IS YOUR UNCERTAINTY MAP WRONG, OR IS ITS TARGET? EXACT DIAGNOSTICS FOR THE TWEEDIE DIAGONAL, AND A GRADIENT-FREE ALTERNATIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models can predict a follow-up medical scan from a baseline, and a clinician reading such a prediction needs a per-voxel map of where it should be trusted. Many of these maps approximate the diagonal of the Tweedie posterior covariance, and they are evaluated by comparing one approximation against another, so it remains unclear whether the estimator or the target is the limiting factor. We therefore compute the exact diagonal on six checkpoints across fourteen model–corpus conditions. A Hutchinson estimator at $M{=}200$ tracks it at rank agreement of at least 0.92 in every condition. However, in four of the fourteen the exact diagonal is anti-correlated with the denoising error, reaching -0.13, so a faithful estimator reproduces that reversal. All four are real-image conditions, and in these checkpoints the reversal does not appear on the models' own samples, so evaluation on model-generated samples gives a favourable reading of this family. From this one may conclude that the limiting factor is the target rather than the estimator.

We then introduce Tweedie Probe-Tangent (T-PT), a gradient-free residual probe that corrupts one model-supported prediction repeatedly and measures the voxel-wise variance of the denoiser's response. T-PT reads a different functional of the same Jacobian, and its exact second-order form ranks identically with the diagonal wherever the diagonal reverses. At the working budget of thirty probes, however, T-PT returns a map too unstable to reproduce that ranking, whilst a Hutchinson estimator at $M{=}5$ is already stable enough to reproduce it; T-PT at that budget is therefore too unstable to be read as evidence against the reversal. Where the diagonal is not reversed, T-PT tracks the diagonal closely in most conditions, at a small fraction of the cost of the exact computation.

We therefore offer T-PT as an instrument rather than as a better approximation, and we evaluate it on clinical follow-up prediction from a single reverse chain, on brain magnetic resonance imaging (MRI) and on lung computed tomography (CT). On brain MRI at full resolution, where every Jacobian-based estimator we test exhausts memory, T-PT leads the twenty-chain Monte-Carlo ensemble on five of the eight endpoints measured inside tissue and trails it on none, at $16\times$ fewer network evaluations. Over the whole volume the ensemble leads instead, on a comparison carried by the background that fills three quarters of it, and a fifty-chain ensemble closes the tissue gap on a thirty-volume subset. On lung CT the ensemble is ahead on every endpoint, whilst a Hutchinson estimate of that diagonal ranks the error well below T-PT. Both estimators lose most of their discrimination in the regions of longitudinal change, which remains an open problem.

1 INTRODUCTION

Diffusion models can predict a follow-up medical image from a baseline scan (Oliveras et al., 2026; Puglisi et al., 2024; Litrico et al., 2026; Liu et al., 2025; Pinaya et al., 2022), but a plausible prediction does not tell a clinician where it should be trusted. The practical object is therefore a spatial uncertainty map: a per-voxel ranking that directs attention toward locations where the prediction is most likely to be wrong. We study such rankings rather than calibrated predictive standard deviations (Sec. 6).

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.

Existing approaches obtain these maps in three ways. Monte-Carlo samplers measure variation across reverse chains and multiply inference cost by the number of chains (Wolleb et al., 2022; Kou et al., 2024): the 20-chain reference used here costs 1,220 network evaluations per volume. Analytical methods learn uncertainty in the noise prediction and propagate it through approximate reverse updates along a single chain (Capellera et al., 2025; 2026; Oliveras et al., 2026). A third family approximates the diagonal of the Tweedie posterior covariance from the denoiser Jacobian (Boys et al., 2024; Peng et al., 2024; Rout et al., 2024; Rissanen et al., 2025; Ou et al., 2025; Schioppa & Salimans, 2026), and its effort has gone into making that diagonal cheaper to estimate. Because the exact diagonal requires one Jacobian–vector product per input coordinate, evaluations normally compare one approximation with another. This leaves a basic ambiguity unresolved: when an uncertainty map fails, is the estimator inaccurate, or is it faithfully estimating an unsuitable target?

We resolve that ambiguity in two stages. First, we treat the exact diagonal as a diagnostic instrument. On six checkpoints small enough for exact computation, one of them a public medical model, fourteen model–corpus conditions in all, we compare it with common approximations and ask whether the diagonal itself ranks denoising error. The approximation is not the bottleneck: Hutchinson estimates (Hutchinson, 1990) track the exact diagonal at rank agreement of at least 0.92 in every condition, while in four conditions the exact target is anti-correlated with the error, reaching -0.13, so that a faithful estimator reproduces the reversal. In these checkpoints, evaluation on model-generated samples does not reveal the reversal observed on real images. Posterior variance need not equal reconstruction error, and the works that construct the Tweedie diagonal mostly serve a sampler rather than claim it ranks error (Sec. 2). But an error-ranking overlay is what a clinician reading a predicted scan needs, and maps of every family are deployed that way (Kou et al., 2024; Wolleb et al., 2022). We therefore ask not whether the field's object was built for that use but whether it would serve it, which is what decides if a better approximation of it is worth building.

Second, we respond to the diagnosis with Tweedie Probe-Tangent (T-PT). T-PT repeatedly corrupts one model-supported clean prediction and measures the voxel-wise variance of the denoiser's residual response. ***Its estimand is a different functional whose ranking follows the diagonal's where the diagonal reverses; at the working budget the probe resolves that ranking only where the diagonal is informative***, most negative correlation -0.016 against the diagonal's -0.13, while Hutchinson at a smaller budget already reproduces the reversal. On longitudinal brain MRI at the full 128^3 resolution, where the Jacobian-based estimators do not fit in memory, a single reverse chain and K perturbations rank the error inside brain tissue as well as or better than a twenty-chain Monte-Carlo ensemble at $16\times$ fewer network evaluations. The natural-image experiments measure the probe against the exact diagonal (Sec. 4); the clinical ones test it at the scale the method is for (Sec. 5).

2 Related work and evaluation gap

Sampling-based uncertainty. Monte-Carlo diffusion methods estimate spatial uncertainty from the spread of independent reverse chains (Wolleb et al., 2022; Kou et al., 2024; Puglisi et al., 2024; Shaw et al., 2025): a natural empirical reference, and ours, but linear in the number of chains. The remaining families target something other than the spatial ranking of one error map, whether uncertainty over networks (Kou et al., 2024; Berry et al., 2024; Jazbec et al., 2025; Gal & Ghahramani, 2016) or coverage (Angelopoulos & Bates, 2023; van Amersfoort et al., 2020), and mostly need extra training or several models.

Analytical variance propagation. Analytical methods propagate uncertainty through the reverse diffusion process. Whereas iDDPM learns reverse-transition variances (Nichol & Dhariwal, 2021), U2Diff learns uncertainty in noise predictions and accumulates it along a DDIM (Song et al., 2021) trajectory, neglecting state–prediction covariance and delaying accumulation to limit variance inflation (Capellera et al., 2025); U2Diffine adds denoiser sensitivity through a first-order expansion with a diagonal Jacobian approximation (Capellera et al., 2026). MUMINS (Oliveras et al., 2026) extends U2Diff to 3-D medical follow-up synthesis along a stochastic reverse chain; App. A establishes that the recursion is expansive, which is why the delayed start is necessary.

Tweedie covariance estimators. Recent methods estimate or approximate posterior covariance from the denoiser Jacobian: a row-sum surrogate from one Jacobian–vector product (Boys et al., 2024), a wavelet diagonal (Peng et al., 2024), an isotropic approximation (Rout et al., 2024), low-rank

running estimates (Rissanen et al., 2025), trained heads (Ou et al., 2025), and DCT block diagonals (Schioppa & Salimans, 2026). Analytic-DPM derives an optimal diagonal under an imperfect mean (Bao et al., 2022), and CA-DPS uses finite differences for inverse problems (Mohajer Hamidi & Yang, 2024). These works improve access to an intractable covariance object, and each serves a sampler, where off-diagonal structure matters because one draws from the matrix. We claim no part of the estimand and no improvement to its approximation; what we add is the exact object it is approximating, and what that object turns out to say.

Algebraically T-PT is a perturbation-based sensitivity estimate, in the family of test-time augmentation for aleatoric uncertainty (Ayhan & Berens, 2018; Wang et al., 2019): repeated corruptions of one input, and the spread of the response. Closest of all, De Vita & Belagiannis (2025) denoise to $\hat{x}_0$, re-noise it M times and take the per-pixel variance of the *scores* ϵ_θ, to guide a sampler and filter samples by FID; we take the variance of the residual $\xi - \epsilon_\theta$ instead, which reads a different functional of the same Jacobian (App. P) and a different map on both clinical tasks (App. R). The corruption is the forward process itself, at the timestep t^* the denoiser was trained on, which gives the residual the geometric reading of Prop. 1; keeping the residual rather than the reconstruction is not a further difference, the two rank identically (App. R). Stanczuk et al. (2024) also perturb a point repeatedly, but apply an SVD to score vectors to estimate one intrinsic-dimension scalar; T-PT retains the per-voxel residual variance and forms no matrix.

3 Diagnosis: is the exact diagonal a useful target?

What is computed. For a denoiser ϵ_θ and a noised state x_t, Tweedie's identity gives the posterior mean $\hat{x}_0(x_t)$ and, differentiating it, the posterior covariance up to schedule factors; the per-voxel map in use is its diagonal, $\sigma_i^2 = \frac{1-\bar{\alpha}}{\sqrt{\bar{\alpha}}}\, \partial \hat{x}_{0i} / \partial x_{ti}$ (App. J). The exact diagonal costs d Jacobian–vector products, one per one-hot tangent; we compute all d: 12,288 for a 64^2 RGB image, 3,072 at 32^2, 12,288, 4,096 and 4,096 in the three latent spaces. Forward-mode automatic differentiation is used throughout, under `no_grad` (App. K explains why none of the three public codebases can be traversed with `torch.func` as released).

Design. Six checkpoints: `inet` and `cifar`, the unconditional ImageNet-64 (Deng et al., 2009) and CIFAR-10 (Krizhevsky, 2009) models of Nichol & Dhariwal (2021); `cond`, the class-conditional ImageNet-64 model of Dhariwal & Nichol (2021); `ffhq` and `church`, latent-diffusion models (Rombach et al., 2022) for FFHQ (Karras et al., 2019) and LSUN churches (Yu et al., 2015); and `brats`, a public MONAI latent-diffusion bundle (Pinaya et al., 2023) for axial MRI brain slices from BraTS dataset (Menze et al., 2015) (App. N). Each is evaluated under up to three corpora: `coco` (held-out photographs (Lin et al., 2014)), `self` (the model's own samples) and `real` (held-out images from the training domain, Imagenette (Howard, 2019) for the ImageNet checkpoints, and unavailable for the latent models). A condition such as `cond/self` therefore names a checkpoint–corpus pair, and fourteen such pairs exist. These checkpoints denoise single images; the baseline–follow-up structure of Sec. 5 plays no part in this track. The three corpora are the experiment, not a robustness check (Sec. 3.3). One hundred images per condition, a seeded random subset; the probe timestep matches the clinical runs' forward marginal scale ($\sigma_{t^*} = 0.32$, from $t^* = 60$ under their $T{=}300$ cosine schedule; Sec. 5), so the two halves of the paper measure at comparable noise. Every estimator's cost is counted in the same unit, forward-equivalent network evaluations, with the conversion fixed in App. J before any score was read. The denoising error e is $|\hat{x}_0(x_{t^*}) - x_0|$ per pixel, averaged over twenty noise draws (App. M).

3.1 Validity of the implied covariance

The bracket $1 - \sqrt{1-\bar{\alpha}}\ \mathrm{diag}(J)$ can be negative. In the one-hot runs it never is on the two pixel-space checkpoints, and it is on up to 5.7% of coordinates on the class-conditional one (Table 7). A negative entry there is not an estimation error, because the diagonal is exact; it says that the covariance implied by the learned denoiser is not a covariance. We clamp at zero and report the clamped fraction (App. M). The probe introduced in Sec. 4 cannot produce this, since it reports a sample standard deviation directly; its clamped fraction is exactly zero.

3.2 Fidelity is not the binding constraint

The Hutchinson estimator converges to the exact diagonal; the convergence, however, buys nothing. On the class-conditional ImageNet checkpoint scored on its own samples (`cond/self`, Table 7), Hutchinson's Spearman correlation with the denoising error rises with $M = 5, 15, 50, 200$ probes as 0.298, 0.359, 0.410, 0.435, approaching the 0.448 that the exact diagonal itself achieves (Table 8). The pattern holds in all fourteen conditions: at $M{=}200$ Hutchinson matches the exact diagonal at rank correlation 0.92–0.996, and its correlation with the error climbs to a limit that is the diagonal's own and no better. Fidelity to the diagonal does not guarantee usefulness of the diagonal: the approximation is loosest on the class-conditional checkpoint, where the diagonal ranks the error best (relative L_2 of 8×10^{-2}), and tightest on CIFAR, where the diagonal is anti-correlated with the error on both real-image corpora (3×10^{-4}). On `cond/self`, T-PT reaches 0.451 at 30 forwards, against roughly $49{,}000$ measured forward-equivalents for the exact object (App. J). The row-sum surrogate fails differently: it does not converge to the diagonal at all, and its correlation with the error is negative or null in ten of the fourteen conditions, reaching -0.232 (Table 7). A better approximation of this diagonal is therefore not a route to a better map.

3.3 The failure mode, and what it depends on

Table 7 reports the Spearman correlation between the exact diagonal and the denoising error in each of the fourteen model–corpus conditions; every 95% interval excludes zero except `ffhq/self`, which straddles it. The four negative rows are the failure mode of Sec. 1 (a map that ranks error-prone voxels below error-free ones), and the table narrows where it lives. Being out of distribution does not explain the sign: Imagenette, the unconditional checkpoint's own training domain, gives -0.131 next to held-out COCO's -0.132, so on the one checkpoint where both corpora exist, the domain leaves the sign untouched. What separates the conditions is real images against the model's own samples: the two real-image corpora agree closely, while on their own output five of the six checkpoints turn clearly positive and the sixth stays non-negative. For the two pixel-space unconditional checkpoints the flip is visible within a single model across all three corpora (`inet`: -0.132, -0.131, $+0.058$; `cifar`: -0.091, -0.077, $+0.029$).

The reversal belongs to one estimand: it is a property of the diagonal, inherited by any estimator of it but not by Monte-Carlo sampling or by the variance recursion, which estimate different objects. On the two clinical checkpoints, where the exact diagonal is out of reach, every computable estimator correlates positively with the error (Table 1).

4 Response: Tweedie Probe-Tangent, a one-chain residual probe

If the target is what fails, a better approximation of it is not the remedy, and the estimator we introduce is not one either: it is a different functional whose second-order expansion ranks with the diagonal where the diagonal reverses. This section defines it, gives it a geometric reading, and returns to the fourteen conditions to show what it delivers at its budget.

4.1 Estimator and cost

Given a model-supported clean state $x_0 \in \mathcal{M}$, fix a probe timestep t^*, draw K independent $\xi_k \sim \mathcal{N}(0, I)$ and form

$$X_{t^*}^{(k)} := \sqrt{\bar{\alpha}_{t^*}}\, x_0 + \sqrt{1 - \bar{\alpha}_{t^*}}\, \xi_k. \tag{1}$$

The *Tweedie residual* is $r_k := \xi_k - \epsilon_\theta(X_{t^*}^{(k)}, t^*, c)$, with c the conditioning where the model has one (Sec. 4.3); the estimator is the per-voxel sample variance across draws,

$$\widehat{\mathrm{Var}}(r_i) := \frac{1}{K-1} \sum_{k=1}^{K} \big(r_k[i] - \bar{r}[i]\big)^2, \qquad \bar{r}[i] := \frac{1}{K} \sum_{k=1}^{K} r_k[i], \tag{2}$$

and the reported map is $\sigma_{\mathrm{TW}}[i] := \sqrt{\widehat{\mathrm{Var}}(r_i)}$. Throughout, K is the number of independent draws $\xi_1, \ldots, \xi_K \sim \mathcal{N}(0, I_d)$ at the single timestep t^*: each costs one evaluation of ϵ_θ and yields one residual $r_k \in \mathbb{R}^d$, and the map is the unbiased sample variance of $\{r_k[i]\}_{k=1}^K$ at every voxel, so K is

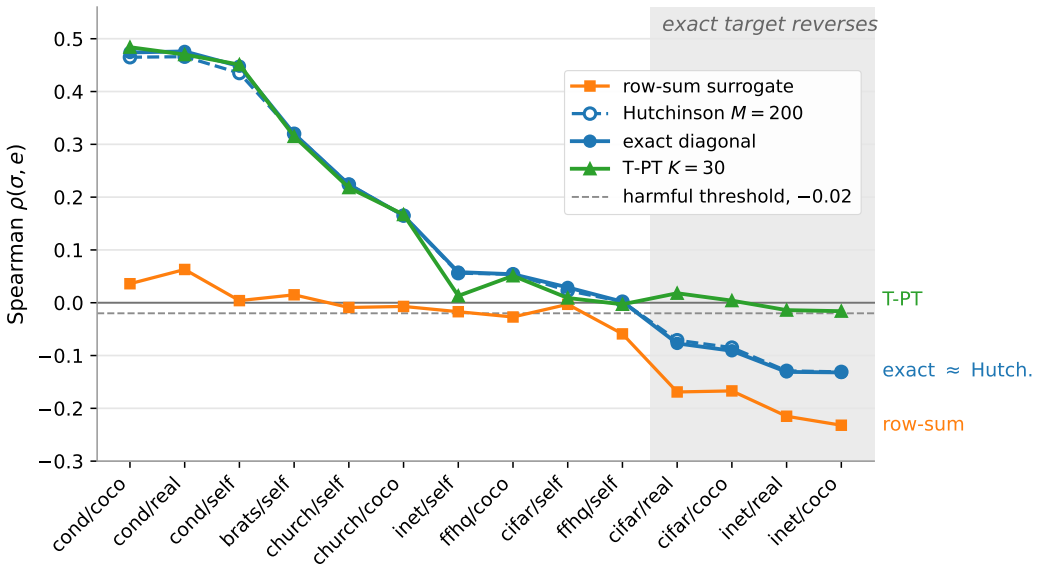


Figure 1: Spearman correlation with the denoising error, per condition, mean over 100 images, ordered by the exact diagonal (Table 7 gives the numbers and the intervals, in its own ordering). In the shaded block the exact diagonal is *anti*-correlated with the error, and Hutchinson at $M{=}200$ follows it there as everywhere else, so what fails is the target and not its approximation. A condition is *harmful* when its interval lies entirely below -0.02 (dashed), a threshold fixed on the synthetic manifolds of App. I before any natural-image score was read: four of fourteen for the exact diagonal, four for Hutchinson, six for the row-sum, and none for T-PT at $K = 30$, whose map runs flat along zero there rather than tracking the diagonal. Sec. 4.4 and App. P show that this flatness is a property of that budget: with 4–16$\times$ the probes T-PT is harmful on the two `inet` conditions too.

at once that variance's sample size and the number of network evaluations the probe adds to its chain. One set of K draws serves every voxel of a map; runs at different K use independent draws. One timestep suffices: the probe reads local sensitivity at one noise scale, where Prop. 1 holds, and the admissible band is wide (App. T). The draws batch into one call; nothing is differentiated, and the estimator runs under `no_grad`. The probe costs one reverse chain of $T/\zeta + 1$ evaluations plus K forwards. On the two clinical tasks of Sec. 5 that is 91 forwards on lung Computed Tomography (CT) at $K{=}30$ and 76 on brain Magnetic Resonance Imaging (MRI) at $K{=}15$, against $J(T/\zeta{+}1) = 1{,}220$ for MC-DDIM at $J{=}20$; App. U sweeps K. We count network evaluations throughout; batching is an implementation option, not a measured speed-up.

4.2 Geometric interpretation

Two measures are in play, and we keep them apart. Under the *joint* law $X_0 \sim p_{\text{data}}$, $\xi \sim \mathcal{N}(0, I)$, conditional expectation is the orthogonal projection in L^2, so $r = \xi - \mathbb{E}[\xi \mid X_t]$ is the component of ξ orthogonal to the square-integrable $\sigma(X_t)$-measurable functions, and $\mathrm{Cov}(r)$ is the covariance of the injected noise that X_t does not identify. The estimator, however, *fixes* x_0 and varies only ξ, so we write $C_t(x_0) := \mathrm{Cov}_\xi\big[\xi - \epsilon^*(X_t, t)\big]$ for that conditional object, with ϵ^* still the population-optimal denoiser under the joint law; the probe estimates $C_t(x_0)$.

Proposition 1 (The residual is the unidentifiable part of the noise)**.** *On a flat manifold $\mathcal{M} = x_0 + T$, $\mathrm{range}(C_t) \subseteq T$ for any tangent law. If the tangent coordinate is Gaussian, $u \sim \mathcal{N}(0, \tau^2 P_T)$, then*

$$C_t(x_0) = \gamma_t^2 \, P_T, \qquad \gamma_t = \frac{\bar\alpha_t \tau^2}{\bar\alpha_t \tau^2 + \sigma_t^2}. \tag{3}$$

The sample covariance of K residuals converges almost surely to $C_t(x_0)$. Proofs and the hypotheses they need are in App. B.

The translation-invariant case, $C_t(x_0) = P_T$ for every t, is the $\tau \to \infty$ limit of Eq. (3), an improper idealisation rather than a law: the geometry fixes the leading eigenspace, while tangent density, noise level, denoiser error and finite K all move the measured eigenvalues (App. I).

Two diagonals, and why they are not the same. The object of Prop. 1 is the diagonal of $C_t(x_0)$, the covariance of the unidentified component of the injected noise, which on a flat manifold is $\gamma_t^2 P_T$, a projection, up to a scalar, whose per-voxel value is the tangent leverage $P_T[i, i]$. The family surveyed in Sec. 1 estimates a different diagonal: that of $\partial \hat{x}_0 / \partial x_t$, the coordinate-wise sensitivity of the posterior mean that Tweedie's identity (Efron, 2011) ties to the full posterior covariance. When the posterior-mean Jacobian is an orthogonal projector (the flat, affine case), the two diagonals carry the same ranking in any tangent frame; beyond that regime they separate, and to second order the probe additionally carries the row energy of the Jacobian (Sec. 4.4).

This distinction is between estimands and is drawn before any measurement: a faithful estimator of $\mathrm{diag}(\partial \hat{x}_0 / \partial x_t)$ must reproduce its sign, while which sign $\mathrm{diag}(C_t)$ takes is an empirical question that Sec. 4.4 answers.

Off the manifold, the leading covariance is governed by the differential of the nearest-point projection, $\mathrm{D}\pi_{\mathcal{M}}(x^*) = (I + hW_\nu)^{-1} P_T$ with W_ν the Weingarten map, so the tangent eigenvalues of C_t become $(1 + h\kappa_i)^{-2}$ in the principal curvatures κ_i: curvature at distance h shrinks the covariance along the directions it bends (Thm. 5, App. C, tested against an exact denoiser in App. H).

Why a geometric quantity should order an error map. C_t describes what the noisy observation fails to identify, not the error of any one reconstruction, and nothing above makes the two equal. A large σ_{TW} marks a voxel where the implied clean prediction is unstable to how the observation was corrupted, and a sampler drawing a clean state can be wrong by more there: that is why we expect the map to order the error, and it fixes no scale, so every endpoint here is rank-based. The probe concerns the model's posterior rather than the anatomy, so a confidently wrong reconstruction is probed as confidently as a correct one (Sec. 6).

4.3 Conditional prediction and the probe point

T-PT's probe (Sec. 4) requires a full clean state $x_0 \in \mathcal{M}$. In the conditional setting, $x_0 = (u_0, v_0)$ splits into a known baseline u_0 and the follow-up v_0 the model is predicting, so half of x_0 is not directly available, and this section constructs a usable probe point. The models here learn the joint state, and at inference the baseline channel is re-injected onto the noised observation at every reverse step (Oliveras et al., 2026).

Under the local-product hypothesis of App. D we model this state as a fibration over plausible baselines: the tangent space splits into horizontal and vertical parts, and for an ideal denoiser conditioned on the clean baseline, measurement injection suppresses the horizontal residual, leaving a progression-conditional component, not "how uncertain is this image" but "given this patient, how plausible is this progression". Our network receives the noised baseline, so for the implemented residual we claim only predominant fibre-alignment, with horizontal contamination from conditioning and model error (Prop. 7, App. D).

The probe point is load-bearing. The training objective constrains the denoiser on and near the training marginal and does not identify its off-support behaviour, so off-support residuals carry no licensed information about $\mathcal{M}$. Setting the unknown channel to $v_0 = 0$ probes a point no training pair ever visited, and the estimator fails *quietly*: the map is smooth, non-negative and spatially varying, yet it correlates with the true error at $\bar{\rho} \approx 0.004$ on lung CT. Using the reconstruction of one additional DDIM chain as the probe, changing nothing else, moves $\bar{\rho}$ to 0.65 (Table 13, App. R). Probe adequacy is therefore a condition we validate against a reference rather than derive from the theory.

4.4 Two estimands that share a reversal

T-PT estimates the diagonal of C_t; the family of Sec. 3 estimates the diagonal of the posterior-mean Jacobian. They are different functionals, separated by Eq. (4) below, but on the four reversed conditions their *rankings* coincide: the second-order map of Eq. (4) ranks identically to the exact diagonal and reverses with it (Table 11). A shared ranking is not a shared estimand, but there it is

enough — a probe that resolved that ranking would reverse too. At $K = 30$ it does not, with no interval in the harmful region, against four of fourteen for the exact diagonal and for Hutchinson at M=200 (Fig. 1). The two therefore target the same ranking here and differ in how hard that ranking is to estimate, which Table 12 measures.

The two are not the same functional of J. Let us write $a = \sqrt{1-\bar{\alpha}_{t^*}} = \sigma_{t^*}$ for the probe's perturbation scale, take $X^{(k)} = \sqrt{\bar{\alpha}}\,\hat{x}_0 + a\,\xi_k$ and $r_k = \xi_k - \epsilon_\theta(X^{(k)}, t^*, c)$, and linearise ϵ_θ about the operating point with Jacobian $J := \partial\epsilon_\theta/\partial x$:

$$r_k \;\approx\; (I - aJ)\,\xi_k - c, \qquad \mathrm{Var}_k(r_i) \;=\; 1 - 2a\,J_{ii} + a^2\,\|J_{i:}\|^2 + O(a^3). \tag{4}$$

At the probe point $a = 0.32$, the quadratic term is not a correction to be dropped, and the $O(a^3)$ remainder is smaller than it by one factor of a only: Eq. (4) identifies which functionals the estimators read, not the size of their difference. In the ideal flat case the tangent projector is $I - aJ$, not J: then $aJ = P_N$, $\|J_{i:}\|^2 = J_{ii}/a$, $\mathrm{Var}_k(r_i) = 1 - a\,J_{ii}$, and the two diagonals rank alike. The separation measured below is therefore a measure of the denoiser's departure from projector structure, consistent with the negative diagonal entries of Sec. 3, which a projector cannot produce. Away from that case the exact diagonal depends only on J_{ii}; the row-sum returns $\sum_j J_{ij}$, in which off-diagonal mass of opposite sign cancels; the probe adds the row energy $\|J_{i:}\|^2$, which on the reversed conditions ranks with J_{ii} and changes nothing there. The row energy alone is a fourth functional, the one a score-variance probe reads (De Vita & Belagiannis, 2025), and it mirrors the diagonal rather than escaping it (App. P).

Where the map resolves, the probe carries more than the diagonal. On twenty 64^3 lung-CT volumes, against a converged Hutchinson reference at M=500, the probe retains $+0.409$ conditioned on the reference and the reference $+0.056$ conditioned on the probe (Table 9); the same reference at M=50 adds only $+0.056$. The clinical map is the one that resolves, so its contrast with the natural-image partials of Table 10 is a probe budget and not two verdicts on the functional (App. P).

The probe does not escape the reversal, and at its working budget does not resolve its own estimand. Eq. (4) is a constant modulated by aJ_{ii}, and at $K = 30$ the sample variance carries a 26% relative error. Where the Jacobian is strong, that resolves the modulation: on `cond/self` two disjoint groups of 30 probes agree at 0.86. Where it is weak, which is where the diagonal reverses, it does not: the same agreement is -0.013 to 0.017 (Table 12). As the budget grows, reproducibility rises on the two `inet` conditions and the anti-correlation emerges with it, -0.016, -0.041, -0.079 at $K = 30, 120, 480$; on `cifar` neither happens at any budget tried, and an approximate attenuation model accounts for all four within 0.02 (App. P). Hutchinson at $M = 5$, about 20 forward-equivalents, is already harmful on all four, so at matched budget one estimator reproduces the reversal, and the other returns a map too unstable to reproduce the ranking it estimates (Table 12).

5 Clinical evaluation: brain MRI and lung CT

5.1 Design

We evaluate two 3-D medical diffusion models from the MUMINS follow-up-prediction pipeline (Oliveras et al., 2026), distinct from the six single-image checkpoints of Sec. 3. Both were retrained on public data using their shared code. Each is a 3-D U-Net (Çiçek et al., 2016) trained with an ℓ_1 noise-prediction loss plus a Gaussian noise-prediction NLL, adapted from U2Diff (Capellera et al., 2025), on a two-channel joint state (Sec. 4.3). One predicts CT lung-nodule progression at 64^3 on the PNG pulmonary-nodule growth dataset, released with Tang et al. (2026), with 130 test pairs from 47 participants; the other predicts MRI brain progression at 128^3 on OASIS-3 (LaMontagne et al., 2019), with 211 scored pairs from 63 participants. Both test splits are patient-disjoint from the training cohorts, so no participant contributes volumes to both. The two models jointly predict the noise ϵ and a per-voxel variance, following the same U2Diff-derived NLL head (Capellera et al., 2025); that variance head feeds the analytical baseline σ_θ below and is not used by T-PT, and the baseline channel u_0 is never predicted, only re-injected at inference (Sec. 4.3). Within each task, every estimator uses the same network. The two tasks use different U-Nets, because the architecture's depth follows the input size.

	Lung CT (PNG, N=130)				Brain MRI (OASIS, N=211)		
	T-PT	MC	Hutch.	Δ [95% CI]	T-PT	MC	Δ [95% CI]
ρ, full volume	0.669	**0.693**	0.404	-0.022 $[-0.037, -0.007]$	0.518	**0.559**	-0.038 $[-0.041, -0.035]$
ρ, tissue	0.670	**0.694**	0.405	-0.022 $[-0.037, -0.007]$	**0.349**	0.337	$+0.016$ $[+0.006, +0.028]$
ρ, interior	0.612	**0.633**	0.326	-0.023 $[-0.043, -0.005]$	**0.334**	0.318	$+0.022$ $[+0.009, +0.037]$
ρ, edges	0.352	**0.528**	0.177	-0.147 $[-0.186, -0.110]$	0.248	0.245	$+0.003$ $[-0.005, +0.010]$
ρ, change region	0.100	**0.265**	0.033	-0.134 $[-0.183, -0.085]$	0.128	0.133	-0.003 $[-0.013, +0.008]$
ρ, partial (grad. rm.)	0.339	**0.409**	0.143	-0.060 $[-0.090, -0.031]$	**0.290**	0.274	$+0.021$ $[+0.009, +0.035]$
AUROC, top-5% error	0.884	**0.922**	0.753	-0.027 $[-0.039, -0.017]$	0.782	0.782	-0.001 $[-0.007, +0.004]$
AUSE ↓	0.152	**0.133**	0.383	$+0.014$ $[+0.004, +0.024]$	**0.306**	0.315	-0.011 $[-0.019, -0.004]$
AURC ↓	0.030	**0.029**	0.047	$+0.001$ $[+0.000, +0.002]$	**0.053**	0.054	-0.001 $[-0.002, -0.000]$
forwards per volume	**91**	1,220	61+200 JVP		**76**	1,220	

Table 1: Clinical results against the common error map. Estimator columns are per-volume means; Δ is T-PT minus MC-DDIM, so on the two rows marked ↓ a positive Δ favours MC-DDIM, with its paired 95% confidence interval resampling participants (63 on OASIS, 47 on lung CT, where the full volume is 99.5% tissue). Hutch. is Hutchinson at M=200 on the probe's own chain, which has no brain column because at 128^3 it exhausts memory (Sec. 5.3); on lung CT, the probe's chain, the reference and the error map share one conditioning (App. M). Bold marks the better estimator where the interval excludes zero. Volume-weighted contrasts: Table 15.

MC-DDIM is the per-voxel standard deviation of J=20 reverse chains, generated for this work in the deployable conditioning arm with twenty distinct seeds. The analytical baseline σ_θ propagates the variance head through the reverse update, with the standard delayed-start mask $\hat{s} = 15$, rechecked on our retrained checkpoints' own held-out validation likelihood following the MUMINS selection protocol. The delay is forced rather than tuned, because within the affine additive class the forcing term cancels from the comparison, so no variance head, injected-noise schedule or additive correction makes the recursion contractive, and a delayed start can only bound the accumulated gain (Prop. 2, App. A). T-PT uses $t^* = 60$, $K = 30$ on lung CT and 15 on brain MRI, and one additional DDIM chain as its probe; both were fixed before any participant-level score was read, the sweeps over t^* and K were run on lung CT, and probe-seed sensitivity is in App. R (Apps. T, U).

Every map is scored against one common error map, the error of the J=20 ensemble-mean prediction, so that the three estimators are ranked against the same target. That map is built from the same twenty chains whose spread is σ_{MC}, which favours the reference; App. R therefore also scores each estimator against the error of the prediction it would accompany at deployment (one chain for T-PT, the J-chain mean for MC-DDIM), where that advantage is absent. The endpoint is per-volume Spearman $\rho(\sigma, e)$: rank, because σ and e live on different scales; per-volume, because pooling lets global intensity differences pass as localisation. Tables write this correlation as ρ and a paired difference of it between two estimators as $\Delta\rho$. A bracket is the 95% interval of the number to its left, never of the row. OASIS additionally uses a prespecified *tissue* endpoint (App. M), with the background being three quarters of a skull-stripped volume. Paired intervals resample *participants* on both tasks, up to 21 pairs from one participant on OASIS, and a volume-level bootstrap covers a known simulated effect 57% of the time against a nominal 95%.

5.2 Results

On brain MRI (211 pairs, 63 participants) each row of Table 1 is a paired per-volume difference with a cluster bootstrap over participants, and the two halves of the table point in opposite directions. Over the whole volume the 20-chain reference is ahead by 0.038 and wins on 210 of the 211 pairs. Inside tissue the ordering reverses: T-PT leads on tissue Spearman, the interior, the gradient-controlled partial, and AUSE (and AURC, Table 15), at 76 forwards against 1,220; on edges, in the change region, and on AUROC, the interval includes zero. Three quarters of a skull-stripped volume is background, where both maps and the error are near zero and agree for free, so which reading matters depends on whether the map is to be used inside the anatomy. The analytical baseline σ_θ (App. R) is not distinguishable from the ensemble over the full brain volume, last by a wide margin inside tissue, and below both on lung CT.

On lung CT, T-PT loses to the ensemble on every endpoint. Against the common error map, the paired tissue difference is -0.022 $[-0.037, -0.007]$ in MC-DDIM's favour, and Table 1 places the refer-

ence's advantage where the structure is: on edges (0.528 against 0.352) and inside the change region (0.265 against 0.100). A second lung-CT checkpoint, retrained without the variance head (App. L), reaches 0.633 against the reference's 0.636, -0.020 $[-0.040, -0.001]$ (Table 19), so the probe needs no variance head to run. At 64^3 the Jacobian family is computable: Hutchinson at $M{=}200$ on the probe's own chain reads 0.405 inside lung tissue against 0.670 ($+0.189$ $[+0.158, +0.222]$) and trails on every row; on twenty volumes the probe retains $+0.409$ conditioned on the converged diagonal, against $+0.056$ the other way (App. O). Scored against its own single-chain prediction, the deployment case, T-PT loses 0.009 inside brain tissue and 0.007 on lung CT, and over the whole brain volume MC-DDIM's own-error figure is 0.552 against T-PT's 0.520 (Table 14). One interpretation, which these experiments do not test and uncropped scans would: a lung-CT volume is 99.5% foreground, so the tissue restriction that reverses the OASIS ordering has almost no background to exclude, and over the full volume OASIS orders the same way (-0.038).

5.3 Cost is a feasibility boundary, not a ratio

At 128^3 on a 40 GB accelerator, run one at a time in clean processes, the row-sum surrogate and Hutchinson at $M{=}50$ and $M{=}500$ each exhaust memory before completing one volume, while T-PT peaks at 33.8 GiB and completes in 66.7 s (Table 5). The three that fail are the Jacobian-based ones: forward-mode differentiation doubles the stored activations, and the probe differentiates nothing. The boundary is that of the released implementations, with activation checkpointing disabled for the tangent pass (App. K); batching tangents would not move it, since each Jacobian–vector product already carries one tangent and the peak is that of a single forward-mode pass.

6 Discussion and limitations

The estimate loses its information where the change is, and so does the reference. The clinical task is to predict how a structure evolves. Restricted to the tissue that actually changed between visits, the top decile of $|x_{\text{target}} - u_0|$ computed without any model output, the OASIS tissue Spearman of 0.349 falls to 0.128, while random tissue subsets of the same size score 0.378, so the drop is specific to where the change is. The 20-chain reference falls to 0.133 (Table 1): the failure belongs to the family, not to its cheapest member. Two untested mechanisms are consistent with it (App. P), one binding every estimator here: a confident prediction receives low variance whether it is right.

Scope. The OASIS volumes were resampled to 1.5 mm isotropic before training, so every clinical number is scoped to that preprocessing (App. P shows the structure it would have erased is still present where σ collapses). Why $\mathrm{diag}(J)$ should be *anti*-correlated with the error rather than merely uninformative remains open (four accounts refuted in App. P). Six checkpoints are not a survey; the clinical evidence is two tasks from the two MUMINS checkpoints, with the functional separation resting on one of them and on the four reversed conditions, and the reference is still gaining tissue discrimination at $J{=}20$, its $J{=}50$ advantage narrowing to $+0.010$ (Table 21). The ranking depends on the probe point lying on the data manifold (Sec. 4.3) and on one timestep. Where the diagonal reverses the $K = 30$ map is barely reproducible, so its absence from the harmful column is a fact about the estimator at that budget, not a guarantee about the map (Table 12); and whether the reversal occurs in clinical checkpoints these tasks cannot answer, the one medical checkpoint on which the diagonal is computable being two-dimensional and unconditional.

7 Conclusion

In four of fourteen conditions the exact diagonal is anti-correlated with the error it is meant to rank, and only exact evaluation shows that the failure belongs to the target. We offer the probe as an instrument rather than as a better approximation, since its second-order ranking follows the diagonal there, and at thirty probes its map is too unstable to reproduce that ranking. What it buys is feasibility at 128^3, where every Jacobian-based estimator we test runs out of memory, and, where its own map resolves, more about the error than a Hutchinson estimate of that diagonal; the changed region remains open.

Reproducibility statement

Every claim is backed by material we release. Apps. B–C prove the affine and curved-support results under the assumptions stated there; App. F derives the schedule-only budget rule, so Fig. 2 is reproducible from the schedule in a few lines. The synthetic manifolds of App. I ship with the code, including training scripts, seeds, and per-number protocols. For the medical experiments, Sec. 5 and the configuration index of App. M (Table 6) record checkpoints, samplers, schedules, probe settings and test splits; App. M defines every mask and metric operationally; App. R states the resampling unit behind every interval and validates it on simulated data. Source code for the estimator, the exact-denoiser and synthetic-manifold checks, the scoring pipeline and the statistical analysis is included in the supplementary material, together with the training configurations and the preprocessing. Weights are not redistributed: the OASIS-3 data-use agreement does not cover derivatives of that cohort, so the brain checkpoints have to be retrained from the released configuration against an approved copy of the data, and we release the recipe rather than the model. No image data are redistributed either; the release documents how to request access and reproduce the preprocessing.

Ethics statement

This work uses two de-identified, publicly released medical imaging datasets under their respective data-use agreements, and introduces no new data collection and no human-subject experimentation. The estimator produces an uncertainty *ranking*, not a calibrated predictive standard deviation, and we say so in the body: it indicates where a model's prediction is least constrained, and it is not validated for, and should not be used for, clinical decision-making. We would flag one risk specifically. An uncertainty map is persuasive because it looks quantitative, and Sec. 4.3 shows an off-manifold probe returning a smooth, plausible map whose correlation with the true error is 0.004. Any deployment therefore needs its own validation against a reference, not the numbers reported here.

Use of AI tools

Large language models were used throughout as a writing and coding assistant: for drafting and revising prose, for implementing, refactoring and debugging analysis code, for drafting proof arguments subsequently checked by the authors, and as an adversarial reader of both the theoretical claims and the experimental protocol. The exact-denoiser validation of App. H was implemented with LLM assistance under author direction.

References

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *Foundations and Trends in Machine Learning*, 16(4): 494–591, 2023.

Murat Seçkin Ayhan and Philipp Berens. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *Medical Imaging with Deep Learning (MIDL)*, 2018.

Fan Bao, Chongxuan Li, Jiacheng Sun, Jun Zhu, and Bo Zhang. Estimating the optimal covariance with imperfect mean in diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*, 2022.

Lucas Berry, Axel Brando, and David Meger. Shedding light on large generative networks: Estimating epistemic uncertainty in diffusion models. In *Uncertainty in Artificial Intelligence (UAI)*, 2024.

Benjamin Boys, Mark Girolami, Jakiw Pidstrigach, Sebastian Reich, Alan Mosca, and O. Deniz Akyildiz. Tweedie moment projected diffusions for inverse problems. *Transactions on Machine Learning Research*, 2024.

Guillem Capellera, Antonio Rubio, Luis Ferraz, and Antonio Agudo. Unified uncertainty-aware diffusion for multi-agent trajectory modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22476–22486, 2025.

Guillem Capellera, Antonio Rubio, Luis Ferraz, and Antonio Agudo. Heteroscedastic diffusion for multi-agent trajectory modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention*, pp. 424–432. Springer, 2016.

Michele De Vita and Vasileios Belagiannis. Diffusion model guided sampling with pixel-wise aleatoric uncertainty estimation. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pp. 3844–3854, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: a large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Jeremy Howard. Imagenette: a subset of ten easily classified classes from ImageNet. https://github.com/fastai/imagenette, 2019.

Michael F. Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics — Simulation and Computation*, 19(2):433–450, 1990.

Metod Jazbec et al. Generative uncertainty in diffusion models. In *Uncertainty in Artificial Intelligence (UAI)*, 2025.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Siqi Kou, Lei Gan, Dequan Wang, Chongxuan Li, and Zhijie Deng. BayesDiff: Estimating pixel-wise uncertainty in diffusion via Bayesian inference. In *International Conference on Learning Representations*, 2024.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Pamela J. LaMontagne, Tammie L.S̃. Benzinger, John C. Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei Vlassenko, Marcus E. Raichle, Carlos Cruchaga, and Daniel Marcus. OASIS-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease. *medRxiv*, 2019.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C.L̃awrence Zitnick. Microsoft COCO: common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Mattia Litrico, Francesco Guarnera, Mario Valerio Giuffrida, Daniele Ravì, and Sebastiano Battiato. Temporally-aware diffusion model for brain progression modelling with bidirectional temporal regularisation (TADM-3D). *Computerized Medical Imaging and Graphics*, 127:102688, 2026.

Qinghui Liu, Elies Fuster-Garcia, Ivar Thokle Hovden, Bradley J. MacIntosh, Edvard O. S. Grødem, Petter Brandal, Carles Lopez-Mateu, Donatas Sederevičius, Karoline Skogen, Till Schellhorn, Atle Bjørnerud, and Kyrre Eeg Emblem. Treatment-aware diffusion probabilistic model for longitudinal MRI generation and diffuse glioma growth prediction. *IEEE Transactions on Medical Imaging*, 44 (6):2449–2462, 2025.

Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015.

Shayan Mohajer Hamidi and En-Hui Yang. Enhancing diffusion models for inverse problems with covariance-aware posterior sampling. *arXiv preprint arXiv:2412.20045*, 2024.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171, 2021.

Anna Oliveras, Roger Marí, Rafael Redondo, Oriol Guardià, Cynthia Ifeyinwa Ugwu, Ana Tost, Bhalaji Nagarajan, Carolina Migliorelli, Vicent Ribas, and Petia Radeva. Mumins: Metadata-conditioned uncertainty-aware medical image next-state synthesis, 2026. URL https://arxiv.org/abs/2609.17169.

Zijing Ou, Mingtian Zhang, Andi Zhang, Tim Z. Xiao, Yingzhen Li, and David Barber. Improving probabilistic diffusion models with optimal diagonal covariance matching. In *International Conference on Learning Representations (ICLR)*, 2025.

Xinyu Peng, Ziyang Zheng, Wenrui Dai, Nuoqian Xiao, Chenglin Li, Junni Zou, and Hongkai Xiong. Improving diffusion models for inverse problems using optimal posterior covariance. In *International Conference on Machine Learning (ICML)*, 2024.

Walter H.L̃. Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F. Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Brain imaging generation with latent diffusion models. In *MICCAI Workshop on Deep Generative Models*, pp. 117–126, 2022.

Walter H. L. Pinaya, Mark S. Graham, Eric Kerfoot, Petru-Daniel Tudosiu, Jessica Dafflon, Virginia Fernandez, Pedro Sanchez, Julia Wolleb, Pedro F. da Costa, Ashay Patel, Hyungjin Chung, Can Zhao, Wei Peng, Zelong Liu, Xueyan Mei, Oeslle Lucena, Jong Chul Ye, Sotirios A. Tsaftaris, Prerna Dogra, Andrew Feng, Marc Modat, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Generative AI for medical imaging: extending the MONAI framework. *arXiv preprint arXiv:2307.15208*, 2023.

Lemuel Puglisi, Daniel C. Alexander, and Daniele Ravì. Enhancing spatiotemporal disease progression models via latent diffusion and prior knowledge (BrLP). In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 173–183, 2024.

Severi Rissanen, Markus Heinonen, and Arno Solin. Free hunch: Denoiser covariance estimation for diffusion models without extra costs. In *International Conference on Machine Learning (ICML)*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Litu Rout, Yujia Chen, Abhishek Kumar, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Beyond first-order tweedie: Solving inverse problems using latent diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Andrea Schioppa and Tim Salimans. Covariance-aware sampling for diffusion models. *arXiv preprint arXiv:2605.13910*, 2026.

Luke Shaw, Abdul-Lateef Haji-Ali, Marcelo Pereyra, and Konstantinos Zygalakis. Bayesian computation with generative diffusion models by multilevel Monte Carlo. *Philosophical Transactions A*, 383(2299):20240333, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Jan P. Stanczuk, Georgios Batzolis, Teo Deveney, and Carola-Bibiane Schönlieb. Diffusion models encode the intrinsic dimension of data manifolds. In *International Conference on Machine Learning (ICML)*, 2024.

X. Tang, Z. Luo, F. Liu, W. Huang, and J. Zou. NGP-Net: A lightweight growth prediction network for pulmonary nodules. *IEEE Transactions on Medical Imaging*, 45(5):2468–2480, 2026.

Joost van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International Conference on Machine Learning (ICML)*, pp. 9690–9700, 2020.

Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019.

Julia Wolleb, Robin Sandkühler, Florentin Bieder, Philippe Valmaggia, and Philippe C. Cattin. Diffusion models for implicit image segmentation ensembles. In *Medical Imaging with Deep Learning*, pp. 1336–1348, 2022.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. LSUN: construction of a large-scale image dataset using deep learning with humans in the loop. arXiv:1506.03365, 2015.

APPENDIX ROADMAP

The appendices fall into four groups. Appendices A–G contain the theory: the DDIM recursion and its delayed start, the proofs of Prop. 2 and Thm. 5, the conditional fibred interpretation, the non-optimal denoiser, the delayed-start rule in closed form, and what a uniform injection term can and cannot do to a spatial ranking. Appendices H and I test those claims where the truth is computable, against an exact denoiser and on synthetic manifolds with known tangent spaces. Appendices J–M give the experimental design: the cost accounting, the forward-mode implementation obstacle, the protocol for the head-free checkpoint, and the operational definition of every mask and metric. Appendices N–U report the per-condition tables behind Secs. 3 and 5, the deployment comparison and its paired tests, and the sensitivity of the estimates to K, to t^* and to the ensemble size J. Each appendix is referenced at the single point in the main text where its conclusion is used, and none needs to be read in full.

Theory and proofs.

A THE DDIM RECURSION, THE DELAYED START, AND THE TOY OVERSHOOT

The body states the consequence; the result itself is this.

Proposition 2 (Expansivity of additive affine covariance accumulation)**.** *Let $\mathcal{T}_s(V) = a_s^2 V + Q_s$ be any* affine *recursion on the accumulated covariance, with $Q_s \succeq 0$ independent of V collecting the variance head, the injected sampler noise and any additive correction. Then $\mathcal{T}_s(V) - \mathcal{T}_s(W) = a_s^2(V - W)$ for all V, W, so the linear part has spectral radius exactly a_s^2,* independently of Q_s*: wherever $|a_s| > 1$, no variance head, no injected-noise schedule and no additive correction makes the recursion contractive, and the cumulative gain over a window is $\prod_s a_s^2$.*

It is stated for the affine additive class, which is the one Capellera et al. (2025) and Oliveras et al. (2026) use, and we claim nothing about state-dependent or Jacobian-linearised recursions. The two equations the proof refers to are

$$V_{s-\zeta} \;=\; \mathcal{T}_s(V_s) \;=\; a_s^2 V_s + Q_s, \qquad Q_s \succeq 0 \text{ independent of } V_s, \tag{5}$$

where $Q_s = b_s^2 \hat{\Sigma}_s + \tilde{\sigma}_s^2 I + C_s$ collects the variance head, the injected sampler noise (with its η^2 factor inside $\tilde{\sigma}_s^2$) and any additive correction C_s, with $b_s = c_t - a_s \sigma_t$; and

$$\mathcal{T}_s(V) - \mathcal{T}_s(W) = a_s^2 \, (V - W) \quad \text{for all } V, W. \tag{6}$$

We use the DDIM formulation of Song et al. (2021) with T steps, cumulative products $\bar{\alpha}_t$ and the cosine schedule of Nichol & Dhariwal (2021): the forward process sends x_0 to $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\, \xi$ with $\xi \sim \mathcal{N}(0, I)$, and $\epsilon_\theta(x_t, t, c)$ is trained to recover ξ. For skip interval ζ, the reverse update is

$$x_{t-\zeta} = \sqrt{\bar{\alpha}_{t-\zeta}}\, \hat{x}_0(x_t) + c_t \epsilon_\theta(x_t, t) + \tilde{\sigma}_t \, \xi_t, \qquad \hat{x}_0 = \frac{x_t - \sqrt{1 - \bar{\alpha}_t}\, \epsilon_\theta}{\sqrt{\bar{\alpha}_t}}, \tag{7}$$

with, following Eq. 16 of Song et al. (2021) and exactly as implemented,

$$\tilde{\sigma}_t^2 = \eta^2 \, \frac{1 - \bar{\alpha}_{t-\zeta}}{1 - \bar{\alpha}_t} \Big(1 - \frac{\bar{\alpha}_t}{\bar{\alpha}_{t-\zeta}} \Big), \qquad c_t = \sqrt{1 - \bar{\alpha}_{t-\zeta} - \tilde{\sigma}_t^2},$$

where $\eta \in [0, 1]$ is the DDIM/DDPM interpolation. Two noise scales must not be conflated: throughout this paper $\sigma_t := \sqrt{1 - \bar{\alpha}_t}$ is the *forward marginal* scale, which is what the geometric theory of Sec. 4 uses, while $\tilde{\sigma}_t$ is the *stochastic-injection* coefficient of the reverse transition. The transition coefficient is $a_s := \sqrt{\bar{\alpha}_{t-\zeta} / \bar{\alpha}_t}$, where s indexes the transition $t \to t - \zeta$. Since $\bar{\alpha}$ increases as t decreases, $|a_s| > 1$ over the chain. Throughout, $\bar{\alpha}_0 = 1$ and $\bar{\alpha}_t \in (0, 1)$ for $t > 0$; ϵ_θ is Borel-measurable and square-integrable, and the accumulated covariance is initialised at $V_T = 0$ and required to stay positive semi-definite (Assumption 4).

Restating Prop. 2 in full: Q_s cancels from Eq. (6), and the rest follows by iterating; the size of the effect it licenses is measured below. The result is stated for the affine additive class, which is the one those implementations use; state-dependent corrections and Jacobian-linearised recursions change the multiplicative operator and fall outside it. App. B gives the proof, the constraint positivity places on C_s, and the scope conditions.

Corollary 3 (Schedule-calibrated delayed start)**.** *For the cosine schedule* $a_s^2 = \bar{\alpha}_{t-\zeta} / \bar{\alpha}_t > 1$ *at* every *transition, so no non-expansive accumulation window exists. Delaying the start bounds the total inflation, which telescopes to*

$$\textstyle\prod_{t \le \hat{s}} a_s^2 = 1 / \bar{\alpha}_{\hat{s}}, \qquad s_\kappa := \max\{\, t \;:\; 1/\bar{\alpha}_t \le \kappa \,\}, \tag{8}$$

so a user-specified amplification budget $\kappa > 1$ *determines the latest admissible start from the schedule alone, in* $\mathcal{O}(T/\zeta)$ *scalar operations with no denoiser call and no data. This replaces an opaque timestep by an interpretable tolerance; it does not make* κ *data-optimal, and delayed start is not the only conceivable algorithmic response (Rem. F). Derivation in App. F.*

At $T = 300$, $\zeta = 5$, an illustrative one-per-cent budget gives $s_\kappa = 16$, close to the 15 selected on validation likelihood; we do not claim Eq. (8) *predicts* that value, since the budget was not fixed independently of it. What the mask buys, and what it cannot buy, is arithmetic on the schedule alone: unmasked accumulation ends at $V \approx 1.5 \times 10^3$ against 2.5×10^{-3} at s_κ, while a mask placed late, at $t = 64$, still inflates by a further $40\times$ (App. B, Fig. 2). Equation (8) configures the chain in Sec. 5 once the budget is chosen.

The distinction is material. On a diffusion model trained on a one-dimensional analytic manifold in $\mathbb{R}^{1024}$, where the true per-pixel variance is measurable by Monte-Carlo, propagating the variance analytically along the same chains under the same independence assumption gives

$$\frac{\mathrm{Var}_{\mathrm{MC}}}{\mathrm{Var}_{\mathrm{analytic}}} = 0.073, \tag{9}$$

an overshoot of $13.7\times$, with the delayed-start mask *already applied* at 60% of that model's schedule. This is not the cost of ignoring the accepted remedy; it is what survives it, consistent with the schedule amplification quantified by Cor. 3. Training protocol, seeds, ground-truth construction and caveats in App. I.

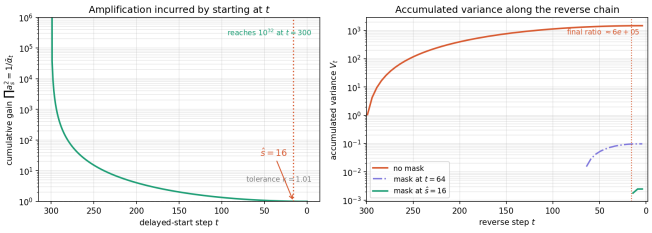


Figure 2: Schedule-induced amplification in the additive variance recursion. **Left:** cumulative gain $\prod a_s^2 = 1/\bar{\alpha}_t$ from starting accumulation at step t; a one-per-cent budget gives $s_\kappa = 16$. **Right:** accumulated variance with no mask, a late mask (t=64), and the schedule-calibrated mask. The mask bounds inflation but cannot make the recursion contractive. Every coefficient comes from the schedule; no model or data is involved.

B Proof of Prop. 2

Assumption 4 (Positivity)**.** We set $\bar{\alpha}_0 = 1$ and $\bar{\alpha}_t \in (0, 1)$ for $t > 0$; ϵ_θ is Borel-measurable and square-integrable. The accumulated covariance is initialised at $V_T = 0$ and is required to remain positive semi-definite at every step, as any object claiming to be a variance must.

Proof. Write the recursion as $\mathcal{T}_s(V) = a_s^2 V + Q_s$ with $Q_s \succeq 0$ independent of V. For any V, W,

$$\mathcal{T}_s(V) - \mathcal{T}_s(W) = \big(a_s^2 V + Q_s\big) - \big(a_s^2 W + Q_s\big) = a_s^2 (V - W),$$

since the forcing term cancels. Hence, in any norm on symmetric matrices, $\|\mathcal{T}_s(V) - \mathcal{T}_s(W)\| = a_s^2 \|V - W\|$, so $\mathcal{T}_s$ is a similarity whose Lipschitz constant equals a_s^2, this being the operator norm and spectral radius of its linear part $D\mathcal{T}_s = a_s^2 I$. This is exact, not a bound, and it does not involve Q_s: the choice of variance head, of injected-noise schedule, or of any additive correction C_s is irrelevant to contractivity.

Composing over a window $[t_1, t_2]$ on which $|a_s| > 1$,

$$\mathcal{T}_{t_2} \circ \cdots \circ \mathcal{T}_{t_1}(V) - \mathcal{T}_{t_2} \circ \cdots \circ \mathcal{T}_{t_1}(W) = \Big(\prod_{s \in [t_1, t_2]} a_s^2 \Big)(V - W),$$

and the gain $\prod_s a_s^2 > 1$ telescopes to $\bar{\alpha}_{t_2}/\bar{\alpha}_{t_1}$ by Eq. (21). Applied to the trajectory itself with $V_T = 0$, the same factor multiplies the accumulated covariance, so V inflates monotonically over the window.

One remark on the hypotheses. Independence of Q_s from V_s is what makes the result exact, and it is not a technicality: a state-dependent correction can restore contractivity while preserving positivity. Taking $J(V_s) = -(a_s^2 - c)V_s$ with $0 < c < 1$ gives $V_{s-\zeta} = c\, V_s + Q_s$, which is a contraction with factor c and maps the positive-semidefinite cone into itself. Such a correction simply lies outside the class of Eq. (5), and it requires a propagation rule that knows a_s and rescales by it rather than a closed-form accumulation. Within the unmodified additive affine rule, delaying the start (Cor. 3) is the remedy in use; Rem. F lists what leaving that rule would buy. □

Remark (What is and is not claimed). The remark is deliberately narrower than "no variance recursion can work". It is exact for the affine family, which is the family those implementations actually use, and it is exact *because* Q_s drops out of Eq. (6): no amount of engineering in the forcing term is relevant to contractivity. Two things fall outside it. A state-dependent correction such as $J(V_s) = -(a_s^2 - c)V_s$ is not an additive forcing independent of V_s: it changes the linear part of the recursion from $a_s^2 I$ to cI, and it does render the map contractive. It does so, however, by subtracting the schedule's own gain, which presupposes knowing a_s and is a rescaling rather than a Jacobian.

Proposition 1, stated in full

Fix t and x_0, and let $r = \xi - \mathbb{E}[\xi \mid X_t]$. The body states the following three claims together; they are separated here and proved in turn.

(i) *(Sampling.)* For fixed t, the sample covariance of K residuals converges almost surely as $K \to \infty$ to the population covariance $C_t(x_0)$.

(ii) *(Flat manifold, translation-invariant prior.)* If $\mathcal{M} = x_0 + T$ is an affine subspace and the law of the tangent coordinate is translation-invariant along T, then the tangent component of ξ is unidentifiable and the normal component is determined, so $C_t(x_0) = P_T$ for every t (the $\tau \to \infty$ limit of Eq. 3).

(iii) *(Flat manifold, proper Gaussian prior.)* If instead $u \sim \mathcal{N}(0, \tau^2 P_T)$, the observation carries partial information about u and the projection shrinks:

$$C_t(x_0) = \left(\frac{\bar{\alpha}_t \tau^2}{\bar{\alpha}_t \tau^2 + \sigma_t^2} \right)^2 P_T. \tag{10}$$

The estimator used in this paper is the diagonal of the sample covariance: a per-voxel scalar, not the full $d \times d$ matrix. Part (ii) is an idealisation; no translation-invariant probability measure exists on a non-compact subspace, and should be read as the improper-prior limit of (iii) rather than as a statement about data. To be explicit about what is and is not proved: $K \to \infty$ buys only (i), convergence to a population covariance, which is not yet a projector; (ii) and (iii) are exact but flat; departures from flatness are the signature of curvature; and Thm. 5 gives the curved extension to second order.

PROOF OF PROPOSITION 1(II) AND (III)

Proof. Split $\mathbb{R}^d = T \oplus N$ orthogonally and write $\xi = \xi_T + \xi_N$, $x = x_0 + u$ with $u \in T$. The forward marginal is $X_t = \sqrt{\bar{\alpha}_t}(x_0 + u) + \sigma_t(\xi_T + \xi_N)$, whose normal component

$$X_t^N = \sqrt{\bar{\alpha}_t}\, x_0^N + \sigma_t \xi_N$$

does not involve u. So ξ_N is a measurable function of X_t and $\mathbb{E}[\xi_N \mid X_t] = \xi_N$: the normal part of the noise is fully identified, and contributes nothing to the residual. Everything turns on the tangent part, where the observation is

$$X_t^T = \sqrt{\bar{\alpha}_t}\, u + \sigma_t \xi_T, \tag{11}$$

a single sum of the unknown position u and the unknown noise $\xi_T \sim \mathcal{N}(0, I_T)$. How much of ξ_T that sum reveals depends entirely on the prior for u.

(ii) Translation-invariant prior. If the law of u is invariant under translations along T, observing Eq. (11) says nothing about ξ_T: any value of ξ_T is consistent with a suitably shifted u, with the same prior weight. The posterior of ξ_T is therefore its prior, $\mathbb{E}[\xi_T \mid X_t] = 0$, and

$$r = \xi - \epsilon^* = \xi_T, \qquad \mathrm{Cov}(r) = \mathrm{Cov}(\xi_T) = P_T,$$

independently of t. The invariance is what does the work, and it is also what makes this an idealisation: Lebesgue measure on T is translation-invariant but infinite, so no proper probability distribution has this property on a non-compact subspace. Read (ii) as the limit of (iii), not as a claim about data.

(iii) Gaussian prior. Take $u \sim \mathcal{N}(0, \tau^2 P_T)$ independent of ξ. Then (u, ξ_T) is jointly Gaussian with X_t^T, and the standard conditioning formula applied to Eq. (11) gives

$$\mathbb{E}[\xi_T \mid X_t^T] = \frac{\sigma_t}{\bar{\alpha}_t \tau^2 + \sigma_t^2} X_t^T.$$

Evaluating the probe at x_0, that is at $u = 0$, gives $X_t^T = \sigma_t \xi_T$ and hence

$$r_T = \xi_T - \mathbb{E}[\xi_T \mid X_t^T] = \frac{\bar{\alpha}_t \tau^2}{\bar{\alpha}_t \tau^2 + \sigma_t^2}\, \xi_T,$$

so $\mathrm{Cov}(r) = \left(\bar{\alpha}_t \tau^2 / (\bar{\alpha}_t \tau^2 + \sigma_t^2)\right)^2 P_T$, which is Eq. (10). The eigenspace is T for every τ and every t; the common eigenvalue is one only in the limit $\tau \to \infty$ or $\sigma_t \to 0$, recovering (ii). □

C Proof of Theorem 5

Theorem 5 (Curved case)**.** *Assume (H1)–(H4) of App. C. Let $x^* \in K$, $x^* = m^* + n^*$ with $m^* = \pi_{\mathcal{M}}(x^*)$, and write $n^* = h\,\nu$ with ν a unit normal. Let $W_\nu := \mathrm{D}\nu|_{T_{m^*}\mathcal{M}}$ be the Weingarten map, with eigenvalues the principal curvatures $\kappa_1, \dots, \kappa_d$ in the direction ν. Then the nearest-point projection satisfies,* exactly in h*,*

$$\mathrm{D}\pi_{\mathcal{M}}(x^*) \;=\; (I + h\,W_\nu)^{-1}\,P_T, \tag{12}$$

and in the regime $\sigma_t \to 0$ with $h = O(\sigma_t)$,

$$C_t(x^*) \;=\; \mathrm{D}\pi_{\mathcal{M}}(x^*)\,\mathrm{D}\pi_{\mathcal{M}}(x^*)^\top \;+\; \mathcal{O}(\sigma_t^2) \;=\; (I + h\,W_\nu)^{-2}\,P_T \;+\; \mathcal{O}(\sigma_t^2), \tag{13}$$

so the eigenvalues of the leading term of C_t restricted to $T_{m^*}\mathcal{M}$ are

$$\frac{1}{(1 + h\,\kappa_i)^2}, \qquad i = 1, \dots, d. \tag{14}$$

Proof in App. C.

Hypotheses. We list them because the theorem is exact only under all four, and because the fourth is the one that fails in practice.

(H1) *Manifold.* $\mathcal{M} \subset \mathbb{R}^N$ is a compact embedded C^4 submanifold of dimension d, without boundary, with positive reach $\tau_{\mathcal{M}} > 0$. (Without boundary is what the Laplace expansion of Step 2 needs: the Gaussian integrals over the tangent chart are complete, the odd moments vanish by parity, and the remainder is $O(\sigma^3)$ rather than the $O(\sigma)$ a truncated domain would leave.) (C^4, rather than C^3, because the proof Taylor-expands the projection to third order; $C^{3,1}$ would also do.) We *assume* this rather than establish it. Intensity normalisation and percentile clipping put the data in a bounded box, which makes boundedness plausible, but boundedness is not compactness without boundary, and the manifold a network represents is not ours to characterise. Positive reach is the load-bearing part: it is what makes the projection unique and the expansion of Step 2 legitimate.

(H2) *Density.* The data measure has a C^3 density φ with respect to the volume measure of $\mathcal{M}$, with locally bounded derivatives, bounded away from zero on a neighbourhood of m^*. The local statement is the one we need and the one that is available: were $\mathcal{M}$ non-compact, a probability density could not be globally bounded below.

(H3) *Tubular neighbourhood.* x^* lies in a compact set K strictly inside the tube of radius $\varrho < \tau_{\mathcal{M}}$, so the projection $m^* = \pi_{\mathcal{M}}(x^*)$ is unique, $n^* = x^* - m^* \in N_{m^*}\mathcal{M}$, and $\delta := \mathrm{dist}(K, \partial\,\mathrm{Tub}_\varrho) > 0$.

(H4) *Optimal denoiser.* $\epsilon_\theta(\cdot, t) = \mathbb{E}[\xi \mid X_t = \cdot]$. App. E treats the failure of this one, which is the only one we cannot arrange.

Step 1: the residual is a posterior-mean displacement. Tweedie's identity applied to $Y = \sqrt{\bar\alpha_t}\,x^* + \sigma_t\xi$, rescaled so that the noised variable reads $Y/\sqrt{\bar\alpha_t} = x^* + \sigma\xi$ with $\sigma = \sigma_t/\sqrt{\bar\alpha_t}$, gives $\epsilon^*(Y) = (Y/\sqrt{\bar\alpha_t} - \mathbb{E}[X_0 \mid Y])/\sigma$ and hence, exactly,

$$r \;=\; \xi - \epsilon^*(Y) \;=\; \frac{\mathbb{E}[X_0 \mid Y] - x^*}{\sigma}. \tag{15}$$

No expansion has been used: the residual *is* the displacement of the posterior mean from the probe, in units of σ. Writing the proof this way removes the need for an asymptotic expansion of the score, which is where an order was lost in an earlier draft of this appendix.

Step 2: a uniform second-order expansion of the posterior mean.

Lemma 6 (Uniform posterior-mean expansion)**.** *Let $K^+ := \{y : \mathrm{dist}(y, K) \le \delta/2\} \subset \mathrm{Tub}_\varrho(\mathcal{M})$, so that perturbations $y = x^* + \sigma\xi$ of any $x^* \in K$ with $\|\sigma\xi\| \le \delta/2$ remain inside the domain of uniformity, the enlargement is what licenses applying the lemma at the perturbed point, not only at x^*. Under (H1)–(H4), with the regularity of φ holding uniformly on a neighbourhood of $\pi_{\mathcal{M}}(K^+)$, there is a locally C^1 vector field $b : K^+ \to \mathbb{R}^N$ such that, uniformly for $y \in K^+$,*

$$m_\sigma(y) := \mathbb{E}[X_0 \mid Y = y] \;=\; \pi_{\mathcal{M}}(y) \;+\; \sigma^2 b(y) \;+\; R_\sigma(y), \qquad \sup_{y \in K^+} \|R_\sigma(y)\| \;\le\; C_{K^+}\,\sigma^3.$$

Proof. By Tweedie's identity, $m_\sigma(y) = y + \sigma^2 \nabla \log q_\sigma(y)$ with $q_\sigma = \rho \operatorname{vol}_{\mathcal{M}} * \mathcal{N}(0, \sigma^2 I)$, so it is enough to prove $\nabla \log q_\sigma(y) = -(y - \pi_{\mathcal{M}}(y))/\sigma^2 + b(y) + O(\sigma)$ uniformly on K^+. Fix $y \in K^+$, set $p = \pi_{\mathcal{M}}(y)$ and $n = y - p = h\nu$, and parametrise $\mathcal{M}$ near p by $m(u) = p + u + \frac{1}{2}\mathrm{I\!I}_p(u,u) + O(\|u\|^3)$, $u \in T_p\mathcal{M}$. With $u = \sigma v$,

$$\|y - m(\sigma v)\|^2 \;=\; \|n\|^2 \;+\; \sigma^2 \langle v,\, G_y\, v\rangle \;+\; O(\sigma^3 \|v\|^3), \qquad G_y := I + h\, W_\nu,$$

the tangent Hessian of the squared-distance function. The point that must not be glossed: h is *fixed* on K^+, not $O(\sigma)$, so the h-dependent quadratic term belongs inside the principal Gaussian, not in the remainder. Positive reach and the strict inclusion $K^+ \Subset \mathrm{Tub}_\varrho(\mathcal{M})$ give uniform bounds $0 < c_- I \preceq G_y \preceq c_+ I$ on K^+ (since $|h\kappa_i| < 1$ there), so Laplace's method applies with the anisotropic Gaussian $e^{-\langle v, G_y v\rangle/2}$ and yields, uniformly in $C^1(K^+)$-norm,

$$q_\sigma(y) \;=\; C_\sigma\, e^{-\|n\|^2/2\sigma^2}\, a_0(y)\, \big(1 + O_{C^1(K^+)}(\sigma)\big), \qquad a_0(y) \;=\; \frac{\varphi(p)}{\sqrt{\det G_y}},$$

with C_σ independent of y; the C^1-uniformity comes from $\mathcal{M} \in C^4$, $\varphi \in C^3$, and Gaussian domination by the lower bound on G_y. Using $\nabla \frac{1}{2}\operatorname{dist}(y,\mathcal{M})^2 = y - \pi_{\mathcal{M}}(y)$ and differentiating the logarithm,

$$\nabla \log q_\sigma(y) = -\frac{y - \pi_{\mathcal{M}}(y)}{\sigma^2} + \nabla \log a_0(y) + O(\sigma),$$

so the claim holds with $b(y) := \nabla \log a_0(y)$, which is C^1 on K^+; multiplying by σ^2 gives the statement. The far part of the manifold integral is harmless: for m with $d_{\mathcal{M}}(p,m) \geq r_0$, positive reach and the strict inclusion $K^+ \Subset \mathrm{Tub}_\varrho(\mathcal{M})$ give $\|y - m\|^2 \geq \|n\|^2 + c r_0^2$ with $c > 0$ uniform on K^+, so its contribution to q_σ and ∇q_σ is $O(e^{-c r_0^2/2\sigma^2})$, beyond every polynomial order.

The formula for b is checkable, and we checked it. On the unit circle with uniform density, $G_y = 1 + h$ and $b(y) = -\frac{1}{2(1+h)}\,\nu$; at $h = 0.15$ this predicts -0.43478. Quadrature at the same point, computed *before* this form of the lemma was written down, gives $(m_\sigma(y) - \pi_{\mathcal{M}}(y))/\sigma^2 \to -0.4348\,\nu$ with vanishing tangential component, agreeing to four decimals, and the residual after subtracting $\sigma^2 b$ scales as $\sigma^{4.5}$, inside the $O(\sigma^3)$ asserted. □

The strengthening matters, and it is worth recording why a bare $O(\sigma^2)$ would not do: dividing by σ in Eq. (15) turns an unstructured $O(\sigma^2)$ remainder into an unstructured *random* $O(\sigma)$ term in r, whose cross-covariance with the leading term need not vanish and could contaminate the covariance at $O(\sigma)$. With the remainder pinned at $O(\sigma^3)$ and the deterministic $\sigma^2 b$ split off, every surviving cross-term is $O(\sigma^2)$.

Interlude: applying the lemma at the random point. Lemma 6 is a deterministic statement on K^+; the proof needs it at the random point $y = x^* + \sigma\xi$. Define $G_\sigma := \{\|\sigma\xi\| \leq \delta/2\}$. On G_σ, $y \in K^+$ and the lemma applies verbatim. On G_σ^c, $\Pr(G_\sigma^c) \leq C e^{-c\delta^2/\sigma^2}$ and the posterior mean is bounded by compactness of the tube, so the complement contributes beyond every polynomial order to the L^2 remainder of Eq. (16) below.

Step 3: expand the projection, not the score. For $x^* \in K$, $\pi_{\mathcal{M}}$ is C^3 and

$$\pi_{\mathcal{M}}(x^* + \sigma\xi) = m^* + \sigma\, A\,\xi + \frac{\sigma^2}{2}\, B[\xi,\xi] + \mathcal{O}(\sigma^3\|\xi\|^3), \qquad A := \mathrm{D}\pi_{\mathcal{M}}(x^*),\; B := \mathrm{D}^2\pi_{\mathcal{M}}(x^*),$$

with $A = (I + hW_\nu)^{-1} P_T$ from Eq. (12). Substituting into Eq. (15) with Lem. 6, and using $b(x^* + \sigma\xi) = b(x^*) + O(\sigma\|\xi\|)$ from the local C^1 regularity of b,

$$r \;=\; -\frac{n^*}{\sigma} \;+\; A\,\xi \;+\; \sigma\Big(\tfrac{1}{2}B[\xi,\xi] + b(x^*)\Big) \;+\; \mathcal{O}_{L^2}(\sigma^2). \tag{16}$$

That the term linear in ξ is the perturbation of the projection is now a consequence of the lemma and the Taylor expansion, not an assertion.

Step 4: take the covariance. In Eq. (16), $-n^*/\sigma$ and $\sigma b(x^*)$ are deterministic and drop out of the covariance. $A\xi$ is odd in ξ and $B[\xi,\xi]$ is even, so their cross-covariance is an odd moment of a centred Gaussian and vanishes; the quadratic term contributes $\mathcal{O}(\sigma^2)$ in its own right, and the L^2 remainder only produces cross-terms of $\mathcal{O}(\sigma^2)$ by Cauchy–Schwarz. Hence, in operator norm,

$$\mathrm{Cov}(r) = A\ \mathrm{Cov}(\xi)\, A^\top + \mathcal{O}(\sigma^2) = (I + hW_\nu)^{-2}\, P_T + \mathcal{O}(\sigma^2),$$

whose tangent eigenvalues are $(1+h\kappa_i)^{-2}$: Eq. (14). Finally, the proof works in the rescaled noise $\sigma = \sigma_t/\sqrt{\bar\alpha_t}$; since $\bar\alpha_t \to 1$ in the small-noise regime, $\sigma^2 = \sigma_t^2/\bar\alpha_t = O(\sigma_t^2)$ and the remainder has the order stated in the theorem. □

A check on the bias, which the covariance does not see. Taking expectations in Eq. (16) gives

$$\mathbb{E}[r] = -\frac{n^*}{\sigma} + \sigma\Big(\tfrac{1}{2}\,\mathrm{tr}\,B + b(x^*)\Big) + \mathcal{O}(\sigma^2),$$

so the order-σ coefficient is *not* curvature alone: it also carries $b(x^*)$, which contains the density gradient. On the unit circle with uniform density the density term vanishes and the coefficient reduces to its curvature part; quadrature there gives $\|\mathbb{E}[r]\| = 1.00\,\sigma$ over $\sigma \in [0.05, 0.30]$ with fitted exponent 1.03, against $|H| = 1$. On a manifold with non-uniform density, this identification would be wrong, and the bias would mix curvature with density drift. We record the check because an earlier version of this proof carried the curvature term at the wrong power of σ, and the bias is the quantity that detects such an error; the covariance does not, since the bias drops out of it.

Remarks on the leading term and the remainder

The body states four caveats in one paragraph; here they are in full.

The projection differential is the object, not its expansion. Writing the first-order truncation $(I - hW_\nu)P_T$, from $(I + hW_\nu)^{-1} = I - hW_\nu + O(h^2)$, instead of Eq. (12) loses an $O(h^2)$ term: on the unit circle at $h = 0.3$ they give 0.592 and 0.490. Since the statement is available exactly, we use the exact one; it is what App. H measures.

Eigenvalues, not eigenspace, at leading order only. The leading term of Eq. (13) is a congruence of P_T, so to that order the range is unchanged and only the eigenvalues move. The $\mathcal{O}(\sigma_t^2)$ remainder is not constrained to respect that, and may both leak into the normal space and rotate the eigenvectors. It is tempting to conclude that a rank statistic cannot see the correction, and that would be wrong: the per-voxel value is $\sigma_i^2 = \sum_j \lambda_j U_{ij}^2$, and reweighting the λ_j *anisotropically* can reorder voxels even with U fixed. Spearman is invariant only under a common positive scalar $\Lambda \mapsto c\Lambda$. What we can say is conditional: if the correction acts approximately as a common scalar across the dominant tangent directions, which holds when the κ_i are comparable, or when one direction dominates U_{ij}^2, the per-voxel ordering is preserved. The observed flatness of the t^* sweep is consistent with that condition; it does not establish it.

The tangent density lives inside the remainder, and dominates it here. Eq. (13) carries no factor γ_t, which looks at first like a conflict with Eq. (10). It is not: for a proper tangent prior of scale τ,

$$1 - \gamma_t^2 \;=\; 1 - \left(\frac{\bar\alpha_t\tau^2}{\bar\alpha_t\tau^2 + \sigma_t^2}\right)^2 \;=\; \frac{2\sigma_t^2}{\bar\alpha_t\tau^2} + \mathcal{O}(\sigma_t^4), \tag{17}$$

so the density shrinkage *is* an $\mathcal{O}(\sigma_t^2)$ term and is already inside the remainder of Eq. (13), with constant $2/(\bar\alpha_t\tau^2)$. The two statements agree; the paper would be wrong only if it attributed the whole remainder to curvature. It does not, and the ordering matters: at $\sigma_{t^*} = 0.32$ with a tangent extent $\tau \sim 1$, Eq. (17) is 0.19, larger than the curvature contribution we are able to name. A measured eigenvalue below one is therefore evidence of a finite tangent prior before it is evidence of curvature.

The remainder is what shrinks. Under the cosine schedule at $T = 300$ our probe sits at $\sigma_{t^*} = 0.32$, which is not an asymptotic regime. App. H tests Eq. (14) against an exactly computed denoiser on a manifold whose projection is closed form: the predicted eigenvalue is tracked to one per cent over a factor of 2.3, and the disagreement scales as σ^2 as claimed. With constants of order one, the first-order statement $C_t = P_T + \mathcal{O}(\sigma_t)$ leaves about a third of the leading term unaccounted for, and Eq. (13) leaves about a tenth. We report the second-order form because it is the one that is quantitatively honest where we actually operate, not because σ_t is small.

What the theorem does not say. Two readings are available, and only one is correct. It does *not* say that a curved manifold gives non-unit eigenvalues: $S_{n^*} := h\,W_\nu$ is linear in n^*, so an on-manifold probe gives P_T to this order however curved $\mathcal{M}$ is. It says that the eigenvalues report curvature *as seen from the offset*. This is why the on-manifold experiments of App. I do not test Eq. (14), and we do not present them as though they did.

D THE FIBRED RESIDUAL, AND WHAT MEASUREMENT INJECTION DOES

Fibration. Add to (H1)–(H4) of App. C:

(H5) *Fibration.* $\pi_0 : \mathcal{M} \to \mathcal{M}_0$, $\pi_0(u, v) = u$, is a C^3 submersion onto the manifold $\mathcal{M}_0 \subset \mathbb{R}^{N'}$ of plausible baselines, $\dim \mathcal{M}_0 = d_0$. For each u the fibre $\mathcal{F}_u = \pi_0^{-1}(u)$ is a smooth submanifold of dimension $d_F = d - d_0$.

The tangent space then splits orthogonally, $T\mathcal{M} = T^{\mathrm{h}} \oplus T^{\mathrm{v}}$ with $T^{\mathrm{v}} = \ker \mathrm{d}\pi_0 = \{0\} \times T\mathcal{F}$, and correspondingly $P_T = P_T^{\mathrm{h}} + P_T^{\mathrm{v}}$ with $P_T^{\mathrm{h}} P_T^{\mathrm{v}} = 0$. The normal splitting does *not* follow from (H5) alone. Under the additional compatible local-product assumption of Prop. 7, the ambient normal bundle admits the corresponding splitting; under that same assumption we write $n^* = n^*_{\mathrm{h}} + n^*_{\mathrm{v}}$ with $\|n^*_{\mathrm{h}}\| = \mathrm{dist}(u^*, \mathcal{M}_0)$ and $\|n^*_{\mathrm{v}}\| = \mathrm{dist}(v^*, \mathcal{F}_{u^*}) + O(\|n^*_{\mathrm{h}}\|)$.

Proposition 7 (Fibred structure of the leading covariance)**.** *Assume (H1)–(H5) and that, near m^*, the manifold, the ambient normal bundle and the data density admit a compatible local product decomposition, with horizontal and vertical projections $P_{\mathrm{h}}, P_{\mathrm{v}}$. Then the leading covariance of Thm. 5, $\mathrm{D}\pi_{\mathcal{M}}(x^*)\,\mathrm{D}\pi_{\mathcal{M}}(x^*)^\top$, is block diagonal with respect to the horizontal–vertical splitting, and the cross-covariance of the two residual components is $\mathrm{Cov}(r_{\mathrm{h}}, r_{\mathrm{v}}) = O(\sigma_t^2)$. Exact vanishing would require an exact product factorisation, and is not claimed for the implemented estimator, whose covariance carries the contamination term E_{fib} below.*

Proof. Under the product assumption P_{h} and P_{v} commute with the Weingarten map, so $\mathrm{D}\pi_{\mathcal{M}} = (I + hW_\nu)^{-1} P_T$ splits blockwise and the congruence $\mathrm{D}\pi_{\mathcal{M}}\,\mathrm{D}\pi_{\mathcal{M}}^\top$ is block diagonal. For the cross term, apply the expansion Eq. (16) projected by P_{h} and P_{v}: after centring, the linear parts are uncorrelated by the block structure, the linear–quadratic cross terms vanish by Gaussian parity, and what survives is $O(\sigma_t^2)$. We state the result for the *covariance*, not the raw cross-moment $\mathbb{E}[r_{\mathrm{h}} r_{\mathrm{v}}^\top]$: both components carry deterministic means ($-n^*_\bullet/\sigma$ and $\sigma b_\bullet$), so the uncentred moment need not be small even when the covariance is. □ □

Remark (Why the extra assumption is not cosmetic). A submersion alone gives the *tangent* splitting $T\mathcal{M} = T^{\mathrm{h}} \oplus T^{\mathrm{v}}$ by taking $T^{\mathrm{h}} = (T^{\mathrm{v}})^\perp \cap T\mathcal{M}$. It does not by itself give a compatible splitting of the *ambient normal* space, nor the distance identities $\|n^*_{\mathrm{h}}\| = \mathrm{dist}(u^*, \mathcal{M}_0)$ and $\|n^*_{\mathrm{v}}\| = \mathrm{dist}(v^*, \mathcal{F}_{u^*}) + O(\|n^*_{\mathrm{h}}\|)$: those need the fibres to sit in a compatible way as the base point moves, which is what a local product structure, or a metric connection with vanishing O'Neill tensor, supplies. We assume it and do not verify it on our data.

Remark (What measurement injection does, and does not, give). Re-projecting the baseline channel onto the noised observation at every reverse step removes the horizontal component of the residual *provided the denoiser is conditioned on the clean baseline* u^*. If it observes only $u_t = \sqrt{\bar\alpha_t} u^* + \sigma_t \xi_u$, it cannot separate u^* from ξ_u exactly, and what survives is a horizontal residual of the same order as the model error. Our sampler injects the measurement but the network is not given u^* as an explicit conditioning input, so the cancellation is approximate. We therefore read

$$r_{\mathrm{obs}} \approx r_{\mathrm{v}} + \Delta^u_\theta$$

as an interpretation supported by the construction, not as an identity.

Three consequences, and one caveat. *Under the exact local-product, clean-conditioning idealisation, the vertical component is fibre uncertainty*: not "how uncertain is this image" but "given this patient, how plausible is this progression". The implemented estimator only approximates that component and may carry horizontal contamination.

The rank counts progression freedom, for the ideal model. Under the product structure and clean-baseline conditioning, the covariance of the observable residual has rank $d_F = \dim \mathcal{F}_{u^*}$: the degrees

of freedom of progression available to this patient. For the implemented estimator $\mathrm{Cov}(r_{\mathrm{obs}}) = \mathrm{Cov}(r_{\mathrm{v}}) + E_{\mathrm{fib}}$ with E_{fib} collecting the horizontal-contamination terms Δ_θ^u of the previous paragraph, and we make no exact-rank claim.

Zero-placeholder probing, in this language. Setting $v_0 = 0$ creates primarily an off-*fibre* displacement, $n_{\mathrm{v}}^* \neq 0$; without an exact product structure it can also excite the horizontal component. The collapse from $\bar{\rho} = 0.65$ to 0.004 in Sec. 4.3 is consistent with this reading, though it does not by itself separate the two contributions.

Caveat. The measurement-injection remark above presumes $u^* \in \mathcal{M}_0$, that is, a baseline typical of the training population. For a patient outside it the injection does not cancel the horizontal drift and the vertical component is contaminated. We have not measured this, and it is the one place where the fibred reading could mislead: an atypical baseline would produce a fibre-uncertainty map that is not one.

E When the denoiser is not optimal

Hypothesis (H4) of App. C asks for the population-optimal denoiser. No trained network satisfies it, and this is the failure that matters in practice, so it is worth writing down exactly rather than absorbing into a remainder. Define the *model error*

$$\Delta_\theta(X_t, t) \;:=\; \epsilon_\theta(X_t, t) - \mathbb{E}[\xi \mid X_t]. \tag{18}$$

Since the residual is linear in the predictor, $r_{\text{real}} = r_{\text{ideal}} - \Delta_\theta$ pointwise, and therefore

$$\mathrm{Cov}(r_{\text{real}}) \;=\; \mathrm{Cov}(r_{\text{ideal}}) + \mathrm{Cov}(\Delta_\theta) - 2\,\mathrm{sym}\;\mathrm{Cov}(r_{\text{ideal}}, \Delta_\theta). \tag{19}$$

Three remarks, in decreasing order of comfort.

The global bias is estimable; the local one is not identified. Averaged over a validation distribution matching the joint data law, the mean residual estimates the global mean model-error bias $-\mathbb{E}[\Delta_\theta]$, so

$$\widehat{\Delta_\theta} \;=\; -\,\mathbb{E}_{(x_0,\xi)\sim\mathcal{D}_{\text{val}}}\big[r\big] \tag{20}$$

can be subtracted from the observed bias to remove a global offset; it does not identify the input-dependent local error $\mathbb{E}[\Delta_\theta \mid x_0]$. Nothing here requires a posterior over weights, which is what Laplace- and ensemble-based corrections need.

The cross term is not. Eq. (19) has a term coupling the ideal residual to the model error, and estimating it needs the ideal residual, which is the unknown. We do not estimate it and we do not assume it is zero. It is the honest content of the statement that σ_{TW} is an ordering rather than a calibrated covariance.

This is what the synthetic manifolds measure. On the toy manifolds of App. I the geometry is closed-form, so departures of the measured spectrum from the leading projector combine finite-K sampling, the finite-noise geometric and density remainder, and denoiser error Δ_θ; the known geometry lets these be bounded or measured separately, but they do not vanish identically. That is the right way to read the second tangent eigenvalue reported there: at $K = 2000$ and $d = 2$ in codimension 1022, finite sampling alone places the two non-zero eigenvalues at 1.03 ± 0.03 and 0.97 ± 0.03, so a measured 0.839 is roughly five standard deviations low and is a real effect, while a measured 1.052 is not distinguishable from sampling and needs no explanation at all.

F The delayed-start mask in closed form

Corollary 3 states that, *given* a user-selected amplification budget κ, the schedule uniquely determines s_κ. This appendix derives that map, and records the two facts the body quotes: that no stable window exists, and that the cumulative inflation telescopes. It replaces an opaque timestep by an interpretable tolerance; it does not eliminate the user's choice.

No stable window exists. Define the linear map on the accumulated covariance $V \mapsto \mathcal{T}_t[V] := a_s^2 V + b_s^2 \hat{\Sigma}_s + \tilde{\sigma}_s^2 I$ implied by any recursion of the form of Prop. 2. By Eq. (6) its derivative is $a_s^2 I$,

so

$$\rho\big(D\mathcal{T}_t\big) = a_s^2 = \frac{\bar{\alpha}_{t-\zeta}}{\bar{\alpha}_t}. \tag{21}$$

It is tempting to place the mask where $\rho \le 1$. That set is empty: $\bar{\alpha}$ increases monotonically as t decreases, so Eq. (21) exceeds 1 at every transition of the reverse chain. For the cosine schedule at $T = 300$, $\zeta = 5$ the homogeneous part is expansive at 100% of steps. This is Prop. 2 restated concretely: a mask cannot make the recursion contractive. What it can do is bound how much the recursion inflates.

The cumulative gain telescopes. Starting accumulation at step $\hat{s}$, the total homogeneous amplification down to $t = 0$ is

$$\prod_{t \le \hat{s}} a_s^2 = \prod_{t \le \hat{s}} \frac{\bar{\alpha}_{t-\zeta}}{\bar{\alpha}_t} = \frac{\bar{\alpha}_0}{\bar{\alpha}_{\hat{s}}} = \frac{1}{\bar{\alpha}_{\hat{s}}}, \tag{22}$$

because every intermediate factor cancels. The delayed-start step is therefore a statement about how much inflation one is willing to tolerate, which is Eq. (8); the tolerance itself remains the user's choice.

Sensitivity to the tolerance. At $T = 300$, $\zeta = 5$: $\kappa = 1.005$ gives $s_\kappa = 11$, $\kappa = 1.01$ gives 16, and $\kappa = 1.02$ gives 24. The value reported in variance-propagating implementations, $\hat{s} = 15$, is close to the one-per-cent inflation point of this schedule. We do not claim Eq. (8) predicted it independently: the one-per-cent budget was not fixed before seeing the published value. What the map offers is $\mathcal{O}(T/\zeta)$ scalar operations, no autograd, no denoiser call, and no data.

Remark (Escapes). Prop. 2 rules out one mechanism, not uncertainty estimation in diffusion models. Four escapes remain. Change the schedule so $|a_s| \le 1$ throughout, which means leaving the standard DDIM/DDPM family and retraining. Leave the affine class for a state-dependent or fully linearised propagation, at the cost of the Jacobian products the closed-form recursion existed to avoid. Track higher moments, which reintroduces the sampling cost. Or stop recursing: with no accumulated covariance there is no operator to misbehave. We take the last (Sec. 4).

Exact and synthetic validation.

G Magnitude and ranking in analytical variance propagation

Consider the zero-Jacobian analytical baseline in its stochastic form (Oliveras et al., 2026): along a DDIM chain with $\eta = 1$, its output separates into a spatially varying contribution from the uncertainty head and a uniform contribution from the injected sampler noise. The deterministic variant of Capellera et al. (2025) ($\eta = 0$) is the special case $\nu_k = 0$ below, in which the second contribution is absent. This decomposition clarifies which parts of the map can affect the spatial ranking.

Index the actual active propagation transitions by $k = 0, \dots, n-1$, in sampling order, and write

$$V_{k+1} = a_k^2 V_k + b_k^2 D_k + \nu_k^2 I, \qquad V_0 = 0.$$

Here a_k, b_k, ν_k are spatially uniform schedule coefficients, D_k is the nonnegative diagonal predicted noise variance along the fixed trajectory, and ν_k is the sampler's injected-noise scale ($\nu_k = 0$ for a deterministic DDIM chain). The delayed start determines the active transitions. A final step that carries variance unchanged is omitted from this index, matching the MUMINS convention.

Proposition 8 (Uniform injection preserves spatial ordering)**.** *For the fixed trajectory and active transitions above, define*

$$w_k = \prod_{j=k+1}^{n-1} a_j^2, \qquad L = \sum_{k=0}^{n-1} w_k b_k^2 D_k, \qquad F = \sum_{k=0}^{n-1} w_k \nu_k^2.$$

Then $V_n = L + FI \succeq FI$. The propagated standard-deviation map $\sqrt{V_n[i,i]} = \sqrt{L_{ii} + F}$ has the same voxel ordering, including ties, as L_{ii}. Consequently, adding the uniform injection contribution cannot change Spearman correlation with a fixed error map or selective prediction based only on the uncertainty ordering, wherever those metrics are defined and ties are handled consistently.

Proof. Repeated substitution in the recursion gives $V_n = \sum_k w_k(b_k^2 D_k + \nu_k^2 I) = L + FI$. All summands of L are positive semidefinite. Since $F \geq 0$ is common to all coordinates, $x \mapsto \sqrt{x + F}$ is strictly increasing on $x \geq 0$ and preserves the ranks. □

This result concerns rank-based evaluation, not calibrated intervals: F can change coverage and interval width. A common positive conversion back to image units likewise preserves ranks. The statement does not compare different samplers or start times: changing η, the active window, or the trajectory can change D_k, b_k, and the temporal weights, and therefore change the ordering. It provides no guarantee that the spatially varying term L ranks reconstruction error well. A spatially constant injection floor by itself cannot explain a loss of ranking performance; that explanation must involve other parts of the estimator or evaluation.

H TESTING THE CURVATURE TERM AGAINST AN EXACT DENOISER

Theorem 5 is the one claim in this paper that our manifold experiments do not touch, because those probe *on* the manifold where $S_{n^*} = 0$ and the second-order statement collapses to the first-order one. This appendix tests it directly, and does so without a trained network: the optimal denoiser is computed by quadrature, so hypothesis (H4) holds by construction, $\Delta_\theta = 0$, and any disagreement is the theorem's own remainder rather than model error.

Setup. $\mathcal{M}$ is the unit circle in $\mathbb{R}^2$ with uniform density, chosen because its projection is closed form: $\pi_{\mathcal{M}}(y) = R\,y/\|y\|$, whose differential at $x^* = (R + h)\hat{n}$ is $[R/(R + h)]P_T$. With the outward unit normal, $W_\nu = R^{-1} I_T$ and $\kappa = 1/R$, so Eq. (12) reads $(1 + h/R)^{-1} P_T$ and reproduces that exactly. The prediction of Eq. (14) therefore carries no free parameter and no sign convention to argue about:

$$\lambda_{\text{tangent}} \;=\; \frac{1}{(1 + h\kappa)^2} \;=\; \Big(\frac{R}{R + h}\Big)^2. \tag{23}$$

This is also the test that distinguishes the exact covariance eigenvalue $(1 + h\kappa)^{-2}$ from the value obtained by squaring the first-order projection $(I - hW_\nu)P_T$, namely $(1 - h\kappa)^2$: at $h = 0.3$ they predict 0.592 and 0.490, a 17% gap that the measurement resolves cleanly in favour of the exact form.

The residual is formed exactly as in the toy scripts: fix x^*, draw ξ, set $X_s = \sqrt{\bar{\alpha}}\,x^* + \sqrt{1 - \bar{\alpha}}\,\xi$, take $r = \xi - \epsilon(X_s)$, with ϵ the exact conditional mean, and both expectations are quadratures rather than Monte Carlo, so no sampling error enters.

h	h/σ	predicted	measured	rel. error
-0.20	-2.0	1.5625	1.5425	-1.28%
-0.10	-1.0	1.2346	1.2206	-1.13%
0.00	0.0	1.0000	0.9898	-1.02%
$+0.10$	$+1.0$	0.8264	0.8188	-0.92%
$+0.20$	$+2.0$	0.6944	0.6886	-0.84%

Table 2: Tangent eigenvalue of $\mathrm{Cov}(r)$ against normal offset, at $\sigma = 0.10$; "predicted" is Eq. (23). The prediction moves over a factor of 2.3 and the measurement follows it to about one per cent throughout, with the sign the theorem requires: probing inside the circle inflates the eigenvalue above one, outside deflates it, and $h = 0$ recovers the flat result. The normal component stays below 1.3×10^{-2}, confirming that the normal direction is identified as the decomposition claims.

The remainder is $O(\sigma^2)$, tested not asserted. Holding $h/\sigma = 1$ so the regime $\|n^*\| = O(\sigma)$ is respected, the relative disagreement divided by σ^2 is -0.87, -0.90, -0.92, -0.94, -0.96 at $\sigma = 0.20, 0.14, 0.10, 0.07, 0.05$. A sixteen-fold change in σ^2 moves the ratio by nine per cent, so the error really does scale as the theorem's remainder, with a constant near -0.9.

Control. The residual could have been quadrature error rather than physics, and at small σ the kernel narrows so discretisation would *grow* as σ falls, not shrink. It shrinks. Refining both rules eightfold, n_θ from 1024 to 8192, Gauss–Hermite order from 48 to 128, leaves the measured

eigenvalue unchanged at 0.818837 to six figures. The quadrature is converged and the one per cent is the second-order term.

I SYNTHETIC VALIDATION ON MANIFOLDS WITH KNOWN TANGENT SPACES

Everything in Sec. 4 rests on Proposition 1, and on medical data that proposition cannot be checked directly: $\mathcal{M}$ is unknown, so P_T has no ground truth to be compared against. The results of Sec. 5 are therefore consistent with the geometric reading but do not establish it. A flat t^* curve, in particular, is equally consistent with normalisation effects, denoiser saturation, or spatial correlation induced by the U-Net.

This appendix removes the ambiguity by running the same estimator on manifolds where P_T is available in closed form: an analytic manifold of dimension one, and an analytic manifold of dimension two with position-dependent curvature. Both use the same 32×32 embedding ($d = 1024$), the same 2-D U-Net (Ho et al., 2020) with adaptive group normalisation (AdaGN) (Dhariwal & Nichol, 2021), $T = 100$ timesteps, and the same hybrid objective $\mathcal{L}_{\text{simple}} + \lambda \mathcal{L}_{\text{vlb}}$ at $\lambda = 10^{-3}$, so what changes between them is the geometry and not the recipe. The objective is of the same family as the medical checkpoints', though trained here from scratch at a much smaller scale; we do not claim the two share hyper-parameters.

Stage	Manifold	$\dim \mathcal{M}$	codim
S1	S^1, one Gaussian blob on a circular orbit	1	1023
S2	T^2, two independent blobs	2	1022

Table 3: The two synthetic regimes. Codimension is severe by design: a rank-1 covariance inside $\mathbb{R}^{1024}$ is not something a normalisation artefact produces by accident.

I.1 $\mathrm{Cov}(r)$ RECOVERS THE TANGENT EIGENSPACE AND EFFECTIVE RANK

The prediction of Proposition 1, read in its broad-prior limit, is that the residual covariance is tangent-supported with the leading eigenspace of the projector onto $T_{x_0}\mathcal{M}$: its spectrum should consist of $\dim \mathcal{M}$ eigenvalues near one and $d - \dim \mathcal{M}$ eigenvalues near zero, and its leading eigenspace should coincide with the analytic tangent space. Both are directly testable in S1 and S2.

	λ_1	λ_2	λ_3	subspace alignment with $T_{x_0}\mathcal{M}$
S1 ($\dim = 1$)	0.9931	0.0118	—	0.9998
S2 ($\dim = 2$)	1.052	0.839	0.018	0.9967
Prop. 1 (broad-prior limit)	1	1 if $\dim \geq 2$, else 0	0	1

Table 4: Spectrum and leading eigenspace of $\mathrm{Cov}(r)$ against the analytic prediction. Alignment is the mean squared canonical correlation between the leading $\dim \mathcal{M}$ eigenvectors and the analytic tangent span. On S1 the eigengap λ_1/λ_2 is 84.2; on S2 the gap between the second and third eigenvalues is a factor of 47, so the intrinsic dimension is read off the spectrum without a threshold being tuned.

The spectrum exhibits the correct effective rank in both cases without a threshold being tuned, and the leading eigenspace agrees with the analytic tangent space to four decimal places on S1 and three on S2. We take this as validation of tangent support, effective rank and leading-eigenspace recovery. It is not evidence that the finite-noise, trained-denoiser covariance equals the exact projector, Prop. 1(iii) itself forbids that at finite τ.

I.2 THE CORRECTION TERMS ARE MEASURABLE, AND NOT ZERO

Proposition 1 is a statement about the limit $K \to \infty$ *and* $\sigma_{t^*} \to 0$ under an optimal denoiser. At finite K, finite noise, and with a trained network, the estimator carries three separable errors,

$$\widehat{C}_{K,t}(x_0) - P_T \;=\; E_K(t) \;+\; E_{\text{geom}}(t) \;+\; E_\theta(t), \tag{24}$$

where E_K is sample-covariance error, $E_{\text{geom}} = \mathcal{O}(\sigma_t^2)$ for an on-manifold probe by Thm. 5, and E_θ is the gap between the trained and the population denoiser. We do not quote an operator-norm rate for E_K, in this codimension that would require a dependence on effective rank we have not derived, and instead measure it by repetition (App. E gives the finite-K eigenvalue spread). On the toys, all three can be measured rather than assumed.

The geometric term is visible in the *mean* residual, not the covariance: on the manifold, $\mathbb{E}[r] = \sigma\, b(x_0) + \mathcal{O}(\sigma^2)$, so the posterior-mean displacement $\sigma\mathbb{E}[r]$ offsets along the mean curvature vector at order σ^2. Sweeping the probe timestep on S1 gives a ratio between the measured and predicted offset that is linear in σ,

$$\frac{\sigma\,|\langle \bar{r}, \hat{H}\rangle|}{\sigma^2 \kappa} \;\approx\; 1.171 - 0.138\,\sigma, \qquad \kappa = 0.497,$$

across $\sigma \in [0.04, 0.81]$. On S2, at a single probe step, the measured offset is 0.0508 against a predicted $\sigma^2|H| = 0.0358$, a ratio of 1.42.

Two things follow. The curvature scaling is correct; the offset tracks $\sigma^2|H|$ over more than a decade of σ, but the extrapolation to $\sigma \to 0$ lands at 1.17 rather than 1. A residual of roughly 17% therefore survives the small-noise limit and is not curvature. We do *not* identify it with E_θ: the experiment cannot separate denoiser approximation from discretisation of the schedule, higher-order geometric terms, and finite-sample error in the covariance, and we interpret the discrepancy as a combination of these. What the measurement does establish is a magnitude; the estimator approximates P_T to within tens of per cent at the noise scales used, not to within per cent, and that magnitude is now on record.

I.3 Naive variance propagation, against ground truth

Proposition 2 argues that the affine variance recursion inflates. On S1, the argument can be checked against a quantity that is otherwise never available: the true per-pixel variance of the reverse chain, obtained by Monte-Carlo over the toy model, where nothing is approximated, and the manifold is known.

Eq. (9) is obtained here, and the protocol matters enough to give in full. We draw $J = 20$ unconditional DDPM samples from the S1 model (seeds 100–119, $T = 100$ ancestral steps). Along each chain, we propagate the diagonal variance with the affine additive recursion of App. A in its stochastic form, as those implementations do along a DDIM chain with $\eta = 1$ (the ancestral DDPM step is that case), using the model's own σ_θ head, under the independence assumption those implementations make. The empirical reference is the per-pixel variance across the J samples; the analytic quantity is the per-pixel propagated variance averaged over the J runs. Eq. (9) is the ratio of their means over the full 32×32 frame.

The single most important detail is that accumulation is *delayed-started at* $t = 60$ *of* 100, so the recursion is already running with the mask that published implementations apply. The overshoot of $13.7\times$ is therefore not what happens if one ignores the accepted remedy; it is what remains after applying it. Cor. 3 says the mask bounds the inflation rather than removing it, and this is the size of what it fails to remove.

Three caveats. $J = 20$ is a small Monte-Carlo estimate: its standard error is immaterial against a factor of 13.7, but it is not zero. The S1 model is unconditional, so the empirical variance is the spread of the generative distribution over the orbit rather than the predictive variance of a conditional forecast, the same object the recursion claims to track in this setting, but not identical to the conditional quantity of Sec. 4.2. And this is one manifold with one seed set; we report the magnitude, not a distribution over regimes.

I.4 On- and off-manifold behaviour

The probe input matters, and the toys say why. Displacing the query point off the manifold along the normal direction and measuring the residual drift $\|\bar{r}\|$ separates cleanly in both stages: by $4.54\times$ on S1 and $2.79\times$ on S2, with no overlap between the on- and off-manifold histograms in either case.

This is the toy version of a failure we later met on real data. Probing at $[x_{\text{baseline}}, 0]$ rather than at a reconstruction places the query far off the learned manifold, and the resulting map carries essentially

no information about the error ($\bar{\rho} = 0.004$). Probing at the model's own reconstruction restores it. The predicate in Prop. 1 that x_0 lie on $\mathcal{M}$ is not a formality; it is the whole method, and S1–S2 show the degradation is monotone in the distance to $\mathcal{M}$ rather than a cliff.

We note, as the medical sections do, that a DDIM reconstruction is only approximately on $\mathcal{M}$: the sampler is approximate, the score imperfect, and $\zeta = 5$ skips. The claim we make is the weaker and correct one, that it lands near enough for the projector reading to survive.

I.5 Reproducing

Both stages (S1 and S2 in App. I) are self-contained scripts of a few hundred lines, train their own model from scratch, run on a single consumer GPU or on CPU in under half an hour, and depend only on `torch` and `numpy`. Seeds are fixed at 42 throughout. Scripts and logs accompany the submission.

Experimental design and implementation.

J Cost accounting

The $\sqrt{\bar{\alpha}}$ in the diagonal of Sec. 3 follows from $\partial \hat{x}_0 / \partial x_t = \bar{\alpha}^{-1/2}\big(I + (1-\bar{\alpha})\nabla^2 \log p_t\big)$ and $\mathrm{Cov}(x_0 \mid x_t) = \frac{1-\bar{\alpha}}{\bar{\alpha}}\big(I + (1-\bar{\alpha})\nabla^2 \log p_t\big)$. Budgets are counted in forward-equivalent network evaluations and were fixed before any score was read. A Jacobian–vector product carries a primal and a tangent, so it costs two forwards of compute in principle; the conversion actually used is the measured wall-time ratio given at the end of this section. One DDIM reconstruction at skip ζ costs $T/\zeta + 1$ evaluations, 61 at $T = 300$, $\zeta = 5$. The MC-DDIM reference with J chains therefore costs $J(T/\zeta + 1) = 1{,}220$ at $J = 20$. T-PT costs $(T/\zeta + 1) + K$: 76 at $K = 15$ and 91 at $K = 30$, or $16\times$ and $13.4\times$ fewer. The row-sum surrogate is one JVP on top of the reconstruction; Hutchinson at M probes is M JVPs; the exact diagonal is d JVPs, one per input coordinate, $12{,}288$ at 64^2 RGB. JVPs are converted to forward-equivalents with a wall-time ratio *measured on each architecture* rather than the theoretical factor of two, which is right about memory and wrong about time, because forward-mode kernels are not the fused ones: $6.5\times$ on the 64^3 clinical model and about $4\times$ on the natural-image checkpoint, which is where the roughly $49{,}000$ forward-equivalents of Sec. 3.2 come from.

estimator	peak memory	wall time	one 128^3 volume
T-PT, K=30	33.8 GiB	66.7 s	completes
row-sum surrogate (one JVP)	$>$ 40 GB	—	out of memory
Hutchinson, M=50	$>$ 40 GB	—	out of memory
Hutchinson, M=500	$>$ 40 GB	—	out of memory

Table 5: Memory feasibility at 128^3 on a 40 GB accelerator (39.5 GiB usable), one estimator per invocation in a clean process, released implementations with activation checkpointing disabled for the tangent pass (App. K). The three Jacobian-based estimators exhaust memory before completing a single volume; the probe differentiates nothing and completes.

K Why forward-mode differentiation cannot traverse the public codebases

Forward-mode differentiation through the released implementations fails before it reaches the denoiser. All three wrap their residual blocks in PyTorch's activation checkpointing, whose `CheckpointFunction` is an old-style autograd function: it defines `forward` and `backward` but no `setup_context`, and `torch.func.jvp` requires the latter to push a tangent through a custom function. The call raises rather than silently returning a wrong tangent, which is the good failure. We disable checkpointing for the probe pass; nothing else about the network changes, and the primal outputs are identical to the released ones. This is an implementation obstacle, not a mathematical one, but it is the reason none of the three can be run as released.

L PROTOCOL FOR THE CHECKPOINT WITHOUT A VARIANCE HEAD

This checkpoint supports the head-free-compatibility claim, so its provenance should not live in a table footnote. Everything below is read from the training script and configuration shipped with the code.

Architecture and data are identical to the hybrid PNG checkpoint (Oliveras et al., 2026), same 3D U-Net (base width 64, multipliers $1, 2, 4, 8$, conditioning dimension 256), same two-channel joint state at 64^3, same training split and preprocessing, same $T = 300$ cosine schedule, with one change: `learned_variance = false`, so the network emits C output channels instead of $2C$ and the $\mathcal{L}_{\text{vlb}}$ term vanishes from the loss, which reduces to the ℓ_1 noise-prediction objective. Training ran 150k steps at batch size 4 on two GPUs with the repository's default optimiser settings; the training log confirms the loss dictionary contains no NLL term. At inference, the sampler is the one used everywhere else ($\zeta = 5$, $\eta = 1$); its $J = 20$ Monte-Carlo reference and the probe's chain both carry the observed blur difference in the conditioning vector, so this experiment also shows the probe under a second conditioning of the same models.

M OPERATIONAL DEFINITIONS OF MASKS AND METRICS

Every quantity below is computed per volume, inside the stated mask, exactly as implemented in the released code; nothing here is a post-hoc idealisation. The three thresholds (foreground fraction, change percentile, edge percentile) are the defaults fixed when the masking module was written, before any masked result was seen, and were not tuned afterwards.

Conditioning arm. The conditioning vector carries a blur difference computed from both scans of the pair. The Monte-Carlo reference and the analytical baseline were generated with that component zeroed, the value a forecasting deployment supplies, and so was the probe's chain on both tasks; every estimator in Table 1 is scored against the same common error map, built from the zero-blur ensemble, so the three are compared under one conditioning.

Clamping. Every estimate of the Tweedie diagonal is clamped at zero before scoring, since the implied variance can be negative; the rate on medical data is reported in App. O.

Masks. *Tissue (foreground):* voxels whose baseline intensity exceeds 2% of the volume's dynamic range, computed from the baseline only. On skull-stripped brain MRI, this recovers the brain ($\approx 24\%$ of voxels); on cropped lung CT, it removes only the air surrounding (99.5% retained), which is why the tissue row there is a consistency check. If the threshold leaves fewer than 100 voxels, the mask falls back to the full volume. The threshold is not load-bearing: rescored at 1% and 5% over the same 211 volumes, the participant-weighted primary contrast moves by $+0.0008$ and -0.0017 from its value at the prespecified 2%, and the gradient-controlled contrast by $+0.0009$ and -0.0017, with intervals of the same width: sign and magnitude are stable across a five-fold change of threshold. *Edges:* tissue voxels in the top 20% of $|\nabla u_0|$, the gradient magnitude by central differences, percentile per volume within tissue. *Interior:* tissue minus edges. *Regions of true change:* tissue voxels in the top 10% of $|x_{\text{target}} - u_0|$, percentile per volume within tissue. This uses the target and is therefore an *oracle evaluation slice*, not a mask available at deployment.

Partial Spearman. Within the mask, rank-transform σ, e and $|\nabla u_0|$; linearly regress the ranks of σ and of e on the ranks of $|\nabla u_0|$; report the Pearson correlation of the two residual vectors. This controls for the monotone component explained by gradient magnitude, no more. A synthetic map built from $|\nabla u_0|$ itself scores 0 under this control, which is the sense, and the only sense, in which "a pure edge tracer scores zero". If either variable is a deterministic function of the control, the statistic is reported as 0 rather than NaN.

Selective-prediction metrics. *AUSE:* remove voxels in decreasing σ in 20 steps over removal fractions 0 to 0.95; at each step record the mean error of the retained set, normalised by the mean error at zero removal; AUSE is the trapezoidal area between this curve and the oracle curve that removes by true error, with the random-ordering area reported as the scale. *AURC:* coverage from 0.05 to 1 in 20 steps retaining voxels in increasing σ; risk is the mean error of the retained set; the

oracle ordering is the attainable floor. *AUROC:* positives are the top 5% of error voxels within the mask, computed by the Mann–Whitney identity.

Configuration index. One place for every number needed to re-run the medical experiments; each entry is the value in the released launch scripts and configs.

Field	PNG (hybrid / head-free)	OASIS
Input resolution	64^3	128^3
Scored split	130 pairs, 47 participants	211 pairs, 63 participants
Forward schedule	cosine, $T = 300$	cosine, $T = 300$
Reverse sampler	DDIM, $\zeta = 5$, $\eta = 1$	DDIM, $\zeta = 5$, $\eta = 1$
MC reference	J=20, seeds 42–61	J=20, seeds 42–61
Self-chain probe	chain seed 42	chain seed 42
Conditioning c	$(\Delta\tau,\ \text{age},\ \Delta b)$: interval, baseline age, blur difference of the pair	
t^*, K	60, 30	60, 15
σ_θ baseline	propagated, $\hat{s}$=15; none for the head-free variant	propagated, $\hat{s}$=15
Preprocessing	source pipeline	source pipeline
Head-free variant	PNG only, retrained without the head (App. L)	—
Code, commit, versions	pinned in the released repository	

Table 6: Configuration index for the medical experiments. The unscored OASIS pairs are duplicated split entries or corrupted files, excluded before any scoring. Masks and metrics are defined operationally above; the statistical protocol is below and in App. R.

Statistical plan. The primary endpoint on OASIS is the participant-weighted paired difference of tissue Spearman: correlation per volume, mean over each participant's volumes, mean over participants, paired bootstrap resampling participants with $B = 2 \times 10^4$ and percentile intervals. The gradient-controlled partial Spearman is the robustness endpoint. Edges, interior, regions of change, AUROC, AUSE and AURC are secondary and unadjusted for multiplicity; their role is to show the direction is consistent, and any one alone should be read as exploratory.

Additional results and robustness checks.

N THE FOURTEEN CONDITIONS

Five natural-image checkpoints against three corpora would be fifteen combinations, and two are absent: `church/real` and `ffhq/real`. That leaves thirteen: `cond`, `inet` and `cifar` on all three corpora each, and `church` and `ffhq` on `self` and `coco`. The fourteenth is `brats/self`.

The medical checkpoint. `brats` is the MONAI model-zoo bundle `brats_mri_axial_slices_generative_diffusion` (Pinaya et al., 2023): a latent diffusion model over axial slices of the BraTS glioma MRI collection (Menze et al., 2015), with a KL autoencoder and a $1\times64\times64$ latent, so the exact diagonal is $4{,}096$ one-hot Jacobian–vector products per slice. It is read exactly as the bundle ships it: its own configuration instantiates both networks, the released weights load with no missing or unexpected key, the bundle carries a single weight set and no EMA copy, and it states no latent scale factor, so latents are taken at the autoencoder's native scale. The corpus is 100 latents drawn from the model's own prior by deterministic DDIM (seed 20260813), scored in latent space without decoding, since a decode–encode round trip would fold the autoencoder's reconstruction error into a measurement that concerns the denoiser's Jacobian. The generator is unconditional: there is no follow-up to predict, so it enters Table 7 and nowhere else. On it, the exact diagonal correlates with the error at $+0.320$ $[+0.298, +0.342]$, with no clamped voxel; Hutchinson at M=200 reproduces it at rank agreement 0.995; T-PT reads $+0.315$ $[+0.292, +0.337]$; and the row-sum $+0.015$. It is the one medical checkpoint in this paper whose weights the reader can download.

checkpoint	corpus	$\rho(\nabla, e)$	exact	Hutch. $M{=}200$	row-sum	T-PT $K{=}30$
cond	coco	0.389	**0.474**	0.465	0.036	**0.484**
cond	real	0.382	**0.476**	0.466	0.063	0.470
cond	self	0.340	**0.448**	0.435	0.004	**0.451**
brats	self	0.297	0.320	0.320	0.015	0.315
church	self	0.130	0.224	0.224	-0.009	0.218
church	coco	0.064	0.165	0.165	-0.007	0.167
inet	self	0.022	0.058	0.056	-0.017	0.013
ffhq	coco	0.006	0.054	0.054	-0.027	0.051
cifar	self	-0.007	0.029	0.023	-0.003	0.009
ffhq	self	-0.024	$0.002^{\dagger}$	$0.002^{\dagger}$	-0.059	-0.003
cifar	real	0.037	$\mathbf{-0.077}$	$\mathbf{-0.071}$	-0.169	$+0.018$
cifar	coco	0.054	$\mathbf{-0.091}$	$\mathbf{-0.085}$	-0.167	$+0.004$
inet	real	0.040	$\mathbf{-0.131}$	$\mathbf{-0.129}$	-0.215	$-0.014^{\ddagger}$
inet	coco	0.066	$\mathbf{-0.132}$	$\mathbf{-0.131}$	$\mathbf{-0.232}$	$-0.016^{\ddagger}$

Table 7: Spearman correlation with the denoising error, per condition, mean over 100 images. Rows are ordered by $\rho(\nabla, e)$, how much of the error is spatial structure. The block below the rule is where the exact diagonal is *anti*-correlated with the error. The Hutchinson column carries the same four harmful conditions as the exact column, with worst value -0.131 and mean over the fourteen 0.128. Every interval in both columns excludes zero except † (`ffhq/self`, $[-0.006, 0.010]$ for each). For T-PT, the only intervals that exclude zero on the negative side are the two marked ‡: `inet/real` $[-0.021, -0.007]$ and `inet/coco` $[-0.023, -0.009]$, on the checkpoint where the diagonal reverses hardest; neither lies entirely below the -0.02 threshold, whereas four exact-diagonal intervals do. Over the ten conditions above the rule, where the diagonal is not reversed, the exact column means 0.225, Hutchinson 0.221 and T-PT 0.218, and eight of the ten T-PT–exact differences lie within 0.010; the two exceptions are `inet/self` and `cifar/self`, where the exact diagonal itself reaches only 0.058 and 0.029. The exact column is clamped at zero wherever the implied variance is negative; the clamped fraction is zero on both pixel-space unconditional checkpoints, below 0.2% on the latent ones, and at most 5.7% on the class-conditional one.

estimator	$\bar{\rho}(\sigma, e)$	forward-equivalents
Hutchinson, $M{=}5$	0.298	≈ 20
Hutchinson, $M{=}15$	0.359	≈ 60
Hutchinson, $M{=}50$	0.410	≈ 200
Hutchinson, $M{=}200$	0.435	≈ 800
exact diagonal	0.448	$\approx 49{,}000$
T-PT, $K{=}30$	0.451	30

Table 8: The M sweep behind Sec. 3.2, on `cond/self` (the same 100 images as Table 7): Spearman correlation with the denoising error as the Hutchinson estimator converges to the exact diagonal. Forward-equivalents follow App. J. The pattern is the same in all fourteen conditions; at $M{=}200$ Hutchinson tracks the exact column of Table 7 to within 0.013 here, and at least 0.92 in rank agreement in every one of the fourteen.

O THE DIAGONAL REFERENCE AGAINST THE PROBE, IN FULL

Table 9 reports the interior mask with clamped voxels excluded, which is the condition least favourable to our claim. The ordering is unchanged in the other two. On the whole volume the probe retains $+0.450$ $[+0.367, +0.529]$ conditioned on the diagonal reference, against $+0.059$ $[+0.033, +0.085]$ the other way, with the row-sum at $+0.347$ and Hutchinson at $+0.024$. Inside the interior with clamped voxels retained the four figures are $+0.420$, $+0.064$, $+0.367$ and $+0.025$. Excluding the clamped voxels moves the reference's own correlation with the error from 0.364 to 0.361, so the clamping is not what holds it back. The clamped fraction is 7.5% of interior voxels on average and 17.4% on the worst of the twenty volumes, with estimation noise contributing (10.3% at $M{=}50$); σ_{TW} never clamps. Twenty volumes from twelve subjects throughout; intervals resample subjects. On the full cohort of 130 volumes, Hutchinson at $M{=}200$ on the probe's own chain (the

Estimator	functional of J	$\rho(\cdot, e)$	partial, given J_{ii}
Hutchinson, $M{=}50$	J_{ii}, unbiased	0.321	$+0.056$ $[+0.041, +0.070]$
Diagonal reference ($M{=}500$)	J_{ii}	0.361	—
Row-sum (Boys)	$\sum_j J_{ij}$	0.474	$+0.353$ $[+0.283, +0.420]$
T-PT, $K{=}30$	J_{ii} and off-diag. mass	**0.539**	$+\mathbf{0.409}$ $[+0.322, +0.489]$
Diagonal reference, given T-PT			$+0.056$ $[+0.032, +0.085]$

Table 9: Partial Spearman correlations with the error map, twenty 64^3 lung-CT volumes from twelve subjects, interior mask, clamped voxels excluded. The diagonal reference is Hutchinson at $M{=}500$; the one-hot exact diagonal is not affordable at this size. Intervals are a bootstrap over subjects. The last row is the reverse conditioning.

Table 1 column) has a split-half agreement of 0.79, clamps 5.5% of voxels, and trails the probe by $+0.189$ $[+0.158, +0.222]$ inside tissue and $+0.142$ $[+0.123, +0.161]$ on the gradient-controlled endpoint, participant-level intervals; its peak memory at 64^3 is $2.1\times$ the probe's and its wall time per volume $20\times$.

condition	exact vs. e	T-PT vs. e	T-PT, given exact	exact, given T-PT
inet/coco	-0.132 $[-0.164, -0.101]$	-0.016 $[-0.023, -0.009]$	$+0.005$ $[+0.002, +0.008]$	-0.131 $[-0.162, -0.100]$
inet/real	-0.131 $[-0.165, -0.098]$	-0.014 $[-0.021, -0.007]$	$+0.005$ $[+0.001, +0.008]$	-0.130 $[-0.164, -0.097]$
cifar/coco	-0.091 $[-0.117, -0.064]$	$+0.004$ $[-0.002, +0.011]$	$+0.007$ $[+0.001, +0.013]$	-0.091 $[-0.117, -0.065]$
cifar/real	-0.077 $[-0.105, -0.049]$	$+0.018$ $[+0.012, +0.024]$	$+0.020$ $[+0.014, +0.025]$	-0.077 $[-0.106, -0.049]$

Table 10: The same partial analysis on the four conditions where the exact diagonal is anti-correlated with the error, 100 images each, full image, intervals, a bootstrap over images. Here the conditioning variable is the one-hot exact diagonal itself, not a Hutchinson reference, and no voxel clamps. The first two columns reproduce Table 7. Removing the probe from the diagonal leaves the diagonal's reversal where it was; removing the diagonal from the probe leaves a small positive partial that excludes zero in every condition. These partials describe the finite-budget maps and do not, by themselves, establish complementary information in the population estimand. They are also not comparable in reliability with the clinical partials of Table 9: the attenuation bound $|\rho| \le \sqrt{\mathrm{rel}}$ places the clinical map's reliability at ≥ 0.45 and these four at ≤ 0.02, so the contrast between the two tables is a probe budget and not two verdicts on the functional.

condition	$a\overline{\|J_{i:}\|^2}/\overline{\lvert J_{ii}\rvert}$	$\rho(\|J_{i:}\|^2, e)$	$\rho(\sigma_{\mathrm{T2}}, e)$	$\rho(\sigma_{\mathrm{TW}}, \sigma_{\mathrm{T2}})$	T-PT, given σ_{T2}
inet/coco	0.34	$+0.135$ $[+0.104, +0.167]$	-0.131 $[-0.163, -0.099]$	0.141	$+0.005$ $[+0.002, +0.007]$
inet/real	0.34	$+0.134$ $[+0.100, +0.167]$	-0.130 $[-0.163, -0.096]$	0.143	$+0.005$ $[+0.001, +0.008]$
cifar/coco	0.28	$+0.093$ $[+0.066, +0.119]$	-0.091 $[-0.117, -0.064]$	0.023	$+0.007$ $[+0.001, +0.013]$
cifar/real	0.28	$+0.079$ $[+0.051, +0.106]$	-0.077 $[-0.105, -0.050]$	0.013	$+0.020$ $[+0.014, +0.026]$
cond/self	0.98	-0.442 $[-0.471, -0.413]$	$+0.431$ $[+0.402, +0.460]$	0.840	$+0.174$ $[+0.159, +0.189]$

Table 11: The row energy, read directly. The same exact-Jacobian pass accumulates $\|J_{i:}\|^2$ alongside J_{ii} on the four reversed conditions and, for contrast, on the self-conditioned ImageNet model (`cond/self`, the best-behaved condition of Table 7); 100 images each, full image, intervals a bootstrap over images. $\sigma_{\mathrm{T2}} = (1 - 2aJ_{ii} + a^2\|J_{i:}\|^2)^{1/2}$ is Eq. (4) evaluated exactly, with $a = \sigma_{t^*}$. The first column is the quadratic term against $\overline{|J_{ii}|}$; the linear term of Eq. (4) carries a further factor of two, so the quadratic term matches it at a column value of 2. On the reversed conditions the row energy is the mirror image of the diagonal ($\rho(\|J_{i:}\|^2, \sigma_{\mathrm{exact}})$ between -0.984 and -0.999), because $\|J_{i:}\|^2$ is dominated there by J_{ii}^2; the second-order map ranks identically to the exact diagonal ($\rho = 1.000$ on all four, 0.952 on `cond/self`) and reverses with it; and the probe retains conditioned on σ_{T2} exactly what it retains conditioned on J_{ii} (Table 10). The $K{=}30$ map correlates with that combination at only 0.01 to 0.14 on the reversed conditions, and Table 12 says why: at that budget the map is barely reproducible there. The low agreement is consistent with this poor reproducibility and cannot, by itself, establish a difference in population rankings. On `cond/self`, where the map is reproducible, it agrees at 0.840 and still retains $+0.174$ beyond it.

Is the neutrality the functional, or the probe budget? A near-zero correlation has two readings: the functional is uninformative where the diagonal reverses, or the map has not been resolved. Table 12 separates them with three budgets, $K = 30$, 120 and 480, on the four reversed conditions and on `cond/self` as the informative control, and with a direct measure of resolution. A *g-probe map* is the statistic of Eq. (2) computed from g of the K draws of one run; the *reliability at g* is the Spearman correlation between the g-probe maps built from draws 1 to g and from draws $g{+}1$ to $2g$, which share the image, the chain and the timestep and nothing else, averaged over the 100 images. It is the reproducibility of the map the body reports when $g = 30$. Under a null of independent coordinates that correlation has standard deviation $1/\sqrt{d}$, 0.018 at 32^2 and 0.009 at 64^2. Real maps are spatially dependent, so this is an order of magnitude for comparison and not an exact threshold; the values below are read as very low reproducibility rather than as tests against it.

The reading is the second. On `cond/self` the 30-probe map is already reproducible (0.86) and the correlation barely moves with budget. On the two `inet` conditions, reliability grows monotonically with g, from 0.014 at 15 to 0.118 at 240, and the correlation follows it towards the exact second-order map's -0.131: -0.016, -0.041, -0.079, each step's interval excluding zero, and the harmful threshold is crossed at $K \geq 120$ on `inet/coco` and at 480 on both. On the two `cifar` conditions, reliability stays at the noise floor through $g = 240$ and the correlation wanders within 0.02 of zero; at $g = 15$ and 30 the two-group agreement is slightly negative, which is what an unreproducible map gives and which we truncate at zero wherever a reliability enters a formula. As a consistency check, classical attenuation, $\rho(\sigma_K, e) \approx \rho(\sigma_{\mathrm{T2}}, e)\,\sqrt{\mathrm{rel}(K)}$ with $\mathrm{rel}(K)$ the Spearman–Brown extrapolation of the two-group agreement at $K/2$, reproduces the twelve points of the four reversed conditions within 0.02 and the three control points within 0.05. It is an approximate model of attenuation, and we do not read an identification of the population functional from its fit. Hutchinson at $M = 5$, about 20 forward-equivalents, is already reproducible enough (0.07–0.34) to be harmful on all four. The $K = 30$ map is therefore not neutral about the reversal; it is too unstable to say anything about it, and Hutchinson at a smaller budget is not.

condition	σ_{T2} vs. e	T-PT, $\rho(\sigma_K, e)$ [95% CI] K=30	K=120	K=480	Hutch. M=5 [95% CI]
cifar/coco	-0.091	$+0.004$ $[-0.002, +0.010]$	-0.014 $[-0.020, -0.008]$	-0.007 $[-0.016, +0.001]$	-0.043 $[-0.057, -0.029]$
cifar/real	-0.077	$+0.018$ $[+0.012, +0.024]$	-0.015 $[-0.021, -0.009]$	-0.009 $[-0.017, -0.001]$	-0.042 $[-0.057, -0.028]$
inet/coco	-0.131	-0.016 $[-0.023, -0.010]$	-0.041 $[-0.052, -0.030]$	-0.079 $[-0.099, -0.059]$	-0.104 $[-0.131, -0.077]$
inet/real	-0.130	-0.014 $[-0.021, -0.007]$	-0.031 $[-0.043, -0.020]$	-0.067 $[-0.087, -0.047]$	-0.101 $[-0.129, -0.074]$
cond/self	$+0.431$	$+0.451$ $[+0.422, +0.479]$	$+0.465$ $[+0.436, +0.493]$	$+0.471$ $[+0.443, +0.499]$	$+0.298$ $[+0.272, +0.326]$

Reliability: rank agreement of two disjoint g-probe maps (reference $1/\sqrt{d}$: 0.018 cifar, 0.009 inet, cond)

	g=15 / 30	g=60 / 120	g=240	Hutch. M=5
cifar/coco	-0.026 / -0.013	-0.001 / 0.015	0.002	0.070
cifar/real	-0.026 / -0.013	-0.002 / 0.014	0.000	0.067
inet/coco	0.014 / 0.017	0.027 / 0.064	0.118	0.340
inet/real	0.014 / 0.017	0.027 / 0.064	0.118	0.340
cond/self	0.755 / 0.858	0.918 / 0.955	0.976	0.121

Table 12: Probe budget on the reversed conditions, with `cond/self` as the informative control. Upper block: correlation with the error at three budgets from one run, each with its own bootstrap interval over the 100 images ($B = 2 \times 10^4$); a bracket belongs to the number on its left. The $K = 30$ column reproduces the exact-Jacobian pass to the last digit. Probe sets at different K are drawn independently. Lower block: reliability of a g-probe map, the Spearman agreement between the maps of two disjoint groups of g probes from the $K = 480$ draws, so $g = 30$ is the reproducibility of the map the body reports. The last column is Hutchinson at $M = 5$, about 20 forward-equivalents, level above and split-half reliability below. Two facts carry the appendix: at matched budget Hutchinson is harmful on all four and the probe on none, and given 4–16$\times$ the probes the probe's `inet` maps become more reproducible and harmful with them.

P THE OPEN QUESTION, AND TWO THINGS THAT ARE NOT CRITERIA

Candidate explanations of the two failure modes, all refuted. Accounts (i), (ii) and the fourth below concern the natural-image sign; account (iii) concerns the clinical change region. (i) That the Tweedie conversion assigns low variance where the denoiser is most responsive, at edges and texture: $\mathrm{diag}(J)$ is in fact *lower* where the intensity gradient is high in thirteen of the fourteen conditions, the opposite

sign, and removing the gradient from both variables leaves the correlation essentially unchanged. (ii) That the natural-image sign was a domain artefact: refuted by the real-image corpora (Sec. 3.3). (iii) That the error inside the change region is less predictable from image structure, so no variance estimate could track it: on lung CT this holds, as $\rho(\nabla, e)$ falls from 0.598 to 0.071; on brain MRI it does not (0.230 to 0.197, interval covering zero), and σ collapses anyway. The drop is not a property of the 10% threshold that defines the region: on lung CT, at the top 5%, 10% and 20% of true change, the probe reads 0.054, 0.100 and 0.233 against MC-DDIM's 0.169, 0.265 and 0.388, paired gaps -0.092, -0.134 and -0.117, every interval excluding zero; on brain MRI the two read 0.101, 0.128, 0.106 and 0.075, 0.133, 0.129, all far below the tissue-wide 0.349 and 0.337, with the ordering between them changing sign across thresholds ($+0.028$, -0.003, -0.029). Two mechanisms remain consistent with the change-region drop and are untested: conditioning on the baseline scan may suppress variance exactly where the anatomy moves, since the model is asked for a follow-up that stays close to what it was shown; and the model may be confident on progression trajectories it has not seen, in which case a confident prediction receives low variance whether or not it is right. The second is not specific to the probe: every sampling-based estimator here, the reference included, reads the model's uncertainty and not its truthfulness (App. F). The volume-level coupling between the two drops is positive on both tasks but its interval covers zero.

A fourth candidate fails as well: the reversal does not coincide with the region where the implied diagonal becomes negative. On the four reversed conditions, the exact diagonal is never clamped (the clamped fraction is 0.000 on every image), whilst the conditions that do clamp are the well-behaved ones, and within those the clamped fraction associates with better ranking, not worse.

A principled reading of the row energy is available, and on natural images it has now been tested. The row energy $\|J_{i:}\|^2$ measures the total sensitivity of output voxel i to perturbations of the input; where the denoiser amplifies its input, a small mismatch between model and data becomes a large reconstruction error, so voxels of high row energy would be voxels where errors concentrate. Read directly (Table 11), the row energy does correlate positively with the error on the four reversed conditions, $+0.079$ to $+0.135$, but only because it mirrors the diagonal there; on the self-conditioned model it reads -0.442 where the diagonal reads $+0.448$. This matters beyond our own estimator, because the row energy is what a probe that takes the variance of the *scores* rather than of the residual reads to second order, as De Vita & Belagiannis (2025) do: such a probe does not escape the reversal; inverts the diagonal on every condition we measured it, and so ranks the error correctly on the conditions where the diagonal fails and incorrectly on those where it works. Their estimate is moreover identified with the Fisher information of the *marginal* noising distribution, an object with no x_0 in it, while the object a per-image map requires is the conditional $C_t(x_0)$ of Sec. 4.2. The row energy is not what the probe reads on either: the probe's partial given the row energy equals its partial given the diagonal in every condition. On lung CT, where the diagonal is only available as a Hutchinson reference, the direct test remains open.

Nor is the ordering of Table 7 a criterion. That table is ordered by $\rho(\nabla, e)$, which costs one forward pass, and the ordering is visibly related to the column beside it: on the five conditions where we first looked, the rank correlation between the two was 1.000; on all fourteen it is 0.640, and on real images alone 0.405. It does not predict the sign either; `inet/coco` has a middling $\rho(\nabla, e)$ and the most negative correlation in the table. We report it as an observation about these fourteen conditions and specifically not as a criterion, because the perfect ordering was an artefact of the five points that produced the hypothesis.

Q Qualitative comparison

Aggregate correlations hide what a clinician would actually look at. Figure 3 shows one test case per dataset: baseline, target, and absolute error on top, the three σ maps as alpha-blended overlays on the target anatomy below, each normalised to its own quantiles. The overlay is deliberate. A σ map is only useful when read *against* the anatomy it annotates; side-by-side heatmaps make that harder than it needs to be.

On lung CT (Fig. 3, left) all three estimators put their mass on the growing nodule and the vessels next to it, and leave the static parenchyma dark. Without the labels, σ_{MC} and σ_{TW} are hard to tell apart, which is what cross-agreement $\rho = 0.79$ looks like on a single volume. On brain MRI (Fig. 3, right) everything is more diffuse: uncertainty follows the cortical ribbon and the ventricular boundaries,

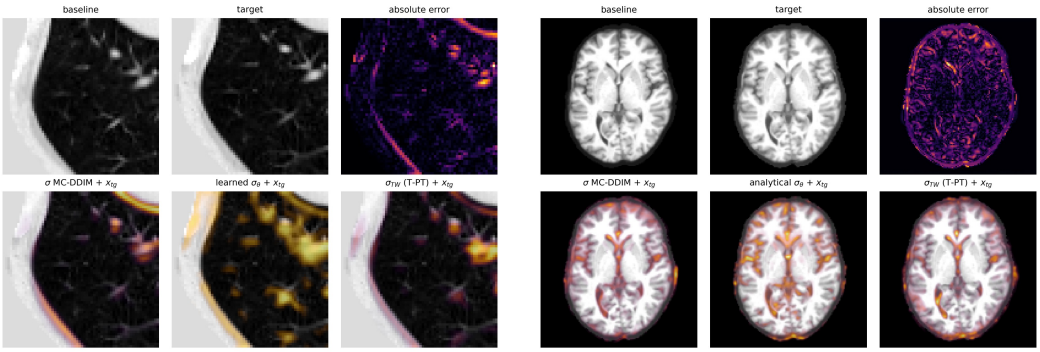


Figure 3: Representative lung-CT sample (**left**) and OASIS-3 transverse slice (**right**). Top row of each: baseline, target, absolute error. Bottom row, overlaid on the target: MC-DDIM σ, the analytical σ_θ, and T-PT σ_{TW}. Each overlay is normalised to its own quantiles, so the panels show each map's spatial ordering and not comparable magnitudes. On lung CT all three concentrate on the nodule region and T-PT reproduces MC-DDIM's spatial pattern (cross-agreement $\rho = 0.79$ on this volume) at $13.4\times$ fewer network evaluations. On brain MRI σ_{TW} tracks the ventricular and cortical boundaries consistently with MC-DDIM ($\rho = 0.70$ here, 0.709 over the cohort) and inside brain tissue it leads the ensemble on five of the eight endpoints of Table 1, the rest unresolved. What no panel shows, and Sec. 6 reports, is that both maps lose most of their information inside the region that actually changed between the two visits.

where skull-stripping and atrophy both introduce ambiguity. Again, the two maps pick out the same structures.

Both samples have per-volume error correlations below their dataset means ($\rho_{\text{TW}} = 0.601$ lung, 0.499 brain), and per-volume cross-agreements close to the dataset means (0.79 vs 0.815; 0.70 vs 0.709): representative volumes, not favourable ones. We picked them for anatomical clarity, not for favourable numbers. Read the figures for spatial behaviour; Table 1 carries the common-error-map comparison; Table 14 reports the deployment variant.

R THE DEPLOYMENT COMPARISON, AND PAIRED TESTS

Table 1 in the body scores every estimator against a single common error map (Table 16 below gives the full per-region and selective-prediction rows), that of the $J = 20$ ensemble mean, so that differences reflect the uncertainty estimator alone. That is the scientific comparison. It is not, however, the comparison a practitioner faces: at deployment T-PT annotates the single chain that produced its probe input, while MC-DDIM annotates its J-chain mean, and a J-chain mean is the better prediction. Table 14 reports that operating comparison for completeness. It conflates estimator quality with prediction quality, which is why it is here and not in the body. *Deployment* qualifies the scoring and nothing else: each map is ranked against the error of the prediction it would accompany rather than against a common reference no practitioner holds; the conditioning is the deployable one on both tasks (App. M).

probe input x_0 (lung CT, N=130)	source	$\bar{\rho}(\sigma_{\text{TW}}, e)$
$[x_{\text{baseline}}, 0]$, zero placeholder in the residual channel	first implementation	0.004
one DDIM reconstruction (deployment probe)	Table 14	0.648

Table 13: The probe point is load-bearing (Sec. 4.3): the same estimator, at the same $t^* = 60$ and $K = 30$, scored against the error of the chain each map accompanies — the deployment scoring defined above, not the common error map — with only the probe input changed. Both rows are per-volume means, so the two are read against each other; the participant-weighted figure for the deployment probe is the 0.613 $[0.576, 0.650]$ of Table 14. A placeholder that no training pair ever visited yields a smooth, non-negative map that carries no information about the error; any model-supported reconstruction restores it.

Dataset	Estimator	Fwd.	$\bar{\rho}(\sigma, e)$ [95% CI]	cross $\leftrightarrow$ MC	Requires
PNG	σ_{MC} (J=20)	1220	0.642 [0.596, 0.687]	$\equiv 1$	—
N=130	σ_θ (analytical, $\hat{s}$=15)	61	0.597 [0.551, 0.641]	—	hybrid loss
	σ_{TW} (T-PT)	**91**	0.613 [0.576, 0.650]	0.790 [0.765, 0.812]	—
OASIS	σ_{MC} (J=20)	1220	**0.552** [0.538, 0.564]	$\equiv 1$	—
N=211	σ_θ (analytical, $\hat{s}$=15)	61	**0.551** [0.539, 0.562]	—	hybrid loss
	σ_{TW} (T-PT)	**76**	0.520 [0.511, 0.529]	0.709 [0.705, 0.712]	—

Table 14: Predictive discrimination in the deployment setting: each estimator is scored against the error of the prediction it would actually accompany, so T-PT is measured against a single chain and MC-DDIM against its J-chain mean. "Fwd." is network evaluations per volume; CI is confidence interval. σ_θ accumulates along its own single chain, so its 61 are that chain itself, with no evaluations beyond it. σ_{MC} is the reference, so its cross-agreement with itself is 1 by construction; entries are participant-weighted with participant-level bootstraps on both tasks. The two σ_θ rows are scored against the common error map (their cross-agreement with MC-DDIM was not computed); on lung CT the analytical baseline sits below both estimators, and T-PT needs no variance head to run, at $13.4\times$ fewer network evaluations than the reference. On brain MRI, over the whole volume, MC-DDIM is ahead of T-PT by 3.2 Spearman points in this self-scored setting (about four against the common error map); Table 1 shows the ordering reversing inside tissue, which is where the background stops carrying the correlation. The tissue rows of both analytical baselines are in Table 17.

Paired tests. Overlapping intervals do not establish that two estimators differ, and non-overlapping ones do not establish by how much. Since every estimator is scored on the same volumes, we bootstrap the per-volume difference directly, preserving the pairing (2×10^4 resamples), and we resample participants rather than volumes as specified in Sec. 5 and validated below on simulated data.

The analytical baseline. σ_θ is produced by a single reverse chain: the sampler accumulates the variance head along the chain with the delayed-start mask at $\hat{s} = 15$, so one chain per volume (seed 42, the first reference seed) is the whole computation. It is scored against the same common error map as every other estimator (Table 17); Table 14 gives its full-volume figure next to the deployment-scored rows. On lung CT the ordering is simple: the variance head is below both estimators on every endpoint, T-PT by $+0.024$ $[+0.015, +0.034]$ in tissue Spearman and the reference by $+0.046$. On brain MRI, it is not. The result is the paper's thesis in one row. Over the full volume the propagated variance is not distinguishable from the J=20 ensemble (0.553 vs 0.559; paired $\Delta = +0.001$ $[-0.004, +0.006]$) while sitting *above* T-PT (0.518). Inside tissue it is last by a wide margin, and the gradient control, AUROC and AUSE agree. Its median voxel sits at 3.0×10^{-3}, of the same order as the schedule's own accumulated injection at the mask, with the anatomy modulating scarcely a factor of two above it. Those magnitudes are a description and not an explanation of the ranking: the injected term enters the recursion as a schedule scalar added identically at every voxel, so it cannot reorder them: no constant additive term across voxels can move a rank correlation (Prop. 8, App. G). What the tissue restriction removes is the agreement that background supplies for free, where map and error are both near zero; whether what remains is dominated by Monte-Carlo noise is a question a split-half of the seeds inside tissue would answer, and we have not run it. At 61 network evaluations against T-PT's 76, the fifteen additional probe forwards buy a participant-weighted $+0.050$ tissue-Spearman contrast.

Deployment without a reference ensemble. The paired comparisons of Tables 1 and 15 score σ_{TW} against the common ensemble-mean error map, which a deployed system does not have. The honest deployment quantity scores the probe against the error of its own reconstruction. Doing so on the same 211 volumes moves the whole-volume Spearman *up*, from 0.518 to 0.523 (per-volume means; the participant-weighted figure of Table 14 is 0.520; $+0.0063$ $[+0.0018, +0.0110]$), and the worst-5% AUROC up from 0.782 to 0.793 ($+0.0092$ $[+0.0061, +0.0125]$), because the probe and the error map now share a reconstruction and agree on the background. Inside tissue it moves *down*, from 0.349 to 0.343 (-0.0092 $[-0.0137, -0.0045]$), and it moves down by a similar amount in the interior (-0.0101), in the regions of change (-0.0138) and on the gradient-controlled endpoint (-0.0089). The primary endpoint therefore costs about one Spearman point when the reference is removed, and the whole-volume figure is the one that should not be trusted: it improves for the same

Metric	participant-wtd. $\bar{\Delta}$ [CI]	volume-wtd. $\bar{\Delta}$ [CI]	T-PT wins
Lung CT *(PNG, N=130, G=47)*			
ρ, full volume	-0.0215 $[-0.0373, -0.0070]$	-0.0242 $[-0.0376, -0.0104]$	36/130
ρ, tissue	-0.0215 $[-0.0368, -0.0069]$	-0.0242 $[-0.0375, -0.0103]$	36/130
ρ, edges	-0.1465 $[-0.1865, -0.1096]$	-0.1760 $[-0.2102, -0.1295]$	17/130
ρ, interior	-0.0234 $[-0.0435, -0.0048]$	-0.0210 $[-0.0413, -0.0062]$	50/130
ρ, regions of change	-0.1337 $[-0.1821, -0.0860]$	-0.1656 $[-0.1945, -0.1171]$	30/130
ρ, partial, $\lvert\nabla u_0\rvert$ ctl.	-0.0595 $[-0.0900, -0.0307]$	-0.0696 $[-0.0930, -0.0417]$	33/130
AUROC, worst 5%	-0.0274 $[-0.0388, -0.0167]$	-0.0376 $[-0.0505, -0.0199]$	28/130
AUSE (lower better)	$+0.0140$ $[+0.0041, +0.0239]$	$+0.0188$ $[+0.0076, +0.0275]$	41/130
AURC (lower better)	$+0.0013$ $[+0.0003, +0.0023]$	$+0.0017$ $[+0.0006, +0.0025]$	41/130
Brain MRI *(OASIS, N=211, G=63)*			
ρ, full volume	-0.0380 $[-0.0415, -0.0345]$	-0.0412 $[-0.0448, -0.0372]$	1/211
ρ, tissue	$+0.0165$ $[+0.0057, +0.0282]$	$+0.0120$ $[+0.0032, +0.0215]$	117/211
ρ, edges	$+0.0027$ $[-0.0047, +0.0102]$	$+0.0030$ $[-0.0053, +0.0115]$	103/211
ρ, interior	$+0.0221$ $[+0.0089, +0.0369]$	$+0.0164$ $[+0.0064, +0.0274]$	120/211
ρ, regions of change	-0.0025 $[-0.0132, +0.0082]$	-0.0049 $[-0.0122, +0.0029]$	99/211
ρ, partial, $\lvert\nabla u_0\rvert$ ctl.	$+0.0212$ $[+0.0090, +0.0350]$	$+0.0167$ $[+0.0070, +0.0270]$	123/211
AUROC, worst 5%	-0.0014 $[-0.0065, +0.0041]$	-0.0004 $[-0.0062, +0.0058]$	82/211
AUSE (lower better)	-0.0108 $[-0.0188, -0.0036]$	-0.0084 $[-0.0151, -0.0024]$	106/211
AURC (lower better)	-0.0010 $[-0.0016, -0.0004]$	-0.0009 $[-0.0015, -0.0004]$	106/211

Table 15: Paired participant-level bootstrap of $\sigma_{\text{TW}} - \sigma_{\text{MC}}$ against the common error map, on both tasks: the 130 lung-CT pairs from 47 participants and all 211 OASIS volumes from the 63 participants of the published test split. Δ is T-PT minus MC-DDIM on every row, so on AUSE and AURC, a positive value favours MC-DDIM. The participant-weighted column averages the difference within a participant before averaging across participants, so a participant contributing 21 of the 211 pairs counts once rather than 21 times; the volume-weighted column averages over volumes directly. Both bootstraps resample participants. The two agree in sign on every row of both tasks. The last column counts the volumes on which T-PT is better, with the direction of each metric taken into account. Tissue Spearman is the pre-designated primary endpoint and the gradient-controlled row its robustness check (App. M); the remaining rows are secondary, unadjusted for multiplicity, and any one of them alone should be read as exploratory. CI is the confidence interval.

reason it was uninformative to begin with. The probe-seed rerun makes the same point from the other side: changing the probe seed from 42 to 777 moves the self-scored whole-volume figure by $+0.08$ and the tissue figure by 0.006 (Table 18; the number we say not to trust is also the only one sensitive to the probe seed.

Probe-seed sensitivity, and the no-residual baseline. The deployment probe chain uses seed 42, which is also one of the twenty reference seeds (42–61). To rule out any dependence between the probe and the reference, we reran the full masked pass with a probe seed outside the reference set, 777 on brain MRI, 123 on lung CT, on both datasets (Table 18). Every primary contrast moves by at most 0.002 on brain MRI and 0.006 on lung CT, with near-identical intervals, and cross-agreement is unchanged (0.709 under both seeds on OASIS, 0.790 vs 0.784 on PNG), so the shared seed neither drives the tissue advantage nor inflates the agreement with the reference. The same rerun carries the no-residual control requested by the construction itself: the per-voxel variance of the raw noise predictions over the same K perturbations, identical probes, no subtraction of the injected noise, which is the statistic of De Vita & Belagiannis (2025) evaluated at our probe point and budget rather than in their pipeline. It correlates with the common error map at -0.010 (OASIS tissue) and -0.019 (PNG tissue), against T-PT's 0.353 and 0.618; the paired gaps are $+0.363$ $[+0.345, +0.381]$ and $+0.637$ $[+0.592, +0.681]$. (The standard-deviation map of $\hat{x}_0$ over the same probes needs no experiment: $\hat{x}_0 = x_0 + \sigma_t/\sqrt{\bar{\alpha}_t}\,(\xi - \epsilon_\theta)$, so it is the T-PT map times a schedule constant, identical in rank.) What carries the signal is the subtraction of the known noise, the Tweedie residual, not the local perturbations.

The pattern is the one Sec. 5.2 describes, now with the resampling unit corrected. Over the whole volume, MC-DDIM is significantly better; inside tissue the ordering reverses, in the interior it reverses by more, and the reversal survives partialling out the baseline intensity gradient, which is the control

	PNG (N=130)		OASIS (N=211)	
	σ_{TW}	σ_{MC}	σ_{TW}	σ_{MC}
ρ, full volume	0.669	**0.693**	0.518	**0.559**
ρ, tissue	0.670	**0.694**	**0.349**	0.337
ρ, edges	0.352	**0.528**	0.248	0.245
ρ, interior	0.612	**0.633**	**0.334**	0.318
ρ, regions of true change	0.100	**0.265**	0.128	0.133
ρ, partial, $\lvert\nabla u_0\rvert$ controlled	0.339	**0.409**	**0.290**	0.274
AUROC, worst 5% voxels	0.884	**0.922**	0.782	0.782
AUSE (lower is better)	0.152	**0.133**	**0.306**	0.315
AURC (lower is better)	0.030	**0.029**	**0.053**	0.054

Table 16: Full per-region and selective-prediction results against the common error map of the $J = 20$ ensemble mean, expanding the compact Table 1 of the body. OASIS entries are per-volume means, shown for scale, following the convention of Table 1; the pre-declared participant-weighted differences and their intervals for OASIS are in Table 15; the PNG paired contrasts are the Δ column of Table 1. Bold marks the better of σ_{TW} and σ_{MC} where the 95% paired interval excludes zero (participant-level resampling on both tasks); intervals are in Table 15. "Tissue" is the foreground mask. On lung CT it covers 99.5% of the volume; the scan is already cropped to the thorax, so foreground and full volume coincide, and the row is a consistency check rather than a result. On brain MRI it covers 24%, and there the two rows differ sharply. The σ_θ columns for both tasks are in Table 17.

that would kill a pure anatomy tracer. Edges, the change region and AUROC on the worst 5% of voxels are the contrasts we cannot resolve.

We report the weighting we do because the split is unbalanced, and it is worth being explicit that the imbalance was working *against* the effect rather than creating it: the participant contributing 21 of the 211 pairs carries a tenth of a volume-weighted estimate, and removing that leverage strengthens the primary and robustness contrasts rather than weakening it. A reader who prefers volume weighting can read the last column and reach the same inferential conclusion on all nine rows: the two weightings agree in sign on every one of them, and where they differ in magnitude, both intervals cross zero.

Why the resampling unit matters. Resampling volumes rather than participants is not conservative in one direction; it is simply wrong, and it is wrong in the direction that manufactures significance. We checked this on simulated data matched to our cluster sizes, 20 participants with 1 to 11 volumes each, a true between-estimator difference of 0.05 carrying a participant-level component, and repeated it 300 times. A volume-level bootstrap covered the true effect in 57% of the replicates against a nominal 95%; the participant-level bootstrap covered 95% (94.7% measured). On a single replicate, the two disagree qualitatively: the volume-level interval $[+0.0147, +0.0358]$ excludes zero, while the participant-level interval on the same data, $[-0.0056, +0.0546]$, contains it. A reader checking our contrasts should confirm that the interval excludes zero under the participant-level resampling, which is the one we report.

Two readings follow, and they pull in opposite directions. T-PT does *not* match MC-DDIM on lung CT: it is significantly worse, by a margin the bootstrap resolves cleanly, and any claim of parity there would be unsupported. On brain MRI it is behind MC-DDIM over the whole volume and ahead of it inside tissue (Table 15). The defensible summary is that T-PT retains most of MC-DDIM's spatial discrimination at a fraction of the cost, and that which wins depends on the dataset *and* on where in the volume one looks.

Head-free compatibility, with the confound removed. Expressed as a fraction of each model's own Monte-Carlo reference, T-PT recovers 95.6% [93.9, 97.9] on the hybrid checkpoint and 99.5% [96.0, 102.4] on the one trained without a variance head, resampling the 47 participants. The difference between the two ratios is +3.9 points $[+0.4, +5.5]$ in favour of the model trained without a variance head. T-PT therefore tracks its reference *at least as closely* without the variance head as with it, which is the strongest form the training-invariance claim could take, and we flag that we

	σ_θ (analyt.)	σ_{TW}	σ_{MC}
Lung CT *(PNG, N=130, G=47)*			
ρ, full volume	0.643	0.669	**0.693**
ρ, tissue	0.644	0.670	**0.694**
ρ, partial, $\lvert\nabla u_0\rvert$ ctl.	0.295	0.339	**0.409**
Network evaluations / volume	61	91	1220
Paired contrasts against σ_θ (participant-weighted, $B = 2 \times 10^4$):			
$\Delta\rho$, tissue	—	$+0.0241$ $[+0.0148, +0.0336]$	$+0.0456$ $[+0.0307, +0.0607]$
$\Delta\rho$, partial (grad. rm.)	—	$+0.0574$ $[+0.0405, +0.0747]$	$+0.1169$ $[+0.0908, +0.1453]$
$\Delta\rho$, full volume	—	$+0.0238$ $[+0.0142, +0.0333]$	$+0.0453$ $[+0.0306, +0.0610]$
ΔAUROC, worst 5%	—	$+0.0501$ $[+0.0446, +0.0557]$	$+0.0775$ $[+0.0639, +0.0923]$
ΔAUSE (lower better)	—	-0.0260 $[-0.0317, -0.0203]$	-0.0400 $[-0.0518, -0.0283]$
Brain MRI *(OASIS, N=211, G=63)*			
ρ, full volume	0.553	0.518	**0.559**
ρ, tissue	0.299	**0.349**	0.337
ρ, partial, $\lvert\nabla u_0\rvert$ ctl.	0.253	**0.290**	0.274
Network evaluations / volume	61	76	1220
Paired contrasts against σ_θ (participant-weighted, $B = 2 \times 10^4$):			
$\Delta\rho$, tissue	—	$+0.0496$ $[+0.0370, +0.0627]$	$+0.0331$ $[+0.0234, +0.0433]$
$\Delta\rho$, partial (grad. rm.)	—	$+0.0355$ $[+0.0229, +0.0486]$	$+0.0142$ $[+0.0060, +0.0224]$
$\Delta\rho$, full volume	—	-0.0372 $[-0.0401, -0.0341]$	$+0.0008$ $[-0.0040, +0.0058]$
ΔAUROC, worst 5%	—	$+0.0274$ $[+0.0223, +0.0325]$	$+0.0288$ $[+0.0229, +0.0350]$
ΔAUSE (lower better)	—	-0.0391 $[-0.0480, -0.0307]$	-0.0283 $[-0.0355, -0.0215]$

Table 17: The analytical variance-propagation baseline on both tasks, from one seed-42 chain per volume and scored against the common J=20 ensemble-mean error map. Estimator rows are per-volume means, following the convention of Table 1; contrast rows are the participant-weighted paired difference (column estimator minus σ_θ), with brackets giving 95% confidence intervals (CI), where positive favours the column estimator except for AUSE, where negative does. AUROC is reported for the worst-5%-error voxels. On lung CT, both estimators beat the variance head on every endpoint, the reference by more. On brain MRI, inside tissue both estimators beat it and T-PT by more; over the whole volume neither does, the T-PT contrast being negative and the MC one crossing zero. In the OASIS regions of change (not shown) no contrast against σ_θ excludes zero in either direction.

		probe seed 42	alternative probe seed
OASIS	$\Delta\rho$, tissue	$+0.0165$ $[+0.0056, +0.0282]$	$+0.0149$ $[+0.0038, +0.0268]$
	$\Delta\rho$, partial (grad. rm.)	$+0.0212$ $[+0.0088, +0.0348]$	$+0.0193$ $[+0.0068, +0.0332]$
	$\Delta\rho$, full volume	-0.0380 $[-0.0415, -0.0346]$	-0.0374 $[-0.0408, -0.0339]$
	$\rho(\sigma_{\text{TW}}, \sigma_{\text{MC}})$	0.709	0.709
PNG	$\Delta\rho$, tissue	-0.0215 $[-0.0372, -0.0069]$	-0.0256 $[-0.0428, -0.0093]$
	$\Delta\rho$, partial (grad. rm.)	-0.0595 $[-0.0896, -0.0306]$	-0.0650 $[-0.0960, -0.0353]$
	$\Delta\rho$, full volume	-0.0215 $[-0.0369, -0.0072]$	-0.0255 $[-0.0433, -0.0095]$
	$\rho(\sigma_{\text{TW}}, \sigma_{\text{MC}})$	0.790	0.784

Table 18: Probe-seed sensitivity: the masked pass run with the deployment probe seed (42) and with a probe seed outside the reference seed set (42–61): seed 777 on OASIS, seed 123 on lung CT, everything else held fixed, on the deployment conditioning and against the same references as Table 1. $\Delta\rho$ is the paired difference, T-PT minus MC-DDIM, of the per-volume Spearman correlation with the common error map; the last row of each block is the Spearman correlation between the two σ maps themselves. The table is read for the seed-to-seed movement only; contrasts are participant-weighted with participant-level intervals ($B = 10^4$) on both tasks. Every primary contrast moves by at most 0.006.

reached it only after removing a confound: scoring T-PT against its own single-chain error while scoring MC-DDIM against the ensemble-mean error produces an apparent 86.6% against 92.6% and reverses the conclusion.

Checkpoint	Contrast Δ	$\bar{\Delta}$ [95% CI]	$P(\Delta > 0)$
PNG hybrid	$\sigma_{\text{TW}} - \sigma_{\text{MC}}$	-0.022 $[-0.037, -0.007]$	$< 10^{-3}$
PNG no-head	$\sigma_{\text{TW}} - \sigma_{\text{MC}}$	-0.020 $[-0.040, -0.001]$	0.02

Table 19: Paired bootstrap on the two lung-CT checkpoints against the error map of each checkpoint's own J=20 ensemble, resampling the 47 participants ($B = 2 \times 10^4$); the conditioning of each experiment is given in App. M and App. L. MC-DDIM is ahead of T-PT by about two Spearman points on both checkpoints; on the one trained without a variance head, the interval only just excludes zero. The sample-level bootstrap used in an earlier draft gave -0.003 $[-0.016, +0.009]$ for that checkpoint, which is the coverage failure App. R documents. The OASIS contrasts are in Table 15.

The mechanism is worth stating plainly, because it is not that the estimator improves. Removing the variance head produces a noisier sampler: mean absolute error rises from 0.064 to 0.069 for the ensemble mean and from 0.064 to 0.086 for a single chain, and σ_{MC} is 2.7× larger at the same J. Both estimators sit lower on that checkpoint: the ensemble at 0.636 and the probe at 0.633 against the ensemble's own reference, and the paired gap between them is -0.020 $[-0.040, -0.001]$. What the experiment establishes is that no part of T-PT depends on the hybrid objective; it neither needs σ_θ to exist nor pays a penalty when it does not.

S What a larger Monte-Carlo ensemble buys

Sub-sampling the complete $J = 20$ ensembles, at $J' \in \{2, 5, 10, 20\}$ on lung CT and $J' \in \{2, 3, 5, 8, 10, 15, 20\}$ on OASIS, and rescoring gives two quantities: discrimination against the error map, and the split-half agreement obtained by splitting the J' chains into two disjoint halves and correlating the two σ maps. The first is how informative the estimate is, the second how reproducible.

	PNG (N=130)		OASIS (N=211)			
J'	$\bar{\rho}(\sigma_{J'}, e)$	split-half	$\bar{\rho}(\sigma_{J'}, e)$	agree w/ J=20	split-half	$\bar{\rho}_{\text{tissue}}(\sigma_{J'}, e)$
2	0.529 (77%)	—	0.485 (91%)	0.668	—	0.139 (43%)
3	—	—	0.521 (98%)	0.757	—	0.195 (60%)
5	0.653 (95%)	0.496	0.529 (99%)	0.827	0.535	0.248 (77%)
8	—	—	0.531 (100%)	0.887	0.665	0.284 (88%)
10	0.676 (98%)	0.768	0.532 (100%)	0.915	0.684	0.296 (91%)
15	—	—	0.532 (100%)	0.964	0.712	0.314 (97%)
20	0.687 (100%)	0.863	0.534 (100%)	1.000	0.740	0.323 (100%)

Table 20: Percentages are relative to $J' = 20$. Each entry averages 20 subsets of size J' drawn at random without replacement, so the curve does not depend on the order of the seeds. Entries are *volume-weighted* means computed by the standalone convergence probe directly in prediction space, so its absolute values are not comparable with the scored pipeline of Tables 1–16; only the percentages, which are internal to this table, are meaningful here. The last column is the finding: full-volume discrimination saturates almost immediately ($J' = 3$ recovers 98%), but that saturation is carried by background. *Inside tissue*, the primary endpoint, the same reference recovers only 60% at $J' = 3$, 91% at $J' = 10$, and is still climbing at $J' = 20$. Among the tested configurations $J' \leq 20$, no smaller ensemble reaches the J=20 tissue discrimination, which is itself still increasing.

Fifty chains: the scores are insensitive to the target, the reference's own spread is not. Table 20 extrapolates; this rescoring measures. We generated thirty further reverse chains per volume, seeds 62–91, under the deployment conditioning, and rescored against the resulting J=50 ensemble: every lung-CT volume, and the first thirty brain volumes, which is what the budget allowed. Two things separate, and only the second was in doubt. The mean absolute error of the ensemble mean is unchanged (0.0734 → 0.0732 on lung CT, 0.0260 → 0.0261 on brain MRI), and the statement the tables actually need holds directly: the probe's map is fixed by construction, and scoring that one fixed map against e_{20} and against e_{50} gives 0.670 → 0.671 on lung CT and 0.356 → 0.356 inside brain tissue (Table 21). Equal correlations with two targets do not make the two targets equal, so this is not a convergence claim about the error map; it is the narrower one the body relies on, that the

discrimination it reports for T-PT is unchanged when the twenty-chain error map is replaced by the fifty-chain one. What is not converged at $J{=}20$ is MC-DDIM's own σ, which improves by $+0.008$ (lung) and $+0.027$ (brain, inside tissue) when it is estimated from fifty chains rather than twenty. The probe's map is unchanged by construction, so the whole movement in Table 21 is the reference improving.

		$J{=}20$		$J{=}50$	
		T-PT	MC	T-PT	MC
Lung CT	ρ, tissue	0.670	0.694	0.671	0.702
$N{=}130$, $G{=}47$	$\Delta\rho$	-0.022	$[-0.037, -0.007]$	-0.025	$[-0.039, -0.012]$
	$\Delta\rho$, interior	-0.023	$[-0.043, -0.005]$	-0.026	$[-0.044, -0.009]$
	$\Delta\rho$, partial (grad. rm.)	-0.060	$[-0.090, -0.031]$	-0.057	$[-0.087, -0.029]$
	ΔAUROC	-0.027	$[-0.039, -0.017]$	-0.033	$[-0.044, -0.022]$
Brain MRI	ρ, tissue	0.356	0.318	0.356	0.345
$N{=}30$, $G{=}5$	$\Delta\rho$		$+0.038$		$+0.010$
	$\Delta\rho$, interior		$+0.046$		$+0.015$
	$\Delta\rho$, partial (grad. rm.)		$+0.043$		$+0.012$
	ΔAUROC		$+0.020$		$+0.000$
	volumes with T-PT ahead		$24/30$		$12/30$

Table 21: Rescoring against a $J{=}50$ reference. $\Delta\rho$ and ΔAUROC are T-PT minus MC-DDIM, as everywhere else; the lung-CT $J{=}20$ intervals are those of Table 1. The lung-CT rows are the full cohort with participant-level intervals; the brain rows are the first thirty volumes, which come from five participants, so they are reported as a convergence check and carry no intervals — resampling five clusters would not mean anything. The brain $J{=}20$ column is recomputed on those same thirty volumes, so the two columns are paired; it is larger there ($+0.038$) than over the full cohort ($+0.016$, Table 1), which is why only the within-table comparison is read. At fifty chains the ensemble costs 3,050 network evaluations per volume against the probe's 76, and inside brain tissue the two are then close on every endpoint, AUROC included. The last row counts volumes on which T-PT has the higher tissue ρ.

T Sensitivity to the probe timestep

The probe timestep is the only continuous knob in T-PT. The sweep below reruns the full lung-CT masked pass of Table 1 at four values of t^*, self-chain probe, $K = 30$, deployment conditioning, everything else fixed; the $t^* = 60$ column is Table 1 itself.

	$t^*{=}15$	$t^*{=}30$	$t^*{=}60$	$t^*{=}120$	MC-DDIM
ρ, tissue	0.663	0.671	0.670	0.669	0.694
ρ, interior	0.619	0.621	0.612	0.606	0.633
ρ, edges	0.244	0.296	0.352	0.406	0.528
ρ, change region	0.042	0.065	0.100	0.145	0.265
ρ, partial (grad. rm.)	0.317	0.334	0.339	0.346	0.409
AUROC, top-5% error	0.854	0.870	0.884	0.897	0.922
$\Delta\rho$, tissue	-0.027	-0.020	-0.022	-0.026	—
$\Delta\rho$, partial (grad. rm.)	-0.093	-0.072	-0.060	-0.048	—

Table 22: Sensitivity to the probe timestep on lung CT ($N{=}130$, $G{=}47$), self-chain probe, $K = 30$, scored against the common error map; absolutes are per-volume means, the paired contrasts are participant-weighted, and every contrast interval excludes zero. Tissue ρ moves by less than 0.01 between $t^* = 15$ and 120, the gradient-controlled partial by 0.03. The regions where the probe trails MC-DDIM, edges and true change, improve monotonically with t^* and at 120 remain well below it. $t^* = 60$ was fixed before any of these runs and is kept.

U How many probes?

The reverse chain dominates the per-volume cost, so K can be swept from the same chain at the price of the extra probe forwards alone. Table 23 does this over the complete scored set of lung CT; the OASIS value of K was fixed by the same reasoning and not swept.

	PNG (N=130)		
K	Fwd.	$\bar{\rho}(\sigma, e)$	% of K=30
5	66	0.567 [0.530, 0.603]	92.4%
10	71	0.595 [0.557, 0.631]	96.9%
15	76	0.604 [0.566, 0.639]	98.4%
30	91	0.613 [0.575, 0.649]	100%

Table 23: Probe size, self-chain probe, $t^* = 60$, lung CT, participant-weighted with participant-level intervals; percentages relative to $K = 30$. Returns flatten quickly: $K = 10$ recovers 96.9% and $K = 15$ recovers 98.4%. The reverse chain dominates the per-volume cost, so the sweep is obtained from the same chain and covers the whole test set rather than a subset. We report $K = 30$ on PNG and $K = 15$ on OASIS, both fixed before any 63-participant score was read; nothing in the method depends on those choices. The maps are scored against the self-chain (deployment) error map, as in Table 14.