READABILITY IS NOT RELIABILITY: SEMANTIC READOUTS IN UNCONDITIONAL DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Semantic probes are widely used to interpret diffusion representations, yet accurate decoding alone does not guarantee that the underlying predictive relationship persists across noise realizations and generative states. In this work, we systematically investigate semantic readout reliability in unconditional diffusion models by distinguishing mere readability from reproducibility under noise resampling and predictive reuse along reverse trajectories. Through theoretical analysis, we characterize why readable representations can yield unreliable readouts: changes in the optimal predictive mapping undermine cross-state reuse, while nuisance variation and insufficient identifiability limit reproducible readout recovery. For linear readouts, we derive exact risk decompositions to quantify decoder mismatch and the $1/K$ nuisance contribution from averaging K independent noise realizations, establishing sufficient conditions for predictive-subspace recovery and local reuse. Empirical evaluations on CelebA-HQ latent diffusion and CIFAR-10 pixel-space diffusion reveal readable representations possessing with unstable readout geometries. Furthermore, an independent analysis of 5,000 reverse trajectories demonstrates that accurate attribute decoding frequently coexists with severe transfer failures of an unchanged readout. Our results expose the critical gap between semantic accessibility and readout persistence, proving that reliable diffusion interpretability demands evidence beyond static probe accuracy.

1 INTRODUCTION

Diffusion models contain hidden representations that encode rich semantic information. Intermediate activations have been reused for correspondence, segmentation, depth, concept localization, and image manipulation(Baranchuk et al., 2022; Hedlin et al., 2023; Yang & Wang, 2023; Tang et al., 2023a; Luo et al., 2023; Zhao et al., 2023; Ke et al., 2024; Kadkhodaie et al., 2025; Sclocchi et al., 2025), and linear probes have become a routine tool for evaluating whether a semantic variable is decodable from a chosen layer and noise level. Yet a high probe score answers only a local question: *can this target be decoded here?* It does not tell us whether the same predictive relationship survives nuisance variation or persists as the generative computation evolves.

This distinction matters because diffusion representations are dynamic: the same semantic variable can be evaluated across different stochastic realizations or different states along the reverse generation process. A readout that succeeds at one state may therefore rely on a transient predictive relationship rather than a persistent one(Huang & Hahn, 2026; Paduraru et al., 2026). For example, two nearby diffusion states may both support accurate head-pose prediction while relying on different predictive mappings, causing a probe trained at one state to fail when transferred to the other. Likewise, a predictive direction identified from one noise realization may not remain consistent under another realization of the same underlying image. In both cases, the target is readable, but the interpretation is fragile. Treating such readouts as persistent semantic representations would overstate what probing establishes.

We therefore study semantic readout reliability: the persistence of an evaluator-defined predictive relationship beyond the representation condition in which it is identified. We distinguish *readability* from two complementary forms of reliability: *noise reproducibility* and *cross-time predictive reuse*.

Readability asks whether a target can be decoded from a given representation state. Noise reproducibility asks whether the same predictive relation survives nuisance variation. Cross-time predic-

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.

tive reuse asks whether a readout remains usable as the representation evolves along the generation process. These properties are distinct and can fail independently.

We combine theoretical analysis with controlled noise resampling and readout transfer along reverse-generation trajectories to explain and evaluate the gap between semantic readability and readout reliability. Analysis reveals that semantic readability and semantic reliability are distinct properties of diffusion representations. Across semantic variables and diffusion settings, we find that readability often emerges before stronger reliability properties: semantic variables can become linearly decodable while their readouts remain unstable under nuisance perturbations or fail to transfer across generation states. In summary, our contributions are as follows:

1. **Theoretical Foundations of Cross-State Reuse:** We theoretically explain the gap between semantic readability and cross-state reuse through an exact risk decomposition and sufficient conditions for local predictive reuse.
2. **Limits of Readout Reproducibility:** We derive an exact $1/K$ nuisance-risk decomposition under repeated noising and establish predictive-subspace recovery guarantees under explicit identifiability conditions.
3. **Empirical Validation of the Readability-Reliability Gap:** We demonstrate that semantic accessibility does not guarantee readout stability. Comprehensive evaluations across latent and pixel-space diffusion models reveal unstable readout geometries despite highly accurate same-state decoding.

2 Related Work

Diffusion representations. Diffusion models are iterative generative systems built from denoising or score estimation (Ho et al., 2020; Song et al., 2021; Rombach et al., 2022). Their hidden states have become useful general-purpose visual descriptors: prior work has used diffusion features for semantic correspondence, segmentation, geometry, and transfer (Hedlin et al., 2023; Zhao et al., 2023; Xiang et al., 2023; Ke et al., 2024). These results establish accessibility of semantic information, but typically evaluate a frozen state or aggregate multiple states. Our focus is different: whether a semantic readout remains reproducible when nuisance noise changes and reusable when the reverse state changes.

Interpretability and concept analysis in diffusion models. Interpretability work has localized concepts through attention, feature manipulation, sparse decompositions, or explicit prototypes (Tang et al., 2023b; Gandikota et al., 2023; 2024; Kim et al., 2025; Surkov et al., 2025; Tinaz et al., 2025). Recent work also studies how interpretable concepts evolve across diffusion states. We complement these efforts by auditing the reliability of a readout rather than claiming to identify a unique intrinsic semantic coordinate system.

Auditing and subspace interpretability. Recent interpretability work increasingly treats representation structure itself as an object of audit. Patronus uses prototypes to expose what, where, and when semantic patterns appear in diffusion generation (Weng et al., 2026); Huang and Hahn recover non-axis-aligned interpretable subspaces in language models (Huang & Hahn, 2026); BEE uses probing and embedding geometry to expose spurious predictive structure missed by conventional evaluation (Paduraru et al., 2026); and block-recurrent analyses interpret layerwise computation through reusable dynamical structure (Jacobs et al., 2026). These works motivate going beyond endpoint accuracy when attaching meaning to an internal representation. Our focus is narrower and complementary: for a fixed evaluator-defined target, we seek to characterize the conditions governing semantic readout reliability under nuisance resampling and state changes.

Probing, subspaces, and representation reliability. Linear and reduced-rank probes are widely used to measure information content, and low-rank regression provides a natural predictive subspace when the target dimension is small (Mukherjee & Zhu, 2011; Denton et al., 2014; Degras et al., 2024). Probe success, however, is not itself evidence of invariance or transportability. Our repeated-noise design turns nuisance reproducibility into an estimable object; our reverse analysis tests predictive reuse directly. This places the work closer to representation auditing than to a claim of mechanistic causality.

3 STATISTICAL FOUNDATIONS OF SEMANTIC READOUT RELIABILITY

Linear probing establishes whether a target is accessible at a particular representation condition. It does not establish whether the resulting readout remains predictive when that condition changes. We therefore ask: *what distinguishes a readable semantic variable from a reliable semantic readout?* Proofs are given in Appendix A.

3.1 PROBLEM FORMULATION: READABILITY VERSUS RELIABILITY

Let e index a representation condition, $h_e \in \mathbb{R}^p$ the corresponding feature, and $y \in \mathbb{R}^q$ a fixed evaluator-defined target. For controlled noising, $y = g(x)$ is a measurement of the underlying image; along a reverse trajectory, it is a measurement of the final generated sample. The target is held fixed across the conditions being compared. Cross-condition reuse is evaluated within a common feature space, with the layer, feature coordinates, and preprocessing held fixed. Semantics refers to the specified target, without assuming a unique internal semantic coordinate system.

We include an intercept through $\phi_e = [1, h_e^\top]^\top$ and consider readouts $\widehat{y} = \Theta^\top \phi_e$. With finite second moments, define

$$\mathcal{L}_e(\Theta) = \mathbb{E}\|y - \Theta^\top \phi_e\|_2^2, \qquad \mathcal{J}_e(\Theta) = \mathcal{L}_e(\Theta) + \lambda \operatorname{tr}(\Theta^\top P \Theta), \tag{1}$$

where $P = \operatorname{diag}(0, I_p)$ leaves the intercept unpenalized and $\lambda \geq 0$ is fixed across conditions. Write $G_e = \mathbb{E}[\phi_e \phi_e^\top]$, $C_e = \mathbb{E}[\phi_e y^\top]$, and $A_e = G_e + \lambda P$. Whenever $A_e \succ 0$, the population ridge readout is $\Theta_e^\star = A_e^{-1} C_e$.

Readability is assessed by the same-condition prediction risk $\mathcal{L}_e(\Theta_e^\star)$ relative to a constant-prediction baseline. *Predictive reuse* is assessed by applying a source readout unchanged at a destination and comparing $\mathcal{L}_d(\Theta_s^\star)$ with $\mathcal{L}_d(\Theta_d^\star)$. Controlled noise reproducibility and reverse-time reuse test different changes of condition; neither implies the other. A reliable semantic readout must retain useful predictive performance under the change being tested.

3.2 A DECOMPOSITION OF READOUT RELIABILITY

Two conditions can each support accurate prediction while requiring different readouts. The following identity makes this mismatch explicit.

Theorem 1 (Readability–reliability decomposition)**.** *For source and destination conditions with $A_s, A_d \succ 0$, let $D_{s,d} = \Theta_s^\star - \Theta_d^\star$. Then*

$$\mathcal{J}_d(\Theta_s^\star) - \mathcal{J}_d(\Theta_d^\star) = \|A_d^{1/2} D_{s,d}\|_F^2. \tag{2}$$

The corresponding prediction-risk difference is

$$\mathcal{L}_d(\Theta_s^\star) - \mathcal{L}_d(\Theta_d^\star) = \|G_d^{1/2} D_{s,d}\|_F^2 - 2\lambda \operatorname{tr}(D_{s,d}^\top P \Theta_d^\star). \tag{3}$$

In particular, when $\lambda = 0$, destination risk is the sum of optimal destination risk and a nonnegative, destination-weighted readout mismatch.

The mismatch depends on the destination feature distribution, rather than Euclidean coefficient distance alone. For example, let y have zero mean and unit variance, $h_s = y$, and $h_d = ay$ with $a \neq 0$. Without regularization, both conditions have zero optimal risk, but transferring the source readout gives destination $R^2 = 1 - (a-1)^2$, which can be arbitrarily negative. Conversely, equally uninformative representations can share the same constant predictor. A small reuse gap is therefore meaningful only together with readability. For ridge fits, Eq. (3) also distinguishes the regularized objective from the unpenalized risk measured by held-out R^2.

From pairwise reuse to a shared readout. A predictor fitted jointly over a trajectory segment answers a different question from a predictor fitted at one source. For weights $\pi_t \geq 0$ with $\sum_{t \in \mathcal{T}} \pi_t = 1$, minimizing $\sum_t \pi_t \mathcal{J}_t(\Theta)$ gives

$$\Theta_{\mathcal{T}}^\star = A_{\mathcal{T}}^{-1} \sum_{t \in \mathcal{T}} \pi_t A_t \Theta_t^\star, \qquad A_{\mathcal{T}} = \sum_{t \in \mathcal{T}} \pi_t A_t. \tag{4}$$

Thus the shared readout is a matrix-weighted barycenter determined by the predictive objectives. Its average regularized excess risk is no greater than that of any single-source readout, although this does not guarantee small error at every state. Shared-suffix performance and single-source transport must therefore be evaluated separately.

What drives readout drift? The normal equations yield

$$D_{s,d} = \underbrace{A_s^{-1}(C_s - C_d)}_{\text{feature–target coupling}} + \underbrace{A_s^{-1}(A_d - A_s)A_d^{-1}C_d}_{\text{feature geometry}}. \tag{5}$$

Readout drift can arise from either changing target associations or changing feature second moments, and poor conditioning can amplify both. This gives a sufficient condition for local reuse.

Corollary 2 (Trajectory-local reuse)**.** *On a trajectory segment parameterized by τ, suppose $a_- I \preceq A_\tau \preceq a_+ I$ with $a_- > 0$, $\|C_\tau\|_F \le c_0$, and $\|A_s - A_d\|_{\mathrm{op}} \le L_A|\tau_s - \tau_d|$, $\|C_s - C_d\|_F \le L_C|\tau_s - \tau_d|$. For $\gamma = L_C/a_- + L_A c_0/a_-^2$,*

$$\|D_{s,d}\|_F \le \gamma|\tau_s - \tau_d|, \qquad \mathcal{J}_d(\Theta_s^\star) - \mathcal{J}_d(\Theta_d^\star) \le a_+\gamma^2|\tau_s - \tau_d|^2. \tag{6}$$

Smooth, well-conditioned moment evolution therefore guarantees local reuse in the regularized objective. The same quadratic guarantee applies directly to prediction risk when $\lambda = 0$; otherwise the correction in Eq. (3) must be retained. This is a sufficient condition, not a claim that diffusion trajectories necessarily satisfy it or that readability must precede reuse.

3.3 Recovering Reliable Readouts from Repeated Perturbations

Cross-time analysis compares changing representation states. To isolate nuisance variation, we now fix a layer and noise level and repeatedly perturb the same image. For $z = \mathrm{Enc}(x)$, let $z_t^{(k)} = \alpha_t z + \sigma_t \epsilon^{(k)}$, with independent $\epsilon^{(k)} \sim \mathcal{N}(0, I)$. Suppressing the fixed layer and timestep, write

$$h^{(k)} = \mu(x) + \xi^{(k)}, \qquad \mu(x) = \mathbb{E}[h^{(k)} \mid x], \qquad \mathbb{E}[\xi^{(k)} \mid x] = 0. \tag{7}$$

Assume the replicates are conditionally independent and identically distributed, with finite second moments, and set $\Omega = \mathbb{E}[\mathrm{Cov}(\xi^{(k)} \mid x)]$. The shared component $\mu(x)$ need not itself be semantic; its predictive relevance is determined by $y = g(x)$. We center features and targets in this controlled setting and omit the intercept.

The observable mean $\bar{h}^{(K)} = K^{-1}\sum_k h^{(k)}$ approximates $\mu(x)$. Define the shared-component oracle risk and its finite-replicate surrogate by

$$\mathcal{R}_\lambda^{(\infty)}(B) = \mathbb{E}\|y - B^\top \mu\|_2^2 + \lambda\|B\|_F^2, \tag{8}$$

$$\mathcal{R}_\lambda^{(K)}(B) = \mathbb{E}\|y - B^\top \bar{h}^{(K)}\|_2^2 + \lambda\|B\|_F^2. \tag{9}$$

For $1 \le r \le \min(p, q)$ and $M > 0$, let $\mathcal{B}_r(M) = \{B : \mathrm{rank}(B) \le r,\ \|B\|_F \le M\}$, and denote global minimizers of the two risks on this set by $B_\infty^\star$ and $B_K^\star$, respectively.

Theorem 3 (Repeated-noise decomposition and oracle-risk recovery)**.** *Under the conditional replicate model, for every B,*

$$\mathcal{R}_\lambda^{(K)}(B) = \mathcal{R}_\lambda^{(\infty)}(B) + \frac{1}{K}\,\mathrm{tr}(B^\top \Omega B). \tag{10}$$

Consequently,

$$0 \le \mathcal{R}_\lambda^{(\infty)}(B_K^\star) - \mathcal{R}_\lambda^{(\infty)}(B_\infty^\star) \le \frac{M^2}{K}\|\Omega\|_{\mathrm{op}}. \tag{11}$$

Equation (10) identifies the effect of repeated averaging: finite replicates add an anisotropic nuisance penalty, which vanishes as K increases. Equation (11) then controls recovery of the oracle objective, without assuming that the oracle coefficients are unique. Increasing K reduces replicate noise; it does not replace independent images, guarantee coefficient convergence, or ensure that the recovered readout is reliable on a single noisy observation.

An observable consequence. For a fixed readout W, evaluated on two independent noise groups of sizes K_A, K_B for the same held-out image,

$$\mathbb{E}\|W^\top \bar{h}_A - W^\top \bar{h}_B\|_2^2 = \left(\frac{1}{K_A} + \frac{1}{K_B}\right) \mathrm{tr}(W^\top \Omega W). \tag{12}$$

The readout must be fixed independently of these evaluation observations. This identity tests the predicted nuisance contribution without observing μ; it does not apply unchanged to two separately refitted readouts.

3.4 Limits of Readout Identification

Recovering predictive risk is weaker than identifying a predictive subspace. We first specify the population object and then state the additional conditions needed to recover it.

The recoverable object. Let $\Sigma = \mathbb{E}[\mu\mu^\top]$, $C = \mathbb{E}[\mu y^\top]$, $A_\lambda = \Sigma + \lambda I \succ 0$, and $D_\lambda = A_\lambda^{-1/2} C$. When the norm bound M is inactive, reduced-rank ridge regression gives

$$B_\infty^\star = A_\lambda^{-1/2}[D_\lambda]_r, \qquad S_\infty^\star = \mathrm{col}(B_\infty^\star), \tag{13}$$

where $[D]_r$ denotes a best rank-r approximation in Frobenius norm. The relevant spectral operator is $D_\lambda D_\lambda^\top = A_\lambda^{-1/2} C C^\top A_\lambda^{-1/2}$; its leading subspace must be mapped back by $A_\lambda^{-1/2}$ to obtain the coefficient subspace. A strict gap $\sigma_r(D_\lambda) > \sigma_{r+1}(D_\lambda)$, with singular values padded by zeros, makes the truncated solution unique. Thus the estimand depends on the target, feature covariance, ridge penalty, and chosen rank.

From risk recovery to subspace recovery. Let $\widehat{B}_{n,K}$ minimize the empirical surrogate risk on $\mathcal{B}_r(M)$ using n independent images, each with K replicates. Assume that, with probability at least $1-\delta$,

$$\sup_{B \in \mathcal{B}_r(M)} \left|\widehat{\mathcal{R}}_\lambda^{(K)}(B) - \mathcal{R}_\lambda^{(K)}(B)\right| \le \Delta_{n,K,\delta}. \tag{14}$$

Assume also that the oracle is locally identifiable: for some $\rho, \kappa > 0$,

$$\mathcal{R}_\lambda^{(\infty)}(B) - \mathcal{R}_\lambda^{(\infty)}(B_\infty^\star) \ge \frac{\kappa}{2}\|B - B_\infty^\star\|_F^2 \quad \text{if } B \in \mathcal{B}_r(M),\ \|B - B_\infty^\star\|_F < \rho. \tag{15}$$

To justify using this local condition, define the separation from competing solutions outside that neighborhood:

$$g_\rho = \inf_{\substack{B \in \mathcal{B}_r(M) \\ \|B - B_\infty^\star\|_F \ge \rho}} \left\{\mathcal{R}_\lambda^{(\infty)}(B) - \mathcal{R}_\lambda^{(\infty)}(B_\infty^\star)\right\} > 0. \tag{16}$$

For a unique oracle on the compact feasible set, this separation is positive whenever the set outside the neighborhood is nonempty.

Theorem 4 (Conditional predictive-subspace recovery)**.** *On the event in Eq.* (14)*, let*

$$\varepsilon_{n,K,\delta} = 2\Delta_{n,K,\delta} + \frac{M^2}{K}\|\Omega\|_{\mathrm{op}}, \qquad e_{n,K,\delta} = \sqrt{2\varepsilon_{n,K,\delta}/\kappa}. \tag{17}$$

If $\varepsilon_{n,K,\delta} < g_\rho$, then $\|\widehat{B}_{n,K} - B_\infty^\star\|_{\mathrm{op}} \le e_{n,K,\delta}$. If, in addition, $B_\infty^\star$ has rank r and $s_r = \sigma_r(B_\infty^\star) > e_{n,K,\delta}$, then $\widehat{S}_{n,K} = \mathrm{col}(\widehat{B}_{n,K})$ has dimension r and

$$\|\sin\angle(\widehat{S}_{n,K}, S_\infty^\star)\|_{\mathrm{op}} \le \frac{e_{n,K,\delta}}{s_r - e_{n,K,\delta}}. \tag{18}$$

Here $\angle(S_1, S_2)$ denotes the diagonal matrix of principal angles.

The guarantee separates three requirements: accurate estimation across images, sufficient noise averaging, and an identifiable rank-r signal. The bound becomes informative only when estimation error is small relative to both the isolation gap and the retained singular value. Risk consistency alone supplies neither condition.

Where identification fails. At pure forward noise ($\alpha_t = 0$), features computed solely from z_t and fixed model inputs are independent of the original image. Hence centered image-specific targets

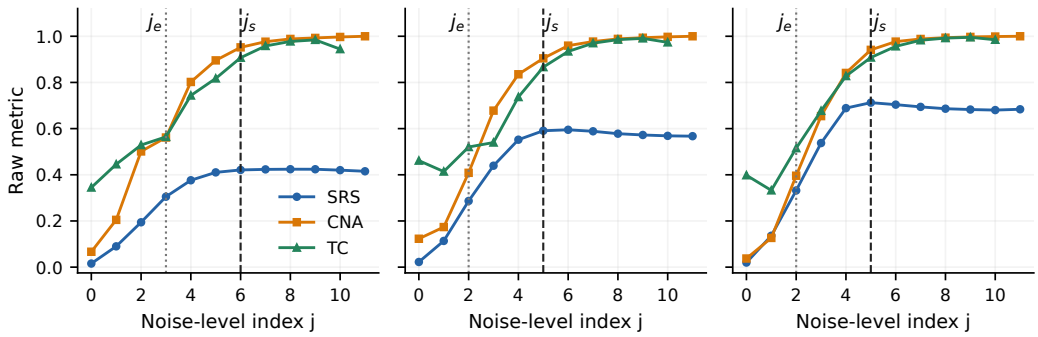


Figure 1: **Readability and geometric reproducibility under controlled noising.** Raw SRS, CNA, and TC at three CelebA-HQ U-Net stages; j increases as noise decreases. Dotted/dashed vertical lines mark readability/joint-stability onsets under the absolute rule in Table 1. TC compares each anchor with the next and is undefined at the last.

satisfy $C = 0$, and no positive-rank predictive subspace is identified. This argument concerns the original image in the controlled experiment; it does not apply to the final generated image, which depends on the initial noise along a reverse trajectory. Separately, a tie at a positive truncation boundary can produce multiple equally optimal subspaces. Neither failure is resolved by increasing K alone.

Implications for reliability measurements. Within fixed coordinates, two estimates close to the same oracle must be close to each other. For adjacent conditions, however, their discrepancy also includes drift between the two oracle subspaces; repeated-noise recovery alone does not guarantee temporal alignment. We therefore report held-out readability, cross-noise alignment, and temporal consistency separately, and evaluate reverse-time reuse through predictions. Under an invertible change $\widetilde{h} = Th$, the transported coefficients $\widetilde{B} = T^{-\top}B$ preserve predictions while Euclidean coefficient angles can change. This is a statement about the fitted predictive function; refitting with an unchanged isotropic ridge penalty need not be invariant. The recovered object is thus an evaluator-defined predictive subspace in a specified representation system, rather than an intrinsic semantic basis.

4 Experimental Protocol

Measurements and Experimental settings. For q-dimensional targets and rank-r orthonormal readout bases, we measure

$$\mathrm{SRS} = \frac{1}{q}\sum_{a=1}^{q}\left(1 - \frac{\sum_i (y_{ia} - \widehat{y}_{ia})^2}{\sum_i (y_{ia} - \bar{y}_a)^2}\right), \quad \mathrm{CNA} = \frac{\|U_A^\top U_B\|_*}{r}, \quad \mathrm{TC}(t,t') = \frac{\|U_t^\top U_{t'}\|_*}{r}. \tag{19}$$

Here $\bar{y}_a$ is the held-out target mean, with positive target variance, and $\|\cdot\|_*$ is the nuclear norm. CNA compares bases fitted from disjoint noise groups; TC compares adjacent noise levels in fixed feature coordinates. Predictive reuse is evaluated separately by transferring a source readout without refitting. We study an unconditional CelebA-HQ latent diffusion model with 3DDFA-V2 pose targets and an unconditional CIFAR-10 pixel-space DDPM(Ho et al., 2020) with class targets.Controlled audits use eight noise realizations per image and noise level, split into two groups of four, over twelve CelebA-HQ and fourteen CIFAR-10 noise levels. Reverse evaluation uses deterministic 100-step DDIM trajectories, fixed final-sample pose targets, and twelve anchors. We compare state-specific, unchanged-source, and shared-suffix readouts, including a pooled-latent baseline.

5 Experiments

We evaluate cross-noise reproducibility, cross-time reuse, and reliability's added information beyond probe accuracy, using controlled noising with fixed images and targets followed by reverse

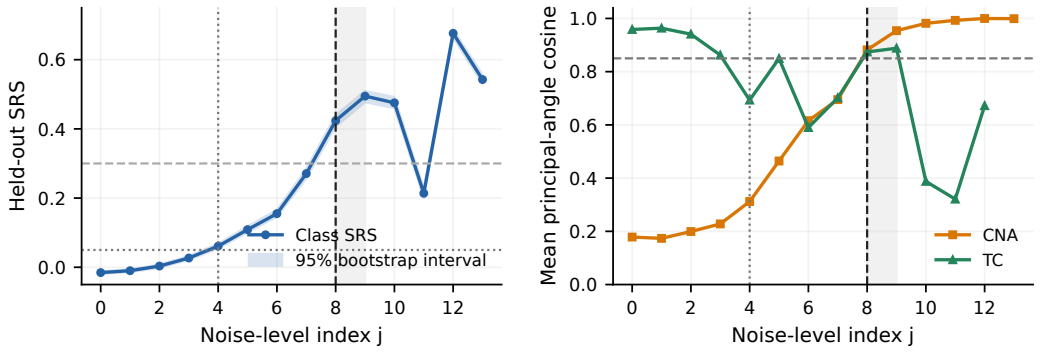


Figure 2: **CIFAR-10 early-up replication.** Left: raw held-out class SRS with 95% intervals from 200 stratified test-image bootstrap resamples. Right: CNA and TC point estimates. Vertical lines mark $j_e = 4$ and $j_s = 8$ and shading marks the first jointly stable two-anchor interval.

Table 1: **CelebA-HQ timing robustness.** Onsets are noise-grid indices $j_e \to j_s$, using thresholds $(0.50; 0.80, 0.90, 0.90)$ (normalized) and $(0.20; 0.30, 0.85, 0.85)$ (raw).

Layer	Normalized	Raw	Bootstrap gap	Norm. sweep	Raw sweep
late-down	$3 \to 7$	$3 \to 6$	3 [3, 3]	144/144	72/72
mid	$3 \to 6$	$2 \to 5$	4 [3, 4]	144/144	108/108
early-up	$3 \to 6$	$2 \to 5$	3 [3, 3]	144/144	144/144

trajectories. Semantic targets provide a common measurement reference for the studied attributes in unconditional diffusion. Implementation details appear in Appendix B.

5.1 Readable is not reproducible

Separation in the measured curves. We first ask whether accurate pose prediction coincides with reproducible readout geometry. Figure 1 shows that these properties develop at different noise levels. In early-up, SRS is already 0.332 at $t = 817$, while CNA and TC are only 0.396 and 0.515. At $t = 545$, the same quantities reach 0.712, 0.941, and 0.908, respectively. Thus, appreciable predictive information is present before the fitted subspace becomes reproducible across noise realizations and adjacent noise levels.

Sensitivity to onset criteria. Table 1 summarizes the separation using persistent threshold crossings. For the three displayed layers, onset gaps remain positive across all 500 bootstrap refits and all resolved threshold configurations. These gaps are threshold-dependent summaries of the curves, with joint stability requiring a higher SRS cutoff than readability.

5.2 Controls and the effect of repeated perturbations

Target correspondence and geometric baselines. Permuting pose labels makes SRS negative at all 60 layer–timestep pairs, compared with positive SRS at 58/60 pairs under the original labels. Learned readouts achieve mean TC 0.738, versus 0.073 for independently sampled, rank- and dimension-matched random subspaces; 54/55 adjacent comparisons exceed the pointwise random 95th percentile (Table 2). These controls support target-specific readability and alignment beyond the matched random baseline.

Finite-replicate behavior. Increasing the number of noise realizations from $K = 1$ to 8 raises mean SRS from 0.3291 to 0.3872. Over $K = 2$ to 8, mean CNA increases from 0.6755 to 0.7443 and mean TC from 0.6994 to 0.7383. This behavior is consistent with reducing the nuisance contribution in Eq. (10).

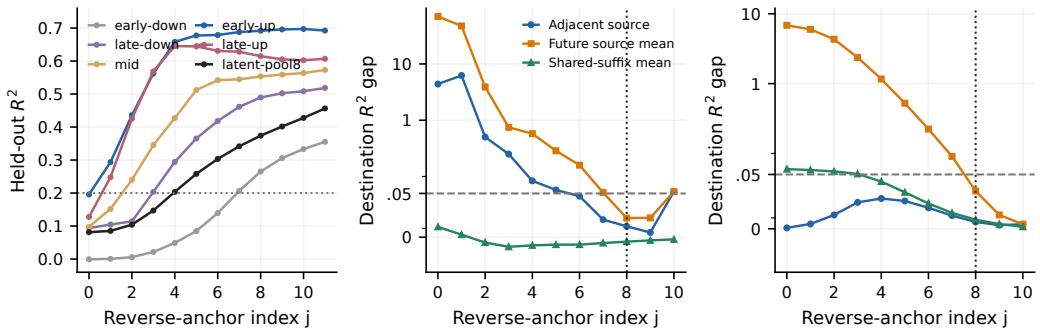


Figure 3: **Independent reverse-trajectory confirmation.** (a) Same-state readability for six representations. (b–c) Adjacent-source, mean future-source, and mean shared-suffix gaps for early-up and pooled latents. Horizontal/vertical lines mark the 0.05 threshold/joint reuse onset at $j = 8$ ($t = 272$).

Table 2: **Controls and replicate ablation.** Random-subspace exceedances are pointwise comparisons. Reverse controls use the independent confirmation data.

Test	Result
Pose-label permutation	Mean SRS $0.3872 \to -0.0048$; positive SRS $58/60 \to 0/60$.
Random subspaces	Mean TC: learned 0.738, random 0.073; $54/55$ exceed the random 95th percentile.
Replicates, $K = 1, 2, 4, 8$	Mean SRS: 0.3291, 0.3505, 0.3700, 0.3872.
Replicates, $K = 2, 4, 8$	Mean CNA: 0.6755, 0.7075, 0.7443; mean TC: 0.6994, 0.7228, 0.7383.
Reverse-label permutation	Maximum held-out $R^2 = -0.0018$.
Destination-trajectory permutation	Maximum held-out $R^2 = -0.0124$.

5.3 Replication on CIFAR-10

We extend the audit to an unconditional CIFAR-10 pixel-space DDPM with class one-hot targets. In early-up, from $j = 9$ to $j = 10$, SRS remains similar ($0.495 \to 0.475$) and CNA increases ($0.954 \to 0.982$), while TC drops sharply ($0.888 \to 0.388$; Figure 2). This threshold-independent contrast shows that readability and strong cross-noise alignment can coexist with weak temporal alignment. Under the nominal rule, only early-up resolves joint stability: readability at $j = 4$ ($t = 3727$) precedes stability at $j = 8$ ($t = 2532$). All threshold configurations with both onsets resolved preserve this ordering. Table 3 also retains unresolved cases. These results extend the readability–reliability distinction to class semantics in pixel-space diffusion.

5.4 Readable is not reusable along the reverse trajectory

Evaluating reuse during generation. We evaluate 5,000 independent deterministic 100-step DDIM trajectories ($4{,}000/500/500$ train/validation/test), with final-sample pose estimates fixed across twelve anchors. Using one training-fitted normalization per representation, we compare state-specific, frozen-source, and shared-suffix readouts, including an 8×8 pooled-latent baseline. Joint reuse requires same-state $R^2 \geq 0.20$ and adjacent-source, mean future-source, and mean shared-suffix gaps ≤ 0.05 for two consecutive anchors. Future-source means include only later readable destinations.

Readable targets do not ensure reusable readouts. All six representations are readable, but only early-up and pooled latents resolve joint reuse (Table 4). Readability begins at $t = 908$ and $t = 636$, respectively; both reach joint reuse at $t = 272$, yielding a seven-anchor gap for early-up (Figure 3). Training-trajectory bootstrap refits preserve this ordering whenever both onsets resolve: $463/500$ for early-up and $500/500$ for pooled latents. Thus, local readability does not ensure predictive reuse during generation.

Table 3: **CIFAR-10 threshold audit.** Onsets are reported as j/t. "Resolved" counts configurations where both onsets are detected. A dash denotes an unresolved nominal joint-stability onset.

Layer	Best SRS	Read onset	Stable onset	Resolved/144	Ordered/resolved
early-down	.221	$5/3534$	–	$0/144$	–
late-down	.430	$4/3727$	–	$24/144$	$24/24$
mid	.425	$4/3727$	–	$24/144$	$24/24$
early-up	.676	$4/3727$	$8/2532$	$64/144$	$64/64$
late-up	.423	$5/3534$	–	$0/144$	–

Table 4: **Reverse-trajectory onsets and refit sensitivity.** Timesteps decrease during generation; dashes denote unresolved point-estimate joint reuse onsets.

Representation	Terminal R^2	Read onset	Joint reuse onset	Both resolved/500
early-down	.355	363	–	$0/500$
late-down	.518	727	–	$29/500$
mid	.573	817	–	$58/500$
early-up	.692	908	272	$463/500$
late-up	.607	908	–	$0/500$
latent-pool8	.456	636	272	$500/500$

Joint fitting and source transfer. The shared early-up readout meets its mean-gap criterion from $t = 908$, while unchanged-source transfer remains unreliable. Thus, a predictor fitted across multiple states can perform well even when a predictor fitted at one state transfers poorly, consistent with Eq. (4). Appendix B.6 reports the empirical risk and decoder-drift decompositions, including the residual cross term needed to relate predictor mismatch to held-out transfer error.

5.5 Reliability beyond probe accuracy

We examine whether reliability measurements inform representation selection over 55 non-endpoint CelebA-HQ candidates. Probe-only selection maximizes validation SRS; reliability-aware policies predict single-noise performance from SRS+CNA or SRS+CNA+TC using leave-one-layer-out models. Adding CNA selects a candidate with higher single-noise R^2 (0.6077 versus 0.5879) and lower cross-noise disagreement (0.0319 versus 0.0590; Appendix Table 14). Adding TC further improves prediction correlations at a small cost in single-noise accuracy. Thus, maximizing validation SRS alone does not identify the most reproducible candidate.

A complementary regression over all 60 candidates confirms CNA's incremental predictive value: adding CNA to validation SRS improves leave-one-layer-out R^2 for all all five cross-noise outcomes (Appendix Table 15).

6 Limitations

The empirical conclusions concern the specified evaluators and linear or reduced-rank ridge readouts in unconditional models. Nonlinear readouts and conditional generation may have different reliability boundaries. The mathematical conditions are stated explicitly. The experiments measure observable reliability rather than certify every population assumption.

7 Conclusion

We distinguish semantic readability, cross-noise reproducibility, and cross-time predictive reuse, linking each to explicit tests. Controlled-noise experiments, CIFAR-10 replication, and independent reverse trajectories show that readable representations need not support reproducible or reusable readouts. Our central conclusion is that *readability is not reliability*: claims of persistence require direct evaluation across noise realizations and generative states, beyond same-state probe accuracy.

REPRODUCIBILITY STATEMENT

The main text states the theoretical assumptions and experimental design; proofs appear in Appendix A. Data splits, checkpoints, feature hooks, random seeds, and readout fitting are documented in Appendices B.1–B.2. Additional results and analysis protocols, including threshold sweeps, bootstrap procedures, controls, and unresolved outcomes, are provided in Appendices B.3–B.7. Appendix B.8 records computational environments and artifact provenance.

ETHICS STATEMENT

The study analyzes pretrained unconditional models and established datasets using frozen evaluators. Pose and other evaluator outputs are used as measurement targets, not as judgments of personal attributes. Dataset and checkpoint usage remains subject to the respective licenses and privacy restrictions.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and LaTeX formatting to improve the readability of this manuscript, as well as interactive review of mathematical exposition, experimental design, and result presentation. The authors independently verified all mathematical statements and proofs, implementation choices, data processing procedures, and reported numerical results, and take full responsibility for the final manuscript.

REFERENCES

Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *International Conference on Learning Representations*, 2022.

David Degras, Thomas Chapalain, and Bertrand Thirion. On the stability of reduced-rank ridge regression. In *European Signal Processing Conference*, 2024.

Emily L. Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun, and Rob Fergus. Exploiting linear structure within convolutional networks for efficient evaluation. In *Advances in Neural Information Processing Systems*, 2014.

Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2426–2436, 2023.

Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. In *European Conference on Computer Vision*, pp. 172–188. Springer, 2024.

Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion. In *Advances in Neural Information Processing Systems*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Xinting Huang and Michael Hahn. Decomposing representation space into interpretable subspaces with unsupervised learning. In *International Conference on Learning Representations*, volume 2026, pp. 51314–51360, 2026.

Mozes Jacobs, Thomas Fel, Richard Hakim, Alessandra Brondetta, Demba Ba, and T Anderson Keller. Block recurrent dynamics in vision transformers. In *International Conference on Learning Representations*, volume 2026, pp. 51617–51642, 2026.

Zahra Kadkhodaie, Stéphane Mallat, and Eero P. Simoncelli. Unconditional CNN denoisers contain sparse semantic representation of images. *arXiv preprint arXiv:2506.01912*, 2025.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9492–9502. IEEE, 2024.

Dahye Kim, Xavier Thomas, and Deepti Ghadiyaram. Revelio: Interpreting and leveraging semantic information in diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4659–4669, 2025.

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36:47500–47510, 2023.

Ashin Mukherjee and Ji Zhu. Reduced rank ridge regression and its kernel extensions. *Statistical Analysis and Data Mining*, 4(6):612–622, 2011.

Cristian D Paduraru, Antonio Barbalau, Radu Filipescu, Andrei Nicolicioiu, and Elena Burceanu. Bridging explainability and embeddings: Bee aware of spuriousness. In *International Conference on Learning Representations*, volume 2026, pp. 27954–27984, 2026.

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

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

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.

Viacheslav Surkov, Chris Wendler, Antonio Mari, Mikhail Terekhov, Justin Deschenaux, Robert West, Caglar Gulcehre, and David Bau. One-step is enough: Sparse autoencoders for text-to-image diffusion models. In *Advances in Neural Information Processing Systems*, 2025.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in neural information processing systems*, 36: 1363–1389, 2023a.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Türe. What the daam: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5644–5659, 2023b.

Berk Tinaz, Zalan Fabian, and Mahdi Soltanolkotabi. Emergence and evolution of interpretable concepts in diffusion models. In *Advances in Neural Information Processing Systems*, 2025.

Nina Weng, Aasa Feragen, and Siavash Bigdeli. Patronus: Interpretable diffusion models with prototypes. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 46923–46953, 2026.

Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Denoising diffusion autoencoders are unified self-supervised learners. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Xingyi Yang and Xinchao Wang. Diffusion model as representation learner. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18892–18903. IEEE, 2023.

Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *IEEE/CVF International Conference on Computer Vision*, 2023.

A Proofs and Additional Theory

A.1 Readability–reliability decomposition

The destination objective expands as

$$\mathcal{J}_d(\Theta) = \mathbb{E}\|y\|_2^2 - 2\,\mathrm{tr}(\Theta^\top C_d) + \mathrm{tr}(\Theta^\top A_d \Theta). \tag{20}$$

Since $A_d\Theta_d^\star = C_d$, substituting $\Theta = \Theta_d^\star + D$ cancels the linear terms and gives

$$\mathcal{J}_d(\Theta_d^\star + D) - \mathcal{J}_d(\Theta_d^\star) = \mathrm{tr}(D^\top A_d D). \tag{21}$$

Subtracting the change in the ridge penalty yields

$$\begin{aligned}\mathcal{L}_d(\Theta_d^\star + D) - \mathcal{L}_d(\Theta_d^\star) &= \mathrm{tr}(D^\top A_d D) - \lambda\,\mathrm{tr}(D^\top P D) - 2\lambda\,\mathrm{tr}(D^\top P \Theta_d^\star) \\ &= \mathrm{tr}(D^\top G_d D) - 2\lambda\,\mathrm{tr}(D^\top P \Theta_d^\star).\end{aligned} \tag{22}$$

Taking $D = D_{s,d}$ proves Theorem 1. In the scalar example, the two unregularized slopes are 1 and $1/a$, whereas the transferred slope remains 1. Its prediction error is $(a-1)^2\mathbb{E}y^2 = (a-1)^2$, giving the stated destination R^2.

A.2 SHARED PREDICTORS AND TRAJECTORY-LOCAL REUSE

Differentiating the weighted objective gives $A_{\mathcal{T}}\Theta^\star_{\mathcal{T}} = \sum_t \pi_t C_t$. Substituting $C_t = A_t\Theta^\star_t$ proves Eq. (4). Statewise application of Theorem 1 gives

$$\sum_t \pi_t \big[\mathcal{J}_t(\Theta) - \mathcal{J}_t(\Theta^\star_t)\big] = \sum_t \pi_t \|A_t^{1/2}(\Theta - \Theta^\star_t)\|_F^2. \tag{23}$$

The left side is minimized by $\Theta^\star_{\mathcal{T}}$, so its value there cannot exceed its value at any source-optimal predictor.

For drift, add and subtract $A_s^{-1}C_d$ and use $A_s^{-1} - A_d^{-1} = A_s^{-1}(A_d - A_s)A_d^{-1}$. The assumptions of Corollary 2 imply

$$\begin{aligned}\|D_{s,d}\|_F &\le \|A_s^{-1}\|_{\mathrm{op}}\|C_s - C_d\|_F + \|A_s^{-1}\|_{\mathrm{op}}\|A_d - A_s\|_{\mathrm{op}}\|A_d^{-1}\|_{\mathrm{op}}\|C_d\|_F \\ &\le \left(\frac{L_C}{a_-} + \frac{L_A c_0}{a_-^2}\right)|\tau_s - \tau_d|. \end{aligned}\tag{24}$$

Then $\|A_d^{1/2}D_{s,d}\|_F^2 \le a_+\|D_{s,d}\|_F^2$ proves Eq. (6). For unpenalized prediction risk under ridge, the available bound is instead

$$\big|\mathcal{L}_d(\Theta^\star_s) - \mathcal{L}_d(\Theta^\star_d)\big| \le a_+\gamma^2|\tau_s - \tau_d|^2 + 2\lambda\|P\Theta^\star_d\|_F\gamma|\tau_s - \tau_d|. \tag{25}$$

Thus the assumptions alone do not establish a purely quadratic held-out prediction-risk gap when $\lambda > 0$.

A.3 REPEATED-NOISE RISK AND SPLIT DISAGREEMENT

Let $\bar\xi^{(K)} = K^{-1}\sum_k \xi^{(k)}$. Conditional centering and independence imply

$$\mathbb{E}[\bar\xi^{(K)} \mid x] = 0, \qquad \mathrm{Cov}(\bar\xi^{(K)} \mid x) = \mathrm{Cov}(\xi^{(1)} \mid x)/K. \tag{26}$$

Expanding $\|y - B^\top\mu - B^\top\bar\xi^{(K)}\|_2^2$, the cross term has zero expectation because $y = g(x)$ and μ are measurable with respect to x. Taking expectations proves Eq. (10). Optimality of $B^\star_K$ and nonnegativity of the nuisance penalty give

$$\begin{aligned}\mathcal{R}_\lambda^{(\infty)}(B^\star_K) &\le \mathcal{R}_\lambda^{(K)}(B^\star_K) \le \mathcal{R}_\lambda^{(K)}(B^\star_\infty) \\ &= \mathcal{R}_\lambda^{(\infty)}(B^\star_\infty) + K^{-1}\,\mathrm{tr}((B^\star_\infty)^\top\Omega B^\star_\infty). \end{aligned}\tag{27}$$

The last term is at most $M^2\|\Omega\|_{\mathrm{op}}/K$, proving Eq. (11). This proof neither uses nor implies uniqueness of a minimizer.

For the split law, $\bar h_A - \bar h_B = \bar\xi_A - \bar\xi_B$ has conditional mean zero and conditional covariance $(K_A^{-1} + K_B^{-1})\,\mathrm{Cov}(\xi^{(1)} \mid x)$. Taking the quadratic expectation after applying $W^\top$ proves Eq. (12). A readout trained on an independent sample can be treated as fixed by conditioning on that training sample. Equal groups of size K give a $2/K$ coefficient; splitting a total budget of K equally gives $4/K$.

A.4 THE REDUCED-RANK ORACLE AND ITS SPECTRAL OPERATOR

Completing the square gives

$$\mathcal{R}_\lambda^{(\infty)}(B) = \|A_\lambda^{1/2}B - D_\lambda\|_F^2 + \mathbb{E}\|y\|_2^2 - \|D_\lambda\|_F^2. \tag{28}$$

Because $A_\lambda^{1/2}$ is invertible, the substitution $H = A_\lambda^{1/2}B$ preserves rank. If the norm constraint is inactive, the best rank-r approximation theorem yields $H = [D_\lambda]_r$, proving Eq. (13). If the rank-r truncation is unique, so is this coefficient matrix. The left singular vectors of D_λ diagonalize $A_\lambda^{-1/2}CC^\top A_\lambda^{-1/2}$, but the coefficient subspace is their image under $A_\lambda^{-1/2}$.

An additional factor Σ_{yy}^{-1} would describe target-whitened associations, as in canonical correlation analysis, and does not follow from the ordinary squared-loss objective used here. Likewise, spectral separation of D_λ and a lower bound on $\sigma_r(B^\star_\infty)$ are distinct conditions: the former selects a unique truncation, while the latter controls sensitivity of the coefficient column space.

A.5 Interpreting the concentration assumption

For population-centered features and targets, define

$$\widehat{G}_K = \frac{1}{n}\sum_i \bar{h}_i^{(K)}\bar{h}_i^{(K)\top}, \quad \widehat{C}_K = \frac{1}{n}\sum_i \bar{h}_i^{(K)} y_i^\top, \quad \widehat{v} = \frac{1}{n}\sum_i \|y_i\|_2^2. \tag{29}$$

Let $G_K = \Sigma + \Omega/K$ and $v = \mathbb{E}\|y\|_2^2$. The population feature–target cross-moment is C for every K. Direct expansion and the Frobenius/operator norm inequalities show that

$$\sup_{B\in\mathcal{B}_r(M)} |\widehat{\mathcal{R}}_\lambda^{(K)}(B) - \mathcal{R}_\lambda^{(K)}(B)| \le |\widehat{v} - v| + 2M\|\widehat{C}_K - C\|_F + M^2\|\widehat{G}_K - G_K\|_{\mathrm{op}}. \tag{30}$$

Thus Eq. (14) can be established from two-sided moment concentration under appropriate distributional assumptions. Its rate is not specified without such assumptions. The independent sampling units are images; the nK feature vectors share image-specific components and are not nK independent image observations. If means are estimated from training data, the corresponding mean-estimation terms must also be controlled. Centering and all other preprocessing are frozen before evaluation.

A.6 Localization and predictive-subspace recovery

On the concentration event, empirical optimality implies

$$\begin{aligned}
\mathcal{R}_\lambda^{(\infty)}(\widehat{B}_{n,K}) &\le \mathcal{R}_\lambda^{(K)}(\widehat{B}_{n,K}) \\
&\le \widehat{\mathcal{R}}_\lambda^{(K)}(\widehat{B}_{n,K}) + \Delta_{n,K,\delta} \\
&\le \widehat{\mathcal{R}}_\lambda^{(K)}(B_\infty^\star) + \Delta_{n,K,\delta} \\
&\le \mathcal{R}_\lambda^{(K)}(B_\infty^\star) + 2\Delta_{n,K,\delta} \\
&\le \mathcal{R}_\lambda^{(\infty)}(B_\infty^\star) + \varepsilon_{n,K,\delta}.
\end{aligned} \tag{31}$$

If $\varepsilon_{n,K,\delta} < g_\rho$, the definition of g_ρ rules out $\|\widehat{B}_{n,K} - B_\infty^\star\|_F \ge \rho$. Local quadratic growth therefore applies and gives

$$\|\widehat{B}_{n,K} - B_\infty^\star\|_{\mathrm{op}} \le \|\widehat{B}_{n,K} - B_\infty^\star\|_F \le \sqrt{2\varepsilon_{n,K,\delta}/\kappa} = e_{n,K,\delta}. \tag{32}$$

Continuity of the objective and compactness of $\mathcal{B}_r(M)$ ensure $g_\rho > 0$ whenever $B_\infty^\star$ is the unique global minimizer and the outer set is nonempty. If it is empty, use $g_\rho = +\infty$.

Write $e = e_{n,K,\delta}$ and $s_r = \sigma_r(B_\infty^\star)$. Singular-value perturbation gives $\sigma_r(\widehat{B}_{n,K}) \ge s_r - e > 0$. Since the estimator has rank at most r, it consequently has rank exactly r. Let $\widehat{B}_{n,K} = \widehat{U}\widehat{\Lambda}\widehat{V}^\top$ be its thin singular value decomposition and $\Pi_\star$ the orthogonal projector onto $S_\infty^\star$. Since $(I-\Pi_\star)B_\infty^\star = 0$,

$$(I - \Pi_\star)\widehat{U} = (I - \Pi_\star)(\widehat{B}_{n,K} - B_\infty^\star)\widehat{V}\widehat{\Lambda}^{-1}. \tag{33}$$

Taking operator norms bounds the largest sine of the principal angles by $e/(s_r - e)$, proving Theorem 4. This proof makes explicit why a local growth assumption cannot be used before localization has been established.

A.7 Geometric reproducibility and failure boundaries

For equal-dimensional subspaces define $d(S_1, S_2) = \|\Pi_{S_1} - \Pi_{S_2}\|_{\mathrm{op}}$, which equals the largest sine of their principal angles. If two noise-group estimates obey $d(\widehat{S}_A, S_\infty^\star) \le \eta_A$ and $d(\widehat{S}_B, S_\infty^\star) \le \eta_B$, the triangle inequality gives

$$d(\widehat{S}_A, \widehat{S}_B) \le \eta_A + \eta_B. \tag{34}$$

For adjacent conditions with potentially different oracles,

$$d(\widehat{S}_t, \widehat{S}_{t'}) \le \eta_t + d(S_t^\star, S_{t'}^\star) + \eta_{t'}. \tag{35}$$

Consequently, reducing estimation errors alone does not control between-condition oracle drift. If $\eta_A + \eta_B < 1$, every cross-noise principal-angle cosine is at least $\sqrt{1-(\eta_A+\eta_B)^2}$, providing a lower bound on their mean alignment.

At the pure controlled forward-noise endpoint, $h = f(\sigma_t\epsilon, t)$ is independent of x when all other model inputs are fixed and independent of x. After centering, $\mu = 0$ and $C = 0$. With $\lambda > 0$, the oracle is the zero matrix; with $\lambda = 0$, coefficient matrices cannot be identified by predictions on the zero shared component. Neither case identifies a positive-rank image-specific predictive subspace. Along reverse generation, however, the final sample is a function of the initial noise and this independence premise does not hold.

B Additional Experimental Details and Results

This appendix documents the protocols behind the controlled CelebA-HQ experiments, CIFAR-10 replication, reverse-trajectory confirmation, and representation-selection analysis. Full-layer results complement the focused main-text figures; source-transfer and shared-readout results are reported separately.

B.1 Models, data, and feature extraction

Controlled CelebA-HQ. We use the unconditional CelebA-HQ latent diffusion model at 256×256 resolution. The checkpoint is `ldm/celeba256/model.ckpt`. Images are encoded once, and independent Gaussian perturbations are applied to the cached latent at each evaluated noise level. The nominal image split is $24{,}000/3{,}000/3{,}000$ for training/validation/test; excluding pose-invalid images leaves $23{,}986/3{,}000/2{,}999$ valid images. Frozen 3DDFA-V2 estimates of yaw, pitch, and roll provide the three targets. They describe the original image throughout the controlled-noise experiment. The primary cache contains eight draws per image and noise level, partitioned into two disjoint groups of four. The main SRS uses the eight-draw mean; CNA compares group-specific fits.

CIFAR-10. We use an unconditional pixel-space DDPM on CIFAR-10. The checkpoint is `cifar10_uncond_50M_500K.pt`, a 32×32 improved-diffusion model with $4{,}000$ cosine-schedule steps, 128 base channels, three residual blocks per level, learned variance, and dropout 0.3. Class-balanced splits contain $5{,}000/1{,}000/2{,}000$ images, with 500/100/200 images per class. The target is the ten-dimensional one-hot class vector. Four draws are stored in each of two disjoint noise groups: eight draws in total. Pooled readability uses their combined mean, whereas CNA compares the two four-draw groups.

Generation and anchors. The reverse experiment uses the same unconditional face checkpoint, deterministic DDIM sampling with an anchored-uniform 100-step schedule and $\eta = 0$, and $5{,}000$ valid confirmation trajectories split $4{,}000/500/500$. Each trajectory retains one final-image pose target across all anchors. Confirmation initial-noise seeds are disjoint from the development trajectories used to establish the analysis protocol. In all reported curves, j increases as the diffusion timestep decreases. The controlled face and reverse analyses use

$$\mathcal{T}_{\text{face}} = (999, 908, 817, 727, 636, 545, 454, 363, 272, 182, 91, 0).$$

The CIFAR-10 grid is

$$\mathcal{T}_{\text{CIFAR}} = (3999, 3968, 3928, 3852, 3727, 3534, 3269, 2932, 2532, 2078, 1584, 1021, 71, 0).$$

The reverse anchors occur at sampler-step indices $0, 9, 18, 27, 36, 45, 54, 63, 72, 81, 90, 99$. Adjacent anchors refer to this evaluation grid, not consecutive sampler steps. Controlled noising holds the original image fixed; reverse evaluation holds the final generated image's target fixed. These are different experimental estimands.

Randomness and evaluator. The main controlled cache uses experiment seed 1234 and noise base seed 2026. CIFAR-10 uses split seed 1234 and noise seed 2026. The reverse full-run master seed is 940260915, with validation/test offsets 10^8 and 2×10^8; invalid pose samples are replaced by subsequent seeds under the generation protocol. Reverse bootstrap, training-label permutation, and destination permutation use seeds 17031, 17041, and 17051. The pose evaluator uses the 3DDFA-V2 `mb1_120x120.yml` configuration and reports angles in degrees.

Table 5: **Feature locations and dimensions after global average pooling.** Face-model paths are relative to `model.diffusion_model`; CIFAR-10 paths are relative to its U-Net.

Stage	Face-model module	p	CIFAR-10 module	p
early-down	`input_blocks.1`	224	`input_blocks.1`	128
late-down	`input_blocks.7`	672	`input_blocks.15`	256
mid	`middle_block`	896	`middle_block`	256
early-up	`output_blocks.3`	672	`output_blocks.5`	256
late-up	`output_blocks.8`	448	`output_blocks.15`	128

Features. Table 5 specifies the five U-Net locations. Global average pooling removes spatial dimensions without concatenating layers. Reverse evaluation additionally average-pools the noisy latent to 8×8 and flattens its three channels, giving 192 features. Features and targets are cached before readout fitting so that comparisons reuse the same image or trajectory identities.

B.2 Readout fitting and measurement definitions

Controlled fits. For CelebA-HQ, ridge and requested rank are selected by mean validation R^2 from $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ and $r \in \{1, 2, 4, 8, 16\}$. Effective rank is capped by the three pose outputs. The archived implementation standardizes features, fits ridge, and truncates its coefficient matrix by ordinary SVD; the selected estimator is then refitted on training and validation images. This is an approximation to objective-aware reduced-rank ridge when truncation is active. The reported CIFAR-10 branch uses $\lambda = 1000$ and effective rank nine, the maximum nonredundant rank of centered ten-class targets. Its recorded ridge search spans 10^{-4} through 10^{3} in powers of ten, with feature standardization and training-plus-validation refitting. The accompanying per-site table records the selected rank and ridge values for all 60 CelebA-HQ sites.

Scoring and geometry. SRS is the arithmetic mean of targetwise held-out coefficients of determination, rather than a pooled sum-of-squares ratio:

$$\mathrm{SRS} = \frac{1}{q} \sum_{a=1}^{q} \left(1 - \frac{\sum_i (y_{ia} - \widehat{y}_{ia})^2}{\sum_i (y_{ia} - \bar{y}_a)^2} \right). \tag{36}$$

Every target variance in this expression must be positive. Before computing geometric metrics, fitted coefficient directions are mapped back through the feature scaler and orthonormalized in the original feature coordinates. CNA compares independently fitted noise-group bases at one site. TC compares bases at adjacent noise levels within a layer.

B.3 Controlled-noise results, controls, and decision rules

Figure 4 gives all five layers, including the two omitted from the focused main-text figure. Figure 5 reports the correspondence and random-geometry controls alongside the replicate ablation.

Correspondence and random geometry. The label-permutation control refits readouts after shuffling pose labels (seed 91001). All 60 SRS values become negative, with mean -0.0048, compared with 58 positive sites and mean 0.3872 under the original labels. The random-subspace control draws 100 rank- and dimension-matched random bases per site. Mean adjacent TC is 0.073 for random subspaces and 0.738 for learned readouts; 54 of 55 learned comparisons exceed their pointwise random 95th percentile. This percentile is not a familywise significance threshold.

Replicate and rank sensitivity. For $K = 1, 2, 4, 8$, mean SRS is $0.3291, 0.3505, 0.3700, 0.3872$. For $K = 2, 4, 8$, mean CNA is $0.6755, 0.7075, 0.7443$ and mean TC is $0.6994, 0.7228, 0.7383$. Table 6 reports the fixed-rank ablation. Higher geometric overlap at rank one accompanies substantially lower SRS, illustrating why geometry must be interpreted with predictive performance.

Exact archived onset rule. For the face timing tables, let s_j, c_j, u_j denote either the raw SRS/CNA/TC curves or their separate within-layer min–max normalizations. Set $m_j = (s_j +$

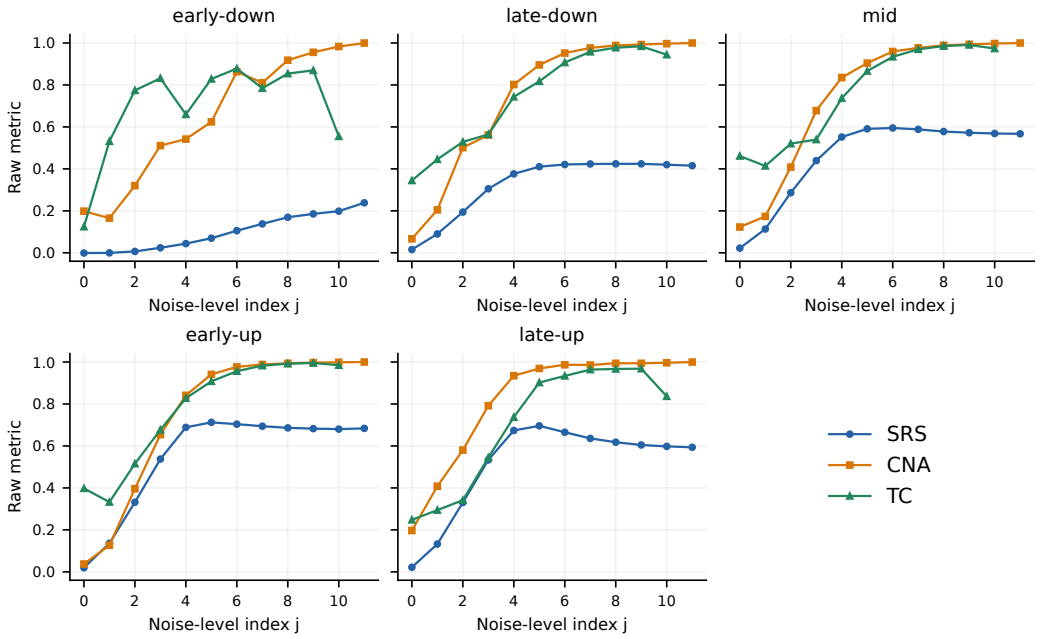


Figure 4: **Complete controlled CelebA-HQ results.** Raw SRS, CNA, and TC for all five U-Net stages. Each panel uses the same noise grid and vertical scale; final-anchor TC is undefined.

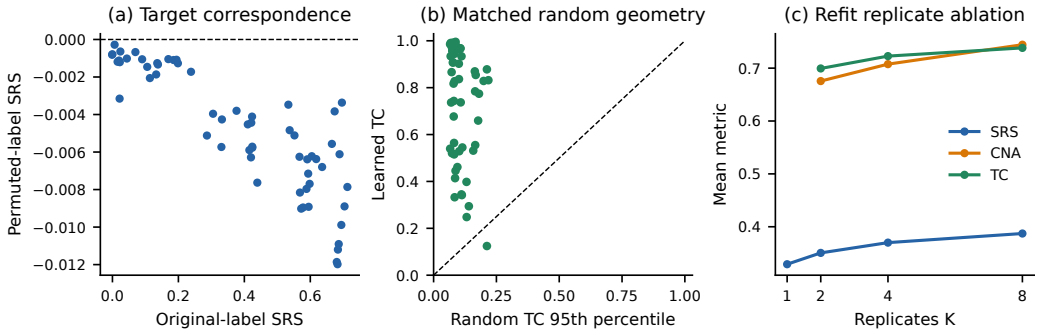


Figure 5: **Controlled-noise controls and replicate sensitivity.** (a) Original versus permuted-label SRS at 60 sites. (b) Learned adjacent TC versus its matched random 95th percentile; the dashed line is equality. (c) Mean refitted metrics as the replicate count increases.

$s_{j+1})/2$ when both entries are finite. Readability is the first j with

$$s_j \geq \theta_e, \qquad m_{j+1} \geq m_j - 10^{-8}, \qquad m_{j+2} \geq m_{j+1} - 10^{-8}, \tag{37}$$

requiring all three moving averages to be defined. Joint stability is the first j for which $s_k \geq \theta_s$, $c_k \geq \theta_c$, and $u_k \geq \theta_t$ at both $k = j, j+1$. Thus the archived face readability rule includes a local trend condition; it is not simply a two-anchor threshold crossing. The nominal threshold tuples $(\theta_e; \theta_s, \theta_c, \theta_t)$ are $(0.50; 0.80, 0.90, 0.90)$ for normalized metrics and $(0.20; 0.30, 0.85, 0.85)$ for raw metrics.

The normalized sweep varies $\theta_e \in \{.4, .5, .6, .7\}$, $\theta_s \in \{.7, .8, .9, .95\}$, and $\theta_c, \theta_t \in \{.8, .85, .9\}$. The raw sweep varies $\theta_e \in \{.20, .25, .30, .35\}$, $\theta_s \in \{.30, .40, .50, .60\}$, and $\theta_c, \theta_t \in \{.85, .90, .95\}$. Both contain 144 configurations per layer. Table 7 includes resolved, tied, and unresolved outcomes. In normalized early-down, 18 of the 36 resolved configurations are tied; the strict-ordering result for the three main-text focus layers does not extend to every resolved configuration of every layer.

Table 6: **Fixed-rank sensitivity.** Means cover 60 sites for SRS/CNA and 55 adjacent pairs for TC. Pose has three outputs, so requested ranks above three are capped.

Requested rank	Effective rank	SRS	CNA	TC
1	1	0.1923	0.7820	0.7583
2	2	0.3395	0.7624	0.7442
3	3	0.3876	0.7324	0.7121
4	3	0.3876	0.7324	0.7121
8	3	0.3876	0.7324	0.7121
16	3	0.3876	0.7324	0.7121

Table 7: **Complete CelebA-HQ threshold sweeps.** Each rule evaluates 144 configurations per layer. Ordered means $j_e < j_s$; ties are reported separately. No resolved configuration reverses the ordering.

Layer	Normalized			Raw		
	Resolved	Ordered	Tied	Resolved	Ordered	Tied
early-down	36	18	18	0	0	0
late-down	144	144	0	72	72	0
mid	144	144	0	108	108	0
early-up	144	144	0	144	144	0
late-up	36	36	0	108	108	0

Controlled uncertainty. The 500 hierarchical bootstrap refits resample images and noise draws while retaining validation-selected hyperparameters. Table 8 gives the complete nominal-rule results; its intervals are conditional on resolution. The bootstrap may resolve an onset absent from the original fit, as occurs in early-down under the raw rule. Larger joint-stability SRS cutoffs impose part of the onset ordering by definition. Threshold gaps therefore summarize the curves and do not independently establish a universal transition.

B.4 Complete CIFAR-10 replication

Figure 6 extends the main early-up panel to all five stages. SRS bands use 200 class-stratified test-image bootstrap samples (seed 1234), keeping fitted predictors fixed; CNA and TC are point estimates. The nominal raw rule is $(0.05; 0.30, 0.85, 0.85)$, with two-anchor persistence for both readability and joint stability. The 144-setting raw grid varies $\theta_e \in \{.03, .05, .08, .10\}$, $\theta_s \in \{.25, .30, .35, .40\}$, and $\theta_c, \theta_t \in \{.80, .85, .90\}$. Joint onsets resolve in $0, 24, 24, 64, 0$ settings for early-down, late-down, mid, early-up, and late-up, respectively; every resolved setting is strictly ordered. Early-up's later TC decline shows that a detected stable interval need not persist.

Target-correspondence control. With predictions fixed, we apply 5,000 test-label permutations (seed 94001), sharing each permutation across all 70 sites. This differs from the face control, which refits on permuted training labels. The 95th percentile of the maximum null SRS across sites is -0.0016353. Table 9 compares the stored SRS bootstrap lower bound with the pointwise and maximum-null cutoffs. None of the five highest-noise sites passes either comparison. These controls address predictive target correspondence, not significance of geometric stability.

B.5 Reverse-trajectory fitting, reuse, and sensitivity

Common coordinates and fitting. For each representation, a single feature mean and scale are estimated from all training trajectories and anchors, then used unchanged at every validation/test state. Scales below 10^{-8} are replaced by one. One ridge value per representation is selected by mean same-state validation R^2 over the twelve anchors from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. All six representations select 1, the upper grid endpoint. Source readouts are fitted on training trajectories only. For a suffix of length L, each stacked training row receives weight $1/L$, preserving the total data weight and the meaning of the ridge penalty across suffix lengths.

Table 8: **Complete controlled-noise bootstrap summaries.** Gaps are median $j_s - j_e$ [95% percentile range], conditional on both onsets resolving. All resolved refits in this table are strictly ordered.

Layer	Norm. resolved	Norm. gap	Raw resolved	Raw gap
early-down	0/500	–	122/500	2 [1, 2]
late-down	500/500	3 [2, 4]	500/500	3 [3, 3]
mid	500/500	3 [3, 3]	500/500	4 [3, 4]
early-up	500/500	3 [3, 3]	500/500	3 [3, 3]
late-up	448/500	3 [2, 3]	500/500	3 [3, 3]

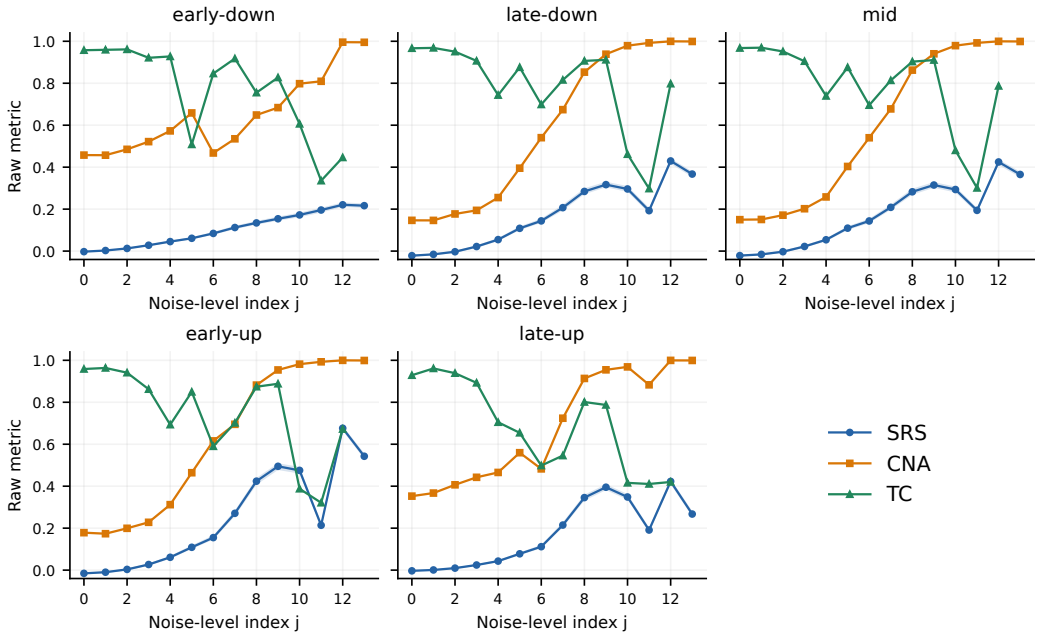


Figure 6: **Complete CIFAR-10 results.** Raw class SRS, CNA, and TC at all five stages. Shading denotes 95% fixed-predictor, stratified test-image bootstrap intervals for SRS; geometric metrics are point estimates.

Functional estimands. Let A_d be the test R^2 of the state-specific readout at destination d, and T_{sd} the test R^2 obtained by applying the source-s readout unchanged at d. Define

$$g_{sd} = A_d - T_{sd}, \qquad g_s^{\text{adj}} = g_{s,s+1},$$
$$g_s^{\text{future}} = \frac{1}{|\mathcal{D}_s|} \sum_{d \in \mathcal{D}_s} g_{sd}, \quad \mathcal{D}_s = \{d > s : A_d \geq 0.20\}, \tag{38}$$
$$g_s^{\text{shared}} = \frac{1}{J-s} \sum_{d=s}^{J-1} \big(A_d - R_d^2(\widehat{\Theta}_{s:J-1})\big), \qquad J = 12. \tag{39}$$

Here $\widehat{\Theta}_{s:J-1}$ is fitted jointly across the suffix. The state-specific reference is an estimated predictor, not a population oracle, so empirical gaps can be negative. Future gaps with no eligible destinations and terminal adjacent gaps are undefined. A terminal-only shared fit matches the state-specific fit and supplies no evidence of cross-time reuse.

Joint reuse requires $A_s \geq 0.20$ and all three gaps at most 0.05 for two consecutive source anchors. Figure 7 gives the full transfer-gap matrices, and Figure 8 separates the three gap curves at every representation. Table 10 reports readability-gated crossings for each component. In early-up, shared fitting meets its criterion at $t = 908$, adjacent transfer at $t = 454$, and future-average transfer at

Table 9: **CIFAR-10 target-correspondence control.** Counts compare the stored SRS bootstrap lower bound with the pointwise or maximum-null 95th percentile from 5,000 shared test-label permutations. Each layer has 14 sites.

Layer	Above pointwise null /14	Above maximum null /14
early-down	13	13
late-down	13	11
mid	13	11
early-up	13	11
late-up	13	13

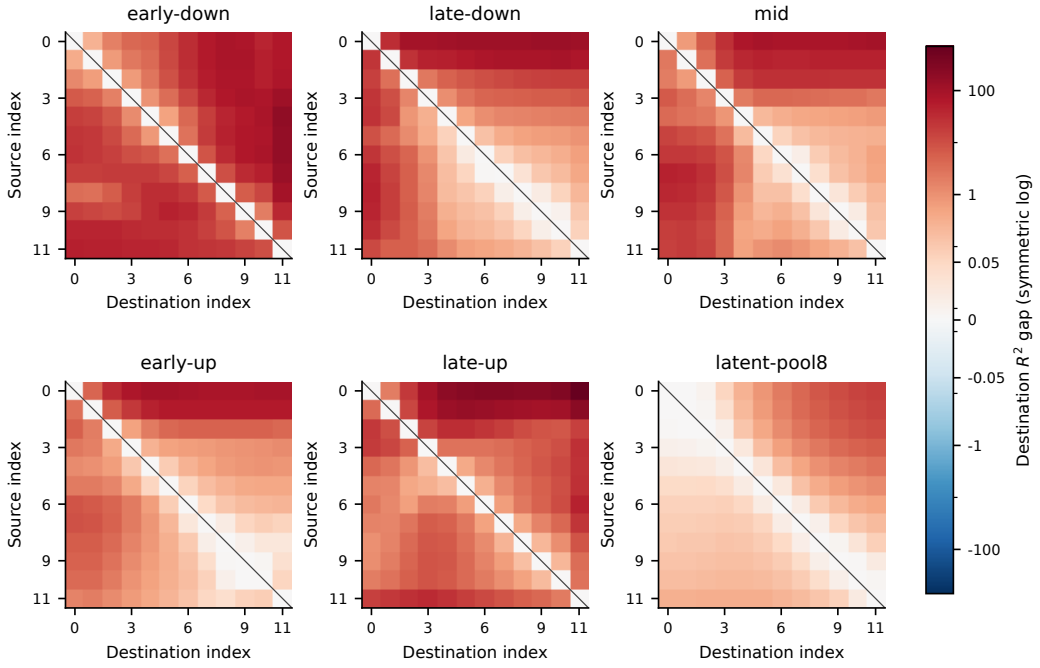


Figure 7: **Complete reverse transfer-gap matrices.** Rows are source anchors and columns are destination anchors; $d > s$ lies above the diagonal. Colors show $A_d - T_{sd}$ on a shared symmetric-log scale, linear within ± 0.05, without clipping. Diagonal entries are zero by construction.

$t = 272$. Late-down and mid pass the adjacent criterion but not the future-average criterion. Thus their unresolved joint onset does not mean that every form of reuse fails.

Pose coordinates and controls. Table 11 separates yaw, pitch, and roll at the terminal anchor. Roll is less readable than yaw and pitch in every representation; the aggregate score should not be interpreted as equal accuracy for all pose components. A training-label permutation yields negative scores for every source-transfer and shared-suffix comparison (maximum -0.0018). Permuting destination-trajectory correspondence also makes every source-transfer score negative (maximum -0.0124). These are single seeded permutation controls, not distributions from repeated permutation tests.

Refit sensitivity and threshold sweep. The 500 bootstrap runs resample complete training trajectories, refit the common normalization and all readouts, and retain the original test set and selected ridge values. Percentile ranges describe fitting sensitivity, not calibrated confidence intervals for the full-data score. For example, early-up's terminal point estimate is 0.692, while its refit percentile range is $[0.619, 0.660]$; a bootstrap training sample contains about 2,528 distinct trajectories on average out of 4,000 draws.

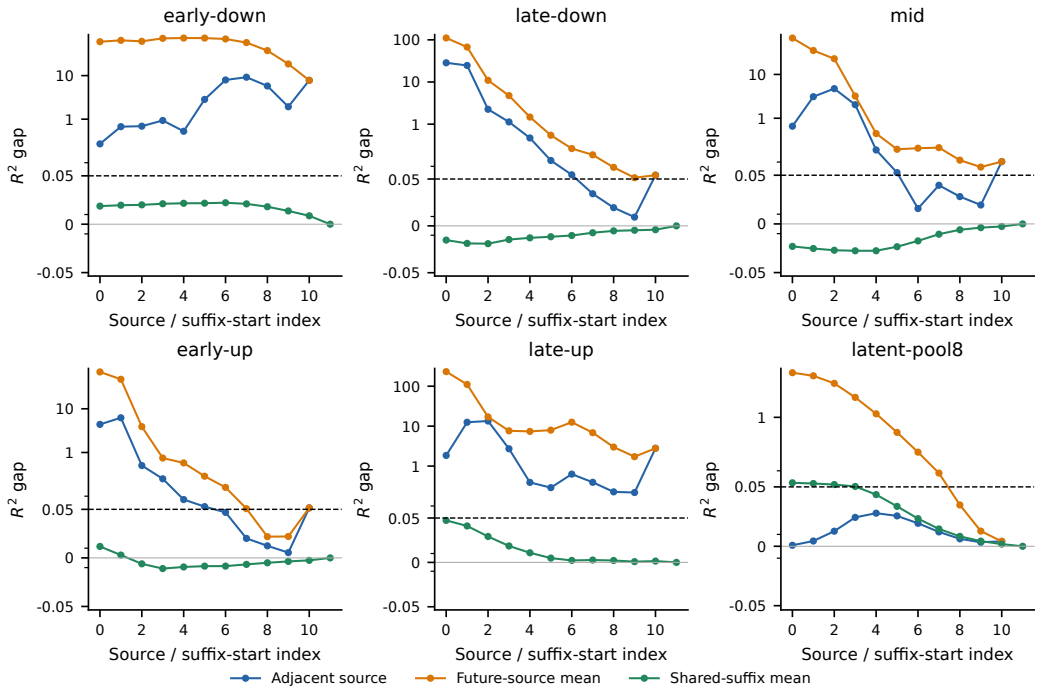


Figure 8: **Separate reuse measurements for all representations.** Adjacent-source, future-source mean, and shared-suffix mean gaps. Dashed lines mark 0.05. Vertical axes use symmetric-log scales with linear regions near zero; panel ranges adapt to the unclipped values.

Table 10: **Separate reverse reuse criteria.** Entries are the first timesteps at which readability ($R^2 \geq 0.20$) and the named gap ceiling (0.05) hold for two anchors. Joint reuse requires all three gaps simultaneously; dashes indicate unresolved crossings.

Representation	Adjacent	Future mean	Shared mean	Joint
early-down	–	–	363	–
late-down	363	–	727	–
mid	454	–	817	–
early-up	454	272	908	272
late-up	–	–	908	–
latent-pool8	636	272	636	272

The frozen threshold sweep uses readability $\{.1, .2, .3\}$ and each gap ceiling $\{.02, .05, .1\}$, giving 81 configurations per representation. The archived sweep changes these crossing thresholds while retaining the future-destination mask computed at the nominal readability cutoff 0.20. Table 12 retains unresolved outcomes for both analyses. Percentile bounds between integer anchor indices arise from interpolation, not extra evaluated states.

B.6 Empirical risk and decoder-drift decompositions

For the held-out destination design Φ_d, set $D = \widehat{\Theta}_s - \widehat{\Theta}_d$ and $E_d = Y - \Phi_d \widehat{\Theta}_d$. With MSE averaged over n images and q targets,

$$\widehat{L}_d(\widehat{\Theta}_s) - \widehat{L}_d(\widehat{\Theta}_d) = \underbrace{\frac{\|\Phi_d D\|_F^2}{nq}}_{Q} \underbrace{- \frac{2\langle E_d, \Phi_d D\rangle_F}{nq}}_{R}. \tag{40}$$

The residual cross term need not vanish for a predictor fitted on different data. This held-out identity is distinct from a quadratic population ridge-objective gap. Across $6 \times 12 \times 12 = 864$ pairs, the

Table 11: **Terminal reverse readability by pose coordinate.** Held-out R^2 at $t = 0$; the mean weights the three coordinates equally.

Representation	Yaw	Pitch	Roll	Mean
early-down	0.616	0.295	0.155	0.355
late-down	0.816	0.626	0.114	0.518
mid	0.837	0.710	0.172	0.573
early-up	0.904	0.837	0.336	0.692
late-up	0.862	0.722	0.236	0.607
latent-pool8	0.814	0.355	0.201	0.456

Table 12: **Reverse refit and threshold sensitivity.** Bootstrap gaps are median anchor differences [95% percentile range], conditional on both onsets resolving. Every resolved bootstrap refit is strictly ordered. Sweeps retain unresolved configurations in the denominator.

Representation	Bootstrap resolved	Bootstrap gap	Sweep resolved	Sweep ordered
early-down	0/500	–	0/81	–
late-down	29/500	$4\,[3.7, 5]$	27/81	27/27
mid	58/500	$5\,[4, 6]$	0/81	–
early-up	463/500	$6\,[5, 7]$	54/81	54/54
late-up	0/500	–	0/81	–
latent-pool8	500/500	$3\,[3, 4]$	81/81	81/81

maximum reconstruction error is 1.31×10^{-10}. Figure 9 compares observed gaps with Q for the 66 strictly forward pairs in early-up and pooled latents; Table 13 retains the residual contribution for every representation.

For augmented training sufficient statistics $\widehat{A}_e = \Phi_{e,\mathrm{tr}}^\top \Phi_{e,\mathrm{tr}} + \lambda P$ and $\widehat{C}_e = \Phi_{e,\mathrm{tr}}^\top Y_{\mathrm{tr}}$, we also evaluate

$$D = \widehat{A}_s^{-1}(\widehat{C}_s - \widehat{C}_d) + \widehat{A}_s^{-1}(\widehat{A}_d - \widehat{A}_s)\widehat{A}_d^{-1}\widehat{C}_d. \tag{41}$$

The maximum reconstruction error is 3.90×10^{-10}. The two drift terms can cancel, so their norms are not additive fractions of transfer failure. These calculations verify the decompositions for the fitted predictors; they do not establish moment smoothness, local identifiability, or causal mediation.

B.7 Reliability prediction and representation selection

Selection protocol. The 55 candidates are five face-model layers at eleven nonterminal controlled-noise anchors. Probe-only selection maximizes validation SRS. Reliability-aware policies fit ordinary least squares to predict single-noise R^2 from SRS+CNA, optionally adding TC. Each candidate is scored by a model trained on the other four layers, with covariate standardization fitted within that fold. Although the candidate's own layer is excluded, other-layer outcomes use the same test images. This is cross-layer analysis, not an independent image-level evaluation of a learned policy.

Evaluation outcomes. Single-noise R^2 averages scores across individual draws. Cross-noise correlation and disagreement compare the same readout applied to two disjoint four-draw means; correlations are computed across images for each pose coordinate and averaged. Squared disagreement and prediction variance use predictions normalized by target standard deviation. Noise-averaging gain is the difference between averaged-input and mean single-draw R^2. Adjacent prediction correlation compares separately fitted state-specific predictors, whereas temporal transfer evaluates one unchanged source predictor at the next anchor.

Incremental information. Table 15 gives all five cross-noise outcomes over 60 sites, including the terminal anchor. Adding CNA improves every outcome, although absolute CV R^2 remains small for disagreement and variance. Table 16 reports all four temporal outcomes over 55 adjacent pairs. TC improves prediction correlation but does not improve prediction of unchanged-readout transfer. These outcomes concern neighboring controlled-noise sites, not the independent reverse trajectories.

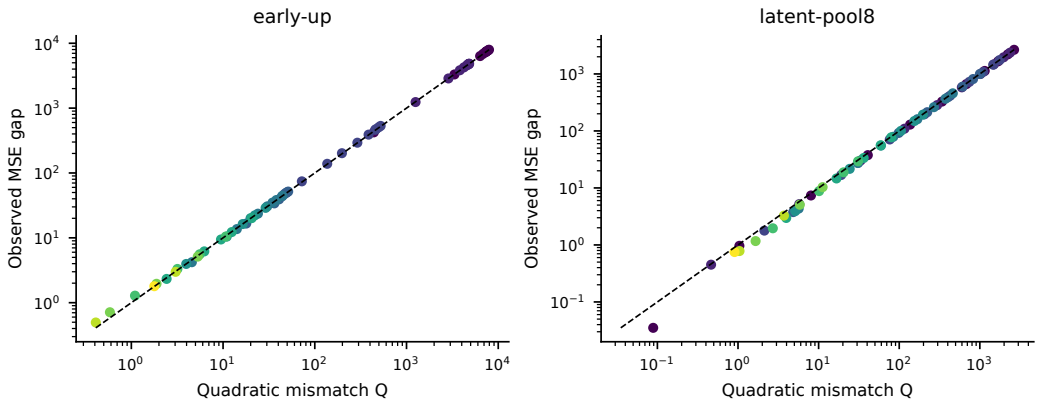


Figure 9: **Predictor mismatch and observed transfer loss.** Each point is a strictly forward source–destination pair. Dashed lines mark equality between quadratic mismatch Q and the observed MSE gap; deviations are the residual term R. Both axes use logarithmic scales.

Table 13: **Empirical decomposition checks.** Maximum errors use all 144 ordered source–destination pairs per representation. Median residual ratios use the 66 strictly forward pairs with nonzero risk gaps.

Representation	Risk identity error	Drift identity error	Median $\lvert R\rvert/\lvert Q+R\rvert$
early-down	2.18e-11	3.89e-10	0.0109
late-down	5.91e-12	1.82e-10	0.0121
mid	4.55e-12	1.91e-10	0.0179
early-up	8.19e-12	2.86e-10	0.0095
late-up	1.31e-10	3.90e-10	0.0066
latent-pool8	3.64e-12	1.40e-13	0.0516

B.8 Computational environment and reproducibility

Feature extraction is cached; readout fitting, permutations, threshold sweeps, and bootstrap analyses reuse these arrays. Table 17 records environments associated with the saved runs, rather than assuming one software stack for every experiment. Confirmation extraction generated 5,003 trajectories to obtain 5,000 valid pose measurements. The recorded worker extraction durations are approximately 2,576 and 2,591 seconds on two GPUs; they exclude later readout and bootstrap analysis and are not total project compute. The supplied result tables retain site identities and unrounded scores, and the source manifest records the configurations and result files used to generate this appendix.

The appendix supplies empirical evidence for the observed readability–reliability separation. It does not claim numerical validation of universal recovery constants or an independent-pool $1/K$ rate, nor independent-image confirmation of the selection policy. Those stronger conclusions require additional experiments.

C Discussion and Scope

What has been established. Semantic accessibility and semantic reliability are distinct. In controlled repeated-noise families, finite replicate averaging has an exact nuisance-variance interpretation and can recover an identifiable predictive subspace under explicit conditions. Along actual reverse trajectories, high same-state readability does not guarantee that a source-trained predictor can be reused later. These statements are stronger than probe accuracy but weaker than a theory of how semantic information is created by training.

Why this matters for interpretability. Interpretability claims often attach meaning to a direction because a probe decodes a concept. A useful interpretation should survive at least the nuisance

Table 14: **Reliability-aware representation selection.**

Policy	t	Val. R^2	Single-noise R^2	Noise corr.	Adj. corr.	Noise disagr.
Probe only	545	**.6801**	.5879	.9401	.9465	.0590
SRS+CNA	454	.6634	**.6077**	.9653	.9675	.0319
SRS+CNA+TC	363	.6385	.6005	**.9776**	**.9801**	**.0197**

Table 15: **Incremental cross-noise reliability prediction.** Leave-one-layer-out regression over 60 sites. The baseline uses validation SRS; the augmented model adds CNA. CV R^2 concerns prediction of each listed outcome.

Outcome	Baseline	+CNA	ΔR^2	MAE reduction
Prediction disagreement	-0.162	0.062	+0.224	+0.0112
Across-noise variance	-0.164	0.058	+0.222	+0.0227
Noise-averaging gain	-0.163	0.058	+0.221	+0.0203
Single-noise R^2	0.167	0.656	+0.489	+0.0910
Cross-noise correlation	0.080	0.929	+0.849	+0.1993

variation and state changes over which it is claimed to describe the model. Noise reproducibility and predictive reuse provide two auditable tests, and the exploratory cross-layer selection analysis shows that these tests can change which representation is chosen in practice.

Unresolved and negative outcomes. Several hidden representations are readable but never satisfy the confirmatory transport rule, and the pooled latent reaches transport no later than the tested U-Net feature. We retain these outcomes. A reliability audit that labeled every readable state stable would collapse back to probing. Likewise, NVC is not semantic-specific under the random-subspace control and is therefore not used alone as a semantic certificate.

What remains open. The semantic targets inherit evaluator error, cross-time reuse is functional rather than causal, and conditional or language-conditioned diffusion may introduce additional condition–representation interactions. Nonlinear readout families and direct empirical checks of the local-identifiability constants are natural extensions. These open questions do not change the estimands studied here, but they delimit where the conclusions should be carried.

D Claim–Evidence Map

Table 16: **Incremental temporal prediction.** Leave-one-layer-out regression over 55 adjacent controlled-noise pairs. The baseline uses the mean and absolute difference of validation SRS at the two anchors; the augmented model adds TC. Positive MAE reduction is better.

Outcome	Baseline	+TC	ΔR^2	MAE reduction
Unchanged-readout transfer drop	-0.490	-0.499	-0.008	-0.0556
Adjacent prediction disagreement	0.645	0.581	-0.064	-0.0029
Adjacent prediction correlation	0.237	0.671	+0.434	+0.0581
Cross-time transfer R^2	-0.492	-0.494	-0.003	-0.0339

Table 17: **Recorded computational environments.** Entries describe the named run records; the initial face environment check does not establish the software version of every later analysis.

Record	Software	Execution details
Face-model environment check	Python 3.8.5; PyTorch 1.11.0; CUDA 11.3; cuDNN 8.2	RTX 4090; primary controlled cache records two device IDs, batch 64, float16 storage.
CIFAR-10 preflight	PyTorch 2.1.1+cu121; NumPy 1.26.4	Batch 32; autocast disabled; float16 feature storage.
Reverse confirmation	Python 3.8.5; PyTorch 1.12.1; CUDA 11.3	Two RTX 4090 GPUs (24 GB each); batch 24 per GPU; autocast extraction, float16 cache; float64 ridge solves.

Question	Primary evidence	Main outcome
Do readable targets have reproducible readouts?	Repeated-noise audit; permutation/random-subspace controls; K-ablation	Target-specific alignment exceeds random baselines; replication improves measured reliability.
Does reliability replicate beyond faces/LDMs?	CIFAR-10 unconditional pixel-space DDPM	Early-up readability at $j = 4$ precedes joint stability at $j = 8$; unresolved layers remain unresolved.
Is a readable relation reusable in generation?	Independent $5{,}000$-trajectory DDIM confirmation	Early-up readability at $t = 908$ precedes joint reuse at $t = 272$.
Does reliability change a practical decision?	Cross-layer representation selection	Cross-noise correlation improves from 0.9401 to 0.9776 versus probe-only selection.

Table 18: Claim–evidence map. Each empirical block tests a distinct aspect of semantic readout reliability.