosine and CKA were $0.916$ and $0.966$, respectively. Corresponding values for the sign-reversed arm were $0.932$ and $0.934$.

## D Validity of the interventions

We assessed intervention sensitivity in an earlier Split-ImageNet-R cohort using the same continual self-supervised protocol and eight class orders. We included one run per method and class order. CaSSLe used its selected strength, $\gamma = 0.05$; other regularized methods used the cohort's common strength, $\gamma = 0.2$. Partial activation patching yielded profiles close to those obtained with full replacement, with pattern cosines of $0.96$ at 25% and $0.98$ at 50%. Method rankings were unchanged (Figure S7a). Fully patching a head with the median effect changed the normalized representation less than augmenting the same image. Even the largest head effect was smaller than the change induced by another image of the same class (Figure S7b). These sensitivity checks do not establish that patched activations follow the natural activation distribution.

## E Supplementary figures

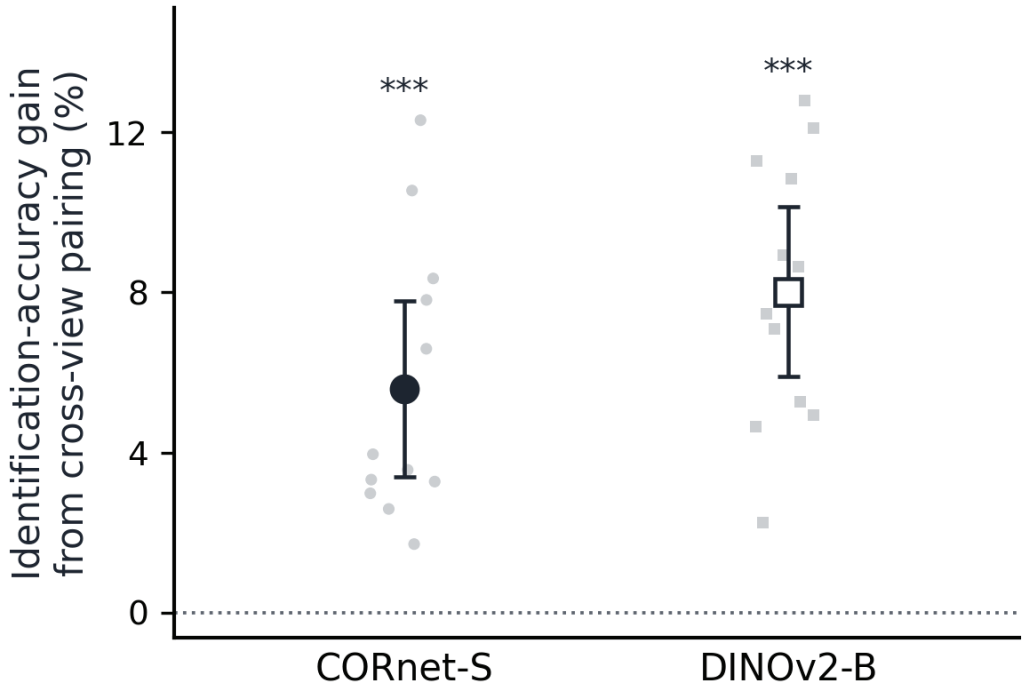



Figure S1: **Cross-view pairing improves identification of learned objects in the controlled-exposure setting.** Gain in identification accuracy on held-out views from cross-view pairing, in %, across twelve stimulus sets per model. Learning effects compare the directly trained copy with the copy that did not see the identity; the plotted contrast subtracts the corresponding effect under the within-view control, which receives identical inputs. Gray dots denote stimulus sets; black symbols show means with 95% $t$ intervals; marks indicate exact sign-flip tests against zero. The corresponding changes in causal-profile and effective-network specificity appear in Figure 1b.

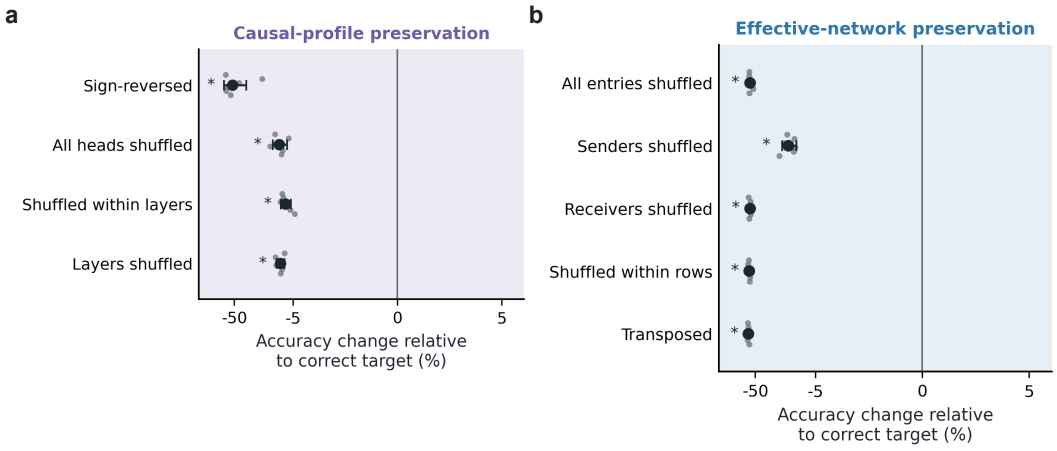



Figure S2: **Effects of every target rearrangement.** Standalone causal-profile (**a**) and effective-network (**b**) objectives at their selected strengths are compared with their correct-target controls. Final accuracy differences are measured in % and paired across Split-ImageNet-R class orders 1–6. Profile targets comprise sign reversal of the standardized profile, shuffling all heads, shuffling heads within layers, and shuffling whole layers. Network targets comprise shuffling all entries, permuting sender rows, permuting receiver columns, shuffling within rows, and transposition; the strength-preserving null appears in Figure 2b. The horizontal axis is symmetric logarithmic. Comparisons with fine-tuning appear in Figures 2b and S3.

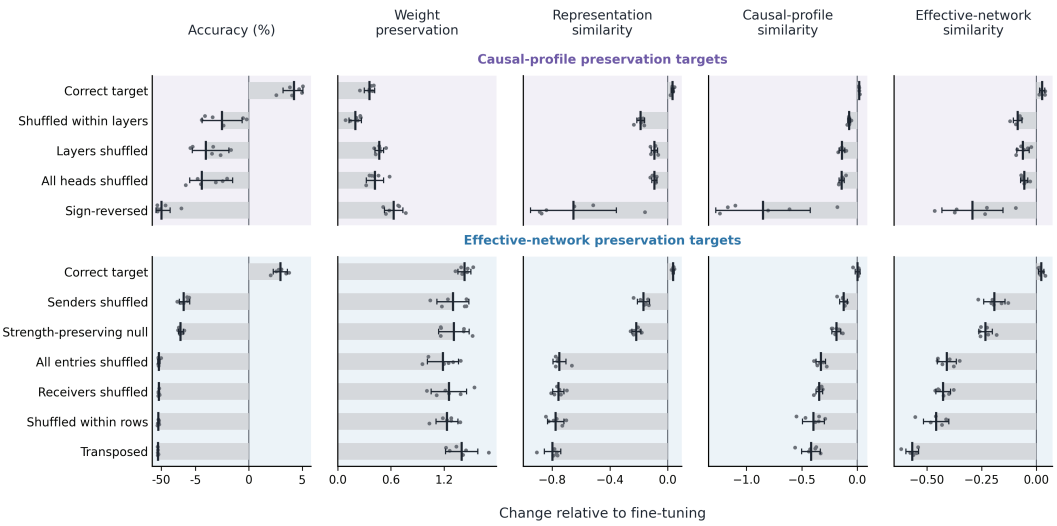



Figure S3: **Preserved measures and accuracy across preservation targets.** Standalone causal-profile ($\gamma = 1$) or effective-network ($\gamma = 4$) preservation with the correct stored target or a rearranged one, on Split-ImageNet-R class orders 1–6. Rows are ranked by final accuracy within each family; columns give accuracy and each measure of Figure 1c for the same runs, all relative to fine-tuning in the same class order. A measure that decreases down the rows in the same order as accuracy follows accuracy across targets. Gray dots denote class orders; bars and black lines show means with 95% intervals. Accuracy uses a symmetric logarithmic axis, linear within $\pm 5\%$.

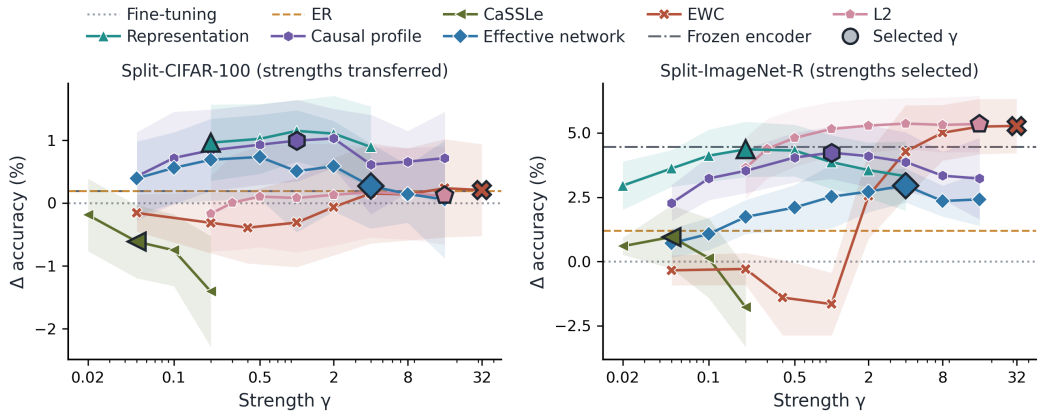



Figure S4: **Preservation-strength sensitivity.** Final accuracy gains over fine-tuning, in %, are plotted against preservation strength $\gamma$. Curves show means with 95% intervals over class orders 1–6. Split-CIFAR-100 appears on the left; strengths in Table 1 were transferred unchanged. Split-ImageNet-R appears on the right; strengths were selected by validation accuracy (Table S2). Enlarged markers with black edges identify the strengths used in the main comparisons. Reference lines show fine-tuning, ER and the frozen pretrained encoder on the same class orders.

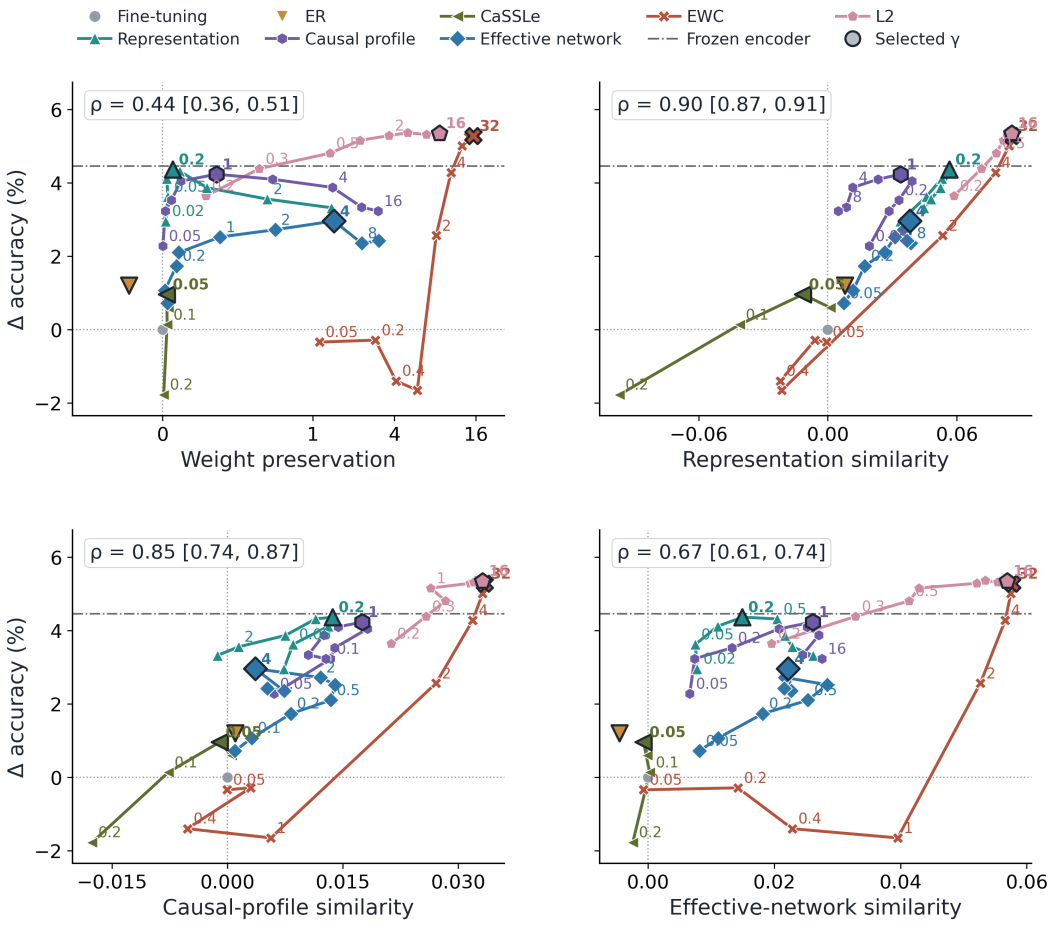



Figure S5: **Preserved measures and accuracy across preservation strengths.** Every method at every tested strength on Split-ImageNet-R development class orders 1–6: the change in each measure of Figure 1c against the change in final accuracy, both relative to fine-tuning in the same class order and averaged over orders. Lines join successive strengths of one method, and numbers give $\gamma$; larger markers with black edges identify the strengths used in the main comparisons. Weight preservation uses a symmetric logarithmic axis, linear within $\pm 1$. $\rho$ denotes the Spearman correlation across all method–strength points, with 95% intervals from resampling class orders. Dash-dot lines show the frozen encoder; dotted lines show fine-tuning.

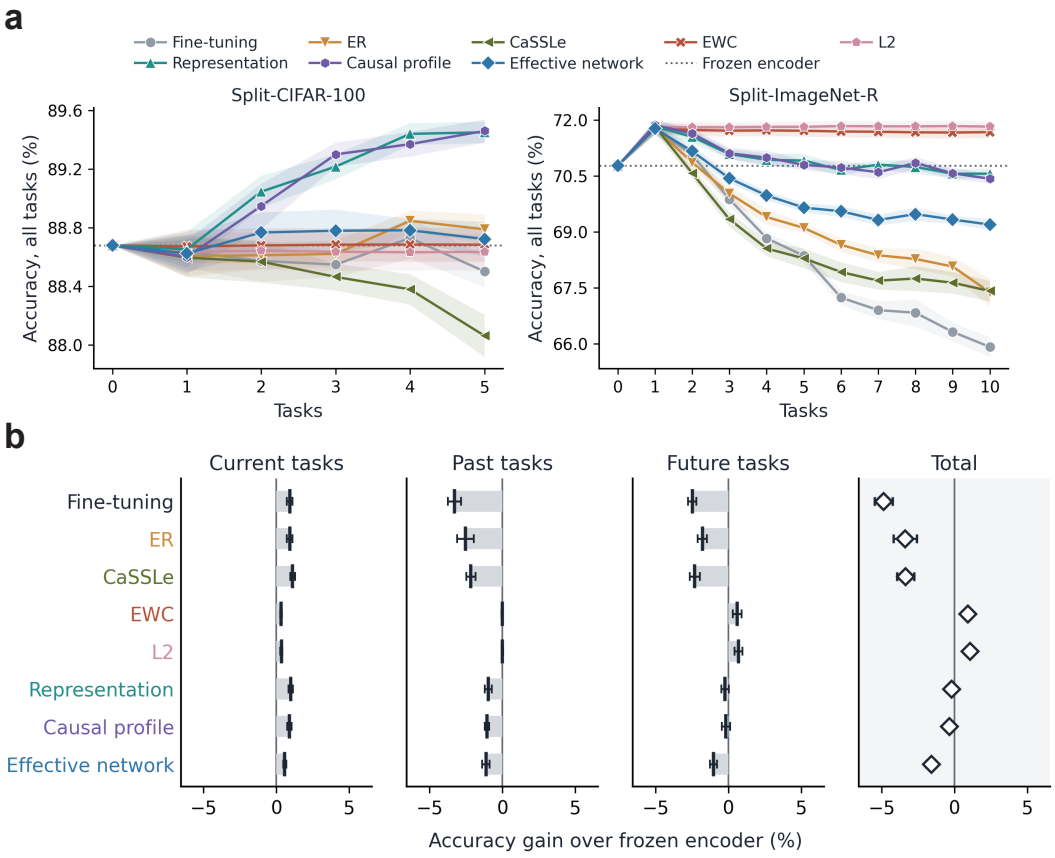



Figure S6: **Learning dynamics beyond Figure 3. (a)** Mean accuracy on the complete task set after every stage, shown as mean ± s.e.m. Split-CIFAR-100 appears on the left and Split-ImageNet-R on the right. Stage 0 denotes the pretrained encoder (dotted reference). The evaluated task set remains fixed across stages, complementing the seen-task average in Figure 3a. **(b)** The decomposition in Figure 3b applied to Split-ImageNet-R. Contributions sum to the final accuracy gain over the frozen pretrained encoder and are measured in %.

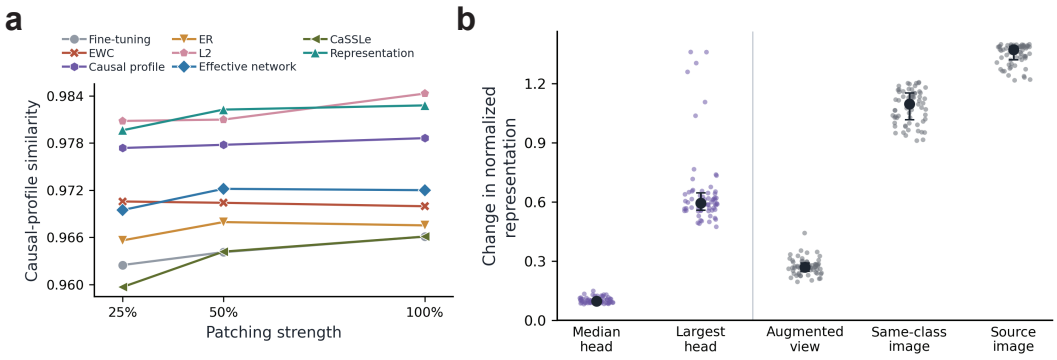



Figure S7: **Validity of the intervention measurements.** Sensitivity checks use an earlier Split-ImageNet-R cohort of the same self-supervised protocol, with DINOv2-B and eight class orders. One run is included per method and class order. CaSSLe uses its selected strength, $\gamma = 0.05$; other regularized methods use the cohort's common strength, $\gamma = 0.2$. **(a)** Causal-profile similarity between final and post-task-1 models, averaged over class orders. Head outputs move 25%, 50% or fully toward their source-image values. **(b)** Normalized final-representation changes under full patching of heads with the median or largest effects among each run's 144 heads (purple). Natural-variation references use an augmented view of the same image, another image of the same class and the source image (gray). Black dots and bars show medians and interquartile ranges across runs.