Early Semantic Commitment in Diffusion Sampling

Anonymous Authors[1]

Abstract

When do diffusion trajectories first reveal their final semantic outcome? Prior work mainly studies population-level transitions, while nearby seeds can still end in different semantic basins. We introduce Trajectory Sensitivity Analysis (TSA): pair nearby noise seeds, run deterministic DDIM, and condition trajectory statistics on whether final outputs are semantically stable or flipped. State-space RMSE separates flip and stable pairs only around $\bar{\alpha}_t \sim 10^{-2}$. This is a delayed signature. An earlier cohort-level signal appears in the denoiser's clean-sample prediction as a transverse secant defect at $\bar{\alpha}_t \approx 1.7 \times 10^{-4}$, about 60 times earlier in signal fraction than the state-space onset. Pairwise predictability emerges later: a leave-seed-out probe becomes useful at larger signal fractions, and its AUC trajectory is predicted by a plug-in Mahalanobis/LDA calculation with RMS residual 0.012. TSA separates visual separability, cohort-level semantic structure, and per-pair predictability, locating the first outcome-conditioned signal in clean-sample response.

1. Introduction

Diffusion models generate samples by iteratively denoising random noise (Ho et al., 2020; Song et al., 2021b). Nearby initial noise vectors can produce final images with different semantic content, even when intermediate states remain visually indistinguishable until late in sampling. We study when information about the final semantic outcome first becomes detectable.

Prior work mostly studies population-level regimes such as noise exploration, class commitment, and sample convergence (Biroli et al., 2024; Sclocchi et al., 2025; Li & Chen, 2024; Raya & Ambrogioni, 2023). Other work studies trajectory sensitivity and semantic latent directions, including CCS, which uses approximate linearity of the initial-noise-to-output map for controlled sampling (Song et al., 2025; Kwon et al., 2023).

Recent early-prediction, intervention, and geometry work studies complementary observables. Probe-style methods predict quality or scene content from early activations, attention maps, or sparse features; intervention methods localize when concepts become editable or locked in; and seed-selection or trajectory-variance methods show that some sample properties can be predicted early (Guo et al., 2026; Cui et al., 2026; Görgün et al., 2025; Tinaz et al., 2025; Aithal et al., 2024; Xu et al., 2025). In parallel, diffusion Fisher, score-spectral, RMT, and latent-geometry work analyze $H_t = -\nabla_x^2 \log p_t$, score-Jacobian spectra, finite-data renormalization, pullback metrics, h-space directions, and Jacobian singular vectors (Wang et al., 2025; Ventura et al., 2025; Wang et al., 2026; Haas et al., 2024; Chen et al., 2024; Dockhorn et al., 2022). Semantic-geometry work also studies supervised attribute directions, paired-prompt score differences, and antithetic noise (Haas et al., 2024; Wang et al., 2023; Jia et al., 2025). TSA is complementary: it combines controlled nearby-seed perturbations, a longitudinal/transverse secant decomposition, and final flip/stable outcome stratification. Under the Tweedie/Fisher linearization, its secant variables become directional finite-difference probes of H_t.

We introduce Trajectory Sensitivity Analysis (TSA): pair nearby noise seeds, perform deterministic DDIM sampling (Song et al., 2021a), and compare intermediate trajectories according to whether their final outputs remain semantically stable or flip. In this work, semantic commitment operationally refers to outcome-conditioned structure. A statistic at sampling time t carries commitment-relevant information if it differs between pairs of trajectories that, under the same deterministic sampler, later produce stable or flipped final outputs. This definition is diagnostic, describing when information about the final outcome becomes detectable. It also separates three concepts that are often conflated: visual discriminability of intermediate states, population-level feature emergence, and per-pair predictability. We show that these occur at different times.

This distinction matters for interpreting the results. The earliest signal we report is a cohort-level outcome signal: flip

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the FoGen Workshop at ICML 2026. Do not distribute.

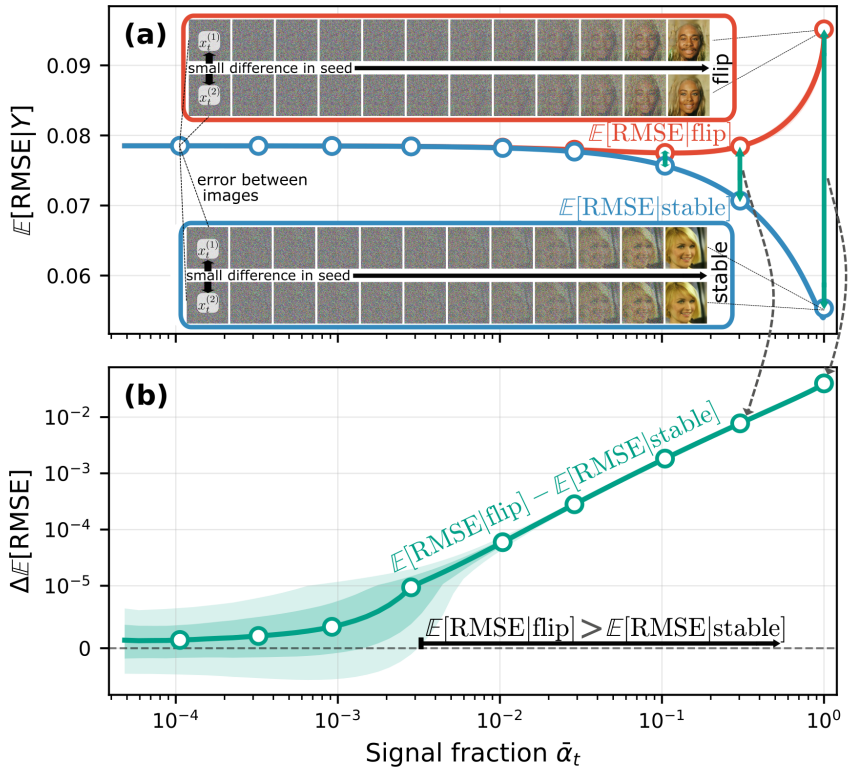

Figure 1. **Trajectory Sensitivity Analysis reveals a delayed state-space signature of semantic outcome. (a)** Two seed pairs with identical angular separation in initial noise are denoised with deterministic DDIM. The flip pair produces semantically distinct final outputs, while the stable pair produces similar outputs. Conditional RMSE trajectories remain close until late in sampling. **(b)** The outcome-conditioned state-space gap $\mathbb{E}[\mathrm{RMSE} \mid \mathrm{flip}] - \mathbb{E}[\mathrm{RMSE} \mid \mathrm{stable}]$ plotted against signal fraction $\bar{\alpha}_t$. The gap becomes positive in the low-signal regime, giving a state-space signature of semantic outcome before final images are visually distinguishable. Shaded regions indicate bootstrap confidence intervals.

and stable groups have different mean trajectories. Pairwise decodability is measured separately in Section 5, where we show that useful prediction appears later.

The state-space RMSE gap between flip and stable pairs appears around $\bar{\alpha}_t \sim 10^{-2}$, before the final images are visually distinguishable. However, RMSE only measures divergence in state space. Semantic content in a diffusion model is more directly exposed through the denoiser's clean-sample prediction. Our central result is that the RMSE gap is preceded by an earlier denoiser-space defect in this clean-sample prediction. This defect decomposes exactly into a transverse and a longitudinal component, and the clean-sample readout amplifies it by a factor of $1/\sqrt{\bar{\alpha}_t}$. Empirically, the transverse defect appears about 60 times earlier in signal fraction than the state-space RMSE gap.

Early structure at the population level does not immediately imply useful per-pair prediction. A leave-seed-out linear probe on the structural defect channels is barely above chance at the earliest transverse onset. It reaches AUC > 0.7 only at later signal fractions. Its full AUC trajectory is predicted by a plug-in Mahalanobis/LDA calculation with an RMS residual of 0.012. This calculation shows that delayed pairwise decodability follows quantitatively from mean/covariance separation. This work makes three contributions:

- **TSA and delayed state-space signature.** We condition paired deterministic DDIM trajectories on the final semantic outcome and show that state-space RMSE carries outcome-conditioned structure before the final images are visually distinguishable.
- **Fisher geometry with clean-sample amplification.** We show that the earliest detectable cohort-level signal is a transverse defect in the denoiser's clean-sample prediction, amplified by the readout factor $1/\sqrt{\bar{\alpha}_t}$ and interpretable as a directional finite-difference probe of the score-Jacobian/Fisher response.
- **Mahalanobis closure of per-pair decodability.** We show that linear probes on the defect channels become predictive only later and that their AUC trajectory is quantitatively explained by Mahalanobis/LDA separation.

2. Method

TSA compares paired sampling trajectories that differ only by a controlled angular perturbation of the initial noise. We change seed direction while keeping its norm fixed, so trajectory differences reflect angular sensitivity.

Paired seeds and deterministic sampling. For a base noise vector $z \sim \mathcal{N}(0, I)$, we construct a paired seed z' at fixed angular separation θ:

$$z' = \|z\| \left(\cos\theta\, \hat{z} + \sin\theta\, \hat{z}_\perp\right), \tag{1}$$

where $\hat{z} = z/\|z\|$ and $\hat{z}_\perp$ is a unit vector drawn uniformly from the subspace orthogonal to $\hat{z}$. This construction yields $\|z'\| = \|z\|$. The two seeds therefore share the same radius $\|z\|$, which is χ_d-distributed because z is Gaussian, and differ only in direction. The angular separation fixes the relative perturbation magnitude independent of dimension and makes θ comparable across models.

Both seeds are passed through the same deterministic DDIM sampler (Song et al., 2021a) with 100 steps and $\eta = 0$, producing paired trajectories (x_t, x'_t) indexed by signal fraction $\bar{\alpha}_t$. We use $\eta = 0$ to remove sampling stochasticity, so trajectory differences are attributable to the controlled initial perturbation. Stochastic samplers require coupled-noise ablations and are discussed as a limitation in Section 6.

Cross-model state-space TSA uses 5,000 pairs per model with $\theta = 10°$, except CIFAR-10 where we use $\theta = 20°$. CelebA robustness analyses use 62,200 DDPM-CelebA pairs at $\theta = 4.5°$. The secant-defect mechanism (Section 4) and Mahalanobis/LDA closure (Section 5) use a separate LDM-CelebA cache with 10 sampling seeds and 1,000 pairs per seed, because these analyses require cached denoiser-response differences $\delta\epsilon_t$ at each timestep.

Trajectory divergence. We measure state-space divergence with root mean squared error:

$$\mathrm{RMSE}(x_t, x'_t) = \sqrt{\frac{1}{d}\sum_{i=1}^{d}(x_{t,i} - x'_{t,i})^2}, \tag{2}$$

where d is the dimensionality of the model state. RMSE is computed in the model's native operating space: pixels for pixel-space models and latent coordinates for latent-space models. Absolute RMSE values are therefore not directly comparable across operating spaces. For cross-model comparisons, we normalize by the initial paired divergence.

Semantic outcomes. Final images are assigned semantic labels using CLIP (Radford et al., 2021). For face experiments, we also replicate the analysis with FairFace (Kärkkäinen & Joo, 2021); robustness to labeler choice is summarized in Section 3 and reported in full in Appendix B. For each binary attribute, we compare positive and negative prompt scores. We convert the two scores to a confidence by applying a two-way softmax over the positive and negative scores and taking the probability of the higher-scoring label.

An attribute is evaluated for a pair only if both final images exceed the confidence threshold τ on that attribute; otherwise the attribute is ignored for that pair. A pair flips on an evaluated attribute if the two retained labels disagree. A pair is a *flip pair* if it flips on at least one evaluated attribute, and a *stable pair* otherwise. Pairs for which no attribute is evaluated are discarded. Unless otherwise stated, we use $\tau = 0.75$. Results are robust to thresholds in the range $\tau \in [0.50, 0.95]$.

For CIFAR-10 and ImageNet, we use class disagreement as a coarser semantic outcome. A pair flips if the predicted final classes differ. Class disagreement is less fine-grained than face-attribute flips, and we use object-domain results as a generality check.

Outcome-conditioned gaps and onset. At each signal fraction, TSA compares flip and stable cohorts. The state-space gap is

$$\Delta E(\bar{\alpha}_t) = \mathbb{E}[\mathrm{RMSE}(x_t, x'_t) \mid \mathrm{flip}] - \mathbb{E}[\mathrm{RMSE}(x_t, x'_t) \mid \mathrm{stable}]. \tag{3}$$

A positive gap means that pairs destined to flip are already farther apart than pairs destined to remain stable.

We use the same convention for all onset claims. Onset is defined as the earliest timestep $t^\star$ such that the lower bound of the 95% bootstrap confidence interval for the flip-stable gap is above zero at $t^\star$ and at the four subsequent DDIM steps. This five-step rule suppresses isolated significant timesteps caused by local bootstrap fluctuations in the trajectory grid. Confidence intervals are computed by percentile bootstrap resampling over trajectory pairs using 500 bootstrap iterations. The same onset rule is applied to the RMSE gap and to the structural defect gaps introduced in Section 4.

We use "clean-sample prediction" to refer to the DDIM $\hat{x}_{0,t}$ prediction in the model's native space: an image for pixel-space models and a latent sample for latent-space models.

3. State-space TSA phenomenon

We first study what can be detected from the sampled state alone. This section examines RMSE in the model state space and treats it as an observable state-space signature of semantic outcome. The mechanism section then shows that the denoiser's clean-sample prediction contains an earlier signal.

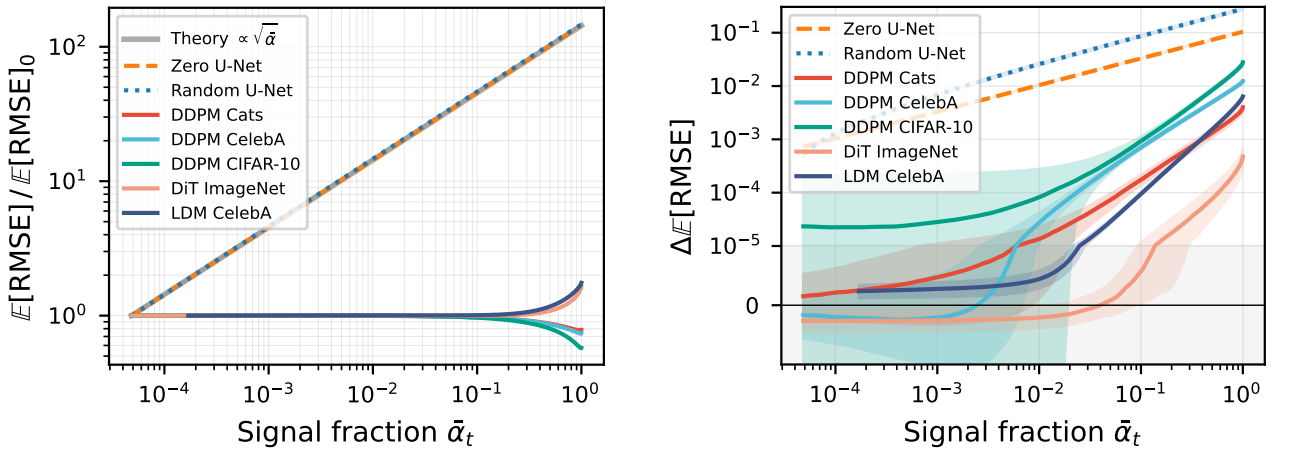

Figure 2. **State-space TSA establishes the state-space phenomenon. (a)** Normalized RMSE between paired trajectories. Null baselines grow as $\sqrt{\bar{\alpha}_t}$, while trained denoisers strongly suppress trajectory divergence. **(b)** Normalized flip-stable RMSE gap across models. Trained models show outcome-conditioned state-space gaps in the low-signal regime; Section 4 shows that this gap is preceded by a clean-sample response defect.

Unconditional trajectory compression. Unconditional trajectory compression provides the baseline for TSA. If trained denoisers behaved like pure diffusion dynamics, nearby seeds would separate according to the growing signal component. Figure 2a compares five trained models against two null baselines. A Zero U-Net returns $\epsilon_\theta(x, t) = 0$ at every step, preserving the DDIM update while removing learned structure. A random U-Net uses untrained weights as a second null baseline.

Both null baselines grow as $\sqrt{\bar{\alpha}_t}$, as expected for pure diffusion dynamics. Trained models behave differently. Across U-Net and Transformer architectures, pixel and latent operating spaces, and face/object domains, learned denoisers strongly suppress paired-trajectory divergence relative to the pure-diffusion null, and in most settings keep trajectories near or below their initial separation by the end of sampling. This compression is the unconditional baseline against which we next compare outcome-conditioned divergence.

Outcome-conditioned gap across models. The key TSA question is whether compressed trajectories differ by final semantic outcome. We therefore compare RMSE between flip and stable cohorts at each signal fraction. Figure 2b shows that the outcome-conditioned RMSE gap is not specific to one architecture or dataset. Across trained models, flip pairs have larger RMSE than stable pairs in the low-signal regime, before final images are visually distinguishable.

The null baselines (dashed and dotted) show a different pattern. Their gaps reflect the geometry of random perturbations crossing semantic decision boundaries and follow $\sqrt{\bar{\alpha}_t}$ diffusive scaling. In trained models, the gap provides a reproducible state-space signature of final semantic outcome.

Robustness. Because TSA depends on semantic labels and seed perturbations, we test the main RMSE gap under alternative labelers, thresholds, perturbation angles, and shuffled labels. CLIP and FairFace produce consistent face-attribute gap curves. Varying the confidence threshold changes sample size but preserves the qualitative structure. Changing angular separation rescales the gap without eliminating it. Shuffling outcome labels destroys the gap. Full robustness plots are reported in Appendix B.

Together, these results establish the RMSE gap as a robust state-space signature of final semantic outcome. They do not explain why the signal appears when it does. The next section shows that the RMSE gap is preceded by an earlier clean-sample response defect.

4. Secant-defect mechanism

State-space RMSE reveals an outcome-conditioned state-space gap, but it is not the earliest detectable signal. The DDIM sampler uses the denoiser through its predicted noise, while semantic content is more directly exposed through the denoiser's clean-sample prediction $\hat{x}_{0,t}$. We therefore compare paired trajectories through the clean-sample response induced by the denoiser. This gives an exact geometric object: the defect between the current state secant and the denoiser-response secant.

Clean-sample state-slaving. Let

$$\delta x_t = x'_t - x_t, \qquad \delta\epsilon_t = \epsilon_\theta(x'_t, t) - \epsilon_\theta(x_t, t). \quad (4)$$

Define the normalized state secant and normalized denoiser-response secant

$$u_t = \frac{\delta x_t}{\|\delta x_t\|}, \qquad v_t = \frac{\sqrt{1-\bar{\alpha}_t}\,\delta\epsilon_t}{\|\delta x_t\|}. \tag{5}$$

The scaling in v_t is chosen so that perfect clean-sample state-slaving corresponds exactly to $v_t = u_t$. In that case, the denoiser response cancels the paired state difference in the clean-sample prediction.

Indeed, the clean-sample prediction induced by the noise predictor is

$$\hat{x}_{0,t} = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(x_t,t)}{\sqrt{\bar{\alpha}_t}}. \tag{6}$$

Taking paired differences gives

$$\delta\hat{x}_{0,t} = \frac{\delta x_t - \sqrt{1-\bar{\alpha}_t}\,\delta\epsilon_t}{\sqrt{\bar{\alpha}_t}} = \frac{\|\delta x_t\|}{\sqrt{\bar{\alpha}_t}}(u_t - v_t). \tag{7}$$

Thus the vector $u_t - v_t$ is not an arbitrary residual: it is exactly the clean-sample response defect, measured relative to the current state separation. Any deviation from clean-sample state-slaving is amplified by $1/\sqrt{\bar{\alpha}_t}$ in clean-sample space. At the observed transverse onset $\bar{\alpha}_t \approx 1.7 \times 10^{-4}$, this factor is about 77; at $\bar{\alpha}_t = 10^{-4}$, it is 100. This amplification explains how a denoiser-response defect can be detectable well before state-space trajectories separate.

This normalization makes the secant defect $u_t - v_t$ a meaningful geometric object. Decomposing the denoiser-response secant relative to u_t separates failures of state-slaving into longitudinal and transverse channels.

Transverse and longitudinal defect. To distinguish the channels through which the clean-sample response defect appears, we decompose v_t into components parallel and perpendicular to the state secant:

$$v_t = R_t u_t + w_t, \qquad R_t = u_t^\top v_t, \qquad u_t^\top w_t = 0. \tag{8}$$

Here R_t is a scalar projection of the denoiser response along the state secant, not a cosine, since v_t is not unit-normalized. The component w_t is the transverse or rotational response.

Substituting Eq. (8) into the defect gives

$$u_t - v_t = (1 - R_t)u_t - w_t. \tag{9}$$

By orthogonality of u_t and w_t, and since $\|u_t\| = 1$,

$$\|u_t - v_t\|^2 = \|w_t\|^2 + (1 - R_t)^2. \tag{10}$$

The total secant defect therefore decomposes exactly into a transverse channel $\|w_t\|^2$ and a longitudinal channel $(1 - R_t)^2$.

Squaring Eq. (7) gives

$$\|\delta\hat{x}_{0,t}\|^2 = \frac{\|\delta x_t\|^2}{\bar{\alpha}_t}\|u_t - v_t\|^2. \tag{11}$$

Substituting Eq. (10) yields the clean-sample amplification identity:

$$\frac{\|\delta\hat{x}_{0,t}\|^2}{\|\delta x_t\|^2} = \frac{\|w_t\|^2 + (1-R_t)^2}{\bar{\alpha}_t}. \tag{12}$$

This equation identifies the readout geometry behind the onset separation. The clean-sample response amplitude factorizes as

$$\frac{\|\delta\hat{x}_{0,t}\|}{\|\delta x_t\|} = \frac{1}{\sqrt{\bar{\alpha}_t}}\sqrt{\|w_t\|^2 + (1-R_t)^2}. \tag{13}$$

Thus early detectability is not due to the readout factor alone. It reflects the product of the clean-sample readout amplification and the growth of the defect channels themselves.

Fisher interpretation. The same variables also admit a local score-geometry interpretation. Under the exact-score Tweedie relation

$$\epsilon_\theta(x_t,t) \approx -\sqrt{1-\bar{\alpha}_t}\,\nabla_x \log p_t(x_t), \tag{14}$$

a first-order expansion of the paired denoiser response gives

$$v_t = \frac{\sqrt{1-\bar{\alpha}_t}\,\delta\epsilon_t}{\|\delta x_t\|} \approx (1-\bar{\alpha}_t)H_t u_t, \qquad H_t = -\nabla_x^2 \log p_t(x_t). \tag{15}$$

Thus the transverse channel rescales to

$$F_t^{\mathrm{trn}} = \frac{\|w_t\|}{1-\bar{\alpha}_t} \approx \|P_{u_t^\perp} H_t u_t\|, \tag{16}$$

while the longitudinal state-slaving defect rescales to

$$D_t^{\mathrm{slave}} = \frac{1-R_t}{1-\bar{\alpha}_t} \approx \frac{1}{1-\bar{\alpha}_t} - u_t^\top H_t u_t. \tag{17}$$

The origin $(D_t^{\mathrm{slave}}, F_t^{\mathrm{trn}}) = (0,0)$ corresponds to perfect clean-sample state-slaving. For learned non-conservative denoisers, these quantities should be read as empirical finite-difference probes of the denoiser/score Jacobian, and as Fisher/Hessian probes under the exact-score approximation.

Onset hierarchy. Figure 3a shows the resulting onset hierarchy under the common onset rule from Section 2. The transverse defect $\Delta\|w_t\|^2$ and total secant defect $\Delta\|u_t - v_t\|^2$ appear first, at $\bar{\alpha}_t \approx 1.7 \times 10^{-4}$. The longitudinal defect $\Delta(1-R_t)^2$ appears later, near $\bar{\alpha}_t \approx 10^{-3}$. The state-space RMSE gap appears last, near $\bar{\alpha}_t \approx 9.5 \times 10^{-3}$.

These are cohort-level statistical onsets under a fixed sample size and onset rule; Section 5 shows that useful pairwise

prediction appears later. The transverse defect therefore precedes the state-space RMSE gap by nearly two orders of magnitude in signal fraction. Equivalently, the RMSE gap is a delayed state-space signature of an earlier clean-sample response defect. We interpret this as a geometric mechanism for the diagnostic signal: we identify the denoiser-response component that precedes later state-space separation, but do not show that manipulating this component steers final outcomes.

5. Population signal precedes pairwise decodability

The onset hierarchy in Section 4 is a cohort-level statement: flip and stable pairs have different mean defect trajectories. It does not imply that individual pairs are classifiable at the earliest onset. This distinction is important because semantic commitment can appear as a weak population signal before it becomes useful for pairwise prediction.

To test pairwise decodability, we train leave-seed-out logistic regressions on the four structural defect scalars at each signal fraction:

$$z_t = \left(R_t,\ \|w_t\|^2,\ (1-R_t)^2,\ \|u_t - v_t\|^2\right). \tag{18}$$

Here "seed" refers to the independent cache seed used to generate a block of paired trajectories. All pairs from the held-out cache seed are excluded from training and used only for testing. The logistic regression is low-dimensional and linear. We use it to study whether the secant-defect channels themselves carry information about the outcome, without using pixels, classifier confidence, or intermediate network activations.

We also compare the observed probe AUC with the prediction of a simple Gaussian model with equal covariance. Let μ_F and μ_S be the means of z_t in the flip and stable cohorts, and let Σ_t be the pooled covariance. Because $\|u_t - v_t\|^2 = \|w_t\|^2 + (1-R_t)^2$, the four-feature covariance is structurally singular: the fourth feature is an exact linear combination of the second and third. We therefore compute the Mahalanobis separation with a ridge-regularized pseudo-inverse Σ_t^+, using $\Sigma_t^+ = (\Sigma_t + \lambda_t I)^\dagger$ with $\lambda_t = 10^{-5}\,\mathrm{tr}(\Sigma_t)/d$. Equivalently, dropping the redundant total-defect feature gives the same curve within numerical precision. The squared Mahalanobis separation is

$$M_t^2 = (\mu_F - \mu_S)^\top \Sigma_t^+ (\mu_F - \mu_S). \tag{19}$$

Under the Gaussian approximation with equal covariance, the AUC of the optimal linear discriminator is

$$\mathrm{AUC}_t \approx \Phi(M_t/\sqrt{2}), \tag{20}$$

where Φ is the standard normal CDF. This plug-in prediction uses the empirical means and covariance of the structural channels at each signal fraction; it is not fitted to the observed AUC trajectory.

Figure 3b resolves the apparent tension between early cohort-level onset and later pairwise predictability. At the transverse onset, the structural defect gap is statistically detectable, but the within-cohort variance remains large, so a single pair is only weakly classifiable. As the signal fraction increases, Mahalanobis separation increases and the logistic regression becomes predictive. The predicted AUC tracks the observed leave-seed-out AUC with an RMS residual of 0.012. Thus, under an equal-covariance Gaussian model, the first two moments of the structural defect channels account for the observed decodability curve to within RMS residual 0.012. Within this low-dimensional probe setup, no additional nonlinear correction is needed to match the measured AUC trajectory.

6. Limitations

TSA establishes outcome-conditioned structure as a diagnostic. It does not test whether intervening on the defect channels would steer outcomes.

Semantic outcomes are operationally defined by classifiers. CLIP and FairFace produce consistent labels and the main TSA gap is robust across labelers, thresholds, and angular separations (Section 3); the present analysis is restricted to binary attributes and class disagreement. We compare to population-level transition work only at the level of signal-fraction scale, and do not measure class-entropy or marginal class-commitment transitions on the same models.

The current evaluation focuses on unconditional generation. Text-conditional diffusion is a natural next setting: the secant-defect identities follow algebraically from deterministic DDIM and should carry over, but whether the onset hierarchy and Mahalanobis/LDA agreement replicate quantitatively under classifier-free guidance remains future work.

7. Conclusion

Trajectory Sensitivity Analysis reveals semantic outcome information along paired diffusion trajectories before final images are visually distinguishable. Different observables reveal commitment at quantitatively distinct signal fractions: the transverse defect at $\bar{\alpha}_t \approx 1.7 \times 10^{-4}$, the longitudinal defect near 10^{-3}, and the state-space RMSE gap near 10^{-2}. The clean-sample readout factor $1/\sqrt{\bar{\alpha}_t}$ amplifies deviations from state-slaving, which exposes the cohort-level signal early and explains the onset hierarchy. Pairwise predictability is delayed further and is quantitatively explained by Mahalanobis/LDA separation in the same defect channels with RMS residual 0.012.

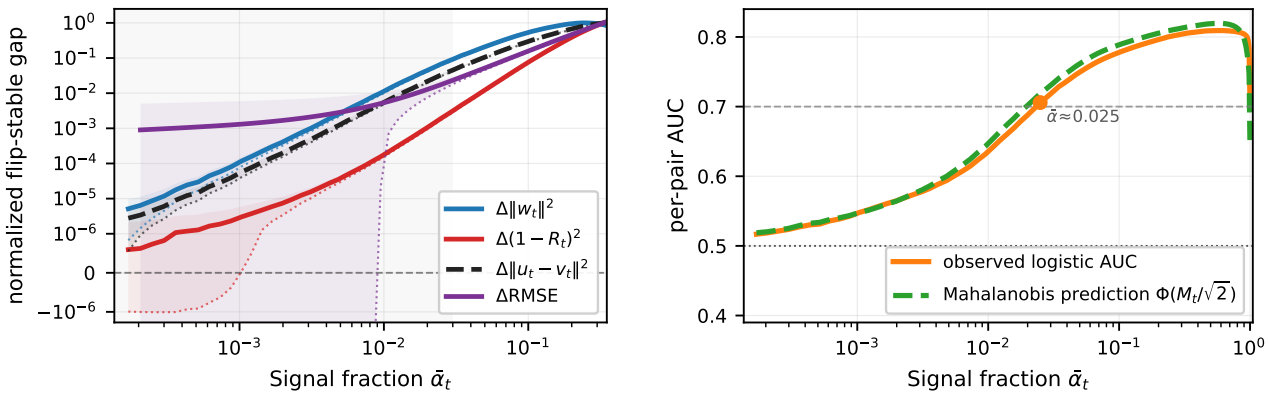

Figure 3. **Secant-defect geometry separates cohort-level onset from pairwise decodability. (a)** Peak-normalized flip-stable gaps for transverse, longitudinal, total secant-defect, and state-space RMSE channels. The transverse and total defects appear first, followed by the longitudinal defect and then state-space RMSE. **(b)** Leave-seed-out probe AUC on the four defect scalars. Pairwise decodability is delayed relative to cohort-level onset and is tracked by the plug-in Mahalanobis/LDA prediction with RMS residual 0.012.

The earliest diagnostic signature of final semantic outcome is therefore a clean-sample response phenomenon, with state-space divergence emerging as its delayed signature. TSA separates three notions of commitment that are often conflated in diffusion sampling: visual separability of intermediate states, cohort-level semantic structure, and pairwise predictability. Under the Tweedie/Fisher interpretation, the transverse channel is a directional finite-difference probe of the denoiser/score geometry, and empirically it is the earliest outcome-conditioned channel in the onset hierarchy.

References

Aithal, S. K., Maini, P., Lipton, Z. C., and Kolter, J. Z. Understanding hallucinations in diffusion models through mode interpolation. In *Advances in Neural Information Processing Systems*, 2024.

Biroli, G., Bonnaire, T., de Bortoli, V., and Mézard, M. Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957, 2024.

Chen, S., Zhang, H., Guo, M., Lu, Y., Wang, P., and Qu, Q. Exploring low-dimensional subspaces in diffusion models for controllable image editing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Cui, B., Huang, B., Ye, Z., Dong, X., Chen, T., Xue, H., Yang, D., Huang, L., Tang, J., and Hong, H. Diffusion probe: Generated image result prediction using CNN probes. *arXiv preprint arXiv:2602.23783*, 2026.

Dockhorn, T., Vahdat, A., and Kreis, K. GENIE: Higher-order denoising diffusion solvers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Görgün, A., Sammani, F., Deligiannis, N., Schiele, B., and Fischer, J. Temporal concept dynamics in diffusion models via prompt-conditioned interventions. *arXiv preprint arXiv:2512.08486*, 2025.

Guo, H., Wei, H., and Jing, B. Toward early quality assessment of text-to-image diffusion models. *arXiv preprint arXiv:2603.02829*, 2026.

Haas, R., Huberman-Spiegelglas, I., Mulayoff, R., Graßhof, S., Brandt, S. S., and Michaeli, T. Discovering interpretable directions in the semantic latent space of diffusion models. In *18th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2024.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Jia, J., Liu, S., Song, B., Yuan, W., Shen, L., and Wang, G. Antithetic noise in diffusion models. *arXiv preprint arXiv:2506.06185*, 2025.

Kärkkäinen, K. and Joo, J. FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021.

Kwon, M., Jeong, J., and Uh, Y. Diffusion models already have a semantic latent space. In *International Conference on Learning Representations*, 2023.

Li, M. and Chen, S. Critical windows: non-asymptotic theory for feature emergence in diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 27474–27498, 2024.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable

visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Raya, G. and Ambrogioni, L. Spontaneous symmetry breaking in generative diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 66377–66389, 2023.

Sclocchi, A., Favero, A., and Wyart, M. A phase transition in diffusion models reveals the hierarchical nature of data. *Proceedings of the National Academy of Sciences*, 122 (1):e2408799121, 2025.

Song, B., Zhang, Z., Luo, Z., Hu, J., Yuan, W., Jia, J., Tang, Z., Wang, G., and Shen, L. CCS: Controllable and constrained sampling with diffusion models via initial noise perturbation. In *Advances in Neural Information Processing Systems*, 2025.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b.

Tinaz, B., Fabian, Z., and Soltanolkotabi, M. Emergence and evolution of interpretable concepts in diffusion models. In *Advances in Neural Information Processing Systems*, 2025.

Ventura, E., Achilli, B., Silvestri, G., Lucibello, C., and Ambrogioni, L. Manifolds, random matrices and spectral gaps: The geometric phases of generative diffusion. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.05898.

Wang, B., Zavatone-Veth, J. A., and Pehlevan, C. A random matrix theory perspective on the consistency of diffusion models. *arXiv preprint arXiv:2602.02908*, 2026.

Wang, F., Yin, H., Zhuang, S., Zhu, H., Li, Y., Qian, L., Zhang, C., Zhao, H., Qian, H., and Li, C. Efficiently access diffusion Fisher: Within the outer product span space. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

Wang, Z., Gui, L., Negrea, J., and Veitch, V. Concept algebra for (score-based) text-controlled generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Xu, K., Zhang, L., and Shi, J. Good seed makes a good crop: Discovering secret seeds in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3024–3034, 2025.

A. Model and experimental details

We evaluate the models and datasets in Table 1. All main experiments use 100-step deterministic DDIM sampling with $\eta = 0$. Cross-model state-space TSA uses 5,000 pairs per model. CelebA robustness analyses use 62,200 DDPM-CelebA pairs at $\theta = 4.5°$. Secant-defect and Mahalanobis/LDA analyses use an LDM-CelebA cache with 10 sampling seeds and 1,000 pairs per seed. For classifier-defined outcomes, we use confidence threshold $\tau = 0.75$ unless otherwise stated.

Table 1. **Models evaluated.** All model-comparison experiments use 100-step deterministic DDIM sampling with $\eta = 0$.

Model	Architecture	Domain	Dataset
DDPM	U-Net (pixel)	Faces	CelebA-HQ
DDPM	U-Net (pixel)	Cats	AFHQ
DDPM	U-Net (pixel)	Objects	CIFAR-10
LDM	U-Net (latent)	Faces	CelebA-HQ
DiT	Transformer (latent)	Objects	ImageNet

Trajectory caches are generated on a single A100 GPU per run; a 1,000-pair deterministic DDIM cache takes approximately 40 minutes, and downstream probe/Mahalanobis analyses run in CPU minutes after caching. All datasets and pretrained checkpoints are used as publicly available research assets under their published terms of use; no dataset is redistributed by this submission.

B. Robustness checks

Figure 4 reports the robustness checks summarized in Section 3. The outcome-conditioned RMSE gap persists across labeler choice, confidence threshold, and perturbation angle, while label shuffling destroys the gap. LPIPS-based cohorts are formed by thresholding final-image perceptual distance and are used only as a robustness check, not as the primary semantic definition.

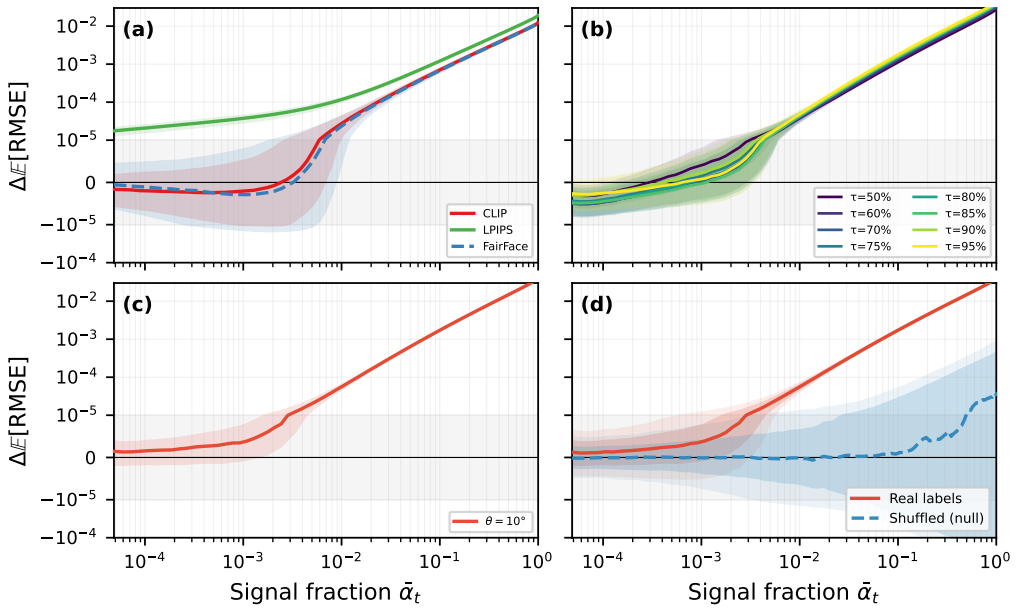

Figure 4. **Robustness of the TSA state-space gap.** The outcome-conditioned gap is robust to classifier source, confidence threshold, and angular separation. The main paper summarizes these conclusions in Section 3.

C. Mahalanobis/LDA calculation

For completeness, we give the calculation behind Eq. (20). Under an equal-covariance Gaussian model for the structural defect features at time t,

$$X \mid Y = 1 \sim \mathcal{N}(\mu_F, \Sigma_t), \qquad X \mid Y = 0 \sim \mathcal{N}(\mu_S, \Sigma_t), \tag{21}$$

the optimal linear score is

$$s(X) = (\mu_F - \mu_S)^\top \Sigma_t^+ X, \tag{22}$$

where Σ_t^+ is the regularized pseudo-inverse used in the main text. Define

$$M_t^2 = (\mu_F - \mu_S)^\top \Sigma_t^+ (\mu_F - \mu_S). \tag{23}$$

For this score, the class-conditional score distributions are one-dimensional Gaussians with variance M_t^2 and mean separation M_t^2, so their standardized separation is M_t. The probability that a random flip score exceeds a random stable score is therefore

$$\mathrm{AUC}_t = \Phi(M_t/\sqrt{2}). \tag{24}$$