SENSITIVITY DECOMPOSITION FOR PREDICTOR-CONDITIONED DIFFUSION SPEECH ENHANCEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion-based speech enhancement systems often combine a deterministic predictor with a generative refinement stage, but it is not clear which component determines how the system responds when test conditions change. We study this question in StoRM, where the predictor output both initializes the reverse diffusion process and acts as its mean-reversion target, while the original noisy mixture is supplied separately to the generative stage. We show that the sensitivity of the enhanced output to a change in noise level can be decomposed exactly into contributions associated with these two conditioning routes. This decomposition reveals a structural asymmetry such that perturbations entering through the predictor affect the reverse process from its initial state, whereas perturbations entering only through score conditioning act through the subsequent dynamics. Our measurements show that the diffusion stage does not simply copy the predictor output; it substantially reshapes perturbations introduced through this route. Nevertheless, the predictor-associated contribution remains strongly aligned with the overall response of the enhanced signal. Extending the analysis from local sensitivities to finite changes in noise level leads to the same conclusion: near the training condition, the predictor route determines the direction of the output change, while the direct noisy-mixture conditioning contributes a substantially smaller, largely complementary component. Replacing the predictor-based reference with the noisy mixture causes the system to track the noisy signal instead of enhancing it, providing an independent intervention on the same mechanism. Together, these results identify the conditioning route that anchors the reverse process as the dominant source of robustness behavior in the model studied here. The analysis extends naturally to other differentiable generative samplers with multiple conditioning inputs.

1 INTRODUCTION

In hybrid speech enhancement, the same observation may enter a generative model through several architectural routes, as an initial state, as the target of a mean-reverting drift, or as an input to a score network. These routes induce different dependencies on the reverse process and are not interchangeable generally. In StoRM (Lemercier et al., 2023), a deterministic predictor Π first maps a noisy mixture y to an estimate $D = \Pi(y)$. The estimate D initializes the reverse process and sets its mean-reversion target, while both D and y condition the score network. Related predictive–generative designs have subsequently appeared in joint decoders (Shi et al., 2024), parallel predictive and score branches (Zhang et al., 2025), and predictor–GAN hybrids (Serbest et al., 2025). The architectural asymmetry of different routes, resulting from non- interchangeability, leads us to study how the output responds to a test-time perturbation distributed across the different conditioning routes.

This distinction matters even when the architecture is treated as a single end-to-end enhancer. A robustness curve can show that quality deteriorates as the noise level changes, but it cannot say whether the deterioration was inherited from the deterministic front end, created by the generative stage, or produced by an interaction between the two. Those explanations lead to different interventions. If the predictor determines the response, more reverse steps or a larger score network will not address the failure; if the generator determines it, improving the predictor alone may be irrelevant.

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.

We begin our study by perturbing the test-time noise amplitude. Let $y^{(M)} = s + \sqrt{M}\, n$, with $M = 1$ denoting the training condition. As M varies, both the predictor output $D_M = \Pi(y^{(M)})$ and the raw mixture change, and the final enhanced waveform changes with them. Existing diffusion theory primarily controls approximation or sampling error for the model as a whole (Song et al., 2021; Chen et al., 2023); it does not separate the response induced by one conditioning route from the response induced by another. We instead differentiate the generator itself, with the Brownian motion component fixed, and ask how the output sensitivity decomposes across its two inputs. Fixing the Brownian realization is important since it turns the stochastic sampler into a deterministic map of its conditioning variables so changes in the output can be assigned to changes in those variables rather than to a change in the sampled noise realization.

Since sensitivity analysis describes the instantaneous response of the enhanced output to an infinitesimal change in the noise level M, whereas robustness experiments compare outputs across finite changes in M, using the derivative at a single value of M to explain such a change would implicitly assume that the channel sensitivities remain approximately constant along the perturbation path. Empirically, this assumption fails even over moderate changes in noise level. We therefore evaluate the decomposition along the path from M_0 to M_1 and integrate it, yielding an exact decomposition of the finite output change. In this form, the contribution of each conditioning channel is allowed to change with M rather than being extrapolated from a single local derivative.

Our proposed decomposition is exact. For the discrete sampler used at inference we find,

$$\frac{\partial \hat{s}^{(M)}}{\partial M} = K^D(M)\frac{\partial D_M}{\partial M} + K^y(M)\frac{\partial y^{(M)}}{\partial M}. \tag{1}$$

Here, $\hat{s}^{(M)}$ denotes the enhanced waveform obtained at perturbation level M, $y^{(M)}$ is the corresponding noisy mixture, and $D_M = \Pi(y^{(M)})$ is the deterministic predictor output. The linear operators $K^D(M)$ and $K^y(M)$ describe how infinitesimal perturbations entering through the predictor and mixture channels, respectively, are transported to the final sampler output. Both operators are generated by the same linearized reverse dynamics, but they differ in their initial conditions. A perturbation of D_M immediately changes the initial state of the reverse sampler, giving $K_0^D = I$, whereas a perturbation of $y^{(M)}$ affects the sampler only through the score-network conditioning and therefore gives $K_0^y = 0$. We refer to D_M as the *anchored* channel and $y^{(M)}$ as the *modulating* channel.

Contributions. We make three contributions. First, we derive a pathwise two-channel sensitivity decomposition for predictor-conditioned diffusion samplers and show that anchored and modulating channels differ through the initial conditions of the same sensitivity recursion (§3). Second, for the OUVE dynamics used by StoRM, we isolate the score-Jacobian quantity that controls departure of the anchor operator from identity and give a step-count-uniform bound (§4). Third, we measure the channel operators by forward-mode automatic differentiation and show empirically that (i) the anchor is strongly transformed rather than simply passed through, (ii) the integrated anchor contribution nevertheless dominates the finite response, and (iii) changing the anchor relocates the behavior of the overall system(§6).

2 Related work

Score-based speech enhancement evolved from single-channel conditional score models (Richter et al., 2023) to hybrid systems that combine a deterministic predictor with stochastic regeneration (Lemercier et al., 2023) and, more recently, other predictive–generative combinations (Shi et al., 2024; Zhang et al., 2025; Serbest et al., 2025). Our question is orthogonal to whether these hybrids improve enhancement quality. Given a trained hybrid, we ask how a controlled test-time perturbation reaches the output through its distinct conditioning slots.

Mathematically, differentiating a parameterized flow yields a variational equation and a variation-of-constants representation (Coddington, 1955; Hartman, 2002; Teschl, 2012); related sensitivity systems are standard in nonlinear and optimal control (Khalil & Grizzle, 2002; Kaufman, 1964) and underpin gradient methods for continuous-depth and stochastic models (Chen et al., 2018; Li et al., 2020). We use this classical machinery diagnostically rather than for training. The new object is the channel split where multiple conditioning inputs share the same linearized dynamics but need not

share the same initial condition. This yields a pathwise attribution statement, whereas convergence analyses for score models characterize the accuracy of the complete learned process (Song et al., 2021; Chen et al., 2023). The distinction also separates our goal from internal activation attribution. We do not ask which neuron or feature is important inside one network. We ask how a specified perturbation reaches the output through named architectural input slots, with all other randomness held fixed. The resulting quantities have a direct bearing on system design with one operator corresponding to perturbation in the anchor slot and another to perturbation in the modulating slot.

3 THE TWO-CHANNEL DECOMPOSITION

3.1 PERTURBATION MODEL AND CONDITIONING CHANNELS

Let $s \in \mathbb{R}^d$ be clean speech and $n \in \mathbb{R}^d$ additive noise. We vary the test input along

$$y^{(M)} = s + \sqrt{M}\, n, \qquad M = (1+\alpha)^2, \tag{2}$$

with $M = 1$ at the training amplitude and (s, n) held fixed. The predictor produces $D_M := \Pi(y^{(M)})$, so

$$\frac{\partial y^{(M)}}{\partial M} = \frac{n}{2\sqrt{M}}, \qquad \frac{\partial D_M}{\partial M} = \nabla\Pi(y^{(M)})\frac{n}{2\sqrt{M}}. \tag{3}$$

We abbreviate these directions by $w := \partial y^{(M)}/\partial M$ and $u := \partial D_M/\partial M$.

StoRM's reverse process is

$$\mathrm{d}X_t^{(M)} = b_\theta(X_t^{(M)}, t, D_M, y^{(M)})\mathrm{d}t + g(t)\mathrm{d}W_t, \qquad t \in [T, 0], \tag{4}$$

with OUVE drift

$$b_\theta(x, t, D, y) = \gamma(D - x) - c\, g(t)^2 s_\theta(x, t, D, y), \tag{5}$$

and terminal condition

$$X_T^{(M)} = D_M + \sigma_T \varepsilon, \qquad \hat{s}^{(M)} := X_0^{(M)}. \tag{6}$$

The predictor estimate therefore affects both the terminal state and the mean-reversion target, in addition to appearing in the score network; the noisy mixture appears only in score conditioning. We call a channel *anchored* if it sets the terminal state and *modulating* otherwise. The terminology refers to how the variable enters the sampler, not to how strongly it ultimately affects the output: an anchored channel can in principle be strongly transformed, and a modulating channel can in principle dominate through the drift. The architecture determines how each channel enters the sampler, and therefore the initial condition of its sensitivity operator. This structural asymmetry is not specific to StoRM and extends to any conditioned sampler in which different inputs enter through different routes.

3.2 DISCRETE SENSITIVITY RECURSION

We analyze the discrete sampler that actually produces the reported outputs. With N Euler–Maruyama steps, $h = T/N$, and $t_k = T - kh$,

$$X_{k+1}^{(M)} = X_k^{(M)} - h\, b_\theta(X_k^{(M)}, t_k, D_M, y^{(M)}) + g(t_k)\Delta W_k,$$

where $X_0^{(M)} = D_M + \sigma_T \varepsilon$ and the increments ΔW_k are fixed across M. Writing $V_k := \partial X_k^{(M)}/\partial M$ gives $V_{k+1} = A_k V_k + B_k^D u + B_k^y w$ with $V_0 = u$, where

$$A_k := I - h\nabla_x b_\theta, \qquad B_k^D := -h\nabla_D b_\theta, \qquad B_k^y := -h\nabla_y b_\theta. \tag{7}$$

where all Jacobians are evaluated along the reverse trajectory.

Lemma 1 (Two-channel decomposition)**.** *Under Assumption 1, define*

$$K_{k+1}^D = A_k K_k^D + B_k^D, \quad K_0^D = I, \qquad K_{k+1}^y = A_k K_k^y + B_k^y, \quad K_0^y = 0. \tag{8}$$

Then $V_k = K_k^D u + K_k^y w$ for every k, and hence

$$\frac{\partial \hat{s}^{(M),h}}{\partial M} = K^D(M)\frac{\partial D_M}{\partial M} + K^y(M)\frac{\partial y^{(M)}}{\partial M}, \qquad K^D := K_N^D,\ K^y := K_N^y. \tag{9}$$

If $\Phi_h(k,j) := A_{k-1}\cdots A_j$ *and* $\Phi_h(k,k) := I$*, then*

$$K^D = \Phi_h(N,0) + \sum_{j=0}^{N-1} \Phi_h(N,j{+}1)B_j^D, \qquad K^y = \sum_{j=0}^{N-1} \Phi_h(N,j{+}1)B_j^y. \tag{10}$$

Proof. Let $\widetilde{V}_k := K_k^D u + K_k^y w$. Equation 8 gives $\widetilde{V}_{k+1} = A_k\widetilde{V}_k + B_k^D u + B_k^y w$ and $\widetilde{V}_0 = u$, matching the recursion and initial condition of V_k and equality follows by induction. Unrolling Eq. 8 yields Eq. 10. □

The same state-transition dynamics act on both channels, while the anchored channel carries the additional initial-state term $\Phi_h(N,0)$. In Eq. 10, $\Phi_h(N,0)$ is the transported effect of moving the starting state while the sums accumulate the effects injected by the drift at each reverse step. The modulating operator contains only the latter making the architectural asymmetry explicit. The operators are utterance- and M-dependent because their Jacobians are evaluated along the corresponding reverse trajectory. Thus K^D and K^y are not global matrices attached once to a trained checkpoint: each utterance and operating point induces its own local operators. Therefore, all empirical operator statistics in our experiments are distributions over utterances.

3.3 Finite changes

Equation 9 is local in M, whereas robustness curves compare outputs across finite changes in noise amplitude. Under the same regularity assumptions the local identity can be integrated without linearizing the operators.

Corollary 2 (Exact finite-change decomposition)**.** *For any* $M_0 < M_1$ *in the operating range,*

$$\hat{s}^{(M_1)} - \hat{s}^{(M_0)} = \int_{M_0}^{M_1} K^D(M)\frac{\partial D_M}{\partial M}\mathrm{d}M + \int_{M_0}^{M_1} K^y(M)\frac{\partial y^{(M)}}{\partial M}\mathrm{d}M. \tag{11}$$

Proof. Assumption 1 makes $M \mapsto \hat{s}^{(M)}$ continuously differentiable. The claim follows from the fundamental theorem of calculus and Lemma 1. □

Corollary 2 turns the local sensitivity decomposition into a decomposition of the finite output change between two noise levels. Each integral accumulates the contribution transmitted through one conditioning route as M varies from M_0 to M_1. Importantly, the predictor sensitivity and the sampler's response to each conditioning channel are both allowed to vary along this path; no single local derivative is extrapolated across the interval. This form also gives a direct way to quantify channel attribution over a finite perturbation. Evaluating both integrals recovers the total change $\hat{s}^{(M_1)} - \hat{s}^{(M_0)}$, and comparing each integral against that total measures how much of it one conditioning route accounts for. In practice the integrals are evaluated numerically on a grid of operating points. We verify convergence by refining this grid, so that numerical integration error can be separated from the contribution assigned to each channel (Appendix A5).

4 The anchor operator

Substituting the OUVE drift into Eq. 7 gives

$$A_k = (1+h\gamma)I + hcg_k^2\nabla_x s_\theta, \quad B_k^D = -h\gamma I + hcg_k^2\nabla_D s_\theta, \quad B_k^y = hcg_k^2\nabla_y s_\theta, \tag{12}$$

where $g_k = g(t_k)$.

Proposition 3 (Deviation identity)**.** *Let* $\Delta_k := K_k^D - I$*. Then*

$$\Delta_{k+1} = A_k\Delta_k + hcg_k^2\left(\nabla_x s_\theta + \nabla_D s_\theta\right), \qquad \Delta_0 = 0. \tag{13}$$

For comparison, the modulating operator obeys

$$K_{k+1}^y = A_k K_k^y + hcg_k^2\nabla_y s_\theta, \qquad K_0^y = 0. \tag{14}$$

Thus the two quantities are propagated by the same linearized state dynamics, but are driven by different score-network Jacobians.

The proof is given in Appendix A10. The identity has a simple interpretation. The mean-reversion term depends only on the difference between the current state and the anchor. An equal perturbation of both therefore leaves this term unchanged; in the absence of the score network, an anchor perturbation would be propagated exactly and $K_k^D = I$ at every step. Departure from pass-through is instead driven by the score network through $\nabla_x s_\theta + \nabla_D s_\theta$. For comparison, the modulating channel is driven by $\nabla_y s_\theta$. The perturbations of either channel propagate through the same linearized reverse dynamics A_k. In particular, if $\nabla_x s_\theta + \nabla_D s_\theta = 0$ and $\nabla_y s_\theta = 0$ along the trajectory, then $K_k^D = I$ and $K_k^y = 0$ for every k.

Corollary 4 (Exact pass-through)**.** *Under the two Jacobian conditions above,* $\partial \hat{s}^{(M)}/\partial M = \partial D_M/\partial M$ *for every sampler step count.*

Next we see that the departure is controlled uniformly in the discretization resolution.

Proposition 5 (Step-count-uniform deviation bound)**.** *Let* $\bar{S} := \sup_k \|\nabla_x s_\theta + \nabla_D s_\theta\|_2$, $L_s := \sup_k \|\nabla_x s_\theta\|_2$, $\bar{G}^2 := \sup_k g_k^2$, *and* $\Lambda := \gamma + cL_s\bar{G}^2$. *Then*

$$\|\Delta_k\|_2 \le c\bar{S}\bar{G}^2 \frac{e^{\Lambda k h} - 1}{\Lambda}, \qquad \|\Delta_N\|_2 \le c\bar{S}\bar{G}^2 \frac{e^{\Lambda T} - 1}{\Lambda}. \tag{15}$$

Consequently $\sigma_{\min}(K^D) \ge 1 - \|\Delta_N\|_2$*; in particular,* K^D *is non-degenerate whenever the upper bound in Eq. 15 is strictly below one. An analogous bound holds for* K^y *with* $\bar{S}$ *replaced by* $\sup_k \|\nabla_y s_\theta\|_2$.

Appendix A1 proves the bound. The role of Proposition 5 is qualitative rather than to predict the measured operator tightly: it shows that refining the sampler does not introduce an uncontrolled growth purely because N increases.

5 Measuring channel influence

The operators K^D and K^y act on waveform-sized vectors and are therefore too large to form explicitly. Fortunately, our analysis only requires their action on particular perturbation directions. We compute these actions using forward-mode automatic differentiation, which evaluates Jacobian–vector products (JVPs) directly without constructing the full Jacobian (Baydin et al., 2018; Griewank & Walther, 2008).

Let $\mathcal{S}(D, y)$ denote the executed reverse sampler when the anchor is D and the noisy-mixture conditioning is y. The terminal noise and Brownian increments are held fixed, so that the derivative measures only the response to the conditioning input. For any direction v,

$$K^D v = \frac{\mathrm{d}}{\mathrm{d}\epsilon}\mathcal{S}(D + \epsilon v, y)\Big|_{\epsilon=0}, \qquad K^y v = \frac{\mathrm{d}}{\mathrm{d}\epsilon}\mathcal{S}(D, y + \epsilon v)\Big|_{\epsilon=0}. \tag{16}$$

Thus $K^D v$ is the change in the enhanced waveform produced by an infinitesimal perturbation of the anchor in direction v, while $K^y v$ is the corresponding response to perturbing the mixture-conditioning channel. Forward-mode differentiation computes these directional derivatives through the sampler directly, rather than approximating them with finite differences (Baydin et al., 2018).

For the perturbation studied in this paper, the relevant directions are

$$u = \frac{\partial D_M}{\partial M}, \qquad w = \frac{\partial y^{(M)}}{\partial M}.$$

Consequently, $K^D u$ and $K^y w$ are precisely the two contributions to the output sensitivity in Eq. 1. We compare their magnitudes and their cosine similarity to determine how strongly each channel contributes and whether the two contributions act in similar or different directions.

We also characterize how the anchor perturbation itself is transformed by the sampler. For a direction v, we report

$$a(v) = \frac{\langle K^D v, v\rangle}{\|v\|^2}, \qquad G(v) = \frac{\|K^D v\|}{\|v\|}, \qquad \rho(v) = \frac{\langle K^D v, v\rangle}{\|K^D v\|\,\|v\|}.$$

Here $a(v)$ measures how much of the original perturbation direction is preserved, $G(v)$ measures its overall change in magnitude, and $\rho(v)$ measures its directional alignment after transport. Exact pass-through, $K^D v = v$, gives $a = G = \rho = 1$. Together these quantities distinguish faithful transport from amplification or rotation of the anchor perturbation. Appendix A4 gives the measurement protocol employed by our experiments.

6 EXPERIMENTS

6.1 EXPERIMENTAL SETUP

We evaluate StoRM on VoiceBank/DEMAND (Valentini-Botinhao, 2017) using an NCSN++ score network and the OU-VESDE schedule (Song et al., 2021). Unless stated otherwise, inference uses the deployed $N = 50$-step Euler–Maruyama sampler. We measure waveform fidelity using scale-invariant signal-to-distortion ratio (SI-SDR) (Roux et al., 2019) and perceptual quality using wideband PESQ (ITU-T P.862.2) (International Telecommunication Union, 2001). Our main comparisons involve a 32-channel StoRM model, a separately trained 128-channel StoRM model, the 32-channel predictor output alone, and a matched 32-channel re-anchored control in which the noisy mixture occupies the anchor slot.

The finite-change and system-level experiments use a fixed set of 22 test utterances. Pointwise operator measurements use 20 of these utterances because two files contain bins at which the amplitude transform used in the sensitivity calculation has a singular derivative. We therefore exclude those files only from analyses that require the corresponding derivative. We report statistics across utterances rather than only population averages because our goal is to explain how individual enhancement trajectories respond to the perturbation. Complete model and sampler configurations are given in Appendix A11.

Our experiments are organized around three objectives. First, we test whether the anchor influences the output primarily through near-identity transport across the reverse process. Second, we evaluate whether the local sensitivity decomposition remains informative for the finite changes in noise level used in robustness evaluation. Third, we intervene on the signal occupying the anchor slot and measure the resulting change in enhancement behavior. Together, these experiments distinguish structural influence of the anchor channel from a simpler correspondence between the predictor and the enhanced output.

6.2 THE ANCHOR IS TRANSFORMED RATHER THAN SIMPLY PASSED THROUGH

We first test if a perturbation introduced through the predictor survives because the reverse sampler behaves approximately as an identity map on that perturbation. If this were the case, we would expect $K^D u \approx u$, where $u = \partial D_M / \partial M$. In the measurements of Section 5, exact pass-through corresponds to unit projection, unit gain, and zero deviation from u.

At the training operating point ($\alpha = 0$) on babble, the 32-channel model does preserve a large component of the anchor perturbation along its original direction. Its projection is

$$a = 0.902 \pm 0.064.$$

Taken alone, this number could suggest near pass-through. The remaining geometry shows otherwise. The transported perturbation has gain 1.311 ± 0.156 and acquires an orthogonal component with norm $0.951\|u\|$. As a result,

$$\frac{\|K^D u - u\|}{\|u\|} = 0.956. \tag{17}$$

Thus, although much of the original perturbation direction survives, the output perturbation is not close to the input perturbation itself. The diffusion stage substantially reshapes the anchor perturbation rather than merely carrying it unchanged to the output.

This observation is important for interpretation. If K^D were close to the identity, a strong correspondence between predictor and enhanced-output robustness could be explained by trivial pass-through of the predictor perturbation. Equation 17 rules out that explanation for the measured direction. The relative decomposition residual of the two-channel factorization in Eq. 9 is 6.3×10^{-5}. We compute

Table 1: Pointwise channel geometry at $\alpha = 0$ on babble. "dev" denotes $\|K^D u - u\|/\|u\|$; "mod. share" is $\|K^y w\|/(\|K^D u\| + \|K^y w\|)$; and "anchor align." is $\cos(K^D u, \partial\hat{s}/\partial M)$.

model	projection a	gain	dev	mod. share	anchor align.
StoRM-32	0.902 ± 0.064	1.311 ± 0.156	0.956	0.156 ± 0.031	0.983 ± 0.008
StoRM-128	0.816 ± 0.096	1.313 ± 0.292	1.045	0.103 ± 0.020	0.993 ± 0.003

this residual using independently evaluated total and channel-wise JVPs, so it serves as a numerical check of the decomposition rather than an algebraic consequence of reusing the same computation.

We next ask how the two conditioning routes combine to produce the total output sensitivity. For the 32-channel model, the modulating contribution $K^y w$ accounts for a relative share of only 0.156 ± 0.031. Moreover, it is nearly orthogonal to the anchor contribution,

$$\cos(K^D u, K^y w) = 0.048 \pm 0.054.$$

Consequently, adding the modulating term changes the total sensitivity mainly in a direction perpendicular to the anchor contribution, and the complete output sensitivity remains strongly aligned with $K^D u$:

$$\cos\left(K^D u, \frac{\partial \hat{s}}{\partial M}\right) = 0.983 \pm 0.008.$$

The anchor therefore determines the dominant direction of the local response, even though the diffusion stage has substantially rotated and amplified the anchor perturbation itself.

To check that this behavior is not a consequence of a score network that is too small to affect the anchor, we repeat the measurement with a separately trained 128-channel StoRM model. Increasing the width does not make anchor transport more identity-like. The relative deviation increases from 0.956 to 1.045, while the modulating share decreases from 0.156 to 0.103 and anchor alignment increases from 0.983 to 0.993 (Table 1). Thus the model with the larger score network transforms the anchor perturbation at least as strongly, while the direction of the total sensitivity is even more tightly associated with the anchor contribution.

We do not interpret two widths as establishing a monotonic relationship between capacity and channel influence. Rather, the wider model provides a targeted falsification of the pass-through explanation: stronger deviation from the identity does not cause sensitivity to migrate toward the modulating channel. Full operator statistics, sampler-step checks, and measurements on additional noise types are provided in A6, and A13.

6.3 Finite-change attribution

A pointwise derivative cannot be extrapolated across the robustness sweep. Using the exact sensitivity at the midpoint of $M \in [0.64, 1.69]$ to predict the finite change gives relative error 1.095 ± 0.433 across the evaluated utterances; the predicted change has norm 1.578 ± 0.401 times the truth and reaches 2.79 on the worst utterance. A relative error above one means that, as a vector prediction of the finite waveform change, the tangent is farther from the truth than predicting no change at all. The failure is therefore not a small higher-order correction. It is the reason we do not interpret the pointwise channel shares as if they were the shares of an entire robustness curve. The derivative itself is computed by automatic differentiation; the error here is solely the use of one local tangent over a finite interval.

We therefore evaluate Corollary 2 directly. Refining the trapezoid grid from 21 to 41 nodes reduces the full decomposition's relative error from 0.040 ± 0.025 to 0.012 ± 0.009, close to the second-order refinement expected for a smooth integrand. The largest 41-node error is 0.029. The reduction by a factor of 3.2 is close to the factor of four expected when halving the grid spacing for the trapezoid rule on a smooth integrand. This refinement behavior is important: without it, a small full-prediction residual could be an accidental cancellation between an incorrect channel split and quadrature error. Instead, the residual decreases with the quadrature resolution while the channel attribution statistic below remains stable. Appendix A5 reports the refinement study.

Both integrals evaluated numerically then provide the finite-range attribution test. Let $T = \hat{s}^{(M_1)} - \hat{s}^{(M_0)}$ be the finite change and let A and B be the anchor and modulating integrals of

Table 2: Re-anchoring moves the system onto the noisy anchor. The first column is the mean SI-SDR difference between the re-anchored output and the measured noisy mixture over the sweep; the remaining columns give enhancement above the noisy mixture at $\alpha = 0$.

noise	re-anchor − noisy	predictor	full system	re-anchor
restaurant	$+0.26$	$+8.94$	$+7.64$	$+0.28$
babble	$+0.03$	$+7.73$	$+6.99$	$+0.05$
car	$+0.01$	$+19.61$	$+17.60$	$+0.01$

Eq. 11 (Appendix A5). On the same window,

$$\frac{\|A\|}{\|T\|} = 0.966 \pm 0.015, \qquad \frac{\|B\|}{\|T\|} = 0.223 \pm 0.030, \qquad \cos(A, T) = 0.974 \pm 0.007. \tag{18}$$

The anchor therefore determines the direction of the finite perturbation response, while the modulating channel contributes a non-negligible, largely transverse remainder. The modulating norm is stable under node refinement (0.2226 ± 0.0295 at 21 nodes and 0.2225 ± 0.0296 at 41), so the attribution number is not an artifact of the grid. The geometry also clarifies what "dominant" means here. The modulating pathway is not negligible, since its contribution over this window is about one fifth of the total change in norm. Rather, the anchor integral is almost the same size as the total and is closely aligned with it, while the remaining component is largely transverse. Over the wider unsaturated range the modulating norm grows and the anchor cosine falls to 0.937, but the anchor remains the larger directional contribution (Appendix A7).

6.4 Re-anchoring and system-level inheritance

The channel decomposition suggests a direct intervention: change which signal occupies the anchor slot. Our matched re-anchored control uses the same 32-channel score architecture and training protocol but anchors the reverse process on the noisy mixture instead of the predictor estimate. Its output tracks that new anchor. Across the sweep, its SI-SDR remains within 0.31 dB of the measured noisy mixture. Table 2 shows the mean difference over the sweep and the enhancement at the training point. The intervention nearly removes the large improvement supplied by the predictor: at $\alpha = 0$ the predictor is $+7.7$ to $+19.6$ dB above the noisy mixture and the full system remains $+7.0$ to $+17.6$ dB above it, whereas the re-anchored control is within 0.28 dB. PESQ gives the same qualitative conclusion (Appendix A2).

The re-anchored system does not carry the per-utterance structure that the predictor and the full system share. Its SI-SDR dispersion across utterances is an order of magnitude smaller and its correlation with the full system at fixed α is near zero. Since the sweep fixes the input SNR, the mixtures at a given α have nearly the same SI-SDR for every utterance, so these two statistics restate the tracking result in per-utterance form. Changing the occupant of the anchor slot is a stronger test than observing that the predictor and full-system curves resemble each other, and we report it as the behavior of one matched control at 32 channels (Appendix A9).

The predictor's response is also visible at the system level. At fixed α, the per-file predictor–system SI-SDR correlation is high on restaurant and babble and weaker on car; pooling per-file points across the sweep gives affine slopes 0.943, 0.972, and 0.889 with $R^2 = 0.9974$, 0.9962, and 0.9811, respectively. Thus the diffusion stage largely preserves the predictor's ordering of which utterances are more or less sensitive to the perturbation. The result is not only driven by the common α trend: at fixed α, median per-file predictor–system correlations are 0.956 on restaurant and 0.971 on babble; car is weaker at 0.759, which is also the condition with the largest modulating share in the cross-noise operator analysis. We treat that agreement as descriptive, not as a law inferred from three noises.

The inheritance persists under PESQ and on the unseen `office` noise, where the pooled slope is 1.00 with $R^2 = 0.98$ (Appendices A2 and A3). These system-level observations are consistent with, but are not substitutes for, the channel decomposition: pooled metric correlations could arise for reasons unrelated to the pathway split. The attribution claim itself comes from Eq. 11 and the deletion experiment above; the system-level curves show that the same qualitative inheritance is visible in standard enhancement metrics.

7 DISCUSSION AND CONCLUSION

Our results separate two properties that are easy to conflate. One is how faithfully the sampler transports a perturbation that enters through a channel. The other is how strongly that channel determines the response of the output. In the systems we study, the anchored channel is transported poorly but determines the response. The sampler moves the anchor perturbation far from pass-through, with $\|K^D u - u\| \approx \|u\|$, and the wider score network moves it further. The anchor contribution is nonetheless aligned with the total output sensitivity, with cosine 0.983 at 32 channels and 0.993 at 128 channels. Over a finite change in noise amplitude, its cosine with the total change is 0.974, and the modulating contribution has about a fifth of the norm of that change. System-level metrics cannot make this distinction on their own. The full system's SI-SDR is a near-affine function of the predictor's, which is consistent with a generative stage that passes the predictor estimate through unchanged and equally with one that transforms it substantially. The operator measurements show that the second holds.

For predictor-conditioned enhancers, this suggests that the signal placed in the anchor slot matters as much as the generative stage when designing for robustness to noise mismatch. In the systems we study, widening the score network changes how the anchor perturbation is transformed without shifting the direction of the response toward the modulating channel. Changes confined to the generative stage may therefore leave the dominant pathway unchanged.

Limitations. The decomposition is an exact identity for any differentiable sampler with distinct conditioning inputs, but the attribution result is empirical and specific to the settings we measured. We study a single architecture on one dataset, perturb only the noise amplitude, and report detailed operator statistics at a 0 dB reference operating point. At 32 channels the generative stage adds little to the predictor in SI-SDR or PESQ, so part of the anchored channel's dominance may reflect a weak generative stage, and two widths do not establish how the attribution changes with capacity. The re-anchored control is trained at a single width. Score-based enhancers anchored on the noisy mixture do enhance at the network sizes used in prior work (Richter et al., 2023), so its failure to enhance characterizes the 32-channel control rather than noisy anchoring in general. The attribution also depends on how the contributions are separated. Integrating along the physical perturbation path does not isolate interactions between the channels. Moreover, the score network is trained only on predictor estimates paired with the mixtures they were computed from, so the separate operators partly reflect how the network extrapolates beyond its training distribution. Finally, the measured operators belong to the deployed 50-step sampler and are not converged in the number of steps. In another architecture the anchored channel could be weak and the modulating channel strong, and nothing in our analysis rules this out. Appendix A12 discusses each point in detail.

Conclusion. Differentiating the executed reverse sampler of a predictor-conditioned diffusion enhancer separates its response to a test-time perturbation into one sensitivity operator per conditioning input. The operators obey the same linearized dynamics and differ only in their initial conditions, and under OUVE dynamics the anchored operator departs from pass-through only through an explicit score-Jacobian term. Integrating the decomposition along the perturbation path attributes finite output changes exactly. Applied to StoRM under noise-amplitude mismatch, the analysis shows that the generative stage substantially transforms the predictor's contribution while the predictor channel still determines the direction of the output change. Natural next steps are to measure the channel operators in systems with stronger generative stages and under other forms of mismatch, such as a change of noise type or reverberation, and to separate interaction effects with path-independent attribution methods. Because the method requires only forward-mode differentiation through the sampler, it applies to other conditional generative samplers with multiple inputs.

AI USE STATEMENT

Generative AI tools were used to implement experimental code from author-specified designs and to draft and edit portions of the manuscript. The research questions, experimental design, choice of analyses, and interpretation of the reported results were developed by the authors. The authors reviewed the AI-assisted work and take responsibility for the final manuscript, code, claims, and reported results.

Reproducibility statement

The mathematical assumptions and complete proofs are provided in the appendix. The appendix also specifies the perturbation construction, operator-measurement protocol, implementation controls for forward-mode differentiation, experimental setup, and quadrature and step-count checks. These details are intended to make the sensitivity measurements and reported ablations reproducible from the corresponding model checkpoints and test data.

References

Atilim Gunes Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research*, 18(153):1–43, 2018. URL http://jmlr.org/papers/v18/17-468.html.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 6572–6583, Red Hook, NY, USA, 2018. Curran Associates Inc.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=zyLVMgsZ0U_.

Earl A Coddington. Theory of ordinary differential equations. *McGraw-Hill*, 1955.

Andreas Griewank and Andrea Walther. *Evaluating derivatives: principles and techniques of algorithmic differentiation*. SIAM, 2008.

Philip Hartman. *Ordinary differential equations*. SIAM, 2002.

International Telecommunication Union. ITU-T Recommendation P.862 (2001) — Perceptual Evaluation of Speech Quality (PESQ), 2001. URL https://www.itu.int/rec/T-REC-P.862-200511-W!Amd2/en. Includes Annex A: Wideband PESQ (P.862.2).

H. Kaufman. The mathematical theory of optimal processes, by l. s. pontryagin, v. g. boltyanskii, r. v. gamkrelidze, and e. f. mishchenko. authorized translation from the russian. translator: K. n. trirogoff, editor: L.. w. neustadt. interscience publishers (division of john wiley and sons, inc. , new york) 1962. viii + 360 pages. *Canadian Mathematical Bulletin*, 7(3):500–500, 1964. doi: 10.1017/S0008439500032112.

Hassan K Khalil and Jessy W Grizzle. *Nonlinear systems*, volume 3. Prentice hall Upper Saddle River, NJ, 2002.

Jean-Marie Lemercier, Julius Richter, Simon Welker, and Timo Gerkmann. Storm: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2724–2737, 2023. doi: 10.1109/TASLP.2023.3294692.

Xuechen Li, Ting-Kam Leonard Wong, Ricky T. Q. Chen, and David Duvenaud. Scalable gradients for stochastic differential equations. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3870–3882. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/li20i.html.

Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, and Timo Gerkmann. Speech enhancement and dereverberation with diffusion-based generative models. 31:2351–2364, June 2023. ISSN 2329-9290. doi: 10.1109/TASLP.2023.3285241. URL https://doi.org/10.1109/TASLP.2023.3285241.

Horn Roger and R Johnson Charles. Topics in matrix analysis, 1994.

Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. Sdr – half-baked or well done? In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630, 2019. doi: 10.1109/ICASSP.2019.8683855.

Sanberk Serbest, Tijana Stojkovic, Milos Cernak, and Andrew Harper. DeepFilterGAN: A Full-band Real-time Speech Enhancement System with GAN-based Stochastic Regeneration. In *Interspeech 2025*, pp. 878–882, 2025. doi: 10.21437/Interspeech.2025-1889.

Hao Shi, Kazuki Shimada, Masato Hirano, Takashi Shibuya, Yuichiro Koyama, Zhi Zhong, Shusuke Takahashi, Tatsuya Kawahara, and Yuki Mitsufuji. Diffusion-based speech enhancement with joint generative and predictive decoders. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12951–12955, 2024. doi: 10.1109/ICASSP48485.2024.10448429.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling Through Stochastic Differential Equations. In *International Conference on Learning Representations*, 2021. URL `https://mlanthology.org/iclr/2021/song2021iclr-scorebased/`.

Gerald Teschl. *Ordinary differential equations and dynamical systems*, volume 140. American Mathematical Soc., 2012.

Cassia Valentini-Botinhao. Noisy reverberant speech database for training speech enhancement algorithms and tts models. 2017.

Jie Zhang, Haoyin Yan, and Xiaofei Li. A composite predictive-generative approach to monaural universal speech enhancement. *IEEE Transactions on Audio, Speech and Language Processing*, 33:2312–2325, 2025. doi: 10.1109/TASLPRO.2025.3577387.

A1 Proof of Proposition 5

Recall the deviation recursion Eq. 13 established in Proposition 3, with $\Delta_0 = 0$.

Proof. Unrolling Eq. 13 from $\Delta_0 = 0$,

$$\Delta_k = \sum_{j=0}^{k-1} \Phi_h(k, j{+}1)\; hc\, g_j^2 \left(\nabla_x s_\theta + \nabla_D s_\theta\right)\Big|_j. \tag{19}$$

For the state-transition product, Eq. 12 gives $\|A_k\|_2 \le 1 + h\gamma + hc\bar{G}^2 L_s = 1 + h\Lambda$, hence $\|\Phi_h(k,j)\|_2 \le (1+h\Lambda)^{k-j} \le e^{\Lambda(k-j)h}$. Taking norms in Eq. 19,

$$\|\Delta_k\|_2 \le c\,\bar{S}\,\bar{G}^2 \sum_{j=0}^{k-1} h\, e^{\Lambda(k-j-1)h} \le c\,\bar{S}\,\bar{G}^2 \int_0^{kh} e^{\Lambda\tau}\mathrm{d}\tau = c\,\bar{S}\,\bar{G}^2\,\frac{e^{\Lambda kh}-1}{\Lambda}, \tag{20}$$

where the middle inequality is the right-endpoint Riemann comparison for the increasing integrand, which is where the ceiling $\bar{G}^2$ is used. Setting $k = N$ and $Nh = T$ gives Eq. 15, and the right-hand side depends on N only through $Nh = T$, hence is uniform in the step count. Finally $\sigma_{\min}$ is 1-Lipschitz in the operator norm (Roger & Charles, 1994), so $\sigma_{\min}(K^D) = \sigma_{\min}(I + \Delta_N) \ge 1 - \|\Delta_N\|_2$. □

A2 Inheritance under PESQ

The system-level analysis of §6.4 uses SI-SDR, which measures waveform fidelity. To test whether the inheritance of predictor behavior by the full system is specific to that metric, we repeat the analysis with wideband PESQ (ITU-T P.862.2), which targets perceived quality. All other settings are unchanged. Figure 1 shows the resulting curves.

Correlation across utterances. At each α we compute the Pearson correlation between predictor and full-system PESQ across utterances, which removes the common trend in α. The median over operating points is 0.622 on restaurant, 0.909 on babble and 0.752 on car. The correlation varies widely with α on restaurant (0.131 to 0.947) and babble (-0.018 to 0.978), and little on car (0.690 to 0.793). On restaurant and babble the mean PESQ of every configuration approaches the lower end of the scale as α increases (Figure 1). Near that limit PESQ differs little between utterances, so a correlation computed across utterances is less informative there. Pooling all utterances and operating points, an affine fit of full-system PESQ on predictor PESQ has slope 0.944, 0.943 and 0.909, with R^2 of 0.954, 0.970 and 0.918. These pooled fits include the common trend in α, so they overstate agreement between individual utterances relative to the fixed-α correlations. Under both measures, the full system follows the predictor under PESQ as it does under SI-SDR, although less closely, most noticeably on restaurant.

Dispersion across utterances. Under SI-SDR, the spread across utterances is an order of magnitude smaller for the re-anchored variant than for the predictor and the full system (Appendix A9). This contrast does not appear under PESQ. On restaurant, the standard deviation across utterances ranges over α from 0.015 to 0.171 for the predictor and from 0.025 to 0.111 for the full system, and the other noise types show similar values for all configurations. The compression of PESQ near its lower limit contributes to this, and we therefore treat the dispersion contrast as a property of SI-SDR.

Score levels. Averaged over all operating points, full-system PESQ exceeds predictor PESQ by 0.050 on restaurant and 0.023 on babble, and falls below it by 0.081 on car. Together with the 1 to 2 dB SI-SDR deficit of the full system relative to the predictor, this shows that at 32 channels the generative stage adds little measurable quality to the predictor estimate on either metric. We discuss the implications for the attribution in Appendix A12. StoRM-128 is included in Figure 1 for reference.

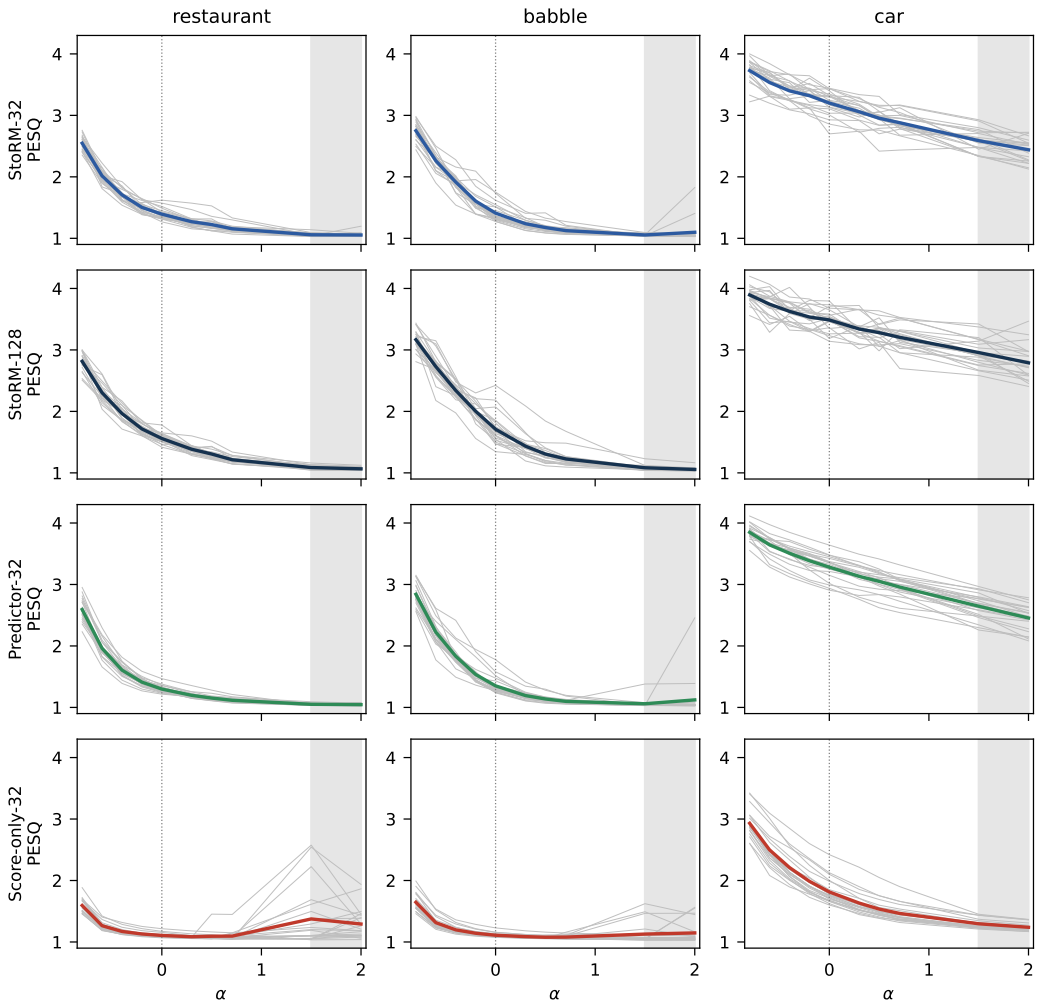


Figure 1: Wideband PESQ against α for each configuration (rows) and noise type (columns). Grey lines show individual utterances and bold lines their mean. The dotted line marks the reference operating point $\alpha = 0$. The shaded region marks operating points at which the stored mixtures clip (Appendix A11). Score-only-32 denotes the re-anchored variant.

Re-anchored variant. The re-anchored variant scores below the predictor by 0.222, 0.347 and 1.330 PESQ on restaurant, babble and car, averaged over operating points. Its correlation with the full system across utterances has medians 0.151, 0.340 and -0.218, compared with 0.622, 0.909 and 0.752 for the predictor. As under SI-SDR, its output shares little of the utterance-dependent structure common to the predictor and the full system.

A3 Out-of-distribution noise

The system-level analysis of §6.4 uses noise types seen during training. To test whether the full system still follows the predictor under a noise type it has not seen, we repeat the analysis with office noise from the VoiceBank/DEMAND test set, which does not occur in the training data. Office noise is stationary. The mixtures use the same 22 utterances, 0 dB reference SNR and StoRM-32 and Predictor-32 checkpoints as the seen-noise experiments (Appendix A11), at 12 operating points in $\alpha \in [-0.8, 2]$. As with the seen noises, the stored mixtures clip at the largest operating points (Appendix A11). The analysis in this appendix compares two enhanced outputs of the same stored

mixture and does not use Eq. 2, so clipping changes the input both systems receive without affecting the comparison, and we include all twelve operating points.

At each α we compute the Pearson correlation between predictor and full-system SI-SDR across utterances, which removes the common trend in α. The correlation ranges from 0.63 to 0.93 over operating points, with median 0.83. This lies between the medians for the seen noises, which are 0.956 on restaurant, 0.971 on babble and 0.759 on car. Pooling all utterances and operating points, an affine fit of full-system SI-SDR on predictor SI-SDR has slope 1.00 and $R^2 = 0.98$, with an intercept of about -2 dB. The full system therefore tracks the predictor on this noise with a nearly constant SI-SDR deficit, as it does on the seen noises, where the pooled slopes are 0.943, 0.972 and 0.889.

This experiment is limited in two respects. It uses a single unseen noise type, and a stationary one, so it does not show that the result extends to nonstationary or otherwise harder mismatched noise. It also does not include the re-anchored variant, so the contrast between the predictor and the re-anchored control in Appendix A9 is established only for the seen noises.

A4 Measurement protocol

Computing the operators. The operators K^D and K^y map perturbations of the conditioning inputs to perturbations of the enhanced output. Each acts on vectors with as many entries as the network's input representation, so forming either one as a matrix is infeasible. Our analysis requires only their action on a small number of specified directions v,

$$K^D v = \frac{\mathrm{d}}{\mathrm{d}\epsilon}\bigg|_{\epsilon=0} \mathcal{S}(D + \epsilon v,\, y), \qquad K^y v = \frac{\mathrm{d}}{\mathrm{d}\epsilon}\bigg|_{\epsilon=0} \mathcal{S}(D,\, y + \epsilon v). \tag{21}$$

Here $\mathcal{S}(D, y)$ denotes the reverse sampler run with terminal state $D + \sigma_T \varepsilon$, mean-reversion target D and score conditioning (D, y), with the predictor stage bypassed. The terminal noise ε and the Brownian increments ΔW_k are sampled once per utterance and reused in every evaluation, so $\mathcal{S}$ is a deterministic function of its inputs. Perturbing the D slot moves the terminal state, the drift target and the score input together, which matches the definition of K^D.

We compute these quantities as Jacobian-vector products (JVPs) using forward-mode automatic differentiation (Baydin et al., 2018), implemented with `torch.func.jvp`. Backpropagation computes vector-Jacobian products $v^\top J$ for a function with Jacobian J. Forward mode computes the column product Jv, which is what $K^D v$ and $K^y v$ require. It propagates the derivative along v through every network evaluation and reverse step during a single run of the sampler, so each run returns the exact derivative of the executed sampler up to floating-point precision, at a cost comparable to a few ordinary sampler runs. Unlike a finite-difference estimate, the result does not depend on a step size.

Resampling layers. The NCSN++ score network changes the resolution of its feature maps with finite impulse response (FIR) filters. The reference implementation provides this operation in two equivalent forms, a custom GPU kernel and a composition of standard tensor operations. The custom kernel defines only the derivative rule used by backpropagation, so forward-mode differentiation cannot pass through it. We therefore use the standard-operation form for all derivative computations. To confirm that this substitution does not alter the model, we verified that the enhanced outputs produced by the two forms are bitwise identical.

Perturbation directions and decomposition residual. The perturbation directions are $u = \partial D_M / \partial M$ and $w = \partial y^{(M)} / \partial M$, represented in the network's input domain after the STFT and amplitude compression. Both are obtained by forward mode, w through the input transform alone and u through the input transform and the predictor. To check Eq. 9 numerically, we also compute the total sensitivity $\partial \hat{s} / \partial M$ by a separate forward-mode run through the complete system, predictor included, starting from the waveform derivative $n/(2\sqrt{M})$. This run perturbs both conditioning inputs at once and never evaluates K^D or K^y individually, so its agreement with $K^D u + K^y w$ is an independent check of the decomposition rather than a restatement of it. We report the relative

decomposition residual

$$\frac{\|\partial \hat{s}/\partial M - K^D u - K^y w\|}{\|\partial \hat{s}/\partial M\|}. \tag{22}$$

Since Eq. 9 holds exactly, this residual measures only floating-point error and would expose implementation errors in either computation.

Reported quantities. To describe how the sampler transforms an anchor perturbation v, we report

$$a(v) := \frac{\langle K^D v,\, v\rangle}{\|v\|^2}, \qquad \text{gain} := \frac{\|K^D v\|}{\|v\|}, \qquad \rho := \frac{\langle K^D v, v\rangle}{\|K^D v\|\,\|v\|}. \tag{23}$$

The projection a is the coefficient of the original direction retained in $K^D v$, the gain is the change in magnitude, and ρ is the cosine of the angle between $K^D v$ and v. Exact pass-through, $K^D v = v$ (Corollary 4), gives $a = \text{gain} = \rho = 1$. Decomposing $K^D v$ into a component along v and a component orthogonal to it, the orthogonal component has norm $\sqrt{\text{gain}^2 - a^2}\,\|v\|$ and the distance from pass-through is $\|K^D v - v\| = \sqrt{\text{gain}^2 - 2a + 1}\,\|v\|$. Both are therefore determined by the projection and the gain.

Input normalization. The reference inference code divides each input waveform by its peak absolute value before the STFT. Because the peak of $s + \sqrt{M}\,n$ changes with M, this step would rescale the input by a different factor at each noise level, and the network input would no longer follow Eq. 2. We therefore compute the normalization constant once, at $\alpha = 0$, and apply the same constant at every M. The network input is then a fixed multiple of $s + \sqrt{M}\,n$.

A5 Numerical integration of the finite-change decomposition

The integrals in Eq. 11 have no closed form, so we evaluate them numerically. For each utterance we compute the integrands $K^D u$ and $K^y w$ at equally spaced values of M on $[0.64, 1.69]$, which corresponds to $\alpha \in [-0.2, 0.3]$, and approximate each integral with the trapezoidal rule. Appendix A7 uses nodes equally spaced in α instead, which samples the same integral differently. This rule replaces the integrand between neighboring grid points by a straight line and sums the resulting areas. We denote the approximations of the anchor and modulating integrals by $\tilde{A}$ and $\tilde{B}$, and compare their sum against the true finite change $T = \hat{s}^{(M_1)} - \hat{s}^{(M_0)}$, obtained by running the sampler at the two endpoints. We report the relative error $\|\tilde{A} + \tilde{B} - T\|/\|T\|$. For comparison, predicting T from a single tangent at the midpoint of the same window gives a relative error of 1.095 ± 0.433 (§6.3).

This error has two possible sources. The first is the trapezoidal approximation. For a smooth integrand its error is proportional to the square of the grid spacing, so halving the spacing should reduce it by roughly a factor of four. The second is any error in the decomposition itself or in its implementation. That error does not depend on the grid and would persist under refinement. Refining the grid therefore distinguishes the two.

Moving from 21 to 41 grid points halves the spacing. Over 21 of the 22 utterances, this reduces the mean relative error from 0.040 ± 0.025 to 0.012 ± 0.009, and the largest error at 41 points is 0.029. The reduction is close to the factor of four expected for the trapezoidal rule, so the remaining error is due to numerical integration rather than to the decomposition. This agrees with the pointwise decomposition residual of 6.3×10^{-5} reported in §6.2. Refinement also leaves the attribution statistic unchanged, with $\|\tilde{B}\|/\|T\| = 0.2226 \pm 0.0295$ at 21 points and 0.2225 ± 0.0296 at 41, so the channel attribution in §6.3 is not an artifact of the grid.

The remaining utterance, has relative error 0.375 at 21 points and 0.148 at 41. Its transport gain $\|K^D u\|/\|u\|$ varies more along the path than any other utterance, with a standard deviation across nodes of 1.49 compared with 0.21 for the others. A more rapidly varying integrand requires a finer grid for the same accuracy, and the error still decreases substantially under refinement, so we attribute it to numerical integration. We exclude this utterance from the error statistics above and retain it in the attribution results of §6.3, where its values lie within the range of the other utterances.

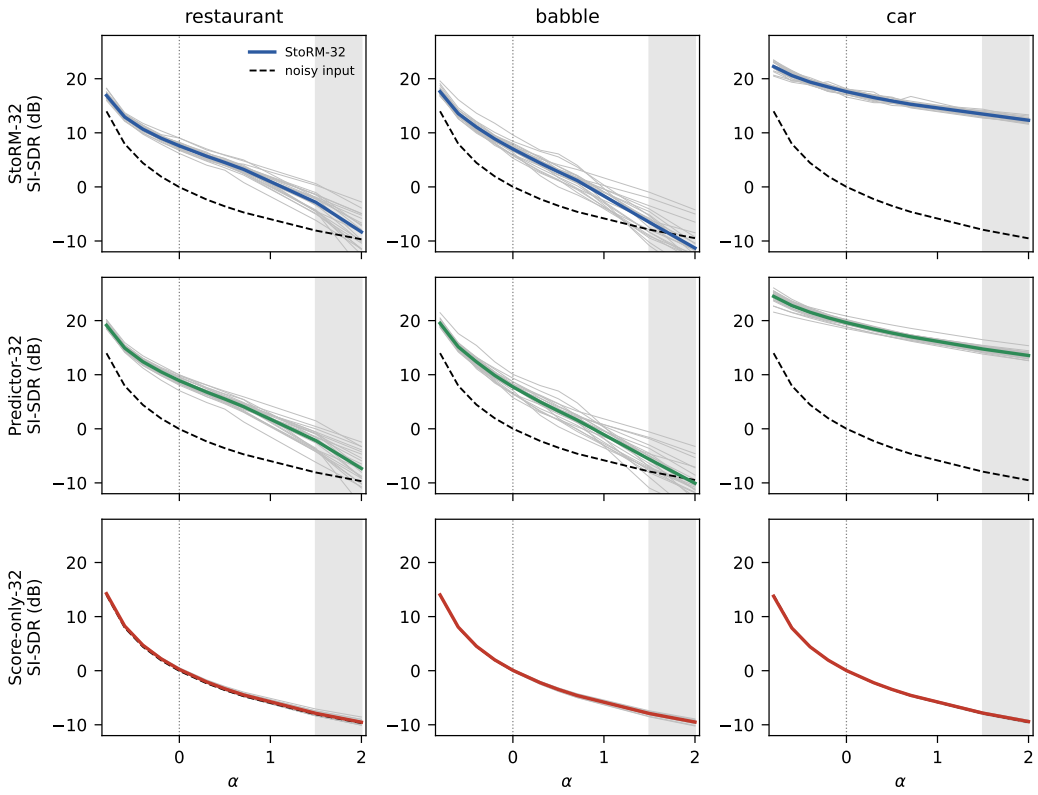


Figure 2: *The re-anchored variant sits on the noisy input line and the predictor supplies the enhancement.* SI-SDR against α for the three 32-channel variants, per noise, with grey per-utterance traces, bold means, and the measured noisy-mixture curve dashed. The dotted vertical line marks the training condition $\alpha = 0$. The re-anchored variant overlaps the noisy curve to within 0.31 dB throughout; the predictor and the full system lie $+7$ to $+20$ dB above it. The shaded region marks the high-amplitude regime affected by saturation; claims that use Eq. 2 are restricted to the verified unsaturated range in Appendix A11.

A6 Dependence on the number of sampler steps

The operators in Eq. 9 are Jacobians of the discrete sampler with N reverse steps, not of the continuous-time reverse SDE. Each step of the recursion in Eq. 13 adds a forcing term proportional to the step size $h = T/N$. The discrete operator therefore only approximates its continuous-time counterpart, with an error that shrinks as N grows, and its measured geometry depends on N. We report it for the $N = 50$ sampler that produces every enhanced output in the paper.

Table 3 shows that this dependence is material. On the 20 utterances at $\alpha = 0$ on babble, reducing the sampler from $N = 50$ to $N = 5$ steps lowers the projection from 0.902 to 0.821 and the orthogonal component from $0.951\|u\|$ to $0.823\|u\|$. A coarse sampler is therefore not a reliable proxy for the deployed one.

To assess how close $N = 50$ is to convergence, we repeated the measurement at $N = 100$ on four utterances. At $N = 100$, the Brownian increments are obtained by subdividing those used at $N = 50$, and the terminal noise ε is shared, so both runs follow the same realization of the reverse process. These four utterances are a subset, so their values at $N = 50$ differ from the 20-utterance means, and only the comparison across step counts within the subset is meaningful. Doubling the step count increases the deviation from pass-through, $\|K^D u - u\|/\|u\|$, from 0.793 to 0.855 and the orthogonal component from $0.777\|u\|$ to $0.839\|u\|$, while the projection is nearly unchanged at 0.840 and 0.833. The operator is thus not converged at $N = 50$, and we do not extrapolate a continuous-time value from two step counts.

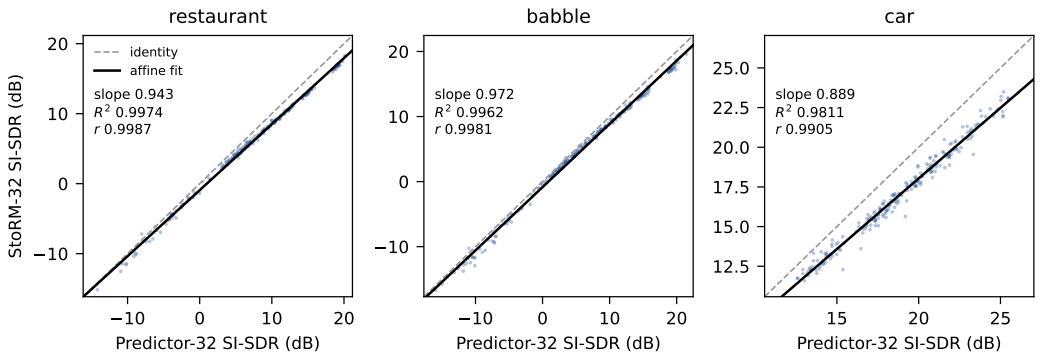


Figure 3: *Per-file inheritance.* StoRM-32 against Predictor-32 SI-SDR, one point per utterance and per α, with the affine fit solid and the identity line dashed. The full system is a near-unit-slope image of the predictor on all three noises, offset below identity by the diffusion's small, near-mean-preserving cost, so it tracks the predictor's per-file ordering rather than reordering it.

Table 3: The anchor operator at $\alpha = 0$ on babble, over twenty utterances at 32 channels unless the row says otherwise, mean plus or minus standard deviation. Exact pass-through (Corollary 4) would give 1 in the first three columns and 0 in the orthogonal column. The measured operator depends on the number of sampler steps, which Appendix A6 examines. The lower block repeats the comparison on four matched utterances so that the step-count trend is read on a common set.

N	projection a	gain	ρ	orthogonal	residual
5	0.821 ± 0.057	1.163 ± 0.056	0.706 ± 0.036	0.823	2.7×10^{-4}
50	0.902 ± 0.064	1.311 ± 0.156	0.693 ± 0.051	0.951	6.3×10^{-5}
128 channels, same twenty utterances, $N = 50$.					
50	0.816 ± 0.096	1.313 ± 0.292	0.638 ± 0.084	1.028	2.0×10^{-4}
Four utterances, matched, at two step counts.					
50	0.840	1.144	0.735	0.777	5.5×10^{-5}
100	0.833	1.182	0.706	0.839	4.5×10^{-5}
At $N = 50$, modulating share 0.156 ± 0.031, $\cos(K^D u, K^y w) = 0.048 \pm 0.054$, $\cos(K^D u, \partial \hat{s}/\partial M) = 0.983 \pm 0.008$.					

This does not weaken the conclusions of §6.2 and §6.3, for two reasons. First, those conclusions concern the sampler that produces the reported outputs, and Eq. 9 is exact for that sampler at every N. Second, the observed change favors the conclusion of §6.2. The deviation from pass-through grows with N, so a finer sampler transforms the anchor perturbation more, not less. On the same subset, the modulating share is 0.176 at $N = 50$ and 0.181 at $N = 100$, and the inter-channel cosine is 0.105 and 0.176. Applying Eq. 24 to these means gives an alignment of the total sensitivity with the anchor contribution of 0.979 at both step counts, so the attribution is unchanged. The bound in Proposition 5 does not grow with N, which rules out unbounded growth of the deviation under refinement, although it does not predict the measured values.

The decomposition residual is 5.5×10^{-5} at $N = 50$ and 4.5×10^{-5} at $N = 100$ on the same subset. Doubling the number of steps, and with it the number of operations over which floating-point error could accumulate, does not increase the residual.

A7 Attribution along the path and over a wider range

Variation along the path. The pointwise balance between the channels is not constant along the perturbation path. On the 81-point grid over $\alpha \in [-0.8, 0.7071]$, the ratio $\|K^y w\|/\|K^D u\|$ has a per-utterance mean between 0.182 and 0.292 and takes values between 0.026 and 0.782 across all

nodes and utterances. The inter-channel cosine $\cos(K^D u, K^y w)$ has a per-utterance mean between 0.082 and 0.167 and ranges from -0.44 to $+0.68$. A decomposition evaluated at a single operating point therefore does not determine the attribution over a finite range, which is consistent with the failure of the single-tangent prediction in §6.3. The finite-change decomposition of Eq. 11 accounts for this variation, because it accumulates each channel's contribution from its value at every point along the path.

Wider range. We repeat the finite-change analysis over $\alpha \in [-0.8, 0.7071]$, which corresponds to $M \in [0.04, 2.91]$ and input SNRs from $+14.0$ to -4.6 dB. This is the full range on which Eq. 2 is verified and the mixtures are unclipped (Appendix A11). Table 4 compares it with the narrow window of §6.3, using the notation of Appendix A5. We did not repeat the grid-refinement study on the wider range. Its integration error of 0.033 is larger than the 0.012 reached at 41 points on the narrow window, since the range is about three times wider, but it remains an order of magnitude below the attribution quantities.

Table 4: Finite-change attribution on the narrow window and on the full verified range. $\tilde{A}$ and $\tilde{B}$ are the numerically integrated anchor and modulating contributions and T is the true finite change. The narrow window uses nodes placed uniformly in M and the full range nodes placed uniformly in α, which is the same integral sampled differently. Values are means over the 22 utterances.

range of α	input SNR (dB)	grid points	$\frac{\|\tilde{A}+\tilde{B}-T\|}{\|T\|}$	$\frac{\|\tilde{A}\|}{\|T\|}$	$\frac{\|\tilde{B}\|}{\|T\|}$	$\cos(\tilde{A}, T)$
$[-0.2, 0.3]$	$+1.9$ to -2.3	41	0.012	0.966	0.223	0.974
$[-0.8, 0.7071]$	$+14.0$ to -4.6	81	0.033	0.913	0.349	0.937

Interpretation. As the range widens, the modulating contribution grows from $0.223\|T\|$ to $0.349\|T\|$ and the anchor contribution becomes less closely aligned with the total change, with $\cos(\tilde{A}, T)$ falling from 0.974 to 0.937. The margin of the anchored channel is therefore largest near the reference operating point. The anchored channel nonetheless remains the dominant contribution over the full verified range, at $0.913\|T\|$ against $0.349\|T\|$ for the modulating channel and closely aligned with the total change. The stored data do not allow us to test whether this ordering persists at larger amplitudes, because the mixtures clip there.

A8 Assumptions

The results of Sections 3 and 4 rely on the following condition. Let $\mathcal{M} = [M_{\min}, M_{\max}] \subset (0, \infty)$ denote the operating range.

Assumption 1 (Regularity)**.**

(i) The drift $b_\theta(x, t, D, y)$ and its Jacobians $\nabla_x b_\theta$, $\nabla_D b_\theta$ and $\nabla_y b_\theta$ are continuous in (x, t, D, y).

(ii) The maps $M \mapsto D_M$ and $M \mapsto y^{(M)}$ are continuously differentiable on $\mathcal{M}$.

(iii) There is a compact set that contains the sampler iterates $X_k^{(M)}$ for every $M \in \mathcal{M}$, every step count N and every $k \le N$.

Conditions (i) and (ii) make each sampler step, and therefore the output $\hat{s}^{(M)}$, a continuously differentiable function of M, which is what Lemma 1 requires. It also gives the continuity of the integrands in Corollary 2. Condition (iii) is used only in Proposition 5. Combined with (i), it bounds the Jacobians of b_θ along every trajectory independently of N. Since $g(t)^2$ is bounded away from zero on $[0, T]$, the score Jacobians and hence the constants $\bar{S}$ and L_s are bounded independently of N as well, which makes the bound in Eq. 15 uniform in the step count.

In our setting, the predictor and the score network are compositions of smooth operations, so (i) holds and the predictor is continuously differentiable. The waveform $s + \sqrt{M}\, n$ is smooth in M on $\mathcal{M}$. The only non-smooth operation is the amplitude compression $|X|^a$ with $a < 1$ applied

before both networks, which is not differentiable where $|X| = 0$. Condition (ii) therefore holds at every time-frequency bin except those at which both s and n vanish. Such bins occur only in digital silence and are treated as described in Appendix A11. We have not verified condition (iii) for all step counts, and it enters only the uniformity statement of Proposition 5.

A9 CHANGING THE ANCHOR

The re-anchored variant replaces the predictor estimate with the noisy mixture in the anchor slot. It is a separately trained 32-channel score network with the same architecture and training protocol as the StoRM score network (Appendix A11). Its output tracks the new anchor, with SI-SDR within 0.31 dB of the noisy mixture across the sweep.

Per-utterance statistics. At each α, the standard deviation of SI-SDR across utterances is 0.03 to 0.36 dB for the re-anchored variant, compared with 0.47 to 3.05 dB for the predictor and 0.41 to 3.03 dB for the full system. Across utterances at fixed α, the correlation between re-anchored and full-system SI-SDR has medians $+0.10$, -0.12 and -0.16 on restaurant, babble and car, with values between -0.30 and $+0.32$. The corresponding medians between the predictor and the full system are 0.956, 0.971 and 0.759. These statistics follow directly from the re-anchored output tracking the noisy mixture. The sweep fixes the input SNR, so at each α the mixtures have nearly the same SI-SDR for every utterance, with scatter below 0.12 dB at $\alpha = 0$ (Appendix A11). An output that follows the mixture inherits this near-uniformity and therefore cannot carry the utterance-dependent structure that the predictor passes to the full system.

Scope. The control is trained at a single width. By the deviation identity in Eq. 13, the departure of an anchored channel from pass-through is governed by the trained score Jacobians $\nabla_x s_\theta + \nabla_D s_\theta$, which depend on the capacity and training of the network rather than on the architecture alone (Corollary 4). Increasing the width of the full system moves K^D further from pass-through while strengthening the anchor's control of the output direction (§6.2), but this does not determine whether a wider re-anchored sampler would continue to track its anchor. Score-based enhancers anchored on the noisy mixture do enhance at the network sizes used by Richter et al. (2023), so this result characterizes the 32-channel control studied here rather than noisy anchoring in general.

A10 PROOF OF THE DEVIATION IDENTITY

Proof. Writing $K_k^D = I + \Delta_k$ in Eq. 8 gives $\Delta_{k+1} = A_k \Delta_k + (A_k + B_k^D - I)$, with $\Delta_0 = K_0^D - I = 0$. By Eq. 12, the mean-reversion terms $h\gamma I$ in A_k and $-h\gamma I$ in B_k^D cancel, so $A_k + B_k^D = I + hcg_k^2(\nabla_x s_\theta + \nabla_D s_\theta)$. Substituting this gives Eq. 13. Equation 14 follows by substituting B_k^y from Eq. 12 into Eq. 8. □

A11 EXPERIMENTAL SETUP

Data. All experiments use VoiceBank/DEMAND at 16 kHz. Test mixtures combine clean utterances from the test set with three noise types seen during training, namely restaurant, babble and car. For each utterance and noise type, mixtures are stored at ten operating points $\alpha \in [-0.8, 2]$, scaled so that $\alpha = 0$ corresponds to an input SNR of 0 dB. This reference level lies below the 2.5 to 17.5 dB range of the test set and at the lower edge of the 0 to 15 dB range used in training. The out-of-distribution analysis in Appendix A3 uses office noise, which does not occur in training, at twelve operating points.

Models. We evaluate four configurations. StoRM-32 and StoRM-128 are StoRM systems whose NCSN++ score networks have 32 and 128 base channels, respectively. In both, the score network receives the current state, the predictor estimate and the noisy mixture as input. Predictor-32 uses the deterministic estimate of the StoRM-32 predictor directly as the enhanced output. The re-anchored variant, labeled Score-only-32 in Figures 1 and 2, is a separately trained 32-channel score network of the same architecture that is conditioned on and initialized from the noisy mixture, with no predictor.

The three 32-channel configurations form a matched-width set and are the basis of the system-level comparisons.

Sampler and input representation. The reverse process follows the OUVE SDE with stiffness $\gamma = 1.5$, $T = 1$, $\sigma_{\min} = 0.05$, $\sigma_{\max} = 0.5$ and diffusion coefficient $g(t)^2 = \sigma_{\min}^2(\sigma_{\max}/\sigma_{\min})^{2t}\, 2\log(\sigma_{\max}/\sigma_{\min})$. Unless stated otherwise, we integrate it with $N = 50$ Euler–Maruyama steps and $c = 1$, which corresponds to the reverse SDE rather than the probability flow ODE. Both networks operate on the complex STFT with an FFT size of 510, a hop of 128 samples and a periodic Hann window, followed by the amplitude compression $X \mapsto \beta|X|^a e^{i\angle X}$ with $a = 0.5$ and $\beta = 0.15$.

Perturbation model. Equation 2 is linear in s and n, so any two stored mixtures of an utterance determine both. We recover s and n from the mixtures at $\alpha = -0.8$ and $\alpha = 0$. The recovered clean signal agrees with the reference clean file to a cosine similarity of 1 at six decimal places, with relative reconstruction error between 1.9×10^{-4} and 3.4×10^{-4}, consistent with 16-bit quantization of the stored files. Mixtures at all other values of M, including the dense grids of §6.3, are synthesized from (s, n) in floating point in the waveform domain and then passed through the STFT and amplitude compression. Synthesizing them in the compressed domain would be incorrect, since the compression is nonlinear. As an independent check, the SI-SDR of the stored mixtures at $\alpha = 0$ is -0.045, $+0.021$ and $+0.011$ dB on restaurant, babble and car, with per-utterance scatter below 0.12 dB, and across the sweep it follows $-20\log_{10}(1+\alpha)$ to within 0.15 dB, as Eq. 2 predicts for a 0 dB reference.

Clipping. The stored mixtures are 16-bit files and saturate at full scale. At $\alpha = 0.7071$, the stored mixtures of four of the 22 utterances reach peak amplitudes of 0.99997 to 1.00000, and larger operating points clip further. The recovery is unaffected, since no utterance exceeds a peak of 0.99 at either $\alpha = -0.8$ or $\alpha = 0$, and the synthesized mixtures are not clipped. The operator measurements and the finite-change analysis use synthesized mixtures with $\alpha \in [-0.8, 0.7071]$. The operating points $\alpha = 1.5$ and $\alpha = 2$ are shaded in Figures 1 and 2, and no conclusion in the paper relies on them.

Utterances. All analyses use a fixed set of 22 test utterances. We report per-utterance statistics rather than averages over the 824-utterance test set, since averaging suppresses the between-utterance variation that the analysis studies. The amplitude compression has derivative $a\beta|X|^{a-1}$, which is unbounded where $|X| = 0$. Two utterances contain digital silence and therefore one time-frequency bin each at which both s and n vanish, about four parts per million of the spectrogram. The true directional derivative at such a bin is zero, and we set it to zero in the operator measurements. The total-sensitivity computation for these two utterances did not apply this correction, so their decomposition residual, modulating share and alignment are undefined. To keep every pointwise quantity in Tables 1 and 3 on a common set, we report all of them over the remaining 20 utterances. Including the two utterances in the quantities that are defined for them changes the mean projection on babble at $\alpha = 0$ from 0.902 to 0.913.

A12 LIMITATIONS

The decomposition in Eqs. 9 and 11 is an exact identity for any differentiable sampler, and its validity does not depend on the experimental choices below. The limitations in this section concern the empirical attribution, meaning the finding that the anchored channel sets the direction of the output change in the systems we study.

Models and data. All experiments use a single predictor-conditioned architecture, StoRM with an NCSN++ score network, trained on VoiceBank/DEMAND. Operator statistics are computed on a fixed set of 22 test utterances rather than the full test set. We measure the operators at two score-network widths. At 32 channels the generative stage adds little to the predictor in either SI-SDR or PESQ, so the dominance of the anchored channel could in part reflect a weak generative stage. The 128-channel model transforms the anchor perturbation more strongly and still attributes the direction of the response to the anchored channel (§6.2). Two widths are not enough to establish

how the attribution changes with capacity, and we do not experiment with a 128-channel predictor with which to separate the contributions of the two stages at that width. Predictor-conditioned systems with a stronger generative stage, or with a different placement of the predictor output, may distribute the response differently.

Perturbation. We perturb only the noise amplitude, along the path $y^{(M)} = s + \sqrt{M}\, n$ with s and n fixed. Other forms of mismatch, such as a change of noise type or the addition of reverberation, perturb the inputs in different directions and may shift the balance between the channels. The reference operating point $\alpha = 0$ corresponds to an input SNR of 0 dB, at the lower end of the SNR range seen in training. Detailed operator statistics are reported at this point for three seen noises (Appendix A13), and the finite-change analysis covers $\alpha \in [-0.8, 0.7071]$. Larger amplitudes are excluded because the stored mixtures clip there (Appendix A11). The out-of-distribution analysis uses a single stationary noise and does not include the re-anchored control.

Numerical approximations. The operators are those of the deployed $N = 50$ sampler. They change substantially between $N = 50$ and $N = 100$ (Appendix A6), so our measurements do not characterize the continuous-time reverse process. The bound in Proposition 5 shows that the deviation from pass-through cannot grow without bound as N increases, but it is not tight, and at the measured deviation it does not guarantee that K^D is non-degenerate. The geometry of K^D is therefore established by measurement alone. All operator statistics are conditional on one realization of the terminal noise and Brownian increments per utterance, and we have not measured how they vary across realizations. Two utterances are excluded from the pointwise statistics and one from the quadrature error statistics, for the reasons given in Appendices A11 and A5.

Table 5: Channel geometry across seen noise types. "dev" is $\|K^D u - u\|/\|u\|$. The last entry compares the measured total alignment with the value predicted from the relative channel norm and inter-channel cosine through Eq. 24.

configuration	proj	gain	ρ	orth	dev	share	$\cos(K^D u, K^y w)$	total / pred
babble, 32	0.902	1.311	0.693	0.951	0.956	0.156	$+0.048$	0.983 / 0.984
restaurant, 32	0.928	1.272	0.731	0.870	0.873	0.241	$+0.043$	0.954 / 0.955
car, 32	0.799	0.970	0.826	0.549	0.585	0.382	$+0.287$	0.894 / 0.893
babble, 128	0.816	1.313	0.638	1.028	1.045	0.103	-0.001	0.993 / 0.993

Interpretation of the attribution. Equation 11 attributes the finite output change by integrating along the physical perturbation path, on which D_M and $y^{(M)}$ vary together. Integrating along a different path between the same endpoints, for example by changing D first and y second, generally assigns different contributions to the two channels. The difference arises from the interaction between the channels, which our analysis does not isolate. A related consideration concerns the training data. The predictor output is a deterministic function of the mixture, so the score network is trained only on conditioning pairs of the form $(\Pi(y), y)$. The combined sensitivity $K^D u + K^y w$ along the perturbation path depends only on the network's behaviour near such pairs. The separate operators K^D and K^y, by contrast, also depend on how the network responds when one input changes while the other is held fixed, a situation not represented in training. The split between the channels is therefore partly determined by how the trained network extrapolates beyond its training distribution.

Re-anchored control. The re-anchored variant is trained at a single width, and at that width its output tracks the noisy mixture. As discussed in Appendix A9, this result characterizes the 32-channel control and does not show that a score-based enhancer anchored on the noisy mixture must fail to enhance.

A13 THE OPERATORS ACROSS NOISE TYPES

The main text reports the detailed operator geometry on babble. We repeat the same measurement at $\alpha = 0$ and $N = 50$ on restaurant and car using the 32-channel model. Table 5 also includes the 128-channel babble result for reference.

Writing $r := \|K^y w\|/\|K^D u\|$ and $c := \cos(K^D u, K^y w)$, the alignment of the total sensitivity with the anchor contribution is

$$\cos\left(K^D u, \frac{\partial \hat{s}^{(M)}}{\partial M}\right) = \frac{1 + rc}{\sqrt{1 + 2rc + r^2}}. \tag{24}$$

Equation 24 reproduces the independently measured alignment to three or four figures in every configuration in Table 5, providing a useful internal consistency check. The anchor remains the larger directional contribution for all three seen noises, although the margin is weaker on car, where the modulating share is 0.382 and the total alignment is 0.894.