Ex 2.1.8: Repulsion tuning

> We tested anti-subspace schedules (trailing timing and strength) to find an operating point that contains the cube-wide drift without reducing selectivity. Holding the repulsion high for longer contains the drift and keeps the margin, which nothing before this did. It's unclear if the response is well-graded.

In ex-2.1.6 we aligned *red* with a direction in the residual stream, but it dragged the whole color space with it. Then in ex-2.1.7 we added a repulsive *anti-subspace* term that pushed everything away from the *red* axis $\hat{v}_{\text{red}}$. It helped, but not enough: the mean *red*-alignment over all colors fell from $\bar{\alpha}$ = 0.53 to 0.33, and the target was < 0.1.

The training trajectories show what happens over the course of a run. While the repulsion outweighs the pull, $\bar{\alpha}$ stays near the control, which is what we want. Then the anneal[1] brings $\lambda_{\bar{s}}$ down to about a tenth of λ_{a}, and $\bar{\alpha}$ starts to climb again. Each condition turned at its own point in the schedule: around epoch 40 on the M1 schedule, and epoch 75 on `span-anti-late`, the arm that stretched it.

The M1 keyframes came from a 5-dimensional autoencoder bottleneck, mapped onto our 100 epochs by fraction of training. So we never had much reason to expect them to suit a 64-dimensional residual stream.

All that `span-anti-late` did was push the end of the anneal from epoch 50 to epoch 90. That improved every statistic the H2 and H4 hypotheses scored: margin 0.46 → 0.60, $\bar{\alpha}$ 0.33 → 0.13, retention across the anneal, and grading. For future experiments, then, should we stretch the anneal, or end it at a higher weight? Each option is a reading of the ex-2.1.7 result, and the rest of this report refers to them by name:

- The timing account: the repulsion has to still be there while the pull is strong enough to move the cube, and a later anneal keeps it there. Stretch the anneal.
- The level account: containment is lost when $\lambda_{\bar{s}}/\lambda_{a}$ falls past the threshold ex-2.1.7 observed, and a later anneal helps only by postponing that. End the anneal at a higher weight.

We test both. The two separate cleanly, because only the level account predicts an interaction: a floor that stays above the threshold never crosses it, so at that floor the timing of the endpoint should stop mattering.

▶ **Notation and glossary…**

Findings

H1 (task cost) — holds. Largest `named_holdout` exact-match gap from the control, across all 8 conditions: 0.0065. Gate: 0.02.

H2 (an operating point exists) — fails, on grading. `end90-hold30`, the preregistered primary condition, meets containment ($\bar{\alpha} = 0.087$, gate 0.1), margin (0.611, gate 0.5) and retention (0.97, gate 0.8). No condition in the grid reaches the grading gate of 0.8 on either track; the best R^2 is 0.78. Neither named partial applies.

H3 (a sustained floor makes the anneal endpoint redundant) — holds. Hold-ratio main effect on $\bar{\alpha}$: +0.126, gate 0.067. Interaction: +0.142, gate 0.133 — a pass by 6%. Both gates are the preregistered values rescaled by this experiment's own seed spread, as the H3 footnote directs.

Containment is now newly reachable. `end90-hold30` is the first condition in M2 to hold $\bar{\alpha}$ near the control while keeping the margin. The best in ex-2.1.7 was 0.13 and its gate went unmet everywhere.

> **How to read this report**
>
> The method, hypotheses and decision thresholds were written and frozen in Git before the experiment ran, and the results below replace the placeholders that stood in for them. Anything conceived after seeing the data goes under *Exploratory analyses*.

Method

Like ex-2.1.7: the word-level `v216` corpus from ex-2.1.3 (216 colors, one token each, d64-L4), and the same corpus, seeds, sequence-level labels, and measurements.

We use the same *anchor* schedule, but vary the *anti-subspace* schedule.

Since we're using sequence-level labels, this experiment is like the `span-*` conditions in ex-2.1.7. The op1-only pull in ex-2.1.7 scored higher, but that option exists only because this synthetic language has a known position carrying the concept. Once we retire that position oracle we have to pull on a span.

The anchor weight stays at $\lambda_{\mathrm{a}} = 0.1$. The ceiling arm in ex-2.1.7 showed selectivity falling at ten times this weight, with no task cost even there. So the previous scoring rung is still the right place to tune the repulsion against.

The sweep grid

Factors:

- **A.** Anneal endpoint

  The epoch at which $\lambda_{\bar{\mathrm{s}}}/\lambda_{\mathrm{a}}$ reaches its hold ratio. Levels: 50, 70 and 90. Endpoint 50 is the M1 keyframe mapped by fraction of training, and 90 is the timing arm from ex-2.1.7. So two conditions repeat settings we have already measured.

- **B.** Hold ratio

  The level it settles at. Levels: 0.03 and 0.30. Ratio 0.03 is the M1 level. Ratio 0.30 is three times the level at which ex-2.1.7 saw containment give way, so on the level account it should never cross the threshold at all.

We leave out endpoint 100: the anneal would run to the end of training, leaving no hold window for the ratio to act on, so that row would cross one factor against nothing and dilute its main effect.

An un-anchored control ($\lambda_{\mathrm{a}} = 0$) runs through the same code, giving 8 conditions and 24 runs in total.

A confound, and its mitigation

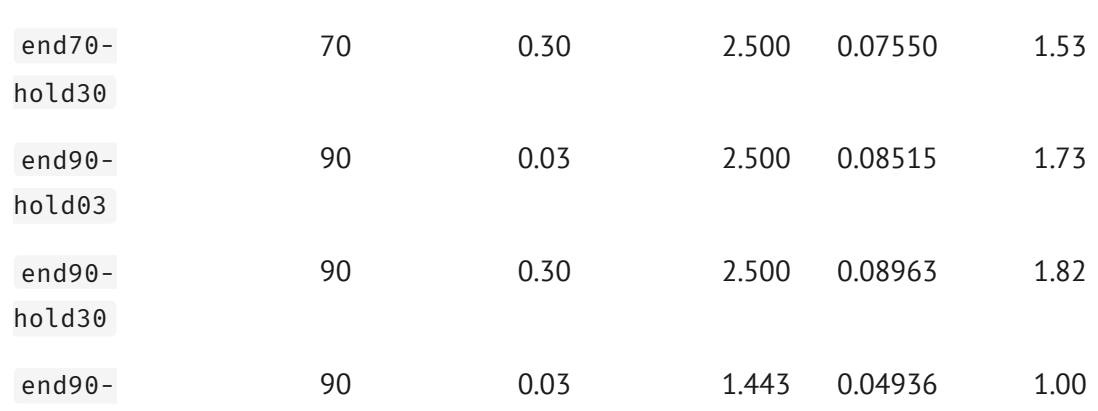
The two factors are not symmetric. Changing the hold ratio barely changes the total repulsion delivered. Changing the endpoint does, because a later anneal spends longer at a high weight. So along factor A, *when* the repulsion acts and *how much* of it there is move together.

> We call that total the *dose*, and measure it as $\int \lambda_{\bar{\mathrm{s}}}(e) \cdot \mathrm{lr}(e)\, de$ over training. The learning rate belongs inside the integral because it multiplies the gradient of every term, and it is itself warming up and annealing across the same window. A weight-only integral would call two schedules equal even when they deliver visibly different pushes.

condition	`anneal_end`	`hold_ratio`	`peak_ratio`	dose	÷ `end50-hold03`
`end50-hold03`	50	0.03	2.500	0.04936	1.00
`end50-hold30`	50	0.30	2.500	0.05774	1.17
`end70-hold03`	70	0.03	2.500	0.06929	1.40
`end70-hold30`	70	0.30	2.500	0.07550	1.53
`end90-hold03`	90	0.03	2.500	0.08515	1.73
`end90-hold30`	90	0.30	2.500	0.08963	1.82
`end90-hold03-dose`	90	0.03	1.443	0.04936	1.00

Across the hold ratio, dose moves by at most a sixth; across the endpoint it rises by about three quarters. So we don't need to dose-match the two levels of factor B. And we can't dose-match the three levels of factor A with the hold ratio without extending past the tested range of ex-2.1.7.[2]

Instead, we add an arm: `end90-hold03-dose` is `end90-hold03` with the opening ratio scaled 2.5 → 1.443, so it delivers the dose of `end50-hold03` on the late schedule. That's still asymmetrical: this arm has less repulsion through the epochs where the anchor is first ramping in. If it tracks `end50-hold03` rather than `end90-hold03`, weak early repulsion may be the cause. This should be visible in the trajectory figure.



The schedules under test. Left: *anti-subspace weight $\lambda_{\bar{\mathrm{s}}}$ over training, for the six factorial conditions and the dose arm.* ***Right:*** *the dose itself, as its integrand $\lambda_{\bar{\mathrm{s}}} \cdot$ lr, so area on the page is repulsion delivered. Hatching shows where the dose arm falls short of the reference early and where it makes that back late. Ghost ink in both panels: the anchor weight λ_{a}; on the right, also `end90-hold03`, the schedule the arm was scaled down from.*

Measurements

Like ex-2.1.7:

- `named_holdout` exact match, against the in-experiment control (H1).
- m_{op1}: the label-affinity-weighted alignment at op1 minus the cube mean, averaged over layers (H2, H3).
- $\bar{\alpha}$: mean alignment over all 216 colors at op1, layer-averaged (H2, H3). This is the quantity the repulsive term is defined on.
- Grading: Spearman ρ against `redness` and Pearson R^2 against `sim¹·⁵` over the 216 colors (H2).
- Per-run margin trajectories, for retention (H2).

Ex-2.1.7 quoted *retention* two ways under one name: the minimum final-over-peak margin across seeds in its Findings, and the mean across seeds in its Arms section. This report uses the **min**, and every later report should. The gate of H4(b) is worded per run ("for every anchored run…"), so the minimum is what it needs.

Hypotheses

Gates and noise floors carry over from ex-2.1.7 with their thresholds and rationale, since we have the same architecture, the same measurement, and the same statistics.

H1: Task cost. Every condition is task-clean, meaning `named_holdout` exact match within 0.02 (absolute) of the seed mean of the in-experiment control.[3] Partial: the `hold03` conditions and the dose arm are clean and a `hold30` condition is not.

H2: An operating point exists. Some task-clean factorial condition meets all four of the selectivity and containment criteria from ex-2.1.7 at once, which no condition in that experiment did: (a) containment, $\bar{\alpha} \leq 0.1$; (b) margin, $m_{\text{op1}} \geq 0.5$; (c) grading, Spearman $\rho \geq 0.8$ against `redness` or Pearson $R^2 \geq 0.8$ against `sim¹·⁵`; (d) retention, every run whose running maximum of m_{op1} reaches 0.2 ending at $\geq 0.8\times$ that maximum. Partial, in decreasing strength: (a) is met at the looser 0.25 (the halving level from ex-2.1.7) while (b)–(d) all hold; or (a) is met at 0.1 but the margin falls between the ex-2.1.7 value of 0.46 and the gate.

H3: A sustained floor makes the anneal endpoint redundant. This is the prediction that sets the level account apart. If containment is lost at a threshold, then the endpoint only has leverage in the row whose floor crosses that threshold. We score H3 on $\bar{\alpha}$ with two criteria, each a difference of means reported with its sign. (a) The hold-ratio main effect,[5] the nine runs at 0.03 minus the nine at 0.30, is at least 0.05. That says the floor does something at all. (b) The interaction is at least 0.1: the end50 − end90 difference within the 0.03 row, minus the same difference within the 0.30 row.[6] Criterion (a) keeps (b) from passing without any containment behind it: a *negative* endpoint effect in the 0.30 row would inflate the interaction on its own.

Two outcomes would fail H3. **1.** The endpoint effect within a row is the three runs at endpoint 50 minus the three at endpoint 90. If it clears 0.05 in both rows while the interaction stays under 0.1, that is the timing account: the repulsion has to be present while the pull is strong, and no floor stands in for it. **2.** If `end90-hold30` alone contains the drift while (a) fails, then each factor is blocked by the other.[4]

We also report, without scoring it, whether the hold-ratio effect is at least as large as the endpoint effect. In the clean outcomes this ordering agrees with the interaction. In the partial ones it can disagree, because a floor that slows the climb without preventing it can come out ahead on the ordering while the interaction stays small. Where they disagree, we trust the interaction.

Reproduction check

condition	m_{op1} here	ex-2.1.7	Δ	ᾱ here	ex-2.1.7	Δ	ᾱ tolerance
`end50-hold03`	0.456	0.46	-0.004	0.326	0.33	-0.004	±0.017
`end90-hold03`	0.598	0.60	-0.002	0.132	0.13	+0.002	±0.012

The two conditions this grid shares with ex-2.1.7, against that report's published seed means. `end50-hold03` *repeats its* `span-anti` *and* `end90-hold03` *its* `span-anti-late`*. The margin tolerance is the carried seed-mean noise floor, 0.032; the ᾱ tolerance is two seed-mean spreads from this experiment's own seeds, in the last column.*

Both conditions reproduce. The margins land 0.004 from the published values against a floor of 0.032, and ᾱ within 0.004, inside the tolerance in either row. Grading agrees as well: `end90-hold03` scores $R^2 = 0.78$ here, the value ex-2.1.7 published for the same condition.

So the two experiments measure the same thing, and this grid can be read against the ex-2.1.7 numbers — which the H2 partials and the H3 power argument both rely on.

Task cost (H1)

condition	seen EM ↑	holdout EM ↑	holdout NLL ↓	`open` dist ↓	Δ holdout EM	H1 gate
`lam0`	1.000 ±0.000	0.999 ±0.002	0.019 ±0.009	0.133 ±0.002	+0.0000	clean
`end50-hold03`	1.000 ±0.000	0.997 ±0.004	0.020 ±0.010	0.135 ±0.006	-0.0013	clean
`end50-hold30`	1.000 ±0.000	1.000 ±0.000	0.015 ±0.000	0.139 ±0.002	+0.0013	clean
`end70-hold03`	1.000 ±0.000	0.992 ±0.004	0.030 ±0.010	0.138 ±0.004	-0.0065	clean
`end70-hold30`	1.000 ±0.000	0.993 ±0.006	0.035 ±0.017	0.143 ±0.006	-0.0052	clean
`end90-hold03`	1.000 ±0.000	0.999 ±0.002	0.021 ±0.002	0.137 ±0.002	+0.0000	clean
`end90-hold30`	1.000 ±0.000	0.992 ±0.004	0.030 ±0.006	0.134 ±0.003	-0.0065	clean
`end90-hold03-dose`	1.000 ±0.000	0.996 ±0.004	0.023 ±0.007	0.136 ±0.004	-0.0026	clean

Behavior by condition. Each cell is the seed mean, with half the seed range beside it. EM is exact match on the answer token, NLL its surprise in nats, and `open` dist the RGB distance from the guessed color to the true mix on pairs with no named answer. Δ holdout EM is the signed gap to the in-experiment control, which the H1 gate scores at 0.02.

H1 holds. The largest gap from the control is 0.0065, at `end70-hold03`, against a gate of 0.02. Every condition solves the task.

The preregistration named the `hold30` row as the place a task cost would show up, since it sustains a repulsive weight no earlier experiment held. It does not: the worst condition in that row is `end90-hold30` at 0.0065, the same size as the gaps elsewhere. Holding the repulsion an order of magnitude above the M1 floor for the rest of training costs nothing the task loss can see.

An operating point (H2)

Four criteria have to hold at once. The table below scores them together, with values that pass their gate in bold.

condition	end	hold	$\bar{\alpha}$ ↓	m_{op1} ↑	ρ ↑	R^2 ↑	retention ↑	H2
`end50-hold03`	50	0.03	0.326 ±0.015	0.456 ±0.027	0.61	0.62	0.72	containment, margin, grading, retention
`end50-hold30`	50	0.30	0.139 ±0.050	**0.597** ±0.061	0.62	0.76	**0.90**	containment, grading
`end70-hold03`	70	0.03	0.249 ±0.027	**0.514** ±0.033	0.57	0.66	0.75	containment, grading, retention
`end70-hold30`	70	0.30	0.104 ±0.027	**0.615** ±0.037	0.56	0.75	**0.95**	containment, grading
`end90-hold03`	90	0.03	0.132 ±0.011	**0.598** ±0.014	0.64	0.78	**0.92**	containment, grading
`end90-hold30`	90	0.30	**0.087** ±0.006	**0.611** ±0.025	0.56	0.77	**0.97**	grading

*The H2 scoring table, one row per factorial condition. **Bold** marks a value that passes its gate: $\bar{\alpha} \le 0.1$, $m_{op1} \ge 0.5$, $\rho \ge 0.8$ or $R^2 \ge 0.8$, retention ≥ 0.8. $\bar{\alpha}$ and m_{op1} carry the seed standard deviation; grading is read off the seed-mean response, and retention is the minimum across seeds. The last column names every criterion a condition misses, or reports that it meets all four. `end50-hold03` and `end90-hold03` are ex-2.1.7's `span-anti` and `span-anti-late`.*

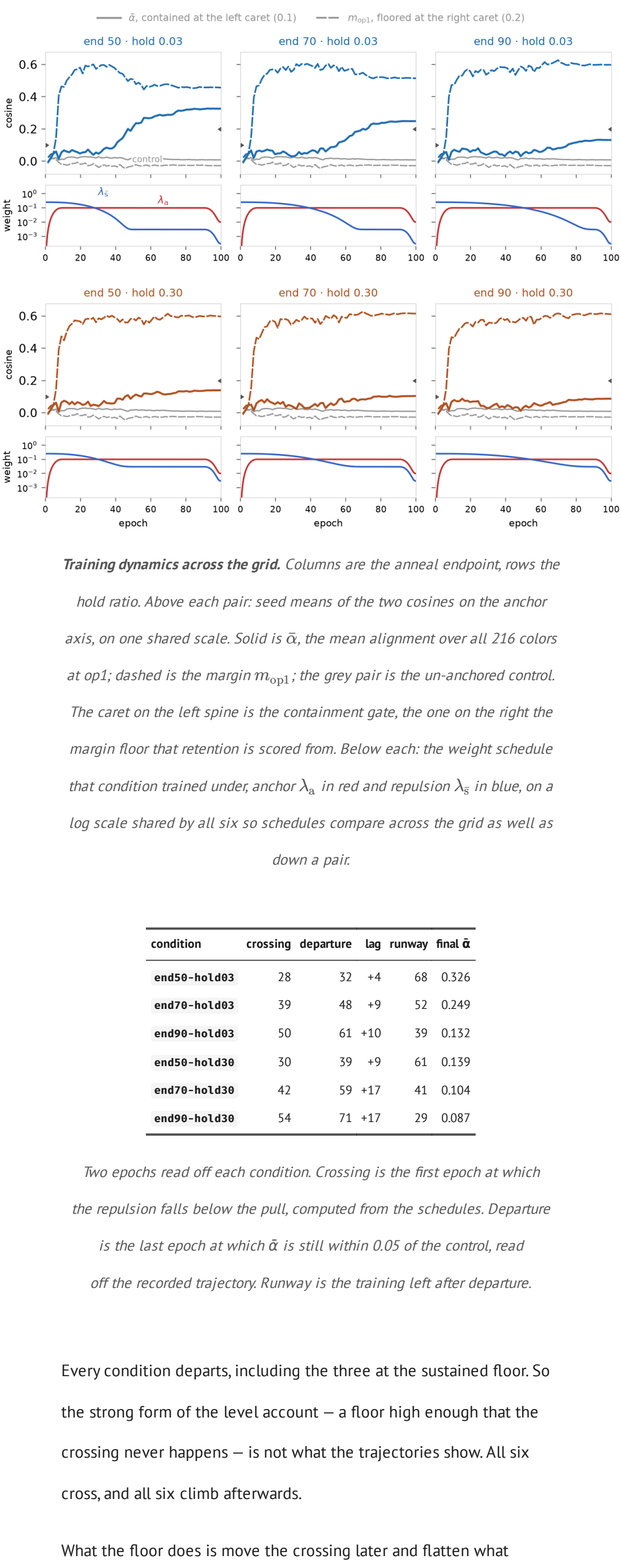
H2 fails, on grading. The condition the preregistration named, `end90-hold30`, meets three of the four criteria: it contains the drift at $\bar{\alpha} = 0.087$ against a gate of 0.1, holds the margin at 0.611 against 0.5, and retains 0.97× its peak against 0.8. But its response is graded at $\rho = 0.56$ and $R^2 = 0.77$, and the gate is 0.8 on either track.

Every condition misses on grading. The best R^2 in the grid is 0.78 (`end90-hold03`) and the best ρ is 0.64 (`end90-hold03`), so no condition comes within 0.02 of passing. Neither named partial applies, since both assume grading holds and vary the containment or margin around it.

The contrary result the preregistration named — containment bought by flattening the response — is not what happened. As the anneal is stretched, $\bar{\alpha}$ falls and the margin rises together: `end50-hold03` sits at $\bar{\alpha}$ 0.326 with margin 0.456, and `end90-hold03` at 0.132 with 0.598. Grading rises with them, from 0.62 to 0.78. So stretching the anneal buys selectivity, and just stops short of the level H2 asked for.

Containment on its own is now reachable, which it was not in ex-2.1.7. One condition of the six falls to $\bar{\alpha} \le 0.1$ (`end90-hold30`), and a second comes within 0.004 of it. That experiment's best was 0.13, and its gate went unmet everywhere. Both of the near conditions are in the `hold30` row.

The endpoint table says where each run finished. The trajectories say how it got there, which separates the two accounts H3 is built on.



***Training dynamics across the grid.** Columns are the anneal endpoint, rows the hold ratio. Above each pair: seed means of the two cosines on the anchor axis, on one shared scale. Solid is $\bar{\alpha}$, the mean alignment over all 216 colors at op1; dashed is the margin m_{op1}; the grey pair is the un-anchored control. The caret on the left spine is the containment gate, the one on the right the margin floor that retention is scored from. Below each: the weight schedule that condition trained under, anchor λ_{a} in red and repulsion $\lambda_{\bar{\mathrm{s}}}$ in blue, on a log scale shared by all six so schedules compare across the grid as well as down a pair.*

condition	crossing	departure	lag	runway	final $\bar{\alpha}$
`end50-hold03`	28	32	+4	68	0.326
`end70-hold03`	39	48	+9	52	0.249
`end90-hold03`	50	61	+10	39	0.132
`end50-hold30`	30	39	+9	61	0.139
`end70-hold30`	42	59	+17	41	0.104
`end90-hold30`	54	71	+17	29	0.087

Two epochs read off each condition. Crossing is the first epoch at which the repulsion falls below the pull, computed from the schedules. Departure is the last epoch at which $\bar{\alpha}$ is still within 0.05 of the control, read off the recorded trajectory. Runway is the training left after departure.

Every condition departs, including the three at the sustained floor. So the strong form of the level account — a floor high enough that the crossing never happens — is not what the trajectories show. All six cross, and all six climb afterwards.

What the floor does is move the crossing later and flatten what follows. Departure tracks the crossing in every condition, and raising the hold ratio pushes both later, the departure by more than the crossing. That is the delay. The flattening shows up when the departure epoch is held roughly fixed and the row changes: `end70-hold30` and `end90-hold03` leave the control within two epochs of each other, and finish at $\bar{\alpha}$ = 0.104 against 0.132. Same runway, different climb rate.

Both effects run the same way, which is why the endpoint and the floor look interchangeable in the endpoint table and are not. The margin, dashed in every panel, climbs early while the repulsion is strong and then holds across the whole anneal; nothing in the grid costs it.

Attribution (H3)

H3 has a footnote that asks for the seed spread before either statistic is used, and allows the gates to move with it. Pooled over the six factorial conditions, the seed spread of $\bar{\alpha}$ is 0.0266, against the ex-2.1.6 control value of 0.02 the gates were sized on. The anchored conditions do not hold it, so the gates move by 1.33×: the main effect is scored at 0.067 and the interaction at 0.133. Those are the numbers used below.

One condition accounts for most of that spread. The rest sit between 0.006 and 0.027, so the rescaling is driven by a single condition rather than by the grid being noisy throughout.

$\bar{\alpha}$ ↓	end 50	end 70	end 90	row mean
hold 0.03	0.326	0.249	0.132	0.235
hold 0.30	0.139	0.104	0.087	0.110
m_{op1} ↑	**end 50**	**end 70**	**end 90**	**row mean**
hold 0.03	0.456	0.514	0.598	0.523
hold 0.30	0.597	0.615	0.611	0.607

The factorial, as condition means on both statistics. Above: $\bar{\alpha}$, which H3 is scored on. Below: the margin, repeated because a factor that contains the drift while losing the margin does not give an operating point.

H3 holds, on both conditions. The hold-ratio main effect is +0.126 against a gate of 0.067, so the floor does something. The interaction is +0.142 against 0.133: the end50 − end90 difference is +0.194 in the `hold03` row and +0.053 in the `hold30` row, so the endpoint has about 3.7 times as much leverage at the low floor as at the high one.

Neither named miss occurred. Miss 1 needed the endpoint effect to clear 0.067 in both rows with the interaction under its gate; the interaction is over it. Miss 2 needed the main effect to fail, and it passes by a factor of 1.9.

However, the interaction only clears its gate by 0.008, about 6% — on a gate that moved because of this experiment's own seed spread. And the level account predicted about +0.20, with the whole endpoint effect inside the low row; the observed +0.142 is below that, because the endpoint still moves $\bar{\alpha}$ by +0.053 in the high row, where that account expects nothing.

So the endpoint keeps some leverage even at the sustained floor. The trajectories show that the floor postpones the crossing and slows the climb after it, but it still crosses. The statistic H3 was built on separates the accounts in the direction it was meant to, and the mechanism behind it is milder than the account it favors.

We also said we would report the ordering of the main effects without scoring it. The hold-ratio effect (+0.126) is larger than the endpoint effect averaged over the two rows (+0.124), which agrees with the interaction.

The dose arm

condition	end	peak ratio	dose	$\bar{\alpha}$ ↓	m_{op1} ↑	R^2 ↑	retention ↑
`end50-hold03`	50	2.500	1.00	0.326 ±0.015	0.456	0.62	0.72
`end90-hold03`	90	2.500	1.73	0.132 ±0.011	0.598	0.78	0.92
`end90-hold03-dose`	90	1.443	1.00	0.185 ±0.019	0.565	0.72	0.88

The dose arm against the two conditions it is built from. Dose is relative to `end50-hold03`, as in the Method table. The arm has that condition's dose and the `end90-hold03` timing, so its position between them is the comparison.

The arm lands at $\bar{\alpha}$ = 0.185, between `end50-hold03` (0.326) and `end90-hold03` (0.132), and about 73% of the way across. So most of what the endpoint buys is timing, and some of it is dose.

That should be taken loosely: this is one condition on three seeds, it scores nothing, and scaling the opening ratio to 1.443 weakens the repulsion through the epochs where the anchor first ramps in, which the Method section named as a confound of its own. The reading it does support is that the endpoint effect is not *merely* a dose effect, because holding the dose fixed leaves most of the gap intact.

Secondary measurements

Carried from ex-2.1.7 without a gate. Two probes read redness from complementary parts of the residual stream at op1: one from the anchor coordinate alone, and one from the other 63 directions. The first starts empty and measures how much redness anchoring put there. The second has an RGB floor it cannot go far below, since redness is a function of color and the task needs the color cube, so it reports whether the cube survived rather than whether the concept moved.

condition	on-axis R^2	off-axis, embed	off-axis, last	vs floor	cube spread
`lam0`	0.015	0.842	0.848	-0.014	0.125
`end50-hold03`	0.483	0.882	0.866	+0.004	0.131
`end50-hold30`	0.543	0.876	0.887	+0.024	0.117
`end70-hold03`	0.477	0.894	0.892	+0.029	0.121
`end70-hold30`	0.504	0.887	0.899	+0.036	0.130
`end90-hold03`	0.561	0.891	0.904	+0.041	0.147
`end90-hold30`	0.512	0.892	0.906	+0.043	0.132
`end90-hold03-dose`	0.528	0.910	0.921	+0.058	0.146

Where redness is readable at op1, and what the color cloud looks like there. On-axis R^2 is redness read from the anchor coordinate alone, averaged over layers. The two off-axis columns are the ridge probe on the other 63 directions, at the token embedding and at the last layer, and the next column is the last-layer value against the RGB floor of 0.863. Cube spread is the mean squared distance of the 216 op1 states from their centroid at the last layer: how much extent the cloud keeps. All are seed means, and none is scored.

The anchor coordinate carries redness in every anchored condition, from $R^2 = 0.48$ to 0.56 (`end90-hold03`), against 0.02 in the control. The spread across the grid is narrow, and it does not order the conditions the way $\bar{\alpha}$ does, so this is not a second containment measurement.

Off the axis, every condition sits within 0.06 of the RGB floor at the last layer, and the anchored conditions all sit slightly above it (0.87–0.92) where the control sits just below (0.85). Above the floor is reachable because the residual stream encodes color nonlinearly and redness is quadratic in RGB, so a linear probe does better there than on raw channels. Nothing here reads as a second copy of the concept living off the axis, and nothing reads as the cube being lost: spread at the last layer runs 0.117–0.147, with the control at 0.125 in the middle of that range.

This is the point ex-2.1.7 made and this grid does not change: containment is not exclusivity. Holding $\bar{\alpha}$ near the control keeps the cube off the anchor direction; it does not stop redness being readable from everywhere else, because the task needs the color cube either way.

Exploratory analyses

Post hoc, and scoring nothing. H2 failed on grading in every condition, so the question is what the response actually looks like.

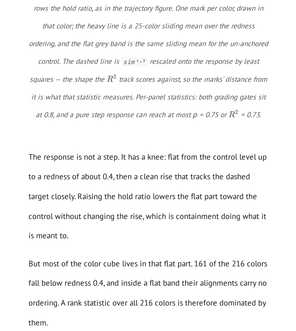
The grading metrics disagree

The grading metrics are meant to rule out a *step*, in which very *red* colors would be aligned but everything else orthogonal. Both metrics cap out on a step, at the best it can score over all step locations: ρ can reach 0.75 and R^2 0.73, both under the 0.8 gates, so no step passes.

Considering those ceilings, the two tracks say opposite things.

- **ρ** ranks colors by `redness`. Every condition scores below what a step would: the best ρ in the grid is 0.64 against a step ceiling of 0.75.
- **R^2** fits values against `sim^0.5` (proportionality). 4 of the six score above the ceiling, up to $R^2 = 0.78$ against 0.73.

Neither actually says whether the response is a grade or a step, so we visualize it below.



Alignment at op1, per color. α_c, the layer mean of $\cos(h, \hat{v}_{\text{hold}})$ averaged over seeds, against the redness of the color. Columns are the anneal endpoint, rows the hold ratio, as in the trajectory figure. One mark per color, drawn in that color; the heavy line is a 25-color sliding mean over the redness ordering, and the flat grey band is the same sliding mean for the un-anchored control. The dashed line is `sim^0.5` rescaled onto the response by least squares — the shape the R^2 track scores against, so the marks' distance from it is what that statistic measures. Per-panel statistics: both grading gates sit at 0.8, and a pure step response can reach at most ρ = 0.75 or $R^2 = 0.73$.

The response is not a step. It has a knee: flat from the control level up to a redness of about 0.4, then a clean rise that tracks the dashed target closely. Raising the hold ratio lowers the flat part toward the control without changing the rise, which is containment doing what it is meant to.

But most of the color cube lives in that flat part. 161 of the 216 colors fall below redness 0.4, and inside a flat band their alignments carry no ordering. A rank statistic over all 216 colors is therefore dominated by them.

condition	all 216	≥ 0.2 (108)	≥ 0.3 (81)	≥ 0.4 (55)	≥ 0.5 (33)
end50·hold03	0.607	0.744	0.766	0.861	0.826
end50·hold30	0.621	0.729	0.775	0.922	0.905
end70·hold03	0.574	0.718	0.744	0.868	0.823
end70·hold30	0.559	0.715	0.785	0.909	0.918
end90·hold03	0.640	0.770	0.800	0.901	0.881
end90·hold30	0.564	0.738	0.785	0.895	0.904

Spearman ρ against `redness`, computed over all 216 colors and then over the colors above each redness cut, with the number of colors remaining in each header. The gate is 0.8.

Above the knee the ordering is good in every condition — ρ runs 0.87 to 0.92 on the 55 colors with redness ≥ 0.4, against a gate of 0.8 — and it climbs monotonically as the flat colors are excluded. So the graded part of the response is graded, but the full-set statistic is dominated by the flat part.

Whether the figure shows *good* grading depends on the shape we ask for, and the two statistics ask for different shapes. ρ names no shape at all: it asks that all 216 colors be ordered by `redness`, with every pair weighted equally, so no part of the cube should sit flat. It reads the flat block as 161 failures of ordering, however clean the rise above the knee is.

R^2 names one shape. `sim^0.5` is itself below 0.05 for three quarters of the colors, so most of the cube should sit flat; where the response does rise, it should follow the similarity curve, which sits between linear and quadratic in the cosine similarity to *red*. Grading has no answer without a target. R^2 makes that choice explicit, and the knee in the figure is roughly the shape it asks for.

What a contained response can score

That raises a question about the gate. If containment collapses the low-redness colors together, and the grading statistic reads their ordering, then they must be in tension. How high could the statistic be, under those constraints?

We build the best response either statistic could hope for: the target shape exactly, above a knee, and contained to near zero below it. This is a property of the two statistics and the 216-color grid, not a measurement of any run.

knee	colors contained	best ρ	best R^2	$\hat{a}$
0	0	0.82	1.00	0.126
0.2	108	0.84	0.98	0.111
0.3	135	0.75	0.98	0.105
0.4	161	0.58	0.94	0.092
0.5	183	0.39	0.86	0.072

The best either grading track can score, for a response that reproduces `sim^0.5` exactly above the knee and sits at the control below it, with the $\hat{a}$ that response implies. A knee of 0 is no containment at all. Both grading gates are 0.8 and the containment gate is $\hat{a} \leq 0.1$. Expectations over the contained block's arbitrary ordering; a pure step reaches ρ = 0.75 and $R^2 = 0.73$ for comparison.

The two tracks come apart. Even with no containment at all, the best ρ a perfect `sim^0.5` response reaches is 0.82, barely over the gate, because `sim^0.5` is not a monotone function of `redness` (they correlate at 0.90, not 1). With the 161 colors below `redness` 0.4 contained, the ceiling falls to 0.58, under the observed values. So past a light knee the ρ track penalizes containment, and at the containment this grid achieves it cannot reach 0.8 however well the response is graded.

R^2 behaves the other way. Its target is near zero for the colors containment collapses, so containment moves the response toward the target rather than away: the ceiling is 0.94 at a knee of 0.4 and 0.86 even at 0.5. That track is reachable, and the best condition in this grid gets 0.78 of it.[7]

Within that family — `sim^0.5` at full amplitude — the ρ gate and the containment H2 also asks for cannot both be met. $\hat{a}$ crosses under 0.1 somewhere between a knee of 0.3 and 0.4, and by there ρ has fallen from 0.75 to 0.58.

That family is narrow, though. ρ ranks colors, so it is blind to the amplitude the $\hat{a}$ gate measures: scale the above-knee response to half of `sim^0.5` at a knee of 0.2, and both gates clear at once, at ρ = 0.81 and $\hat{a} = 0.055$. What the two gates jointly exclude is a response that is both steeply graded and full-amplitude on the reds, which is narrower than the gates being incompatible. Either way the ρ gate is doing something the preregistration did not say it was doing, and a future design should settle the amplitude question before reusing it.

The R^2 shortfall is in the response itself: against an achievable 0.94, 0.78 is a real gap. Setting a realistic R^2 gate for future work would take more than this one grid.

Discussion

The operating point to carry forward is `end90-hold30`: $\bar{\alpha} = 0.087$, margin 0.61, retention 0.97, task-clean. `end90-hold03` grades slightly better, but $\bar{\alpha}$ decides it: $\bar{\alpha}$ predicts how much of the color cube an edit along the anchor axis would drag with it, and the two conditions differ by 1.5× on it, against a 0.01 difference on the reachable grading track. Anything that pulls without a position oracle, and any intervention experiment, wants the smaller spread of side-effects.

The M1 keyframes were fractions of training, and those did not transfer. The trajectories suggest carrying something dimensionless instead: $\bar{\alpha}$ departs the control shortly after the repulsion falls below the pull, and the floor sets how fast it climbs after that crossing. If that mechanism holds elsewhere, the design variables are the crossing epoch and the ratio of the floor to the containment threshold. This grid measures one architecture on one task, so that is a reading to test; the reproduction check here shows the kind of evidence that would count.

Three things this experiment does not settle. Containment is not exclusivity: redness stays as readable off the anchor axis as it is in the control, down at the floor the RGB task itself sets. So an edit to the axis moves less of the cube, but the axis is not the only place *red* lives.

The grading gate needs a decision before the next report scores it: the ρ track cannot reach its gate alongside full-amplitude containment, and one grid is not enough to say where a realistic R^2 gate sits.

And the best floor here is also the highest floor tested, so whether a higher or flat-held repulsion does better, or starts to cost the margin, is outside the range this grid measured.

---

1. An anneal here is a schedule that ramps a loss weight down gradually over training. ↩

2. Reaching the endpoint-90 dose from endpoint 50 would need \(\\lambda_{\\bar{\\text{s}}}/\\lambda_\\text{a} \\approx 1.18\), which would hold the repulsion above the anchor weight for half of training. We could do that, but that's a different experiment. ↩

3. Ex-2.1.7 found no task cost anywhere, including at ten times this anchor weight, so this is a standing gate rather than an open question. We state it because a sustained repulsive floor is the one thing in this grid that has not been tried. ↩

4. Suppose the `hold03` row repeats its ex-2.1.7 span of 0.33 to 0.13. Then (a) can only fail if the mean of the 0.30 row is nearly as high, and with `end90-hold30` contained that in turn requires a wide endpoint spread inside the 0.30 row itself, of the same sign as the spread in the 0.03 row. Such a spread pulls the interaction toward zero, so this miss would take (b) down with (a). ↩

5. A main effect is the average change from moving one factor, averaged over the levels of the other. ↩

6. Start from the ex-2.1.6 control spread of about 0.02 for $\bar{\alpha}$. The hold-ratio effect compares nine-run means, so its noise is near $0.02\sqrt{2}/\sqrt{9} \approx 0.009$, and 0.05 is about five times that. The interaction is a difference of two differences of three-run condition means, with noise near $0.02 \cdot 2/\sqrt{3} \approx 0.023$, so 0.1 is about four times that. The endpoint effect within one row is a difference of three-run means, with noise near $0.02\sqrt{2}/\sqrt{3} \approx 0.016$, so 0.05 is about three times that where named miss 1 reads it. What the level account expects for the interaction is about 0.20, twice the gate: the full 0.33 to 0.13 span in the 0.03 row and none of it in the 0.30 row. Before reading either statistic, we compute the spread from the seeds of this experiment and check that the anchored conditions hold it. If they do not, the observed spread replaces the control spread and the gates move with it. They are the only thresholds in this report allowed to move. ↩

7. M1 scored proportionality with a Pearson R^2 as well, but of a different quantity: reconstruction error under intervention against $\cos^2$ similarity to *red* (D1.3), where the square is what the projection predicts. Here it is training-time alignment against `sim^1.5`, an exponent chosen as a shape between linear and quadratic rather than derived. Same statistic, different response variable and target, so M1's values are not a benchmark for this gate. ↩