Ex 2.1.11: Ablations and a survey for the D2.1 operating point

> A survey to close out D2.1. The schedules and weights in the recipe were inherited piece by piece and never tuned together. We ablate first by replacing schedules with constants, and dropping training epochs. Then we run a Sobol search over what is left (anchor weight, pooling temperature, repulsion dose).
>
> The anchor schedule and half the epochs could go. The anti-subspace schedule could not: a constant delivering the same total dose loses grading. The search maps a wide feasible plateau and proposes a stronger, softer point on its edge.

Observations

The reference recipe reproduces. `ref` re-runs the primary condition of ex-2.1.10 through the code path of this experiment. Every statistic is within one per-run standard deviation: m_line 0.4204 against 0.4202 (σ 0.0088), group contrast 0.8562 against 0.8542 (σ 0.0057), holdout exact match 0.9961 against 0.9948 (σ 0.0087).

The anchor schedule can go. `flat-anchor` stays within band on all five decision statistics, and no run latches (m_line +0.0039, band 0.0144). Rule 1: the survey runs a flat anchor weight.

The anti-subspace schedule cannot. Both flat brackets come out worse. `anti-hold` loses the group contrast (−0.4667, band 0.0093), grading (−0.2199, band 0.0392), and containment (+0.1275, band 0.0346), and two of its three runs latch onto `+`. `anti-peak` keeps containment and latches nothing, but it loses grading (−0.1283, band 0.0392), m_line (−0.0282, band 0.0144), and a little contrast (−0.0174, band 0.0093). Rule 2: the survey searches the schedule family, branch A.

Half the training is enough. `short` stays within band on every statistic (m_line +0.0014), and costs 0.0013 on the task against its own control. Rule 4: every survey trial runs 50 epochs.

Nothing resolves on the interpolation shape. `linear` stays within band throughout; its largest excursion is grading, at −0.0248 against a band of 0.0392. Rule 5: keep minimum-jerk, and the question retires.

What matters about the anti-subspace schedule is its shape, not its dose. `anti-mid` is a constant that matches the dose of the schedule, and it still loses grading (−0.1022, band 0.0392), m_line, and contrast. That is the pattern the shape reading of rule 6 predicted, so branch A is confirmed. The flat pairing `anti-mid-flat` fails the same way against `flat-anchor`, so flattening the anti term costs about the same with either anchor shape.

A quarter of the training is too little. `short25` fails on m_line (−0.0268, band 0.0144), containment, and contrast. The predicted failure, in which the un-anchored control loses the task, did not happen: the task gate holds, using 0.0026 of its 0.02. Rule 8: 50 epochs stands.

The survey maps the trade the calibration predicted. 38 of 64 trials are feasible. m_line rises with λ_a until the task gate binds (every task failure at λ_a ≥ 0.38), and with τ until grading and contrast reach their floors. The latch veto accounts for the cold edge, with all ten vetoes at τ ≤ 0.12. The anti-subspace axes organize little.

Proposed operating point. `t00` sets λ_a 0.557, τ 0.265, and an anti peak of 1.376 with anneal end 0.537. It reaches m_line 0.534 ± 0.004 over five seeds, against 0.420 ± 0.005 for the reference recipe. It is at the grading floor (r^2 0.690 against 0.682), with four more trials within one band spanning λ_a 0.28–0.94. D2.2 confirms at fresh seeds before any of these numbers is quoted.

> **How to read this report**
>
> This is a survey, not a hypothesis-scoring experiment. It preregistered a *search plan* and scores nothing. The plan covers the space to search, the sampling rule, the objective and its constraints, the seed budget, and the decision rules that stage the work. It was frozen before the runs, and amended once mid-experiment under rules that were themselves frozen before their arms ran (the amendment section records how). Results replaced the placeholders in place, so each analysis section reads as prediction then observation, and anything conceived after seeing the data appears under *Exploratory analyses*, marked as post hoc.
>
> Nothing this report finds may be quoted as a result. It proposes an operating point. The next preregistered experiment adopts that point, re-measures it at fresh seeds, and reports the survey value next to the confirmed one. The difference between the two is the winner's-curse correction: the best point in a search looks better than it really is,

> because part of what put it on top was luck.

Hyperparameters in D2.1 have history but little justification. The anchor schedule (warmup, hold, then anneal at the end down to a floor) and the anti-subspace schedule (2.5 → 0.3 of the anchor weight) were inherited from M1, mapped onto 100 epochs by fraction of training. Those keyframes were tuned for a 5-dimensional autoencoder bottleneck and are now applied to a 64-dimensional residual stream.[1]

The anchor weight λ_a = 0.1 is the rung ex-2.1.6 scored at, chosen to sit safely inside the region where the task is unharmed. Ex-2.1.7 showed that selectivity is gone by λ_a = 1, but didn't show where it begins to degrade.

τ = 0.1 is the one rung of the coarse ladders that stayed graded in every seed: the rungs below it latch per run (ex-2.1.9, ÷3 and ÷10), and the rung above it loses the group contrast and the grading in a single step (ex-2.1.10, ×5). So we don't know whether the primary is on a plateau or a knife edge.

We ablate first and search second, because deleting a dimension is cheaper than searching it. If a constant can replace a schedule, four parameters go away; if the testbed converges in 50 epochs, every later trial is half the cost. Before either, a calibration stage fixes the noise floors that decide what "no difference" means, and settles the shape of the objective.

Out of scope: the position-dropout mechanism raised in the ex-2.1.9 review, which is not part of the recipe yet; and the grammar, corpus, labeller, and architecture, which stay fixed at the ex-2.1.10 primary throughout.

▶ **Glossary...**

Method

The testbed is the ex-2.1.10 primary, unchanged: the word-level `v216` corpus (216 colors, one token each, d64-L4), the either-slot labeller, the mellowmax-pooled anchor term over the four span roles, and the anti-subspace term. Under the either-slot labeller, each operand draws with probability $\text{redness}^8 \cdot 0.04$, and a line is labeled if either operand draws.

The reference recipe is the operating point of that experiment: $\lambda_a = 0.1$, $\tau = 0.1$, anchor schedule with warmup over 0–10% and anneal over 90–100% down to a 0.1 floor, and anti-subspace 2.5 → 0.30 ($\lambda_{\bar{s}}/\lambda_a$) by 90%. Schedule keyframes are given as fractions of training, so the half-length condition runs the same shape.

Measurements

Every run, ablation or trial, records the same slim set of measurements, all of them defined in earlier experiments: the three behavior sets (holdout exact match is the task gate), the alignment map on the probe lines (m_line, $\bar{\alpha}$ at op1, r^2 against the $\text{sim}^{1.5}$ target), the softmin weight profiles (the group contrast and the latch detector), and the training trajectory of m_line (retention).

That costs one eval pass per run. We do not run the readability, geometry, and leakage passes that a scoring experiment carries. Those belong to the confirming experiment, which measures one point properly rather than eighty points thinly.

Search plan

Calibration has already run; it reads only published results, and its numbers are frozen into the design constants. The ablation stage runs nine conditions and applies the decision rules below, which fix the recipe, the length, and the space for the survey. The survey stage then draws its trials from a rule that is deterministic given this document.

The decision rules are the only branch points; everything downstream of the data is fixed.

Calibration: noise floors

Ex-2.1.10 ran nine seeds on its primary, which is the reference recipe here, so the per-run spread of every statistic at the operating point is already in the store. The table below re-derives the floors from the published arrays. The design constants freeze the same numbers (`NOISE_RUN`), and the ablation bands follow as $2\sigma \cdot \sqrt{(2/3)}$ for a three-seed against three-seed comparison.

The margin statistics are tight compared to the effects worth chasing; the τ ladder in ex-2.1.10 spans several standard deviations on m_line. That is what lets us run survey trials at a single seed. Containment is the opposite, with a standard deviation about 40% of its own mean, so it enters the survey as a constraint rather than something we rank on.

These floors carry a caveat: they are measured in the graded regime, at $\tau = 0.1$. At $\tau \lesssim 0.03$ the pull latches from run to run, the statistics become bimodal, and no per-run standard deviation describes them. So the survey keeps τ above that regime and uses a per-run latch veto instead.

statistic	mean	per-run σ	frozen σ	band (3 v 3)
`m_line`	+0.4202	0.0088	0.0088	0.0144
`alpha_op1`	+0.0505	0.0212	0.0212	0.0346
`retention`	+0.9882	0.0061	0.0061	0.0100
`r2_sim`	+0.7856	0.0236	0.0240	0.0392
`contrast`	+0.8542	0.0057	0.0057	0.0093
`holdout_em`	+0.9948	0.0087	0.0087	0.0142

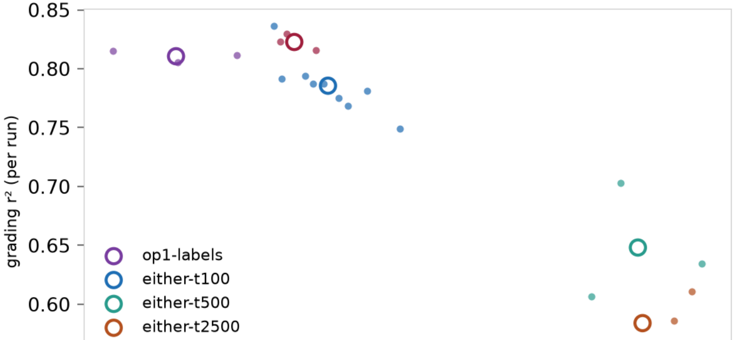
Noise floors at the operating point. *Per-run mean and sd over ex-2.1.10's nine-seed primary, re-derived from its published arrays; the frozen σ column is the value the design constants carry, and the band is the smallest seed-mean difference the ablation stage may call a difference ($2\sigma\cdot\sqrt{(2/3)}$, three seeds a side).*

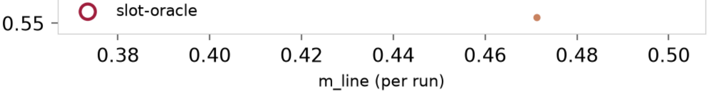
Calibration: the shape of the objective

Across the anchored operating points already in the store, margin and containment don't trade: their correlation is negative on both available readings (computed live below the figure), meaning the better operating points improved both at once. The trade that does bind is margin against grading, and it runs along the τ axis: softening τ raises m_line while the per-color grading and the group contrast collapse, so selectivity smears out into "everything red-ish moves".

That structure decides the shape of the search. A two-way trade would need a Pareto front,[2] but objectives that move together can be constraints around a single scalar. So the objective is one number: maximize m_line, subject to the task gate, containment, retention, grading, contrast, and the latch veto, with the values under *Objective, constraints, and rounds* below.

A search on margin alone would walk to the smeared end of the τ range, and the grading and contrast constraints are what keep it out. We still report the m_line against r^2 plane for every trial, so if the interesting points turn out to sit on the constraint boundary, the map will show it.





The trade the constraints must hold: margin buys smear at soft τ. *Per-run grading r^2 against per-run m_line for ex-2.1.10's five anchored conditions (● = one run, ○ = the condition's seed mean). Softening τ (0.1 → 0.5 → 2.5, left to right) buys margin and spends grading; ranking on m_line alone would pick the right-hand end. The survey caps that walk with the grading and contrast constraints instead of a second objective.*

Margin–containment correlation across the stored anchored operating points: r = -0.19 over the 10 ex-2.1.9/ex-2.1.10 points (m_span basis), and r = -0.82 over all 17 points including the ex-2.1.8 grid (m_op1 basis, the statistic all three share). No trade on either reading; both improve together along the design line. For scale, the primary's nine runs put per-run grading at 0.79 ± 0.024 against 0.58 for the softest arm.

Calibration: the τ range

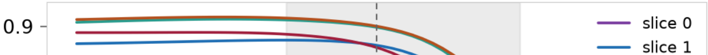
τ only matters through the weights it produces, so we set its bounds on the weight scale: the share of the pooled pull held by the leading span position, computed from the stored alignment profiles of the primary.

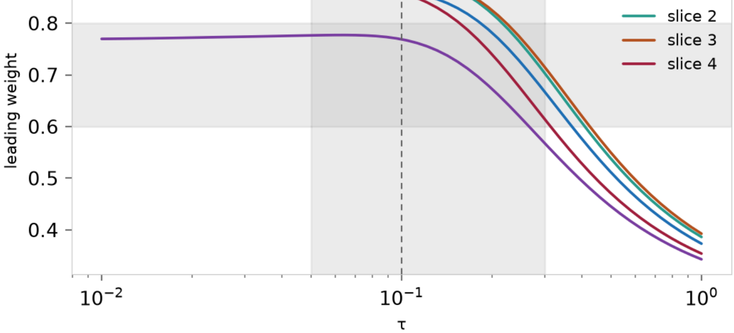
The bounds [0.05, 0.30] carry that leading weight from 0.78 down to

0.57 at the embedding slice, and from 0.90 to 0.67 averaged over the deeper ones. So the survey range spans the target band from ex-2.1.9 – 0.6 to 0.8 – near enough end to end at the embedding slice, which is where that band was set.

Below τ ≈ 0.05 the curve is flat: it saturates at 0.78 rather than climbing to 1, because the profile the pooling is applied to has a spread that a colder temperature runs out of room to sharpen. Nothing is given up by holding the lower bound above the latching regime. The upper bound stops short of τ = 0.5, which the trade figure above shows is already smeared.

One thing this curve cannot do is predict a run trained at a different τ. It applies a hypothetical temperature to the profile of a model trained at 0.1, so it says what the pooling would weight given that model, and the latching at τ ≲ 0.03 is ex-2.1.9's observation from runs rather than anything visible here.





Where the survey's τ bounds sit on the weight scale. *The label-affinity-weighted share of softmin weight held by the leading span role, per residual slice, on the primary's seed-mean alignment profile — ex-2.1.9's π, which is also what the latch veto reads. The lowest curve is the embedding slice, where ex-2.1.9 set its band; it passes through τ = 0.1 at 0.77, reproducing that experiment's published value. Vertical shading: the survey range [0.05, 0.30]; horizontal shading: ex-2.1.9's 0.6–0.8 target band; dashed line: the reference τ = 0.1. The flat left-hand end is the profile's own spread, which a colder temperature runs out of room to sharpen.*

Calibration: doses and the flat arms

One flat arm is enough to ablate the anchor schedule, because flat at the peak and flat at matched dose are the same thing here. The schedule delivers 97% of the flat-at-peak dose, and that 3% difference is well inside the noise floor, so we don't have to pick which quantity to match on.

What the flat arm does change is timing. The ramp carries 7.3% of the dose across the same window in which ex-2.1.9 found the pooled pull commits, by epoch 8, inside the 10-epoch warmup. The `AnchorSpec` docstring says the ramp is there so the anchor arrives with the optimizer rather than ahead of it.

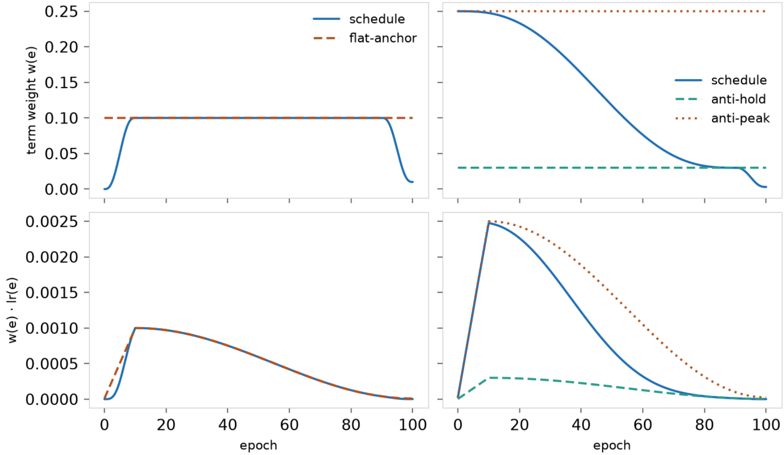
So we don't expect a dose effect from `flat-anchor`. Full λ_a against a still-random model risks latching, so the reading will come from the latch veto and the deep-slice weight profile.

No constant can match the anti-subspace term on dose. Its schedule spans an 83× range of weights, and the constant that matched its dose would sit at 1.7× the anchor peak for the whole of training, which is a different regime; the dose arm in ex-2.1.8 documented that trap.

So two arms bracket it instead: flat at the hold ratio (0.17× the dose of the schedule) and flat at the peak ratio (1.41×). We gate the effect of the schedule on the peak-matched arm, and read the dose alignment as a stated secondary. The conditions table below reports the delivered dose and centroid for every arm.

anchor term

anti-subspace term



What each arm delivers, when. Top row: *the term weights, λ_a(e) (left) and λ_s̄(e) (right). The flat arms really do hold a constant weight.* ***Bottom row:*** *the effective weights, w(e) · lr(e), which is what the dose integrals measure. The rise and fall shared by every curve is the LR schedule.* ***Left:*** *the anchor schedule and `flat-anchor` differ visibly only during warmup, and their doses agree to 3%.* ***Right:*** *`anti-hold` and `anti-peak` bracket the schedule from below (0.17× its dose) and above (1.41×). No constant matches its dose while staying in the regime.*

run (cond-seed)	op1	+	op2	=	max syntax ↓
`pool-t010-s0`	0.14	0.85	0.00	0.00	**0.85**
`pool-t010-s1`	0.13	0.87	0.00	0.00	**0.87**
`pool-t010-s2`	0.98	0.01	0.01	0.00	0.01
`pool-t030-s0`	0.03	0.97	0.00	0.00	**0.97**
`pool-t030-s1`	0.97	0.02	0.00	0.00	0.02
`pool-t030-s2`	0.97	0.02	0.01	0.00	0.02

`pool-t100-s0`	0.93	0.02	0.02	0.04		0.04
`pool-t100-s1`	0.93	0.03	0.01	0.02		0.03
`pool-t100-s2`	0.93	0.03	0.02	0.02		0.03

The latch detector, calibrated on ex-2.1.9. *Deep-slice (post-attention) mean softmin weight per span role, per stored run of the pooled conditions (with seed shown as a suffix). Latched runs put ≥ 0.85 on* `+`*; healthy runs ≤ 0.03 on any syntax role (bold = over the 0.5 veto). The threshold falls in a wide empty range, so the veto is insensitive to where in it we put the bar.*

Ablation conditions

There are nine conditions, three seeds each, all at the reference λ_a and τ under the either-slot labeller. `ref` re-runs the primary recipe from ex-2.1.10 through the code path of this experiment, so every comparison is between runs of identical code. `lam0` and `short-lam0` are the task-cost controls, one at each length. `flat-both` combines the two flat simplifications. We need it because an arm that changes one schedule at a time cannot show an interaction between the two; we read it only if decision rules 1 and 2 both simplify. The table reports dose and centroid for each arm, so the outcomes can tell us which of the two was doing the work.

Four more conditions were added once these had settled, under rules frozen in turn before they ran; they are set out under *An amendment to the search plan* below, with the reasoning that motivated them.

condition	one change	epochs	anchor dose	centroid	anti dose	centroid
`lam0`	un-anchored control	100	0.0000	–	0.0000	–
`ref`	the ex-2.1.10 primary recipe, re-run	100	0.0490	34.9	0.0896	26.5
`flat-anchor`	λ_a constant from step 0; no ramp, anneal or floor	100	0.0505	34.1	0.0896	26.5
`anti-hold`	λ_s̄/λ_a constant at the hold ratio (0.30)	100	0.0490	34.9	0.0151	34.1
`anti-peak`	λ_s̄/λ_a constant at the peak ratio (2.5)	100	0.0490	34.9	0.1262	34.1
`flat-both`	flat anchor and flat anti (hold ratio) together	100	0.0505	34.1	0.0151	34.1
`linear`	every minimum-jerk ramp/anneal made linear	100	0.0487	35.0	0.0842	28.1
`short`	the whole recipe at 50 epochs, keyframes scaled	50	0.0245	17.4	0.0448	13.3
`short-lam0`	50-epoch task-cost control	50	0.0000	–	0.0000	–

The ablation conditions and their delivered invariants. *Dose is* $\int w \cdot \mathrm{lr}\,\mathrm{d}e$ *over that condition's training; centroid is the dose-weighted mean epoch. Doses are reported, not matched.*

Decision rules

Two names recur from here on: **branch A** keeps the anti-subspace term's schedule shape, sampling its peak and anneal endpoint; **branch B** replaces it with one flat ratio. Rule 2 below picks between them, and the next section gives each branch's own dimensions.

Rendered from the design constants (`experiment.DECISION_RULES`), so the frozen text has one home:

Frozen before any ablation run. "Within band" means: for every statistic in DECISION_STATS, |seed-mean difference from `ref`| < equiv_band(stat); and the task gate holds (holdout EM within TASK_GATE of the same-length control); and no run latches a syntax role (LATCH_PI). Any resolved regression keeps the incumbent recipe; ties break toward the simpler recipe only when nothing resolves.

Resolution caveat, known going in: five statistics read at ±2 sd across six comparison arms makes a spurious "resolved" call somewhere likely even under no true effect. Every such error lands on the conservative side, keeping the incumbent recipe or the longer run, so the stage is biased toward keeping complexity, never toward adopting a simplification the data doesn't support. Read a lone marginal excursion with that in mind.

1. `flat-anchor` within band, or better on m_line → the survey runs the flat anchor (four schedule parameters deleted). Otherwise it keeps the schedule, also deleting the parameters (kept fixed by inheritance, now with evidence the shape earns its place). Predicted failure mode, from the ex-2.1.9 commit window: full λ_a on a still-random model latches a syntax role, so the latch veto, and which role leads, is the reading that matters.
2. `anti-hold` or `anti-peak` within band or better → the anti schedule collapses to a flat ratio, and the survey searches

SPACE_ANTI_FLAT (branch B; prefer the passing level nearer `anti-hold`, the lower dose). Both resolved worse → the shape matters; the survey searches the schedule family, SPACE_ANTI_SCHED (branch A).

3. Rules 1 and 2 are one-factor readings, so their simplifications are adopted together only if `flat-both` is also within band. If each flat arm passes alone and `flat-both` fails, the schedules interact: keep the anchor schedule and adopt the flat anti, which preserves branch B and costs the survey nothing (rule 1 fixes the anchor parameters either way). `flat-both` is read only when rules 1 and 2 both simplify.
4. `short` within band (against `ref`, task-gated against `short-lam0`) → the survey runs at EPOCHS_SHORT, halving every trial. Otherwise 100.
5. `linear` within band → shapes are interchangeable; keep minimum-jerk everywhere and stop carrying the question. Resolved worse → keep minimum-jerk, note the shape as load-bearing. Either way no survey dimension: this arm only retires a todo item.

Survey space

Two dimensions are always sampled, both on a log scale. λ_a spans [0.02, 1.0]: from a fifth of the rung we scored at, below which the ladder in ex-2.1.6 suggests the pull is weak, up to the weight of the ex-2.1.7 ceiling arm, where selectivity was already gone. So we believe the optimum is somewhere inside. τ spans [0.05, 0.30], on the weight-scale argument above.

The anti-subspace dimension depends on decision rule 2. If a flat ratio is enough (branch B), it is one dimension: λ_s̄/λ_a ∈ [0.10, 3.0], log, running from a third of the hold level up past the peak. If the shape of the schedule matters (branch A), the family keeps that shape with the hold ratio fixed at 0.30, and we sample two dimensions: peak ratio ∈ [0.75, 4.0] (log) and anneal endpoint ∈ [30%, 95%] of training.

For branch A the analysis reports marginals in derived coordinates, delivered dose and timing centroid, rather than in the sampled parameters. Those parameters form a correlated ridge, since the endpoint is effectively the dose axis (ex-2.1.8). Sampling still happens in the parameter box, because every point in it is a valid schedule, whereas a box in (dose, centroid) has corners we can't reach.

Trials come from a scrambled Sobol set, which gives even one-dimensional marginals at this budget with no two trials aligned on any axis. The scramble seed is frozen in the design constants, so the run rebuilds the trial list below identically.

We don't use Bayesian optimization: it returns a single best point where D2.2 needs a map, and given the noise in the containment constraint it would chase seed luck.

▶ **Branch A trial list (schedule retained)...**

▶ **Branch B trial list (flat anti)...**

***The frozen trial lists**, scrambled-Sobol points at seed 0, shown for review before any run: 64 for branch A's four dimensions, 32 for branch B's three. Branch A adds the derived (relative dose, centroid) coordinates its marginals will be reported in; relative dose is the trial's* $\int(\lambda_{\bar{s}}/\lambda_a)\cdot\mathrm{lr}\,\mathrm{d}e$ *over the operating point's.*

Objective, constraints, and rounds

Objective (from `experiment.OBJECTIVE`): rank feasible trials by seed-mean m_line, largest first. m_line is the one statistic tight enough to rank on, so everything else is a constraint:

constraint	bar	source
task cost	holdout EM within 0.02 of the same-length control	ex-2.1.6 on
containment	$\bar{\alpha}$ at op1 ≤ 0.1	ex-2.1.8 on
retention	final m_line ≥ 0.8 × running peak (floor 0.2)	ex-2.1.8 on
grading	per-run r^2 within 0.1 of `ref`'s seed mean	ex-2.1.9 on, relative form
contrast	group contrast ≥ 0.1	ex-2.1.10's partial bar

latch	no run with deep-slice syntax weight > 0.5	calibrated above

Rounds. Round 1 runs every trial at seed 0, which the noise floors allow. The 6 feasible trials with the highest m_line are promoted to 5 seeds. If fewer are feasible, we promote all of them; if none are, we publish the infeasibility map and propose nothing.

The proposed operating point is the promoted trial with the highest seed-mean m_line that is still feasible on seed means; the latch veto stays per-run. Promotion only partly corrects the winner's curse.

Stopping rule. Fixed budget, two rounds, no adaptive continuation. Trials that diverge or fail are published as failures rather than resampled. Memoization keeps them anyway, and a complete table makes the map from a search trustworthy.

Budget

9 ablation conditions × 3 seeds = 27 runs, plus the amendment's 4 × 3 = 12, then at most 64 + 6 × 4 = 88 survey runs: about 127 training runs in the worst case.

The design budgeted roughly 25 minutes of L4 per run, inherited from ex-2.1.10, which carried a much heavier eval.

Each stage launches with `--max-containers 8` and a `--budget` sized from its run count. Memoization means the promoted round reruns only the new seeds, and `ctx.map` with `allow_partial=True` lets diverged trials land as data rather than failures.

The ablation stage's runs took 0.7 to 3.6 minutes of L4 each, training and eval together, because the slim eval drops the readability, geometry and leakage passes and the corpus is prepared once for the whole DAG; all 27 settled in about 25 minutes of wall clock at eight containers. The full DAG — all 127 training runs plus the prep, eval and publish steps — billed about $3.60 of compute, nearly all of it L4, in 2.8 hours of wall clock.

Ablations

Nine conditions, three seeds each. Each subsection reads one arm against `ref` on the five decision statistics, and the rule it feeds was fixed before the run.

condition	seeds	m_line ↑	alpha_op1 ↓	retention ↑	r2_sim ↑	contrast ↑	holdout EM ↑	latched ↓
`lam0`	3	+0.0239	-0.0415	+0.5717	+0.0011	-0.0364	1.0000	1
`ref`	3	+0.4204	+0.0390	+0.9897	+0.7823	+0.8562	0.9961	0
`flat-anchor`	3	+0.4243	+0.0505	+0.9975	+0.8184	+0.8573	0.9974	0
`anti-hold`	3	+0.3025	+0.1665	+0.9986	+0.5625	+0.3895	0.9987	2
`anti-peak`	3	+0.3922	+0.0105	+0.9915	+0.6541	+0.8388	0.9922	0
`flat-both`	3	+0.3391	+0.1851	+0.9108	+0.7004	+0.3552	0.9987	0
`linear`	3	+0.4098	+0.0515	+0.9923	+0.7575	+0.8580	0.9935	0
`short`	3	+0.4218	+0.0262	+0.9968	+0.8065	+0.8491	1.0000	0
`short-lam0`	3	+0.0260	+0.0304	+0.4094	+0.0311	-0.0373	0.9987	1
`anti-mid`	3	+0.4048	+0.0114	+0.9885	+0.6802	+0.8427	0.9948	0
`anti-mid-flat`	3	+0.4040	+0.0165	+0.9921	+0.6859	+0.8465	0.9974	0
`short25`	3	+0.3936	+0.0800	+0.9938	+0.7875	+0.8340	0.9935	0
`short25-lam0`	3	+0.0281	-0.0110	+0.5336	+0.0647	-0.0218	0.9961	0

Every arm, every decision statistic. *Seed means over three runs per condition. Arrows give the direction that counts as better; containment (`alpha_op1`) is the one that improves downward. The subsections below take the differences from `ref` and read each against its band.*

Is the anchor schedule needed?

No. `flat-anchor` holds λ_a at 0.1 from step 0, with no ramp, no end anneal, and no floor. Every decision statistic is inside its band, and no run latches. The largest excursion is grading, +0.0361 against a band of 0.0392, and it falls on the better side, so at three seeds a side this says the schedule is not doing anything we can resolve rather than that it does nothing. Rule 1 fires: the survey runs a flat anchor weight.

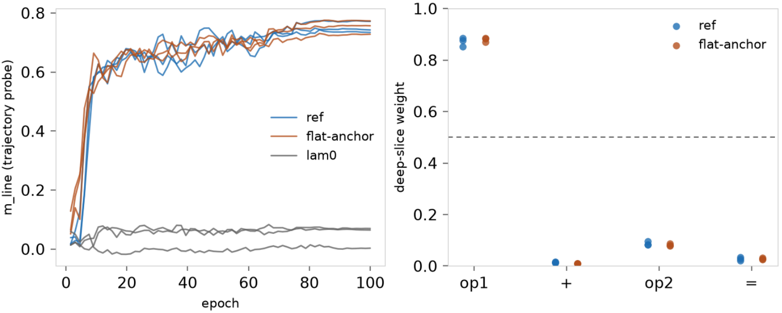
We had predicted the other outcome. Ex-2.1.9 found that the pooled pull commits by epoch 8, inside the 10-epoch warmup, and the `AnchorSpec` docstring says the ramp exists so that the anchor arrives with the optimizer rather than ahead of it. Applying full weight to a still-random model looked like a way to latch a syntax role. It isn't: all six runs put 0.85–0.89 of their deep-slice weight on op1, and the flat arm reaches its margin sooner and then holds it (retention 0.9975).

The end anneal is the half we were least willing to lose, since M1/ex-2.9.3 found that withdrawing protection entirely lets the task loss reclaim the axis. Flattening does not withdraw it, because the weight stays at full strength through the last step. So that lesson is untouched here.

statistic	**ref**	**band**	**flat-anchor**
`m_line` ↑	+0.4204	0.0144	+0.0039

`alpha_op1` ↓	+0.0390	0.0346	+0.0115
`retention` ↑	+0.9897	0.0100	+0.0078
`r2_sim` ↑	+0.7823	0.0392	+0.0361
`contrast` ↑	+0.8562	0.0093	+0.0010
holdout EM − control ↓	0.0039	0.0200	0.0026
latched runs ↓	0	–	0

`flat-anchor` against `ref`. *Seed-mean differences, three runs a side. Bold marks a difference the band can resolve; ✓ and ✗ say which side of `ref` it landed on. The band is 2σ·√(2/3) from the nine-seed noise floors of ex-2.1.10, and the two gate rows carry absolute bars rather than bands.*



What the ramp buys. Left: *the m_line trajectory of every run, each on its own epoch axis. It sits above the endpoint value in the tables: the trajectory measures one fixed partner line per color, where the endpoint eval averages all 27. That makes it a cheaper instrument, but a consistent one within a run, which is what retention compares.* ***Right:*** *the post-attention mean softmin weight on each span role, one ● per run. The dashed line is the latch veto at* $\pi = 0.5$*; a run above it on `+` or `=` has committed to a syntax role. The* *reference and flat arms differ only over the warmup window, so a difference here says when the pull arrives rather than how much of it arrives.*

Does the anti-subspace schedule beat a constant?

It comes out ahead of both constants.

`anti-hold` holds the low ratio for all of training, which removes the early repulsion (the schedule's min-jerk anneal is slow to start, so its ratio still sits at 2.47 of an eventual 0.30 at epoch 10). The pull then goes where ex-2.1.8 said it would: two of three runs put more than half their deep-slice weight on `+` and trip the latch veto, and the third sits at 0.482, just under. Containment, grading, and the group contrast all come out worse, the contrast by fifty times its band.

`anti-peak` holds the high ratio instead: op1 keeps 0.86–0.88 of the deep-slice weight, and the contrast holds at 0.839, which is 0.0174 under `ref`: a small drop, but nearly twice the band, so it resolves. The cost is grading, with r^2 falling from 0.78 to 0.65 and m_line falling with it. Repulsion that never comes down leaves the color response less graded, which is the failure the constraints exist to catch.

So both ends of the schedule seem to be required: the peak keeps the pull off the syntax roles while the model is still random, and the anneal down to the hold ratio leaves the response graded once it has committed. That is one reading — both brackets change the dose along with the shape, and the amendment below adds the arm that separates the two. On what ran here, rule 2 sends the survey to branch A, the schedule family, with the hold ratio fixed and the peak and the anneal endpoint sampled.

statistic	ref	band	anti-hold	anti-peak
`m_line` ↑	+0.4204	0.0144	**-0.1179 ✗**	**-0.0282 ✗**
`alpha_op1` ↓	+0.0390	0.0346	**+0.1275 ✗**	-0.0285
`retention` ↑	+0.9897	0.0100	+0.0089	+0.0018
`r2_sim` ↑	+0.7823	0.0392	**-0.2199 ✗**	**-0.1283 ✗**
`contrast` ↑	+0.8562	0.0093	**-0.4667 ✗**	**-0.0174 ✗**
holdout EM − control ↓	0.0039	0.0200	0.0013	0.0078
latched runs ↓	0	–	**2**	0

The two flat brackets against `ref`. *Read as the table above. The arms bracket the dose of the schedule rather than matching it: `anti-hold` delivers 0.17× and `anti-peak` 1.41× of what the schedule delivers. So an arm that holds every statistic while delivering a different dose tells us the dose was not the active ingredient.*

Do the flat arms compose?

`flat-both` fails on all five statistics. Rule 3 never reads it: the rule fires only when rules 1 and 2 both simplify, and rule 2 did not, so nothing here changes the survey either way. The arm was insurance against an outcome that did not happen, and it cost three runs.

Two runs put their deep-slice weight on `=` and op2, and the third on `+`. So the failure is not stable across seeds — and it is not quite `anti-hold`'s failure either: that arm sent all three of its runs to `+` and kept retention within band, where this one regresses it. What survives the comparison is that the hold-level anti term fails with either anchor shape; whether the anchor's shape also steers where the pull settles is a three-seed observation with nothing riding on it.

Rule 1 (flat anchor) and rule 2 (branch A, scheduled anti) together give a pairing that already ran: `flat-anchor` only flattened the anchor term, leaving the anti-subspace term on its schedule, so the survey's recipe is that tested condition, not a combination assembled after the fact from separate arms.

statistic	ref	band	flat-both
`m_line` ↑	+0.4204	0.0144	**-0.0813 ✗**
`alpha_op1` ↓	+0.0390	0.0346	**+0.1461 ✗**
`retention` ↑	+0.9897	0.0100	**-0.0789 ✗**
`r2_sim` ↑	+0.7823	0.0392	**-0.0819 ✗**
`contrast` ↑	+0.8562	0.0093	**-0.5010 ✗**
holdout EM − control ↓	0.0039	0.0200	0.0013
latched runs ↓	0	–	0

The composition against* `ref`*. *Read as the tables above. Rule 3 does not read this arm: rules 1 and 2 did not both simplify, so there is no composition to adopt; the numbers are shown either way, since a* *survey that publishes only the arms it acted on is not a complete map.*

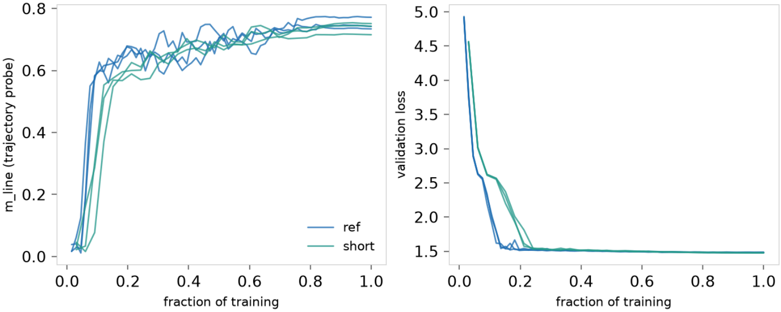
Is 100 epochs buying anything?

Nothing we can measure. `short` runs the whole recipe in 50 epochs with every keyframe scaled, and stays within band on all five statistics (m_line +0.0014 against a band of 0.0144), at a task cost of 0.0013 against its own control. On the shared clock below, the two lengths trace the same curves.

Ex-2.1.6 suggested as much when its margins went flat from about epoch 50. That was read off runs scored for something else; here it is measured directly. Rule 4: the survey runs at 50 epochs, halving its cost.

statistic	ref	band	short
`m_line` ↑	+0.4204	0.0144	+0.0014
`alpha_op1` ↓	+0.0390	0.0346	-0.0127
`retention` ↑	+0.9897	0.0100	+0.0071
`r2_sim` ↑	+0.7823	0.0392	+0.0241
`contrast` ↑	+0.8562	0.0093	-0.0072
holdout EM − control ↓	0.0039	0.0200	0.0013
latched runs ↓	0	–	0

`short` against `ref`. *Read as the tables above, with one difference: the task gate compares* `short` *against* `short-lam0`*, the un-anchored control at the same length, so that whatever* *the shorter run costs the task is not charged to the anchor.*



A hundred epochs against fifty, on a shared clock. *Both panels put fraction of training on the horizontal axis, so the two lengths run the same schedule shape and land at the same place on it.* ***Left:*** *m_line, measured with the cheaper one-partner-per-color probe, which is why it sits above the endpoint values in the tables.* ***Right:*** *validation loss. If the second half of training consolidates anything, it would show as the hundred-epoch curves pulling away over the right half of the left panel.*

Does the anneal shape matter?

No. Straightening every ramp and anneal in both terms leaves all five statistics within band, the largest excursion being grading at −0.0248 against a band of 0.0392. Minimum-jerk stays, now on evidence rather than by inheritance from M1, and the question retires.

Rule 5 was written to keep minimum-jerk whichever way this arm resolved. Unlike the flat-arm rules, swapping the interpolation shape deletes no survey dimension, so a pass here only closes the question of whether the shape matters — it is not a case for adopting the simpler curve.

A straight line and a smoothstep both cover half their window, so the anchor dose barely moves. The anti-subspace ratio interpolates over 90 epochs, and over that span the two shapes differ enough to change its dose by 6%. No statistic resolved, so near the operating point the repulsion is insensitive to dose changes of that size — worth holding beside the previous section, where what the flat brackets changed was mostly something other than dose.

statistic	ref	band	linear
`m_line` ↑	+0.4204	0.0144	-0.0106
`alpha_op1` ↓	+0.0390	0.0346	+0.0125
`retention` ↑	+0.9897	0.0100	+0.0026
`r2_sim` ↑	+0.7823	0.0392	-0.0248
`contrast` ↑	+0.8562	0.0093	+0.0018
holdout EM − control ↓	0.0039	0.0200	0.0065
latched runs ↓	0	–	0

`linear` against `ref`. *Read as the tables above. Straightening the ramps moves the anchor dose by under a percent, since a smoothstep and a straight line both cover half their window. It moves the anti-subspace dose by 6%, because that term interpolates over 90 epochs, and the two shapes differ* *most over the early part of that window, where the learning rate is highest.*

An amendment to the search plan

Everything above ran under frozen rules. This section adds four conditions and three more rules, also frozen before any of *them* ran. Rules 1–5 stand exactly as they fired; the amendment extends the plan where round 1 turned out not to answer the question we had asked it.

What round 1 could not settle. Rule 2 read two constants, at 0.3 and 2.5, and both resolved worse than `ref`, which we read as the schedule earning its shape at both ends. But those two arms differ from `ref` in two ways at once: in shape, and in dose (0.17× and 1.41× the reference).

Alternatively, total dose may have an interior optimum with the shape incidental. The seed means fit that too. Containment moves monotonically with dose across the three arms, and `ref` sits between the two brackets while beating both on the other three statistics.

The design anticipated this middle level and set it aside. The frozen note under `ABLATION_CONDITIONS` says the anti term "cannot be dose-matched flat (the match would sit at 1.7× the anchor peak for all of training, a different regime)".

That was a reasonable call with no data yet, but after round 1 this level separates two live readings rather than adding one more point in a range.

The four conditions. `anti-mid` holds λ_s̄/λ_a constant at 1.77, the level that matches the reference dose, so it differs from `ref` in shape alone. `anti-mid-flat` pairs that constant with the flat anchor. Since `ref` and `flat-anchor` have already run, the four cells complete a 2 × 2 in (anchor shape) × (anti shape), which is what it takes to read an interaction.

`short25` halves the adopted length again, and `short25-lam0` is its task-cost control, since un-anchored holdout EM is itself a function of length.

Why this does not spend the preregistration. A survey scores nothing. Its output is a proposal, which the next preregistered experiment confirms at fresh seeds, so there is no verdict here for a late change to manufacture. The risks specific to this experiment type are selective reporting and quoting the numbers from a survey as results, and adding arms touches neither, as long as every arm is published.

What protects the rest is ordinary freezing, done a second time: the rules below were fixed before the runs, they name their predictions, and rules 1–5 keep their own record. `decide` reads round 1 alone whatever else has landed, so the amendment shows up as a diff rather than an overwrite.

One cost, uncorrected. Rule 6 is a third constant tested against a rule that simplifies on a tie, so it is a third chance to simplify by luck. We are leaning on the survey contract rather than on the risk being small: a spurious pass buys a slightly wrong search space, and D2.2 re-measures the operating point at fresh seeds. Branch A is the more expensive branch, so a spurious pass would displace the incumbent that costs *more* to search.

condition	what it changes	epochs	anti dose vs `ref`	centroid
`anti-mid`	λ_s̄/λ_a constant at the dose-matched level (1.77)	100	1.00	0.34
`anti-mid-flat`	the same constant, paired with the flat anchor	100	1.00	0.34
`short25`	the whole recipe in a quarter of the epochs	25	0.25	0.26
`short25-lam0`	un-anchored control at 25 epochs	25	–	–

The amendment's conditions*, three seeds each. Dose is* $\int \lambda_{\bar{s}} \cdot \mathrm{lr}\,\mathrm{d}e$ *divided by the same integral for the reference schedule, at the same length; centroid is where that dose is delivered, as a fraction of training.* `anti-mid` *and* `anti-mid-flat` *match the reference dose by construction, and differ from it in when the repulsion arrives: the reference delivers it at a centroid of 0.27, a constant at 0.34. That is the contrast rule 6 turns on. The two short arms carry the reference shape unchanged, so their doses scale with length rather than moving independently.*

The rules (from `experiment.AMENDMENT_RULES`, frozen before the runs).

6. `anti-mid` against `ref`. Within band, or better → the dose is the active ingredient, the anti schedule collapses to a flat ratio, and the survey searches branch B with 1.77 as its reference level. Resolved worse → the shape is load-bearing, and branch A is confirmed rather than merely un-overturned. This comparison holds the anchor schedule fixed at `ref`'s, so a regression here isolates the anti-subspace shape from any interaction with the anchor's shape; rule 7 tests that interaction separately, and confirming branch A does not wait on it.

7. `anti-mid-flat` against `flat-anchor`. Rules 1 and 6 are one-factor readings, so both simplifications are adopted together only if the pair holds. This is the interaction gate of rule 3, now at a level that might pass it. If rule 6 simplifies and this arm does not, the two simplifications interact the way rule 3 describes: we keep the anchor schedule, since rule 1's parameters are fixed either way, and adopt the flat anti at the level rule 6 already showed works alone.

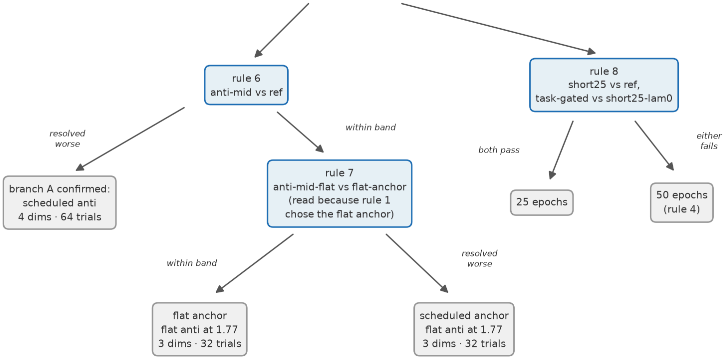
8. `short25` within band against `ref`, task-gated against `short25-lam0` → the survey runs at 25 epochs. Otherwise the 50 of rule 4 stands.

What each reading predicts. A constant at 1.77 delivers about 0.7× the repulsion of the schedule over the window where the latch is decided, and four to six times its late repulsion over the window where grading is learned. So the shape reading predicts m_line and grading below band with containment intact, a weakened blend of both round-1 failures. The dose reading predicts every statistic within band.

Grading is the statistic to watch: it is what the M1 schedules bought, and it is the one both round-1 brackets lost.

`flat-both` already speaks to half of this. It holds the anchor at full strength from step 0 and still fails the way `anti-hold` does, so the flat anti fails at the low end for some reason other than the anchor being weak early. The high end is where the 2 × 2 is still open, and rule 7 closes it.

For rule 8, the trap runs the opposite way to the usual one. The margin peaks near epoch 10 and drifts down over the following forty, so a shorter run can score *better* on m_line while being the less converged model. The task gate and grading decide this arm; an m_line improvement on its own does not carry it.



How rules 6–8 compose. *Each arrow is one frozen comparison, read once the amendment arms settle. The left lane decides what the survey searches: rule 6 sets the anti-subspace branch, and rule 7 — read only if rule 6 simplifies — decides whether the flat anchor and the flat anti are adopted together. The right lane decides how long each trial trains, and composes with any left-lane outcome. Grey boxes are outcomes; each names the surviving configuration and the trial budget its dimension count buys.*

Shape or dose? (rules 6 and 7)

Shape. `anti-mid` matches the dose of the schedule to within rounding and still comes out worse on three of the five statistics: grading falls 0.1022 against a band of 0.0392, m_line 0.0156 against 0.0144, and the group contrast 0.0135 against 0.0093. Containment and retention stay within band, containment on the better side.

That is what the shape reading predicted, a weakened blend of the two round-1 failures with containment intact, and not what the dose reading predicted. Rule 6 confirms branch A, so the peak and anneal endpoint of the schedule stay dimensions of the search.

Grading was the statistic to watch, and it moved the most, in the direction the frozen prediction gave a mechanism for. A constant at the matched level delivers four to six times the late repulsion of the schedule over the window where grading is learned, and the response ends less graded, as it did in both round-1 brackets.

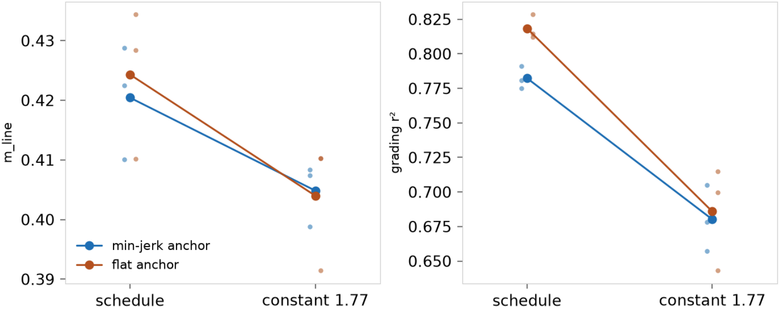
Rule 7 is unread, because its gate needs rule 6 to simplify. The arm ran anyway, and it completes the 2 × 2. Against `flat-anchor`, `anti-mid-flat` shows the same three regressions (grading −0.1325, m_line −0.0203, contrast −0.0108). In the panel below the two anchor-shape lines move nearly in parallel, so flattening the anti term costs about the same with either anchor shape.

statistic	**ref**	**band**	**anti-mid**
`m_line` ↑	+0.4204	0.0144	**-0.0156 ✗**
`alpha_op1` ↓	+0.0390	0.0346	-0.0275
`retention` ↑	+0.9897	0.0100	-0.0012
`r2_sim` ↑	+0.7823	0.0392	**-0.1022 ✗**
`contrast` ↑	+0.8562	0.0093	**-0.0135 ✗**
holdout EM − control ↓	0.0039	0.0200	0.0052
latched runs ↓	0	–	0

`anti-mid` against `ref`. *Read as the round-1 tables. The arm matches the reference dose to 0.3% by construction, so what the deltas measure is the shape alone.*

statistic	flat-anchor	band	anti-mid-flat
`m_line` ↑	+0.4243	0.0144	**-0.0203 ✗**
`alpha_op1` ↓	+0.0505	0.0346	-0.0340
`retention` ↑	+0.9975	0.0100	-0.0055
`r2_sim` ↑	+0.8184	0.0392	**-0.1325 ✗**
`contrast` ↑	+0.8573	0.0093	**-0.0108 ✗**
holdout EM − control ↓	0.0026	0.0200	0.0026
latched runs ↓	0	–	0

`anti-mid-flat` against `flat-anchor`. *The frozen comparison for rule 7. Both arms hold the anchor flat, so the anti term is again the only difference.*



The 2 × 2 of (anchor shape) × (anti term). *Seed means (●, joined) and per-run values (the smaller dots of the same color) for m_line and grading* r^2 *at the four corners. An interaction would show as the two lines diverging; instead they move nearly in parallel on both panels, so the* *cost of flattening the anti term does not depend on the anchor's shape.*

Can we halve the budget again? (rule 8)

No, but not for the predicted reason. The prediction named the task gate as the binding constraint, on the grounds that at 25 epochs the un-anchored control might stop saturating holdout EM. The control is fine: `short25-lam0` holds 0.9961, and the task cost of `short25` is 0.0026 against a gate of 0.02.

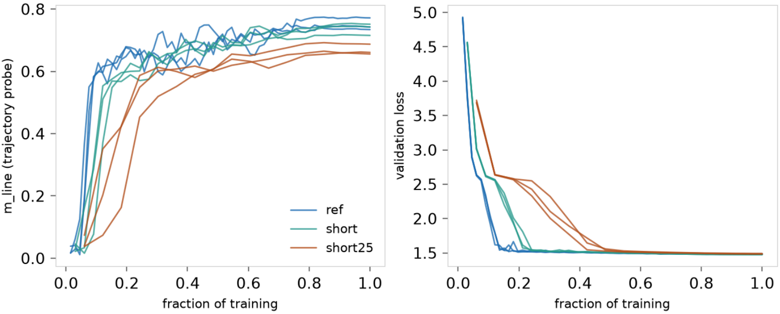
What fails is the anchored recipe itself: m_line drops 0.0268 against a band of 0.0144, containment worsens by 0.0410 against 0.0346, and the contrast drops 0.0222 against 0.0093, with grading within band. The trap the rule named, where a shorter run scores *better* on m_line while being the less converged model, did not appear either; the shorter run scores worse outright.

The trajectories below show how the halving stopped working. On the shared fraction-of-training clock, `ref` and `short` trace the same curves. But `short25` starts far behind, and spends the rest of the run catching up: at a fifth of training its margin probe reads half of what `ref` reads at the same point, and its validation loss is a full nat higher. Perhaps the shorter schedule would work with higher term weights, but that's outside the scope of this survey.

Fifty epochs leaves slack after convergence, and twenty-five leaves none. Rule 8: the survey runs at the 50 epochs set by rule 4.

statistic	**ref**	**band**	**short25**
`m_line` ↑	+0.4204	0.0144	**-0.0268 ✗**
`alpha_op1` ↓	+0.0390	0.0346	**+0.0410 ✗**
`retention` ↑	+0.9897	0.0100	+0.0041
`r2_sim` ↑	+0.7823	0.0392	+0.0051
`contrast` ↑	+0.8562	0.0093	**-0.0222 ✗**
holdout EM − control ↓	0.0039	0.0200	0.0026
latched runs ↓	0	–	0

`short25` against `ref`. *Read as the tables above; as with `short`, the task gate compares against the un-anchored control at the same length (`short25-lam0`), so convergence cost is not charged to the anchor.*



Twenty-five epochs against fifty and a hundred, on a shared clock. *Axes as in the `short` figure above: fraction of training horizontally, the cheap one-partner margin probe on the left, validation loss* *on the right. The two longer lengths trace the same curves; the quarter-length runs spend nearly half their budget still converging, and the margin they reach afterwards levels off below the others.*

What the survey runs

The rules that touch the survey resolve to one recipe and one box. The anchor weight is flat at the sampled λ_a; the anti-subspace term keeps its schedule, with the hold ratio fixed at 0.3 and its peak and anneal endpoint sampled; every trial runs 50 epochs.

dimension	low	high	scale
`lam`	0.02	1	log
`tau`	0.05	0.3	log
`anti_peak_ratio`	0.75	4	log
`anti_anneal_end_frac`	0.3	0.95	lin

That is branch A of the frozen plan, so the trial list is the one shown under *The survey space* above: 64 scrambled-Sobol points at seed 0, run at one seed each, with the top 6 feasible trials promoted to 5. The sampled coordinates are unchanged by the length rules, and so is the relative dose, which is normalized at the run's own length. The centroid column of that table is in epochs at 100, so at 50 epochs each trial's centroid scales down with it.

Two of the ablations paid for themselves in survey cost rather than in dimensions. `short` halves every trial, and `flat-anchor` removes four parameters that would otherwise have been carried untested.

The landscape

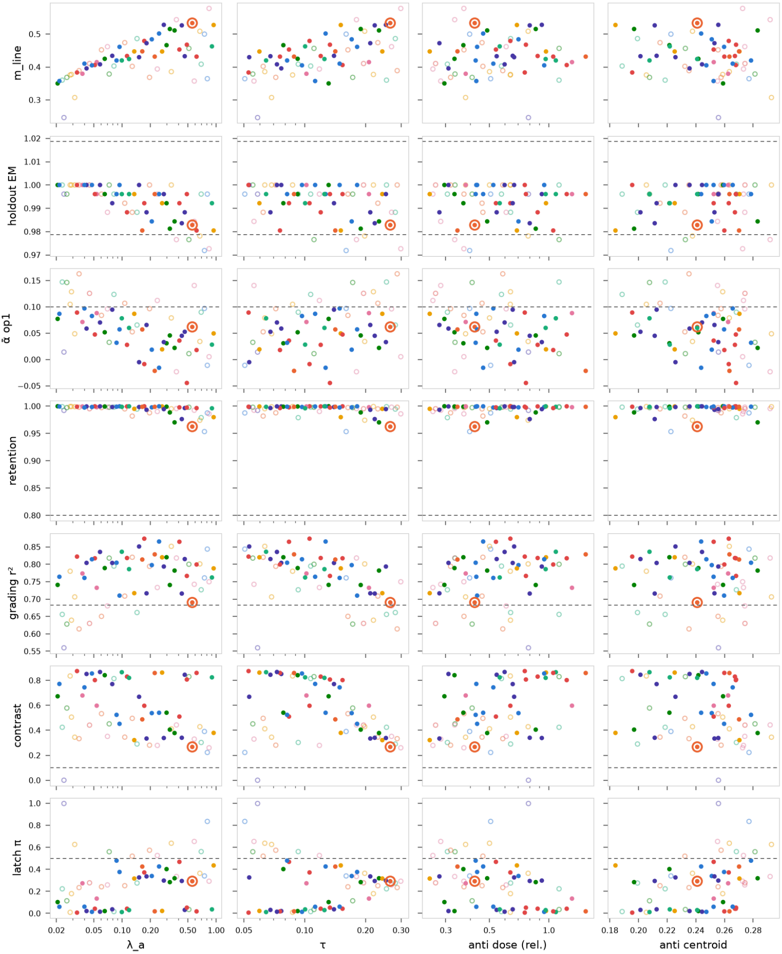
All 64 trials ran; 38 are feasible on the final reading and 26 are not. No constraint is decorative: the latch veto removes 10 trials, containment 10, grading 9, the task gate 5, and the contrast floor 1 (a trial can fail several).

The failures organize by region rather than scattering. The latch veto owns the cold edge: every vetoed trial has τ ≤ 0.12. Grading and containment fail mostly where the pull is weak (λ_a ≲ 0.1), where the response is too control-like to show graded structure. The task gate binds only where the pull is strong: every task failure has λ_a ≥ 0.38.

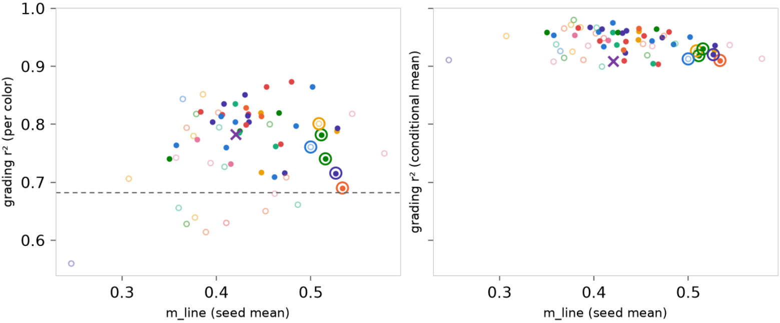
The feasible region is the interior left over, and it is wide on every axis. Across it the marginals repeat the calibration's trade at survey scale: m_line rises with λ_a until the task gate cuts in and rises with τ across the whole sampled range, while grading and the contrast fall along the same τ axis toward their floors.

The anti-subspace coordinates barely organize anything: the plateau's top spans most of the sampled range of both (its extent is quantified under *Proposed operating point*). That is the same insensitivity the amendment measured at the operating point, now visible across the box.

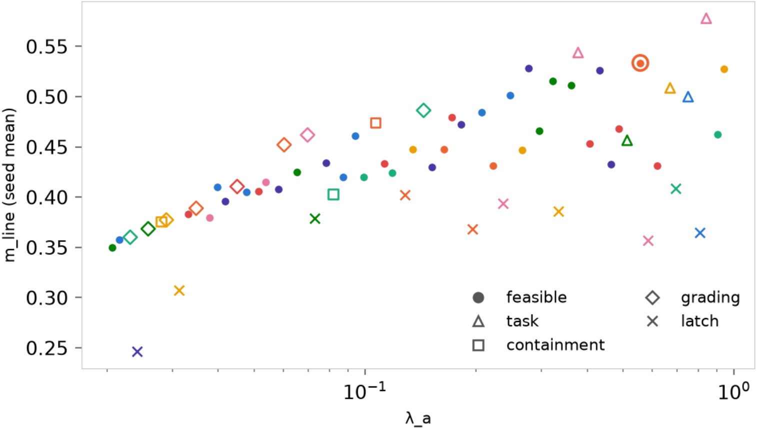
▶ Table of all trial hyperparameters...



The marginals: every statistic against every axis. *One dot per trial (seed means; five seeds for promoted trials, one otherwise), colored by an eight-hue* *categorical palette hashed from trial id (repeating past eight with no order to the repeats, so a dot's identity reads off position and color together): ● solid = feasible, ○ open circle = infeasible, ⦿ ring = the proposed trial* `t00`*. Rows are the objective and the six constraints, with each constraint's gate dashed — a band for the task row, a ceiling for containment and latch, a floor for the rest. Columns are the two sampled scalars and the two derived anti coordinates (dose relative to the reference schedule at 50 epochs; centroid as a fraction of training). The map to read off: m_line climbs with λ_a and with τ; the τ axis spends grading and contrast as it climbs; the latch veto owns the cold-τ edge; the task gate cuts in at strong λ_a; and the anti axes organize little, which is the amendment's insensitivity result seen across the whole box.*



The margin–grading plane, in both readings. *Per-trial seed means, in the trial colors used throughout this section: ● filled = feasible, ○ open = infeasible, ⦿ ring = the six promoted trials, × = the ablation stage's* `ref`*.* ***Left:*** *the per-color r^2 the grading constraint reads, with the floor (*`ref` *− 0.1) dashed; the* *cloud slopes down to the right — the τ walk trades grading for margin — and the constraint is what stops the objective from following it further.* ***Right:*** *the conditional-mean* r^2 *over redness levels, which stays high across the same trials. The pair separates graded on average from graded color by color: what the soft-τ end loses is per-color fidelity, not the average dose–response shape.*



The λ_a marginal on its own. *m_line against λ_a, in the trial colors used throughout this section: ● = feasible, ⦿ ring = the proposed* `t00`*, and each infeasible trial takes the shape of its first failed constraint — △ task, □ containment, ◇ grading, × latch (the legend keys the shapes, in gray). The shape ex-2.1.7 predicted shows up as a rise to a flat top: the margin climbs for a decade of λ_a and stops gaining around 0.3–0.6, while the task gate (△) starts to bind above 0.4 — the two highest margins in the whole survey belong to trials the task gate rejected, so the task, and nothing else, caps the strong end. At the weak end grading (◇) and containment (□) fail together where the pull is too weak to organize the response.*

Proposed operating point

point	λ_a	τ	peak	anneal end	dose	m_line ↑	alpha_op1 ↓	retention ↑	r2_sim ↑	contrast ↑	holdout EM ↑
proposed (`t00`)	0.557	0.265	1.376	0.537	0.42	0.5336 ± 0.0039	0.0619 ± 0.0095	0.9626 ± 0.0027	0.6902 ± 0.0107	0.2672 ± 0.0025	0.9828 ± 0.0039
`ref` (recipe)	0.1	0.1	2.5	0.9	1.00	0.4204 ± 0.0051	0.0390 ± 0.0122	0.9897 ± 0.0035	0.7823 ± 0.0139	0.8562 ± 0.0033	0.9961 ± 0.0050

The proposed operating point beside the recipe it would replace. *Seed means ± the frozen per-run noise floor over √n (five seeds for the proposal, three for `ref`). The coordinate columns are the sampled values plus the relative anti dose; `ref`'s dose is 1 by definition. These numbers are proposals: D2.2 adopts the point and re-measures it at fresh seeds before any of them is quoted.*

The frozen rule proposes `t00`, the promoted trial with the highest seed-mean m_line that stays feasible on every constraint. Relative to the recipe it is a stronger pull with softer pooling and less repulsion: λ_a rises from 0.1 to 0.56, τ softens from 0.1 to 0.27, and the anti term keeps the shape of the schedule at 0.42× the reference dose.

The margin it buys is 0.534, against 0.420 for `ref`. Containment, retention and the task gate pass with room. Grading is 0.008 above its floor, a third of the per-run noise of that statistic, and it is the closest pass among the checks on the proposal.

The proposal is on a plateau rather than a peak. 5 trials land within one m_line band (0.019, five seeds against one) of the proposal, spanning λ_a 0.28–0.94, τ 0.15–0.27, peak ratio 1.4–3.7 and relative dose 0.42–0.92.

The runners-up among them hold 0.03–0.10 more grading at nearly the same margin (the plane figure above). So what the search found is the plateau together with the constraint that bounds it, and the proposal is one end of that plateau.

trial	m_line, seed 0	m_line, 5 seeds	change	feasible at 5 seeds
● `t00`	0.5646	0.5336	-0.0310	✓
○ `t24`	0.5567	0.5087	-0.0480	✗ task
● `t12`	0.5518	0.5111	-0.0406	✓
● `t48`	0.5516	0.5266	-0.0251	✓
○ `t56`	0.5414	0.5000	-0.0414	✗ task
● `t44`	0.5338	0.5155	-0.0183	✓

The promotion round is the winner's curse, measured. *The six trials promoted on their seed-0 m_line, in promotion order. Every one fell back on re-seeding, so the trials were promoted partly on merit and partly on a lucky seed, and two crossed the task gate on their five-seed means and dropped out of feasibility. The five-seed mean of the proposal already carries this correction, while the single-seed trials around it on the plateau do not. That is one more reason the numbers here are proposals for D2.2 rather than results.*

The handoff. D2.2 adopts this operating point and re-measures it at fresh seeds before quoting any number from this section. The survey values stand beside the confirmed ones there, and the difference between them is the remaining winner's-curse correction.

The statistic to watch at confirmation is grading. The proposal passes its floor by 0.008, so a modest unlucky draw at fresh seeds would put it under. If that happens, the place to step back to is the plateau, where the runners-up hold more grading at nearly the same margin; the exploratory section names them.

Exploratory analyses

Post hoc: noticed after seeing results, not part of the frozen plan.

The latch veto fires on an un-anchored control

One run in each un-anchored control trips the veto: `lam0-s2` puts 0.55 of its deep-slice weight on `+`, and `short-lam0-s0` puts 0.52 there, both above the 0.5 threshold. The detector is reading a profile that only means something when there is a pull. With λ_a = 0, the softmin weights describe whatever alignment structure the task happens to leave behind, and there is no pull that could have latched.

The veto is calibrated on the anchored runs of ex-2.1.9, and the controls were never going to be judged by it, so nothing here is affected.

It does bear on the survey. λ_a runs down to 0.02, a fifth of the scored rung, and a trial that weak may look like the control rather than like a pull that committed. Such a trial would be marked infeasible for a reason other than the one the veto is named for.

It would fail the contrast constraint too, so the error goes in the safe direction: we would reject a trial that was never going to be proposed. We are leaving the veto frozen and reporting the λ_a of every vetoed trial, so the two readings can be told apart in the map.

Where the vetoes actually landed

The vetoes from the survey sort by τ rather than by λ_a. All 10 vetoed trials sit at τ ≤ 0.12, the cold end of the box. 3 of them are the weak-pull corner the paragraph above anticipated (λ_a ≤ 0.07; one, `t29` at π = 1.00, also fails grading and contrast, which is the control-like signature).

The other 7 have healthy pulls, λ_a 0.13–0.81. Their π values, 0.52 to 0.83, fall inside the range ex-2.1.9's calibration found empty, with latched runs at ≥ 0.85 and healthy runs at ≤ 0.03. Cold pooling concentrates the softmin weight, so a modest alignment with a syntax role reads as a large π there.

Whether these are true latches, or the detector reading far from its calibration point, cannot be settled from this data. The error direction is safe either way, since every vetoed trial is in the τ region whose feasible neighbors score well below the plateau. The frozen noise floors carry the same cold-regime caveat, which is one more reason for D2.2 to stay off the τ ≲ 0.1 edge unless it brings a recalibrated detector.

Naming the step-back points

The objective cannot rank the 4 promoted trials that are still feasible on five-seed means. Their m_line values span only 0.022, or 2.0 resolution bands at five seeds a side.

The same trials span 0.092 of grading (3.0 bands) and 0.138 of contrast (19 bands). So the frozen argmax ordered them by differences near its own noise floor, and the places where they differ measurably are the statistics it treats as constraints.

	trial	λ_a	τ	dose	m_line ↑	r2_sim ↑	contrast ↑	r^2 − floor
●	`t00`	0.557	0.265	0.42	0.534 ± 0.004	0.690 ± 0.011	0.267 ± 0.003	+0.008
●	`t48`	0.431	0.222	0.80	0.527 ± 0.004	0.716 ± 0.011	0.340 ± 0.003	+0.033
●	`t44`	0.322	0.189	0.86	0.516 ± 0.004	0.740 ± 0.011	0.405 ± 0.003	+0.058
●	`t12`	0.361	0.233	0.49	0.511 ± 0.004	0.782 ± 0.011	0.378 ± 0.003	+0.100

The plateau's confirmed members. *The promoted trials still feasible on five-seed means, ranked as the frozen rule ranks them. Coordinates as sampled (dose relative to the reference schedule at the same length); statistics are seed means ± the frozen per-run floor over $\sqrt{5}$. The last column is each trial's margin above the grading floor the feasibility check reads. A post-hoc re-cut of numbers already published above; no new runs.*

Two step-back points stand out:

- `t48` is within one band of `t00` on m_line and on grading, so the search cannot tell the two apart on either. It holds 0.07 more contrast (10 bands), and its grading margin is 1.4 per-run σ, against 0.3 for `t00`.

- `t12` is the knee. Giving up 0.022 of m_line (2.0 bands) brings grading back to the level of the reference recipe: 0.782 against 0.782 for `ref`, a floor margin of 4.1 σ. It also holds 0.11 more contrast than `t00`.

The margin matters because two of the six promoted trials crossed the task gate on their five-seed means. A grading pass of 0.3 σ per run faces the same draw again at the fresh seeds in D2.2.

None of this reopens the frozen rule, and `t00` stands as the proposal from the survey. The recommendation goes instead to the preregistration for D2.2, which is not yet frozen: adopt `t00`, `t12`, and `t48` as candidate points, re-measure all three at fresh seeds, and freeze there a selection rule that weighs grading and contrast alongside the margin.

Discussion

The shape of the schedule transfers from M1, and the dose does not: a constant that delivers the same integrated repulsion loses the grading the schedule protects, with either anchor shape. The schedule on the anchor term, which M1 carried for training stability, resolves nothing here, and neither does the interpolation shape.

Of everything inherited from the autoencoder recipe, the one load-bearing piece of scheduling is *when the repulsion arrives*: strong while the pull is committing, reduced while grading is learned. That reading now stands on a one-variable comparison rather than on brackets that moved two things at once.

Within this testbed, the map says the recipe does not need careful tuning. The feasible region spans a decade of λ_a and most of the τ box, the top of the m_line landscape is a plateau rather than a spike, and the anti-subspace coordinates barely matter within the schedule family.

Each edge of the region belongs to a different constraint: the latch veto at cold τ, grading and contrast at soft τ, the task gate at strong λ_a, and grading and containment where the pull is too weak. So the box a future search inherits is drawn by measured boundaries rather than by guesses.

The survey format did what it was adopted for. The amendment could be added mid-experiment without spending the preregistration, because the frozen rules, their predictions, and the record-as-a-diff made the addition inspectable. Publishing every trial is what makes the plateau readable as a result.

Two frictions are worth carrying to the next survey plan. First, a scalar objective walks to its constraint boundary. The proposal is on the grading floor with a pass margin a third of the per-run noise of that statistic, so a plan that proposes a single point should either freeze how close to a constraint that point may be, or expect the confirming experiment to step back onto the plateau.

Second, ranking on single-seed values promoted two trials that the task gate rejected at five seeds. The gate width (0.02) is about twice the per-run noise of holdout EM (0.0087), so a seed-0 pass near the boundary carries little information. The curse correction absorbed both, which is what promotion is for, but a future plan could rank on constraint margins as well as on the objective.

---

1. The residual stream is the running vector of activations that each transformer layer reads from and writes back into. ↩

2. The set of points at which no objective can be improved without giving up some of another. It is a curve to choose along, rather than a single winner. ↩