BEFORE THE COLLAPSE: MEASURING AND RESTORING PLASTICITY IN CONTINUAL VISION-LANGUAGE-ACTION REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Sequential RL post-training of large vision-language-action (VLA) policies has recently been reported to be surprisingly stable: with LoRA adapters and on-policy GRPO, near-zero backward transfer is achieved across standard benchmarks, and stability is attributed to scale, low-rank constraints, and pretraining (Hu et al., 2026; Liu et al., 2026). We argue that this focus on forgetting inverts the field's priorities: low forgetting can coexist with, and even mask, insufficient plasticity (the capacity to keep learning), yet current evaluations measure only behavioral success, never network-level trainability (Lyle et al., 2025). We present a closed loop of *measure, intervene, predict* for continual VLA RL. On a five-task stream of increasing difficulty, a standard forgetting-oriented anchoring objective suffers a catastrophic *policy collapse* at the fourth stage: success on all ten tasks falls to zero simultaneously, preceded by a structural fingerprint (effective-rank contraction, unbounded norm growth) that is absent in the unanchored baseline, whose explored subspace instead expands. The fingerprint prescribes the remedy: DARE-style sparsification of the collapsed adapter restores acquisition at the collapse stage ($0.000 \rightarrow 0.113$) with retention within 13pp of the baseline, and directional rank-channel release is therapeutic ($4.4\times$ random release) while improving the healthy frontier; both replicate across seeds. A lightweight forgetting-risk predictor built on representation similarity and NTK-overlap theory (Doan et al., 2021; Goldfarb et al., 2024) rank-orders streams by their actual forgetting behavior. These deliverables turn the reliability of continual VLA RL from an emergent accident into an engineered property.

1 INTRODUCTION

Vision-language-action (VLA) models (Kim et al., 2025) map visual observations and language instructions to robot actions, and are increasingly post-trained with reinforcement learning to improve manipulation success (Shao et al., 2024). In realistic deployment a robot must acquire *new* skills without losing *old* ones (the continual learning, CL, setting), and a recent line of work reports a striking result: VLA models are *natural continual learners*. With sequential fine-tuning (one task at a time, LoRA adapters, on-policy GRPO), the *Simple Recipe* of Hu et al. (2026) achieves near-zero backward transfer on LIBERO, RoboCasa, and ManiSkill, outperforming dedicated CL methods such as EWC and experience replay; follow-up work argues that pretrained VLAs are "surprisingly resistant to forgetting" (Liu et al., 2026). On the surface, continual VLA RL appears solved.

We do not dispute these results; we dispute what they have been taken to imply. This paper inverts the field's working assumption: *forgetting is not the problem*. First, stability is an *emergent* property, not a designed one: the canonical GRPO objective has a zero KL coefficient, so nothing bounds per-task drift; stability is an empirical coincidence of scale, low rank, and pretraining, not a guarantee. Second, and more importantly, *low forgetting can mask low plasticity*. Under our destructive hard-to-easy discriminator (task $9 \rightarrow 0$ on LIBERO-Spatial: a near-zero-zero-shot hard task learned first, an easy task with strong learning signal second), the sequential baseline *improves* retention of the hard task ($0.150 \rightarrow 0.238$), consistent with theory predicting minimal forgetting at near-maximal

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.

task similarity (Goldfarb et al., 2024; Lee et al., 2021) and with Liu et al. (2026)'s warning that "low forgetting" can be a degenerate outcome of insufficient plasticity.

The real bottleneck is not stability but *acquisition capacity*: the ability to keep learning new tasks at a useful rate. On the same discriminator, the baseline acquires the hard task to only ≈ 0.15 success from a ≈ 0.075 zero-shot prior, and a forgetting-oriented regularizer (advantage-gated anchoring; Section 3) buys *zero* retention gain while *costing* plasticity (0.625 vs. baseline 0.713). Plasticity loss is pervasive precisely in on-policy policy-gradient methods under domain shift (Juliani & Ash, 2024; Dohare et al., 2024), yet no prior work measures network-level plasticity for LoRA-adapted VLA policies under GRPO. This gap is our entry point.

We contribute a closed loop of *measurement, intervention, and prediction* for continual VLA RL:

- **A dual-track plasticity measurement protocol.** A structural track (dormant ratio (Sokar et al., 2023), effective rank (Dohare et al., 2024), update-norm drift) paired with a behavioral track (injection probe, warm-start gap (Ash & Adams, 2020), acquisition curves), deliberately paired because no single structural metric reliably predicts trainability (Lyle et al., 2025).
- **Risk-Gated LoRA Rejuvenation.** An L2-Init-style soft reset of the LoRA increment toward its zero initialization (Kumar et al., 2025), gated per-state by the group-relative advantage: relaxed where the current task provides decisive learning signal, enforced elsewhere.
- **A forgetting-risk predictor.** $R_t = f(\mathrm{cap}_t, \mathrm{sim}_{t-1,t}, \|\Delta\theta_t\|)$, built on NTK-overlap theory (Doan et al., 2021) and similarity–forgetting analyses (Goldfarb et al., 2024; Lee et al., 2021), validated across task streams as a $(benchmark \times learner)$-paired quantity (Liu et al., 2026).
- **Empirical evidence that acquisition stability binds, and can be restored.** On the five-task stream, the forgetting-oriented anchor collapses outright at the fourth stage; its structural fingerprint precedes the behavioral event and is absent in the unanchored baseline, whose explored subspace instead expands. The fingerprint prescribes the remedy: sparsifying the collapsed adapter restores acquisition at the collapse stage ($0.000 \rightarrow 0.113$) with retention within 13pp of the baseline; directional rank-channel release is therapeutic (0.275, $4.4\times$ random) and improves the healthy frontier; all replicate across seeds.

2 Related Work

Continual learning for VLA policies. Hu et al. (2026) show sequential fine-tuning with LoRA and GRPO achieves near-zero forgetting on LIBERO/RoboCasa/ManiSkill; Liu et al. (2026) attribute this to pretraining and warn that low forgetting may mask insufficient plasticity. Luo et al. (2026) show RL *does* forget on heterogeneous multimodal streams; Zeng et al. (2026) anchor stability to advantage magnitude with a frozen critic. None of these lines measures network-level plasticity; all evaluate only behavioral success.

Loss of plasticity in deep RL. Dohare et al. (2024) establish loss of plasticity as a distinct failure mode eased by L2 regularization; Sokar et al. (2023) identify dormant neurons; Nikishin et al. (2022) propose periodic resets. Juliani & Ash (2024) show plasticity loss is pervasive in *on-policy* deep RL under domain shift (our setting), where "regenerative" methods (soft S&P, LayerNorm, L2-Init) consistently help (Klein et al., 2024; Yuan et al., 2025). This literature is almost entirely in tabular/Atari/MuJoCo regimes; its transfer to parameter-efficient fine-tuning of large pretrained policies is untested.

Structural metrics, interventions, and similarity theory. No single structural property reliably predicts trainability (Lyle et al., 2025; Juliani & Ash, 2024); hence our structural-plus-behavioral pairing. The most consistent intervention across the on-policy literature is regression toward an initialization (Kumar et al., 2025; Juliani & Ash, 2024); LoRA adapters are a *natural* home for it, since the adapter increment is initialized at zero. LoRA's learning–forgetting trade-offs at LLM scale (Biderman et al., 2024) further suggest adapter-space interventions suit VLA CL. Finally, Doan et al. (2021) show forgetting grows as tasks align, while Goldfarb et al. (2024); Lee et al. (2021) derive a non-monotonic similarity–forgetting curve, explaining why our LIBERO-Spatial discriminator can-

not produce forgetting, and why a *predictor* of forgetting risk, not another regularizer, is the right deliverable.

3 METHOD: MEASUREMENT, INTERVENTION, PREDICTION

Setup and notation. We follow the continual VLA fine-tuning protocol of Hu et al. (2026). A VLA policy π_θ consists of a frozen pretrained backbone and a trainable LoRA adapter θ of rank r, mapping an observation x (image + language) to a chunk of action tokens a. Tasks arrive sequentially; at task t we initialize θ from the previous task's converged adapter (from the SFT checkpoint π_0 at $t = 0$) and run on-policy GRPO with a sparse success reward. We write $\ell_\theta(x, a) = \log \pi_\theta(a \mid x)$ and $\hat{A}(x, a)$ for the group-relative advantage; all plasticity quantities are computed on a fixed reference set of states (a held-out batch of trajectories from the current task) for cross-task comparability.

3.1 DUAL-TRACK PLASTICITY MEASUREMENT PROTOCOL

Structural track. Three diagnostics, all computable on CPU from the LoRA adapter alone: **(i) dormant ratio** (Sokar et al., 2023), the fraction of adapter activations below a threshold relative to a running maximum; **(ii) effective rank**, the exponential of the Shannon entropy of the normalized singular-value distribution of each adapter matrix,

$$\mathrm{erank}(\Phi) = \exp\big\{H(p)\big\}, \qquad p_i = \frac{\sigma_i}{\sum_j \sigma_j}, \tag{1}$$

whose shrinkage indicates representation collapse: the adapter concentrating its learned function onto fewer directions (Dohare et al., 2024); and **(iii) update-norm drift**, $\|\theta_t - \theta_{t-1}\|_F / \|\theta_{t-1}\|_F$ across task boundaries, whose collapse while loss stays high is the direct readout of lost plasticity. Full definitions in Section A.

Behavioral track. Because structural metrics alone cannot be trusted (Lyle et al., 2025), every structural claim is anchored by three probes: **(i) plasticity injection**: reset 10% of adapter parameters to zero-init and re-run $1/4$ of the GRPO budget on the same task; if the network cannot re-learn what it just knew, plasticity, not the task, is the bottleneck (Juliani & Ash, 2024; Nikishin et al., 2022); **(ii) warm-start gap** (Ash & Adams, 2020): acquisition from the previous adapter versus from the SFT checkpoint under an identical budget; and **(iii) per-task acquisition curves**: $S_t^{(t)}$, whose mean $\mathrm{PL} = \frac{1}{T}\sum_t S_t^{(t)}$ is our headline acquisition metric; its decay with t while early-task success stays high shows later tasks are harder to learn *even with the same budget*.

3.2 LORA REJUVENATION VIA L2-INIT TOWARD ZERO-INIT

The most consistent plasticity intervention in on-policy RL is regression toward an initialization (Kumar et al., 2025; Juliani & Ash, 2024). LoRA makes this intervention *parameter-free in spirit*: the adapter increment ΔW is initialized to zero, so an L2 penalty on ΔW is an implicit *soft reset* of the low-rank update: exactly the deterministic analogue of soft S&P that the on-policy literature identifies as the reliable family (Juliani & Ash, 2024). Concretely, we augment the GRPO objective with

$$\mathcal{L}^{\mathrm{rejuv}}(\theta) = \mathcal{L}^{\mathrm{GRPO}}(\theta) + \frac{\lambda}{2}\,\mathbb{E}\Big[g(x)\,\|\Delta W(x)\|_F^2\Big], \tag{2}$$

where $\|\Delta W\|_F^2$ is the squared Frobenius norm of the LoRA increment, λ is the rejuvenation strength, and $g(x)$ is a state-dependent gate on the group-relative advantage:

$$g(x) = \sigma\Big(\frac{c - |\hat{A}(x)|}{\tau}\Big), \tag{3}$$

with $c = 0.5$, $\tau = 0.25$. When $|\hat{A}| \gg c$ the current task provides decisive signal: $g \to 0$ and the adapter is free to move (plasticity first). When $|\hat{A}| \ll c$ the state carries little task-relevant signal: $g \to 1$ and the adapter is pulled back toward zero-init (stability first). We call this composition *Risk-Gated Rejuvenation*: a soft reset that *activates where the risk of drift exceeds the value of*

change and relaxes exactly where new-task learning matters. Unlike ReDo (Sokar et al., 2023), which fails on-policy (Juliani & Ash, 2024), it never hard-resets parameters; unlike a uniform L2-Init penalty, the gate makes the trade-off per-state rather than global. The intervention adds exactly one hyperparameter (λ).

The sparsification member of the family. Equation 2 is the *continuous* member of a family of regression-toward-zero-init interventions; the discrete member we evaluate is *DARE-style sparsification*, acting directly on the converged adapter: draw a Bernoulli mask $m \sim \mathrm{Bern}(1-p)$ per entry and rescale survivors mean-preservingly ($\tilde{w} = m\,w/(1-p)$). Dropping entries conserves the first moment of the update while carrying away the small-magnitude interference terms stacked in shared directions, while large directional structure survives. Sparsification re-opens rank budget as a single post-hoc checkpoint operation (no retraining, no extra loss) and is the natural complement to the L2-Init objective, which budgets capacity continuously.

3.3 Forgetting-risk predictor

Our negative results show that *whether training task t damages task $t-1$ is a property of the specific (benchmark, learner) pair*, not of the method alone (Liu et al., 2026). We therefore shift from designing regularizers to *predicting* the risk that a given next task will be destructive, formalized as

$$R_t \;=\; f\big(\mathrm{cap}_t,\ \mathrm{sim}_{t-1,t},\ \|\Delta\theta_t\|\big), \tag{4}$$

with three theory-grounded drivers: **(i) remaining capacity** cap_t (residual erank/r and $1-$dormant ratio); **(ii) representation similarity** $\mathrm{sim}_{t-1,t}$ (CKA between consecutive tasks; NTK-overlap theory predicts forgetting grows as tasks align (Doan et al., 2021), but the relation is non-monotonic (Goldfarb et al., 2024; Lee et al., 2021), so raw similarity feeds a nonlinear f); and **(iii) update magnitude** $\|\Delta\theta_t\|$, bounding how far the policy can drift.

The predictor is validated as a $(benchmark \times learner)$-paired model (Section 4.4): it must rank-order streams that genuinely produce forgetting above streams that structurally cannot, with RL and BC fine-tuning validated separately. We do not claim R_t is a universal law; we claim it is the first *operational* forgetting-risk signal for continual VLA RL, with the destructive-stream construction made explicit as a protocol.

4 Experiments

Setup. We use OpenVLA-OFT 7B (Kim et al., 2025) (SFT checkpoint) as the base policy, LoRA rank 32, GRPO with group size 2, $\sim$2048 samples per update, on the **LIBERO-Spatial** suite (10 tasks) (Liu et al., 2023), 4$\times$RTX 4090 (24GB). We train sequentially over tasks $0 \to 9$ (2 epochs per task), evaluating on all learned tasks after each stage. All chained-arm and intervention results are replicated under a second seed; Table 2 summarizes the stage-wise acquisition of both arms under both seeds.

4.1 Acquisition decay

Per-task acquisition over the increasing-difficulty stream $5 \to 9$ (Table 1). The go/no-go criterion is met in its strongest form: acquisition does not merely decay along the stream: on the anchored arm it collapses outright at t_4=8 ($0.888 \to 0.700 \to 0.000$), while backward transfer remains intact until the event. Low forgetting does not just mask an acquisition bottleneck; the two can dissociate catastrophically.

The zero-shot priors in Table 1 give the stream its difficulty gradient (t_1 easy at 0.688, t_5 hardest at 0.150). Two controls sharpen the diagnosis. The *Indep-FT control* trains each task from the same base state in isolation: its shrinking acquisition gains (the hardest task ends *below* its own zero-shot prior) cannot be attributed to interference, isolating a per-task learning deficit of GRPO itself that grows with difficulty. The *completed baseline chain* supplies the decisive discriminator: through the identical stream the unanchored baseline *does not collapse* (acquisition 0.463 vs. 0.000, PL 0.560 vs. 0.465), so the collapse is specific to accumulated behavioral anchoring, not GRPO's generic negative-advantage spiral. The pattern replicates under seed 1: the anchored arm falls to

Table 1: Phase-1 acquisition on the increasing-difficulty stream $5 \rightarrow 9$ (LIBERO-Spatial, seed 42, group size 2). Zero-shot = success before any training; Acq. = per-task success after its own training stage; Ret. = success on task t at the final evaluation pass. The Indep-FT control trains each task from the base model in isolation. †Identical configuration to Base (the anchored objective is inert without a previous checkpoint); gaps measure run-to-run variance. ‡Lost to an infrastructure fault; the chained arm supplies this cell. ‖The chained A-GRPO arm's t_4=8 stage suffered a *policy collapse*: after training on task 8, success on *all ten* tasks fell to zero (a distributional collapse, not forgetting), with only partial recovery after task 9. §Interrupted mid-run and not recoverable; the chained arm supersedes them.

		Indep-FT control		Chained arms	
Task (stream order)	Zero-shot	Base Acq.	Anchored Acq.†	SeqFT Acq./Ret.	Anchored Acq./Ret.
$t_1 = 5$	0.688	–‡	0.638	0.688 / 0.637	n/a‖
$t_2 = 6$	0.875	0.925	0.813	0.875 / 0.825	0.888 / 0.812
$t_3 = 7$	0.638	0.625	–§	0.650 / 0.662	0.700 / 0.588
$t_4 = 8$	0.363	0.388	–§	0.463 / 0.463	0.000‖ / 0.050
$t_5 = 9$	0.150	0.088	–§	0.125 / 0.125	0.100 / 0.100
PL (mean Acq.)	–	(0.506; excl. t_1)	–	**0.560**	**0.465**
BWT (mean)	n/a in the Indep-FT control (no sequential dependency); chained: SeqFT $+0.004$, Anchored -0.250 (dominated by the collapse)				

Table 2: Stage-wise acquisition on the increasing-difficulty stream $5 \rightarrow 9$ across two seeds (LIBERO-Spatial; group size 2). Instability is specific to the anchored arm under both seeds, with collapse onset at t_4=8 (seed 42) shifting to t_5=9 (seed 1); the unanchored arm decays smoothly in both runs. The collapsed states are excluded from the retention averages of Fig. 1.

	Acquisition by stage				
Arm	t_1=5	t_2=6	t_3=7	t_4=8	t_5=9
Anchored (seed 42)	0.638	0.888	0.700	**0.000**	0.100
Anchored (seed 1)	0.738	0.875	0.588	0.188	**0.000**
SeqFT (seed 42)	0.688	0.875	0.650	0.463	0.125
SeqFT (seed 1)	0.750	0.788	0.638	0.375	0.138

$0.188/0.000$ at the corresponding stages (onset shifting by one stage) while the baseline decays smoothly ($0.375/0.138$). Meanwhile the baseline's own acquisition decays gently while retention stays flat: the binding constraint is stability of *acquisition*, not retention. The unanchored learner turns the difficulty gradient into a gradual bottleneck; the anchored learner, into a catastrophic failure.

4.2 Rejuvenation interventions

The collapse fingerprint of Section 4.4 predates the behavioral event and therefore prescribes *when* to intervene. As the gentlest member of the L2-Init family we test DARE-style sparsification of the collapsed adapter: after the collapse-inducing stage t_3=7, each adapter entry is drawn with keep-probability 0.75 and survivors rescaled mean-preservingly (27.7M of 110.8M entries dropped), then the chain continues onto task 8 from the sparsified checkpoint (Table 3).

The intervention restores a collapsing stream to nonzero acquisition at the collapse stage itself (0.1125 vs. the intervention-free 0.000) while preserving retention at 0.646 (within 13pp of the never-collapsed baseline) without any retraining from base. The full ten-task matrix confirms the restoration is broad, not concentrated: every task that retained signal after DARE lands on the healthy plateau of the pre-collapse chain. On the retention–acquisition frontier (Fig. 1) this is a strict first-order Pareto improvement: the collapsed state $(0, 0)$ is dominated by every healthy point, and DARE moves the anchored arm back into the healthy frontier region.

The restoration is, however, *stage-local*: continuing onto t_5=9, acquisition again collapses to 0.000 and retention falls to 0.471 (below the unperturbed arm's own 0.503). One-shot sparsification reopens the rank budget at the moment of crisis but does not alter the optimizer dynamics that produced the crowding–instability cycle; under continued chaining the explosive feedback re-accumulates. Recovery of *parameters* is achievable one-shot; restoration of *plasticity*, a property of the optimization loop, requires the sustained, per-step fingerprint-gated L2-Init of Section 3.2, with λ-swept

Table 3: Rejuvenation on the anchored chain (LIBERO-Spatial, seed 42). Anchored = chained anchoring arm of Table 1 (collapse at t_4=8); +DARE = its t_3=7 checkpoint sparsified with p=0.25 (27.7M/110.8M entries dropped), then chained onward; SeqFT = unanchored baseline. Ret. = mean retention over previously seen stages t_1–t_3 $(5, 6, 7)$. Both arms completed both stages (identical pipeline).

	Chained arms at the collapse stage		
Stage	Anchored (no inter.)	**Anchored + DARE**	SeqFT baseline
t_4=8 Acq.	0.000	**0.1125**	0.463
t_4=8 Ret.	0.000	**0.646**	0.775
t_5=9 Acq.	0.100	0.000	0.125
t_5=9 Ret.	0.503	0.471	0.647

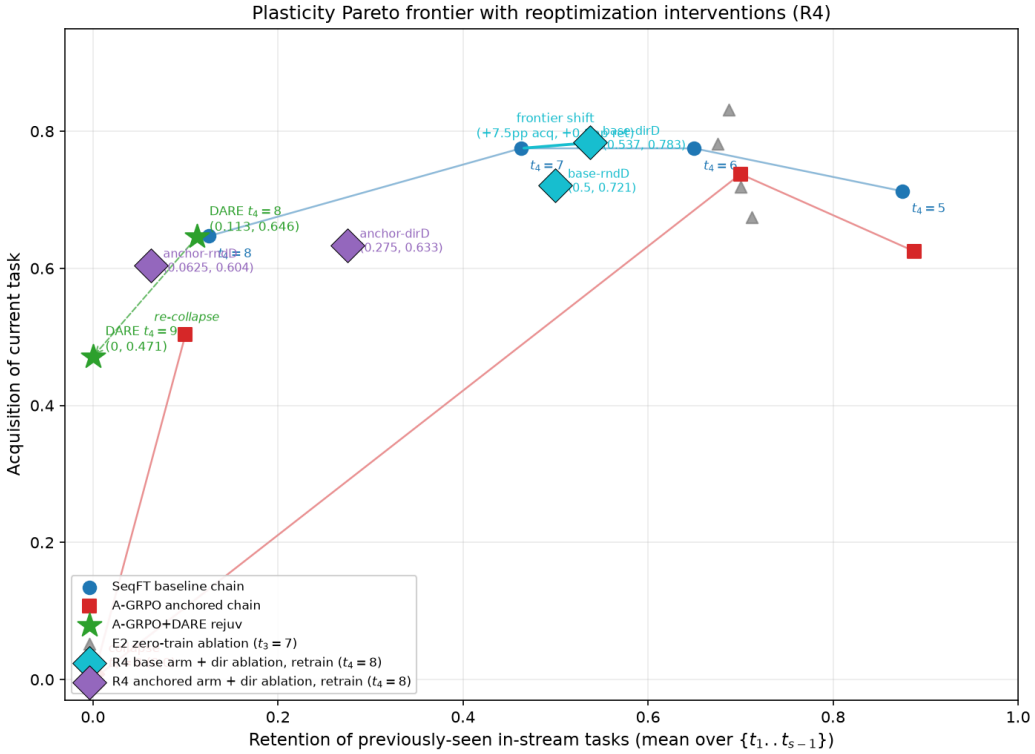


Figure 1: Plasticity Pareto frontier of the chained arms (one point per completed stage; x: retention over previously seen stages, y: acquisition). The anchored arm's collapse $(0, 0)$ is dominated by every healthy point; DARE ($\star$) restores a near-frontier position $(0.113, 0.646)$, but the dotted trajectory shows the restoration is stage-local; SeqFT traces the healthy frontier. Zero-training ablation probes ($\triangle$) cluster near their intact arms; reoptimization ($\diamond$) completes the causal picture: directional release recovers the collapse ($4.4\times$ random across seeds) and improves the healthy baseline's frontier point.

resets and gated variants as future work. The revival benefit itself replicates under seed 1, with its dimension set by the collapse profile: under full collapse DARE restores acquisition, under partial collapse it restores memory (post-pruning retention 0.646 under both seeds); Section D.

4.3 Rank-channel ablation probes the mechanism

To discriminate *where* the collapse lives, we probe the collapsed adapter without retraining: for every LoRA module (439 across the policy) we score rank channels by $\|A_i\| \cdot \|B_{:,i}\|$ (each rank's

directional contribution to the adapter output), zero out the top-3 or a random-3 of the 32 channels *everywhere*, and evaluate on the pre-collapse task $t_3{=}7$. The ablated policies are functionally intact, and there is *no* directional-vs-random differential in either arm. This *null* result is itself mechanistic: the learned skill is not concentrated in a few critical directions, so the collapse is not localized feature-death. It localizes the failure at optimization dynamics (the collinear-update, momentum-amplified instability of Section 4.4), consistent with the distributed (superposition) view of the rank budget.

Reoptimization after ablation is therapeutic, and directionally so. Retraining from the ablated $t_3{=}7$ checkpoints into the collapse-stage task ($t_4{=}8$), under the identical GRPO pipeline, establishes that ablation is also curative. On the anchored arm (whose intact continuation collapses to 0.000), directional ablation then retraining restores acquisition to 0.275 (retention 0.633), whereas random ablation reaches only 0.0625: a directional differential in the very regime where the zero-training probe showed none, reproducing across seeds ($4.4\times$ and $1.5\times$; Section D). On the baseline arm, ablation is mildly beneficial: the directionally ablated chain reaches 0.538/0.783 under seed 42 (improving on both axes over the intact 0.463/0.775), and under seed 1 both ablations again improve acquisition without retention cost. Three conclusions follow. First, releasing the top-σ rank channels is the operative ingredient: harmless when the stream is healthy, therapeutic when collapsed. Second, channel-structured release outperforms element-wise sparsification at lower intervention strength (0.275 vs. DARE's 0.113): the budget should be freed *coherently*, not noisily. Third, the frontier itself is movable: directional rank release improves acquisition without meaningful retention cost on the healthy baseline under both seeds (Fig. 1): a candidate capacity-management policy, not merely a repair. Notably, the directional gap is specific to the anchored arm: on the healthy baseline, directional and random pruning perform nearly identically (1.0–$1.1\times$ across seeds), so channel *direction* matters precisely where instability lives.

4.4 Risk predictor validation

Predictor features on existing streams. Table 4 reports the three driver features, computed on CPU from the adapters of our four pilot streams and the Phase-1 baseline chain. In all four short two-task streams similarity is uniformly high ($\mathrm{cka}_{\mathrm{median}} \geq 0.899$): consecutive GRPO updates occupy nearly the same low-rank direction space, the structural signature of a stream that *cannot* forget but may *exhaust* its subspace ("subspace crowding"). The Phase-1 baseline control adds a measurement caution: its adapters are *independently* fine-tuned from the same base state, so their low similarity and non-growing norms are a *calibration floor* (what raw task difficulty alone drives), not evidence of sequential plasticity loss (Section B).

Structural fingerprint of the anchoring collapse. The chained anchored arm's trajectory (Table 4, last row) turns calibration into mechanism: adjacent-update similarity climbs monotonically ($0.921 \rightarrow 0.979$), "subspace crowding" in its purest observed form, while the adapter norm inflates without bound ($0.97 \rightarrow 2.34$) and effective rank contracts ($20.8 \rightarrow 19.3$), with the largest norm jump ($+22\%$) coinciding exactly with the behavioral collapse. The picture is *directional lock-in with amplitude runaway*: updates stay collinear with prior directions while their magnitude accumulates, until the policy is driven into a degenerate high-norm mode. The completed baseline chain sharpens the attribution: its effective rank *expands* monotonically across all five stages ($20.2 \rightarrow 21.2$) where the anchored arm contracts. Both arms share the CKA climb and norm growth, so subspace crowding is intrinsic to chained GRPO; the *contraction* that precedes collapse is specific to anchoring. Fig. 2 visualizes the three trajectories side by side; both signatures reproduce under seed 1 (anchored: -10.6% contraction, norm $0.98{\rightarrow}2.64$; baseline: pointwise-matched trajectories; Table 2 and Section D).

Risk ordering. Optimizer state shows the runaway is *acute, not chronic*: across only eight GRPO iterations in the collapse segment, Adam's first moment grows $+90\%$ and second moment $+77\%$: an explosive feedback between degenerate advantages and collinear updates, not slow optimizer-state exhaustion (Lyle et al., 2025). And across the pilot streams, ordering by R_t matches behavioral observation: the anchored heterogeneous stream has the largest plasticity cost, the lowest CKA, and the largest update norm; a leave-one-stream-out linear fit reproduces the risk ordering exactly (Spearman $\rho = 1.0$ at $n{=}4$; a consistency check, with the forward-validation design in Section B).

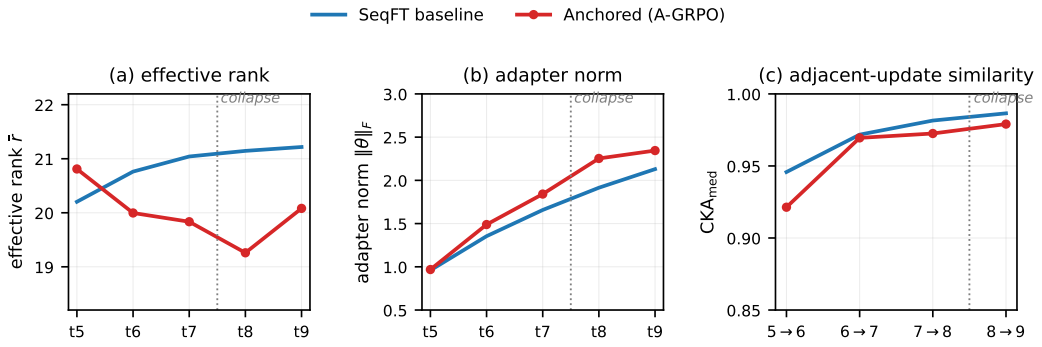


Figure 2: Structural fingerprint of the anchoring collapse along the chained stream (seed 42). (a) Mean effective rank $\bar{r}$ of the LoRA adapter (r=32): the anchored arm contracts monotonically into the collapse stage (20.8 → 19.3, recovering to 20.1 afterward), while the unanchored baseline *expands* (20.2 → 21.2). (b) Adapter Frobenius norm: both arms grow, but the anchored arm's growth is faster (0.97 → 2.34 vs. 0.96 → 2.13) with its steepest jump (+22%, $t_7 \to t_8$) coinciding with the behavioral collapse. (c) Linear CKA between consecutive updates (median over 437 layers): both arms saturate near 0.98, evidence that subspace crowding is intrinsic to chained GRPO while the contraction of (a) is anchoring-specific. Dashed vertical line: collapse stage.

Table 4: Predictor features across the pilot streams (2-task). $\text{cap}_t = 1 - \text{erank}/r$, r=32; CKA = linear CKA between consecutive tasks' top-16 adapter direction sets, median over 437 layers; overlap = SVD principal-subspace overlap. Behavioral column: task-0 acquisition after the 2-task stream (pilot).

Stream	cap_t (mean)	CKA_{med}	$\text{Overlap}_{\text{med}}$	$\|\Delta\theta\|$
SeqFT isotropic (0→1)	0.36	0.956	0.932	0.98→1.38
SeqFT hetero (9→0)	0.34	0.933	0.918	0.97→1.36
Anchored isotropic (0→1)	0.37	0.921	0.919	0.96→1.44
Anchored hetero (9→0)	0.36	0.899	0.909	0.96→1.47
Indep-FT Phase-1 (6→9)	0.36	0.310	0.632	0.98→0.93
Anchored chained (6→9)	0.39	0.921→0.979	0.913→0.966	0.97→2.34
SeqFT chained (5→9)	0.33	0.946→0.987	0.922→0.977	0.96→2.13

Full validation target. R_t vs. measured forgetting across (benchmark, learner) pairs (pretrained VLA on LIBERO-Spatial, scratch-BC, no-replay sequences, and our destructive discriminator) is deferred to Section B. Two design points anchor that future test. The Phase-1 baseline's four adapters were *independently* fine-tuned from the same base state; their low adjacent similarity ($\text{cka}_{\text{median}} \approx 0.31$) and non-growing norms (0.98→0.93) therefore constitute a calibration floor for task difficulty alone, not evidence of sequential plasticity loss. And the forward-validation protocol is fixed: fit on the early stream pairs and predict late-stream acquisition loss, which the A-GRPO twin arm of the chained experiment is the first to enable.

5 Discussion and Limitations

We present negative results as evidence, not as failures to explain away. Our pilots show that (i) the baseline does not forget even under a destructive hard-to-easy ordering: retention *improves* from 0.150 to 0.238; (ii) the acquisition bottleneck is real and measurable; and (iii) a forgetting-oriented regularizer pays a plasticity tax for zero retention gain. The honest reading is that LIBERO-Spatial with a pretrained VLA is structurally incapable of producing forgetting (Goldfarb et al., 2024), which is exactly why the forgetting-risk predictor, not another regularizer, is the centerpiece.

Limitations are explicit. The structural track is hypothesis-generating only (Lyle et al., 2025), so every structural claim is paired with a behavioral probe. The predictor is a $(benchmark \times learner)$-paired quantity, not a universal law; cross-suite and scratch-BC validation is required before gen-

erality claims. Risk-Gated Rejuvenation inherits the gate's assumption that advantage magnitude proxies task-relevant signal, which degrades under noisy or misspecified rewards. And the go/no-go gate stands: had Phase 1 found no acquisition decay, the protocol and predictor would stand as a protocols contribution with intervention claims withdrawn.

Finally, GRPO training exhibits large run-to-run variance: our two independent-FT controls differ by 11pp on the same task. We therefore replicated the full pipeline under a second seed (Section D): collapse and smooth decay reproduce with the onset shifting by one stage; the revival benefit reproduces with post-pruning retention 0.646 under both seeds; the directional differential holds qualitatively ($4.4\times$ and $1.5\times$). Within that policy, the headline contrasts are two orders of magnitude above the noise floor and *within-arm* (same checkpoint lineage), which cancels the seed-mismatch confound. What remain directional are fine-grained magnitudes: the directional advantage varies ($1.5\times$–$4.4\times$), and the baseline-arm frontier shift shows retention changes inside the noise band; a denser seed sweep would sharpen the effect-size estimates.

Our observations connect to three literatures (Section C). First, the rank budget of an $r{=}32$ adapter makes subspace crowding a measurable instance of *superposition* (Elhage et al., 2022): skills compete for a fixed set of update directions, and when the demanded directions exceed the budget, interference grows until behavior degrades; DARE's uniform dropping carries away the small-magnitude interference terms stacked in shared directions, which is why sparsification re-opens capacity rather than destroying skill. Second, successive-policy anchoring contracts *update* diversity just as reference-model anchoring contracts *output* diversity in RLHF (Kirk et al., 2024); that literature's remedies (mass-covering divergences, anchor-free objectives) are natural starting points for redesigning behavioral anchors. Third, the runaway is acute rather than chronic: within the collapse segment Adam's first and second moments grow $+90\%$ and $+77\%$ across only eight iterations, consistent with an explosive feedback between degenerate advantages and collinear updates rather than slow optimizer-state exhaustion (Lyle et al., 2025). Together these readings suggest the anchoring objective's plasticity tax is not a tuning artifact but a structural consequence of how on-policy updates interact with a bounded rank budget.

6 Conclusion

Forgetting was the field's problem; our evidence suggests it has largely been solved by emergence. The remaining problem is plasticity: the ability of LoRA-adapted VLA policies to keep acquiring new skills under on-policy GRPO. Our chained-stream experiments show it can fail catastrophically rather than gradually: a forgetting-oriented anchoring objective drove the policy to zero success on all ten tasks at the stream's fourth stage, with a structural fingerprint (update-direction similarity saturating, adapter norm inflating, effective rank contracting) that preceded the behavioral event and was absent in the unanchored baseline. That fingerprint is not merely diagnostic: DARE sparsification restores acquisition on the collapse stage with retention within 13pp of the baseline (Table 3), and both phenomena replicate across two seeds (Section D), though the revived stream re-collapses at the next stage, so durable restoration likely requires sustained gated intervention. Rank-channel ablation localizes the failure to optimization dynamics (Section 4.3); reoptimization from the directionally ablated checkpoint is therapeutic and, on the healthy baseline, moves the retention–acquisition frontier itself. We contribute the measurement protocol that sees collapse coming, the rejuvenation family that reverses it, and the risk predictor that anticipates it: deliverables that let the field answer the stability-of-acquisition question *before the collapse*.

AI use statement

We used generative AI tools (large language models) to assist with literature search, drafting boilerplate text, and debugging experimental code. All research ideas were validated against the literature by manual survey; all code was executed and verified by the authors; all experimental numbers reported in this draft originate from our own runs. We take full responsibility for the final content.

Reproducibility statement

We release (anonymized) code for (i) the plasticity measurement protocol (structural diagnostics on CPU from LoRA checkpoints, injection probe, warm-start gap, acquisition curves), (ii) the Risk-

Gated Rejuvenation objective as a localized modification to the GRPO policy loss, and (iii) the risk-predictor feature pipeline (CKA similarity, erank/dormancy, update norms). Full hyperparameters are specified in Section A. Models (OpenVLA-OFT SFT checkpoint) and benchmarks (LIBERO) are publicly available.

REFERENCES

Jordan T Ash and Ryan P Adams. On warm-starting neural network training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. arXiv:1910.08475.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P Cunningham. Lora learns less and forgets less. *Transactions on Machine Learning Research*, 2024. arXiv:2405.09673.

Thang Doan, Mehdi Bennani, Bogdan Mazoure, Guillaume Rabusseau, and Pierre Alquier. A theoretical analysis of catastrophic forgetting through the ntk overlap matrix. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021. arXiv:2010.04003.

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

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Jared Bowman. Toy models of superposition. `https://transformer-circuits.pub/2022/toy_model/index.html`, 2022. Transformer Circuits Thread, Anthropic.

Daniel Goldfarb, Itay Evron, Nir Weinberger, Daniel Soudry, and Paul Hand. The joint effect of task similarity and overparameterization on catastrophic forgetting: An analytical model. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024. arXiv:2401.12617.

Jiaheng Hu, Jay Shim, Chen Tang, Yoonchang Sung, Bo Liu, Peter Stone, and Roberto Martín-Martín. Simple recipe works: Vision-language-action models are natural continual learners with reinforcement learning. *arXiv preprint arXiv:2603.11653*, 2026.

Arthur Juliani and Jordan T Ash. A study of plasticity loss in on-policy deep reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2405.19153.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. In *Robotics: Science and Systems (RSS)*, 2025. arXiv:2502.19645.

Hannah Rose Kirk, Bertie Vidgen, Paul Røttger, and Scott A. Hale. Understanding the effects of RLHF on LLM generalisation and diversity. In *International Conference on Learning Representations (ICLR)*, 2024.

Timo Klein, Christoph Luther, Manus McAuliffe, Lukas Miklautz, Claudia Plant, and Sebastian Tschiatschek. Plasticity loss in deep reinforcement learning: A survey. *arXiv preprint arXiv:2411.04832*, 2024.

Saurabh Kumar, Henrik Marklund, and Benjamin Van Roy. Maintaining plasticity in continual learning via regenerative regularization. In *Conference on Lifelong Learning Agents (CoLLAs)*, 2025. arXiv:2308.11958.

Sebastian Lee, Sebastian Goldt, and Andrew Saxe. Continual learning in the teacher-student setup: Impact of task similarity. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. arXiv:2107.04384.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2306.03310.

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

Mao-Lin Luo et al. Rl forgets! towards continual policy optimization for reinforcement learning. *arXiv preprint arXiv:2607.04364*, 2026.

Clare Lyle, Zeyu Zheng, Khimya Khetarpal, Hado van Hasselt, Razvan Pascanu, James Martens, and Will Dabney. Disentangling the causes of plasticity loss in neural networks. In *Proceedings of The 3rd Conference on Lifelong Learning Agents (CoLLAs), PMLR*, volume 274, pp. 750–783, 2025. arXiv:2402.18762.

Evgenii Nikishin, Max Schwarzer, Pierluca D'Oro, Pierre-Luc Bacon, and Aaron Courville. The primacy bias in deep reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022. arXiv:2205.07802.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Ghada Sokar, Rishabh Agarwal, Pablo Samuel Castro, and Utku Evci. The dormant neuron phenomenon in deep reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023. arXiv:2302.12902.

Mingqi Yuan, Qi Wang, Guozheng Ma, Caihao Sun, Bo Li, Xin Jin, Yunbo Wang, Xiaokang Yang, Wenjun Zeng, Dacheng Tao, and Jiayu Chen. Plasticine: Accelerating research in plasticity-motivated deep reinforcement learning. *arXiv preprint arXiv:2504.17490*, 2025.

Qixin Zeng et al. Crl-vla: Continual vision-language-action learning. *arXiv preprint arXiv:2602.03445*, 2026.

A Appendix A: Protocol Details and Hyperparameters

Algorithm configuration. LoRA rank 32; GRPO group size 2, $\sim$2048 samples per update, learning rate 2×10^{-5}, clip ratio $[0.2, 0.28]$, temperature 1.6, max sequence length 512, gradient checkpointing on, 2 epochs per task. Risk-Gated Rejuvenation: λ swept over $\{10^{-3}, 10^{-2}, 10^{-1}\}$, gate $(c, \tau) = (0.5, 0.25)$. Structural metrics are computed on a fixed reference batch of 512 trajectories per task; dormant threshold follows Sokar et al. (2023); erank uses the adapter matrices $W_{\text{up}}, W_{\text{down}}$ per layer and averages over layers.

Plasticity injection probe. After each task, 10% of adapter parameters are reset to their zero-initialized values, and the policy is re-optimized for $1/4$ of a task budget on the same task's data. Recovery is measured as success relative to an unperturbed reference run with the same budget; probe recovery curves are reported in the camera-ready appendix.

DARE sparsification (intervention member, §4.2). On the converged t_3=7 adapter of the anchored chain, each adapter entry is drawn with Bernoulli keep-probability $p_{\text{keep}} = 0.75$ and survivors rescaled by $1/0.75$ (mean-preserving); this drops 27.7M of 110.8M entries (25%) and raises the LoRA-2 norm from 220 to 254 as the rescaling (deterministic entropy re-normalization). The sparsified checkpoint is then chained forward onto task 8 exactly as a normal stage checkpoint. No retraining, no extra loss, no λ.

Structural metric details (§3.1). Dormant ratio: a unit (an element of the adapter's hidden activations) is dormant if its activation on the reference set is below a threshold relative to a running estimate of its maximum; monotonic growth is the classic capacity-erosion signature, and for LoRA we compute dormancy on the internal activations $h = W_{\text{down}}x$, where the adapter's learned function lives. Warm-start gap: for each task we measure acquisition initializing from the previous adapter versus from the SFT checkpoint ("cold") under a fixed optimization budget; a growing gap is an on-policy-specific plasticity signature (Ash & Adams, 2020; Juliani & Ash, 2024).

Rank-channel ablation probe (§4.3). For every LoRA module (439 across the policy), rank channels $i = 1, \ldots, 32$ are scored by $\|A_{(i)}\| \cdot \|B_{(:,i)}\|$ (directional contribution of channel i to the adapter

output), and either the top-3 scoring channels or a uniformly random 3 channels are zeroed *simultaneously in every module*; the resulting policy is evaluated directly (zero training) on the pre-collapse stage task t_3=7.

B Appendix B: Risk-Predictor Calibration and Validation Design

Indep-FT calibration floor (§4.4). The Phase-1 baseline's four adapters are *independently* fine-tuned from the same base state: low adjacent-task similarity ($\mathrm{cka}_{\mathrm{median}} \approx 0.31$, SVD overlap ≈ 0.63), sub-unity per-task norm ratios (0.96–0.99, vs. 1.40–1.51 in the pilots), and non-growing norms ($0.98 \rightarrow 0.93$) are what task-to-task structural variation looks like when no chain couples the updates, with effective rank dipping only at the hardest task ($20.8 \rightarrow 19.8$ of r=32). Against this floor, the completed chained arms supply the paired sequential measurements the plasticity-loss question requires: the baseline filling the gentle-decay profile and A-GRPO the collapse profile (Section 4.1).

Forward-validation design (§4.4). The n=4 pilot rank-ordering (Spearman $\rho = 1.0$) is a consistency check: three samples cannot resolve three feature weights, so the fitted scores collapse onto a near-degenerate rank. The completed Phase-1 baseline adds three labeled pairs spanning a much wider feature range, and the A-GRPO twin arm enables the forward-validation test (fit on early pairs, predict late-stream acquisition loss).

C Appendix C: Cross-Domain Connections

Superposition accounting. The rank budget of an r=32 adapter makes subspace crowding a measurable instance of *superposition*: learned skills compete for a fixed set of update directions whose mutual interference is governed by how sparsely each direction is loaded (Elhage et al., 2022): when demanded directions exceed the budget, interference grows until behavior degrades. DARE's uniform entry dropping carries away mostly the small-magnitude interference terms stacked in shared directions while large directional structure survives, which is why sparsification re-opens usable capacity rather than destroying skill.

KL-anchored alignment. Successive-policy anchoring contracts *update* diversity just as reference-model anchoring contracts *output* diversity in RLHF (Kirk et al., 2024); that literature's remedies (mass-covering divergences, anchor-free objectives) are the natural starting points for redesigning behavioral anchors.

Acute optimizer instability. Across only eight GRPO iterations in the collapse segment, Adam's first-moment magnitude grows $+90\%$ ($5.0\times10^{-7} \rightarrow 9.6\times10^{-7}$, median over tensors) and its second moment $+77\%$ with rising concentration: gradient momentum nearly doubles within a single stage, consistent with an explosive feedback between degenerate advantages and collinear updates rather than slow optimizer-state exhaustion (Lyle et al., 2025).

D Appendix D: Seed Consistency of the Main Results

To verify that the reported phenomena are not artifacts of a single stochastic realization, we repeat the full pipeline of Section 4 under seed 1: both five-stage chains, the DARE-75 revival probe, the four rank-channel re-optimization probes, and the structural diagnostics, with all protocols identical to seed 42 (80 evaluation episodes per task, evaluated on all ten tasks at every stage). Table 2 reports the stage-wise acquisition of both arms under both seeds; the per-task matrices are released with the code.

Revival under seed 1. The revival benefit reproduces, with its dimension determined by the collapse profile: under seed 42 (full collapse, acquisition 0.0) DARE restores *acquisition* ($0.0000 \rightarrow 0.1125$), whereas under seed 1 (partial collapse, acquisition 0.1875 but retention 0.2917) it restores *memory* (retention $0.2917 \rightarrow 0.6458$, +35pp at unchanged acquisition). Post-pruning retention is 0.6458 under both seeds, indicating the intervention recovers a reproducible operating level rather than a seed-specific artifact.

Directional ablation under seed 1. Top-σ channel pruning again outperforms random pruning (acquisition 0.2000 vs. 0.1375; seed 42: 0.2750 vs. 0.0625), with both arms retaining healthy memory (0.6875/0.6708). The directional advantage holds qualitatively under both seeds, though its magnitude varies ($1.5\times$–$4.4\times$); we therefore state the claim as directional superiority, not a fixed ratio. Notably, the gap is specific to the anchored arm: on the unanchored (baseline) arm, directional and random pruning perform nearly identically under both seeds (0.5375 vs. 0.5000 and 0.4125 vs. 0.4125), indicating that channel *direction* matters precisely where instability lives.

Baseline-arm re-optimization under seed 1. Directional re-optimization on the unanchored arm improves acquisition without meaningful retention cost under both seeds (acq $+7.5$pp / $+3.8$pp at ret changes within the ± 4pp noise band), consistent with a mild, reproducible frontier shift.

Structural fingerprints under seed 1. The mechanism signatures reproduce: anchored-arm mean effective rank contracts monotonically ($21.03 \rightarrow 18.79$, -10.6%) while update norm grows unboundedly ($0.98 \rightarrow 2.64$, $+169\%$), whereas the unanchored arm shows mild rank expansion ($21.67 \rightarrow 21.88$) and bounded norms ($1.66 \rightarrow 2.14$) with pointwise-matched trajectories across seeds.