RECOVERABLE BUT NOT PRESERVED: HIDDEN FORGETTING IN CONTINUAL SELF-SUPERVISED MEDICAL PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual self-supervised pretraining is a natural strategy for keeping medical foundation encoders current, yet its standard evaluation protocol, fine-tuning the updated encoder on a source task, can itself recover degraded representations and thereby obscure underlying forgetting. Continuing CheXWorld, a JEPA-style chest-radiograph ViT, on RSNA Pneumonia leaves NIH ChestX-ray14 fine-tuning performance nearly unchanged (79.6–80.1 mAUROC), while a frozen linear probe on the same checkpoints loses 1.6 points. These trends diverge consistently across three training seeds and individual pathologies: Nodule, Pneumothorax, and Emphysema lose 3.5–4.2 probe points that fine-tuning largely recovers. This degradation occurs even though RSNA images are drawn from NIH and grows to 12.5 points after a single PadChest hop. We then ask what mechanisms preserve frozen representational accessibility. Prediction surprise, used to gate a fast/slow EMA dual memory, is a valid *plasticity* signal, tracking which samples induce larger representation changes ($\rho = 0.51$), but per-sample target arbitration retains no more than a constant gate matched to its mean. More broadly, methods whose targets track the learner single, dual, gated, or rate-modulated converge to similar retention (77.8–78.2 mAUROC), whereas methods that explicitly constrain the learner toward the pre-update model retain up to 80.0 mAUROC in single-seed comparisons; EWC exhibits the same trend as regularization strength increases. Finally, changing only the probe early-stopping rule shifts the estimated retention by 2.4 points. Together, these results distinguish recoverability from preservation and plasticity from vulnerability, and motivate a reporting protocol for evaluating continually updated encoders.

1 INTRODUCTION

Foundation encoders are increasingly updated rather than retrained. In medical imaging, new scanners, sites, and populations continually produce unlabeled data that can be absorbed through additional self-supervised pretraining (Gururangan et al., 2020; Cossu et al., 2024). The evaluation question looks simple: *after the update, did the model keep what it knew?*

The answer depends on what "keep" means. The common test fine-tunes the entire updated encoder on a source task, changing the representation it is meant to assess. High post-update accuracy is then consistent with source information having remained accessible (*preservation*) or with supervised optimization having rebuilt it afterwards (*recoverability*). We measure preservation without updating the encoder, principally with a refitted linear probe, and use the term operationally rather than as a claim of irreversible information loss. The distinction matters whenever an updated encoder is used frozen, shared by several heads, or served as a reusable foundation representation. "Hidden forgetting" denotes a decrease in frozen accessibility that an evaluation based only on fine-tuning would miss.

Using CheXWorld (Yue et al., 2025), we show that the two evaluations reach opposite conclusions on the same checkpoints (Figure 1). We then follow four dissociations, each of which breaks a scalar "stability plasticity" view into quantities that must be measured separately:

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.

- **Recoverability vs. preservation.** Along an RSNA continuation, the NIH probe falls from 79.62 to 78.06 mAUROC while fine-tuning stays within 79.63–80.13; the two trends oppose each other in all three training seeds, and fine-tuning largely restores the very pathologies with the largest probe losses.
- **Shift vs. vulnerability.** Hidden forgetting does not require a distribution shift: RSNA consists of reannotated NIH images and is nearly indistinguishable from NIH in representation space, yet it costs more frozen accessibility than the more distant COVIDx. A single PadChest hop costs 12.52 points without a matching gain on its own task.
- **Adaptation signal vs. preservation mechanism.** Prediction surprise has reproducible pathology structure and tracks which images move, but routing targets by surprise retains no more than a mean-matched constant: a signal for *where to learn* is not a mechanism for *what to keep*.
- **Tracking vs. anchoring references.** Every target we test that is an average of the current learner ends within a 0.44-point band, whatever its timescale or gating. The methods that retain more are those that constrain the learner toward a fixed copy of the pre-update model.

Contributions. (1) A replicated demonstration that full fine-tuning and frozen probing disagree about retention during continual SSL, down to individual pathologies, with evidence that the probe decline is not a readout artifact (§4). (2) A three-domain stress test, including single-hop controls, showing that frozen-accessibility loss is neither explained by representation-space shift nor tied to target-domain benefit (§5). (3) A controlled dual-memory intervention (SGDM) with a mean-matched control, separating an active plasticity signal from a retention benefit and showing that timescale separation alone does not anchor a representation (§6–7). (4) A reporting protocol for continually updated encoders (Table 3), motivated by a case in which a probe's early-stopping rule alone reversed a retention conclusion.

2 Related work

Measuring forgetting in representations. Davari et al. (2022) show with linear probes that representations can forget less than task performance suggests, and Ramasesh et al. (2021) localize forgetting to specific layers using representational similarity. Kumar et al. (2022) show that full fine-tuning distorts pretrained features, which is the mechanism that makes it a recovery procedure. Liu et al. (2026) report that apparently forgotten skills in vision–language–action models are rapidly recovered by fine-tuning. We study the complementary failure mode in continual *self-supervised* pretraining, where a frozen, reusable encoder is the point of the exercise: fine-tuning reports no loss while frozen accessibility declines.

Continual self-supervised and domain-adaptive pretraining. Self-supervised objectives forget less than supervised ones in several continual settings (Madaan et al., 2022; Hu et al., 2022; Cossu et al., 2024), although non-stationary streams remain challenging (Purushwalkam et al., 2022). Domain-adaptive pretraining improves in-domain transfer (Gururangan et al., 2020), and radiograph encoders are increasingly pretrained at scale on heterogeneous sources (Pérez-García et al., 2025). Mitigation methods either constrain parameters (EWC; Kirkpatrick et al., 2017), constrain outputs or features toward the previous model (LwF, CaSSLe, PFR; Li & Hoiem, 2017; Fini et al., 2022; Gomez-Villa et al., 2022), or replay earlier data (Rolnick et al., 2019). We use EWC, CaSSLe, and PFR as reference methods.

Fast and slow memories. Complementary Learning Systems theory motivates separate fast acquisition and slow consolidation (McClelland et al., 1995), and DualNet instantiates fast and slow learners (Pham et al., 2021). EMA teachers are central to self-distillation and JEPA-style SSL (Tarvainen & Valpola, 2017; Grill et al., 2020; Assran et al., 2023). Our intervention is deliberately minimal, two gradient-free EMA copies of one learner, so that it tests whether timescale separation alone provides a preservation reference.

3 SETTING AND EVALUATION CONTRACTS

Model and continuation. We start from CheXWorld's released epoch-229 ViT-Base checkpoint, which was pretrained on approximately 0.5M frontal radiographs from MIMIC-CXR, NIH ChestX-ray14, and CheXpert (Yue et al., 2025). CheXWorld is a JEPA-style model (Assran et al., 2023) with a context encoder, a predictor, and an EMA target encoder. We retain its self-supervised objectives and masking configuration. Continuation uses batch size 64, base learning rate 10^{-3}, two warm-up epochs, and weight decay increasing from 0.04 to 0.4. We reproduce CheXWorld's reported downstream results to within 0.41 points on four benchmarks (Appendix A).

The sequential stream is RSNA Pneumonia (26,684 images, 30 epochs), COVIDx CXR-4 (67,863 images, 20 epochs), and PadChest (94,407 images, 20 epochs) (Shih et al., 2019; Wu et al., 2023; Bustos et al., 2020). Single-hop COVIDx and PadChest controls start directly from the original checkpoint.

RSNA as a near-zero-shift control. RSNA consists of reannotated NIH images (Shih et al., 2019). The RSNA hop therefore continues pretraining on source-distribution data, and it is almost indistinguishable from NIH in the original representation space (domain-classifier accuracy 0.599; Appendix G). We treat this as a feature of the design: frozen-accessibility loss on this hop cannot be attributed to acquisition shift. Some RSNA images may coincide with NIH evaluation images; such unlabeled exposure would be expected to favor rather than harm their accessibility, making the observed decline conservative, but it precludes a claim of independent external validation for this hop.

Evaluation contracts. NIH ChestX-ray14 is the source benchmark (Wang et al., 2017); scores are macro-averaged AUROC (mAUROC, %), and differences are in points. We evaluate every checkpoint under two contracts:

- **Frozen probe (preservation).** A linear head is refitted on the frozen encoder: 30 epochs, two warm-up epochs, early-stopping patience 15, and test score at the best-validation epoch. The original checkpoint scores 79.62 in the reference fit and 79.61 ± 0.02 over three probe fits.
- **Full fine-tuning (recoverability).** All encoder layers are trainable, with 30 epochs, two warm-up epochs, patience 15, base learning rate 10^{-4}, uniform layer-wise decay, and minimum learning rate zero. We report the test score at the best-validation checkpoint. The original checkpoint scores 79.64.

The fine-tuning recipe is a single untuned recipe applied identically to every checkpoint. It is weaker than CheXWorld's original 50-epoch recipe, which reaches 83.24 in our reproduction (Appendix A). All claims compare changes *within* a contract, never absolute scores across contracts. A stronger recipe has more capacity to rebuild source features, so we would expect it to mask the probe decline at least as well; we have not verified this (§8).

Three secondary diagnostics characterize representation change without replacing task evaluation: debiased linear CKA to the original encoder (Kornblith et al., 2019), label-based retrieval precision@5, and a frozen-head output divergence (Appendix D). Neither geometric nor output drift alone implies worse task discrimination.

Replication. The RSNA audit evaluates seven seed-0 checkpoints and three training seeds at epochs 4, 19, and 29 under both contracts. SGDM ablations use three training seeds; we report the sample standard deviation across seeds, not patient-level intervals. EWC, CaSSLe, PFR, the chain, the single-hop controls, and the momentum sweep are single-seed, and we use them for direction and magnitude rather than ranking.

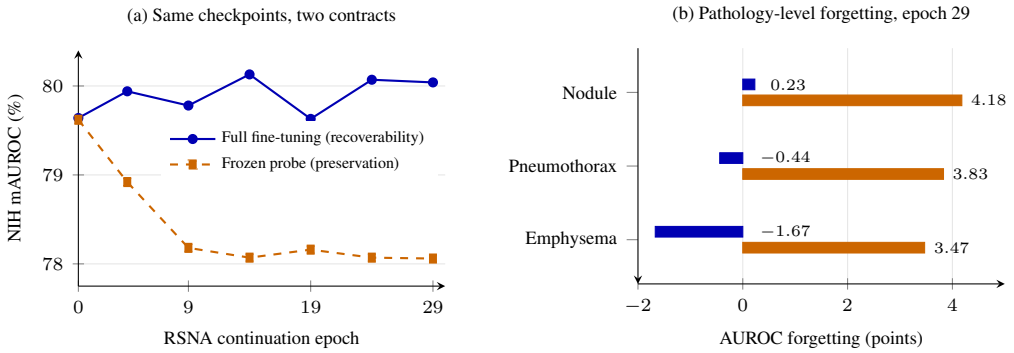

Figure 1: **Fine-tuning reports no forgetting while the frozen encoder loses source accessibility.** (a) NIH performance of the same seed-0 RSNA checkpoints under full fine-tuning (encoder updated during evaluation) and a refitted frozen probe (encoder fixed). (b) Epoch-29 per-pathology forgetting relative to the original checkpoint (three-seed means; positive = worse) under the probe (orange) and fine-tuning (blue). The three pathologies with the largest probe losses are recovered wholly or largely by fine-tuning. Values in Tables 5 and 18.

Table 1: NIH source-task evaluation across three RSNA training seeds (mean ± s.d., mAUROC). The original checkpoint scores 79.62 (probe) and 79.64 (fine-tuning). Per-seed values are in Appendix C.

Contract	Original	Epoch 4	Epoch 19	Epoch 29
Frozen probe	79.62	78.89 ± 0.04	78.00 ± 0.14	77.90 ± 0.15
Full fine-tuning	79.64	79.91 ± 0.48	80.03 ± 0.51	80.22 ± 0.40

4 Recoverability is not preservation

4.1 Fine-tuning and frozen probing disagree on the same checkpoints

Figure 1 (left) shows the seed-0 audit. Full fine-tuning fluctuates between 79.63 and 80.13 without a sustained decline. The frozen probe falls from 79.62 to 78.06, and 1.44 of the 1.56-point loss is already present at epoch 9: most of the damage occurs early. Table 1 shows that the disagreement replicates. From epoch 4 to 29, the paired probe changes are -0.86, -1.02, and -1.08 points in seeds 0, 1, and 2, while the fine-tuning changes are $+0.10$, $+0.52$, and $+0.30$. The directions disagree within every seed, so the conclusion does not depend on pooling seeds or on a shared baseline. Fine-tuning's between-seed spread (s.d. 0.40–0.51) exceeds these gains, so we read it as "no decline", not as improvement.

The disagreement is sharper within pathologies. At epoch 29, Nodule, Pneumothorax, and Emphysema lose 4.18, 3.83, and 3.47 probe points. Under fine-tuning, their forgetting is $+0.23$, -0.44, and -1.67 (Figure 1, right). A practitioner who accepted this update on fine-tuned scores would deploy a frozen encoder several points worse on exactly these findings. We establish an evaluation discrepancy, not the mechanism of recovery: fine-tuning may reuse retained features or learn new ones.

4.2 The decline is not an artifact of the readout

A frozen probe is itself an optimization, so the decline could in principle reflect under-fitted readouts. Four observations argue against this. **(i) Budget.** Every epoch-29 probe behind a three-seed result (Tables 1 and 8) ran its full 30-epoch budget. **(ii) Precision.** Refitting the probe on the original checkpoint varies by only ± 0.02, two orders of magnitude below the 1.56-point decline. **(iii) The same probe can register retention.** Applied to continued encoders trained with a previous-model anchor, the identical procedure returns 79.79 (CaSSLe) and 79.99 (PFR), at or above the original 79.62 (§7). The procedure is therefore not biased downward for continued encoders in general.

Table 2: Sequential chain (seed 0). CKA to the original measures cumulative change; stepwise CKA compares with the preceding stage. Single-hop controls and shift measures are in Appendix G.

Stage	NIH probe	CKA to orig.	Step CKA	P@5
Original	79.62	1.000	—	0.199
RSNA, ep. 29	78.00	0.746	0.746	0.194
COVIDx, ep. 19	78.41	0.697	0.705	0.187
PadChest, ep. 19	67.85	0.258	0.280	0.152

(iv) Structure. The deficit is concentrated in the same pathologies across seeds and arms: Emphysema is among the three largest probe losses in all eleven arms of Appendix H. Optimization noise would not be expected to produce such a consistent pattern.

The readout still matters, and in one case it decided the conclusion. PFR's initial epoch-29 probe stopped early at probe epoch 15, having selected probe epoch 0, and scored 77.64; refitting the unchanged encoder without premature stopping gives 79.99, a 2.35-point swing from apparent loss to apparent retention (Appendix I). A retention number is a property of an encoder *and* a readout, so the readout's budget and selection rule belong in any retention report (Table 3). Equal budgets do not guarantee equal convergence, so our conclusions are specific to the stated probe procedure.

4.3 Geometry reorganizes early; neighborhoods remain comparatively stable

On the same trajectory, CKA to the original encoder falls from 1.000 to 0.792 by epoch 4, reaches 0.723 at epoch 14, and partially recovers to 0.742 at epoch 29. Retrieval P@5 remains within 0.191–0.199 (Appendix D). The update thus reorganizes global geometry, in step with the early probe loss, without collapsing label-based neighborhoods, consistent with source information becoming less linearly accessible rather than erased. Reduced CKA alone does not establish information loss.

5 Preservation, shift, and plasticity decouple across domains

In the single-seed sequential chain, the NIH probe moves from 79.62 to 78.00 after RSNA, 78.41 after COVIDx, and 67.85 after PadChest (Table 2). The RSNA endpoint comes from a separate seed-0 launch and differs from the audit endpoint by 0.06 points. The PadChest step removes 10.56 points and coincides with a collapse of CKA to the original (0.258) and lower retrieval precision. We do not interpret the small COVIDx increase as a replicated improvement.

Order is not required for the severe loss. Starting directly from the original checkpoint, COVIDx reaches 78.70 at epoch 19 (-0.92), whereas PadChest reaches 67.10 (-12.52). Preceding hops are therefore unnecessary for the PadChest loss. Both runs use 20 epochs, but PadChest is larger and so receives more updates. The control therefore does not separate shift magnitude from update count or preprocessing.

Shift does not order vulnerability. In the original encoder's representation space, RSNA is nearly indistinguishable from NIH ($\mathrm{MMD}^2 = 0.0071$; domain-classifier accuracy 0.599), whereas COVIDx is clearly separable (0.0504; 0.871). Yet at matched epoch 19, RSNA costs more frozen accessibility (-1.46) than COVIDx (-0.92). PadChest is the most distant on every measure and by far the most destructive. With three domains we claim no shift–forgetting law. The observations do show that hidden forgetting arises without acquisition shift, and that representation-space distance is not a sufficient screen for a safe update.

Preservation cost is not paid back as plasticity (exploratory). On their own target tasks, COVIDx continuation gains 2.96 test AUROC points, whereas PadChest is 0.85 below its baseline at epoch 9 and approximately returns to baseline by epoch 19. These are single-run deltas under task-specific configurations, so we treat them as exploratory. They suggest that the update that removes the most source accessibility need not buy the most target benefit. Source retention and target utility must therefore be reported separately.

6 SURPRISE LOCATES PLASTICITY BUT DOES NOT PRESERVE

6.1 DIFFICULTY IS CLASS-STRUCTURED BUT DOES NOT IDENTIFY VULNERABILITY

With the original model frozen, we compute per-image JEPA prediction loss over 277,551 NIH and CheXpert images (mean 0.369, coefficient of variation 0.279, versus 0.274 in a 1% subsample). Dataset identity shifts the scale (NIH 0.308, CheXpert 0.396), so we standardize within each dataset. Edema has the highest mean standardized loss and "No Finding" the lowest, in both corpora (NIH $+0.733/-0.179$; CheXpert $+0.283/-0.549$). Prediction loss thus carries reproducible pathology structure, but it is an imperfect proxy for vulnerability: the largest probe losses under continuation fall on the lower-loss classes Emphysema, Pneumothorax, and Nodule (Appendix H), whereas Edema, the highest-loss class, has median forgetting rank four across arms. These are associations rather than diagnostic or causal statements, and a complete class-level predictive analysis is left to future work.

6.2 SURPRISE-GATED DUAL MEMORY

SGDM tests whether a surprise signal, routed through two memory timescales, can supply what a single EMA target lacks. It replaces the single EMA target with fast and slow copies of the context-encoder parameters θ. The encoder, predictor, data, and base SSL losses are unchanged (Figure 2):

$$\theta_f \leftarrow m_f\theta_f + (1-m_f)\theta, \qquad \theta_s \leftarrow m_s\theta_s + (1-m_s)\theta, \tag{1}$$

with $m_f = 0.9$ and $m_s = 0.999$. A sample's predictor output p_i is compared with its slow-memory target h_i^s:

$$L_i^{\text{surp}} = \|p_i - \text{sg}(h_i^s)\|_2^2, \qquad g_i = \sigma\left(\text{clip}\left(\frac{L_i^{\text{surp}} - \mu}{s}, -6, 6\right)\right), \tag{2}$$

where sg is stop-gradient and μ, s are running location and scale statistics (momentum 0.999, carried across hops). No gradient passes through g_i. Measuring surprise against slow memory avoids a circular dependency with the mixed target

$$h_i^{\text{target}} = g_i h_i^f + (1-g_i)h_i^s. \tag{3}$$

Target arbitration (Eq. 3) changes *who teaches* each sample while both memories keep fixed rates. **Write-rate modulation** instead uses the batch-mean gate $\bar{g}$ to change *how fast* each memory is overwritten:

$$\theta_s \leftarrow \theta_s + (1-m_s)\max(1-\bar{g}, \epsilon)(\theta - \theta_s), \qquad \theta_f \leftarrow \theta_f + (1-m_f)\max(\bar{g}, \epsilon)(\theta - \theta_f), \tag{4}$$

with floor $\epsilon = 0.05$. Dual memory adds no trainable parameters but stores one additional target encoder and requires one additional target forward pass (0.700–0.760 s/step across the fixed, adaptive, and write-rate arms; Appendix E).

6.3 ADAPTIVE TARGET ARBITRATION FAILS A MEAN-MATCHED CONTROL

The critical control for an adaptive gate is a constant gate fixed at the adaptive gate's own mean. Any benefit of adaptivity must exceed this control, which has identical average target mixing. The adaptive gate reaches 77.78 ± 0.11 at epoch 29, versus 77.93 ± 0.05 for the mean-matched constant ($g = 0.4947$), and it is numerically lower at all three measured checkpoints (Figure 3; per-epoch values in Table 8). Uniform mixing (77.86 ± 0.30) and vanilla (77.90 ± 0.15) are indistinguishable from both. This is evidence against a retention benefit of this arbitration rule, not an equivalence test for all adaptive mixtures. A limited fast-momentum sweep ($m_f \in \{0.8, 0.9, 0.95\}$: 77.88, 77.82, 78.15; seed 0) stays within the same band (Appendix K).

Write-rate modulation helps modestly, but its source is unresolved. Write-rate modulation reaches 78.22 ± 0.13, a gain of 0.32 points over vanilla and 0.44 over target arbitration. Every write-rate seed exceeds every vanilla seed at epochs 19 and 29, by margins of at least 0.07 and 0.05 points (Table 9), and its seed-0 CKA is higher (0.814 vs. 0.742 for vanilla). However, Eq. 4 also slows both memories on average, so the gain cannot yet be attributed to surprise-dependent timing. Appendix J specifies the fixed-rate control, matched to the same average write coefficients, that separates the two explanations. The ambiguity does not affect the target-arbitration result, whose memory rates are fixed.

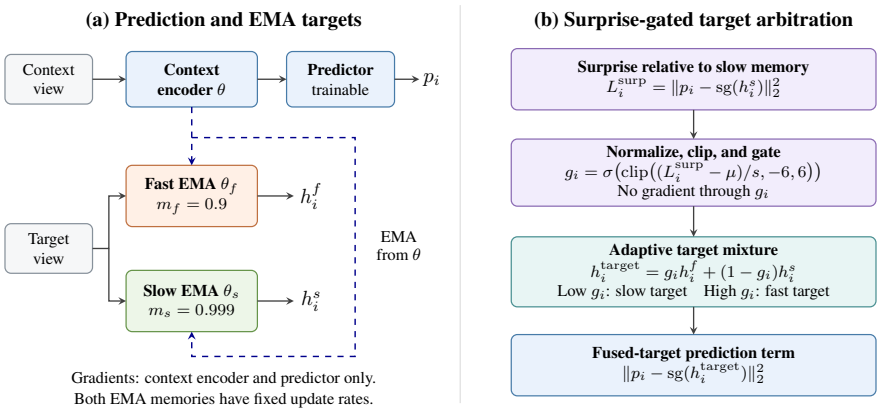

Figure 2: **SGDM target arbitration.** (a) The context encoder and predictor produce p_i; fast and slow EMA copies of θ produce targets h_i^f, h_i^s. Solid arrows: forward computation; dashed: EMA parameter updates. (b) Prediction error against the slow target sets a normalized, clipped gate that weights the target mixture. No gradient passes through either memory or the gate; other base SSL objectives are unchanged. The write-rate variant (Eq. 4) instead uses the batch-mean gate to modulate the two EMA rates.

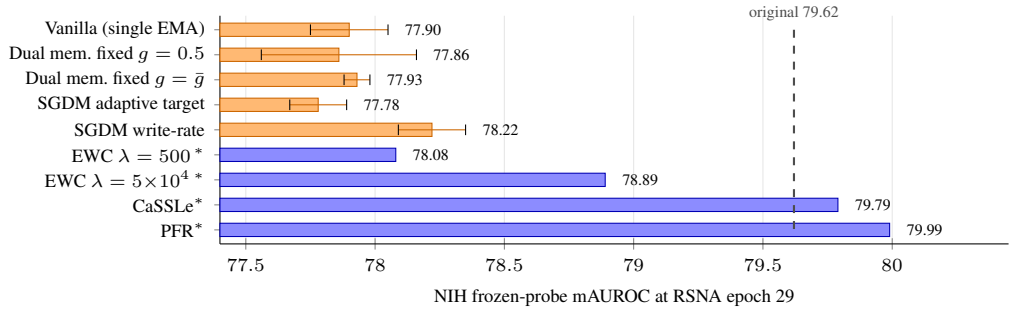

Figure 3: **Targets that track the learner retain similarly; references anchored to the pre-update model retain more.** Orange: arms whose target is an EMA of the current learner (three-seed mean $\pm$ s.d.). Blue: methods that penalize departure from a fixed copy of the pre-update model (*single seed; vanilla seed 0 reaches 78.06). EWC's effect grows with anchor strength. The dashed line marks the original checkpoint.

6.4 The gate is active and informative about where representations move

The adaptive gate is not degenerate. Its standard deviation is about 0.217, fewer than 0.2% of samples reach the clamp, and it maintains a persistent image-level ordering during training. Pneumonia-positive RSNA exposures receive a higher mean gate than negatives (0.577 vs. 0.471). Across the 26,684 unique RSNA images, the per-image mean gate correlates with embedding change between epochs 4 and 29 (Spearman $\rho = 0.508$), and mean drift increases from 0.214 to 0.280 from the lowest to the highest gate quartile. The association is contemporaneous and non-causal, and images from the same patient may be dependent. Surprise therefore identifies *where* representations move, yet using it to decide which memory teaches does not change *how much* source accessibility is kept. Plasticity allocation and preservation are distinct problems.

Direct probes of the slow memory itself (two seeds) show an advantage of 0.57–0.84 points over the context encoder at epoch 4, narrowing to about 0.03 by epoch 29. This is the signature of a lagged copy: a slow EMA of a drifting learner delays drift but ultimately inherits it.

Table 3: A reporting protocol for continually updated encoders, derived from the failure modes in this paper.

Question / use case	Report	Failure it prevents
Encoder reused frozen or by new heads	Refitted probe on original *and* updated checkpoints; identical, stated budget, stopping, and selection	Hidden forgetting (§4.1); readout artifacts (§4.2)
Existing deployed head	The unchanged head applied to the updated encoder	Silent head incompatibility
Encoder re-fine-tuned per task	Full fine-tuning, one recipe for all checkpoints, labelled recoverability	Mistaking recovery for preservation
Value of the update	Target-task gain, separate from source retention	Conflating plasticity with preservation (§5)
Adaptive update rule	A constant control matched to the rule's average	Crediting adaptivity for average effects (§6.3)

7 What preserves? Tracking versus anchoring references

The results of §6 suggest a simple reading. Every target we tested is an exponential average of the current learner: vanilla's single EMA, fixed or adaptive dual memories, a rate-modulated pair, and fast-momentum variants. All end between 77.78 and 78.22 at epoch 29 (Figure 3, orange). Changing the timescale, the mixture, or its adaptivity moves retention by at most 0.44 points. The reference methods differ in kind: they penalize departure from a *fixed* copy of the pre-update model. They retain markedly more in seed 0. CaSSLe and PFR reach 79.79 and 79.99, above the original 79.62 and 1.7–1.9 points above vanilla's same-seed 78.06. EWC shows a dose response, from 78.08 at $\lambda = 500$, inside the tracking band, to 78.89 at $\lambda = 5 \times 10^4$. CKA agrees in direction: PFR and CaSSLe stay at 0.930 and 0.913 to the original, versus 0.742 for vanilla (Appendix D).

These comparisons are single-seed for the anchored methods, so they establish direction and magnitude rather than a statistically reliable ranking. They are, however, large relative to the tracking arms' between-seed spread (s.d. ≤ 0.30). They support a design principle that the negative SGDM results only implied: *adaptation signals decide where to learn; preservation needs a reference that does not move with the learner.* A slower copy of a drifting learner is still a drifting learner. Two caveats bound the principle. First, anchoring results depend on a correct readout: PFR's retention was invisible under a prematurely stopped probe (§4.2). Second, we have not tested anchors under the largest shift. There, a whole-chain run with adaptive target gating reaches only 68.68 at PadChest epoch 19, 10.94 points below the original, and resetting its gate statistics before PadChest changes early calibration without improving the endpoint (68.52; Appendix F). Because this chain's recipe is not matched closely enough to the vanilla chain, we report it as an absolute outcome rather than a ranking. Whether a fixed anchor or replay (Rolnick et al., 2019) remains effective under a PadChest-scale shift is the most important open question our results raise.

8 Discussion and limitations

Evaluate the deployment contract. Full fine-tuning, a refitted frozen probe, and an unchanged head answer different operational questions, and continual-update evaluations should reflect how the encoder will ultimately be used. Table 3 summarizes the reporting protocol motivated by our experiments. Across these settings, we find that fine-tuning can mask changes visible under frozen evaluation, probe optimization can materially affect estimated retention, source retention need not track target-task gain, and adaptive update rules should be compared against controls matched to their average behavior. Together, these observations argue for separating recoverability, preservation, and target-domain utility rather than treating them as interchangeable measures of continual adaptation.

Scope. Our experiments study a controlled setting built around a ViT-Base JEPA-style radiograph encoder with NIH as the source benchmark. The central probe–fine-tuning dissociation and principal SGDM comparisons are evaluated across three training seeds, while the sequential-domain analysis, anchored reference methods, and momentum sweep are used primarily to characterize trends and effect magnitudes. The same fine-tuning recipe is applied uniformly across continual checkpoints to support within-protocol comparisons; evaluating additional adaptation recipes may further clarify

the relationship between preservation and recoverability. Across domains, differences in dataset size and preprocessing accompany distributional differences, so these factors are not fully disentangled. Finally, the target-task and write-rate analyses are exploratory, and the observed association between surprise and representation change should not be interpreted causally. Broader evaluation across self-supervised objectives, architectures, and imaging domains would establish how widely these phenomena generalize.

9 Conclusion

Continual self-supervised updates can leave source-task performance recoverable by fine-tuning while reducing what the frozen encoder makes accessible. In CheXWorld, probe and fine-tuning trends oppose each other within each of three training seeds and within individual pathologies. The effect appears without acquisition shift and becomes severe under PadChest. Prediction surprise is a valid signal for where representations move, but routing targets by it retains no more than its mean, and every learner-tracking target we tested retains similarly. References anchored to the pre-update model retain more. Measuring the frozen encoder directly, documenting the readout, separating source retention from target utility, and matching adaptive rules to their averages make these distinctions visible.

Ethics statement

This work uses publicly available, de-identified chest-radiograph datasets under their respective access conditions; no new human-subjects data were collected. It studies representation learning and evaluation, not clinical validation. Prediction loss and gate values are not diagnostic scores, and the evaluated methods are not validated for clinical deployment. Our central finding, that a standard evaluation can conceal degradation of a frozen encoder, is relevant to the safe maintenance of deployed medical models.

Reproducibility statement

Section 3 specifies the model, continuation sequence, and both evaluation contracts; Equations 1–4 define the dual-memory interventions. Appendices B–C give per-checkpoint and per-seed scores from which every paired change in §4.1 can be recomputed; Appendix I documents probe budgets and the PFR re-evaluation; Appendix J specifies the matched-rate control. We will provide an anonymized code repository with the submission and release the full repository upon acceptance.

AI use statement

Large language models assisted with manuscript editing, analysis-script drafting, literature checks, and consistency checks among reported results. The authors are responsible for the final manuscript and for verification of its results and citations.

References

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. PadChest: A large chest X-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797, 2020.

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, 2024.

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, pp. 16712–16721, 2022.

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*, pp. 9621–9630, 2022.

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*, pp. 3866–3876, 2022.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360, 2020.

Dapeng Hu, Shipeng Yan, Qizhengqiu Lu, Lanqing Hong, Hailin Hu, Yifan Zhang, Zhenguo Li, Xinchao Wang, and Jiashi Feng. How well does self-supervised pre-training perform with streaming data? In *International Conference on Learning Representations*, 2022.

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.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529, 2019.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017.

Huihan Liu, Changyeon Kim, Bo Liu, Minghuan Liu, and Yuke Zhu. Pretrained vision-language-action models are surprisingly resistant to forgetting in continual learning. *arXiv preprint arXiv:2603.03818*, 2026.

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.

James L. McClelland, Bruce L. McNaughton, and Randall C. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457, 1995.

Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Matthew P. Lungren, et al. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 7:119–130, 2025.

Quang Pham, Chenghao Liu, and Steven C. H. Hoi. DualNet: Continual learning, fast and slow. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Senthil Purushwalkam, Pedro Morgado, and Abhinav Gupta. The challenges of continuous self-supervised learning. In *European Conference on Computer Vision*, 2022.

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

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.

George Shih, Carol C. Wu, Safwan S. Halabi, et al. Augmenting the National Institutes of Health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, 2019.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2097–2106, 2017.

Yifan Wu, Hayden Gunraj, Chi-en A. Tai, and Alexander Wong. COVIDx CXR-4: An expanded multi-institutional open-source benchmark dataset for chest X-ray image-based computer-aided COVID-19 diagnostics. *arXiv preprint arXiv:2311.17677*, 2023.

Yang Yue, Yulin Wang, Chenxin Tao, Pan Liu, Shiji Song, and Gao Huang. CheXWorld: Exploring image world modeling for radiograph representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20778–20788, 2025.

A CheXWorld reproduction and fine-tuning recipes

We evaluated the released CheXWorld checkpoint on a single NVIDIA A10G (24 GB) with batch size 64, using the original downstream fine-tuning recipe (up to 50 epochs). Table 4 compares the results with the original paper. The continual-checkpoint audit instead uses one shorter, untuned recipe for every checkpoint (Section 3): 30 epochs, two warm-up epochs, patience 15, base learning rate 10^{-4}, uniform layer-wise decay, minimum learning rate zero, all layers trainable, and test score at the best-validation checkpoint. Its original-checkpoint NIH score (79.64) is therefore not interchangeable with the reproduction's 83.24. The audit recipe was chosen for identical treatment of every audited checkpoint rather than for peak accuracy, and all audit conclusions are within-recipe changes.

Table 4: Reproduction of CheXWorld downstream results under the original fine-tuning recipe.

Dataset	**Metric**	**Paper**	**Ours**	Δ	**Epochs**
ShenZhen	AUROC	98.88	98.81	-0.07	37/50
RSNA	Accuracy	75.03	74.94	-0.09	38/50
NIH	mAUROC	83.58	83.24	-0.34	25/50
CheXpert	mAUROC	89.63	89.22	-0.41	20/50

B Complete seed-0 audit

Both contracts start from the same continually pretrained checkpoint at each time point. Full fine-tuning changes the encoder during evaluation; the probe leaves it frozen.

Table 5: NIH retention across the RSNA continual-pretraining trajectory (seed 0). Full fine-tuning shows no sustained decrease, while the frozen probe declines, mostly by epoch 9.

RSNA epoch	**Full fine-tune**	**Frozen probe**	**Δ probe** (vs. ep. 0)
0 (orig.)	79.64	79.62	—
4	79.94	78.92	-0.70
9	79.78	78.18	-1.44
14	80.13	78.07	-1.55
19	79.63	78.16	-1.46
24	80.07	78.07	-1.55
29	80.04	78.06	-1.56

C PER-SEED RESULTS

Tables 6 and 7 add training seeds 1 and 2 to the seed-0 trajectory. Summaries are mean $\pm$ sample standard deviation over the displayed per-seed values; the paired changes in Section 4.1 are differences of these values.

Table 6: Frozen-probe NIH retention across three training seeds. The decline replicates in every seed.

Seed	**Ep. 4**	**Ep. 19**	**Ep. 29**	$\Delta_{4\to29}$
0	78.92	78.16	78.06	-0.86
1	78.89	77.97	77.87	-1.02
2	78.85	77.88	77.77	-1.08
Mean $\pm$ s.d.	78.89 ± 0.04	78.00 ± 0.14	77.90 ± 0.15	

Table 7: Full fine-tuning NIH performance across three training seeds. No seed declines from epoch 4 to 29; intermediate values fluctuate.

Seed	**Ep. 4**	**Ep. 19**	**Ep. 29**	$\Delta_{4\to29}$
0	79.94	79.63	80.04	$+0.10$
1	79.42	79.86	79.94	$+0.52$
2	80.37	80.60	80.67	$+0.30$
Mean $\pm$ s.d.	79.91 ± 0.48	80.03 ± 0.51	80.22 ± 0.40	

Table 8: NIH frozen-probe retention during RSNA continuation (mAUROC; original 79.62). Upper block: tracking-memory arms, mean $\pm$ s.d. over three training seeds. Lower block: methods anchored to the pre-update model, seed 0 (vanilla seed 0 at epoch 29: 78.06). PFR uses the full-budget probe (§4.2).

	Arm	**Epoch 4**	**Epoch 19**	**Epoch 29**
tracking	Vanilla (single EMA)	78.89 ± 0.04	78.00 ± 0.14	77.90 ± 0.15
	Dual memory, fixed $g = 0.5$	79.06 ± 0.12	77.97 ± 0.24	77.86 ± 0.30
	Dual memory, fixed $g = 0.4947$ (mean-matched)	79.06 ± 0.04	78.07 ± 0.11	77.93 ± 0.05
	SGDM, adaptive target gate	78.90 ± 0.18	78.02 ± 0.14	77.78 ± 0.11
	SGDM, write-rate gate	78.88 ± 0.12	78.37 ± 0.22	78.22 ± 0.13
anchored	EWC, $\lambda = 500$	79.16	78.08	78.08
	EWC, $\lambda = 5 \times 10^4$	79.25	78.83	78.89
	CaSSLe (projector)	79.21	—	79.79
	PFR (projector)	79.40	—	79.99

Table 9: Per-seed NIH frozen-probe scores for vanilla and write-rate gating. The minimum write-rate score exceeds the maximum vanilla score by 0.07 points at epoch 19 and 0.05 at epoch 29. These observed ranges do not establish statistical significance or isolate a surprise-specific benefit.

Epoch	Seed	Vanilla	Write-rate	Gain
19	0	78.16	78.62	0.46
19	1	77.97	78.23	0.26
19	2	77.88	78.26	0.38
29	0	78.06	78.36	0.30
29	1	77.87	78.18	0.31
29	2	77.77	78.11	0.34

D REPRESENTATION DIAGNOSTICS

CKA and retrieval use a fixed random subset of 5,000 NIH validation images. CKA compares each checkpoint with the original encoder. Retrieval measures label-based neighborhood quality at each checkpoint; it does not compare neighbor identities across checkpoints, and we report both raw precision@5 and its gain over random retrieval. The frozen-head divergence (KL) applies the original checkpoint's NIH probe head, unchanged, to the original and continued encoders and measures the divergence between their predicted label probabilities over the full NIH validation set ($n = 11{,}212$). It measures compatibility with a fixed readout, not task discrimination, and we do not base retention conclusions on it.

Table 10: Per-checkpoint representation diagnostics along the RSNA trajectory (seed 0), relative to the original encoder.

Epoch	CKA	P@5	Gain@5	KL	Highest-KL classes
0 (orig.)	1.000	0.199	0.080	0.000	—
4	0.792	0.199	0.080	0.782	Hernia, Cardiomegaly, Pl. thickening
9	0.743	0.196	0.077	1.463	Pneumothorax, Pl. thickening, Emphysema
14	0.723	0.191	0.072	1.336	Pneumothorax, Pneumonia, Hernia
19	0.732	0.192	0.073	0.973	Hernia, Pneumothorax, Pneumonia
24	0.741	0.195	0.076	0.684	Hernia, Cardiomegaly, Pneumothorax
29	0.742	0.193	0.074	0.647	Hernia, Emphysema, Pneumothorax

Jaccard overlap of the 100 highest-KL images is 0.12 between epochs 4 and 24 and 0.77 between epochs 24 and 29, so the images whose outputs move most change during the early reorganization and stabilize later. In the sequential chain, stepwise KL is 0.59 (RSNA), 0.94 (COVIDx), and 6.49 (PadChest), in line with the CKA collapse at the PadChest hop.

Table 11 gives CKA and KL for the anchored CaSSLe and PFR runs. Both keep CKA much closer to the original than vanilla does (0.913 and 0.930 vs. 0.742 at epoch 29), consistent with their higher probe retention, yet both show larger frozen-head KL. A projector-based feature constraint can therefore preserve linearly accessible information while shifting the outputs of a fixed head. This is another instance of the contracts disagreeing, and the reason Table 3 lists fixed-head compatibility separately.

Table 11: CKA and frozen-head KL to the original encoder for the single-seed CaSSLe and PFR runs.

	PFR		CaSSLe	
Epoch	**CKA**	**KL**	**CKA**	**KL**
4	0.896	2.838	0.866	1.408
9	0.898	2.676	0.869	1.555
14	0.904	2.797	0.879	1.652
19	0.917	2.629	0.897	1.321
24	0.928	1.929	0.910	1.122
29	0.930	1.854	0.913	1.043

E BASELINE SWEEPS, LOSSES, AND COST

Table 12: Final JEPA losses for the continuation arms without an additional continual-learning penalty. "Final step" is the last optimization step; the epoch-29 running average is the reported final loss. Different target constructions preclude treating the ordering as a measure of source preservation.

Arm	**Final step**	**Epoch-29 running average**
Vanilla	0.3393	0.3486
Fixed gate	0.3511	0.3518
Adaptive target gate	0.3461	0.3510
Write-rate gate	0.3506	0.3541

For the anchored methods, the optimization loss combines the JEPA term with the continual-learning penalty. The main-text PFR run ends at 0.3762 combined loss (penalty 0.0182; JEPA-only 0.3580), and CaSSLe at 0.3718 (penalty 0.0095; JEPA-only 0.3623). Their JEPA-only losses are within 0.014 of the tracking arms, so the anchors retain source accessibility without a large sacrifice of the self-supervised objective on RSNA. Table 13 reports the strength sweeps used to characterize each penalty; combined losses there are not comparable with the JEPA-only values in Table 12.

Table 13: Penalty-strength sweeps (ViT-Base, batch 64, 30 epochs). Combined loss = JEPA + continual-learning penalty.

Method	λ	**Train time**	**Combined loss**	**Final penalty**
EWC	500	1:59:28	0.3366	0.0075
EWC	5000	2:03:50	0.3503	0.0317
EWC	50000	1:56:48	0.3435	0.0409
PFR	0.1	2:24:40	0.2535	0.0068
PFR	1.0	1:20:55	0.3762	0.0182
PFR	10.0	2:26:40	0.3296	0.0380
CaSSLe	0.1	2:26:39	0.2631	0.0115
CaSSLe	1.0	3:12:08	0.3718	0.0095
CaSSLe	10.0	2:24:47	0.3914	0.0958

EWC shows a same-seed dose response in retained source performance (Table 14). CaSSLe at $\lambda = 10$ exhibits a large combined-loss spike, whereas PFR's bounded cosine penalty remains stable over the sweep. Seed-0 step times are 0.700 s for the fixed-gate dual memory, 0.735 s for adaptive target gating, and 0.760 s for write-rate gating; these are single-run measurements without variability estimates.

Table 14: Single-seed NIH frozen-probe retention for two EWC strengths.

EWC strength	Ep. 4	Ep. 19	Ep. 29
$\lambda = 500$	79.16	78.08	78.08
$\lambda = 50{,}000$	79.25	78.83	78.89

F CHAIN AND STRESS-TEST DETAILS

COVIDx ends with training loss 0.330 and is still slowly decreasing at epoch 19. PadChest plateaus earlier, with the global-layout loss component approximately twice the local component. The PadChest download contained 1,880 missing and 17 corrupt files, which were excluded before training. One gate log from an interrupted early launch was excluded from analysis.

The vanilla chain in Table 2 continues PadChest from the COVIDx checkpoint. Its RSNA endpoint (78.00) and the audit's (78.06) come from separate launches.

In the adaptive-target PadChest chain, epoch-level mean gate values are 0.240, 0.536, 0.462, 0.488, and 0.482 at epochs 0, 1, 2, 9, and 19. Resetting the running surprise statistics before PadChest lowers the epoch-0 gate mean to 0.164, as intended, after which the trajectories largely reconverge from epoch 2. The reset run reaches 68.52 at epoch 19 versus 68.68 with carry-over: the reset changes early gate behavior without improving endpoint retention. This single-run comparison is not an equivalence test.

G DOMAIN SHIFT AND SINGLE-HOP CONTROLS

Table 15 reports NIH retention when each incoming domain is used as a single hop from the original checkpoint. Table 16 reports three distances between NIH and each domain in the frozen original encoder's embedding space. With three domains, these values characterize the observed regimes rather than estimating a general shift–forgetting relation.

Table 15: Single-hop NIH frozen-probe retention from the original checkpoint (79.62), seed 0. RSNA values are from Table 5. The sequential chain reaches 67.85 at PadChest epoch 19.

Domain	Ep. 9	Δ	Ep. 19	Δ
RSNA	78.18	-1.44	78.16	-1.46
COVIDx	78.32	-1.30	78.70	-0.92
PadChest	66.22	-13.40	67.10	-12.52

Table 16: Representation-space shift between NIH and each incoming domain, with single-hop forgetting at epoch 19. PadChest separates from the other domains on every measure. RSNA is the closest domain yet loses more than COVIDx.

Domain	MMD^2	Fréchet	Domain-clf. acc.	Δ probe (ep. 19)
RSNA	0.0071	0.025	0.599	-1.46
COVIDx	0.0504	0.185	0.871	-0.92
PadChest	0.9175	1.514	1.000	-12.52

H PER-PATHOLOGY FORGETTING ACROSS ARMS

Table 17 reports per-pathology probe forgetting for eleven arms, averaging seeds where available. Positive values denote worse discrimination than the original checkpoint. The arms share evaluation data and include related hyperparameter variants, so counts are descriptive rather than independent replications.

Emphysema is among the three largest probe losses in all eleven arms and ranks first in nine; Pneumothorax is among the three largest in ten; Nodule in nine. The anchored methods (CaSSLe, PFR) reduce these losses by roughly an order of magnitude. Hernia shows elevated fixed-head KL but improved refitted-probe AUROC in every arm, so fixed-head output drift and refitted discrimination need not agree. Probability shifts, readout mismatch, and the class's small sample size remain possible explanations. The cross-class rarity–KL association ($\rho = -0.42$, $p = 0.14$ across 14 classes; $\rho = -0.27$ without Hernia) does not establish a causal role for rarity.

Table 17: Per-pathology AUROC forgetting at epoch 29 relative to the original checkpoint (positive = forgotten; rank among 14 pathologies in parentheses, 1 = most forgotten). n: training seeds averaged. PFR uses its full-budget probe.

Arm	n	**Hernia**	**Emphysema**	**Pneumothorax**	**Nodule**
Vanilla	3	-0.67 (14)	$+3.47$ (3)	$+3.83$ (2)	$+4.18$ (1)
EWC, $\lambda = 500$	1	-1.18 (14)	$+3.17$ (2)	$+2.17$ (3)	$+3.88$ (1)
EWC, $\lambda = 5 \times 10^4$	1	-2.34 (14)	$+2.90$ (1)	$+1.97$ (3)	$+2.44$ (2)
CaSSLe	1	-1.94 (14)	$+0.45$ (1)	$+0.25$ (3)	$+0.25$ (4)
PFR	1	-1.03 (14)	$+0.45$ (1)	$+0.09$ (2)	-0.80 (11)
Fixed $g = 0.5$	3	-0.92 (14)	$+4.21$ (1)	$+2.96$ (3)	$+3.73$ (2)
Fixed $g = \bar{g}$	3	-1.33 (14)	$+4.79$ (1)	$+3.35$ (2)	$+3.13$ (3)
Adaptive target gate	3	-0.65 (14)	$+4.74$ (1)	$+3.19$ (3)	$+3.28$ (2)
Adaptive target gate, $m_f = 0.8$	1	-1.44 (14)	$+3.88$ (1)	$+2.01$ (5)	$+2.84$ (2)
Adaptive target gate, $m_f = 0.95$	1	-2.37 (14)	$+3.62$ (1)	$+2.98$ (2)	$+2.21$ (3)
Write-rate gate	3	-1.90 (14)	$+4.16$ (1)	$+2.55$ (3)	$+2.65$ (2)

Table 18: Per-pathology forgetting for the vanilla runs at epoch 29 (three-seed mean) under the frozen probe and under full fine-tuning (positive = forgotten). The pathologies most forgotten under the probe show much smaller losses, or gains, once the encoder is fine-tuned.

Pathology	**Frozen probe**	**Full fine-tune**
Nodule	$+4.18$	$+0.23$
Pneumothorax	$+3.83$	-0.44
Emphysema	$+3.47$	-1.67
Consolidation	$+1.03$	$+0.73$
Hernia	-0.67	-3.45

I PROBE EARLY STOPPING AND THE PFR RE-EVALUATION

The probe uses a two-epoch learning-rate warm-up and early-stopping patience of 15 epochs. For encoders trained with a feature-distillation anchor, validation performance can be high after the first probe epoch and then dip during warm-up; if the dip outlasts the patience window, optimization stops before recovering. Probe epoch 0 here denotes the first trained epoch, not an untrained head.

PFR's epoch-19 and original epoch-29 probes, and CaSSLe's epoch-19 probe, stopped at probe epoch 15 after selecting probe epoch 0. Both epoch-4 probes ran to completion and selected probe epoch 29. CaSSLe's epoch-29 probe exceeded its initial best at epoch 15, reset patience, and completed 30 epochs, reaching approximately 83.0 validation and 79.79 test mAUROC. Re-running PFR's epoch-29 probe without premature stopping yields 83.09 validation and 79.99 test mAUROC, versus 77.64 under the prematurely stopped evaluation of the same encoder.

The main text uses full-budget endpoints throughout. Every epoch-29 probe behind the three-seed rows of Table 8 ran all 30 epochs, so this failure mode does not affect the vanilla or SGDM endpoints. The epoch-19 PFR and CaSSLe values remain early-stopped; we list them in Table 19 for completeness but do not use them for comparison.

Table 19: Test mAUROC for the single-seed anchored baselines. †: early-stopped probe that selected probe epoch 0 (not used for comparison). The prematurely stopped PFR endpoint is shown in parentheses.

Method	Epoch 4	Epoch 19	Epoch 29
CaSSLe	79.21	$77.31^{\dagger}$	79.79
PFR	79.40	$77.35^{\dagger}$	79.99 (early-stopped: $77.64^{\dagger}$)

J GATE SANITY CHECKS AND THE MATCHED-RATE CONTROL

The uniform fixed-gate arm uses $g \equiv 0.5$; the mean-matched arm uses $g \equiv 0.4947$, the mean recorded gate of the adaptive arm. Neither varies with samples. Across adaptive runs the gate remains non-degenerate ($g_{\mathrm{std}} \approx 0.217$), fewer than 0.2% of samples hit the clamp, and per-image mean gate closely tracks the underlying surprise. In the full RSNA population, the highest-mean-gate images remain near 0.9–1.0 for most of training while the lowest remain near 0.05–0.15, a persistent image-level ordering rather than a noisy gate. These checks establish implementation health, not improved retention.

The decisive control for the write-rate intervention is a fixed two-memory system with the same average effective write coefficients but no batch-dependent adaptivity. Under Eq. 4, the coefficients at update t are

$$\alpha_s(t) = (1 - m_s)\max(1 - \bar{g}_t, \epsilon), \qquad \alpha_f(t) = (1 - m_f)\max(\bar{g}_t, \epsilon),$$

so the matched control uses $\alpha_s^\star = T^{-1}\sum_t \alpha_s(t)$ and $\alpha_f^\star = T^{-1}\sum_t \alpha_f(t)$, i.e., $m_s^\star = 1 - \alpha_s^\star$ and $m_f^\star = 1 - \alpha_f^\star$. With $\bar{g} \approx 0.495$, these are approximately $m_s^\star \approx 0.9995$ and $m_f^\star \approx 0.95$. The control should keep the same target construction, initialization, update order, and training budget. An arithmetic mean does not reproduce the full time-varying EMA kernel, so a null result would indicate that adaptivity adds nothing beyond average slowdown at first order.

K ADDITIONAL INTERVENTION OBSERVATIONS

Varying the fast-memory momentum over $\{0.8, 0.9, 0.95\}$ gives final NIH probe scores of 77.88, 77.82, and 78.15 in seed 0, all within the tracking-memory band of Figure 3. Without matched fixed-gate controls at each momentum, this sweep does not establish a negative result for target arbitration at every momentum. The adaptive-target PadChest chain reaches 66.82 at epoch 9 and 68.68 at epoch 19. The slow-memory probe advantage over the context encoder is 0.57–0.84 points at epoch 4 and about 0.03 at epoch 29 (two seeds); we infer from this only that the slow memory lags the learner, not a permanent-preservation or information-transfer mechanism.