What Time Is It? How Data Geometry Makes Time Conditioning Optional for Flow Matching

Alec Helbling[1,2] **Sebastian Gutierrez Hernandez**[1] **Benjamin Hoover**[1]
Duen Horng Chau[1] **Parikshit Ram**[2]

[1]Georgia Tech [2]IBM Research

Abstract

Diffusion and flow models generate data by learning to denoise, conventionally with explicit time conditioning that tells them the noise level. Recent work shows that they can be trained without time conditioning, challenging the standard view that time conditioning is needed to denoise effectively. Why does this work? We decompose the time-blind flow matching loss into two sources of irreducible error: a *coupling variance*, from ambiguity in how noise and data are paired, and a *time-blindness gap*, the extra cost of ignoring time. We show that high-dimensional geometry keeps this gap small. When data concentrates near a k-dimensional subspace, the noise in the remaining directions acts as a clock: under a spiked Gaussian model with independent coupling, a closed-form estimator recovers t from a single noisy sample at rate $O(1/\sqrt{d-k})$, and away from a critical time the time-blindness gap is asymptotically negligible relative to the coupling variance. We validate the estimator on real data and find, across CIFAR-10, CelebA-HQ, and FFHQ, that changing the coupling affects flow-matching loss substantially more than removing time conditioning, while removing time conditioning consistently improves mean FID. These results explain why time-blind flow matching works and show that the main practical lever for flow-matching loss is the choice of coupling, not explicit time conditioning.

1 Introduction

Generative models produce new samples, such as images, from random noise. Flow matching [17, 1] trains a neural network to predict how a noisy sample should move toward data, using a velocity field $v_\theta(z, t)$ conditioned on interpolation time t. The time variable is usually treated as essential, under the intuition that conditioning on t seems necessary to disambiguate between velocity targets when at the same spatial location. From this perspective, a time-blind vector field $f_\theta(z)$ should be fundamentally disadvantaged, since it must assign a single velocity to locations that may be reached by many trajectories at different times (Figure 2).

Recent work challenges this view. The performance of time-unconditional generative models often barely degrades at all when explicit conditioning is removed and sometimes even improves [28, 31], and blind denoising analyses show that the noise level can sometimes be inferred from the noisy observation itself under low-dimensional data structure [13]. But for flow matching, this is not the whole story. Flow matching has another source of irreducible ambiguity: the coupling between noise and data during training. Even if time is known, many coupled pairs can pass through the same interpolant and induce different velocities. Thus the central question is not if a time-blind model can infer t, but rather if the residual cost of hiding t is large compared to the coupling-induced error that all flow matching models already pay.

The time-blind flow matching loss decomposes into a *learnable approximation* term, a *coupling variance* that remains even for an optimal time-conditioned predictor, and a *time-blindness gap*

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

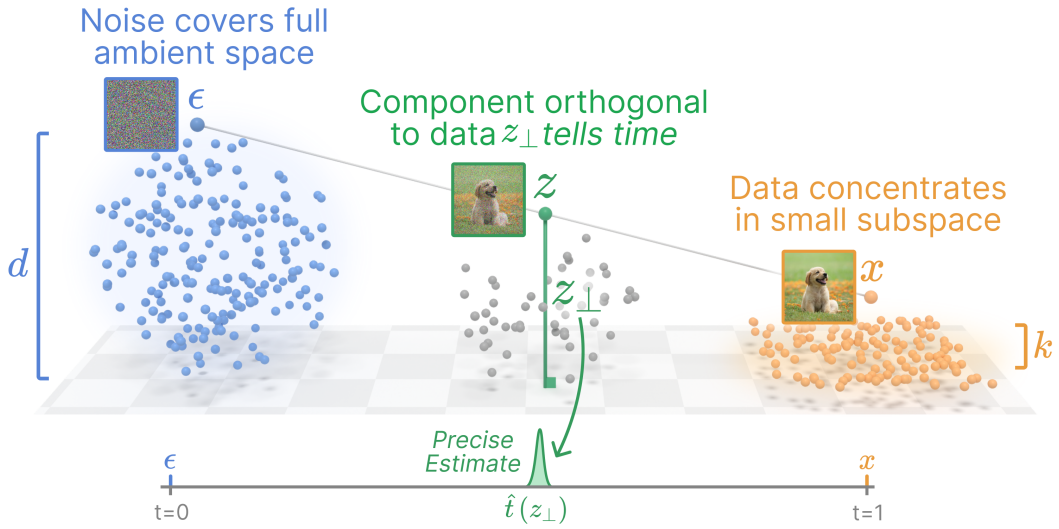


Figure 1: **Time-blind flow matching works because data and noise occupy subspaces of very different sizes, leaving a large residual noise subspace whose variance encodes the interpolation time.** (Left) Noise dominates in a large subspace: the source $\varepsilon \sim \mathcal{N}(0, I_d)$ is isotropic over the full ambient dimension d. (Right) Data concentrates in a small subspace: the target distribution is supported on a k-dimensional signal subspace with $k \ll d$. (Center) Identify time from noise-subspace statistics: the interpolant $z = (1-t)\varepsilon + tx$ inherits the dimensionality mismatch, so its component $z_\perp$ orthogonal to the data subspace is supplied almost entirely by the noise term and is approximately Gaussian with variance directly tied to t. Estimating that variance recovers t from a single observation z with no time conditioning required.

caused by averaging over unobserved times. The gap measures exactly how much harder time-blind training is than standard flow matching, and we show that in practical settings it is small. When data concentrates in a low-dimensional subspace, the interpolant's component orthogonal to that subspace is supplied almost entirely by the Gaussian source, and its magnitude changes predictably with t, a built-in statistical clock for the hidden time. We formalize this with an explicit estimator whose error shrinks with the residual ambient dimension, making the cost of hiding time negligible compared with the coupling variance. In our experiments, coupling has the larger effect on flow-matching loss, while removing time conditioning consistently improves mean FID. The latter is a separate empirical finding that our loss decomposition does not explain.

In summary, our contributions are as follows:

1. **An explanatory loss decomposition for time-blind flow matching.** We decompose the time-blind flow matching (FM) loss into a learnable error, an irreducible coupling variance shared with standard FM, and an additional time-blindness gap unique to time-blind flows (Proposition 1).

2. **A constructive time-identification statistic for noisy interpolants.** Under a spiked Gaussian data model with independent coupling, on an invertible clock branch away from its critical point, the residual-subspace energy of a single interpolant z yields a closed-form estimator of t with error $O_p(1/\sqrt{d-k})$ (Theorem 2).

3. **A relative bound showing that coupling dominates the cost of hiding time.** Under the same assumptions, combining the decomposition with the identifiability rate, we prove that the time-blindness gap is smaller than the coupling variance by a factor $O(1/(d-k))$ (Theorem 3).

4. **Empirical validation on real data and trained flow models.** On CIFAR-10, CelebA-HQ, and FFHQ, the residual-subspace statistic recovers interpolation time at the predicted rate (Section 5.2), and trained models show that coupling has a much larger effect on flow-matching loss than explicit time conditioning. Removing time conditioning nevertheless yields lower mean FID in all six dataset–coupling comparisons (Section 5.1). Probing and subspace-corruption diagnostics further show that time is decodable from learned features and localize information supporting that decoding to residual input directions (Section 5.3).

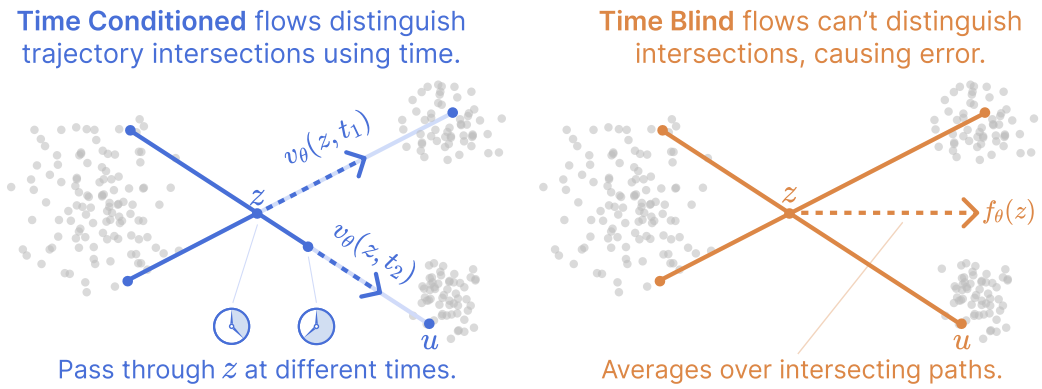


Figure 2: **Time-conditioned flows disambiguate trajectory intersections using t. Time-blind flows must commit to a single velocity at each location, averaging over conflicting trajectories.** (*Left*) A time-conditioned model $v_\theta(z, t)$ observes t and can assign different velocities at the same spatial location z for different times; here, $v_\theta(z, 0.3)$ and $v_\theta(z, 0.7)$ point along distinct trajectories that pass through z. (*Right*) A time-blind model has no t to disambiguate with, so it must commit to a single $f_\theta(z)$ that averages over all paths through z.

2 Problem Setup

In this section we present the theoretical formulation for understanding the core problems faced by conventional flow matching models, and more specifically by time-blind models that drop the time conditioning.

Flow matching. Flow matching (FM) [17, 1] learns a time-conditioned velocity field $v_\theta(x, t)$ transporting a noise distribution p_0 to a data distribution p_1 over $t \in [0, 1]$. Given a coupling $(\varepsilon, x) \sim \pi$ with $\varepsilon \sim p_0$ and $x \sim p_1$, the linear interpolant defines a conditional path

$$z = (1 - t)\,\varepsilon + t\,x \tag{1}$$

with constant conditional velocity $u = x - \varepsilon$. The model is trained to regress $v_\theta(z, t)$ onto u:

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{t,\,(\varepsilon,x)\sim\pi}\left[\|v_\theta(z, t) - u\|^2\right]. \tag{2}$$

Time-blind flow matching. We consider a time-blind vector field $f_\theta(z)$ trained against the same interpolant-based regression target $u = x - \varepsilon$. Given $(\varepsilon, x) \sim \pi$ and the interpolant $z = (1-t)\varepsilon + tx$ with $t \sim \mathcal{U}[0, 1]$, the time-blind loss is

$$\mathcal{L}_{\mathrm{TB}} = \mathbb{E}_{t,\,(\varepsilon,x)\sim\pi}\left[\|f_\theta(z) - u\|^2\right]. \tag{3}$$

The key question is: what is the cost of dropping the time conditioning that FM relies on?

3 An Explanatory Decomposition of the Time-Blind Loss

The time-blind loss $\mathcal{L}_{\mathrm{TB}}$ mixes together errors of very different origins: some that any flow matching model must pay regardless of time conditioning, some that are unique to dropping t, and some that the practitioner can drive away by training a better model. Pulling these apart is what lets us pose the central question of the paper precisely: how does the extra cost of hiding t compare to the irreducible cost that flow matching pays anyway? Proposition 1 makes this separation exact.

3.1 The Three-Term Decomposition

We consider arbitrary time-blind vector fields $f : \mathbb{R}^d \to \mathbb{R}^d$. Throughout, unsubscripted $\mathbb{E}$ denotes expectation over the full joint $(\varepsilon, x, t) \sim \pi \otimes \mathcal{U}[0, 1]$.

Proposition 1 (Anatomy of the Time-Blind Loss). *For any predictor $f : \mathbb{R}^d \to \mathbb{R}^d$, the time-blind loss decomposes as*

$$\mathcal{L}_{\mathrm{TB}}(f) \;=\; \underbrace{\mathbb{E}\big[\|f(z) - \mathbb{E}[u \mid z]\|^2\big]}_{\text{I. Learnable error}} \;+\; \underbrace{\mathbb{E}\big[\mathrm{Var}(u \mid z, t)\big]}_{\text{II. Coupling variance } \mathcal{C}_\pi} \;+\; \underbrace{\mathbb{E}\big[\mathrm{Var}\big(\mathbb{E}[u \mid z, t] \,\big|\, z\big)\big]}_{\text{III. Time-Blindness Gap } \Delta}. \tag{4}$$

(Proof in Appendix A.1.)

The decomposition is exact, holds for any f, and requires no assumptions on the coupling or model class. Each term has a distinct origin and a distinct lever, which we now unpack in turn.

I. Learnable error. This is the squared distance between the predictor and the optimal time-blind predictor $f^*(z) := \mathbb{E}[u \mid z]$. It is the only term that depends on the choice of predictor, and the only one the practitioner can drive to zero: with sufficient capacity and data, $f_\theta \to f^*$ and this term vanishes.

II. Coupling variance. This is the irreducible variance $\mathcal{C}_\pi := \mathbb{E}[\mathrm{Var}(u \mid z, t)]$ of the regression target after conditioning on *both* z and t. This floor exists even if we grant the model the very information we are trying to hide: fixing (z, t) does not fix the underlying pair (ε, x), since $z = (1-t)\varepsilon + tx$ is one affine constraint on two variables. Distinct pairs that route through the same (z, t) generally carry different velocities, and the best any predictor can do is predict their conditional mean $v^*(z, t) := \mathbb{E}[u \mid z, t]$, leaving residual variance $\mathcal{C}_\pi$. This floor is paid by the optimal time-*conditioned* predictor in standard flow matching too (Lemma A.1), and the time-blind model must at least inherit it. The size of $\mathcal{C}_\pi$ is controlled entirely by the coupling π: independent couplings permit many distinct (ε, x) to route through the same (z, t) with disagreeing velocities, while structured couplings such as mini-batch optimal transport pair nearer points and tighten the conditional. This is the lever we manipulate in Section 4.2.

III. Time-blindness gap. This is the *additional* price $\Delta := \mathbb{E}[\mathrm{Var}(v^*(z, t) \mid z)]$ paid for not observing t, on top of the coupling variance. A time-blind predictor cannot assign different velocities to the same location at different times (Figure 2), so at its optimum it must average v^* over the posterior $t \mid z$: $f^*(z) = \mathbb{E}[v^*(z, t) \mid z]$. The gap Δ measures how much v^* varies across this posterior — if z nearly determines t, the gap collapses. Since $\Delta \geq 0$ always (Corollary A.2), this term is unique to time-blind models and makes the time-blind optimum at least as hard as conventional flow matching.

4 Data Geometry Mitigates the Time-Blindness Gap

Proposition 1 reduces the puzzle to a comparison between two irreducible terms: the time-blindness gap Δ and the coupling variance $\mathcal{C}_\pi$. This section shows that, in high-dimensional low-rank regimes, the hidden time is already encoded in the geometry of the interpolant. In particular, the residual subspace provides a high-dimensional clock that identifies t from z alone. Once t is nearly determined by z, explicitly conditioning on t adds little information, so the time-blindness gap is small. Under the spiked covariance model, this intuition yields the asymptotic comparison $\Delta/\mathcal{C}_\pi = O(1/(d-k))$, leaving the coupling variance as the dominant loss floor.

4.1 The Noise Subspace is a High-Dimensional Clock

Suppose the data $x \in \mathbb{R}^d$ concentrates near a k-dimensional signal subspace with $k \ll d$. Let $U_\perp \in \mathbb{R}^{d \times (d-k)}$ be an orthonormal basis for the orthogonal complement of this subspace, and let $P_\perp = U_\perp U_\perp^\top$ be the projection preserved by that basis. The key observation is that $P_\perp z$ removes most of the data signal while retaining the ambient Gaussian source noise. Thus the residual energy $\|P_\perp z\|^2$ changes predictably with the interpolation time t.

Spiked covariance model. To make this mechanism explicit, we analyze a standard spiked covariance model [12, 6, 30]

$$x \sim \mathcal{N}(0,\, U_x \Lambda U_x^\top + \sigma^2 I_d),$$

where $U_x \in \mathbb{R}^{d \times k}$ spans the signal subspace, Λ contains the excess variances in the signal directions, and $\sigma^2 I_d$ is an isotropic residual floor. This model idealizes the empirical observation that natural

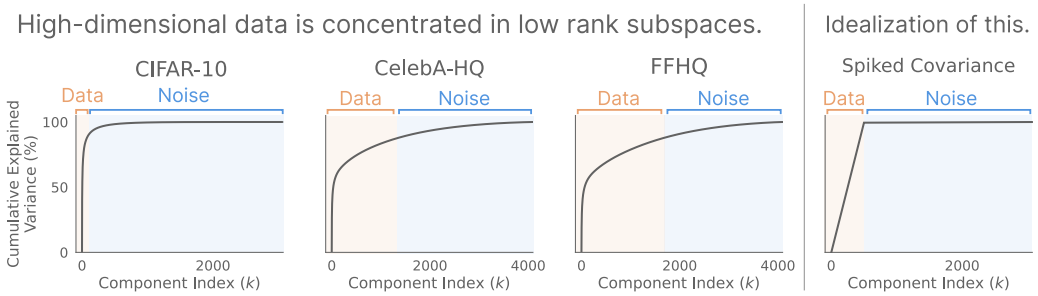


Figure 3: **Real data concentrates in a low-rank subspace and leaves a sizable residual subspace.** Cumulative explained variance captured by the top-k principal components on each real dataset and on the spiked covariance model used in Theorem 2. Pixel-space CIFAR-10 shows dramatic low-rank concentration; on the FFHQ and CelebA-HQ VAE latents variance is more spread out across components but still concentrates most of its mass in a small fraction of them, leaving a large residual subspace ($d-k$ in the thousands at the 80% threshold shown). The spiked model is one tractable instance of the broader condition that $d-k$ is large enough for the noise-subspace clock.

data often has most of its variance in a low-dimensional subspace while retaining small residual variance in the remaining directions (Figure 3).

The residual variance clock. With independent Gaussian source and target samples and the linear interpolant, projection onto the orthogonal complement of U_x eliminates the spike directions. In each residual coordinate, $P_\perp z$ combines an ambient Gaussian-noise coordinate with variance 1, scaled by $(1-t)$, and a residual data coordinate with variance σ^2, scaled by t. Since these terms are independent and centered, their variances add:

$$\mathrm{Var}[(P_\perp z)_j] = (1-t)^2\,\mathrm{Var}(\varepsilon_j) + t^2\,\mathrm{Var}(x_j) = (1-t)^2 + t^2\sigma^2.$$

Therefore,

$$P_\perp z \;\sim\; \mathcal{N}\big(0,\; \sigma_\perp^2(t)\, I_{d-k}\big), \qquad \sigma_\perp^2(t) \;:=\; (1-t)^2 + t^2\sigma^2. \tag{5}$$

We call $\sigma_\perp^2(t)$ the *variance clock*. A high-dimensional isotropic Gaussian places almost all its mass on a thin shell of squared radius $(d-k)\,\sigma_\perp^2(t)$, with relative fluctuations of order $1/\sqrt{d-k}$, so observing the empirical radius of $P_\perp z$ pinpoints the variance and hence t. This gives a closed-form estimator on any branch where the clock is invertible:

$$\widehat{\sigma}_\perp^2 \;:=\; \frac{1}{d-k}\,\|P_\perp z\|^2, \qquad \hat{t}(z) \;:=\; (\sigma_\perp^2)^{-1}\big(\widehat{\sigma}_\perp^2\big). \tag{6}$$

Because the shell concentrates at rate $1/\sqrt{d-k}$, this single-sample estimator recovers t with vanishing error as the residual dimension grows, as shown in Theorem 2.

Theorem 2 (Identifiability of t from z; informal). *Under the spiked Gaussian model with independent coupling and linear interpolation, on a fixed invertible clock branch bounded away from its critical point, the estimator $\hat{t}(z)$ in* (6) *satisfies*

$$\hat{t}(z) \;=\; t \;+\; O_p\!\left(\frac{1}{\sqrt{d-k}}\right).$$

(Formal statement and proof in Appendix B.2.)

Thus, removing explicit time conditioning does not necessarily remove time information. In high-dimensional low-rank regimes, the interpolant z itself contains an implicit clock, so a sufficiently expressive time-blind predictor can infer the relevant time from geometry rather than from an external conditioning variable. This conclusion is restricted to the stated model and a known branch bounded away from the critical point: $\sigma_\perp^2(t)$ decreases until its critical point $t^* = 1/(1+\sigma^2)$ and increases thereafter, so the variance alone identifies t only on intervals where this map is one-to-one. In our datasets, t^* lies near the data endpoint, so the descending branch covers nearly the full interpolation path (Table C.2).

Beyond isotropic spectra. The residual-clock mechanism does not rely on an exactly isotropic residual spectrum. For anisotropic Gaussian residuals, concentration is governed by the effective rank $r_{\mathrm{eff}}(\Sigma_{P,t}) := (\operatorname{tr}\Sigma_{P,t})^2/\operatorname{tr}(\Sigma_{P,t}^2)$, where $\Sigma_{P,t} := (1-t)^2 I + t^2\Sigma_P$, which measures the number of projected directions that contribute appreciable variance. Thus the role played by $d-k$ in the spiked model is played more generally by $r_{\mathrm{eff}}(\Sigma_{P,t})$ (Proposition B.1, Appendix B.4). We focus on the spiked model in the main text because it yields a closed-form inverse clock and a transparent comparison between the time-blindness gap and coupling variance (Theorem 3). Empirically, PCA-fitted residual spectra on real data have sufficiently large effective rank for the same prediction to track the observed estimator-error distributions (Section 5.2).

4.2 The Time-Blindness Gap is Negligible Relative to Coupling Variance

Theorem 2 shows that t can be recovered from z alone at rate $O_p(1/\sqrt{d-k})$. This directly controls the time-blindness gap in Proposition 1. Indeed, if t is nearly determined by z, then the time-aware optimal predictor $v^*(z,t) = \mathbb{E}[u \mid z,t]$ varies little over the posterior uncertainty $t \mid z$. Equivalently, explicitly revealing t adds little information beyond what is already present in z. Under the spiked covariance model, this intuition yields a quantitative comparison between the extra cost of hiding time and the coupling variance already paid by time-conditioned flow matching.

Theorem 3 (Cost of Time-Independence; informal). *Under the spiked Gaussian model with independent coupling and linear interpolation, on a fixed invertible clock branch bounded away from its critical point, the time-blindness gap Δ and coupling variance $\mathcal{C}_\pi$ satisfy*

$$\frac{\Delta}{\mathcal{C}_\pi} \;=\; \frac{\mathbb{E}\big[\mathrm{Var}\big(\mathbb{E}[u \mid z,t] \mid z\big)\big]}{\mathbb{E}\big[\mathrm{Var}(u \mid z,t)\big]} \;=\; O\!\left(\frac{1}{d-k}\right).$$

(Formal statement and proof in Appendix B.7.)

The rate is set by the noise-subspace dimension alone: holding k fixed and growing d gives the $O(1/d)$ scaling, while holding $d-k$ fixed and growing k leaves the ratio bounded by a constant. This separates our account from purely ambient-dimension explanations of time identifiability [28], and we validate both regimes empirically in Sections 5.4 and 5.1.

Coupling as the dominant lever for flow-matching loss. The coupling variance $\mathcal{C}_\pi$ is not specific to time-blind models: by Lemma A.1, it is the irreducible floor paid even by the optimal time-conditioned predictor $v^*(z,t) = \mathbb{E}[u \mid z,t]$. Theorem 3 shows that the additional price of removing time conditioning is lower order in the high-dimensional low-rank regime. Thus, once time is identifiable from z, reducing $\mathcal{C}_\pi$ should matter more than restoring explicit time conditioning. Our experiments investigate mini-batch optimal transport [29, 9], which approximates the population OT map [3] by solving a linear assignment within each training batch. The decomposition remains exact under this coupling, but the independent-coupling relative bound does not apply: as coupling variance approaches zero, the ratio need not remain small.

5 Experiments

Our experiments evaluate the loss decomposition and time-identifiability predictions, alongside sample quality. Section 5.1 shows approximately additive effects of coupling and time conditioning on measured flow-matching loss, with coupling having the larger effect. Separately, removing time conditioning consistently improves mean FID across the evaluated settings; this sample-quality benefit is not predicted by the loss decomposition. Section 5.2 confirms that the noise-subspace estimator $\hat{t}(z)$ matches its predicted error distribution on synthetic and real data. Section 5.4 shows that on the spiked Gaussian, the time-blindness gap is governed by the noise-subspace dimension $d-k$ rather than d alone, and is dwarfed by the coupling variance across the (d,k) plane.

We evaluate on two complementary settings: (1) **Real image datasets.** CIFAR-10 [16] in pixel space (d=3,072) with an ADM UNet, and CelebA-HQ [14] and FFHQ [15] in Stable Diffusion VAE [25] latent space (d=4,096) with a SiT-B/2 backbone [20], spanning two modalities and two architectures so that observed effects are not artifacts of either. (2) **Synthetic spiked covariance.** The Gaussian model of Section 4.1 with Λ eigenvalues spaced linearly in $[1,10]$ and $\sigma^2 = 0.1$, where varying d and k directly sweeps the noise-subspace dimension $d-k$ that governs the rate in Theorem 2. (Details in Appendix C.1.)

Table 1: **Flow-matching MSE and FID over three random seeds.** Entries are mean ± standard deviation across seeds; lower is better for both metrics. Each model is evaluated under its training coupling. MSE uses the CIFAR-10 test split and the training image sets with fresh noise and times for CelebA-HQ and FFHQ. FID uses clean-FID-50k. Evaluation details are in Appendix C.2.

	CIFAR-10		CelebA-HQ		FFHQ	
Method	MSE ↓	FID ↓	MSE ↓	FID ↓	MSE ↓	FID ↓
Time-conditioned, naive coupling	0.17026 ±0.00002	5.415 ±0.019	0.5991 ±0.0030	11.15 ±0.28	0.6711 ±0.0003	30.01 ±0.16
Time-conditioned, OT coupling	0.14530 ±0.00006	5.249 ±0.165	0.5591 ±0.0002	11.04 ±0.08	0.6342 ±0.0002	29.52 ±0.17
Time-blind, naive coupling	0.17065 ±0.00004	5.226 ±0.080	0.6101 ±0.0012	9.86 ±0.09	0.6743 ±0.0007	29.25 ±0.16
Time-blind, OT coupling	0.14557 ±0.00001	5.136 ±0.123	0.5670 ±0.0001	10.28 ±0.05	0.6374 ±0.0004	29.11 ±0.02

5.1 The Decomposition is Additive and Coupling is the Dominant Lever

We train flow matching models on each dataset, varying (i) whether the model conditions on time t or receives $t = 0$ at every step, and (ii) whether the coupling is independent or mini-batch OT (exact EMD solver). Each model is evaluated under its own training coupling, so the measured loss reflects the irreducible floor of the coupling it was trained with (details in Appendix C.2). We report mean and standard deviation over three random seeds for each configuration. Table 1 supports the following observations:

Coupling and time conditioning appear to be distinct, approximately additive levers. If coupling and time conditioning act independently, switching from independent to OT coupling should produce a similar improvement with and without time conditioning. Across all three datasets, switching from independent to OT coupling produces nearly the same improvement in both the time-conditioned and time-blind rows. This suggests that improved coupling does not simply compensate for missing time information, and that removing time conditioning does not erase the benefit of OT coupling.

Coupling has the larger effect on measured loss. Switching to OT reduces mean MSE by approximately 0.025 on CIFAR-10, 0.040–0.043 on CelebA-HQ, and 0.037 on FFHQ. Removing time conditioning increases mean MSE by only 0.00027–0.00039, 0.0079–0.0110, and 0.0032, respectively. These differences are consistent with a small cost of hiding time (Section 4.2), although measured losses also include learnable approximation error and do not isolate the irreducible terms.

Removing time conditioning improves mean FID across the evaluated settings. Time-blind models achieve lower mean FID than their time-conditioned counterparts in all six dataset–coupling comparisons, despite slightly higher flow-matching MSE. The effect of OT coupling on FID is not uniformly beneficial: on CelebA-HQ, it worsens FID for the time-blind model. We therefore distinguish the larger coupling effect on flow-matching loss from the sample-quality benefit of removing time conditioning (Table 1, Appendix C.2). Our loss decomposition does not explain the FID improvement.

5.2 Validating Time Identifiability on Real Data

Theorem 2 predicts that the interpolation time t is recoverable from z alone at rate $O(1/\sqrt{d-k})$, with errors approximately Gaussian. We construct the noise-subspace estimator $\hat{t}(z)$ from the proof and compare its empirical error distribution against the theoretical Gaussian on the synthetic spiked covariance model and on each of our three real datasets, fitting the spiked covariance via PCA at the 95% explained-variance threshold for the theoretical prediction.

Fitting the spiked covariance to real data. Given a dataset we compute the sample covariance $\hat{\Sigma}$ and eigendecompose it as $\hat{\Sigma} = V\,\mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_d)\,V^\top$ with eigenvalues sorted descending. At a chosen rank k we set the signal subspace U_x to the top-k eigenvectors, the isotropic floor $\sigma^2 = \frac{1}{d-k}\sum_{j>k}\hat{\lambda}_j$ to the mean of the trailing eigenvalues, and the spike eigenvalues $\Lambda_i = \max(\hat{\lambda}_i - \sigma^2,\, 0)$

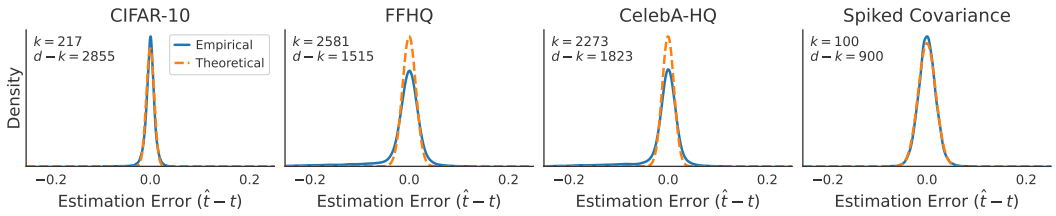


Figure 4: **Empirical estimates of t on real datasets are precise and match the theoretically predicted bound.** Distribution of estimation error $\hat{t} - t$ for the noise-subspace estimator across four settings: a synthetic spiked covariance model, CIFAR-10 (pixel space), FFHQ (VAE latent), and CelebA-HQ (VAE latent), with k chosen at the 95% explained-variance threshold per dataset. Empirical (solid) closely matches the theoretical Gaussian prediction (dashed) from the spiked covariance bound (Theorem 2).

for $i \leq k$, so that $U_x \Lambda U_x^\top + \sigma^2 I_d$ recovers the empirical top-k spectrum and collapses the residual tail into a single isotropic term. These parameters enter only the theoretical Gaussian prediction; the data x in $z = (1-t)\varepsilon + tx$ are actual samples drawn from the dataset, not from this Gaussian.

The experiments yield three observations (details in Appendix C.3):

Predictions are tight in absolute terms. With $d - k$ in the thousands, the empirical estimates track closely around the true time on every panel. The empirical histograms are sharply peaked at zero and the spread shrinks as $d - k$ grows, consistent with the $O(1/\sqrt{d-k})$ rate predicted by Theorem 2.

The closed-form estimator achieves the predicted rate on both synthetic and real data. Under the spiked Gaussian where Theorem 2's assumptions hold exactly, the empirical histogram of $\hat{t} - t$ tracks the theoretical Gaussian closely. The same prediction transfers to real, non-Gaussian data: fitting the spiked model via PCA yields a theoretical Gaussian that matches the empirical error distribution on CIFAR-10, FFHQ, and CelebA-HQ. For these real datasets, this agreement is empirical evidence rather than a guarantee: the theorem assumes Gaussian data and a known invertible clock branch. The residual dimension remains in the thousands even for the flatter VAE-latent spectra (Figure 3).

Approximate invertibility is not a major concern in practice. Because t^* sits close to 1 on every dataset, the descending-branch inverse covers nearly the full interpolation path; even when t is sampled uniformly over $[0, 1]$, the mean absolute error remains small on every dataset (0.006 on CIFAR-10, 0.021 on FFHQ, 0.018 on CelebA-HQ; Appendix C.4).

5.3 Time Information in Learned Representations

We probe frozen intermediate features of trained time-blind models to test whether they retain time information and which input directions support its decoding. Across CIFAR-10, CelebA-HQ, and FFHQ, we find:

A small MLP probe accurately estimates time across models and datasets. We train a three-layer MLP to predict t from frozen model features. Pearson correlations are 0.998 on CIFAR-10 and 0.999 on CelebA-HQ and FFHQ, with mean absolute errors 0.013, 0.010, and 0.010, respectively. Thus, time remains readily decodable from the learned representations despite the absence of explicit time conditioning.

Residual-subspace corruption disrupts time decoding more than signal-subspace corruption. We form a second interpolant at an independently sampled time t', holding (ε, x) fixed, and replace progressively larger subspaces of z_t with components of $z_{t'}$. Replacing the lowest-eigenvalue PCA directions first degrades probe accuracy earlier than replacing the highest-eigenvalue directions first on all three datasets (Figure 5). At 40% corruption, residual-direction replacement also degrades decoding more than a random subspace of equal dimension. Residual corruption increases velocity-prediction loss as well (Appendix D).

These diagnostics show decodable time information and sensitivity to residual perturbations, not that the velocity predictor uses a residual-energy clock. The interventions introduce distribution shift and are not norm-matched. Appendix D provides the architecture, training protocol, controls, and detailed results.

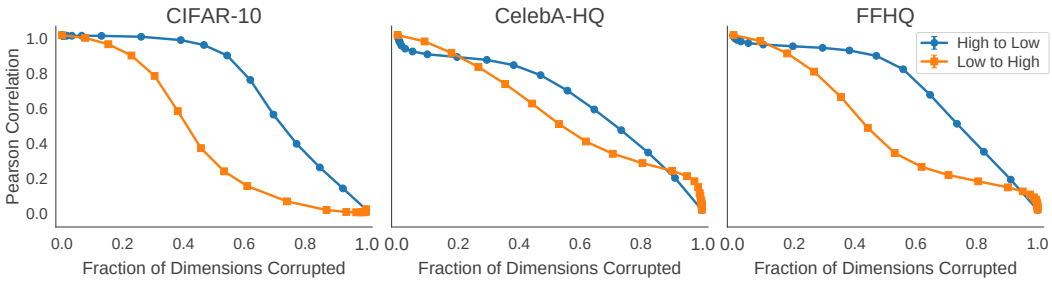


Figure 5: **Time decoding is more sensitive to residual-first than signal-first corruption.** Pearson correlation between probe-predicted and true time as an increasing fraction of input PCA directions is replaced by a wrong-time interpolant sharing the same source–target pair. Low-to-high eigenvalue order replaces residual directions first; high-to-low replaces signal directions first. These curves measure auxiliary-probe sensitivity, not whether the velocity predictor uses the proposed clock.

5.4 Closed-Form Validation of the Variance Decomposition

To isolate the dimensional scaling predicted by our theory from confounds of training and finite-sample noise, we evaluate the three loss terms in an idealized spiked-covariance setting where each admits a closed-form or low-variance numerical estimator. Our analysis makes three sharp predictions about how the ambient dimension d and signal dimension k shape the three terms of Proposition 1. We test all three on the spiked Gaussian (details in Appendix C.5). Figure 6 reports each along a complementary slice of the (d, k) grid.

Shrinking the noise subspace inflates the gap. At fixed d, growing k shrinks the $(d-k)$-dimensional noise subspace through which t leaks into z, so the time-blindness gap should increase. At $d = 1024$, sweeping k increases the gap (Figure 6, left).

Enlarging the ambient space collapses the gap. At fixed k, growing d enlarges the noise subspace and Theorem 3 predicts the ratio $\Delta/\mathcal{C}_\pi$ scales as $1/(d-k)$. At $k = 64$, sweeping d drives the gap toward zero while Terms I and II stay bounded (Figure 6, right).

Coupling variance dominates the gap across the plane. On the bulk of the (d, k) plane (anywhere off the $k \approx d$ diagonal), the coupling variance should dwarf the gap. The heatmap (center) is light everywhere except along the diagonal, and in both slice plots the coupling curve sits well above the gap throughout.

6 Related Work

Noise-level estimation and blind denoising. Recovering the noise level from a corrupted observation is a classical problem in image processing, with estimators based on local image statistics, patch structure, and low-rank covariance structure [27, 19, 5]. Modern blind denoisers instead learn this behavior from data, including convolutional, self-supervised, and real-image blind-denoising methods [10, 2, 32]. These works motivate the possibility that an explicit noise or time label may be redundant when the corrupted sample itself contains enough information to infer it.

Blind generative models and intrinsic dimension. Recent work has extended this question to generative modeling. Kadkhodaie et al. [13] show that blind denoising diffusion models can follow an implicit noise schedule under low intrinsic dimension, building on the view that natural images concentrate near low-dimensional geometric structure [11, 23, 4]. Related perspectives connect denoising to score estimation and empirical Bayes identities [21, 8, 24], and to autonomous or time-unconditional generative dynamics [26, 31]. Our setting differs from these works because flow matching contains an additional ambiguity: even if time is known, the target velocity remains random through the coupling between source and target samples.

Noise conditioning, residual dimension, and coupling in flow matching. Closest to our motivation, Sun et al. [28] analyze why noise-unconditional denoising and flow-based generative models

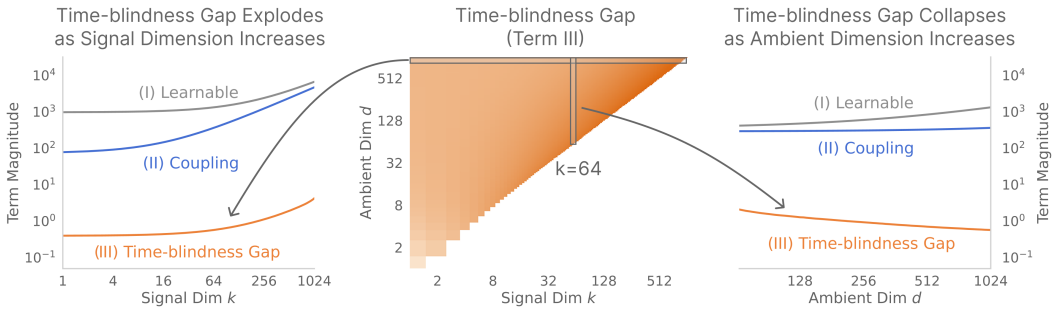


Figure 6: **The time-blindness gap is governed by $d - k$ and dwarfed by the coupling variance.** (Left) Growing k at fixed $d = 1024$ inflates both the gap and the coupling loss. (Center) Heatmap of the gap magnitude across the (d, k) plane. (Right) Growing d at fixed $k = 64$ collapses the gap.

can remain effective, arguing that the posterior over the hidden noise level concentrates in high ambient dimension. We refine this dimensional picture for low-rank data in flow matching: the identifiable clock lives in the residual noise subspace orthogonal to the data, so the relevant dimension is $d - k$, not d alone. This distinction matters when the intrinsic dimension grows with the ambient dimension. In parallel, prior work has shown that flow-matching performance depends strongly on the source–target coupling, with improved couplings based on rectified flows, optimal transport, and minibatch optimal transport [18, 29, 9]. Our contribution is to disentangle these effects: we compare the time-blindness gap against the irreducible coupling variance and show why coupling can remain the dominant lever for flow-matching loss even when time is hidden.

7 Discussion

Our results explain how time can remain identifiable without explicit conditioning. Coupling has the larger effect on measured flow-matching loss, while dropping time consistently improves mean FID in our experiments; explaining the latter remains open.

Scope and limitations. The loss decomposition is exact for square-integrable targets. The identifiability and relative-bound results require the stated Gaussian model, independent coupling, linear paths, and a known clock branch away from its critical point. Anisotropic Gaussian concentration is treated through effective rank in Appendix B.4. Nonlinear schedules, non-Gaussian sources, and relative guarantees under near-OT couplings remain open; experiments on real data do not establish those extensions.

Acknowledgments

Supported in part by an NSF Graduate Research Fellowship, NSF #2403297 and #2502793, and gifts from Google, Amazon, Meta, NVIDIA, Avast, Fiddler Labs, and Bosch.

References

[1] Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants, 2023. URL https://arxiv.org/abs/2209.15571.

[2] Joshua Batson and Loic Royer. Noise2Self: Blind denoising by self-supervision. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 524–533. PMLR, 2019. URL https://proceedings.mlr.press/v97/batson19a.html.

[3] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991. doi: https://doi.org/10.1002/cpa.3160440402. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.3160440402.

[4] Bradley C. A. Brown, Anthony L. Caterini, Brendan Leigh Ross, Jesse C. Cresswell, and Gabriel Loaiza-Ganem. Verifying the union of manifolds hypothesis for image data. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Rvee9CAX4fi.

[5] Guangyong Chen, Fengyuan Zhu, and Pheng Ann Heng. An efficient statistical method for image noise level estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 477–485, 2015.

[6] Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. URL https://arxiv.org/abs/2302.07194.

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

[8] Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. doi: 10.1198/jasa.2011.tm11181.

[9] Kilian Fatras, Thibault Sejourne, Nicolas Courty, and Rémi Flamary. Unbalanced minibatch optimal transport; applications to domain adaptation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3186–3197. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/fatras21a.html.

[10] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1712–1722, 2019.

[11] Olivier J. Hénaff, Johannes Ballé, Neil C. Rabinowitz, and Eero P. Simoncelli. The local low-dimensionality of natural images. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6626.

[12] Iain M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, 2001. doi: 10.1214/aos/1009210544.

[13] Zahra Kadkhodaie, Aram-Alexandre Pooladian, Sinho Chewi, and Eero Simoncelli. Blind denoising diffusion models and the blessings of dimensionality, 2026. URL https://arxiv.org/abs/2602.09639.

[14] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation, 2018. URL https://arxiv.org/abs/1710.10196.

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019. URL https://arxiv.org/abs/1812.04948.

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

[17] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

[18] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. URL https://arxiv.org/abs/2209.03003.

[19] Xinhao Liu, Masayuki Tanaka, and Masatoshi Okutomi. Single-image noise level estimation for blind denoising. *IEEE Transactions on Image Processing*, 22(12):5226–5237, 2013. doi: 10.1109/TIP.2013.2283400.

[20] Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers, 2024. URL https://arxiv.org/abs/2401.08740.

[21] Koichi Miyasawa. An empirical bayes estimator of the mean of a normal population. *Bulletin of the International Statistical Institute*, 38(4):181–188, 1961.

[22] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11410–11420, 2022.

[23] Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=XJk19XzGq2J.

[24] Martin Raphan and Eero P. Simoncelli. Least squares estimation without priors or supervision. *Neural Computation*, 23(2):374–420, 2011. doi: 10.1162/NECO_a_00076.

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

[26] Mojtaba Sahraee-Ardakan, Mauricio Delbracio, and Peyman Milanfar. The geometry of noise: Why diffusion models don't need noise conditioning, 2026. URL https://arxiv.org/abs/2602.18428.

[27] Dong-Hyuk Shin, Rae-Hong Park, Seungjoon Yang, and Jae-Han Jung. Block-based noise estimation using adaptive Gaussian filtering. *IEEE Transactions on Consumer Electronics*, 51 (1):218–226, 2005.

[28] Qiao Sun, Zhicheng Jiang, Hanhong Zhao, and Kaiming He. Is noise conditioning necessary for denoising generative models?, 2025. URL https://arxiv.org/abs/2502.13129.

[29] Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport, 2024. URL https://arxiv.org/abs/2302.00482.

[30] Binxu Wang and John J. Vastola. The unreasonable effectiveness of gaussian score approximation for diffusion models and its applications, 2024. URL https://arxiv.org/abs/2412.09726.

[31] Runqian Wang and Yilun Du. Equilibrium matching: Generative modeling with implicit energy-based models, 2025. URL https://arxiv.org/abs/2510.02300.

[32] Kai Zhang, Yawei Li, Jingyun Liang, Jiezhang Cao, Yulun Zhang, Hao Tang, Deng-Ping Fan, Radu Timofte, and Luc Van Gool. Practical blind image denoising via swin-conv-unet and data synthesis. *Machine Intelligence Research*, 20(6):822–836, 2023. doi: 10.1007/s11633-023-1466-0.

Appendix Contents

A Proofs for Section 3: Isolating Sources of Error **14**

A.1 Anatomy of the Time-Blind Loss 14

A.2 Time-Blind Flow Matching is at Least as Hard as Flow Matching 14

A.3 Relating Irreducible Variance and the Straightness of Paths 15

B Proofs for Section 4: Closing the Two Sources of Irreducible Error **15**

B.1 Baseline Spiked Model and Notation 15

B.2 Identifiability under Spiked Covariance 16

B.3 Closed-Form Inversion of the Variance Clock 17

B.4 Effective-Rank Generalization 18

B.5 Absolute Bound on the Time-Blindness Gap 20

B.6 Independent Coupling Has Extensive Variance 22

B.7 Relative Cost of Time-Independence 23

C Experimental Details **23**

C.1 Datasets and Architectures 23

C.2 Loss Decomposition Validation 23

C.3 Empirical Validation of Time Identifiability 25

C.4 Time-Localized Estimation Error and the Variance-Clock Critical Point 28

C.5 Closed-Form Variance Decomposition on the Spiked Gaussian 29

D Probe Subspace-Corruption Diagnostic **31**

A Proofs for Section 3: Isolating Sources of Error

A.1 Anatomy of the Time-Blind Loss

To isolate the effect of the coupling cleanly, we analyze the time-blind loss over all vector fields $f : \mathbb{R}^d \to \mathbb{R}^d$. We combine the standard bias–variance decomposition with the law of total variance to obtain a single three-term decomposition.

Proposition 1 (Anatomy of the Time-Blind Loss). *For any predictor $f : \mathbb{R}^d \to \mathbb{R}^d$, the time-blind loss decomposes as*

$$\mathcal{L}_{\mathrm{TB}}(f) \;=\; \underbrace{\mathbb{E}\big[\|f(z) - \mathbb{E}[u \mid z]\|^2\big]}_{\textit{I. Learnable error}} \;+\; \underbrace{\mathbb{E}\big[\mathrm{Var}(u \mid z, t)\big]}_{\textit{II. Coupling variance } \mathcal{C}_\pi} \;+\; \underbrace{\mathbb{E}\big[\mathrm{Var}\big(\mathbb{E}[u \mid z, t] \,\big|\, z\big)\big]}_{\textit{III. Time-Blindness Gap } \Delta}. \tag{4}$$

(Proof in Appendix A.1.)

Proof. **Step 1: Bias–variance decomposition.** By the standard bias–variance decomposition (conditioning on z), the cross term $\mathbb{E}[(f(z) - \mathbb{E}[u \mid z])^\top (\mathbb{E}[u \mid z] - u) \mid z] = 0$ vanishes, giving

$$\mathcal{L}_{\mathrm{TB}}(f) = \underbrace{\mathbb{E}[\|f(z) - \mathbb{E}[u \mid z]\|^2]}_{\text{Reducible (bias)}} + \underbrace{\mathbb{E}[\mathrm{Var}(u \mid z)]}_{\text{Irreducible (variance)}}. \tag{7}$$

Step 2: Law of total variance. We now split the irreducible term $\mathbb{E}[\mathrm{Var}(u \mid z)]$ by conditioning further on t. By the law of total variance,

$$\mathrm{Var}(u \mid z) = \mathbb{E}[\mathrm{Var}(u \mid z, t) \mid z] + \mathrm{Var}(\mathbb{E}[u \mid z, t] \mid z).$$

Taking the outer expectation over z:

$$\underbrace{\mathbb{E}[\mathrm{Var}(u \mid z)]}_{\text{Irreducible variance}} = \underbrace{\mathbb{E}[\mathrm{Var}(u \mid z, t)]}_{\text{Coupling variance } \mathcal{C}_\pi} + \underbrace{\mathbb{E}[\mathrm{Var}(\mathbb{E}[u \mid z, t] \mid z)]}_{\text{Time-Blindness Gap } \Delta}.$$

Since $u = x - \varepsilon$, conditioning on (z, t) is equivalent to FM conditioning on (z_t, t) with the same target, so $\mathcal{C}_\pi = \mathbb{E}[\mathrm{Var}(u_t \mid z_t, t)]$. Substituting into (7) gives the three-term decomposition. □

For use throughout the appendix, write

$$\Delta \;:=\; \mathbb{E}[\mathrm{Var}(\mathbb{E}[u \mid z, t] \mid z)], \qquad \mathcal{C}_\pi \;:=\; \mathbb{E}[\mathrm{Var}(u \mid z, t)],$$

matching the underbrace labels above. Step 1 shows that the irreducible variance $\mathbb{E}[\mathrm{Var}(u \mid z)] = \mathcal{C}_\pi + \Delta$ depends only on the coupling π, so the only way to reduce it is to change the coupling. Step 2 further splits this into $\mathcal{C}_\pi$ and the time-blindness gap Δ. The next subsection formalizes the comparison with flow matching.

A.2 Time-Blind Flow Matching is at Least as Hard as Flow Matching

The FM loss admits an analogous bias–variance decomposition. The coupling variance term (II) in Proposition 1 corresponds exactly to the FM irreducible error floor.

Lemma A.1 (Bias–Variance Decomposition of FM Loss). *Let $z_t = (1 - t)\varepsilon + tx$ and $u_t = x - \varepsilon$ under a fixed joint distribution of (ε, x), and let $v : \mathbb{R}^d \times [0, 1] \to \mathbb{R}^d$ be any predictor. Then the FM loss decomposes as*

$$\mathcal{L}_{\mathrm{FM}}(v) := \mathbb{E}[\|v(z_t, t) - u_t\|^2] = \underbrace{\mathbb{E}\big[\|v(z_t, t) - \mathbb{E}[u_t \mid z_t, t]\|^2\big]}_{\textit{Reducible (bias)}} + \underbrace{\mathbb{E}[\mathrm{Var}(u_t \mid z_t, t)]}_{\textit{Irreducible (variance)}}.$$

In particular, $\min_v \mathcal{L}_{\mathrm{FM}}(v) = \mathbb{E}[\mathrm{Var}(u_t \mid z_t, t)]$*, achieved by* $v^*(z_t, t) = \mathbb{E}[u_t \mid z_t, t]$.

Proof. The proof is identical to Step 1 of Proposition 1, with conditioning on (z_t, t) instead of z alone. □

Combining the FM floor with the three-term decomposition immediately gives:

Corollary A.2 (Time-blind irreducible error $\geq$ FM). *The irreducible time-blind error is at least as large as the FM irreducible error:*

$$\min_f \mathcal{L}_{\mathrm{TB}}(f) \;=\; \mathcal{C}_\pi + \Delta \;\geq\; \underbrace{\mathcal{C}_\pi}_{\min_v \mathcal{L}_{\mathrm{FM}}(v)} .$$

The gap $\Delta \geq 0$ is the cost of time-independence: it vanishes exactly when the time-aware optimal velocity $\mathbb{E}[u \mid z, t]$ is, almost surely, already determined by z.

When the data distribution has low intrinsic dimensionality k relative to the ambient dimension d, Δ is bounded by $O(1/(d-k))$ times $\mathcal{C}_\pi$; we formalize this in Appendix B.7 (Theorem B.5). Crucially, this geometric mitigation leaves $\mathcal{C}_\pi$ unaddressed: it remains an independent, controllable source of difficulty that motivates the OT coupling strategy developed in this work.

A.3 Relating Irreducible Variance and the Straightness of Paths

Definition 1 (Straightness [18]). *Let z_t be the trajectory of an ODE with velocity field v and endpoints $(\varepsilon, x) \sim \pi$. The* straightness defect *is*

$$S \;:=\; \int_0^1 \mathbb{E}\big[\|(x-\varepsilon) - v(z_t, t)\|^2\big]\, dt. \tag{8}$$

$S = 0$ iff all trajectories are constant-velocity straight lines, in which case single-step Euler integration is exact.

Applied to the optimal time-blind velocity $f^*(z) = \mathbb{E}[u \mid z]$, the straightness defect equals the irreducible variance:

$$S_{\mathrm{auto}} \;:=\; \int_0^1 \mathbb{E}\big[\|(x-\varepsilon) - f^*(z_t)\|^2\big]\, dt \;=\; \mathbb{E}\big[\mathrm{Var}(u \mid z)\big], \tag{9}$$

since $f^*(z_t) = \mathbb{E}[u \mid z_t]$ makes the integrand independent of t. By the law of total variance (Proposition 1),

$$S_{\mathrm{auto}} \;=\; \underbrace{\mathcal{C}_\pi}_{S_{\mathrm{FM}}} + \Delta, \tag{10}$$

where $S_{\mathrm{FM}} = \int_0^1 \mathbb{E}[\|(x-\varepsilon) - v^*(z_t, t)\|^2]\, dt$ is the straightness defect of Liu et al. [18] applied to the optimal FM velocity, and Δ is the time-blindness gap.

B Proofs for Section 4: Closing the Two Sources of Irreducible Error

Proposition 1 establishes that the irreducible error of time-blind flow matching is strictly higher than that of FM, raising the natural question of why time-blind models work at all in practice. Extending the analysis of Kadkhodaie et al. [13] from the diffusion setting to linear interpolants, we resolve this puzzle by showing that the interpolation time t is almost entirely determined by the observation z alone, so conditioning on t provides little additional information and the gap $\Delta = \mathbb{E}[\mathrm{Var}(\mathbb{E}[u \mid z, t] \mid z)]$ becomes negligible relative to the irreducible coupling variance $\mathcal{C}_\pi$.

Mechanism: concentration reveals time. The observation $z_t = (1-t)\varepsilon + t\,x$ mixes noise and signal with weights determined by t. In high dimensions, simple statistics of z_t (such as its projected norm) concentrate sharply around deterministic functions of t. If this concentration is sufficiently strong, t can be recovered from z alone, so conditioning on t provides little additional information.

B.1 Baseline Spiked Model and Notation

The headline chain of this appendix, spiked identifiability (Subsection B.2) $\Rightarrow$ absolute gap bound (Subsection B.5) $\Rightarrow$ relative cost (Subsection B.7), is built on a single baseline regime that we collect once here for reuse.

Spiked Gaussian model. The data follow a spiked Gaussian:

$$x \sim \mathcal{N}(0, \Sigma), \qquad \Sigma = U_x \Lambda U_x^\top + \sigma^2 I_d, \tag{11}$$

where $U_x \in \mathbb{R}^{d\times k}$ spans the k-dimensional signal subspace, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ collects the signal variances with $\lambda_{\max} := \max_i \lambda_i < \infty$, and $\sigma^2 > 0$ gives isotropic residual variance in the orthogonal complement. This is a standard tractable proxy for the broader empirical observation that natural high-dimensional data concentrate near a low-dimensional subspace, used throughout high-dimensional statistics [12] and the theoretical analysis of diffusion models [6, 30]. Because all results are rotationally invariant, we may take Σ diagonal in its eigenbasis with per-coordinate variances $S_i = \lambda_i$ for $i \le k$ and $S_i = \sigma^2$ for $i > k$.

Linear interpolant. We observe

$$z_t = (1-t)\,\varepsilon + t\,x, \qquad \varepsilon \sim \mathcal{N}(0, I_d), \qquad t \sim \mathrm{Unif}(I), \tag{12}$$

where $I \subset (0,1)$ is a closed interval avoiding $t^* = 1/(1+\sigma^2)$.

Noise-subspace statistic. Let $P_\perp \in \mathbb{R}^{d\times d}$ project onto the $(d-k)$-dimensional orthogonal complement of the signal subspace. The key statistic is

$$\widehat{\sigma}^2_\perp(z) \;:=\; \frac{1}{d-k}\,\|P_\perp z\|^2, \tag{13}$$

which estimates the deterministic noise-subspace variance

$$\sigma^2_\perp(t) \;:=\; (1-t)^2 + t^2\sigma^2. \tag{14}$$

We also write $r_i(t) = (1-t)^2 + t^2 S_i$ for the per-coordinate interpolant variance and $\alpha_i(t) := (tS_i - (1-t))/r_i(t)$ for the per-coordinate optimal slope.

Noise-subspace time estimator. Inverting the variance clock $\sigma^2_\perp$ yields the noise-subspace time estimator used throughout the appendix:

$$\hat{t}(z) \;:=\; \big(\sigma^2_\perp\big)^{-1}\Big(\Pi_{\sigma^2_\perp(I)}\big(\widehat{\sigma}^2_\perp(z)\big)\Big), \tag{15}$$

where $\Pi_{\sigma^2_\perp(I)}$ denotes projection onto the closed interval $\sigma^2_\perp(I) \subset \mathbb{R}$. The projection ensures $\hat{t}(z) \in I$ even when $\widehat{\sigma}^2_\perp(z)$ leaves the range of $\sigma^2_\perp$ on I (a positive-probability event, since the underlying chi-square statistic is unbounded), and only sharpens the bounds below since projection cannot increase distance to any point in the interval.

Constants convention. All implicit constants in the $O(\cdot)$ statements that follow depend only on the time interval I, the noise floor σ^2, and the bounded signal spectrum $\lambda_{\max}$, not on d or k.

B.2 Identifiability under Spiked Covariance

This appendix restates Theorem 2 from the main text in full formal detail and provides the complete proof. Working in the baseline spiked Gaussian model (Subsection B.1), the concentration argument here takes a particularly clean form: projecting onto the noise subspace yields an isotropic Gaussian, so the projected norm follows a standard χ^2_{d-k} distribution (up to scaling), and the estimation rate scales with the noise-subspace dimension $d-k$.

Theorem 2 (Restated). *Under the baseline spiked model of Subsection B.1, the estimator*

$$\hat{t}(z) \;=\; (\sigma^2_\perp)^{-1}\big(\widehat{\sigma}^2_\perp(z)\big)$$

satisfies $\hat{t}(z) = t + O_p\big(1/\sqrt{d-k}\big)$ *for* t *bounded away from* $t^* = 1/(1+\sigma^2)$.

Proof. Write

$$\sigma^2_\perp(t) := (1-t)^2 + t^2\sigma^2.$$

We show that the projected noise-subspace variance acts as a high-dimensional clock for the interpolation time. The proof has three steps: projection, concentration, and inversion.

Step 1: Projecting onto the noise subspace produces an isotropic Gaussian clock. Let $P_\perp$ be the orthogonal projector onto the complement of the k-dimensional signal subspace. Projecting the interpolant gives

$$P_\perp z_t \;=\; (1-t)\,P_\perp \varepsilon \;+\; t\,P_\perp x.$$

Under the spiked covariance model,

$$P_\perp \varepsilon \sim \mathcal{N}(0, I_{d-k}), \qquad P_\perp x \sim \mathcal{N}(0, \sigma^2 I_{d-k}),$$

and these two projected terms are independent. Therefore, conditional on t,

$$P_\perp z_t \sim \mathcal{N}\big(0,\, \sigma_\perp^2(t)\, I_{d-k}\big).$$

Thus

$$\widehat{\sigma}_\perp^2(z) \;:=\; \frac{1}{d-k}\,\|P_\perp z\|^2$$

is the empirical second moment of $d-k$ independent Gaussian coordinates, each with variance $\sigma_\perp^2(t)$. This projected variance is the statistic from which we recover time.

Step 2: The projected empirical variance concentrates around $\sigma_\perp^2(t)$. From the Gaussian representation above,

$$\widehat{\sigma}_\perp^2(z) \;\overset{d}{=}\; \sigma_\perp^2(t)\,\frac{Y}{d-k}, \qquad Y \sim \chi^2_{d-k}.$$

By standard chi-square concentration,

$$\frac{Y}{d-k} \;=\; 1 \;+\; O_p\!\left(\frac{1}{\sqrt{d-k}}\right).$$

Therefore,

$$\widehat{\sigma}_\perp^2(z) \;=\; \sigma_\perp^2(t)\left(1 + O_p\!\left(\frac{1}{\sqrt{d-k}}\right)\right).$$

Since $\sigma_\perp^2(t)$ is bounded on any compact interval $I \subset (0,1)$, this gives

$$\widehat{\sigma}_\perp^2(z) \;=\; \sigma_\perp^2(t) \;+\; O_p\!\left(\frac{1}{\sqrt{d-k}}\right).$$

Step 3: Inverting the variance clock transfers the rate to t. The derivative of the variance curve is

$$\frac{d}{dt}\sigma_\perp^2(t) \;=\; -2(1-t) + 2t\sigma^2,$$

and vanishes only at $t^* = 1/(1+\sigma^2)$. Since I is closed and $t^* \notin I$, $|(\sigma_\perp^2)'(t)|$ is bounded below on I, so $\sigma_\perp^2$ is monotone there and its inverse is Lipschitz on $\sigma_\perp^2(I)$. Define

$$\hat{t}(z) \;:=\; (\sigma_\perp^2)^{-1}\big(\widehat{\sigma}_\perp^2(z)\big).$$

Since $(\sigma_\perp^2)^{-1}\big(\sigma_\perp^2(t)\big) = t$, Lipschitzness of the inverse gives

$$|\hat{t}(z) - t| \;\le\; C\,\big|\widehat{\sigma}_\perp^2(z) - \sigma_\perp^2(t)\big|.$$

Using Step 2,

$$\hat{t}(z) - t \;=\; O_p\!\left(\frac{1}{\sqrt{d-k}}\right).$$

Equivalently,

$$\hat{t}(z) \;=\; t \;+\; O_p\!\left(\frac{1}{\sqrt{d-k}}\right).$$

This proves the claimed identifiability rate. □

B.3 Closed-Form Inversion of the Variance Clock

The proof of Theorem 2 characterizes the estimator $\hat{t}(z) = (\sigma_\perp^2)^{-1}(\hat{\sigma}_\perp^2)$ abstractly via Lipschitzness of the inverse on a compact interval avoiding t^*. For the linear interpolant the inversion has a closed form, which is the formula evaluated in all experiments.

Inverting the variance clock. Expanding $(1-t)^2$, collecting like powers, and applying the quadratic formula:

$$\begin{aligned}\sigma_\perp^2(t) &= (1-t)^2 + t^2\sigma^2 \\ &= 1 - 2t + t^2 + t^2\sigma^2 \\ &= (1+\sigma^2)\,t^2 - 2t + 1,\end{aligned}$$

so setting $\sigma_\perp^2(t) = \hat{\sigma}_\perp^2$ and moving everything to one side gives the standard form $(1+\sigma^2)\,t^2 - 2t + (1-\hat{\sigma}_\perp^2) = 0$, whose two roots are

$$t_\pm \;=\; \frac{2 \pm \sqrt{4 - 4(1+\sigma^2)(1-\hat{\sigma}_\perp^2)}}{2(1+\sigma^2)}. \tag{16}$$

Branch selection. The clock $\sigma_\perp^2(t)$ is a strictly convex parabola attaining its global minimum

$$\sigma_\perp^2(t^*) \;=\; \frac{\sigma^2}{1+\sigma^2} \qquad \text{at} \qquad t^* \;=\; \frac{1}{1+\sigma^2}.$$

Each value $\hat{\sigma}_\perp^2 \in (\sigma^2/(1+\sigma^2),\, 1]$ has two preimages: t_- on the descending branch $[0, t^*]$ and t_+ on the ascending branch $[t^*, 1]$. On any working interval $I \subset [0, t^*]$ (the regime used throughout our experiments, where $\sigma^2 \ll 1$ pushes t^* close to 1) the inverse selects the descending branch, equivalently the minus sign:

$$\hat{t}(z) \;=\; \frac{2 - \sqrt{4 - 4(1+\sigma^2)(1-\hat{\sigma}_\perp^2)}}{2(1+\sigma^2)}. \tag{17}$$

This is the closed-form realization of the inverse $(\sigma_\perp^2)^{-1}$ used in Step 3 of the proof of Theorem 2.

B.4 Effective-Rank Generalization

The spiked proof above relies on a specific structural fact: projecting onto the noise subspace yields an isotropic Gaussian, so the projected norm $\|z_\perp\|^2$ is a standard χ^2_{d-k} up to the scalar $\sigma_\perp^2(t)$. Real data covariances rarely exhibit such a sharp signal/noise split; e.g. the latent covariances of trained VAEs typically decay smoothly over several orders of magnitude. We now show that the same project–concentrate–invert argument carries through for arbitrary PSD Σ and any orthogonal projector P, with the isotropic χ^2_{d-k} replaced by a weighted chi-square whose effective sample size is the *effective rank* of the projected interpolant covariance.

Effective dimension of the observation. The strength of the concentration argument depends on how many independent directions contribute to the statistic. In the isotropic setting of Appendix B.2, concentration improves with the ambient dimension of the projected subspace, yielding the familiar $1/\sqrt{n}$ scaling for an n-dimensional average. In the anisotropic setting, however, different directions contribute unequally, and the relevant notion of dimension is no longer the ambient n but the *effective rank* of the underlying covariance: for a PSD matrix $A \in \mathbb{R}^{n\times n}$ with eigenvalues $\{\lambda_i\}$,

$$r_{\text{eff}}(A) \;:=\; \frac{(\operatorname{tr} A)^2}{\operatorname{tr}(A^2)} \;=\; \frac{\left(\sum_i \lambda_i\right)^2}{\sum_i \lambda_i^2}.$$

Unlike the algebraic rank, the effective rank quantifies how evenly variance is distributed across eigen-directions. When all eigenvalues are equal, $r_{\text{eff}}(A) = n$; in highly anisotropic regimes it can be much smaller, reflecting a collapse onto a few dominant directions. It therefore plays the role of an intrinsic statistical dimension governing concentration rates, replacing the ambient dimension in heterogeneous settings.

Setup. Let $x \sim \mathcal{N}(0, \Sigma)$ with $\Sigma \in \mathbb{R}^{d\times d}$ an arbitrary PSD covariance, $z_t = (1-t)\varepsilon + tx$ with $\varepsilon \sim \mathcal{N}(0, I_d)$, and let P be an orthogonal projector of rank m with orthonormal basis $Q \in \mathbb{R}^{d\times m}$. Write $\Sigma_P := Q^\top \Sigma Q$ for the induced covariance on $\operatorname{range}(P)$, $\bar{\mu}_P := \operatorname{tr}(\Sigma_P)/m$ for its spectral mean, and $\Sigma_{P,t} := (1-t)^2 I_m + t^2 \Sigma_P$ for the projected interpolant covariance, whose effective rank we denote $r_{\text{eff}}(P,t) := r_{\text{eff}}(\Sigma_{P,t})$. As in the spiked case, the variance-to-time map $t \mapsto \sigma_t^2(\bar{\mu}_P) := (1-t)^2 + t^2\bar{\mu}_P$ has a unique critical point $t^*_{\bar{\mu}_P} := 1/(1+\bar{\mu}_P)$.

Proposition B.1 (Effective-rank identifiability). *Under the setup above, on any closed interval avoiding $t^*_{\bar{\mu}_P}$, the estimator $\hat{t}(z)$ obtained by inverting $\sigma_t^2(\bar{\mu}_P) = \|Pz\|^2/m$ on the relevant branch satisfies*

$$\hat{t}(z) \;=\; t \;+\; O_p\Big(r_{\text{eff}}(P,t)^{-1/2}\Big)\,.$$

Proof. The argument follows the same project–concentrate–invert structure as Theorem 2; only the concentration step changes.

Step 1: Weighted-Gaussian projection. Since $P = QQ^\top$, $\|Pz_t\|^2 = \|Q^\top z_t\|^2$, and $Q^\top z_t = (1-t)Q^\top\varepsilon + t\,Q^\top x$ with $Q^\top\varepsilon \sim \mathcal{N}(0, I_m)$ and $Q^\top x \sim \mathcal{N}(0, \Sigma_P)$ independent. Rotating to the eigenbasis of Σ_P decomposes $Q^\top z_t$ into independent Gaussian coordinates with per-coordinate variance $\sigma_{t,i}^2 := (1-t)^2 + t^2\mu_i$, where $\{\mu_i\}_{i=1}^m$ are the eigenvalues of Σ_P. Writing $\chi^2_{1,i}$ for i.i.d. χ^2_1 variates,

$$\|Pz_t\|^2 \;\overset{d}{=}\; \sum_{i=1}^m \sigma_{t,i}^2\,\chi^2_{1,i}.$$

This is the spiked Step 1 with weights $\sigma_{t,i}^2$ in place of a common scalar $\sigma_\perp^2(t)$, and $\Sigma_{P,t}$ has eigenvalues exactly $\{\sigma_{t,i}^2\}_{i=1}^m$.

Step 2: Concentration via the effective-rank identity. Let $\hat{q} := \|Pz\|^2/m$. Using $\mathbb{E}[\chi^2_1] = 1$ and $\mathrm{Var}(\chi^2_1) = 2$,

$$\mathbb{E}[\hat{q}] \;=\; \tfrac{1}{m}\sum_i \sigma_{t,i}^2 \;=\; \sigma_t^2(\bar{\mu}_P), \qquad \mathrm{Var}(\hat{q}) \;=\; \frac{2}{m^2}\sum_i (\sigma_{t,i}^2)^2.$$

Recognising $\sum_i \sigma_{t,i}^2 = \operatorname{tr}\Sigma_{P,t} = m\,\sigma_t^2(\bar{\mu}_P)$ and $\sum_i(\sigma_{t,i}^2)^2 = \operatorname{tr}(\Sigma_{P,t}^2)$, the definition $r_{\text{eff}}(P,t) = (\operatorname{tr}\Sigma_{P,t})^2/\operatorname{tr}(\Sigma_{P,t}^2)$ gives $\sum_i(\sigma_{t,i}^2)^2 = m^2\sigma_t^2(\bar{\mu}_P)^2/r_{\text{eff}}(P,t)$, so

$$\mathrm{Var}(\hat{q}) \;=\; \frac{2\,\sigma_t^2(\bar{\mu}_P)^2}{r_{\text{eff}}(P,t)}.$$

The effective rank acts as an effective sample size for the heteroskedastic average $\hat{q}$: it equals m when all $\sigma_{t,i}^2$ are equal and degrades smoothly under spectral anisotropy. Chebyshev's inequality applied to $\hat{q}$ yields

$$\hat{q}(z) \;=\; \sigma_t^2(\bar{\mu}_P) \;+\; O_p\left(\frac{\sigma_t^2(\bar{\mu}_P)}{\sqrt{r_{\text{eff}}(P,t)}}\right).$$

This is the only step where the proof diverges from Theorem 2: the textbook χ^2_{d-k} rate $1/\sqrt{d-k}$ is replaced by the weighted-chi-square rate $1/\sqrt{r_{\text{eff}}(P,t)}$, with the latter reducing to the former when the weights are constant.

Step 3: Lipschitz inversion. The variance-to-time map $t \mapsto \sigma_t^2(\bar{\mu}_P)$ has derivative $-2(1-t)+2t\bar{\mu}_P$, vanishing only at $t^*_{\bar{\mu}_P}$. On any closed interval I with $t^*_{\bar{\mu}_P} \notin I$, $|(\sigma_t^2(\bar{\mu}_P))'|$ is bounded below, so the inverse is Lipschitz on $\sigma_t^2(\bar{\mu}_P)(I)$ with constant C depending only on I and $\bar{\mu}_P$. The argument from Step 3 of Theorem 2 then gives

$$|\hat{t}(z) - t| \;\le\; C\,\big|\hat{q}(z) - \sigma_t^2(\bar{\mu}_P)\big| \;=\; O_p\Big(r_{\text{eff}}(P,t)^{-1/2}\Big)\,,$$

which is the claimed rate. □

Corollary B.2 (Spiked recovery). *Theorem 2 is the case of Proposition B.1 with $\Sigma = U_x\Lambda U_x^\top + \sigma^2 I_d$ and $P = P_\perp$.*

Proof. Under the spiked covariance and $P = P_\perp$, $P_\perp U_x = 0$, so the induced covariance collapses to the isotropic residual $\Sigma_P = \sigma^2 I_{d-k}$. All eigenvalues μ_i equal σ^2, so $\bar{\mu}_P = \sigma^2$, $t^*_{\bar{\mu}_P} = 1/(1+\sigma^2) = t^*$, and $\Sigma_{P,t} = \sigma_\perp^2(t)\,I_{d-k}$. Constant eigenvalues give $r_{\text{eff}}(P,t) = (d-k)^2/(d-k) = d-k$, and the Proposition's rate becomes the $O_p(1/\sqrt{d-k})$ bound of Theorem 2. □

Scope of the spiked model in the rest of the paper. Proposition B.1 establishes that the identifiability mechanism is not specific to isotropic residuals; the spiked model provides the cleanest closed-form constants but no essential structure. The remainder of the paper retains the spiked model for two distinct reasons. First, the propagation from identifiability to the time-blindness gap Δ (Section B.5) requires closed-form per-coordinate fourth-moment calculations that the spiked structure makes available; the corresponding anisotropic calculation introduces additional weighted-moment bookkeeping without changing the qualitative conclusion that $\Delta/\mathcal{C}_\pi$ vanishes with the residual-subspace dimension. Second, the empirical predictor in Section 5.2 uses the spiked-derived $1/\sqrt{d-k}$ rate because it has a single interpretable parameter (k at a fixed explained-variance threshold) that we report transparently per dataset, and the empirical fit on VAE latents with visibly anisotropic spectra (Figure 4) shows the spiked prediction is already tight without the effective-rank refinement. Proposition B.1 is therefore best read as a robustness statement about the identifiability mechanism, not as a competing predictor or an empirical hypothesis.

B.5 Absolute Bound on the Time-Blindness Gap

Proposition B.3 (Absolute time-blindness gap bound). *Under the spiked covariance model with* $t \sim \mathrm{Unif}(I)$*,*

$$\Delta \;=\; O\!\left(\frac{d}{d-k}\right), \tag{18}$$

where $\Delta := \mathbb{E}_z\big[\mathrm{Var}_{t|z}\big(\mathbb{E}[u \mid z,t]\big)\big]$.

Proof. Let

$$m \;:=\; d-k, \qquad q(t) \;:=\; \sigma_\perp^2(t) \;=\; (1-t)^2 + t^2\sigma^2, \qquad \widehat{q}(z) \;:=\; \frac{\|P_\perp z\|^2}{m}.$$

The proof has two steps. First, we upper-bound the time-blindness gap by the error of a concrete time-blind predictor that estimates t from z (a brief inverse-clock argument turns this into the error of the noise-subspace variance estimate $\widehat{q}$). Second, we bound the resulting weighted variance-estimation error.

Step 1: Reduce the gap to time-estimation error. Recall that

$$\Delta \;=\; \mathbb{E}\Big[\big\|v^*(z,t) - \mathbb{E}[v^*(z,t) \mid z]\big\|^2\Big],$$

where $v^*(z,t) = \mathbb{E}[u \mid z,t]$ is the optimal time-aware predictor.

The conditional expectation $\mathbb{E}[v^*(z,t) \mid z]$ is the best predictor of $v^*(z,t)$ among predictors that only depend on z. Therefore, for any such predictor $g(z)$,

$$\Delta \;\le\; \mathbb{E}\big[\|v^*(z,t) - g(z)\|^2\big].$$

We use the particular predictor

$$g(z) \;:=\; v^*(z, \hat{t}(z)),$$

where $\hat{t}(z) \in I$ is the time estimate constructed in Step 2. This gives

$$\Delta \;\le\; \mathbb{E}\big[\|v^*(z,t) - v^*(z,\hat{t}(z))\|^2\big].$$

It remains to control how much $v^*(z,t)$ changes when t is replaced by $\hat{t}(z)$. In the diagonal Gaussian model,

$$v_i^*(z,t) \;=\; \alpha_i(t)\, z_i, \qquad \alpha_i(t) \;=\; \frac{tS_i - (1-t)}{(1-t)^2 + t^2 S_i}.$$

On I, the functions α_i are uniformly Lipschitz across all coordinates, since $S_i \in [\sigma^2, \lambda_{\max}]$ and the denominator is bounded away from zero. Thus there is a constant $L = L(I, \sigma^2, \lambda_{\max})$ such that

$$|\alpha_i(t) - \alpha_i(s)| \;\le\; L\,|t-s| \qquad \text{for all } i \text{ and } s,t \in I.$$

Therefore,

$$\|v^*(z,t) - v^*(z,s)\|^2 \;=\; \sum_{i=1}^{d} \big(\alpha_i(t) - \alpha_i(s)\big)^2 z_i^2 \;\le\; L^2\,\|z\|^2\,(t-s)^2.$$

Taking $s = \hat{t}(z)$ yields

$$\Delta \;\le\; L^2\,\mathbb{E}\big[\|z\|^2\,(\hat{t}(z)-t)^2\big]\,. \tag{19}$$

Since I is closed and $t^* = 1/(1+\sigma^2) \notin I$, $|q'(t)|$ is bounded below on I, so q^{-1} is Lipschitz on $q(I)$. Apply this with $\hat{t}(z)$ from (15): since projection onto $q(I)$ cannot increase distance to $q(t) \in q(I)$, there is a constant $C = C(I, \sigma^2)$ such that

$$|\hat{t}(z) - t| \;\le\; C\,|\widehat{q}(z) - q(t)|.$$

Therefore

$$\mathbb{E}\big[\|z\|^2(\hat{t}(z)-t)^2\big] \;\le\; C^2\,\mathbb{E}\big[\|z\|^2(\widehat{q}(z)-q(t))^2\big]\,. \tag{20}$$

Step 2: Bound the weighted variance-estimation error. Condition on t. In the noise subspace,

$$P_\perp z_t \;\sim\; \mathcal{N}(0, q(t) I_m),$$

so

$$\widehat{q} \;=\; q(t)\,\frac{Y}{m}, \qquad Y \sim \chi^2_m.$$

For readability, write $q := q(t)$, so

$$\widehat{q} - q \;=\; \frac{q}{m}(Y-m).$$

Split the energy into signal and noise components:

$$\|z\|^2 \;=\; \|P_x z\|^2 + \|P_\perp z\|^2.$$

We bound the two resulting terms separately.

For the signal component, $\widehat{q}$ depends only on $P_\perp z$, and under the diagonal spiked Gaussian model with independent ε, x, the projected blocks $P_x z$ and $P_\perp z$ are independent conditional on t. Hence

$$\mathbb{E}\big[\|P_x z\|^2(\widehat{q}-q)^2 \mid t\big] \;=\; \mathbb{E}[\|P_x z\|^2 \mid t]\cdot\mathbb{E}[(\widehat{q}-q)^2 \mid t].$$

The first factor is $O(k)$, since the k signal coordinates have uniformly bounded variance on I. The second factor is

$$\mathbb{E}[(\widehat{q}-q)^2 \mid t] \;=\; q^2\,\mathrm{Var}(Y/m) \;=\; \frac{2q^2}{m} \;=\; O\!\left(\frac{1}{m}\right),$$

because $q(t)$ is bounded on I. Therefore,

$$\mathbb{E}\big[\|P_x z\|^2(\widehat{q}-q)^2 \mid t\big] \;=\; O\!\left(\frac{k}{m}\right).$$

For the noise component, use

$$\|P_\perp z\|^2 \;=\; qY, \qquad \widehat{q}-q \;=\; \frac{q}{m}(Y-m).$$

Then

$$\mathbb{E}\big[\|P_\perp z\|^2(\widehat{q}-q)^2 \mid t\big] \;=\; \frac{q^3}{m^2}\,\mathbb{E}[Y(Y-m)^2].$$

By standard chi-square moment bounds,[1]

$$\mathbb{E}[Y(Y-m)^2] \;=\; O(m^2).$$

Since $q(t)$ is uniformly bounded on I, the noise contribution is

$$\mathbb{E}\big[\|P_\perp z\|^2(\widehat{q}-q)^2 \mid t\big] \;=\; O(1).$$

Combining the signal and noise contributions,

$$\mathbb{E}\big[\|z\|^2(\widehat{q}-q)^2 \mid t\big] \;=\; O\!\left(\frac{k}{m}\right) + O(1) \;=\; O\!\left(1+\frac{k}{m}\right) \;=\; O\!\left(\frac{d}{d-k}\right).$$

The bound is uniform over $t \in I$, so averaging over $t \sim \mathrm{Unif}(I)$ gives

$$\mathbb{E}\big[\|z\|^2(\widehat{q}(z)-q(t))^2\big] \;=\; O\!\left(\frac{d}{d-k}\right).$$

Combining this with (19) and (20) proves $\Delta = O(d/(d-k))$. □

[1] Indeed, for $Y \sim \chi^2_m$, $\mathbb{E}[(Y-m)^2] = 2m$ and $\mathbb{E}[(Y-m)^3] = 8m$, so $\mathbb{E}[Y(Y-m)^2] = 2m^2 + 8m$.

B.6 Independent Coupling Has Extensive Variance

The coupling variance $\mathcal{C}_\pi$ is the irreducible regression floor of (time-conditioned) flow matching: it is paid by even the optimal time-aware predictor and depends only on the choice of coupling π (Lemma A.1). Under independent coupling and the baseline spiked Gaussian model (Subsection B.1), $\mathcal{C}_\pi$ scales linearly with the ambient dimension; the proof reduces to a per-coordinate Gaussian closed form.

Proposition B.4 (Independent coupling has extensive variance). *Under the spiked Gaussian model with independent coupling and $t \sim \mathrm{Unif}(I)$,*

$$\mathcal{C}_\pi \;=\; \Omega(d).$$

Proof. **Coordinate reduction.** Recall that, working in the eigenbasis of Σ, we write

$$x_i \sim \mathcal{N}(0, S_i),$$

where S_i denotes the i-th eigenvalue of the data covariance Σ. Under the spiked model,

$$S_i = \begin{cases} \lambda_i + \sigma^2, & i \le k, \\ \sigma^2, & i > k. \end{cases}$$

Because the independent coupling makes x and ε independent, and because both are diagonal in this basis, the coordinates decouple. Hence it suffices to compute the one-dimensional conditional variance and then sum over coordinates.

Gaussian conditioning. Fix a coordinate i. Conditional on t, the pair

$$u_i = x_i - \varepsilon_i, \qquad z_i = (1-t)\varepsilon_i + t x_i$$

is jointly Gaussian. Since $x_i \sim \mathcal{N}(0, S_i)$ and $\varepsilon_i \sim \mathcal{N}(0,1)$ are independent,

$$\mathrm{Var}(u_i \mid t) = S_i + 1,$$
$$\mathrm{Var}(z_i \mid t) = (1-t)^2 + t^2 S_i,$$

and

$$\mathrm{Cov}(u_i, z_i \mid t) = \mathrm{Cov}(x_i - \varepsilon_i, (1-t)\varepsilon_i + t x_i) = tS_i - (1-t).$$

Since (u_i, z_i) is jointly Gaussian conditional on t, the scalar Gaussian conditioning formula gives

$$\mathrm{Var}(u_i \mid z_i, t) = \mathrm{Var}(u_i \mid t) - \frac{\mathrm{Cov}(u_i, z_i \mid t)^2}{\mathrm{Var}(z_i \mid t)}.$$

Substituting the above expressions gives

$$\mathrm{Var}(u_i \mid z_i, t) = (S_i + 1) - \frac{(tS_i - (1-t))^2}{(1-t)^2 + t^2 S_i}.$$

A direct simplification yields

$$\mathrm{Var}(u_i \mid z_i, t) = \frac{S_i}{(1-t)^2 + t^2 S_i}.$$

Uniform lower bound and summation. Since $S \mapsto S/((1-t)^2 + t^2 S)$ is increasing for each fixed $t \in I$, and since $S_i \ge \sigma^2 > 0$, we have

$$\frac{S_i}{(1-t)^2 + t^2 S_i} \ge \frac{\sigma^2}{(1-t)^2 + t^2 \sigma^2}.$$

Because $I \subset (0,1)$ is closed and bounded, the right-hand side has a positive minimum on I. Define

$$c_0 := \min_{t \in I} \frac{\sigma^2}{(1-t)^2 + t^2 \sigma^2} > 0.$$

Therefore, uniformly over all coordinates i and all $t \in I$,

$$\mathrm{Var}(u_i \mid z_i, t) \ge c_0.$$

Thus,

$$\mathcal{C}_\pi = \sum_{i=1}^{d} \mathbb{E}_t[\mathrm{Var}(u_i \mid z_i, t)] \ge \sum_{i=1}^{d} c_0 = c_0 d.$$

Hence $\mathcal{C}_\pi = \Omega(d)$.

□

B.7 Relative Cost of Time-Independence

Combining the two preceding propositions: the time-blindness gap Δ is bounded relative to the coupling variance $\mathcal{C}_\pi$ by $O(1/(d-k))$. This is the headline result, equivalent to Theorem 3 in the main text.

Theorem B.5 (Relative cost of time-independence). *Under the spiked Gaussian model with independent coupling and $t \sim \mathrm{Unif}(I)$,*

$$\frac{\Delta}{\mathcal{C}_\pi} \;=\; O\!\left(\frac{1}{d-k}\right). \tag{21}$$

In particular, holding k and the spectral parameters fixed, the ratio is $O(1/d)$.

Proof. By Proposition B.3, there is a constant $C_1 = C_1(I, \sigma^2, \lambda_{\max})$ such that

$$\Delta \;\le\; C_1 \,\frac{d}{d-k}.$$

By Proposition B.4, $\mathcal{C}_\pi \ge c_0\, d$ for a constant $c_0 = c_0(I, \sigma^2, \lambda_{\max}) > 0$. Dividing,

$$\frac{\Delta}{\mathcal{C}_\pi} \;\le\; \frac{C_1\, d/(d-k)}{c_0\, d} \;=\; \frac{C_1/c_0}{d-k} \;=\; O\!\left(\frac{1}{d-k}\right),$$

where the d factor in the numerator and denominator cancels, leaving the rate determined entirely by the noise-subspace dimension $d-k$. □

The bound is constructive: Proposition B.3 upper-bounds the optimal time-blind predictor by exhibiting one specific g^* that uses the noise-subspace estimator; the actual optimum $\bar{v}(z) = \mathbb{E}[u \mid z]$ can only do better. The ambient dimension d enters only through the energy normalization $\mathbb{E}[\|z\|^2] = \Theta(d)$ on both sides of the ratio and cancels; the statistical content is the $O(1/(d-k))$ rate, controlled by the noise-subspace dimension alone. Holding k fixed and varying d gives the $O(1/d)$ ratio scaling; holding $d-k$ fixed and varying k leaves the ratio bounded by a constant, the prediction that pulls the bound apart from the per-dimension account of Sun et al. [28] (Section 6). Note that the absolute gap $\Delta = O(d/(d-k))$ does not itself vanish when $d-k$ is comparable to d: it remains $O(1)$, and what vanishes is the ratio. Methods that suppress the coupling variance directly (optimal transport, mini-batch OT, rectified flows; Section 4.2) drive the dominant Term II down, leaving Δ as the only remaining residual.

C Experimental Details

C.1 Datasets and Architectures

Our experiments cover three datasets with two distinct backbones. CIFAR-10 is modeled in pixel space with an ADM UNet; CelebA-HQ and FFHQ are modeled in the latent space of a pretrained VAE using a SiT transformer. We document each setup below; later experiment subsections reference the relevant setup rather than redefining it.

Backbones. For CIFAR-10 we use an ADM UNet [7] with 128 base channels, channel multipliers $[1, 2, 2, 2]$, 2 residual blocks per resolution, multi-head attention at 16×16 resolution (4 heads, 64 channels per head), and dropout 0.1, totaling approximately 35.7M parameters. For CelebA-HQ and FFHQ we use SiT-B/2: a transformer of depth 12, hidden size 768, 12 attention heads, and patch size 2, acting on $16 \times 16 = 256$ tokens after encoding 256×256 images to $4 \times 32 \times 32$ latents with the Stable Diffusion 1.5 VAE (`stabilityai/sd-vae-ft-ema`, latents cached once before training).

C.2 Loss Decomposition Validation

Setup. We train four flow matching models per dataset in a 2×2 factorial design, varying time conditioning (on/off) and coupling (independent/OT), on CIFAR-10, CelebA-HQ, and FFHQ. Within each dataset, all four variants share the identical architecture and hyperparameters; the only differences are (i) whether the coupling is independent or mini-batch OT (exact EMD `pot_emd`) and (ii) whether

Table C.1: Per-dataset training hyperparameters.

Setting	CIFAR-10	CelebA-HQ	FFHQ
Architecture			
Backbone	ADM UNet ($\approx$35.7M)	SiT-B/2	SiT-B/2
Image space	pixel	VAE latent (SD 1.5)	VAE latent (SD 1.5)
Resolution / d	$32 \times 32 \times 3$, 3,072	$4 \times 32 \times 32$, 4,096	$4 \times 32 \times 32$, 4,096
Optimization			
Optimizer	AdamW	AdamW	AdamW
Learning rate	2×10^{-4}	1×10^{-4}	1×10^{-4}
Gradient clip	1.0	1.0	1.0
Warmup steps	5,000	5,000	5,000
EMA decay	0.9999	0.9999	0.9999
Magnitude function	`truncated`, $a = 0.8$	—	—
Training			
Iterations	200,000	100,000	100,000
Batch size	128 per GPU	256 global	256 global
Hardware	1$\times$ H200	1$\times$ H200	1$\times$ H200
Sampling			
Sampler	NAG	Euler	Euler

the model conditions on t or receives $t = 0$ at every step. The "time-blind" variant zeros the time input fed to the network at every step while keeping the time-embedding layers in the architecture, so the parameter count is identical across all four models. Per-dataset architectures, preprocessing, and training hyperparameters are specified in Appendix C.1; we summarize the cells below.

CIFAR-10. Pixel-space ADM UNet (Table C.1) trained for 200,000 steps with per-GPU batch size 128. OT coupling uses exact EMD with $B = 128$.

CelebA-HQ. SiT-B/2 in the Stable Diffusion VAE latent space (Table C.1) trained for 100,000 steps with global batch size 256. OT coupling uses exact EMD with $B = 256$.

FFHQ. SiT-B/2 in the Stable Diffusion VAE latent space (Table C.1), with the same protocol as CelebA-HQ: 100,000 steps, global batch size 256, exact EMD with $B = 256$.

MSE evaluation protocol. We compute the mean MSE under EMA weights in eval mode as a measurement of the converged training-loss floor of each cell. The measured quantity is the irreducible loss floor of the training distribution: a property of the joint distribution over (x, ε, t), not a generalization claim. To ensure a fair comparison, each model is evaluated under the coupling it was trained with: OT models (A, C) are evaluated with the same `pot_emd` permutation applied per eval batch (using the B matching training), while independent models (B, D) use random pairing.

CIFAR-10. CIFAR-10 has an official train/test partition (inherited from its classification-dataset origin), so we evaluate on its 10,000-image test split (78 batches of 128 samples). The train-vs-test loss gap is ≤ 0.0015, well within the per-batch standard error, so this distinction does not affect the decomposition.

CelebA-HQ and FFHQ. These benchmarks are introduced in the literature as generative-model benchmarks without a canonical train/test partition [14, 15], so following the convention of those benchmarks we evaluate on the same VAE-latent image set used for training, with fresh Gaussian noise $\varepsilon \sim \mathcal{N}(0, I)$ and uniform $t \sim \mathcal{U}[0, 1]$ drawn for each batch (78 batches of 256 samples).

FID evaluation protocol. For each of the four cells per dataset we generate 50,000 samples by integrating the trained vector field with 250 Euler steps from $t = 0$ to $t = 1$ using the EMA weights, and report clean-FID [22]. All four cells (A, B, C, D) per dataset use the identical sampler, sample count, and reference set; the only differences across cells are the model weights themselves. Per-dataset scoring details follow.

CIFAR-10. Samples are scored directly in pixel space against the standard CIFAR-10 training-set reference statistics.

CelebA-HQ and FFHQ. Samples are produced in VAE-latent space and decoded back to 256×256 pixels with the same `stabilityai/sd-vae-ft-ema` VAE used at training time before scoring against each dataset's full real distribution as the reference, following the convention of those benchmarks [14, 15].

Results. Table 1 reports mean $\pm$ standard deviation over three random seeds for each dataset and configuration. On CIFAR-10, switching to OT reduces mean MSE by 0.02496 for time-conditioned models and 0.02508 for time-blind models; removing time conditioning increases mean MSE by 0.00039 under independent coupling and 0.00027 under OT. The corresponding coupling reductions are 0.0400 and 0.0431 on CelebA-HQ, compared with time-conditioning effects of 0.0110 and 0.0079. On FFHQ, the coupling reduction is 0.0369 and the time-conditioning effect is 0.0032 for both comparisons. Thus, coupling has the larger effect on measured MSE across all three datasets. Time-blind models nevertheless achieve lower mean FID in all six dataset–coupling comparisons. The reported standard deviations describe variability across seeds; we do not infer statistical significance for each individual FID comparison from these summaries. Each measured MSE includes approximation error, so these contrasts are not direct estimates of the irreducible coupling variance or time-blindness gap.

C.3 Empirical Validation of Time Identifiability

We validate the theoretical predictions of Theorem 2 in four settings: (i) a fully synthetic spiked covariance model where all parameters are controlled exactly, and (ii–iv) real datasets (CIFAR-10 in pixel space, $d = 3072$; FFHQ and CelebA-HQ in VAE latent space, $d = 4096$), where the spiked covariance is fitted via PCA and used only for the theoretical prediction, not for data generation.

C.3.1 Common Estimator

In both settings, we observe interpolants of the form $z_i = (1 - t_i)\,\varepsilon_i + t_i\,x_i$, where $\varepsilon_i \sim \mathcal{N}(0, I_d)$ is the source noise and x_i is drawn from a distribution with covariance $\Sigma = U_x \Lambda U_x^\top + \sigma^2 I_d$. The interpolation times are sampled as $t_i \sim \mathrm{Uniform}[0, 1]$ over the full path; we report the mean absolute error $\mathrm{MAE} := \frac{1}{N}\sum_i |\hat{t}_i - t_i|$, which absorbs both the in-regime ($t < t^*$) error and any branch-mismatch contribution near the right endpoint.

Estimator. Given z_i alone, we estimate t_i via the noise-subspace statistic. Let $U_\perp \in \mathbb{R}^{d\times(d-k)}$ span the orthogonal complement of the signal subspace spanned by U_x, let $P_\perp := U_\perp U_\perp^\top$ denote the orthogonal projector onto the noise subspace, and define

$$\hat{\sigma}_\perp^2 \;:=\; \frac{1}{d-k}\,\|P_\perp z_i\|^2.$$

In the noise subspace, the interpolant has variance $\sigma_\perp^2(t_i) := (1 - t_i)^2 + t_i^2\,\sigma^2$, so $\hat{\sigma}_\perp^2$ concentrates around this deterministic function of t_i.

C.3.2 Discarded Samples

We compute $\hat{t}_i$ via the closed-form inversion (17) derived in Appendix B.3. The descending-branch minus sign applies in all four settings: Table C.2 reports per-dataset $\sigma_\perp^2$ and t^*, with $t^* \geq 0.89$ throughout, so the descending branch covers nearly the full $[0, 1]$ sampling range. The only finite-sample wrinkle is the discriminant, which can become negative. A short rearrangement of the discriminant in (16) gives

$$4 - 4(1 + \sigma^2)(1 - \hat{\sigma}_\perp^2) \;=\; 4\big[(1 + \sigma^2)\,\hat{\sigma}_\perp^2 - \sigma^2\big],$$

which is negative precisely when $\hat{\sigma}_\perp^2 < \sigma^2/(1 + \sigma^2) = \sigma_\perp^2(t^*)$, i.e. when the empirical noise-subspace variance falls below the global minimum of the clock and therefore lies outside the range of $\sigma_\perp^2$. This can only occur as a finite-sample fluctuation: the chi-square statistic in (22) occasionally dips below its mean far enough to push the estimate into the unattainable region. We discard such samples; in practice fewer than 0.1% are affected.

C.3.3 Theoretical Prediction

The proof of Theorem 2 establishes the rate $\hat{t} - t = O_p(1/\sqrt{d-k})$ via Chebyshev's inequality, but for direct comparison with experiments we want a quantitative reference distribution rather than just a rate. We derive one below in three short moves: the exact variance of the noise-subspace statistic, its propagation through the inverse map, and the aggregation across sampled times.

Variance of the clock statistic. Conditioned on t_i, the components of $U_\perp^\top z_i$ are i.i.d. $\mathcal{N}(0, \sigma_\perp^2(t_i))$, so $(d-k)\,\hat{\sigma}_\perp^2/\sigma_\perp^2(t_i) \sim \chi^2_{d-k}$. The variance of the chi-square distribution then gives the exact conditional variance

$$\mathrm{Var}\big(\hat{\sigma}_\perp^2 \mid t_i\big) \;=\; \sigma_\perp^4(t_i) \cdot \frac{2}{d-k}. \tag{22}$$

For large $d-k$, the central limit theorem additionally implies $\hat{\sigma}_\perp^2$ is approximately Gaussian. Equation (22) is the only stochastic input we need; everything that follows propagates it deterministically through the inverse map.

Delta-method propagation to time. Let $g := (\sigma_\perp^2)^{-1}$ denote the inverse map sending $\hat{\sigma}_\perp^2$ to $\hat{t}_i$. Its derivative at the true variance is

$$g'\big(\sigma_\perp^2(t_i)\big) \;=\; \frac{1}{\dot{\sigma}_\perp^2(t_i)}, \qquad \dot{\sigma}_\perp^2(t) \;=\; -2(1-t) + 2t\sigma^2.$$

A first-order Taylor expansion of g around $\sigma_\perp^2(t_i)$ gives the standard delta-method approximation

$$\hat{t}_i - t_i \;\approx\; g'\big(\sigma_\perp^2(t_i)\big) \cdot \big(\hat{\sigma}_\perp^2 - \sigma_\perp^2(t_i)\big),$$

and squaring then taking variances yields

$$\mathrm{Var}\big(\hat{t}_i - t_i \mid t_i\big) \;\approx\; \underbrace{\frac{1}{\big[-2(1-t_i) + 2t_i\sigma^2\big]^2}}_{\left[g'(\sigma_\perp^2(t_i))\right]^2} \cdot \underbrace{\sigma_\perp^4(t_i) \cdot \frac{2}{d-k}}_{\mathrm{Var}(\hat{\sigma}_\perp^2 \mid t_i)}. \tag{23}$$

Because g is a smooth map applied to a CLT-regime statistic, $\hat{t}_i - t_i$ inherits the Gaussian approximation of $\hat{\sigma}_\perp^2$ to leading order. This is the per-sample reference distribution we will compare against the empirical estimator.

Dataset-level prediction. To produce a single reference curve per dataset, we average (23) over the N sampled times $\{t_i\}_{i=1}^N$, excluding samples within 0.05 of the critical point $t_{\mathrm{crit}} = 1/(1+\sigma^2)$ (where g' nearly vanishes and the first-order Taylor expansion breaks down). This yields the dataset-level standard deviation

$$\bar{\sigma}_{\mathrm{theory}} \;=\; \sqrt{\frac{1}{N}\sum_{i=1}^{N} \mathrm{Var}\big(\hat{t}_i - t_i \mid t_i\big)}, \tag{24}$$

and we plot $\mathcal{N}(0, \bar{\sigma}^2_{\mathrm{theory}})$ against the empirical histogram of $\hat{t}_i - t_i$. Equation (24) is the prediction tested in Figure 4.

C.3.4 Setting 1: Synthetic Spiked Covariance

We sample $N = 10{,}000$ independent triples $(\varepsilon_i, x_i, t_i)$ with $\varepsilon_i \sim \mathcal{N}(0, I_d)$, $x_i \sim \mathcal{N}(0, U_x \Lambda U_x^\top + \sigma^2 I_d)$, and $t_i \sim \mathrm{Uniform}[0,1]$. Here $U_x \in \mathbb{R}^{d\times k}$ is a random orthonormal basis (obtained via QR decomposition of a Gaussian random matrix), $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ with eigenvalues spaced linearly between 1.0 and 10.0, and $\sigma^2 = 0.1$.

To verify the $1/\sqrt{d-k}$ scaling, we fix $k = 10$ and sweep the ambient dimension across $d \in \{20, 50, 100, 200, 500, 1000\}$, giving noise subspace dimensions $d-k \in \{10, 40, 90, 190, 490, 990\}$. For each d, we compute $\hat{t}_i$ for all N samples and record the empirical standard deviation of $\hat{t}_i - t_i$.

Results (Figure 4, panels a–b). Panel (a) shows the empirical standard deviation versus $d - k$ alongside the theoretical prediction; the two curves are nearly indistinguishable, confirming that the delta method approximation is tight. Panel (b) shows the full distribution of estimation errors for the setting $k = 10$, $d - k = 990$: the empirical KDE closely tracks the theoretical Gaussian $\mathcal{N}(0, \bar{\sigma}^2_{\text{theory}})$, with the empirical distribution being slightly more concentrated than the theoretical prediction. This is expected, as the theoretical variance is a first-order upper approximation and the actual estimator benefits from higher-order concentration.

C.3.5 Settings 2–4: Real Datasets with Fitted Covariance Prediction

The synthetic setting validates the theory in the exact model for which it was derived. We now ask a stronger question: does the Gaussian-derived theoretical prediction remain accurate when applied to *real, non-Gaussian data*? To test this, we use actual images (or VAE latents) as the data distribution x_i and only use the Gaussian approximation for the theoretical variance prediction, not for data generation. We evaluate on CIFAR-10 (pixel space, $d = 3072$), FFHQ (VAE latents, $d = 4096$), and CelebA-HQ (VAE latents, $d = 4096$).

Common procedure. For each dataset we compute the sample mean μ and sample covariance $\hat{\Sigma} = \frac{1}{n-1}\sum_i (x_i - \mu)(x_i - \mu)^\top$ over the full set of n available samples (CIFAR-10 in pixel space, CelebA-HQ and FFHQ in VAE-latent space), and form the eigendecomposition $\hat{\Sigma} = V\,\mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_d)\,V^\top$ with eigenvalues sorted in decreasing order. Given a signal rank k, we approximate the data covariance with a spiked model by setting the signal subspace $U_x = V_{:,1:k}$, the spike eigenvalues $\Lambda = \mathrm{diag}(\hat{\lambda}_1 - \sigma^2, \ldots, \hat{\lambda}_k - \sigma^2)$ (clipped to be non-negative), and the isotropic noise variance $\sigma^2 = \frac{1}{d-k}\sum_{j=k+1}^{d} \hat{\lambda}_j$. These parameters are used solely for the theoretical variance prediction; the data itself is not assumed to be Gaussian. We then sample $N = 20{,}000$ triples $(\varepsilon_i, x_i, t_i)$ with $\varepsilon_i \sim \mathcal{N}(0, I_d)$, x_i drawn with replacement from the dataset (and centered by subtracting μ), and $t_i \sim \mathrm{Uniform}[0, 1]$, form the interpolant $z_i = (1 - t_i)\,\varepsilon_i + t_i\, x_i^{\text{centered}}$, and apply the noise-subspace estimator using the PCA-derived $U_\perp$.

The representative signal rank k for the per-dataset error distribution plots (Figure 4) is chosen at the 95% cumulative-explained-variance threshold, which falls in the regime where the spectrum transitions from rapid to gradual decay (Figure 3). The scaling sweep additionally varies $k \in \{10, 50, 100, 200, 500, 1000, 2000\}$ within each dataset to trace the full range of noise-subspace dimensions, with σ^2 recomputed from the residual spectrum at each k.

CIFAR-10. We load all $50{,}000$ training images and flatten each $32 \times 32 \times 3$ image to a $d = 3{,}072$-dimensional vector (pixel values in $[0, 1]$). The 95% explained-variance threshold corresponds to $k = 217$, which is the rank used for the error-distribution panel; the mean absolute error is 0.0062.

FFHQ. We use VAE latents at $d = 4{,}096$ (Stable Diffusion VAE, $4 \times 32 \times 32$); see Table C.1 for the encoding setup. The 95% explained-variance threshold corresponds to $k = 2{,}581$; the mean absolute error is 0.0213.

CelebA-HQ. We use VAE latents at $d = 4{,}096$ from the same Stable Diffusion VAE used for FFHQ; see Table C.1. The 95% explained-variance threshold corresponds to $k = 2{,}273$; the mean absolute error is 0.0175.

Results (Figure 4). Figure 4 shows the error distribution at the 95% explained-variance rank for all four settings. Despite using real, non-Gaussian data, the empirical distributions closely match the Gaussian-derived theoretical prediction in every case, providing empirical agreement beyond the Gaussian model rather than a theoretical guarantee. The scaling sweep (not shown in the main figure) further confirms the $1/\sqrt{d-k}$ rate across all datasets. Notably, the VAE latent distributions (FFHQ, CelebA-HQ) have much flatter spectra than CIFAR-10 pixels: 95% explained variance requires $k \approx 2{,}581$ and $k \approx 2{,}273$ respectively, compared to $k = 217$ for CIFAR-10 (Figure 3). This means the VAE spreads variance across more components, leaving a smaller effective noise subspace. Despite this, the estimator remains accurate in all cases, likely because the projection onto a high-dimensional noise subspace has a CLT-like averaging effect regardless of the marginal distribution of x_i.

C.4 Time-Localized Estimation Error and the Variance-Clock Critical Point

Theorem 2 requires staying away from the critical point $t^* = 1/(1+\sigma^2)$ at which the scalar residual-variance clock $\sigma_\perp^2(t) = (1-t)^2 + t^2\sigma^2$ has zero derivative. This subsection probes how this caveat manifests on real datasets by (a) locating t^* from the fitted residual variance per dataset, and (b) sweeping the noise-subspace estimator across $t \in [0,1]$ and reporting binned mean absolute error.

Setup. For each of CIFAR-10, CelebA-HQ, and FFHQ we run PCA on the data, pick k at the 95% cumulative-explained-variance threshold, and fit $\hat{\sigma}^2_{\perp,\mathrm{data}} = \mathrm{mean}(\hat{\lambda}_{j>k})$ from the trailing eigenvalues, giving the predicted clock and its critical point. We then sample $t \sim \mathrm{Uniform}[0,1]$ (with extra mass at $t=1$ to populate the right edge), draw a real datapoint x and noise $\varepsilon \sim \mathcal{N}(0, I_d)$, form the interpolant $z = (1-t)\varepsilon + tx$, and run the noise-subspace estimator $\hat{t}(z)$ from (6) via the closed-form descending-branch inverter of Appendix B.3 (the minus-sign root of the quadratic). On CelebA-HQ and FFHQ this includes $t > t^*$, where the always-descending-branch convention returns the wrong root by construction; the consequences are quantified in the second result below. We use 10^5 samples per dataset and report the binned conditional mean absolute error $\mathbb{E}[|\hat{t} - t| \mid t]$.

Dataset	d	k	$d-k$	$\sigma_\perp^2$	$t^* = 1/(1+\sigma_\perp^2)$
Synthetic	—	10	—	0.1000	0.9090
CIFAR-10	3072	99	2973	0.0064	0.9937
CelebA-HQ	4096	1571	2525	0.1079	0.9026
FFHQ	4096	1900	2196	0.1182	0.8943

Table C.2: Per-dataset residual-subspace variance $\sigma_\perp^2$ and induced critical point t^*. $\sigma_\perp^2$ is the mean of the bottom $d-k$ PCA eigenvalues of the data covariance, with k chosen at the 95% cumulative-variance threshold.

Results. The experiment makes two predictions, one about the geometry of the variance clock and one about the empirical failure mode of the estimator.

The clock's critical point is dataset-determined and pushed to the data endpoint by small residual variance. *Hypothesis:* since $t^* = 1/(1+\hat{\sigma}^2_{\perp,\mathrm{data}})$ is a monotone function of the fitted residual variance, datasets with small $\hat{\sigma}^2_{\perp,\mathrm{data}}$ should place t^* near the data endpoint $t=1$, while datasets with flatter residual spectra (larger $\hat{\sigma}^2_{\perp,\mathrm{data}}$) should move t^* into the interior. *Validation:* the fitted critical points (Table C.2, Figure C.1) span CIFAR-10 $t^* \approx 0.994$ down to FFHQ $t^* \approx 0.894$: the pixel-space CIFAR-10 has small residual variance and t^* at the boundary, while the VAE-latent datasets have flatter residual spectra and t^* in the interior.

For $t > t^*$ the descending-branch inverter returns the reflection $2t^* - t$, producing a quantitatively predictable rise in error. *Hypothesis:* the variance clock $\sigma_\perp^2(t) = (1+\sigma^2)(t-t^*)^2 + \sigma_\perp^2(t^*)$ is a parabola symmetric around t^*, so for any $\hat{\sigma}_\perp^2 \geq \sigma_\perp^2(t^*)$ the two preimages are $t_\pm = t^* \pm \delta$ with $\delta = \sqrt{(\hat{\sigma}_\perp^2 - \sigma_\perp^2(t^*))/(1+\sigma^2)}$. The minus-root inverter returns t_-, so when the true $t > t^*$ the estimator returns the reflection $2t^* - t$ rather than t, giving error

$$|\hat{t} - t| \;\approx\; 2(t - t^*) \quad \text{for } t > t^*.$$

We therefore predict three things on the binned-MAE plot: (i) error stays small for $t < t^*$, (ii) error rises with slope 2 starting near $t = t^*$, and (iii) the right-edge peak equals $2(1-t^*)$ — predicted to be ≈ 0.012 for CIFAR-10, ≈ 0.194 for CelebA-HQ, and ≈ 0.212 for FFHQ. *Validation:* on CelebA-HQ and FFHQ (Figure C.2) the rising-error region begins near the dataset-specific t^* rather than at the right endpoint, with peak magnitudes matching the dataset-dependent $2(1-t^*)$ prediction. CIFAR-10's predicted peak (0.012) is small enough that the branch-mismatch effect is subdominant; any observed residual rise on CIFAR-10 instead reflects model-misspecification: as $t \to 1$ the marginal of z_t approaches the data marginal, which is non-Gaussian and not perfectly described by the spiked-covariance fit used to compute $\sigma_\perp^2(t)$.

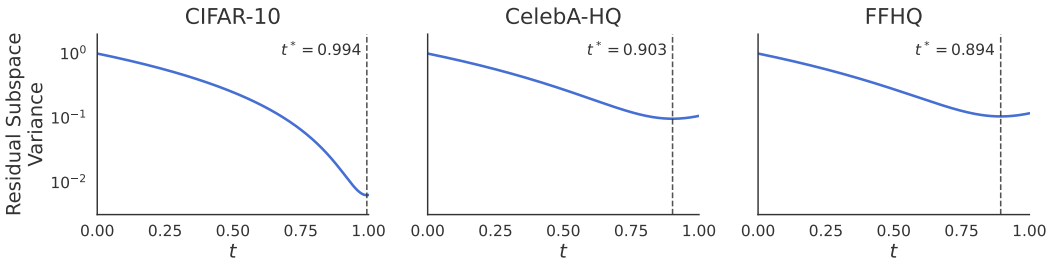


Figure C.1: **Dataset-specific residual-variance clocks.** For each dataset we estimate the residual-subspace data variance $\hat{\sigma}^2_{\perp,\text{data}}$ after fixing the PCA signal dimension k, and plot $\sigma^2_\perp(t) = (1-t)^2 + t^2\,\hat{\sigma}^2_{\perp,\text{data}}$. Dashed lines mark the predicted critical point $t^* = 1/(1+\hat{\sigma}^2_{\perp,\text{data}})$, where the scalar variance statistic has zero derivative. Small residual variance (CIFAR-10) places t^* near the data endpoint $t = 1$; flatter residual spectra (CelebA-HQ, FFHQ, VAE latents) move t^* into the interior.

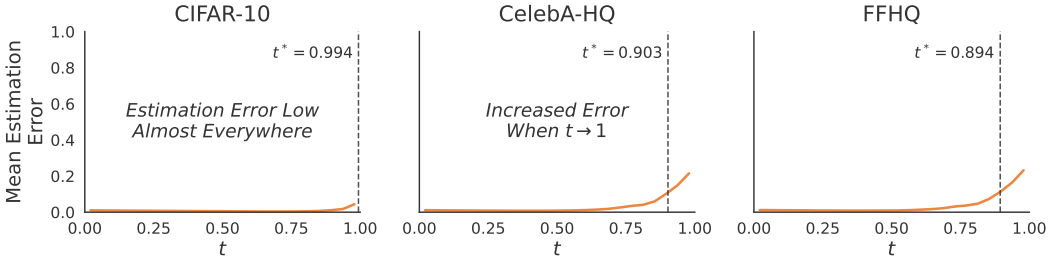


Figure C.2: **Estimator error rises for $t > t^*$, where the always-minus inverter returns the reflection $2t^* - t$ rather than t.** We bin samples by true interpolation time t and report mean absolute error $\mathbb{E}[|\hat{t} - t| \mid t]$. The closed-form inverter (Appendix B.3) always returns the descending-branch root, so for $t > t^*$ it returns $2t^* - t$ and incurs error $2(t - t^*)$. Dashed lines mark the dataset-specific t^*. On CelebA-HQ ($t^* \approx 0.903$) and FFHQ ($t^* \approx 0.894$) the empirical rise begins near t^* with right-edge magnitude $\approx 2(1 - t^*)$, matching the prediction. CIFAR-10 ($t^* \approx 0.994$) has almost no ascending-branch region, so its residual right-edge rise reflects other effects, most plausibly model misspecification of the spiked-covariance Gaussian fit as the marginal becomes data-like.

C.5 Closed-Form Variance Decomposition on the Spiked Gaussian

This experiment underlies Figure 6. It validates Theorem 3 (equivalently, Theorem B.5) in a setting where the ambient and signal dimensions can be controlled directly, so the predicted separation of scales between the coupling variance (Term II) and the time-blindness gap (Term III) is observable without confounds from model capacity or data distribution.

Setup. We use the spiked Gaussian model throughout: $\varepsilon \sim \mathcal{N}(0, I_d)$, $x \sim \mathcal{N}(0, \Sigma)$ with $\Sigma = \text{diag}(\lambda, \ldots, \lambda, \sigma^2, \ldots, \sigma^2)$ (the first k coordinates are signal with variance λ, the remaining $d-k$ are isotropic noise with variance σ^2), $z = (1-t)\varepsilon + tx$, and $u = x - \varepsilon$. Default parameters are $\lambda = 10$, $\sigma^2 = 0.01$, with $t \sim \text{Uniform}[\tau, 1-\tau]$, $\tau = 0.15$. Because Σ is diagonal and ε, x are independent, the pairs (z_i, u_i) are jointly Gaussian and independent across coordinates, so all quantities of interest reduce to sums of per-coordinate integrals and admit either closed forms or low-variance numerical estimators.

Term I: learnable error. The learnable error is the variance of $u = x - \varepsilon$ with no conditioning, which has the closed form

$$\text{tr}(\text{Cov}(u)) \;=\; \text{tr}(\Sigma + I_d) \;=\; k(\lambda+1) + (d-k)(\sigma^2+1).$$

This is the only one of the three terms that admits a fully closed-form integral, since it does not depend on t.

Term II: coupling variance $\mathcal{C}_\pi$. Per coordinate, joint Gaussianity of (z_i, u_i) gives the standard regression-variance closed form

$$\mathrm{Var}(u_i \mid z_i, t) \;=\; (S_i + 1) - \frac{(tS_i - (1-t))^2}{v_i(t)}, \qquad v_i(t) \;:=\; (1-t)^2 + t^2 S_i,$$

with $S_i \in \{\lambda, \sigma^2\}$ depending on whether i is a signal or noise coordinate. This expression does not depend on z_i (Gaussian–Gaussian conditioning), so summing over coordinates produces a deterministic function of t. We integrate it against $t \sim \mathrm{Unif}[\tau, 1-\tau]$ by the trapezoidal rule on a 2000-point grid to obtain $\mathcal{C}_\pi$.

Total time-blind variance $\mathbb{E}_z[\mathrm{tr}(\mathrm{Var}(u \mid z))]$. This is the irreducible error when t is not observed. Unlike $\mathcal{C}_\pi$ it does not reduce to a closed-form integral, because $\mathrm{Var}(u \mid z)$ depends on the full posterior $p(t \mid z)$. We estimate it with a hybrid Monte-Carlo / grid scheme.

We draw $N = 2 \times 10^5$ outer samples (t, ε, x) with $t \sim \mathrm{Unif}[\tau, 1-\tau]$, $\varepsilon \sim \mathcal{N}(0, I_d)$, and $x \sim \mathcal{N}(0, \Sigma)$, and form $z = (1-t)\varepsilon + tx$. For each such z, we evaluate the posterior on a 2000-point grid $\{t'\}$ in $[\tau, 1-\tau]$ via

$$p(t' \mid z) \;\propto\; \prod_i \mathcal{N}\big(z_i;\, 0,\, v_i(t')\big).$$

From this posterior, we compute the conditional mean of each coordinate of u as

$$\mathbb{E}[u_i \mid z] \;=\; z_i \sum_{t'} \alpha_i(t')\, p(t' \mid z), \qquad \alpha_i(t') \;:=\; \frac{t' S_i - (1-t')}{v_i(t')},$$

and the conditional second moment as

$$\mathbb{E}[u_i^2 \mid z] \;=\; \sum_{t'} \big[\mathrm{Var}(u_i \mid z, t') + (\alpha_i(t')\, z_i)^2\big]\, p(t' \mid z).$$

The per-coordinate conditional variance follows by subtraction:

$$\mathrm{Var}(u_i \mid z) \;=\; \mathbb{E}[u_i^2 \mid z] - \big(\mathbb{E}[u_i \mid z]\big)^2.$$

Summing across coordinates and averaging across the N samples of z gives the estimator of $\mathbb{E}_z[\mathrm{tr}(\mathrm{Var}(u \mid z))]$. All arithmetic is performed in double precision to avoid numerical cancellation when Term III is orders of magnitude smaller than Term II.

Term III: time-blindness gap Δ. The time-blindness gap is recovered by the law of total variance as the residual between the total time-blind variance and the coupling variance: $\Delta = \mathbb{E}_z[\mathrm{tr}(\mathrm{Var}(u \mid z))] - \mathcal{C}_\pi$. This decomposition lets us extract Δ without needing a separate Monte-Carlo estimator for it.

Sweep. We sweep both axes of the (d, k) grid simultaneously: $k \in \{1, 2, 4, \ldots, 1024\}$ and $d \in \{2, 4, 8, \ldots, 1024\}$ (log-spaced powers of two), restricted to $k < d$ since the construction requires a non-empty noise subspace. At each (d, k) we recompute all three terms from scratch. From this (d, k) table we extract three slices for Figure 6: the column at fixed $d = 1024$ sweeping k (left panel), the entire grid as a heatmap of Term III (center panel), and the row at fixed $k = 64$ sweeping d (right panel).

Results (Figure 6). The right panel reproduces the asymptotic gap-collapse: at $k = 64$ fixed, Term III decays from ~ 2 at the smallest admissible d to below ~ 0.5 by $d = 1024$, while Terms I and II stay bounded over the same sweep. The left panel shows the complementary regime: at $d = 1024$ fixed, all three terms grow with k, but Term III grows fastest because the noise subspace $d - k$ is shrinking, sharpening the prediction that the variance-based identifiability mechanism degrades as $k \to d$. The center heatmap shows both regimes simultaneously: Δ is small along the upper-right edge ($k \ll d$) and large along the diagonal ($k \approx d$), tracing the same geometry across the (d, k) plane. Together, these slices give a quantitative signature of Theorem 3.

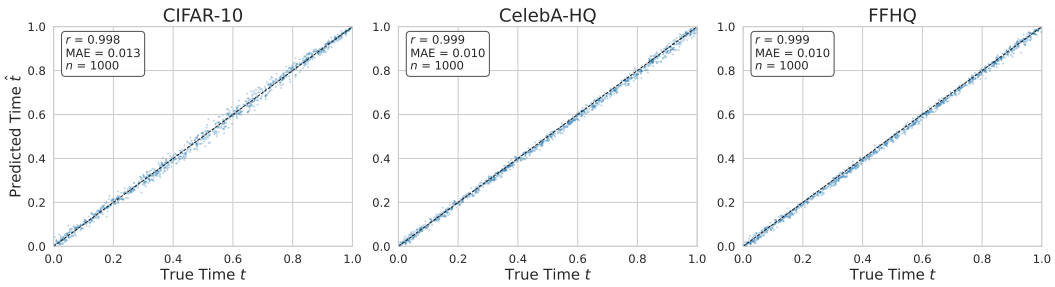


Figure D.1: **Held-out accuracy of the time-prediction probe.** Scatter of true interpolation time t vs. probe prediction $\hat{t}$ on held-out samples for each dataset. The probe is highly accurate on all three datasets, indicating that the model's internal representation contains a near-deterministic encoding of t.

D Probe Subspace-Corruption Diagnostic

We test whether time information is available in frozen time-blind model features and how decoding responds to input-subspace corruption. We also measure changes in velocity-prediction loss. These diagnostics do not establish that the velocity predictor uses the residual-energy estimator constructed by our theory.

Probe architecture. The probe is a 3-layer MLP with hidden size 512, SiLU activations, and a sigmoid head producing $\hat{t} \in (0, 1)$. The underlying generative model is frozen throughout; only the probe is trained. The bottleneck features it reads from differ by backbone:

CIFAR-10. We extract the CFM UNet's middle-block bottleneck and flatten it to a 4,096-dimensional feature vector.

CelebA-HQ and FFHQ. We mean-pool the middle SiT block to obtain a 768-dimensional feature vector.

Training. At each step we sample a batch of training images x, fresh Gaussian noise $\varepsilon \sim \mathcal{N}(0, I_d)$, and times $t \sim \mathrm{Uniform}[0, 1]$, and form the interpolant $z = (1-t)\varepsilon + tx$ in the model's natural input space (pixels for CIFAR-10 and VAE latents for CelebA-HQ and FFHQ). We pass z through the frozen backbone with a dummy $t = 0$ supplied as the time input (the underlying model is time-blind), extract the chosen bottleneck features, and train the probe to minimize $\mathrm{MSE}(\hat{t}, t)$ with Adam at learning rate 10^{-3}, batch size 256, for 3,000 steps. The same training protocol is used for all three datasets.

Held-out accuracy. Figure D.1 shows the per-dataset scatter of true t versus predicted $\hat{t}$ on a held-out evaluation set (10,000 samples). The probe is highly accurate in all three datasets, with Pearson correlation $r = 0.998$ and mean absolute error 0.013 on CIFAR-10, and $r = 0.999$ and mean absolute error 0.010 on CelebA-HQ and FFHQ. This shows that the bottleneck features of trained time-blind models contain sufficient information for the probe to recover t at high precision, the prerequisite for the corruption-based intervention below.

Corruption protocol. To localize where in the input spectrum the probe's predictions draw from, we progressively corrupt PCA dimensions of z_t with content from a wrong time t'. Given a sample (x, ε, t) and an independently drawn wrong time $t' \sim \mathrm{Unif}[0, 1]$, form the clean interpolant $z_t = (1-t)\varepsilon + tx$ and the wrong-time interpolant $z_{t'} = (1-t')\varepsilon + t'x$. For a corruption fraction $f \in [0, 1]$ and rank $k = \lfloor f\, d \rfloor$, let V_k denote either the leading k columns of the data PCA basis V (*high-to-low*, signal-subspace directions first) or the trailing k columns (*low-to-high*, noise-subspace directions first), and form

$$z^{\mathrm{corr}}(t', f) \;=\; z_t \;+\; V_k V_k^{\top}\big(z_{t'} - z_t\big),$$

which replaces the chosen k-dimensional PCA subspace of z_t with the corresponding subspace of $z_{t'}$. The probe is evaluated on z^{corr}: if the probe relies on the corrupted directions, $\hat{t}$ tracks t' rather than t and the correlation collapses; if it relies on the uncorrupted directions, the correlation remains high.

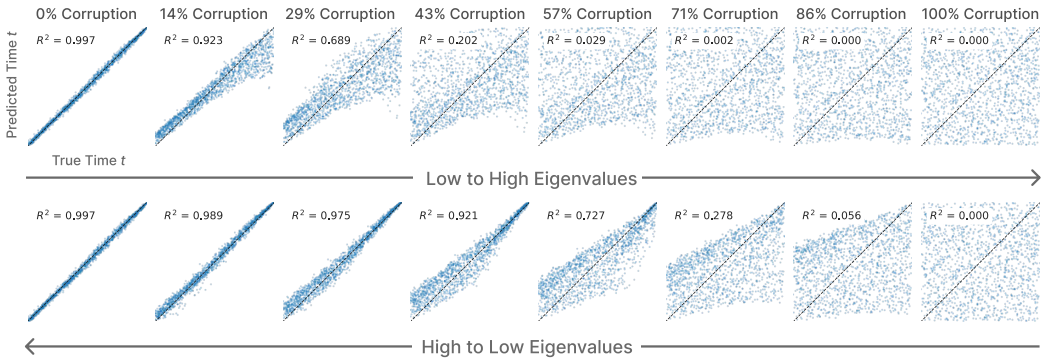


Figure D.2: **Eigenvalue-ordered corruption at progressive fractions.** Probe predictions $\hat{t}$ against true t. **Top:** low-to-high eigenvalues, corrupting residual directions first. **Bottom:** high-to-low eigenvalues, corrupting signal directions first. Columns show eight evenly spaced fractions from zero to full replacement; the displayed percentages are rounded.

The PCA basis $V \in \mathbb{R}^{d\times d}$ is computed from up to $50{,}000$ training samples per dataset. We report two complementary views of probe accuracy as f sweeps $[0, 1]$: the aggregate Pearson correlation $r(\hat{t}, t)$ as a single curve per ordering, and the underlying scatter grid at progressive fractions.

Aggregate correlation. *Hypothesis:* if the probe reads t from the noise-subspace (low-eigenvalue) component of z_t, the high-to-low sweep should leave $r(\hat{t}, t)$ near baseline until corruption reaches noise-subspace directions, while the low-to-high sweep should collapse r early. *Validation:* on all three datasets (Figure 5), the low-to-high curve drops sharply within the first $\sim$30% of the sweep while the high-to-low curve stays near baseline until the sweep reaches noise-subspace dimensions.

Per-panel resolution. The Pearson summary in Figure 5 aggregates each fraction into a single number; we additionally plot the underlying scatter grid of $\hat{t}$ against t at progressive fractions, reading R^2 off each panel directly. We render a $2 \times N$ grid with $N = 8$: rows are the two corruption orderings and columns are corruption fractions $f \in \mathrm{linspace}(0, 1, N)$, where $f = 0$ corresponds to no corruption (probe baseline) and $f = 1$ corresponds to fully replacing z_t with $z_{t'}$. Each cell runs the trained probe on $2{,}000$ trials with independently sampled (x, ε, t, t') and plots $\hat{t}$ against t, with $R^2 = r(\hat{t}, t)^2$ shown per panel. This grid illustrates how the probe responds across corruption fractions; its rounded labels include 43%, not the exact 40% setting used for the separate control experiment in Table D.1.

Random-subspace control. We repeat the replacement intervention using a random orthonormal subspace of the same dimension instead of leading or trailing PCA directions. Table D.1 reports squared Pearson correlation at 40% corruption. On every dataset, residual-direction corruption degrades time decoding more than the random control, while signal-direction corruption degrades it less. Here r^2 denotes squared correlation, not the coefficient of determination computed from residual sums of squares.

Table D.1: Squared Pearson correlation between predicted and true time at 40% of input dimensions corrupted. These are diagnostic results, separate from the three-seed model ablation in Table 1.

Dataset	Clean	Signal	Residual	Random
CIFAR-10	0.998	0.924	0.288	0.589
CelebA-HQ	0.998	0.669	0.437	0.538
FFHQ	0.999	0.822	0.304	0.550

Velocity-output sensitivity. At 40% residual-direction corruption, flow-matching loss increases from 0.171 to 0.411 on CIFAR-10, from 0.611 to 0.842 on CelebA-HQ, and from 0.676 to 0.942 on FFHQ. These values come from the corruption diagnostic, not the three-seed averages in Table 1.

Signal-direction corruption produces a comparable, slightly larger loss increase. Thus the velocity predictor is sensitive to residual-direction perturbations, but the experiment does not establish preferential reliance on those directions or their use for inferring time.

Interpretation and limitations. The clean and wrong-time interpolants share both x and ε; only t' is drawn independently. Nevertheless, replacing subspace components changes the input distribution. Perturbation energy is not explicitly matched across signal, residual, and random subspaces, so these are dimension-matched rather than norm-matched controls. Together, the results support time decodability in learned representations and greater sensitivity of the auxiliary probe to residual directions. They do not identify the statistic computed by the velocity predictor. Raw-input and untrained-feature probes and norm-matched output interventions would provide further controls.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Each contribution stated in the introduction maps to a labeled section: Proposition 1, Theorem 2, Theorem 3, and Section 5.1.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 4 discusses the linear-interpolant scope and Proposition B.1 establishes the effective-rank generalization beyond the spiked model; the Discussion notes the open extension to nonlinear data geometry and other schedules.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Assumptions are stated in each theorem statement; full proofs are given in Appendices A and B.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Full experimental details (datasets, architectures, optimizer, hyperparameters, sampling, evaluation) are in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: All datasets are public, and the setup details needed to reproduce the main results are given in Appendix C.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Splits, architectures, and training/evaluation hyperparameters are reported in Appendix C.1 and the per-experiment subsections of Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Empirical standard deviations and theoretical Gaussian predictions for the noise-subspace estimator are reported in Appendices C.3 and C.5. Table 1 reports mean $\pm$ standard deviation of MSE and FID over three random seeds for every dataset–configuration pair. These standard deviations describe seed variability, not confidence intervals.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute (single NVIDIA H200 GPU, 100,000 iterations, global batch size 256) is reported in Appendix C.1. Synthetic experiments are CPU-light.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics: only public datasets are used, no human subjects are involved, and no new high-risk assets are released.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The contributions are foundational analysis with no direct path to specific societal impacts beyond those inherent to generative modeling broadly.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No pretrained models, datasets, or other high-risk artifacts are released.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets, backbones, and external code (cited in Appendix C.1) are used within their published terms.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new datasets or pretrained models.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not part of the core methodology; any use was confined to writing assistance.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.