# When Does Loss Normalization Preserve Stationarity in Consistency Learning?

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

## Abstract

Consistency models are commonly trained with pseudo-Huber losses, whose gradients normalize each sampled correction before averaging. Normalization can move a model away from its intended flow map even when the expected unnormalized gradient vanishes there. We characterize when it preserves this stationary state. For a fixed joint law with integrable Jacobian-weighted corrections, stationarity holds at every positive floor, or smoothing constant, exactly when the Jacobian-weighted correction has zero mean conditional on the correction magnitude. This radial-balance condition requires cancellation within each group of equal correction magnitude, so it is stronger than cancellation of the overall mean. A smooth Gaussian mixture violates it at the exact map, and one of three networks trained to approximate this map meets prespecified criteria for accuracy and for separating the normalized field from the linear one. For fixed features and quadratic error, we derive sufficient conditions under which one step reduces the error, while a step of the same size from the same state under the reversed correction law increases it; on 16 fresh feature seeds, every resolved forecast agrees with the computed outcome. Across six paired tuning seeds from one pretrained diffusion model, two-step Fréchet inception distance favors zero-floor normalization over a squared-loss control with matched update lengths; three earlier continuation studies give mixed results that depend on the update policy.

## 1 Introduction

Consider a flow from a standard Gaussian to a mixture with weights 80% and 20%, centered at $+1$ and $-4$. Halfway along the path, a point with equal component likelihoods keeps posterior weights 80/20. At the exact flow map, the corrections at this point average to zero, because the more frequent positive corrections balance the rarer, larger negative ones. Normalization discounts large magnitudes and breaks this cancellation, whereas symmetric corrections would cancel under the same odd score. Section 5 shows that this conditional imbalance can also produce a nonzero parameter update.

Pseudo-Huber (PH) losses and tangent normalization are standard in consistency training (Song and Dhariwal, 2024; Geng et al., 2025; Lu and Song, 2025), where averaging sampled data–noise corrections estimates the marginal transport of the intended map (Kim et al., 2026; Lee et al., 2025). Under loss normalization, stationarity requires the expected semi-gradient to cancel before optimizer processing. By the normalization floor, we mean the constant that softens the division near a zero correction. We show that, for a fixed joint law with integrable Jacobian-weighted corrections, stationarity at every positive floor is equivalent to *radial balance*: conditional on the correction magnitude, the Jacobian-weighted correction averages to zero. Radial balance follows from sign symmetry of the Jacobian-weighted contributions given the correction magnitude, and it implies cancellation of the unnormalized parameter field; neither converse holds. Three equal-length contributions at the vertices of an equilateral triangle centered at zero are balanced without being centrally symmetric, while the mixture example cancels in mean without being balanced. The criterion follows from classical transform uniqueness, and when balance fails, analyticity makes the cancellation floors isolated (Theorem 2).

A conditional imbalance affects training when it survives projection through the model Jacobian, and its effect on the error also depends on the initial approximation and step size. We derive exact effective-floor identities from the backward graph (Section 3) and characterize stationarity of the resulting



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

expected parameter field (Section 4). Of three supervised approximations, one meets prespecified criteria for near-oracle accuracy and for separating normalized from linear fields (Sections 5–6.1). For a frozen-feature model, we show how the response to approximation error can prevent reflection of the oracle drift from reversing an update's benefit, and derive conditions under which this reversal does occur (Section 7). Finally, trained-state measurements and image experiments assess how normalization changes updates and sample quality in the tested models (Section 8). The practical lesson is to check the executed update field before relying on stationarity of the target flow; our results give no general preference for normalized or linear training.

## 2 Related work

Robust losses can shift statistical targets, and pseudo-Huber weighting in consistency training is already recognized (Sun et al., 2020; Gneiting, 2011; Patterson et al., 2022; Geng et al., 2025). ACT-Diffusion also notes that consistency gradients can remain nonzero at distributional agreement (Kong et al., 2024). We study the executed semi-gradient: Proposition 1 identifies graph settings that give equal gradients, and Theorem 2 characterizes stationarity at every positive floor for a fixed joint law. Transform uniqueness and orthogonal projection are classical tools; applied to this law, they identify the cancellation a linear transport argument needs to carry over to normalized updates.

**Relation to flow-map analyses.** Li et al. (2025) establish sampling bounds under approximation to squared-loss conditional-regression targets, and Stable Consistency Tuning averages over conditional noise to construct variance-reduced targets (Wang et al., 2024). Both are mean-based, which motivates checking whether cancellation survives nonlinear scoring. The linear cancellations in Kim et al. (2026), the fixed-metric expansion of Truncated Consistency Models (Lee et al., 2025), and the inverse-gap limits of Align Your Flow (Sabour et al., 2025) rest on different graph and scaling premises (Table 3). Adaptive Discretization for Consistency Models combines adaptive gaps with normalized losses (Bai et al., 2025), and its loss-ratio denominator also changes the outer weight. Learned coupling changes the correction law (Silvestri et al., 2025), while composite weighted temporal-difference objectives change the loss (Ying et al., 2026). Applying our criterion to these training systems would require control of how their correction law and objective evolve.

Distributional guarantees involve further conditions. Dou et al. (2024) analyze global minimizers of Wasserstein consistency objectives, and Jain et al. (2025) derive Kullback–Leibler divergence bounds for multistep sampling with renoising and bound the map approximation under an assumed local distillation error. Related flow-map and sampling analyses impose approximation, identifiability, and regularity premises (Lyu et al., 2024; Zhang and Zhou, 2025; Boffi et al., 2025). Goldstein et al. (2026) characterize functional stationary points for a distinct stop-gradient construction under differentiability and time-sampling assumptions. We locate our fixed-law stationarity criterion within this literature in Appendix B.

## 3 The executed update and its effective scale

The update field is determined by the backward graph, including which targets are detached. Let $z = (x, t)$ be an observed condition and $V \mid z$ a conditional velocity with mean $v(z)$. For a smooth map $f$, define

$$C_f(V; z) = \partial_t f(z) + D_x f(z)V. \tag{1}$$

A data-directed marginal flow map satisfies $\mathbb{E}[C_f \mid z] = 0$ and, in our time convention (prior at zero, data at one), the endpoint condition $f(x, 1) = x$. For a finite network, stationarity additionally requires projecting into parameter space and integrating over $z$.

For a gap $\delta > 0$, put $f_0 = f_\theta(x, t)$, $f_- = f_\theta(x - \delta V, t - \delta)$ and $C_\delta = (f_0 - f_-)/\delta$. Consider the actual backward graph

$$y = (1 - a)\,\mathrm{sg}(f_0) + a\,\mathrm{sg}(f_-), \qquad H_c(r) = \sqrt{\|r\|^2 + c^2} - c, \qquad a, c > 0. \tag{2}$$

The residual is $r = f_0 - y$. Here sg keeps a prediction's value but blocks its derivative, and the linear control uses the residual itself as its score, as squared error does. Because the target is frozen during differentiation but recomputed whenever the model changes, the learning field is a semi-gradient; it need not be the gradient of any fixed scalar objective on the shared student–teacher parameters.

The current prediction in the anchor must use the same realization as the differentiable prediction. Coefficients, gaps, and outer weights remain fixed during this backward operation.

**Proposition 1** (Exact effective-floor identity)**.** *Let $J = \partial_\theta f_0$. The loss floor $c$ becomes the effective correction floor $\tau = c/(a\delta)$:*

$$g_{a,c} = J^\top \frac{C_\delta}{\sqrt{\|C_\delta\|^2 + \tau^2}} = J^\top \frac{f_0 - f_-}{\sqrt{\|f_0 - f_-\|^2 + (c/a)^2}}. \tag{3}$$

*Hence $(a, c)$ and $(1, c/a)$ give equal gradients when every other input, target prediction, random realization, and outer weight is shared. At $c = 0$, the same identity holds for nonzero residuals, and every positive $a$ cancels.*

*Proof.* The residual value is $a(f_0 - f_-)$, but its backward derivative is $J$. Substitution into $\nabla_r H_c = r/\sqrt{\|r\|^2 + c^2}$ proves the claim. Differentiating the residual value after deleting stop-gradient annotations would incorrectly differentiate through the target. □

**Optimizer and trajectory equivalence.** Equal expected fields additionally require the same joint law and integrability. Equal gradients give equal processed steps only when optimizer states, hyperparameters, and processing match, and equality of mean gradients alone need not survive clipping or Adam. Equal trajectories further require coupled complete states and randomness, including loss-driven schedules, since the forward losses can differ. We prove these transfer rules and give counterexamples in Appendix B.1.

**Sample, conditional, and parameter fields.** Define $\psi_\tau(u) = u/\sqrt{\|u\|^2 + \tau^2}$ and $S_\tau(u) = \tau\psi_\tau(u)$. We call one coupled realization $J^\top\psi_\tau(C_\delta)$, as in Proposition 1, a *sample gradient*. In the differential limit, averaging over $V \mid z$ at a fixed condition gives the *conditional field* $B_\tau(f; z) = \mathbb{E}[\psi_\tau(C_f) \mid z]$. For deterministic $J(z)$, projecting and averaging over $z \sim \mu$ gives the *parameter field* $\mathbb{E}_\mu[J(z)^\top B_\tau(f; z)]$, and parameter stationarity means that this field is zero; projection can cancel a nonzero conditional field. For random Jacobians, $J$ stays inside the joint expectation (Theorem 10).

We take these expectations with the model, teacher, and sampling law fixed. An exponential moving average (EMA) teacher can preserve the samplewise algebra yet change the transport limit; Table 3 summarizes how far the identity and its population limit carry over to other graphs. For finite $\tau > 0$ and $\mathbb{E}[\|V\| \mid z] < \infty$, the conditional field has the weighted-velocity form

$$B_\tau = \mathbb{E}[w_f \mid z]\big(\partial_t f + D_x f\, v_{w_f}\big), \tag{4}$$

$$w_f = (\|C_f\|^2 + \tau^2)^{-1/2}, \qquad v_{w_f} = \frac{\mathbb{E}[w_f V \mid z]}{\mathbb{E}[w_f \mid z]}. \tag{5}$$

The effective velocity depends on the residual. Preservation requires $D_x f(v_{w_f} - v) = 0$, which can hold even when $v_{w_f} \neq v$. Under conditional symmetry and a unique complete characteristic flow, the intended map is the only $C^1$ endpoint-preserving map with a vanishing conditional field (Proposition 4).

With fixed $a, c > 0$, $\tau = c/(a\delta)$ grows as $\delta \to 0$. Under reciprocal-gap weighting, weighted uniform integrability, and the stated convergence assumptions, the field approaches the linear field multiplied by $a/c$ (Appendices A.3–A.4). Scaling the loss floor as $c = a\delta\tau_0$ instead retains a nonlinear score.

## 4 Stationarity across all positive floors

We now vary the floor, holding the model, teacher, and sampling law fixed, and ask whether the model remains stationary throughout the normalization family. The criterion follows from uniqueness of the generalized Stieltjes transform of order one half (Gomilko and Tomilov, 2023, Section 2). Analyticity further implies that, when radial balance fails, the floors at which cancellation occurs are isolated.

**Theorem 2** (All-floor preservation)**.** *Fix a joint law, a finite nonnegative random variable $R$, and a vector $H$ with $\mathbb{E}\|H\| < \infty$. Then*

$$\mathbb{E}\left[\frac{H}{\sqrt{R^2 + \tau^2}}\right] = 0 \quad \textit{for every } \tau > 0 \quad \Longleftrightarrow \quad \mathbb{E}[H \mid R] = 0 \quad \textit{a.s.}$$

*Vanishing on any nonempty open interval is sufficient. Otherwise, the zero floors are isolated in* $(0, \infty)$ *and have measure zero. The law of* $(H, R)$ *remains fixed as the floor varies.*

**Radial balance.** Group the contributions $H$ by their correction magnitude $R$; for parameter fields, a group collects the Jacobian-weighted corrections $H = J^\top C$ that share the unprojected magnitude $R = \|C\|$. Normalization multiplies every contribution in a group by the same factor $1/\sqrt{R^2 + \tau^2}$, so changing the floor changes the relative weights of the groups. Cancellation survives every positive floor exactly when the mean contribution is zero conditional on almost every magnitude. A zero overall mean $\mathbb{E}[J^\top C]$ can instead arise through cancellation between groups, which reweighting can undo; we measure this effect at a chosen floor in Section 6.1.

The equilateral triangle has a single radius and zero mean, so its score cancels at every floor. Contributions at distinct radii can cancel at one particular floor while violating radial balance.

For the intended conditional map at fixed $z$, we set $H = U = D_x f_*(V - v)$ and $R = \|U\|$; for the parameter field, we set $H = J^\top C$ and $R = \|C\|$, retaining their dependence. We prove the criterion and the isolated-zero result in Appendix A.1, which also gives a local extension by continuity.

A sufficient test at a specified floor follows from Equation (12): if $m_3 = \mathbb{E}[U\|U\|^2 \mid z] \neq 0$ and $M_5 = \mathbb{E}[\|U\|^5 \mid z] < \infty$, then $B_\tau \neq 0$ whenever $\tau^2 > 3M_5/(4\|m_3\|)$. Asymmetry alone is inconclusive when this bound fails.

## 5 A SMOOTH COUNTEREXAMPLE TO STATIONARITY PRESERVATION

At its exact data-directed map, the Gaussian mixture from the introduction violates radial balance, although its densities are smooth and its linear conditional correction vanishes. Let $Z \sim N(0, 1)$ and, independently, $Y \sim 0.8N(1, 1) + 0.2N(-4, 1)$. Put $X_t = (1 - t)Z + tY$ and $V = Y - Z$. Let $F_t$ and $\rho_t$ be the cumulative distribution function and density of $X_t$. The data-directed quantile flow map is

$$f_*(x, t) = F_1^{-1}(F_t(x)), \qquad f_*(x, 1) = x. \tag{6}$$

Since all densities are smooth and positive, the continuity equation gives $(f_*)_t = -v(f_*)_x$ with $(f_*)_x = \rho_t(x)/\rho_1(f_*(x, t)) > 0$, and hence $\mathbb{E}[C_{f_*} \mid x, t] = 0$.

**Proposition 3** (Smooth nonstationarity witness)**.** *At* $t = 1/2$*,* $x = -3/4$*, the map in Equation* (6) *has zero linear conditional correction but a strictly positive normalized-sign correction. Consequently,* $B_\tau(f_*; x, t) > 0$ *for all sufficiently small positive* $\tau$*. There exists a smooth endpoint-preserving scalar parameter perturbation and an observed-condition distribution with positive density near this point for which the integrated normalized semi-gradient at* $f_*$ *is nonzero.*

*Proof.* The two component likelihoods for $X_{1/2} = -3/4$ are equal, so the posterior component weights remain $0.8, 0.2$. A conditional Gaussian calculation gives $V \mid x, t \sim 0.8N(1, 2) + 0.2N(-4, 2)$, with variances given as the second arguments. Its mean is zero. Since $(f_*)_x > 0$, $\operatorname{sign}(C_{f_*}) = \operatorname{sign}(V)$. Writing $\Phi$ for the standard Gaussian CDF,

$$b = 0.8[2\Phi(1/\sqrt{2}) - 1] + 0.2[2\Phi(-4/\sqrt{2}) - 1] > 0. \tag{7}$$

For an analytic positivity bound, monotonicity of the Gaussian density on the positive axis yields $2\Phi(1/\sqrt{2}) - 1 \geq e^{-1/4}/\sqrt{\pi}$. The second bracket is greater than $-1$, so $b > 0.8e^{-1/4}/\sqrt{\pi} - 0.2 > 0$. There is no atom at zero; bounded convergence proves the small-positive-$\tau$ claim. For any such fixed $\tau$, continuity provides an interior neighborhood with positive $B_\tau$. Choose a nonnegative smooth bump $h$ supported there and not identically zero. The family $f_\theta = f_* + \theta h$ preserves the endpoint boundary. For a condition density $q$ positive on that support, the integrated normalized semi-gradient at $\theta = 0$ is $\int qhB_\tau \, dz > 0$. □

As Figure 1 shows, the finite-gap signal depends on how the floor scales with the gap. Existence follows from the analytic proof, and quadrature checks the finite-gap calculation (values and grid details in Table 5).

Conditionally symmetric corrections cancel because the score is odd, and corrections built from the exact marginal velocity vanish pointwise at the oracle; at a finite Euler gap, however, the marginal velocity alone leaves discretization error, so a pointwise zero control needs an exact finite-flow target.

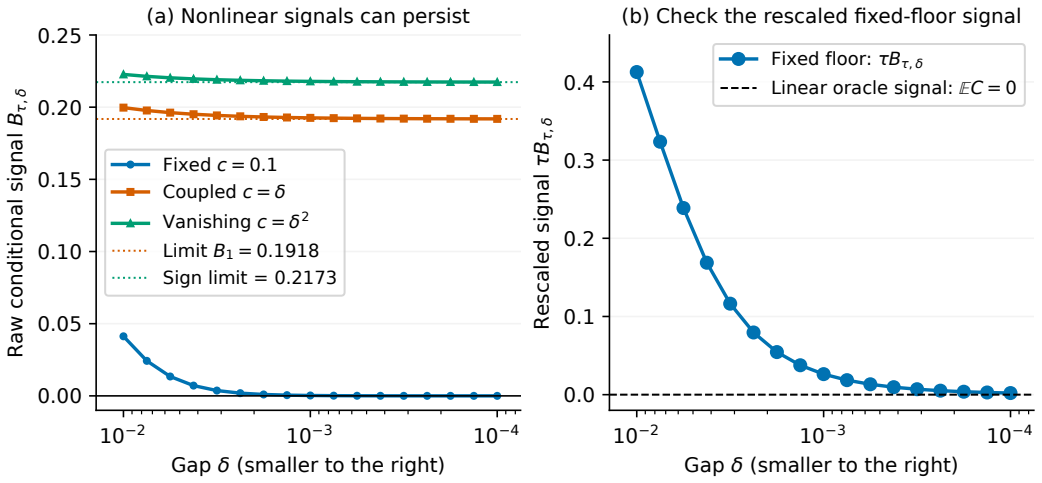



Figure 1: **The signal at the exact map depends on how the floor scales with the gap.** At the smooth witness of Section 5, with $a = 1$: (a) a coupled floor ($c = \delta$) or faster-vanishing floor ($c = \delta^2$) retains a nonzero normalized limit, whereas a fixed floor lets the raw signal vanish; (b) after the rescaling of Theorem 7, the fixed-floor signal approaches the linear oracle signal, which is zero. Deterministic quadrature at one witness; values and resolution details are in Table 5.

**Proposition 4** (A natural preservation and uniqueness case)**.** *Suppose that at every $z$, $V = v(z) + \varepsilon_z$ with $\varepsilon_z$ conditionally centrally symmetric and integrable. For every $C^1$ map and finite $\tau(z) > 0$,*

$$B_\tau(f; z) = 0 \quad \Longleftrightarrow \quad f_t + D_x f\, v = 0.$$

*If $v$ admits a unique complete $C^1$ characteristic flow to time 1, the unique $C^1$ endpoint-preserving solution is $f(x,t) = \Phi_{t,1}(x)$.*

*Proof.* Put $b = f_t + D_x f\, v$ and $U = D_x f\, \varepsilon_z$. Symmetry and oddness give $\mathbb{E}\psi_\tau(U) = 0$. The score has a positive-definite Jacobian, with smallest eigenvalue $\tau^2/(\|u\|^2 + \tau^2)^{3/2}$. For $b \neq 0$, therefore $b^\top[\psi_\tau(b+u) - \psi_\tau(u)] > 0$ pointwise and $b^\top B_\tau > 0$. Thus, a vanishing conditional field is equivalent to the transport equation. The endpoint condition and the uniqueness of the characteristic flow determine its solution. □

Symmetry of the oracle correction alone ensures only that the oracle is stationary. Proposition 4 assumes symmetric velocity fluctuations at every condition and obtains the stronger equivalence for every candidate map.

## 6 From Conditional Discrepancy to Parameter Departure

For a deterministic Jacobian and condition law $\mu$, the parameter field $g = \mathbb{E}_\mu[J^\top B_\tau]$ depends on the projection of the conditional score onto the model's tangent space. Jiang and Zou (2026) also use parameter-space geometry in their memorization analysis. Their theoretical model studies the spectrum of a local squared consistency objective with shared random features. Our projection, by contrast, uses the live Jacobian of the semi-gradient with detached targets.

**Proposition 5** (Finite-tangent criterion and one-step departure)**.** *Fix a probability measure $\mu$, a deterministic square-integrable model Jacobian $J(z) \in \mathbb{R}^{d\times p}$, and $B(z) \in L^2(\mu; \mathbb{R}^d)$. Set $Tu = J(\cdot)u$, $K = \mathbb{E}_\mu J^\top J$, and $g = \mathbb{E}_\mu J^\top B$. Then*

$$g \neq 0 \quad \Longleftrightarrow \quad \|P_{\mathrm{range}(T)} B\|_{L^2}^2 = g^\top K^\dagger g > 0. \tag{8}$$

*In particular, $B \neq 0$ and $\|Tu - B\|_{L^2} \leq \kappa \|B\|_{L^2}$ for some $u$ and $\kappa < 1$ imply $\|g\| \geq (1-\kappa)\|B\|_{L^2}^2/\|u\| > 0$. If the model's $L^2(\mu)$ Taylor remainder along $-\eta g$ is bounded by $M\eta^2\|g\|^2/2$, an actual deterministic step satisfies*

$$\|f_{\theta-\eta g} - f_\theta\|_{L^2} \geq \eta\sqrt{g^\top K g} - \tfrac{1}{2}M\eta^2\|g\|^2. \tag{9}$$

*For $g \neq 0$ this is positive for sufficiently small $\eta$; the remainder vanishes when the updated parameterization is affine.*

Equation (8) follows from the standard orthogonal-projection identity (Appendix A.2). The criterion depends on both the model tangent and the sampling measure; for example, $B$ can be orthogonal to every tangent even when $K$ is nonsingular. For random Jacobians, the relevant object is the joint field $G_\tau = \mathbb{E}[wJ^\top \psi_\tau(C) \mid \mathcal{S}]$ at a fixed complete state $\mathcal{S}$.

Consider two deterministic updates from the same state along the PH and linear expected parameter fields $g_P, g_L$, with a shared fixed linear preconditioner $A$ and step size $\eta$. Their first-order output difference is $-\eta TA(g_P - g_L)$. Its $L^2(\mu)$ norm measures functional separation; the Taylor remainder is bounded in Appendix A.2. Distinct parameter fields can produce the same first-order output change. We have not related this separation to subsequent sample quality.

We measure these fields with Jacobian products, fitting projection directions on one sample and evaluating them on independent data; for batch-dependent networks, the independent units are complete batches. Estimators of the fields and of radial moments, with their uncertainty and sample requirements, are given in Appendix B.2. A projection interval that contains zero leaves cancellation unresolved.

### 6.1 A STANDALONE LEARNED WITNESS

For the smooth mixture problem, we use supervised training to fit three standalone networks to the exact map and its derivatives. We then freeze the networks. We compare $g_L = \mathbb{E}[J^\top C]$ with $g_P = \mathbb{E}[J^\top \tau \psi_\tau(C)]$ at $\tau = .1$, where the factor $\tau$ matches the derivative of the linear score at zero. For a common fixed $\tau > 0$, this factor preserves the zeros and direction of the PH field and its collinearity with the linear field. It changes magnitudes, PH–linear contrasts, and the diagnostic's separation thresholds. Evaluation uses an independently sampled truncated quantile law, with separate samples for calibration, direction fitting, and evaluation (Appendix C.1).

Table 1: Standalone learned witness, all three seeds. Map and derivative entries are normalized root-mean-square error (RMSE) upper bounds. $U_L$ bounds the full linear-field norm; $\ell_P$ bounds an independently fitted, fixed-sign PH projection from below; $S_{\mathrm{cal}}$ is an independently calibrated root-mean-square scale of oracle sample contributions with the learned Jacobian. Bounds use one approximate simultaneous family. The four gates are map $\leq .001$, derivative $\leq .01$, $U_L/S_{\mathrm{cal}} \leq .001$, and $\ell_P/U_L > 5$. Ratios are rounded.

| Seed | Map | Derivative | $U_L/S_{\mathrm{cal}}$ | $\ell_P/U_L$ | All gates |
|---|---|---|---|---|---|
| 42 | .000554 | .013701 | $7.76\times10^{-5}$ | 4.71 | No |
| 123 | .000455 | .010958 | $8.30\times10^{-5}$ | 5.21 | No |
| 777 | .000418 | .008506 | $4.98\times10^{-5}$ | 7.05 | Yes |

All three networks meet the tolerance for a small linear field, and each has a detectably nonzero PH projection, but only seed 777 passes all four gates. The linear fields are small but themselves detectably nonzero. When we keep the learned Jacobian and replace only the correction by its oracle value, the PH projection changes by a small but detectable amount (Appendix C.1); this comparison measures a single projection in the fixed parameterization.

**Supporting oracle controls.** At the smooth oracle, a residual multilayer perceptron receives a nonzero integrated normalized update, while linear and exact-flow controls stay stationary. This response persists across ten hidden-network initializations at $\tau = .1, 1$ under a nonlocalized condition law, and one step agrees with the tangent prediction within $10^{-12}$. Bias-free, nonzero-readout, and orthogonal-tangent variants probe how the response depends on the projection (Appendix C.2).

## 7 CONTROLLED INTERVENTION AND CONDITIONS FOR IMPROVEMENT

Away from the exact map, the update combines the oracle drift with a response to approximation error. Their relative magnitudes determine whether reversing the oracle drift turns a beneficial update

into a harmful one. To vary the direction of the drift in the smooth mixture problem of Section 5, we replace $V$ by

$$V_\kappa = v + S(V - v), \qquad \Pr(S = 1) = (1 + \kappa)/2, \qquad \kappa \in \{-1, 0, 1\}.$$

Here $S \in \{-1, 1\}$ is conditionally independent of $V$. The mean velocity remains $v$, and at the oracle $C_*(V_\kappa) = SC_*(V)$, which preserves correction magnitudes and multiplies the expected normalized parameter field by $\kappa$. Reflection therefore reverses the oracle drift and symmetrization cancels it; each intervention defines its own fixed law.

Our original experiment compares raw PH and condition-only steps in $f_\theta = f_* + F^\top\theta$, with 16 frozen features; the condition-only control averages the PH weight at each $(x, t)$ and thus removes its dependence on the sampled correction. The anticipated reversal from benefit to harm fails at the original error scale: PH beats this control in both designed orientations, and one orientation still deteriorates from initialization (Table 6). In all 16 cases with an adverse oracle slope, the finite-error response outweighs that slope even in the small-step limit. All 1,344 branches of the eight-seed experiment, together with a decomposition of the finite-error response, are reported in Appendices C.3 and C.4.

**Additional conditions for improvement.** Let $g_{\mathrm{PH}}$ and $g_C$ denote the PH and condition-only expected parameter fields. We evaluate their raw steps by $Q(\theta) = \theta^\top M\theta$, where $M = \mathrm{mean}_{\mathrm{test}} FF^\top$, write $\langle x, y\rangle_M = x^\top My$ with induced seminorm $\|x\|_M$, and set $\theta = \varepsilon u$, where $u$ specifies the error direction and $\varepsilon > 0$ scales its size. At the exact map, the field is $b_\kappa = g_{\mathrm{PH}}(0)$.

The constants $L_P, L_C$ of Appendix C.4 bound the response to approximation error: $\|g_{\mathrm{PH}}(\varepsilon u) - b_\kappa\|_M \le \varepsilon L_P$ and $\|g_C(\varepsilon u)\|_M \le \varepsilon L_C$. Our sufficient conditions require the error direction to align strongly with the oracle field, relative to this finite-error response, and the step to be short enough. With $U = \|u\|_M$, $B = \|b_\kappa\|_M$ and the alignment $\beta = \langle u, b_\kappa\rangle_M$ of the error direction with the oracle field, they read

$$\beta > \varepsilon U(L_P + L_C), \qquad 0 < \eta < \frac{2\varepsilon[\beta - \varepsilon U(L_P + L_C)]}{(B + \varepsilon L_P)^2}. \tag{10}$$

These conditions guarantee both $Q(\varepsilon u - \eta g_{\mathrm{PH}}) < Q(\varepsilon u - \eta g_C)$ and $Q(\varepsilon u - \eta g_{\mathrm{PH}}) < Q(\varepsilon u)$, and for $\kappa = \pm 1$, reflecting $\kappa$ at the same state and step reverses both inequalities (Corollary 16). We determine the constants without observing the proposed step. When the conditions fail, the sign remains unresolved; at fixed $\eta$, arbitrarily reducing $\varepsilon$ can instead expose the quadratic cost of the nonzero oracle step. We extend the sufficient condition to finite-sample uncertainty in Appendix C.5. It does not cover trajectories with normalized step lengths or adaptive optimizers.

**Validation on new feature seeds.** We committed the oracle quantities and forecasts before evaluating 8,640 prescribed updates across 16 new feature seeds. The sufficient conditions correctly identify 1,479 joint improvements. Sharper Taylor bounds also predict cases of improvement and harm, and every resolved prediction agrees with the computed outcome. The complete results, including unresolved cases, are reported in Appendix C.4. A separate post-run check finds that all 1,479 reflected counterparts worsen both comparisons. Cases within a seed are dependent, and the cases covered by the sufficient conditions overlap the Taylor results.

## 8 Evidence across trained models and finite schedules

In trained models, we measure expected update fields, finite-gap approximations, and image quality (Table 2), aggregating scores as specified in Appendix B.4.

**Trained-state fields and finite gaps.** To test whether normalization changes the direction of the expected update, we evaluate held-out projections at 240 states trained with the linear score. We detect PH-minus-linear differences at all 240 states and non-collinear fields at 234 of them (Table 7). At one image-model checkpoint, grouping corrections by magnitude also reveals radial imbalance. A second diagnostic at that checkpoint uses its archived training loss and compares PH with a linear score of equal slope at zero; this contrast remains unresolved in the measured parameter subspace (Appendices D.3 and D.6).

Table 2: Results and their experimental or theoretical conditions. Numerical cases are clustered within shared seeds.

| Result | Conditions | Evidence |
|---|---|---|
| Stationarity and field changes | Fixed integrable law; all-floor criterion, fixed-floor tests | Exact criterion and counterexample; one of three trained networks passes all near-oracle gates (Appendix C.1). |
| Local improvement or harm | Known oracle, frozen features, quadratic risk; alignment, error and step bounds | All resolved joint sign forecasts agree with outcomes; the complete results include inconclusive cases (Section 7). |
| Image quality | Shared initializer within each series; matched-step and common-learning-rate policies | Six-seed diffusion-to-consistency tuning favors PH; three earlier continuations show policy-dependent outcomes (Appendix D.5). |

Finite-gap tests ask whether the implemented field agrees with its differential reference. For each of 24 eligible source states, a selector chooses one configuration of gap, floor, and teacher, and independent evaluation leaves all 24 selected configurations unresolved under the prescribed tolerance. In every case, the lower bound on the norm of the reference field is zero, which removes the relative allowance, while the upper bound on the difference exceeds the absolute tolerance (Appendix D.4).

**Image quality.** From one pretrained diffusion initializer, we train PH at zero floor ($c = 0$) and a squared-loss control with matched realized update lengths on six paired tuning seeds. The primary endpoint, two-step Fréchet inception distance (FID) evaluated on 50,000 generated images, is lower for PH in every pair, with a mean linear-minus-PH difference of 2.33 (95% paired-$t$ interval $[1.60, 3.06]$). A separate, consistently evaluated 10,000-image endpoint shows that both policies improve from initialization. In this comparison, each arm keeps its own adaptive moments, and a positive-floor sweep with a slope-matched linear control and shared fixed preconditioning remains untested. Three earlier continuation designs yield policy-dependent outcomes (Appendix D.5).

## 9 Scope and Limitations

Our stationarity criterion applies at a fixed joint law of corrections and Jacobians. Proposition 6 shows that a projected field contrast persists along a changing student–teacher trajectory if the accumulated state variation is sufficiently small. We have not verified its premises along the image-model runs. The learned witness meets an approximate oracle criterion; convergence under consistency training remains open.

The risk bounds require a known oracle, frozen features, and a quadratic evaluation loss, and applying them to adaptive optimizers or image quality would require additional analysis. Whether, and how much, the loss of stationarity contributes causally to the observed quality differences remains unresolved. We check numerical integration by refinement, without certified error enclosures. Intervals at trained states are approximate and conditional on the fitted states and directions; Appendix C.5 gives finite-sample bounds under explicit sampling assumptions. All runs within an image series share one pretrained initializer. The six-seed comparison uses $c = 0$, which lies outside the positive-floor premise of the all-floor theorem, and we amended one early EMA gate after the first seed (Appendix D.5.1).

We recommend checking the executed field with the teacher, floor, and optimizer settings used in training before drawing conclusions from target-flow stationarity. With fixed $a, c > 0$ and reciprocal-gap weighting, the field approaches $(a/c)\mathbb{E}[J^\top C]$ as $\delta \to 0$ under the assumptions of Appendix A.4 on correction and Jacobian convergence, weighted uniform integrability, and teacher tracking. For an EMA teacher, tracking must be checked against the shrinking gap; a tracking error of smaller order than the gap suffices. At finite gaps, field agreement and sample quality need separate assessment. The present results do not establish a general preference for normalized or linear training.

## AI USE STATEMENT

Generative AI tools assisted literature discovery, code improvement, text editing, and proofreading. The authors remain responsible for all claims and artifacts.

## REPRODUCIBILITY STATEMENT

Appendix A gives the assumptions and proofs for the theoretical results, and Appendix B specifies the computational graph and measurement procedures. Appendices C and D describe the experimental protocols, including data, model and optimizer settings, evaluation procedures, and statistical analyses. They distinguish predictions made before evaluation from retrospective analyses and report unsuccessful and inconclusive outcomes. Supporting experimental materials are available upon request, including the materials for the six-seed matched-step tuning study in Appendix D.5.1.

## REFERENCES


Jiayu Bai, Zhanbo Feng, Zhijie Deng, Tianqi Hou, Robert C. Qiu, and Zenan Ling. Adaptive Discretization for Consistency Models. NeurIPS, 2025. https://arxiv.org/abs/2510.17266.

Nicholas M. Boffi, Michael S. Albergo, and Eric Vanden-Eijnden. Flow map matching with stochastic interpolants: A mathematical framework for consistency models. TMLR, 2025. https://arxiv.org/abs/2406.07507.

Zehao Dou, Minshuo Chen, Mengdi Wang, and Zhuoran Yang. Theory of Consistency Diffusion Models: Distribution Estimation Meets Fast Sampling. ICML, PMLR 235:11592–11612, 2024. https://proceedings.mlr.press/v235/dou24a.html.

Zhengyang Geng, Ashwini Pokle, Weijian Luo, Justin Lin, and J. Zico Kolter. Consistency Models Made Easy. ICLR, 2025. https://proceedings.iclr.cc/paper_files/paper/2025/hash/bb166dd4de5dba363bf1023eb956a826-Abstract-Conference.html.

Tilmann Gneiting. Making and Evaluating Point Forecasts. Journal of the American Statistical Association, 2011. https://arxiv.org/abs/0912.0902.

Mark Goldstein, Anshuk Uppal, Raghav Singhal, Aahlad Manas Puli, and Rajesh Ranganath. Flow Map Learning via Non-Gradient Vector Flow. ICLR, 2026. https://proceedings.iclr.cc/paper_files/paper/2026/hash/f535fbd110d5b460c9955bde27b548bb-Abstract-Conference.html.

Alexander Gomilko and Yuri Tomilov. To the theory of generalized Stieltjes transforms. arXiv:2210.03474v2, 2023. https://arxiv.org/abs/2210.03474v2.

Nishant Jain, Xunpeng Huang, Yian Ma, and Tong Zhang. Multi-Step Consistency Models: Fast Generation with Theoretical Guarantees. arXiv:2505.01049v2, 2025. https://arxiv.org/abs/2505.01049v2.

Bingqing Jiang and Difan Zou. On the Memorization of Consistency Distillation for Diffusion Models. arXiv:2604.23552v1, 2026. https://arxiv.org/abs/2604.23552v1.

Youngjoong Kim, Duhoe Kim, Woosung Kim, and Jaesik Park. Stabilizing Consistency Training: A Flow Map Analysis and Self-Distillation. arXiv:2601.22679v2, 2026. https://arxiv.org/html/2601.22679v2.

Fei Kong, Jinhao Duan, Lichao Sun, Hao Cheng, Renjing Xu, Hengtao Shen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. ACT-Diffusion: Efficient Adversarial Consistency Training for One-Step Diffusion Models. CVPR, pp. 8890–8899, 2024. https://arxiv.org/abs/2311.14097.

Sangyun Lee, Yilun Xu, Tomas Geffner, Giulia Fanti, Karsten Kreis, Arash Vahdat, and Weili Nie. Truncated Consistency Models. ICLR, 2025. https://openreview.net/forum?id=ZYDEJEvCbv.

Gen Li, Zhihan Huang, and Yuting Wei. Towards a Mathematical Theory for Consistency Training in Diffusion Models. AISTATS, PMLR 258:1621–1629, 2025. https://proceedings.mlr.press/v258/li25c.html.

Cheng Lu and Yang Song. Simplifying, Stabilizing and Scaling Continuous-Time Consistency Models. ICLR, 2025. https://arxiv.org/abs/2410.11081.

Junlong Lyu, Zhitang Chen, and Shoubo Feng. Sampling is as Easy as Keeping the Consistency: Convergence Guarantee for Consistency Models. ICML, PMLR 235:33664–33685, 2024. https://proceedings.mlr.press/v235/lyu24b.html.

Andrew Patterson, Victor Liao, and Martha White. Robust Losses for Learning Value Functions. arXiv:2205.08464, 2022. https://arxiv.org/abs/2205.08464.

Amirmojtaba Sabour, Sanja Fidler, and Karsten Kreis. Align Your Flow: Scaling Continuous-Time Flow Map Distillation. NeurIPS, 2025. https://arxiv.org/abs/2506.14603.

Marvin Schmitt, Valentin Pratz, Ullrich Köthe, Paul-Christian Bürkner, and Stefan T. Radev. Consistency Models for Scalable and Fast Simulation-Based Inference. NeurIPS, 2024. https://arxiv.org/abs/2312.05440.

Gianluigi Silvestri, Luca Ambrogioni, Chieh-Hsin Lai, Yuhta Takida, and Yuki Mitsufuji. VCT: Training Consistency Models with Variational Noise Coupling. ICML, PMLR 267:55657–55683, 2025. https://proceedings.mlr.press/v267/silvestri25a.html.

Yang Song and Prafulla Dhariwal. Improved Techniques for Training Consistency Models. ICLR, 2024. https://arxiv.org/abs/2310.14189.

Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive Huber Regression. Journal of the American Statistical Association, 2020. https://arxiv.org/abs/1706.06991.

Fu-Yun Wang, Zhengyang Geng, and Hongsheng Li. Stable Consistency Tuning: Understanding and Improving Consistency Models. arXiv:2410.18958v3, 2024. https://arxiv.org/abs/2410.18958v3.

Qizhen Ying, Yangchen Pan, Victor Adrian Prisacariu, and Junfeng Wen. Temporal Difference Learning for Diffusion Models. arXiv:2606.15048, 2026. https://arxiv.org/abs/2606.15048.

Yuchen Zhang and Jian Zhou. Inverse Flow and Consistency Models. ICML, PMLR 267:77152–77176, 2025. https://proceedings.mlr.press/v267/zhang25dm.html.

# A Proofs and assumptions

## A.1 Stationarity across floors

**Proof of Theorem 2.** We prove the order-one-half case of generalized Stieltjes-transform injectivity on which Theorem 2 relies; see Gomilko and Tomilov (2023, Section 2). The reverse implication follows from the tower property. For necessity, fix a coordinate $j$ and define the finite signed measure $\nu(A) = \mathbb{E}[H_j \mathbf{1}_{\{R^2 \in A\}}]$. At an interior point $s_0 = \tau_0^2$, all derivatives of $F(s) = \int (s+x)^{-1/2}\, d\nu(x)$ vanish; differentiation under the integral is justified because $\nu$ has finite total variation and the integrand has bounded derivatives near $s_0 > 0$. Thus $\int (s_0+x)^{-k-1/2}\, d\nu(x) = 0$ for every integer $k \geq 0$. Pushing the weighted measure $(s_0+x)^{-1/2} d\nu(x)$ forward under $y = (s_0+x)^{-1}$ and extending it by zero at $y = 0$ yields a finite signed measure on the compact interval $[0, 1/s_0]$ whose polynomial moments all vanish, so it is zero by the density of polynomials. Inverting the positive weight on bounded $x$ sets and exhausting $[0, \infty)$ then gives $\nu = 0$. Since the coordinate was arbitrary, this proves the claim.

A fixed outer weight can be absorbed into $H$ when integrability holds.

**Isolated cancellation floors.** Under the hypotheses of the main-text criterion, put $F(s) = \mathbb{E}[H(R^2+s)^{-1/2}]$ and fix $s_0 > 0$. On $|h| \leq q s_0$ with $q < 1$, the binomial expansion in $h/(s_0 + R^2)$ is absolutely dominated by a constant times $\|H\|/\sqrt{s_0}$, so termwise integration makes every coordinate of $F$ real analytic. If $\mathbb{E}[H \mid R] \neq 0$ on a set of positive probability, the all-floor criterion shows that at least one coordinate is not identically zero, and the identity theorem for analytic functions makes its zeros isolated. The common zeros of the vector field are therefore isolated and finite on compact positive-floor intervals; by compact exhaustion they are countable and so have Lebesgue measure zero. This measure-zero statement refers to floors at the fixed imbalanced law. The zeros may accumulate at zero or infinity, and a nonzero field need not be appreciably large.

**A finite-floor certificate.** At the intended map, $\mathbb{E}U = 0$. Set $m_3 = \mathbb{E}[U\|U\|^2]$ and $M_5 = \mathbb{E}\|U\|^5 < \infty$. Equation (12) and the reverse triangle inequality give

$$\|\tau B_\tau\| \geq \frac{\|m_3\|}{2\tau^2} - \frac{3M_5}{8\tau^4}.$$

When $m_3 \neq 0$, the right side is positive for $\tau^2 > 3M_5/(4\|m_3\|)$; below this threshold the bound is inconclusive. Estimated moments require uncertainty control; parameter fields additionally require the projection condition below.

**Local persistence.** Suppose that $(H_\xi, R_\xi)$ are almost surely continuous at $\xi_0$ and that $\|H_\xi\|$ has a common integrable envelope on a neighborhood of $\xi_0$. Dominated convergence then makes $G(\xi, \tau) = \mathbb{E}[H_\xi/\sqrt{R_\xi^2 + \tau^2}]$ continuous for $\tau \geq \tau_0/2 > 0$. If $G(\xi_0, \tau_0) \neq 0$, some neighborhood satisfies $\|G(\xi, \tau)\| \geq \|G(\xi_0, \tau_0)\|/2$, and along a training trajectory the bound holds while the state stays in that neighborhood. For the parameter field, the premise is that $H = J^\top C$ remains imbalanced after projection.

**Proposition 6** (Persistence of a projected field contrast)**.** *Let $(\mathcal{X}, d)$ be a metric space of complete training states, and let $D(\xi) \in \mathbb{R}^p$ be a well-defined expected parameter-field contrast. Fix a unit vector $u$ and suppose $u^\top D(\xi_0) \geq m > 0$. If*

$$|u^\top [D(\xi) - D(\xi')]| \leq L d(\xi, \xi')$$

*throughout a region containing $\xi_0, \ldots, \xi_K$, then for every $k \leq K$,*

$$u^\top D(\xi_k) \geq m - L \sum_{j<k} d(\xi_{j+1}, \xi_j).$$

*For $L > 0$, cumulative displacement at most $m/(2L)$ preserves a projection of at least $m/2$. For $L = 0$ the initial lower bound persists.*

*Proof.* Telescope the projected field differences and apply the Lipschitz bound. □

For normalization, we take $D(\xi) = \mathbb{E}_\xi[wJ^\top\{S_\tau(C) - C\}]$ with a common scalar $\tau > 0$ and the same correction, Jacobian, outer weight and sampling law in both fields. The state includes student, teacher, buffers, schedule and sampling law, together with the optimizer state that governs their evolution. For an EMA teacher, $\bar\theta_{j+1} = \beta_j\bar\theta_j + (1-\beta_j)\theta_{j+1}$ gives

$$\|\bar\theta_{j+1} - \bar\theta_j\| = (1-\beta_j)\|\theta_{j+1} - \bar\theta_j\|, \qquad 0 \le \beta_j \le 1.$$

A product-state metric includes this increment and the changes in the student and other state components. Estimated initial margins require uncertainty control. The bound concerns a fixed field projection, not two diverging trajectories; its variation premise remains unverified for our image-model runs.

### A.2 Parameter projection

Because $T : \mathbb{R}^p \to L^2(\mu)$ has finite rank, its range is closed, and the normal equations give $P_{\mathrm{range}(T)}B = TK^\dagger g$, whose squared norm is $g^\top K^\dagger g$. Since $\ker K = \ker T$ and $g = T^*B$ is orthogonal to this kernel, we have $g \in \mathrm{range}(K)$; hence $g \neq 0$ implies both $g^\top K^\dagger g > 0$ and $g^\top K g > 0$. For the sufficient approximation condition, Cauchy–Schwarz yields $u^\top g = \langle Tu, B\rangle \ge (1-\kappa)\|B\|^2$. The one-step lower bound then follows from the model's Taylor remainder and the reverse triangle inequality.

Two equally weighted conditions with $B = (1, -1)$ and constant tangent $J = (1, 1)$ have $g = 0$ despite nonzero conditional scores and a positive scalar Gram matrix.

**Functional effect size of a field contrast.** At a common state, we fix the evaluation law $\mu$ and the prediction map, including its evaluation mode and buffers. Given deterministic fields $g_P, g_L$ and a shared fixed linear preconditioner $A$, let $\theta_P = \theta - \eta A g_P$ and $\theta_L = \theta - \eta A g_L$, and suppose that along both steps the $L^2(\mu)$ Taylor remainder obeys $\|f_{\theta+h} - f_\theta - Th\|_{L^2} \le M\|h\|^2/2$. Subtracting the two expansions gives

$$\|f_{\theta_P} - f_{\theta_L} + \eta TA(g_P - g_L)\|_{L^2} \le R_\eta := \tfrac{1}{2}M\eta^2(\|Ag_P\|^2 + \|Ag_L\|^2).$$

The actual separation therefore differs from $\eta\|TA(g_P - g_L)\|_{L^2}$ by at most $R_\eta$. Because $TA$ can annihilate the parameter contrast, that contrast alone gives no positive lower bound. For sampled gradients, the bound applies to the realized steps under the same remainder assumption, but the norm of an expected contrast need not equal the expected norm of the stochastic separation. The estimate covers one step with common preconditioning.

### A.3 Gap and floor limits

**Theorem 7** (Joint conditional limit)**.** *Fix $z$ and suppose $C_\delta \to C_f$ almost surely under its conditional law. For deterministic positive $\tau_\delta$:*

1. *If $\tau_\delta \to \tau_0 \in (0, \infty)$, then $\mathbb{E}\psi_{\tau_\delta}(C_\delta) \to B_{\tau_0}$.*
2. *If $\tau_\delta \to 0$ and $\Pr(C_f = 0 \mid z) = 0$, then the limit is $\mathbb{E}[C_f/\|C_f\| \mid z]$.*
3. *If $\tau_\delta \to \infty$ and the family $\{\|C_\delta\|\}$ is uniformly integrable, then $\tau_\delta\mathbb{E}\psi_{\tau_\delta}(C_\delta) \to \mathbb{E}[C_f \mid z]$.*

*For a fixed finite deterministic $J(z)$, multiply these limits by $J(z)^\top$. Theorem 10 gives the integrated stochastic extension.*

*Proof.* In the first two cases the scores converge almost surely and have norm at most one, so bounded convergence applies. In the third, $S_\tau(u) = \tau\psi_\tau(u)$ converges along the sequence to $C_f$ almost surely and $\|S_{\tau_\delta}(C_\delta)\| \le \|C_\delta\|$. Uniform integrability and convergence in probability imply convergence in $L^1$. The fixed finite Jacobian is a continuous linear map. □

On an event where $C_f = 0$, the vanishing-floor score depends on the relative rates of $C_\delta$ and $\tau_\delta$. For the growing-floor limit, uniform integrability follows from a common integrable envelope or from a uniform moment bound of order greater than one. These are tail conditions, imposed in addition to smoothness at a fixed positive floor.

**Proposition 8** (Separate signal errors)**.** *If the right-hand side is finite, then conditionally on $z$,*

$$\|\mathbb{E}S_\tau(C_\delta) - \mathbb{E}C_f\| \le \mathbb{E}\|C_\delta - C_f\| + \frac{\mathbb{E}\|C_\delta\|^3}{2\tau^2}. \tag{11}$$

*At $\mathbb{E}C_f = 0$, a finite fifth moment also gives*

$$\mathbb{E}S_\tau(C_f) = -\frac{\mathbb{E}[C_f\|C_f\|^2]}{2\tau^2} + R_\tau, \qquad \|R_\tau\| \le \frac{3\mathbb{E}\|C_f\|^5}{8\tau^4}. \tag{12}$$

*Proof.* Use $0 \le 1 - (1+s)^{-1/2} \le s/2$ for $s \ge 0$, add and subtract $\mathbb{E}C_\delta$, and apply the triangle inequality. Taylor's formula for $(1+s)^{-1/2}$ has a remainder bounded by $3s^2/8$ on $s \ge 0$, proving the expansion after multiplication by $C_f$ and integration. □

With $a, c > 0$ fixed and third moments uniformly bounded, normalization contributes an $O(\delta^2)$ error, although discretization may still contribute $O(\delta)$ under suitable regularity. A coupled floor $c = a\delta\tau_0$ instead retains the nonlinear score. Rescaling changes update speed, and condition-dependent rescaling can also change integrated cancellation.

**Corollary 9** (Include the executed outer weight)**.** *For a deterministic external weight $\lambda_\delta$, the executed field is $\lambda_\delta J^\top \psi_{\tau_\delta}(C_\delta)$. In particular, if $\lambda_\delta = 1/\delta$ and $a, c$ are fixed positive constants, then*

$$\frac{1}{\delta} J^\top \mathbb{E}\psi_{c/(a\delta)}(C_\delta) = \frac{a}{c} J^\top \mathbb{E}S_{c/(a\delta)}(C_\delta) \longrightarrow \frac{a}{c} J^\top \mathbb{E}C_f. \tag{13}$$

*The conditional limit uses Theorem 7; integration over conditions additionally requires a dominating bound including $J$. A condition-dependent gap does not change the factor $a/c$ when this ratio is constant.*

*Proof.* The identity follows from $\tau/\delta^{-1} = c/a$; then apply the rescaled conditional limit. □

In both the modified Easy Consistency Tuning (ECT) graph of Appendix B and the published CMPE objective (Schmitt et al., 2024), reciprocal-gap weighting supplies this rescaling. With a coupled floor $c = a\delta\tau_0$, however, the field is $\delta^{-1}J^\top B_{\tau_0,\delta}$, where $B_{\tau,\delta}(f; z) = \mathbb{E}[\psi_\tau(C_\delta) \mid z]$, and it may diverge. The raw, rescaled and reciprocal-gap-weighted fields therefore have different limits, and even a constant gradient scale can change finite-step adaptive updates.

**Assessing a usable gap.** Appendix B.3 gives a finite-state error budget and tolerance test; Appendix D.4 reports its evaluation.

**A sufficient envelope at a fixed network.** Consider a fixed finite tanh network $f_\theta(x,t) = x + (1-t)N_\theta(x,t)$ on an interior compact time interval. Integrating its bounded input derivatives along the backward segment gives $\|C_\delta\| \le K_\theta(1 + \|V\|)$, and its parameter Jacobian satisfies $\|J(x,t)\| \le K_\theta(1 + \|x\|)$. A $(1+\epsilon)$ moment of $(1 + \|X\|)(1 + \|V\|)$ therefore supplies weighted uniform integrability at fixed $a, c$ with reciprocal-gap weight, and the Gaussian-mixture paths satisfy these moment conditions. Extending the argument along training also requires parameter containment, a common sampling coupling and the stated convergence assumptions; stochastic layers, moving teachers and endpoint branches must be checked separately.

### A.4 Stochastic limits and teacher tracking

**Theorem 10** (Joint stochastic field, without factorization)**.** *On one probability space containing conditions, network randomness and teacher state, suppose $C_n \to C$ almost surely, $\tau_n \to \infty$ almost surely, and $M_n = (\lambda_n/\tau_n)J_n \to M$ almost surely. If $\{\|M_n\|_{\mathrm{op}}\|C_n\|\}$ is uniformly integrable, then*

$$\mathbb{E}[\lambda_n J_n^\top \psi_{\tau_n}(C_n)] \longrightarrow \mathbb{E}[M^\top C]. \tag{14}$$

*No independence is required. At fixed $a, c > 0$ and $\lambda_n = 1/\delta_n$, the limit is $(a/c)\mathbb{E}[J^\top C]$ when $J_n \to J$ and the weighted integrability condition holds.*

*Proof.* Write the integrand as $M_n^\top S_{\tau_n}(C_n)$. It converges almost surely to $M^\top C$ and its norm is bounded by $\|M_n\|_{\mathrm{op}}\|C_n\|$. Uniform integrability yields $L^1$ convergence. □

Table 3: Scope of the backward graph. Samplewise identities assume real arithmetic; trajectory equivalence requires coupled complete states and update rules (Section 3). Published objectives are compared at equation level; implementation checks are specified below.

| Setting | Exact identity | Population-limit status |
|---|---|---|
| Inspected modified ECT | Yes: shared current anchor, detached previous output, coupled randomness | Interior gaps only; endpoint replacement, numerical arithmetic and stochastic coupling require separate checks. |
| CMPE, Eqs. 8–9 (Schmitt et al., 2024) | Same-weight detached teacher, radial pseudo-Huber, $a = 1$; adding the specified anchor gives the floor control | Reciprocal-gap weight is explicit; regularity and weighted integrability remain assumptions. |
| TCM, Appendix D (Lee et al., 2025) | $a = 1$ for its pseudo-Huber specialization | Fixed-metric expansion is an agreement case; a jointly shrinking floor needs separate analysis. |
| EMA / solver target | Algebra survives for $y = (1 - a)\,\mathrm{sg}(f_0) + a\,\mathrm{sg}(T)$ if $T$ has no other $a$ dependence | Self-map interpretation requires teacher mismatch control. |
| SCT (Wang et al., 2024) | Conditional-noise averaging changes target construction; graph analysis required | Averaging before the target differs from averaging scored corrections. |
| ACT-Diffusion (Kong et al., 2024) | Adds a discriminator-dependent adversarial field | Stationarity depends on the sum of the consistency and adversarial fields. |
| AYF, LCM, other full graphs | Separate graph analysis required | Teacher targets, derivatives, reductions and schedules must be verified individually (Appendix B). |

The dependence between $J$ and $C$ decides whether the field cancels. If $J = C = \pm 1$ with equal probability, for example, then $\mathbb{E}C = 0$ but $\mathbb{E}[JC] = 1$. Cancellation is guaranteed if $J$ is measurable with respect to a sigma-field $\mathcal{H}$ with $\mathbb{E}[C \mid \mathcal{H}] = 0$.

For $\tau_n \to \tau_0(\omega) \in (0, \infty)$, assume that $K_n = \lambda_n J_n \to K$ almost surely and that $\{\|K_n\|_{\mathrm{op}}\}$ is uniformly integrable; because the scores are bounded, the field converges to $\mathbb{E}[K^\top \psi_{\tau_0}(C)]$. For $\tau_n \to 0$, assume also that $\Pr(C = 0) = 0$; the limit is then $\mathbb{E}[K^\top C/\|C\|]$. In both cases, pointwise convergence and uniform integrability give $L^1$ convergence. Random positive floors and weights are held fixed during differentiation.

**Teacher mismatch and practical schedules.** Let $e_\delta = f_- - T_-$ be the difference between the same-network and teacher predictions, so that $C_\delta^T = C_\delta + e_\delta/\delta$. At fixed $a, c$ with reciprocal-gap weight, the teacher-induced field difference is at most

$$\frac{a}{c}\mathbb{E}\left[\|J_\delta\|_{\mathrm{op}}\frac{\|e_\delta\|}{\delta}\right]. \tag{15}$$

Because the derivative of $S_\tau$ has eigenvalues at most one, the resulting Lipschitz bound inside the expectation proves Equation (15), and a weighted teacher lag of $o(\delta)$ suffices. Suppose now that $e_\delta/\delta \to b$ almost surely, $J_\delta \to J$, and the weighted correction families are uniformly integrable, with fixed $a, c > 0$ and $\tau_\delta = c/(a\delta)$. The reciprocal-gap limit then becomes $(a/c)\mathbb{E}[J^\top(C + b)]$, so a lag of the order of the gap can contribute $(a/c)\mathbb{E}[J^\top b]$. Solver error, augmentation and dropout coupling require analogous control along the schedule.

**Teacher tracking.** For $\bar\theta_k = \beta\bar\theta_{k-1} + (1 - \beta)\theta_k$ with $0 \le \beta < 1$, let $d_k = \|\theta_k - \bar\theta_k\|$ and bound parameter increments by $u$. The recurrence $d_k \le \beta(d_{k-1} + u)$ gives $d_k \le \beta^k d_0 + \beta u/(1 - \beta)$. If the predictor is $L$-Lipschitz and the Jacobian is bounded by $M$, with buffers and randomness coupled, the field discrepancy is at most $(a/c)MLd_k/\delta_k$. The sufficient condition $d_k = o(\delta_k)$ remains untested; a fixed learning rate or decay alone does not determine this relative rate.

## B EXECUTED GRAPH AND MEASUREMENT METHODS

Even when the samplewise algebra holds, independent noise in the two branches can prevent an $O(\delta)$ residual. This branch coupling differs from the data–noise coupling learned by VCT (Silvestri et al., 2025).

**Relation to other objectives.** The references in Section 2 analyze squared-regression targets (Li et al., 2025, Eq. 20, Theorem 3.3), Wasserstein discrepancies (Dou et al., 2024, Eqs. 11–16), assumed distillation accuracy (Jain et al., 2025, Eq. 12, Theorem 3.5), conditional-noise target averaging (Wang et al., 2024, Eqs. 10–14), and adversarial-plus-consistency losses (Kong et al., 2024, Section 3.3). Each analysis requires assumptions beyond a zero expected semi-gradient. The curvature calculation in Jiang and Zou (2026, Eqs. 5.1–5.3) uses full-gradient feature differences, whereas our projection uses detached targets.

**Modified ECT graph used for the identity analysis.** The modified ECT source we inspected (commit `5981c6e2138f1`; full hash in the supporting experimental materials) samples a lognormal $t$ and a stage-dependent $r$ and shares augmentation and Gaussian noise between $x_t = y + t\epsilon$ and $x_r = y + r\epsilon$. It computes the live $D_t$ once, and the same online network evaluates $D_r$ without gradients and with restored dropout RNG. The target is $aD_r + (1 - a)\,\mathrm{sg}(D_t)$, and per-image pseudo-Huber normalization carries the external weight $1/(t - r)$. The archived source-loss recipe audited in Appendix D.5 uses a different outer weight, $256/t$. Evaluation uses the EMA parameters. At $r = 0$ the clean image replaces $D_r$; this endpoint replacement and the sanitization of nonfinite values lie outside the smooth interior analysis. The toy problem has its data endpoint at 1, whereas this implementation has its clean endpoint at 0. The samplewise identity uses $\delta = t - r, V = \epsilon$.

The equivalent-floor replay restores the parameter, buffer, teacher, optimizer, EMA, scheduler and RNG states, couples every data, time, noise and mask draw, and keeps the same detached current anchor and outer weight. Coordinatewise pseudo-Huber admits the same floor substitution in each coordinate, but extending the radial population criterion to this score requires a separate analysis.

### B.1 FROM EQUAL GRADIENTS TO EQUAL UPDATES

**Proposition 11** (Pipeline transfer)**.** *Suppose two graphs produce the same unweighted gradient $q$ almost surely under a common sampling law. With detached scalar weights $w_A, w_B$, their weighted gradients agree exactly when $(w_A - w_B)q = 0$ almost surely, and, provided $w_A q, w_B q \in L^1$, their expected fields agree exactly when $\mathbb{E}[(w_A - w_B)q] = 0$. A shared deterministic optimizer map at the same complete state takes equal gradient inputs to equal updates. Matching initial complete states and every subsequent transition input gives equal trajectories.*

*Proof.* Subtract the weighted gradients and take expectations. Equal inputs to a deterministic map have equal outputs; induction through the common state transition gives trajectory equality. □

Beyond the replay states listed above, match accumulation groups, clipping order and loss-scaling state; check loss-driven schedules because forward losses may differ. Equal means need not survive nonlinear processing: a gradient equal to 2 with probability $1/3$ and $-1$ otherwise has mean zero, but after unit clipping its mean is $-1/3$. A fixed linear preconditioner preserves equal means, and an invertible one preserves the zero set.

In the all-floor criterion, a floor-independent detached weight can be absorbed into $H = wJ^\top C$ when integrability holds, and separable weights $w_\tau = b(\tau)w$ with deterministic nonzero $b(\tau)$ leave the zero criterion unchanged. Arbitrary sample-dependent floor weights, by contrast, change the joint law and require a new analysis; differentiated weights contribute product-rule terms.

### B.2 INDEPENDENT FIELD AND RADIAL-MOMENT MEASUREMENTS

Freeze a complete state $\mathcal{S}$ and its sampling law. A realization $\omega$ includes inputs, times, augmentation, masks and coupled teacher randomness. Define

$$Y_\tau = wJ^\top \psi_\tau(C), \qquad G_\tau = \mathbb{E}[Y_\tau \mid \mathcal{S}]. \tag{16}$$

For $b = w\psi_\tau(C) \in L^2$ and square-integrable $J$, the joint-space operator $\widetilde{T}u = J(\omega)u$ gives $G_\tau = \widetilde{T}^* b$ and

$$\|P_{\mathrm{range}(\widetilde{T})} b\|_{L^2}^2 = G_\tau^\top (\mathbb{E} J^\top J)^\dagger G_\tau.$$

This follows from Proposition 5 on the joint probability space. The diagnostic below estimates scalar projections without constructing the Gram matrix.

**Proposition 12** (Independent-replicate diagnostic)**.** *For iid copies $Y_1, \dots, Y_n$ with finite second moment, $\widehat{G} = n^{-1} \sum_i Y_i$ is unbiased and*

$$\mathbb{E}\|\widehat{G} - G_\tau\|^2 = \frac{\mathbb{E}\|Y_\tau - G_\tau\|^2}{n}. \tag{17}$$

*Fit unit directions independently of evaluation. Conditional on that fitting sample, let $m$ direction–floor pairs and their common evaluation count $n$ be fixed. If their scalar scores $X_{ik} = u_k^\top Y_{\tau_k,i}$ satisfy justified almost-sure bounds $|X_{ik}| \le L_k$, then with probability at least $1 - \alpha$, simultaneously,*

$$|\bar{X}_k - u_k^\top G_{\tau_k}| \le r_k, \qquad r_k = L_k \sqrt{2 \log(2m/\alpha)/n}. \tag{18}$$

*Thus $\|G_{\tau_k}\| \ge (|\bar{X}_k| - r_k)_+$. A known bound $\mathbb{E}\|Y_\tau\|^2 \le V$ alternatively gives*

$$\Pr\{\|\widehat{G} - G_\tau\| > \sqrt{V/(n\alpha)}\} \le \alpha. \tag{19}$$

*Proof.* Independence cancels cross terms in the centered squared mean. Hoeffding's inequality for scores in $[-L_k, L_k]$ and a union bound give the scalar radii; Cauchy–Schwarz gives the norm lower bound. Markov's inequality applied to the mean-square identity gives the vector bound. □

We compute $Y_\tau$ by differentiating $\langle f_\theta, \mathrm{sg}(w\psi_\tau(C))\rangle$ through the current prediction and obtain scalar projections from Jacobian–vector products (JVPs). Sampling, weights, floors, directions and multiplicity must be fixed before evaluation. The Hoeffding and variance certificates need population bounds covering the Jacobian as well as the score, and substituting a sample maximum requires additional tail assumptions. The empirical Student intervals used below are approximate.

With training-mode BatchNorm, the independent unit is a complete batch together with its averaged semi-gradient. Persistent buffers must be restored for each measurement while within-batch dependence is retained; switching to evaluation mode changes the estimand. Measurements taken along a training trajectory involve evolving states and dependent observations. At finite gaps the measured object uses $C_\delta$, and comparing it with the infinitesimal field also requires

$$\|G_{\tau,\delta} - G_\tau\| \le \tau^{-1} \mathbb{E}[|w| \|J\|_{\mathrm{op}} \|C_\delta - C\|].$$

Away from an oracle, approximation error can produce a nonzero field. Paired controls measure the score contrast, but testing non-collinearity also requires directional information. Detection power depends on how well the fitted direction aligns with the signal, and estimating detection frequencies across trained states requires independent training replications.

**Radial moments.** For floor-independent $H = wJ^\top C$ and $R = \|C\|$, fit directions $u_k$ and radial intervals $I_k$ independently. Estimate

$$m_k = \mathbb{E}[u_k^\top H \mathbf{1}_{\{R \in I_k\}}], \qquad \widehat{m}_k = n^{-1} \sum_i u_k^\top H_i \mathbf{1}_{\{R_i \in I_k\}}.$$

**Proposition 13** (Radial moment detection)**.** *Under $\mathbb{E}\|H\| < \infty$, radial balance makes every such moment zero. Conversely, vanishing coordinate moments on all radial intervals implies $\mathbb{E}[H \mid R] = 0$ almost surely. A finite collection detects some violations but cannot establish balance. For $K$ fitted scores bounded by $L$, the simultaneous radius is $r = L\sqrt{2\log(2K/\alpha)/n}$; if some $|m_k| \ge \gamma$, then $n > 8L^2 \log(2K/\alpha)/\gamma^2$ suffices for detection on the coverage event.*

*Proof.* Use the tower property and uniqueness of a finite signed measure from its values on intervals. Apply Proposition 12; $r < \gamma/2$ implies $|\widehat{m}_k| - r > 0$. □

Power drops when bins carry little mass or directions are poorly aligned. To reconstruct drift at a particular floor, one sums estimates of the weighted bin contributions $\mathbb{E}[u^\top H \mathbf{1}_{\{R \in I\}} / \sqrt{R^2 + \tau^2}]$ over a partition; multiplying separate bin averages of numerator and weight is generally biased. Nonzero radial moments can still cancel at a specific floor.

### B.3 Finite-schedule error and acceptance

Fix a complete state, a coupled sampling law and a differential reference field.

**Proposition 14** (Finite-schedule field error)**.** *Let $\tau = c/(a\delta) > 0$, $D_\delta = C_\delta + e_\delta/\delta$ and*

$$M_\delta = (\lambda/\tau)J_\delta, \quad F_\delta = \mathbb{E}[M_\delta^\top S_\tau(D_\delta)], \quad F_* = \mathbb{E}[M_*^\top C].$$

*The quantities may be random and coupled. Assume the fields and following nonnegative terms are integrable. With $\rho = \|D_\delta\|/\tau$, set*

$$\begin{aligned} A_{\mathrm{norm}} &= \mathbb{E}[\|M_\delta^\top D_\delta\|\{1-(1+\rho^2)^{-1/2}\}], \\ A_{\mathrm{gap}} &= \mathbb{E}\|M_\delta^\top (C_\delta - C)\|, \\ A_{\mathrm{teacher}} &= \mathbb{E}\|M_\delta^\top e_\delta/\delta\|, \\ A_{\mathrm{scale}} &= \mathbb{E}\|(M_\delta - M_*)^\top C\|. \end{aligned} \tag{20}$$

*Then*

$$\|F_\delta - F_*\| \le A_{\mathrm{norm}} + A_{\mathrm{gap}} + A_{\mathrm{teacher}} + A_{\mathrm{scale}}, \qquad A_{\mathrm{norm}} \le \tfrac{1}{2}\mathbb{E}[\|M_\delta^\top D_\delta\|\rho^2]. \tag{21}$$

*Proof.* For $h = (1+\rho^2)^{-1/2}$, the samplewise difference is

$$(h-1)M_\delta^\top D_\delta + M_\delta^\top (C_\delta - C) + M_\delta^\top e_\delta/\delta + (M_\delta - M_*)^\top C.$$

Take expectations and use the triangle inequality and $0 \le 1 - h \le \rho^2/2$. □

The exact attenuation term needs only a first moment of the projected correction, but its weighted tail must be controlled even when the median $\rho$ is small. Because the triangle inequality discards cancellations, the budget can be conservative. With fixed $a, c$, reciprocal-gap weighting and the same Jacobian, we have $M_\delta = M_* = (a/c)J$, so the scale term vanishes. Choosing $M_* = M_\delta$ makes the current linear field the reference. Under the coupling $c = a\delta\tau_0$, the score stays fixed and nonlinear.

**Coupled batches.** For per-example losses, set $M_{\delta,i} = (\lambda_i/\tau_i)J_i$, where $J_i$ differentiates output $i$ through the whole batch, and $h_i = (1 + \|D_{\delta,i}\|^2/\tau_i^2)^{-1/2}$. A valid batch attenuation term is

$$A_{\mathrm{norm}}^{\mathrm{batch}} = \mathbb{E}\left\| B^{-1}\sum_i (h_i - 1)M_{\delta,i}^\top D_{\delta,i} \right\| \le \mathbb{E}B^{-1}\sum_i (1-h_i)\|M_{\delta,i}^\top D_{\delta,i}\|.$$

The other terms are norms of the averaged batch vectors. The same proof then applies, retaining within-batch cancellation and the individual attenuation factors $h_i$, and independent batches serve as replicates.

Along a smooth coupled path $k(s) = f(x - sV, t - s)$ we have $C = -k'(0)$ and $\|C_\delta - C\| \le \delta \sup_{[0,\delta]} \|k''(s)\|/2$. Automatic differentiation can instead provide the paired infinitesimal correction. Endpoint-directed experiments use $k_+(s) = f(x + sV, t + s)$ and $C_+ = -k'_+(0)$. All batch inputs must be shifted together, with masks and buffers preserved. Bounding remainder and roundoff error takes more than agreement at two finite gaps.

Prespecify tolerances $\epsilon_{\mathrm{abs}}, \epsilon_{\mathrm{rel}} \ge 0$. Obtain an upper bound $U$ on the budget and lower bound $L$ on $\|F_*\|$ with joint coverage $1 - \alpha$. Then

$$U \le \epsilon_{\mathrm{abs}} + \epsilon_{\mathrm{rel}} L \tag{22}$$

certifies field tolerance on that event; near a zero reference, use absolute error. The certificate requires tail bounds, independent selection and evaluation, and multiplicity control; approximate intervals give approximate acceptance. A paired bound on the full vector may be sharper. A significant projection can refute closeness, whereas small projections and cases that fail this sufficient rule remain unresolved. Appendix D evaluates trajectories and image quality directly.

### B.4 Aggregation and natural asymmetry

Per-example, mean-residual and batch-RMS normalization are distinct executed objectives. At unit outer weight, absorbing residual-to-prediction scale into $J_i$, their gradients are

$$\begin{aligned} g_{\text{example}} &= B^{-1} \sum_i J_i^\top \psi_\tau(C_i), \\ g_{\text{mean}} &= \bar{J}^\top \psi_\tau(\bar{C}), \\ g_{\text{rms}} &= \frac{B^{-1} \sum_i J_i^\top C_i}{\sqrt{B^{-1} \sum_i \|C_i\|^2 + \tau^2}}. \end{aligned} \tag{23}$$

Here $\bar{J} = B^{-1} \sum_i J_i$ and $\bar{C} = B^{-1} \sum_i C_i$; targets are detached. For each fixed batch law, the all-floor criterion applies with the corresponding numerator and radius, and examplewise normalization uses an example selected uniformly from the random batch. Training-mode BatchNorm changes the joint law but keeps the specified examplewise, mean-residual or batch-RMS reduction.

**Executed reduction.** The retained PH studies apply radial scores to each example and then average the parameter contributions. Each norm sums the squared output coordinates before the square root is taken; for images, the sum runs over channels and pixels. Detached-score surrogates use the specified raw field without differentiating through its score, and for batch-dependent networks the complete batch remains the independent unit. The standalone learned-witness diagnostic uses $\tau\psi_\tau$ to match the linear slope, the stochastic-field measurements use $\psi_\tau$, and the image experiments differentiate the per-image loss with a detached teacher. Appendices C–D specify their outer weights, teacher conventions and optimizer policies.

## C Oracle witnesses and one-step improvement

### C.1 A standalone trained network near the oracle

For the standalone Gaussian-mixture approximation, independently sample $t \sim \text{Unif}[.2, .8]$ and $q \sim \text{Unif}[.01, .99]$, then set $x = F_t^{-1}(q)$. Evaluation uses this truncated quantile law; the full linear field is checked independently of training.

**Training and frozen fields.** We trained standalone 2–32–32–1 tanh networks with three seeds (42, 123, 777); each has 1185 trainable parameters and predicts $f_\theta(x,t) = x + (1-t)N_\theta(x/4, (t-.5)/.3)$. The oracle enters only as supervision: training minimizes the summed MSE of the map and of its $x$- and $t$-derivatives against the known quantile map on a $128 \times 32$ grid. We ran 12,000 Adam updates with batch size 1024 and learning rate .001, followed by 100 L-BFGS blocks of up to 20 internal iterations. Among the scheduled saves, we kept the checkpoint that minimizes the same loss on 2048 separate validation conditions; this choice is independent of the PH diagnostic and final evaluation. Equilibrium under consistency training remains untested.

At each frozen checkpoint, $C_\theta = f_{\theta,x} V + f_{\theta,t}$ and $J_\theta = \partial_\theta f_\theta$. We compare raw semi-gradients

$$g_L = \mathbb{E}[J_\theta^\top C_\theta], \qquad g_P = \mathbb{E}[J_\theta^\top \tau \psi_\tau(C_\theta)], \qquad \tau = .1.$$

The two scores have the same slope at zero. For a common fixed $\tau > 0$, a slope-matched raw-gradient step with learning rate $\eta$ equals a raw-PH step with learning rate $\eta\tau$, but this equivalence does not automatically extend to adaptive optimizers. If $\tau$ varies across sampled conditions, multiplying by it reweights the integrated field and need not preserve its zeros or direction. At zero floor, this scaling argument yields no slope-matched score. The vector–Jacobian product (VJP) differentiates the live prediction with the correction detached, and the graph is the differential backward-teacher limit of the same frozen network, with no finite teacher gap, dropout or BatchNorm. We compute the conditional Gaussian-mixture weights and the mean velocity $v(x,t)$ analytically, so the linear field uses $J_\theta^\top (f_{\theta,x} v + f_{\theta,t})$ exactly. For PH, we use 128 independent standard Gaussian draws per condition together with their antithetic partners, and sum exactly over mixture components. On the same antithetic draws the linear score gives that exact conditional average, and the contrasts share conditions and randomness. The PH computation has no deterministic quadrature error, although oracle evaluation and derivatives retain floating-point error.

We fix $S_{\mathrm{cal}} = [\overline{\|J_\theta\|^2 \mathbb{E}(C_*^2 \mid x,t)}]^{1/2}$, with $C_* = f_{*,x}(V - v)$, on an independent calibration bank of 2048 conditions, and we use another 2048 conditions to fix a unit direction $u$ along the estimated PH field, including its sign. Final evaluation uses 64 independent batches of 256 conditions for each seed. The approximate Student intervals (63 degrees of freedom) form one 95% Bonferroni family of $3(1185 + 7) = 3576$ scalar means, covering every linear coordinate, three PH-related projections and four squared-error and reference moments. The norm bound for the full linear field follows from the simultaneous coordinate box. For the RMSE upper bounds, we divide the upper bound on the numerator moment by the positive lower bound on the denominator before taking the square root. Map normalization uses $\mathbb{E}f_*^2$, and joint derivative normalization uses $\mathbb{E}(f_{*,x}^2 + f_{*,t}^2)$. Field norms and separation refer to this fixed network parameterization.

**Prespecified outcome and all three results.** Success requires all four gates in Table 1. The tolerances were fixed before evaluation; no threshold or checkpoint changed after final evaluation. Seed 42 misses the separation and derivative gates; seed 123 misses the derivative gate. The linear-

Table 4: Standalone trained-network diagnostic. Bounds use the same approximate simultaneous family; all three seeds are retained. $U_L$ bounds the full linear norm, $\ell_P$ bounds the PH projection from below. Both RMSE columns give normalized upper bounds. Only seed 777 passes all four prespecified gates.

| Seed | $U_L$ | $\ell_P$ | Map RMSE upper | Deriv. RMSE upper | $S_{\mathrm{cal}}$ | All gates |
|---|---|---|---|---|---|---|
| 42 | .00314145 | .01478892 | .00055402 | .01370128 | 40.4883 | No |
| 123 | .00316240 | .01646856 | .00045487 | .01095761 | 38.0997 | No |
| 777 | .00226018 | .01592450 | .00041805 | .00850560 | 45.3468 | Yes |

norm *lower* bounds are .00147468, .00120539 and .00117397: all three linear fields, including that of the successful seed 777, are distinguishable from zero.

**Oracle-correction control.** Keeping the trained $J_\theta$, we replace only the diagnostic correction by $C_*$ and denote the resulting PH field by $g_{P,*}$. In seed order, the simultaneous intervals for $u^\top(g_P - g_{P,*})$ are $[4.339, 7.589] \times 10^{-5}$, $[2.353, 5.823] \times 10^{-5}$ and $[1.543, 3.102] \times 10^{-5}$, compared with PH projection means of .01587, .01788 and .01714. Each projected difference is small but detectably nonzero. Establishing equivalence would require a prespecified margin and an assessment of the full vector. The approximate intervals are conditional on the fitted directions and states, and estimating prevalence would require a larger cohort.

### C.2 Supporting oracle controls

Table 5: Numerical values at the smooth witness $x = -0.75$, $t = 0.5$ (Section 5, Figure 1), with $a = 1$. Float64 conditional integrals with grid-refinement checks of numerical resolution.

| Quantity | Value |
|---|---|
| Normalized-sign limit $b$ | 0.21733545 |
| Analytic lower bound $0.8e^{-1/4}/\sqrt{\pi} - 0.2$ | 0.1515 |
| Coupled-floor limit $B_1$ ($c = \delta$) | 0.19181129 |
| Linear oracle signal $\mathbb{E}[C_{f_*} \mid x, t]$ | 0 |
| Conditional parameter derivative, $f_\theta = f_* + \theta(1 - t)$ | 0.09590565 |
| Gaps per floor family (including three anchors) | 17 |
| Post-hoc grid-refinement cells | 42 |

The float64 conditional integration check covers 180 cells across ten conditions, three gaps, two coefficients and three floor families. The largest residual in inverting the cumulative distribution function is $7.94 \times 10^{-15}$, and refinement changes the mean by at most $3.58 \times 10^{-9}$. These refinement errors measure numerical resolution; certified integration bounds remain unavailable.

For the network controls, $f_\theta = f_* + (1 - t)N_\theta(x, t)$, where $N_\theta$ is a 2–32–32–1 tanh network with a zero final layer. The nonlocalized test samples $t \sim \mathrm{Unif}[.2, .8]$ and $x \mid t \sim \rho_t$. For each of $\tau = .1, 1$, ten fixed hidden-network initializations are evaluated on quantile–time grids $16 \times 8$,

$32 \times 12$ and $64 \times 16$. On the finest grid, the normalized gradient norms lie in $[.02926, .04095]$ and $[.02519, .03494]$, respectively, and refinement changes them by at most $4.51\%$, within the $5\%$ threshold. The linear-oracle and centered-Gaussian controls remain stationary. A single gradient-descent step of size $10^{-3}$ gives a map root-mean-square error of $1.80 \times 10^{-5}$–$3.91 \times 10^{-5}$ and agrees with the tangent prediction to within $10^{-12}$; at initialization, 33 readout directions are active. The bias-free and fixed nonzero-readout variants pass the same refinement criterion in all 60 comparisons, whereas an orthogonal tangent cancels the signal. All of these controls embed the oracle in the prediction; standalone learned approximations are evaluated in Appendix C.1.

### C.3 Controlled radial-law intervention

The intervention in Section 7 preserves the conditional mean velocity and oracle correction magnitudes, giving $G_\kappa = \kappa G_1$. The magnitude law can change away from the oracle. At the oracle, raw correction covariance is preserved; symmetrization can change normalized-score covariance because its mean changes.

**Model, graph and controls.** The trainable model is $f_\theta = f_* + F^\top \theta$, with

$$F_j(x,t) = \frac{1-t}{\sqrt{16}} \tanh(W_{xj} x/3 + W_{tj} t + b_j), \quad j = 1, \dots, 16.$$

The weights and biases are independent standard normal draws at seeds 73101–73108 and remain frozen; only the 16 coefficients $\theta$ change. Their correction is

$$C_\theta = (f_*)_x (V_\kappa - v) + (F_t + F_x V_\kappa)^\top \theta.$$

The backward-neighbor graph gives $r = f(x,t) - \operatorname{sg} f(x - \delta V, t - \delta) \sim +\delta C$, so $\theta' = \theta - \eta \mathbb{E}[F \operatorname{score}(C)]$. The experiment uses this differential semi-gradient, with its sign checked against small backward gaps. This semi-gradient need not be stable in every tangent direction; a smaller error than a control can still mean slower deterioration from initialization.

With $w = (C_\theta^2 + .1^2)^{-1/2}$, the four scores are $C_\theta$, $wC_\theta$, $m(t)C_\theta$ and $m(z)C_\theta$, where $m(z) = \mathbb{E}[w \mid z]$ and $m(t)$ further averages over the prescribed equal-weight spatial quantiles at that time. Both means are recomputed at the current parameters of each branch. The condition-only control is stronger than the time-only one: the raw expected PH–condition-only difference is $\mathbb{E}[F\{\mathbb{E}[wC \mid z] - \mathbb{E}[w \mid z]\mathbb{E}[C \mid z]\}]$. All fields use the same learning rate without step normalization, so a zero field produces no displacement.

**Prespecified matrix and forecasts.** Training uses $t \in \{.25, .5, .65\}$ and spatial quantiles $.1, .2, \dots, .9$, giving 27 conditions. Calibration uses times $.3, .45, .6$ and test uses $.275, .475, .625$; both use ten quantiles $.05, .15, \dots, .95$. The three disjoint deterministic grids evaluate performance at fixed conditions. The reported risk is $Q(\theta) = \operatorname{mean}_{\text{test}}(F^\top \theta)^2$.

Let $b = G_{\mathrm{PH},1}(0)$ and $M = \operatorname{mean}_{\text{cal}} FF^\top$. The initial errors are zero, the two signs of $b$, the two signs of an $M$-orthogonal direction, and two seeded random directions; every nonzero start is normalized to calibration RMSE $.05$. The engineered pair tests alignment with the known drift. Even an orthogonal initial direction can have a nonzero PH–control contrast at finite error or with a finite step.

The matrix crosses eight feature seeds, seven starts, three values of $\kappa$, two learning rates $.0025, .01$ and four arms, giving 1,344 branches and 26,880 updates. The one-step PH–condition-only contrast at $.0025$ is primary; 20 steps, the larger rate and other comparators are secondary. For each seed, we saved the forecast

$$-\nabla Q_{\text{cal}}(\theta)^\top (G_{\mathrm{PH}} - G_{\text{control}})$$

before evaluating that seed's trajectories on the test grid, and treated absolute slopes below $10^{-8}$ as unresolved. No seed or endpoint was selected on the basis of its outcome. The primary designed-pair interaction, the aligned-minus-opposed PH–condition-only contrast, is negative for $\kappa = 1$ and positive for $\kappa = -1$ in all eight blocks. Both individual contrasts, however, are negative, so the anticipated sign reversal of relative benefit fails at this endpoint. Appendix C.4 then varies the error and step size to identify conditions for reversal.

Table 6: Controlled-law experiment ($10^{-5}$ MSE units; $\eta = .0025$). Descriptive means over eight frozen-feature seeds, each drawing the features independently, on a fixed 30-condition test grid; no population confidence intervals. $\Delta_h = Q_{\mathrm{PH,h}} - Q_{\mathrm{condition,h}}$, so negative values favor PH. One step is primary and 20 steps secondary. Error directions are fixed using the $\kappa = 1$ oracle drift.

| $\kappa$ | Initial error | $\Delta_1$ | $\Delta_{20}$ | $Q_{\mathrm{PH},20} - Q_0$ |
|---|---|---|---|---|
| $+1$ | Aligned | $-0.446$ | $-9.006$ | $-2.450$ |
| $+1$ | Opposed | $-0.028$ | $-0.604$ | $+5.907$ |
| $-1$ | Aligned | $-0.029$ | $-0.606$ | $+5.909$ |
| $-1$ | Opposed | $-0.446$ | $-9.004$ | $-2.451$ |

**Numerical integration.** Conditional Gaussian components are integrated in float64 over $\pm 12$ standard deviations with composite Gauss–Legendre rules, refined through orders 16, 32, 64 and 128 to a relative field-change threshold of $10^{-7}$; this refinement is an empirical precision check. Finite backward-gap checks at $10^{-4}, 10^{-5}, 10^{-6}$ verify the sign: the error at the last gap is below $10^{-4}$ and at least ten times smaller than at the first. A failed numerical pilot is retained in the development record. The eight feature seeds are the independent units, and the branch contrasts are descriptive paired means.

### C.4 Sufficient conditions for one-step improvement

We derive conditions under which one raw step reduces the fixed-grid quadratic risk relative to both initialization and the condition-only step, for a fixed correction law and frozen features.

Let $M = \mathrm{mean}_{\mathrm{test}} FF^\top \succeq 0$ and write $\langle x, y\rangle_M = x^\top My$, $\|x\|_M = \sqrt{x^\top Mx}$. Expectations defining fields and constants are over the training-condition grid and its conditional velocity law; $M$ specifies the separate, fixed evaluation grid. At each training condition $z = (x, t)$, put

$$A = F_t + vF_x, \quad H = V - v, \quad C_0 = S(f_*)_x H, \quad L_\kappa = A^\top u + S(F_x^\top u)H.$$

Then $C_{\varepsilon u} = C_0 + \varepsilon L_\kappa$, $\mathbb{E}[C_0 \mid z] = 0$ and $\mathbb{E}[C_{\varepsilon u} \mid z] = \varepsilon A^\top u$. Conditional independence of $S$ and $V$ gives

$$q_u(z) := \mathbb{E}[L_\kappa^2 \mid z] = (A^\top u)^2 + (F_x^\top u)^2 \,\mathrm{Var}(V \mid z),$$

which does not depend on $\kappa$. For a fixed floor $\tau > 0$, define constants before taking the proposed update:

$$L_P = \tau^{-1}\mathbb{E}_z[\|F\|_M \sqrt{q_u}], \qquad L_C = \tau^{-1}\mathbb{E}_z[\|F\|_M |A^\top u|]. \tag{24}$$

They are finite on the prescribed grid with the Gaussian-mixture conditional moments. More generally, their finiteness is an explicit assumption. The live-output feature Jacobian must remain fixed and deterministic; the correction is detached in the semi-gradient.

**Proposition 15** (Oracle alignment, error size and step size)**.** *Let $Q(\theta) = \|\theta\|_M^2$, $\theta = \varepsilon u$, $\varepsilon > 0$, and $b_\kappa = g_{\mathrm{PH}}(0)$. With $L_P, L_C$ from* (24)*, define $U, B, \beta$ as in Section 7. If* (10) *holds, the PH step strictly decreases $Q$ from its initial value and achieves strictly smaller $Q$ than the condition-only step at the same $\eta$.*

*Proof.* The pseudo-Huber score $\psi(c) = c/\sqrt{c^2 + \tau^2}$ is $1/\tau$-Lipschitz. The triangle inequality in the $M$-seminorm and conditional Cauchy–Schwarz yield

$$\|g_{\mathrm{PH}}(\varepsilon u) - b_\kappa\|_M \le \frac{\varepsilon}{\tau}\mathbb{E}_z[\|F\|_M \mathbb{E}(|L_\kappa| \mid z)] \le \varepsilon L_P.$$

Since $w(c) = (c^2 + \tau^2)^{-1/2} \le 1/\tau$ and the conditional mean correction is $\varepsilon A^\top u$, $\|g_C(\varepsilon u)\|_M \le \varepsilon L_C$. Consequently

$$\langle \varepsilon u, g_{\mathrm{PH}} - g_C\rangle_M \ge \varepsilon\beta - \varepsilon^2 U(L_P + L_C), \qquad \|g_{\mathrm{PH}}\|_M \le B + \varepsilon L_P.$$

The exact quadratic-risk identity gives

$$\begin{aligned} Q(\varepsilon u - \eta g_{\mathrm{PH}}) - Q(\varepsilon u - \eta g_C) &= -2\eta\langle \varepsilon u, g_{\mathrm{PH}} - g_C\rangle_M + \eta^2(\|g_{\mathrm{PH}}\|_M^2 - \|g_C\|_M^2) \\ &\le -2\eta\varepsilon[\beta - \varepsilon U(L_P + L_C)] + \eta^2(B + \varepsilon L_P)^2 < 0. \end{aligned}$$

For the absolute change, omit $g_C$; its upper bound is no larger because $L_P \le L_P + L_C$. The same step condition therefore also implies $Q(\varepsilon u - \eta g_{\mathrm{PH}}) < Q(\varepsilon u)$. □

**Corollary 16** (Conditional benefit-to-harm reversal)**.** *Suppose Proposition 15 holds for one sign $\kappa \in \{-1, 1\}$. Replacing $\kappa$ by $-\kappa$ while keeping $\theta = \varepsilon u$ and $\eta$ unchanged makes the PH step strictly worse than both the reflected condition-only step and the initial state.*

*Proof.* Reflection negates $b_\kappa$ and preserves $B, L_P, L_C$; moreover $L_P \geq L_C$ because $q_u \geq (A^\top u)^2$. Write $m = \beta - \varepsilon U(L_P + L_C) > 0$ for the favorable law's margin, and $g_P^-, g_C^-$ for the fields recomputed under the reflected law. Then

$$Q(\varepsilon u - \eta g_P^-) - Q(\varepsilon u - \eta g_C^-) \geq 2\eta\varepsilon m - \eta^2\varepsilon^2 L_C^2 > 0,$$
$$Q(\varepsilon u - \eta g_P^-) - Q(\varepsilon u) \geq 2\eta\varepsilon(\beta - \varepsilon U L_P) + \eta^2 \|g_P^-\|_M^2 > 0.$$

The first strict sign follows from $\varepsilon L_C \leq B + \varepsilon L_P$ and the step bound in (10); the second follows directly from $m > 0$. The condition-only comparator is recomputed after reflection; its bound is unchanged. □

**Scope of the sufficient region.** Fix a direction with $\beta > 0$ and $B > 0$. For sufficiently small $\varepsilon$, alignment holds and a positive admissible step exists. Near zero, the displayed bound scales as $\eta = O(\varepsilon)$ with a restricted coefficient, and choosing $\eta = o(\varepsilon)$ is asymptotically sufficient. At $\varepsilon = 0$, however, the PH update incurs $\eta^2\|b_\kappa\|_M^2$ while the condition-only field is zero, so a fixed nonzero step cannot yield absolute improvement arbitrarily close to the oracle. Outside this possibly conservative region the sign remains unresolved; the result applies only to the specified quadratic risk and update rule.

**Sharper predictions from the oracle.** To assess conservatism without fitting to finite-error outcomes, we also expand the fields at the oracle. Let $c_* = (f_*)_x H$, $w_* = w(c_*)$ and $\psi'_* = \tau^2(c_*^2 + \tau^2)^{-3/2}$. The oracle derivatives are

$$P_\kappa = \mathbb{E}_z\left[F\left\{\mathbb{E}[\psi'_* \mid z]A^\top + \kappa\mathbb{E}[\psi'_* H \mid z]F_x^\top\right\}\right],$$
$$T_C = \mathbb{E}_z[F\mathbb{E}[w_* \mid z]A^\top].$$

Use $\widehat{g}_P = b_\kappa + \varepsilon P_\kappa u$ and $\widehat{g}_C = \varepsilon T_C u$. The global derivative bounds $\sup|\psi''| = 48/(25\sqrt{5}\,\tau^2)$ and $\sup|w'| = 2/(3\sqrt{3}\,\tau^2)$ imply

$$\|g_P - \widehat{g}_P\|_M \leq e_P = \frac{24\varepsilon^2}{25\sqrt{5}\,\tau^2}\mathbb{E}_z[\|F\|_M q_u],$$
$$\|g_C - \widehat{g}_C\|_M \leq e_C = \frac{2\varepsilon^2}{3\sqrt{3}\,\tau^2}\mathbb{E}_z[\|F\|_M |A^\top u|\sqrt{q_u}].$$

For either arm $j$, writing $\widehat{Q}_j = \|\varepsilon u - \eta\widehat{g}_j\|_M^2$, its risk differs by at most

$$R_j = 2\eta\|\varepsilon u - \eta\widehat{g}_j\|_M e_j + \eta^2 e_j^2.$$

Thus $\widehat{Q}_P + R_P < \widehat{Q}_C - R_C$ suffices for relative improvement, whereas $\widehat{Q}_P + R_P < Q(\varepsilon u)$ suffices for absolute improvement. Both bounds can be evaluated before taking the proposed step and hold for exact defining expectations. The numerical implementation adds a margin based on quadrature refinement; its integration error lacks a rigorous enclosure.

**New-seed numerical validation.** We froze 16 new feature seeds, 104201–104216, which exclude development seed 104299 and all earlier feature cohorts. The model is unchanged, with width 16 and floor .1, and we kept the original training, calibration and test grids. We used four directions: aligned, opposed and the two seeded random directions. As in the original experiment, each base direction $u$ has calibration RMSE .05, so $\varepsilon = 1$ retains that error scale. For each direction, $\kappa \in \{-1, 0, 1\}$ and

$$\varepsilon \in \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, .01, .03, .1, .3, 1\}, \qquad \eta \in \{.0025, .01\varepsilon, .1\varepsilon, \varepsilon, 10\varepsilon\}.$$

This gives 1,728 field cases and 8,640 proposed raw updates. The feature seeds are the independent replications; grid cases within a seed are dependent. Directions use training and calibration quantities, and both directions and endpoints are fixed independently of the evaluation matrix $M$.

For each seed, oracle quantities, constants and predictions were checksummed before computing its full PH and condition-only fields and both actual risks for every proposed step. No direction, amplitude, step or seed was replaced after observing outcomes.

The sufficient region (10) is nonempty for every seed; all 1,479 accepted updates improve both comparisons. A post-run check of Corollary 16 on the same saved grid, with the initial state and step held fixed, finds that all 1,479 reflected counterparts worsen both comparisons. The table also reports oracle-Taylor forecasts: harm requires reversed, strictly separated bounds; overlapping bounds are unresolved. Joint harm means worsening against both initialization and the condition-only step.

| Pre-step diagnostic | Forecast | Cases | Confirmed | Inconclusive |
|---|---|---|---|---|
| Sufficient conditions | Both improve | 1,479 | 1,479 | — |
| Taylor bounds | Both improve | 2,143 | 2,143 | — |
| Taylor bounds | Both worsen | 2,563 | 2,563 | — |
| Taylor bounds | No joint sign | 3,934 | — | 3,934 |

The sufficient-condition row overlaps the Taylor rows. Joint comparisons are against initialization and condition-only, so a jointly inconclusive case can still have a resolved individual comparison. Both risk bands contain their respective outcomes for all 8,640 updates. At $\varepsilon = 1$, neither diagnostic identifies jointly improving updates despite observed improvements, and the original failed reversal is unaffected.

**Numerical controls.** Oracle fields and derivatives are computed with float64 quadrature of orders 32, 64, 128 and 256 and refinement thresholds $10^{-12}$ and $10^{-9}$, while the constants come from analytical mixture moments. Taylor radii include ten times a propagated proxy for refinement error, and risk separation must exceed $10^{-12}Q(\varepsilon u) + 10^{-18}$. These thresholds were set before evaluation, and integration error is assessed by refinement. An independent adaptive quadrature and a 24-case/120-step audit agree to within $2.26 \times 10^{-17}$ in fields and $6.60 \times 10^{-17}$ in risks and contrasts. For the first new seed at aligned error $\varepsilon = 10^{-6}$, the fixed step .0025 worsens risk, whereas the prescribed step $\eta = \varepsilon$ improves both comparisons. Restart and forecast-order checks also pass.

**Relation to the original failed prediction.** Fix the original state $\theta$, law $\kappa$ and step $\eta$. Set $d_\kappa(\theta) = g_{\mathrm{PH},\kappa}(\theta) - g_{C,\kappa}(\theta)$ and $b_\kappa = d_\kappa(0)$; here $g_{C,\kappa}(0) = 0$. The quadratic risk gives the exact decomposition

$$\begin{aligned}\Delta Q_\kappa = &-2\eta\langle\theta, b_\kappa\rangle_M - 2\eta\langle\theta, d_\kappa(\theta) - b_\kappa\rangle_M \\ &+ \eta^2\big(\|g_{\mathrm{PH},\kappa}(\theta)\|_M^2 - \|g_{C,\kappa}(\theta)\|_M^2\big).\end{aligned} \tag{25}$$

Here $\Delta Q_\kappa = Q(\theta - \eta g_{\mathrm{PH},\kappa}) - Q(\theta - \eta g_{C,\kappa})$. The terms are oracle alignment, finite-error response and finite-step curvature. Reflection negates $b_\kappa$; it need not negate either remaining term at fixed nonzero $\theta$.

Across the $8 \times 2 \times 2$ saved seed–orientation–reflection cases, the full local slope of the contrast is negative in all 32. In each of the 16 cases where the oracle slope $-2\langle\theta, b_\kappa\rangle_M$ is positive, the finite-error slope is negative and larger in magnitude. The failure is therefore already present as $\eta \to 0$ at the original error, independently of the quadratic step term.

For the original mixture label $L$ and PH weight $w$, the law of total covariance gives

$$d_\kappa(\theta) = \mathbb{E}_z[F\,\mathrm{Cov}(w, C_\theta \mid z)] = d_W(\theta) + d_B(\theta),$$

where $d_W = \mathbb{E}_z[F\mathbb{E}[\mathrm{Cov}(w, C_\theta \mid z, L) \mid z]]$ and $d_B = \mathbb{E}_z[F\,\mathrm{Cov}(\mathbb{E}[w \mid z, L], \mathbb{E}[C_\theta \mid z, L] \mid z)]$. In a retrospective virtual replacement, we subtract $d_B(\theta) - d_B(0)$, or analogously the within-component change, from the PH field while keeping its oracle field and the condition-only comparator. A seed counts as reversing only if reversal occurs in both designed orientations, for the local slopes and for steps at .0025, .01 (sign threshold $10^{-10}$). By this criterion, reversal occurs in 0/8 original seeds, 4/8 after removing the between-component change and 8/8 after removing the within-component change. Because these virtual field replacements also change the quadratic step term, the counts leave causal attribution unresolved. A separate equal-norm policy reverses the unmodified comparison in 7/8 seeds.

### C.5 FINITE-SAMPLE SUFFICIENT CONDITIONS

The original experiment uses grid-defined $M$ and analytically computed $L_P, L_C$. The following extension adds sampling uncertainty through confidence sets.

**Proposition 17** (Sufficient conditions with uncertainty)**.** *Retain the fixed-feature assumptions of Proposition 15. Suppose estimates $\widehat{b}$, $\widehat{M} \succeq 0$ and nonnegative bounds satisfy, on a joint event $\mathcal{E}$ of probability at least $1-\alpha$,*

$$\|b_\kappa - \widehat{b}\|_2 \le r, \quad \|M - \widehat{M}\|_{\mathrm{op}} \le \rho, \quad L_P \le \overline{L}_P, \quad L_C \le \overline{L}_C.$$

*Set $M_+ = \widehat{M} + \rho I$, $\Lambda = \lambda_{\max}(M_+)$, and*

$$\begin{aligned}
\overline{U} &= \sqrt{u^\top M_+ u}, \qquad \overline{B} = \sqrt{\widehat{b}^\top M_+ \widehat{b}} + \sqrt{\Lambda}\, r, \\
\underline{\beta} &= u^\top \widehat{M}\widehat{b} - \|\widehat{M}u\|_2 r - \rho\|u\|_2(\|\widehat{b}\|_2 + r), \\
\underline{m} &= \underline{\beta} - \varepsilon\overline{U}(\overline{L}_P + \overline{L}_C).
\end{aligned}$$

*On $\mathcal{E}$, the conditions*

$$\underline{m} > 0, \qquad 0 < \eta < \frac{2\varepsilon \underline{m}}{(\overline{B} + \varepsilon \overline{L}_P)^2} \tag{26}$$

*guarantee that the exact expected-field PH step has smaller $Q_M$ than both initialization and the condition-only step.*

*Proof.* On $\mathcal{E}$, $M \preceq M_+$, so $U \le \overline{U}$ and the triangle inequality gives $B \le \overline{B}$. Writing $e = b_\kappa - \widehat{b}$, the identity $u^\top M b_\kappa - u^\top \widehat{M}\widehat{b} = u^\top \widehat{M} e + u^\top (M - \widehat{M}) b_\kappa$ gives $\beta \ge \underline{\beta}$. Thus the lower margin and upper denominator imply (10). Positive definiteness of $M$ is unnecessary. □

The probability of issuing a false certificate for an expected-field raw step is at most $\alpha$; failure to pass is inconclusive. This probability is unconditional over the estimation sample. Coverage conditional on passing, and guarantees for minibatch or adaptive-optimizer steps, require additional analysis.

**A computable bounded-score construction.** Let $p$ be the number of frozen features. For iid oracle draws $X_i = F(z_i)\psi_\tau(C_{0,i})$, $\widehat{b} = n_b^{-1}\sum_i X_i$. Justified coordinate bounds $|F_j(z)| \le a_j$ give, by Hoeffding and a union bound,

$$|b_{\kappa,j} - \widehat{b}_j| \le c_j := a_j\sqrt{2\log(2p/\alpha_b)/n_b}, \qquad r = \Big(\sum_j c_j^2\Big)^{1/2}.$$

The Gaussian velocity is unbounded, but $|\psi_\tau| \le 1$. On the original fixed grid, take $\widehat{M} = M$, $\rho = 0$, and exact $L_P, L_C$. A sharper use of the same coordinate box is

$$\underline{\beta} = u^\top M\widehat{b} - \sum_j |(Mu)_j| c_j, \qquad \overline{B} = \|\widehat{b}\|_M + \min\{\sqrt{\lambda_{\max}(M)}\|c\|_2, \sqrt{c^\top |M| c}\},$$

where $|M|$ is entrywise absolute value. These bounds also imply (26) with $\overline{U} = U$.

With known $M, L_P, L_C$, one coordinate-box event covers all prescribed directions and steps. Allocate error probability across seeds and budgets; oracle and floating-point errors require separate control, and population risk requires a separate estimate.

**Retrospective finite-sample application.** We replace only the oracle integral for $b_1$ by iid samples from the specified condition grid and Gaussian conditional velocity law, retaining the known oracle, grid-defined $M$ and analytical $L_P, L_C$. The sampling budgets and analysis are fixed before these new draws; the original outcomes are already known. With $p = 16$, we use $a_j = \max_z |F_j(z)|$ over the complete finite training grid and replace $\log(2p/\alpha_b)$ by $\log(2 \cdot 16 \cdot p \cdot 4/.05)$ to cover all 16 seeds and four nested budgets simultaneously. The same box covers all prescribed directions, errors and steps; $b_\kappa = \kappa b_1$ handles reflection without additional estimation.

| Samples per seed | $10^3$ | $10^4$ | $10^5$ | $10^6$ |
|---|---|---|---|---|
| Updates passing the conservative condition | 0 | 133 | 1,105 | 1,270 |

Among the same 8,640 updates, every accepted case belongs to the original 1,479 sufficient-condition predictions and improves both comparisons. Every seed has passing cases at $10^5$ and $10^6$ samples. Doubling $L_P, L_C$ gives counts 0, 121, 996 and 1,092 as a deterministic sensitivity check. The oracle, grid $M$ and moment bounds remain known; repeated-sampling coverage and quadrature certification were not evaluated.

# D Experimental protocols and complete cohort results

## D.1 Executed graphs and deterministic cohorts

**Tasks and models.** The toy cohorts use $X_t = (1-t)Z + tY$, with independent $Z \sim N(0, I_d)$ and $t \sim \mathrm{Unif}[.05, .95]$. In the product task, the coordinates of $Y$ are independent draws from $.8N(1/\sqrt{5}, 1/5) + .2N(-4/\sqrt{5}, 1/5)$, with variances as second arguments. The full-rank task is a $2d$-component Gaussian mixture constructed from means along seeded orthogonal directions, unequal weights $.9/d$ and $.1/d$ for opposite directions, and a shared diagonal covariance; the mixture is centered and scaled to $\mathbb{E}\|Y\|^2 = d$. The task construction and cohort code are included in the supporting experimental materials available upon request. Cohort A crosses these two tasks with $d \in \{8, 32\}$ and tanh or SiLU activations, using two hidden layers of width 256 in $f_\theta(x,t) = x + (1-t)N_\theta(x,t)$. Its PDE-risk is $d^{-1}\mathbb{E}\|f_{\theta,t} + D_x f_\theta v\|^2$.

**Teacher orientation.** Cohorts A and B use the detached teacher $f(x + \delta V, t + \delta)$ with endpoint $f(x, 1) = x$, whereas the main graph uses $f(x - \delta V, t - \delta)$. The two corrections converge to $-C_f$ and $C_f$, respectively, so odd scores share the limiting zero set. Finite-gap fields and dynamics depend on teacher orientation; these measurements apply to the endpoint-directed graph. Backward-teacher stability and the frozen-feature conditions of Section 7 require separate tests.

**Scores and reduction.** Cohort A compares $\tau\psi_\tau(C)$ with the linear score $C$; both cohorts use $\tau = .1\sqrt{d}$ and average parameter contributions over examples. In cohort A, the PH loss is $\tau\sqrt{\|r\|^2 + (\delta\tau)^2}$ and the linear loss is $\overline{\|r\|^2}/(2\delta)$, where $r = \delta C$. In cohort B, two independent conditional draws give weights $w_j = \tau/\sqrt{\|C_j\|^2 + \tau^2}$; PH uses $(w_1C_1 + w_2C_2)/2$, while condition-only uses $(w_1C_2 + w_2C_1)/2$. This cross-weighting estimates the conditional mean-weight control.

**Detection and short continuations.** Cohort A consists of 480 residual-MLP training runs with a fixed budget of 10,000 steps, arranged as 240 paired blocks in eight prespecified task/dimension/activation strata. The frozen directional diagnostic detects 72 cases ($15.0\%$; paired stratified bootstrap 95% interval $[13.54, 16.46]\%$), which falls short of the proposed majority endpoint. All 72 detections occur in linear-trained models; under this diagnostic, the PH-trained arm has none. Every model improves the specified PDE-risk estimate, and the linear branch has lower final risk in 214/240 pairs. Collinearity and radial balance remain unresolved for the nondetections in this plane, and we did not certify stationarity at these finite-budget states.

Cohort B continues 160 prespecified states for 100 updates with four paired streams per state, giving 640 PH/condition-only comparisons in 80 independent source blocks. The mean PH-minus-condition-only risk is $5.3343 \times 10^{-6}$ (stratified block-bootstrap 95% interval $[4.6301, 6.0329] \times 10^{-6}$). Mean risk changes are $-7.8007 \times 10^{-5}$ for PH and $-8.3342 \times 10^{-5}$ for condition-only, so both arms improve, condition-only slightly more. Under its chosen reference, a secondary mean-field decomposition attributes more of the local contrast to scale than to direction. Because step norms vary in this comparison, the causal contribution of each component remains unresolved.

## D.2 Joint stochastic fields at 240 trained states

For each of twelve configurations, we trained twenty sources for 10,000 linear squared-error updates at learning rate $3 \times 10^{-4}$. Validation risk, measured with 32 batches per run, decreases in every run,

Table 7: Joint stochastic-field cohort: 20 independently trained sources per row. Counts report detected PH fields, differences from linear fields, and non-collinearity in the fitted plane. Detection frequencies describe these configurations; interval coverage is assessed per state.

| Task | Layer mode | Width | PH | Difference | Direction |
|---|---|---|---|---|---|
| Product | Base | 288 | 20 | 20 | 20 |
| Product | Dropout | 288 | 20 | 20 | 20 |
| Product | BatchNorm | 288 | 20 | 20 | 20 |
| Product | Base | 992 | 20 | 20 | 15 |
| Product | Dropout | 992 | 20 | 20 | 19 |
| Product | BatchNorm | 992 | 20 | 20 | 20 |
| Full-rank | Base | 288 | 20 | 20 | 20 |
| Full-rank | Dropout | 288 | 20 | 20 | 20 |
| Full-rank | BatchNorm | 288 | 20 | 20 | 20 |
| Full-rank | Base | 992 | 20 | 20 | 20 |
| Full-rank | Dropout | 992 | 20 | 20 | 20 |
| Full-rank | BatchNorm | 992 | 20 | 20 | 20 |

and the average reduction per configuration ranges from 9.03% to 30.91%. The recipe, including the learning-rate adjustment for BatchNorm, was fixed during development.

Both tasks have $d = 32$ in this cohort. The residual predictor has two tanh hidden layers of width 288 or 992; dropout has probability .1 after each hidden activation, and BatchNorm precedes the activation. Training uses SGD and batches of 256. The validation risk here is the mean squared finite-gap correction per coordinate, rather than cohort A's PDE-risk.

At each final state, we selected two probe directions on 32 fitting batches, independently of the 2048 evaluation batches of size 256. The endpoint-directed correction uses gap $10^{-3}$, floor $\tau = .1$, one mask realization shared by the current network and the teacher, and the complete training-mode batch graph. Approximate Student intervals at a per-state family level of 95% detect a nonzero PH projection and a nonzero PH-minus-linear projection at every state. With an approximate 99% jackknife interval, the cross-batch determinant statistic detects non-collinearity in the fitted plane at 234/240 states (Table 7).

The determinant uses products from different batches to exclude within-batch covariance, with delete-one jackknife errors. A significant projected determinant establishes non-collinearity; the six nondetections leave the full fields unresolved. Intervals are conditional on each state's independently fitted probes, and raw differences include scale and approximation error. Because the graphs and estimators differ, this cohort's 97.5% detection rate cannot replace cohort A's 15%.

**Calibration and uncertainty.** Monte Carlo checks on finite banks of 2048 batches passed the registered raw-field and null-rejection tolerances, but the direction-power criteria failed in some development configurations. The directions may be nonorthogonal, and near-zero fields can make the determinant approximations degenerate. Across the twelve empirical-bank calibrations, the largest simultaneous 99% upper bounds are 6.283% for raw-field null rejection and 1.876% for collinear-direction null rejection, and the smallest lower bound on raw-field shift power is 99.836%. Coverage outside these empirical gradient banks remains untested.

### D.3 Radial structure at a real image-model checkpoint

At one public pretrained consistency-model (ECM) checkpoint, we split 512 batches (8192 examples) into 64 independent groups. The coarse graph, which differs from the archived source recipe, uses $t \to t/2$, weight $2/t$ and whole-image residual norms. We measure moments of $wJ^\top r$ in eight radial bins and sixteen parameter blocks, where $w = 2/t$ is the outer weight, $r$ the residual and $J$ the Jacobian of the live prediction. The combined global/block test rejects zero binwise moments (minimum adjusted $p = .002$), conditional on sign symmetry, which is unverified. Separately, fifteen of the sixteen projections fitted on A0 exclude zero under approximate simultaneous intervals (Figure 2).

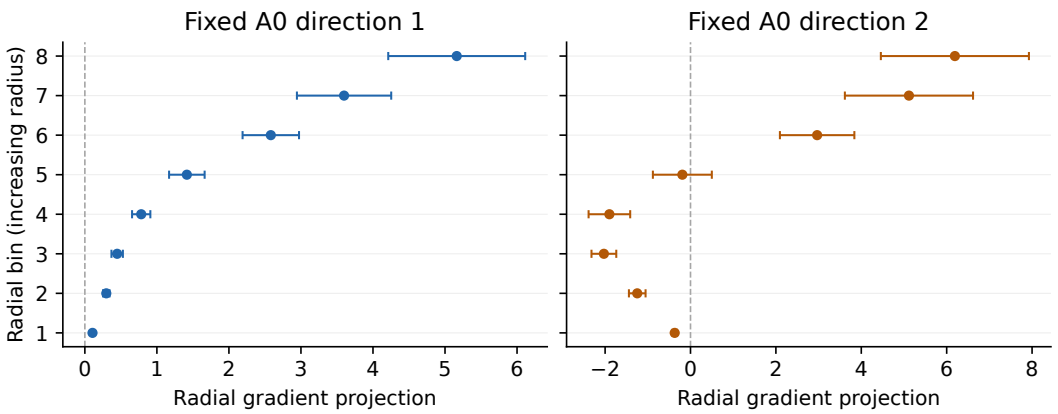



Figure 2: Eight radial bins and two directions fixed on a separate A0 sample at one public ECM checkpoint. Bars are approximate Student intervals with Bonferroni correction for sixteen projections (family 95%), using 64 groups and 8192 examples. Fifteen exclude zero. The projection directions need not be orthogonal.

A stress check with 5000 trials per law passed the registered 7.5% upper-bound tolerance for Gaussian, symmetric sparse and skew sparse null laws; an earlier, smaller audit had failed. The observed rejection counts are 233, 223 and 140 out of 5000, with simultaneous 99% upper bounds of 5.60%, 5.38% and 3.55%. Validity at the 5% level under other laws, and sign symmetry of real gradients, remain unverified.

### D.4 Finite-gap agreement and short trajectories

The prespecified matrix evaluates 24 gap/floor/teacher configurations at each of 48 source states, and the independent phases A1 and A2 each use 2048 whole batches per state. Under the registered tolerance (absolute plus 5% relative), A2 classifies 110/1152 conditions as close, 499 as incompatible and 543 as unresolved; these grid conditions share the 48 source states. With the current teacher and the quantile floor, 2, 32 and 40 of the 48 states are close at gaps .01, .003, .001, respectively. The current-teacher, fixed-floor setting at .001 is close in 36/48, and no lag-10 setting is close.

A1 accepts 80 configurations across 24 states; the independent A2 phase classifies 48 of these as close, 32 as unresolved and none as incompatible. The other 24 states have no A1-accepted candidate. The fixed tie-break selects one candidate per eligible state, preferring larger gaps. All 24 selected candidates are unresolved on independent A2 evaluation. Confidence calculations carry conditional Gaussian-group and approximate-CLT limitations, with families defined per state and per phase.

All selected candidates use the current teacher and the quantile floor: 16 use gap .01, four .003 and four .001. In every selected A2 report, the lower bound on the reference norm is zero and the upper bound on the difference exceeds $\epsilon_{\mathrm{abs}}$, which prevents acceptance. Ten of the 24 point estimates of the difference norm fall below $\epsilon_{\mathrm{abs}}$, but their upper bounds do not. All 24 lower bounds on the difference are positive, yet none establishes incompatibility with the prescribed tolerance. A1 screening is governed by the triangle-inequality budget, whereas A2 uses a direct bound on the field difference, so the unresolved A2 decisions cannot be attributed solely to the A1 budget. Lag-10 mismatch cannot explain this current-teacher subset. The reports do not establish that additional samples would certify the selected settings.

The short continuation study covers twelve source states, SGD and Adam, and two streams per optimizer, giving 48 blocks and 120 branches of 100 updates. Six states have an A1-selected candidate; the others have only baseline and reference branches. After averaging streams within each state, the selected policy has a smaller mean squared output discrepancy from the reference than the baseline in one of the six eligible states and a larger one in the other five, for both optimizers. For SGD, the cohort mean falls from $5.580 \times 10^{-8}$ to $3.147 \times 10^{-8}$, driven by the single improving state; for Adam, it rises from $3.010 \times 10^{-6}$ to $3.134 \times 10^{-6}$.

### D.5 Image quality across tuning and continuation designs

Six-seed EDM tuning and three earlier ECM continuation designs each use their own initializer, paired control and evaluation protocol; absolute scores are comparable only within a protocol.

#### D.5.1 Six-seed matched-step tuning from EDM

**Design.** With six seeds (101–106), we tuned the public EDM VP diffusion checkpoint on unconditional CIFAR-10 using the one-hour ECT recipe of Geng et al. (2025): DDPM++ ECMPrecond, batch 128, RAdam at $10^{-4}$ ($\beta = (.9, .999)$), dropout $0.2$, EMA $0.9993$ initialized from the EDM weights, lognormal $t$ ($-1.1$, 2), $q = 256$ with sigmoid mapping, no augmentation, and 8000 updates ($1.024$ Mimg, ECT's first evaluation point). Write $t_-$ for the earlier teacher time and $r$ for the image residual. For each seed, we trained two arms from the same state: native examplewise normalization with $c = 0$ (loss $\|r\|/(t - t_-)$) and a linear-correction control (loss $\frac{1}{2}\|r\|^2/(t - t_-)$) with the same detached teacher. At every update the two arms share the batch, $t$, noise and dropout masks; we verified this pairing, and exact resumption, before running the series. After each RAdam step, we rescaled the realized parameter displacement of the linear arm in float64 to the norm of the PH displacement (relative tolerance $5 \times 10^{-4}$, bisection fallback, unmatched steps retained and recorded). All 48,000 linear updates meet the tolerance, the smallest cosine between intended and realized linear steps is $0.99996$, and neither arm produces a nonfinite gradient. Each arm accumulates optimizer moments from its own gradients, and matching displacement norms still leaves differences in direction, preconditioning and per-example weighting. The reported arithmetic difference from the ECT recipe is the use of TF32 matrix operations. We chose $c = 0$ after a residual check: at $q = 256$, 82% of residual norms at a trained ECM state lie below $0.1$, so a positive floor of $0.1$ would weaken normalization for many examples. This zero-floor score, however, lies outside the positive-floor premise of the stationarity theorem.

**Evaluation and registration.** We evaluated the EMA weights under the ECT one-step ($t_{\max} = 80$) and two-step (intermediate $t = 0.821$) samplers, with common latents and the Inception detector used throughout. Seeds, endpoints, analysis and the interpretation rule were fixed before the first update. The primary endpoint is the paired difference, linear minus PH, in two-step FID-50k against the standard CIFAR-10 training statistics at update 8000, analyzed by a paired $t$ interval and a sign-flip test that enumerates all $2^6$ sign assignments. The sign-flip test is exact only if the paired differences are independently sign-symmetric under the null, an assumption beyond pairing, and the paired-$t$ interval likewise relies on its sampling assumptions for six seeds. Secondary endpoints are one-step FID-10k, FID and unbiased polynomial kernel inception distance (KID) against the 10,000 CIFAR-10 test images, and FID-10k at updates 2000 and 4000. No choice in this tuning study used the test images, although images 2000–9999 also appeared in the radial diagnostic. As an end-to-end check, the PH arm reaches the published one-hour quality regime (two-step FID 2.68–2.76; ECT reports that this recipe matches consistency distillation at 2.93).

We replaced one pre-registered early gate after the first seed. The gate required EMA FID-10k below 20 at update 2000, but at that point the EDM weights still made up 25% of the EMA (EDM two-step FID-10k 162.1); online weights at the same state gave $8.05$ (PH) and $11.29$ (linear). Training, evaluation and analysis were unchanged, and at update 8000 the EDM coefficient in the EMA is $0.4\%$.

| Seed | 101 | 102 | 103 | 104 | 105 | 106 |
|---|---|---|---|---|---|---|
| PH, two-step FID-50k | 2.756 | 2.679 | 2.706 | 2.707 | 2.732 | 2.724 |
| Linear, two-step FID-50k | 4.625 | 5.790 | 4.431 | 4.980 | 5.995 | 4.467 |
| Linear minus PH | 1.869 | 3.111 | 1.724 | 2.274 | 3.263 | 1.742 |

Table 8: Matched-step tuning from one EDM initializer, update 8000; lower FID is better. Mean difference 2.33 (SD $0.69$, 95% paired-$t$ interval $[1.60, 3.06]$), positive in all six seeds (two-sided sign-flip $p = 1/32$, the smallest attainable value; validity requires the sign-symmetry assumption stated in the text).

**Outcome.** The primary interval excludes zero in favor of PH (Table 8), and both arms improve on the secondary baseline comparison (Table 9), which satisfies the registered interpretation rule.

| Endpoint (EMA) | PH | Linear | Difference [95% interval] |
|---|---|---|---|
| Two-step FID-10k, update 2000 | 48.41 | 49.47 | 1.06 [0.24, 1.89] |
| Two-step FID-10k, update 4000 | 9.56 | 10.70 | 1.14 [0.83, 1.46] |
| Two-step FID-10k, update 8000 | 4.81 | 7.07 | 2.26 [1.54, 2.98] |
| One-step FID-10k, update 8000 | 9.30 | 10.13 | 0.83 [0.43, 1.23] |
| Test FID, two-step, update 8000 | 4.88 | 7.20 | 2.32 [1.59, 3.06] |
| Test KID, two-step, update 8000 | 0.00100 | 0.00250 | 0.00151 [0.00077, 0.00225] |

Table 9: Secondary endpoints, means over six seeds; differences are linear minus PH and are positive in all six pairs at each endpoint. These secondary intervals are interpreted descriptively; no simultaneous 95% coverage across endpoints is claimed. Update-2000 values reflect substantial EMA lag from EDM initialization. The EDM initializer has two-step FID-10k $162.1$.

The secondary FID-10k improves from initialization and from update 4000 to 8000 in all twelve seed–arm pairs; improvement from initialization is measured by FID-10k, whereas the primary between-arm endpoint uses FID-50k. PH endpoints vary little across seeds (SD $0.03$), and most of the variation in the contrast comes from the linear arm. The six tuning seeds share one pretrained model, dataset, resolution and horizon at $c = 0$. The arms also differ in direction, second-moment preconditioning and effective per-example weights, so the contribution of the loss of stationarity to the quality difference remains unresolved.

### D.5.2 Earlier continuation designs from a consistency checkpoint

These three series start from a shared public consistency-model initializer and average per-image PH losses or squared residuals with detached teachers, taking whole-image norms before batch averaging. Their initializer, horizon, optimizer and smoothing floor differ from those of the diffusion-to-consistency study. The KID values below are unbiased degree-three polynomial estimates, computed with common latent draws and 100 paired subsets of 1000 images; these subsets quantify metric variability within each trained model.

**Original matched-step continuation.** Three independent 200-update tuning streams initialize 750 paired continuation steps with Adam, PH at $c = .1$, teacher time $t/2$ and weight $2/t$. Each pair shares the model/optimizer/EMA state and the data, time, noise and dropout draws at the start. Subsequent moments follow each arm's own gradients, while the linear arm's realized step norm matches that of PH within relative tolerance $5 \times 10^{-4}$, using float64 reconstruction if necessary. The online source and endpoint models generate 10,000 images, and the real pool is 8000 CIFAR-10 test images, also used in the radial diagnostic.

| Source | Initial KID | PH | Linear |
|---|---|---|---|
| 0 | .007007 | .016098 | .022352 |
| 1 | .006829 | .015997 | .023533 |
| 2 | .006936 | .015461 | .023281 |

Both arms deteriorate from their own source. PH has lower KID in all three pairs, with mean linear-minus-PH difference $.007203$.

**Learning-rate development search.** From the public initializer with fresh Adam, PH ($c = .1$) and linear runs use four rates and three paired streams of 200 updates, giving 24 endpoints. The teacher time is $t/2$ and the weight $2/t$; step norms are unmatched. EMA endpoints generate 10,000 images, and the real pool is training indices 45,000–47,499, reused from development and possibly seen in pretraining. The baseline KID is $.002273480$.

| Learning rate | PH mean KID | Linear mean KID |
|---|---|---|
| $3 \times 10^{-7}$ | .002308424 | .002295030 |
| $10^{-6}$ | .002365174 | .002329348 |
| $3 \times 10^{-6}$ | .002455405 | .002380664 |
| $10^{-5}$ | .002481006 | .002423620 |

All 24 endpoints deteriorate; PH loses all 12 paired rate/stream comparisons. The frozen rule selects the lowest tested rate for both arms, without expanding the grid to obtain a favorable outcome.

**Independent recipe diagnostic.** Seven recipes are run on three new paired streams of 200 updates with batch 32, rate $3 \times 10^{-7}$ and fresh Adam, using $(\beta_1, \beta_2) = (.9, .999)$, $\epsilon = 10^{-8}$ and zero decay. The coarse recipe uses teacher time $t/2$ and weight $2/t$, whereas the source recipe uses $t_- = t[1 - (1 + 8\,\mathrm{sigmoid}(-t))/256]$ and weight $256/t$; reconstruction matches the archived source loss and gradient. A bridge keeps the source mapping and $c = .03$ but replaces its weight by $1/(t - t_-)$. Step norms are unmatched, and the original optimizer moments are unavailable.

Training uses CIFAR-10 indices 0–44,999; evaluation uses indices 47,500–49,999, which were unused in the rate search but possibly used in pretraining. Each endpoint generates 10,000 one-step images at $t_{\max} = 80$. EMA weights with decay .999 are primary, and online weights are a prespecified sensitivity check. A second latent/subset panel evaluates all seven recipes for stream 0 on the same real pool.

| Recipe | EMA mean KID | Online mean KID |
|---|---|---|
| Baseline | .002040505 | .002040505 |
| PH coarse, $c = .1$ | .002062315 | .002364199 |
| PH coarse, $c = .03$ | .002062300 | .002364814 |
| PH source, $c = .1$ | .002038747 | .002021750 |
| PH source, $c = .03$ | .002039443 | .002032546 |
| PH source, $c = .03$, bridge | .002040242 | .002035214 |
| Linear coarse | .002051699 | .002276678 |
| Linear source | .002038720 | .002009054 |

All 21 endpoints and 52 evaluations completed, and a zero-learning-rate control preserves weights and KID exactly. Source recipes improve on coarse ones in every stream for both losses and both evaluation weights, and the second panel agrees. The coarse-recipe deterioration persists across floors; the bridge produces a smaller effect, but the causal contributions remain unresolved. Primary EMA comparisons of the source recipes have mixed signs across streams, and panel 1 reverses the stream-0 $c = .03$ contrast. With online weights, linear is better in all three source-recipe pairs at either floor. The EMA retains $.999^{200} \approx .819$ of its initial parameter coefficient, whose effect on KID is nonlinear. These design-dependent rankings do not resolve the contribution of the loss of stationarity.

### D.6 Native image-model field decomposition

Here we measure the frozen field of the archived source loss at the public ECM initializer, whereas Appendix D.3 uses the coarse graph. The checkpoint's EMA weights serve as the fixed current network $D_\theta$, and the detached teacher is that same network with shared dropout randomness and restored buffers; no EMA update is performed. For the source stage 7 recipe, define

$$s(t) = 1 + 8\,\mathrm{sigmoid}(-t), \qquad \delta = \frac{ts(t)}{256}, \qquad \lambda = \frac{256}{t}, \qquad c = .03,$$

with $\log t \sim N(-1.1, 2^2)$, without tail clipping. For $x_t = y + t\epsilon$, $\epsilon \sim N(0, I)$, write

$$r = D_\theta(x_t, t) - D_\theta(x_t - \delta\epsilon, t - \delta), \quad C_\delta = r/\delta, \quad C = D_x D_\theta\, \epsilon + \partial_t D_\theta, \quad J = \partial_\theta D_\theta(x_t, t).$$

The positive JVP includes the time-dependent preconditioning. The image norm precedes the square root and batch averaging. With $h = c/\sqrt{\|r\|^2 + c^2}$, the raw PH field is $F_{\mathrm{PH}} = \mathbb{E}[(s/c)J^\top h C_\delta]$. The primary source differential reference is $F_{\mathrm{src}} = \mathbb{E}[(s/c)J^\top C]$, giving the signed decomposition

$$\begin{aligned} F_{\mathrm{PH}} - F_{\mathrm{src}} &= v_N + v_G, \\ v_N &= \mathbb{E}[(h-1)(s/c)J^\top C_\delta], \qquad v_G = \mathbb{E}[(s/c)J^\top (C_\delta - C)]. \end{aligned}$$

The shared teacher and primary reference make teacher and outer-scale mismatch exactly zero. The separately fixed canonical reference $F_{\mathrm{can}} = \mathbb{E}[J^\top C/c]$ adds $v_S = \mathbb{E}[((s-1)/c)J^\top C]$. These expected signed components underlie Proposition 14. Projection and averaging allow cancellation, so their norms can be smaller than the bounds $A_j = \mathbb{E}\|\text{sample component}\|$ in Equation (20).

**Independent measurement.** Fitting uses 32 batches of 16 images to fix an orthonormal basis for the span of the three estimated component means, and the retained rank is three. Evaluation uses 384 new batches of 16, arranged in 48 independent groups of eight batches. Images are sampled with replacement, with independent randomness, from CIFAR test indices 0–1999 for fitting and 2000–9999 for evaluation. Despite earlier use of this dataset, fitting and evaluation are independent for this measurement. Engineering checks cover archived loss/gradient reconstruction, JVP agreement with finite differences, decomposition closure and fresh-process replay. These float32 checks assess numerical resolution; certified bounds over the entire time law remain unavailable.

Approximate Student intervals (47 degrees of freedom) use one 95% Bonferroni family for 36 coordinates: twelve retained fields/contrasts (including the structural-zero teacher vector) in three fixed directions. For coordinate radii $r_j$, projected norm bounds are $[\max(0, \|\widehat{v}\| - \|r\|), \|\widehat{v}\| + \|r\|]$. Table 10 reports norms after projection and averaging of the vectors.

Table 10: Native source-graph decomposition at one checkpoint. Norms and simultaneous approximate bounds are restricted to the independently fitted three-dimensional plane. Structural zeros are exact for the specified graph.

| Component or contrast | Norm estimate | Bounds |
|---|---|---|
| Normalization attenuation $v_N$ | 451.7 | $[0, 1504.2]$ |
| Finite-gap error $v_G$ | 27.65 | $[0, 131.7]$ |
| Teacher mismatch | 0 | structural zero |
| Primary outer-scale mismatch | 0 | structural zero |
| Canonical outer-scale term $v_S$ | 329.2 | $[0, 826.2]$ |
| $F_{\mathrm{PH}} - F_{\mathrm{src}}$ | 432.8 | $[0, 1457.1]$ |
| $F_{\mathrm{PH}} - F_{\mathrm{can}}$ | 331.5 | $[0, 935.3]$ |
| $F_{\mathrm{PH}} - F_L$ (ordinary linear) | 120.7 | $[14.51, 226.89]$ |

Neither prespecified criterion establishes attenuation dominance. The lower bound on the attenuation norm exceeds neither the gap-term upper bound nor, for the canonical reference, the combined gap and scale upper bounds. The point estimates also cancel: $\|\widehat{v}_N + \widehat{v}_G\| = 432.8$ versus $\|\widehat{v}_N\| + \|\widehat{v}_G\| = 479.4$, and adding $v_S$ gives 331.5 versus 808.6. The ordinary linear field $F_L = \mathbb{E}[\lambda J^\top r]$ differs detectably from PH in this plane, but the contrast includes the global slope factor $1/c$:

$$F_{\mathrm{PH}} - F_L/c = v_N, \qquad F_{\mathrm{PH}} - F_L = v_N + (1/c - 1)F_L.$$

The slope-matched contrast and attenuation dominance remain unresolved within the fitted plane.