WHEN CAN DIFFUSION DENOISERS SHARE A REPRESENTATION ACROSS NOISE LEVELS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models solve denoising problems across noise levels, yet typically reuse a single time-conditioned network. Motivated in part by latent diffusion, we ask when the data distribution itself justifies sharing a common representation across noise levels. For Gaussian diffusion, we establish an exact characterization: if the Bayes-denoiser Jacobian is diagonalizable in the same fixed frame at every input at just one positive noise level, then the clean distribution admits an affine orthogonal independent-component representation. Conversely, this structure guarantees a common coordinatewise Bayes-denoising frame across all noise levels. Beyond exact independence, we show that total correlation equals both the noise-integrated excess Bayes risk of coordinatewise denoising and the generative KL discrepancy of an ideal product model. We distinguish this factorization cost from the additional cost of sharing a frame across noise levels. We derive sufficient finite-sample learning bounds for shared and untied frames, and show why approximate positive-noise Bayes-denoiser separability need not imply approximate clean independence in a prescribed frame. Experiments with independently trained denoisers and learned VAE-family and CIFAR-10 representations support these predictions, including the finite-sample crossover between shared and noise-specific representations.

1 INTRODUCTION

Diffusion models often use one time-conditioned network across noise levels, but architectural weight sharing alone does not establish that a common representation is statistically appropriate. We ask: *when does the data distribution itself admit a frame that is optimal across the denoising path?*

We study this question through the Gaussian channel $Y_t = X + tG$, with $G \sim \mathcal{N}(0, I_n)$ and $t > 0$, and its Bayes-optimal denoiser $f_t^\star(y) = \mathbb{E}[X \mid Y_t = y]$. The Bayes denoiser matters beyond minimizing denoising error: through Tweedie's identity, it also determines the score used by the Gaussian reverse process (Anderson, 1982; Song et al., 2020). There is a quantitative reason to learn these denoisers accurately. The likelihood-weighted bound of Song et al. (2021, Theorem 1), specialized to this Gaussian channel, gives

$$\mathrm{KL}(P_\tau \| \widehat{P}_\tau) \le \mathrm{KL}(P_T \| \pi_T) + \int_\tau^T \frac{\mathbb{E}\big[\|\widehat{f}_t(Y_t) - f_t^\star(Y_t)\|^2\big]}{t^3}\, dt, \tag{1}$$

under the corresponding regularity conditions, for $0 < \tau < T$. Here P_t denotes the law of Y_t, π_T is the initialization law, and $\widehat{P}_\tau$ is the law produced by the learned continuous-time reverse process. The bound has a simple message: to control generative discrepancy, we need to control both the initialization error and the accumulated error in approximating the Bayes denoisers across noise levels. Rather than asking whether a particular architecture shares hidden features, we ask when the optimal denoisers themselves admit one common separable frame.

Our main result gives an exact characterization. Fix an orthonormal frame $U \in \mathbb{R}^{n\times k}$. If, at a single positive noise level t_0, the Bayes-denoiser Jacobian satisfies $Df_{t_0}^\star(y) \in \{UDU^\top : D \text{ diagonal}\}$ for every y, then the clean law must have an affine orthogonal independent-component representation $X = b + U\Psi$, with mutually independent coordinates of Ψ and $b \perp \mathrm{range}(U)$. The converse also

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

holds. The key implication is cross-noise: the one-noise Jacobian condition forces factorization of the smoothed density, Gaussian deconvolution transfers this factorization to the clean law, and the resulting independent coordinates are then denoised coordinatewise at *every* noise level. Thus the all-input Jacobian structure at one noise level can reveal a representation shared across the full diffusion path.

A complementary motivation from latent diffusion. Latent diffusion denoises representations learned by autoencoders, often with VAE-based encoders (Rombach et al., 2022). A variance-preserving Gaussian observation $\alpha_s X + \sigma_s G$, with $\alpha_s > 0$, rescales to our channel at $t = \sigma_s/\alpha_s$ (Appendix A.3). This motivates asking how learned-latent distributions affect a common denoising frame. Factorized conditional VAE encoders do not generally imply independent coordinates in the *aggregate* latent distribution, and we do not assume practical latent diffusion networks satisfy our exact fixed-frame hypothesis.

Away from exact independence, we obtain a quantitative but asymmetric picture. For any fixed orthogonal frame, the total correlation of the transformed clean distribution equals the noise-integrated excess Bayes risk of denoising its coordinates independently, and also the KL error of the corresponding product-of-marginals generative model. We then separate this *factorization cost* from the additional cost of forcing the same frame across noise levels. Although tying may introduce approximation bias, pooling observations across noise levels yields a smaller estimation-complexity term in our sufficient learning bounds for a correctly specified model. Our controlled experiments separately exhibit an empirical crossover between estimation error and the approximation cost of sharing under misspecification.

In a prescribed frame, the reverse approximate implication fails without further assumptions. A positive-noise Bayes denoiser can be arbitrarily close, even at the Jacobian level, to that of an independent model while the clean distribution retains nonvanishing dependence in that frame. Gaussian smoothing can therefore hide fine-scale dependence: near-separability is a useful diagnostic of shared geometry, but not a distribution-free certificate of approximate clean independence in the prescribed frame.

Contributions. Our main contributions are:

1. **One-noise characterization.** A fixed Bayes-Jacobian frame at one positive noise level is equivalent to an affine orthogonal independent-component model for the clean law, and therefore implies the same coordinatewise Bayes frame at every noise level. We also give an intrinsic formulation through pairwise commutation of Bayes Jacobians.
2. **Approximate and finite-data theory.** Total correlation equals the noise-integrated regret of coordinatewise Bayes denoising and the KL error of the associated product-of-marginals model. We isolate the additional cost of cross-noise tying, derive sufficient finite-sample learning bounds for shared and untied frames, and show that approximate positive-noise separability does not in general imply approximate clean independence in a prescribed frame.
3. **Empirical tests of sharing.** Controlled and learned-denoiser experiments recover common frames without architectural sharing, demonstrate the finite-sample crossover, and connect latent factorization to cross-noise geometry in dSprites and CIFAR-10 representations.

2 Related Work

Gaussian denoising, separability, and ICA. Tweedie's identity links Gaussian posterior means to scores (Efron, 2011), and Bayes-denoiser Jacobians to conditional covariance (Dytso et al., 2023). Hessian criteria for function factorization and recovery of factorizing directions are classical (Lin, 1997; Theis, 2004); related work studies Gaussian estimation and ICA recovery (Gribonval & Machart, 2013; Comon, 1994; Yeredor, 2000; Arora et al., 2012; Belkin et al., 2013). Our result concerns a different implication: a fixed Jacobian frame of the *Bayes denoiser* at *one positive noise level*, across every input, characterizes independence of the *clean law* (including affine rank-deficient support) and yields a common denoising frame at *every noise level*.

Diffusion geometry and structured diffusion. Recent work studies the geometry of learned denoisers and scores, finding low-dimensional or semantically meaningful Jacobian directions together with substantial input dependence of their eigenbases (Kadkhodaie et al., 2024; Kouzelis et al., 2024). Other theoretical work shows that diffusion learners can exploit low-dimensional or factorized structure statistically (Boffi et al., 2025; Kwon et al., 2026; Fan et al., 2025). We instead ask when the Bayes denoising family itself admits the *same* representation across noise levels, and what is gained or lost by enforcing that sharing.

Information–estimation identities and latent factorization. The I–MMSE and mismatched-estimation identities connect Gaussian estimation error to relative entropy (Guo et al., 2005; Verdú, 2010); likelihood-weighted diffusion results give generative KL bounds and, for diffusion-consistent exact scores, identities (Song et al., 2021; Kadkhodaie et al., 2024). We apply this connection to the product of transformed marginals: the resulting equality interprets total correlation as accumulated coordinatewise Bayes-denoising regret. Related latent factorization objectives include FactorVAE and β-TCVAE (Kim & Mnih, 2018; Chen et al., 2018); unlike semantic disentanglement, our concern is the denoising consequence of statistical dependence.

3 One-Noise Geometry Characterizes Cross-Noise Sharing

We begin with the population question underlying representation sharing. Let $X \in \mathbb{R}^n$ have finite second moment and consider the Gaussian channel $Y_t = X + tG$, with $G \sim \mathcal{N}(0, I_n)$ and $t > 0$, whose Bayes-optimal denoiser is $f_t^\star(y) = \mathbb{E}[X \mid Y_t = y]$. If p_t denotes the density of Y_t, Gaussian smoothing yields the classical identities

$$\nabla \log p_t(y) = \frac{f_t^\star(y) - y}{t^2}, \qquad Df_t^\star(y) = \frac{1}{t^2}\,\mathrm{Cov}(X \mid Y_t = y). \tag{2}$$

Thus every Bayes Jacobian is symmetric positive semidefinite. More importantly, equation 2 connects the local geometry of the denoiser to the distribution of the noisy observation.

Fixed denoising frames. Let $U \in \mathrm{St}(n,k) := \{U \in \mathbb{R}^{n\times k} : U^\top U = I_k\}$ be an orthonormal frame, and define $\mathcal{A}_U := \{UDU^\top : D \in \mathbb{R}^{k\times k} \text{ diagonal}\}$. The condition $Df_t^\star(y) \in \mathcal{A}_U \qquad \forall y$ means that the denoiser may respond differently at different inputs, but its local action never mixes the k directions of U. A denoising path with

$$f_t(y) = b + Um_t(U^\top y), \qquad [m_t(z)]_i = m_{i,t}(z_i), \tag{3}$$

for some $b \perp \mathrm{range}(U)$ therefore has the same separable representation at every noise level: $U^\top$ encodes into fixed coordinates, each coordinate is processed independently, and U maps back to the ambient space. Differentiating equation 3 gives $Df_t(y) \in \mathcal{A}_U$.

The nontrivial question is the converse: if this differential structure is observed at only *one* noise level, does it reveal anything about the clean distribution or about other noise levels?

Theorem 3.1 (One-noise characterization)**.** *Fix $t_0 > 0$ and $U \in \mathrm{St}(n,k)$. The following are equivalent:*

- *(i) $X = b + U\Psi$ a.s. for some $b \in \mathrm{range}(U)^\perp$ and $\Psi = (\Psi_1, \ldots, \Psi_k)$ with mutually independent coordinates;*
- *(ii) $Df_{t_0}^\star(y) \in \mathcal{A}_U$ for every $y \in \mathbb{R}^n$.*

Under either condition, $f_{t_0}^\star$ has the coordinatewise form equation 3.

The forward implication is immediate. Under (i), $U^\top Y_t = \Psi + t\varepsilon$, so the independent prior and Gaussian likelihood factorize coordinatewise, and hence so does the posterior mean.

Conversely, (ii) implies that $f_{t_0}^\star$ is constant along directions orthogonal to $\mathrm{range}(U)$ and that each tangent component depends only on its corresponding coordinate. By Tweedie's identity equation 2, the score of p_{t_0} is therefore coordinatewise in the same frame. Integrating the score gives a noisy law with independent tangent coordinates and an independent Gaussian normal component. Since

the Gaussian characteristic function is everywhere nonzero, Gaussian deconvolution transfers this factorization back to the clean law, giving (i). See Appendix A.1 for the full proof (including $k < n$) and Appendix A.2 for corollary 3.2.

Gaussianity plays two distinct roles here. It provides the posterior-mean/score identity, and it makes exact deconvolution injective. The latter step will also explain why the corresponding *approximate* converse is unstable in section 4.

Corollary 3.2 (One noise level determines the full path)**.** *If condition* (ii) *of Theorem 3.1 holds for some $t_0 > 0$, then for every $t > 0$,*

$$f_t^\star(y) = b + U m_t(U^\top y), \qquad [m_t(z)]_i = \mathbb{E}[\Psi_i \mid \Psi_i + t\varepsilon_i = z_i].$$

Hence the same frame U is Bayes-optimal across the entire Gaussian denoising path.

Importantly, corollary 3.2 is not a direct transport statement between noise levels. Rather, the geometry at t_0 reveals the independent-component structure of the clean law, which guarantees coordinatewise Bayes denoising in the same frame at every Gaussian noise level.

Consequence for reverse diffusion. By Tweedie's identity, corollary 3.2 also applies to the score used in reverse diffusion. At every noise level, the score can be expressed using the same frame U, with separate one-dimensional functions along its directions and a known Gaussian component in the orthogonal complement when $k < n$. With the exact score and matching noisy initialization, the ideal reverse process recovers the clean distribution in the zero-noise limit.

A frame-free formulation. The characterization above assumes a candidate frame U. At the Bayes level the frame can instead be detected intrinsically from the Jacobian family. Let $J(y) = Df_{t_0}^\star(y), \qquad S = \operatorname{span}\{\operatorname{range}(J(y)) : y \in \mathbb{R}^n\}$. If $\dim S = k$ and $[J(y), J(y')] = 0 \qquad \forall y, y'$, then the real symmetric matrices $J(y)$ admit a common orthonormal eigenbasis on S. Consequently there exists $U \in \mathrm{St}(n, k)$ for which $J(y) \in \mathcal{A}_U$ for every y, and theorem 3.1 applies.

This gives an intrinsic exact certificate: pairwise commutation of Bayes Jacobians implies an independent-component clean law and a common frame across all noise levels (proof: Appendix B.1–B.2). For learned denoisers we use the corresponding commutation residual only as an empirical diagnostic. Small commutators do not quantitatively certify independence in a prescribed frame; the frame-optimized case is not resolved by our counterexample in section 4.

4 Approximate Sharing: Regret, Tying Cost, and Its Limits

The characterization in theorem 3.1 is exact, whereas learned representations are never perfectly independent. We therefore ask two approximate questions: how costly is coordinatewise denoising in an approximately factorial frame, and can near-separability of a positive-noise denoiser certify approximate independence of the clean law? The two directions behave very differently.

Coordinatewise denoising regret. Let $Z \in \mathbb{R}^k$, $V \in O(k)$, $W = V^\top Z$, and $Z_t = Z + t\varepsilon$ with $\varepsilon \sim \mathcal{N}(0, I_k)$. Besides the Bayes denoiser $f_t^\star(z) = \mathbb{E}[Z \mid Z_t = z]$, consider the coordinatewise denoiser

$$g_{V,t}(z) = V m_{V,t}(V^\top z), \qquad [m_{V,t}(w)]_i = \mathbb{E}[W_i \mid W_i + t\varepsilon_i = w_i], \tag{4}$$

which is the Bayes denoiser for the product of the true one-dimensional marginals of W. Its excess Bayes risk is

$$\begin{aligned} \mathcal{R}_V(t) &:= \mathbb{E}\|Z - g_{V,t}(Z_t)\|^2 - \mathbb{E}\|Z - f_t^\star(Z_t)\|^2 \\ &= \mathbb{E}\|f_t^\star(Z_t) - g_{V,t}(Z_t)\|^2 \geq 0. \end{aligned} \tag{5}$$

Thus $\mathcal{R}_V(t)$ is precisely the price of discarding cross-coordinate information in frame V. Write $\mathrm{TC}_V(t) := \mathrm{TC}(W + t\varepsilon)$ for the total correlation after Gaussian smoothing.

Theorem 4.1 (Total correlation equals accumulated denoising regret)**.** *For every $0 < t_0 < T < \infty$ for which the quantities are finite,*

$$\mathrm{TC}_V(t_0) - \mathrm{TC}_V(T) = \int_{t_0}^{T} \frac{\mathcal{R}_V(t)}{t^3}\, dt. \tag{6}$$

If additionally $\mathbb{E}\|Z\|^2 < \infty$ *and* $\mathrm{TC}(V^\top Z) < \infty$*, then*

$$\mathrm{TC}(V^\top Z) = \int_0^\infty \frac{\mathcal{R}_V(t)}{t^3}\, dt. \tag{7}$$

Hence total correlation is exactly the accumulated Bayes regret of coordinatewise denoising, with low-noise errors weighted more heavily by t^{-3}. The proof specializes the Gaussian mismatched-estimation identity (Verdú, 2010): for $P = \mathcal{L}(W)$ and $Q = \prod_i P_{W_i}$, Gaussian convolution gives $\mathrm{KL}(P_t\|Q_t) = \mathrm{TC}_V(t)$, while relative de Bruijn and Tweedie's identities yield $\frac{d}{dt}\mathrm{KL}(P_t\|Q_t) = -\mathcal{R}_V(t)/t^3$ (Appendix C.1).

Generative interpretation. Let $Q_V := V_\#(\prod_i P_{W_i})$. The path $\{g_{V,t}\}_{t>0}$ is the exact Bayes path of this *single* product-of-marginals law. An ideal reverse process driven by its score and initialized from its matching noisy law therefore generates Q_V. Combining this fact with eq. (7) yields

$$\mathrm{KL}(P_Z\|Q_V) = \mathrm{TC}(V^\top Z) = \int_0^\infty \frac{\mathcal{R}_V(t)}{t^3}\, dt. \tag{8}$$

The exact equality uses a diffusion-consistent Bayes path, not an arbitrary learned denoiser. In general, likelihood-weighted denoising error instead gives a generative KL *bound*, with initialization and endpoint conditions (Song et al., 2021; Kadkhodaie et al., 2024); see Appendix C.3. Proof of the product-law generative identity and further noise-schedule consequences appear in Appendices C.2 and C.1.

Implications for learned latent representations. FactorVAE and β-TCVAE encourage low aggregate total correlation (Kim & Mnih, 2018; Chen et al., 2018). By eqs. (7) and (8), reducing this quantity in a fixed frame reduces integrated coordinatewise Bayes regret and the ideal product-model KL discrepancy, without guaranteeing better neural latent diffusion performance. Appendix I.1 tests the regret identity on learned VAE-family representations. The additional cost of tying frames across noise levels is considered next.

Factorization cost versus tying cost. Theorem 4.1 measures the cost of coordinatewise processing in a fixed frame, not the additional cost of reusing one frame across noise levels. Over $[t_0, T]$, define

$$\begin{aligned} \mathcal{B}_{\mathrm{share}} &:= \inf_{V\in O(k)} \int_{t_0}^T \frac{\mathcal{R}_V(t)}{t^3}\, dt, \qquad \mathcal{B}_{\mathrm{untied}} := \int_{t_0}^T \frac{\inf_{V\in O(k)} \mathcal{R}_V(t)}{t^3}\, dt, \\ \Delta_{\mathrm{tie}} &:= \mathcal{B}_{\mathrm{share}} - \mathcal{B}_{\mathrm{untied}} \geq 0. \end{aligned} \tag{9}$$

Here $\mathcal{B}_{\mathrm{untied}}$ is the coordinatewise approximation error even when each noise level chooses its own frame, whereas Δ_{tie} isolates the extra approximation cost of sharing. Both vanish under the exact setting of theorem 3.1. More generally, $\Delta_{\mathrm{tie}} > 0$ can coexist with a finite-sample advantage for sharing because one common frame is easier to estimate than many unrelated ones; section 5 provides sufficient bounds for the estimation-complexity component of this tradeoff. Unlike a fixed frame, noise-dependent optimal frames need not define the Bayes path of a single law, so Δ_{tie} is not itself an exact generative KL gap.

Why the approximate converse fails. Small total correlation therefore guarantees small integrated coordinatewise regret. The reverse implication is fundamentally unstable: Gaussian smoothing can suppress fine-scale dependence arbitrarily strongly at every fixed positive noise level.

Consider the smooth two-dimensional family $p_m(x_1, x_2) = \varphi(x_1)\varphi(x_2)\big[1+\epsilon\sin(mx_1)\sin(mx_2)\big]$, with $0 < \epsilon < 1$ and φ the standard Gaussian density.

Proposition 4.2 (No distribution-free fixed-frame approximate converse)**.** *Both marginals of* P_m *are standard Gaussian and* $\inf_{m\geq 1} \mathrm{TC}(P_m) > 0$*. Nevertheless, for every fixed* $0 < \tau < T < \infty$*,*

$$\begin{aligned} \sup_{t\in[\tau,T]} \sup_y \left\| f^\star_{m,t}(y) - \frac{y}{1+t^2} \right\| &\longrightarrow 0, \\ \sup_{t\in[\tau,T]} \sup_y \left\| Df^\star_{m,t}(y) - \frac{1}{1+t^2} I \right\|_{\mathrm{op}} &\longrightarrow 0 \end{aligned} \tag{10}$$

as $m \to \infty$*.*

The total-correlation lower bound concerns the displayed coordinates; the construction does not establish a uniform lower bound after optimizing over orthogonal frames.

The dependence is hidden because Gaussian smoothing attenuates the oscillatory interaction by $\exp\{-m^2t^2/(1+t^2)\}$, which vanishes exponentially in m^2 whenever t is bounded away from zero. Full bounds and the no-modulus corollary are in Appendix C.4–C.5; Appendix C.6 shows that even cross-input, cross-noise commutators can vanish. Thus genuinely dependent clean laws can have positive-noise Bayes paths arbitrarily close, even at the Jacobian level, to that of an independent Gaussian model.

The distinction is *which noise scales are observed*: under theorem 4.1, clean total correlation equals all-noise integrated regret, but near-separability at noise bounded away from zero need not imply small clean total correlation. Approximate commutation and one-noise transfer are geometric diagnostics, not quantitative independence certificates in a prescribed frame.

5 Why Sharing Can Help at Finite Sample

The previous section isolated the approximation cost Δ_{tie} of reusing one frame across noise levels. With finite data, however, sharing can still be preferable because all observations contribute to estimating the same representation rather than several unrelated ones.

A solvable shared-frame model. Let $U_\star \in \mathrm{St}(n,k)$ and $X = U_\star \Psi$ with $\Psi = aS + \xi$, where $S_i \overset{\mathrm{iid}}{\sim} \mathrm{Unif}\{-1,+1\}$, $\xi \sim \mathcal{N}(0, \beta I_k)$, and $a, \beta > 0$. Since the coordinates of Ψ are independent, theorem 3.1 gives the same Bayes frame $U_\star$ at every noise level. The corresponding Bayes denoising path is available in closed form (Appendix D.2), making this model a convenient setting in which to isolate the statistical benefit of estimating one common frame.

Assume a and β are known and only the frame is learned. For i.i.d. training triples with $\mathsf{T} \sim \nu$, $Y = X + \mathsf{T}G$, and $G \sim \mathcal{N}(0, I_n)$, let $\mathcal{L}_\nu(U) = \mathbb{E}\|X - f_U(Y,\mathsf{T})\|^2$. The true frame is well specified, $f_{U_\star}(Y,\mathsf{T}) = \mathbb{E}[X \mid Y, \mathsf{T}]$. Writing $d_{n,k} = nk - k(k+1)/2$ for the dimension of $\mathrm{St}(n,k)$ gives the following finite-sample guarantee.

Theorem 5.1 (Finite-sample bound for shared-frame learning)**.** *For fixed a, β, k, let $\widehat{U}$ be an empirical-risk minimizer over a suitable finite operator-norm net of $\mathrm{St}(n,k)$ using N independent noisy training triples. With probability at least $1-\delta$,*

$$\mathcal{L}_\nu(\widehat{U}) - \mathcal{L}_\nu(U_\star) = \widetilde{O}_{a,\beta,k}\left(\sqrt{\frac{d_{n,k} + \log(1/\delta)}{N}} + \frac{1 + d_{n,k} + \log(1/\delta)}{N}\right), \tag{11}$$

uniformly over the noise-level distribution ν. The full statement and proof are in Appendix E; the untied-frame comparison is in Appendix E.5.

Shared versus untied representations. For L noise levels under the same total budget N, a shared learner estimates one Stiefel frame, whereas an untied learner estimates L unrelated frames. The corresponding parameter complexity therefore grows from $d_{n,k}$ to $Ld_{n,k}$. Ignoring logarithmic and lower-order terms, the leading estimation contributions scale as $O(\sqrt{d_{n,k}/N})$ for sharing and $O(\sqrt{Ld_{n,k}/N})$ for the untied model. Thus sharing pools information across noise levels instead of paying representation complexity separately at each level. This is a sufficient upper-bound comparison rather than a minimax separation, and N counts independent noisy triples. Since the bound controls risk under the training noise law ν, it does not by itself control the likelihood-weighted full-path error needed for a generative KL bound.

Combined with eq. (9), this suggests a finite-data tradeoff: sharing can reduce representation-estimation error but may incur a persistent tying bias. When $\Delta_{\mathrm{tie}} = 0$, sharing incurs no additional approximation cost, though the upper bounds do not guarantee a finite-sample advantage for every estimator. Under misspecification, the two effects can reverse the empirical ordering as sample size grows; section 6 tests both regimes.

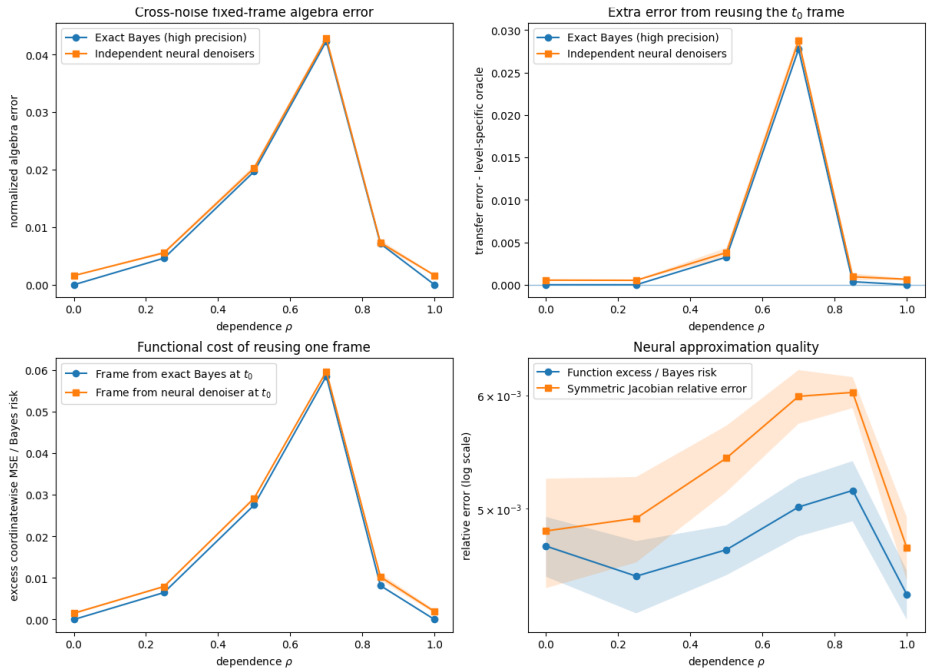


Figure 1: **Cross-noise sharing without shared weights.** A frame recovered only at $t_0 = 0.5$ transfers almost perfectly at the factorial endpoints but degrades when no common factorizing frame exists. Learned denoisers closely track the exact Bayes reference; the same pattern appears in excess algebra error and functional excess MSE. Learned-Jacobian algebra diagnostics use symmetrized network Jacobians. Bands are 95% intervals over ten outer seeds.

6 Experiments

We test common-frame recovery without weight sharing, the finite-sample tradeoff under correct specification and misspecification, and the relationship between latent dependence and denoising geometry. Full protocols and robustness checks are in Appendices F–I.

6.1 Independent Denoisers Recover the Statistical Shared Frame

We use the Rademacher–Gaussian model from section 5, with dependence parameter ρ chosen so that $\rho = 0$ and $\rho = 1$ are factorial in two different orthogonal frames, while intermediate values admit no common factorizing frame. Separate denoisers are trained at $t \in \{0.2, 0.5, 0.8, 1.1\}$ with independent initializations, training streams, no time input, and no shared parameters. A frame is fitted only at $t_0 = 0.5$ and then frozen across the other noise levels; the exact Bayes posterior provides a reference. Bayes-approximation and Jacobian errors are checked across all 240 independently trained denoisers (Appendix F.1–F.6).

Figure 1 follows the population prediction. The learned cross-noise algebra error is only 1.58×10^{-3} at $\rho = 0$ and 1.64×10^{-3} at $\rho = 1$, where the exact Bayes error is numerically zero, but rises to 4.28×10^{-2} at $\rho = 0.70$, matching the Bayes value 4.24×10^{-2}. The effect is also functional: freezing the one-noise frame costs 5.97% of Bayes risk at $\rho = 0.70$, versus only 0.15% and 0.19% at the factorial endpoints. Thus the shared geometry appears without architectural sharing precisely in the regimes where the data distribution admits it.

6.2 Sharing Trades Estimation Error Against Tying Bias

We next test the finite-data mechanism of section 5. In the correctly specified Rademacher–Gaussian model, a shared estimator fits one frame from all N training triples, whereas an untied estimator

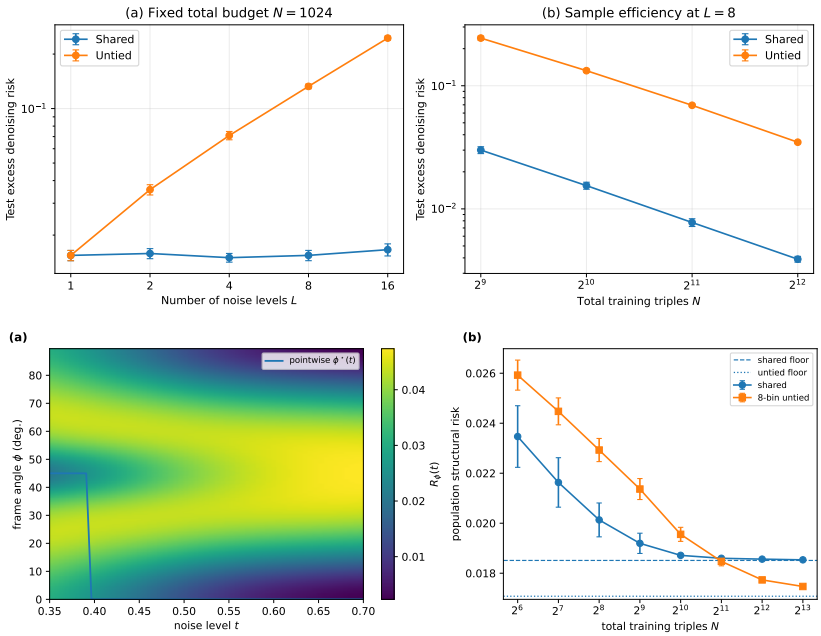


Figure 2: **The statistical benefit and approximation cost of sharing. Top:** under correct specification, sharing pools observations across noise levels while the untied estimator divides the same total budget among multiple frames. **Bottom:** under misspecification, sharing wins at small sample size but eventually loses when its persistent tying bias dominates.

divides the same total budget across L independently estimated frames. At $N = 1024$, shared excess risk stays near 0.015–0.017 as L grows from 1 to 16, while untied risk rises from 0.0154 to 0.2451. Fixed-per-level and i.i.d.-noise controls support this result; untied training loss is no higher in $591/600$ comparisons (Appendix G.3–G.5).

To test misspecification, we use a fixed clean distribution with structure at two orientations, so the preferred coordinatewise frame changes with noise level. Over $t \in [0.35, 0.70]$, the normalized, grid-restricted tying penalty is approximately 1.45×10^{-3}; it vanishes on control windows where the optimal frame stays fixed (Appendix H.3–H.5), with the full sweep in Appendix H.6.

The crossover in Figure 2 is direct: sharing performs better at $N = 64$ and remains advantageous through $N = 1024$; the difference is statistically indistinguishable from zero around $N = 2048$; by $N = 4096$ the untied estimator is better (0.01773 versus 0.01856). The 100-seed population-risk evaluation excludes test-set noise from this estimation-versus-tying crossover (Appendix H.4–H.6).

6.3 Learned Image Representations

VAE-family representations. Across 20 ten-dimensional dSprites representations from VAE, β-VAE, FactorVAE, and β-TCVAE models, the empirical aggregate posterior is represented as a finite diagonal-Gaussian mixture, making its corrupted posterior mean and Jacobian exact for the plug-in distribution; thus these theory-facing quantities contain no learned-denoiser approximation error (Appendix I.2–I.5). For every representation seed, the measured total-correlation drop over $t \in [0.2, 1.1]$ closely matches the independently integrated denoising regret predicted by theorem 4.1. Across runs, the Bayes-Jacobian commutator tracks clean plug-in total correlation strongly (pooled Spearman 0.983), while one-noise basis transfer is less monotone (0.687), consistent with the asymmetry in section 4; complete results and mixture-size checks are in Appendix I.6–I.8.

Controlled factorization in CIFAR-10 latents. We use standardized 32-dimensional latents from a convolutional autoencoder and compare the original empirical latent law with marginal-matched coordinate-shuffled controls, before and after a planted random rotation. For each condition, five

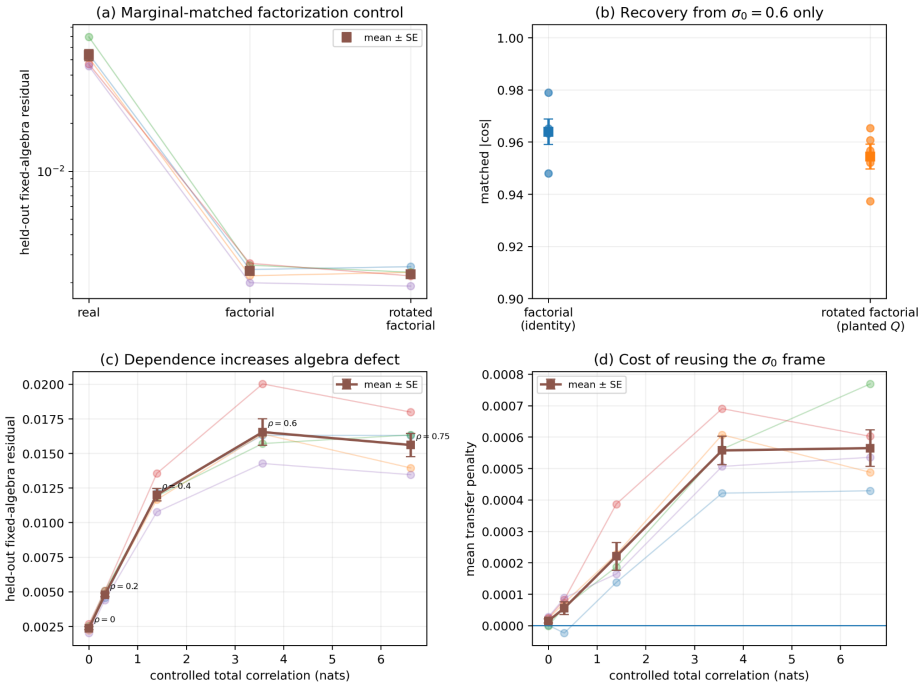


Figure 3: **Controlled factorization in learned CIFAR-10 latents.** Coordinate shuffling sharply reduces symmetrized fixed-frame algebra error; a frame fitted at one noise level aligns with the coordinate directions and a planted random rotation; introducing controlled dependence increases both the fixed-frame defect and one-noise transfer error. In panels (c–d), the horizontal axis is nominal total correlation for the ideal Gaussian-copula intervention. Error bars are standard errors across five independently trained denoisers.

time-conditioned denoisers are trained; a frame is fitted only at $\sigma_0 = 0.6$ and frozen across the other noise levels (Appendix I.9–I.12).

As shown in Figure 3, at $\sigma_0 = 0.6$, the fixed-frame residual falls from 0.05356 ± 0.00427 for the original latents to 0.002372 ± 0.000119 for the factorial control and 0.002258 ± 0.000101 after the planted rotation. The recovered frames align with the planted reference directions (matched cosine 0.9639 ± 0.0049 and 0.9544 ± 0.0048, respectively), and their cross-noise transfer error is far below that of the original latent law. Starting from the shuffled control, a Gaussian-copula dependence intervention raises the fixed-frame residual from 0.00236 at $\rho = 0$ to 0.01654 at $\rho = 0.6$ before saturation. Because these CIFAR denoisers share time-conditioned weights, this tests the predicted geometry in a learned representation, not architecture-free emergence of sharing (Appendix I.13–I.18). In light of proposition 4.2, we interpret the dependence intervention as a controlled geometric response, not as a distribution-free quantitative converse.

7 Discussion and Conclusion

We asked when the data, rather than the architecture, justify sharing a denoising frame. For Gaussian noise, a fixed Bayes-Jacobian frame at every input at one positive noise level characterizes an affine orthogonal independent-component clean law and hence a common frame throughout the denoising path. In a chosen frame, total correlation equals integrated coordinatewise Bayes regret and the ideal product-model KL discrepancy. Sharing frames can reduce finite-sample estimation error but may incur additional tying bias. Near-separability at positive noise does not generally certify approximate clean independence in a prescribed frame. Our theory concerns Bayes denoisers, orthogonal frames, and isotropic Gaussian corruption; its finite-sample bounds are not minimax guarantees. Experiments on learned latents support the predicted geometry without establishing improved end-to-end generation. Quantitative stability after optimizing over frames remains open.

AI USE STATEMENT

Generative AI tools assisted with manuscript editing, brainstorming research directions, developing mathematical arguments and proof strategies, reviewing proofs, providing feedback on experimental design and analysis, and writing and debugging experimental code. The authors independently examined and verified the mathematical arguments, reviewed and validated the experimental code and results, and take full responsibility for the final manuscript.

REFERENCES

Brian D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, May 1982. ISSN 0304-4149. doi: 10.1016/0304-4149(82)90051-5.

Sanjeev Arora, Rong Ge, Ankur Moitra, and Sushant Sachdeva. Provable ICA with unknown gaussian noise, with implications for gaussian mixtures and autoencoders. In F. Pereira, C.J. Burges, L. Bottou, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Mikhail Belkin, Luis Rademacher, and James Voss. Blind signal separation in the presence of gaussian noise. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pp. 270–287. PMLR, 2013.

Nicholas M. Boffi, Arthur Jacot, Stephen Tu, and Ingvar Ziemann. Single hidden layer diffusion models provably learn simple low-dimensional structure. In *International Conference on Learning Representations*, 2025.

Ricky T. Q. Chen, Xuechen Li, Roger B Grosse, and David K. Duvenaud. Isolating Sources of Disentanglement in Variational Autoencoders. In *Neural Information Processing Systems*, pp. 2610–2620, 2018.

Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.

Alex Dytso, H. Vincent Poor, and Shlomo Shamai. Conditional mean estimation in gaussian noise: A meta derivative identity with applications. *IEEE Transactions on Information Theory*, 69(3): 1883–1898, 2023.

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

Jianqing Fan, Yihong Gu, and Ximing Li. Optimal estimation of a factorizable density using diffusion models with ReLU neural networks. *arXiv preprint arXiv:2510.03994*, 2025.

Rémi Gribonval and Pierre Machart. Reconciling "priors" and "priors" without prejudice? In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Dongning Guo, Shlomo Shamai, and Sergio Verdú. Mutual information and minimum mean-square error in gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1261–1282, 2005. doi: 10.1109/TIT.2005.844072.

Zahra Kadkhodaie, Florentin Guth, Eero P Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *The Twelfth International Conference on Learning Representations*, 2024.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pp. 2649–2658, 2018.

Theodoros Kouzelis, Emmanouil Plitsis, Mihalis Nicolaou, and Yannis Panagakis. Enabling local editing in diffusion models by joint and individual component analysis. In *35th British Machine Vision Conference 2024, BMVC 2024, Glasgow, UK, November 25-28, 2024*. BMVA, 2024.

Hyeok Kyu Kwon, Dongha Kim, Ilsang Ohn, and Minwoo Chae. Nonparametric estimation of a factorizable density using diffusion models. *Journal of Machine Learning Research*, 27(22): 1–125, 2026.

Juan Lin. Factorizing multivariate function classes. In M. Jordan, M. Kearns, and S. Solla (eds.), *Advances in Neural Information Processing Systems*, volume 10. MIT Press, 1997.

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

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34:1415–1428, 2021.

Fabian J. Theis. A new concept for separability problems in blind source separation. *Neural Computation*, 16(9):1827–1850, 2004. doi: 10.1162/0899766041336404.

Sergio Verdú. Mismatched estimation and relative entropy. *IEEE Transactions on Information Theory*, 56(8):3712–3720, 2010. doi: 10.1109/TIT.2010.2050800.

Arie Yeredor. Blind source separation via the second characteristic function. *Signal Processing*, 80 (5):897–902, 2000. doi: 10.1016/S0165-1684(00)00062-1.

A Proof of the One-Noise Characterization

This section proves theorem 3.1 and the full-path consequence stated in corollary 3.2. We include the complete argument because the rank-deficient case $k < n$ requires keeping track of both the active subspace $\mathrm{range}(U)$ and its orthogonal complement.

A.1 Proof of theorem 3.1

Proof. We prove the two implications separately.

(i)⇒(ii). Assume $X = b + U\Psi \qquad$ a.s., where $b \in \mathrm{range}(U)^\perp$ and $\Psi_1, \dots, \Psi_k$ are mutually independent.

Complete U to an orthogonal matrix $[U\ U_\perp] \in O(n)$. Writing $b_\perp := U_\perp^\top b$, the Gaussian observation at noise level t_0 becomes, in rotated coordinates, $U^\top Y_{t_0} = \Psi + t_0\varepsilon, \qquad U_\perp^\top Y_{t_0} = b_\perp + t_0\varepsilon_\perp$, where $\varepsilon \sim \mathcal{N}(0, I_k), \qquad \varepsilon_\perp \sim \mathcal{N}(0, I_{n-k})$, and the two Gaussian blocks are independent.

The normal observation $U_\perp^\top Y_{t_0}$ contains no information about Ψ. Moreover, because the prior law of Ψ is a product and the Gaussian likelihood factorizes coordinatewise, the posterior distribution of Ψ also factorizes. Hence $f^\star_{t_0}(y) = b + Um(U^\top y)$, where $m_i(z_i) = \mathbb{E}[\Psi_i \mid \Psi_i + t_0\varepsilon_i = z_i]$. Differentiating gives

$$Df^\star_{t_0}(y) = U\ \mathrm{diag}\big(m_1'(u_1^\top y), \dots, m_k'(u_k^\top y)\big)\, U^\top \in \mathcal{A}_U.$$

which proves (ii).

(ii)⇒(i). Now assume $Df^\star_{t_0}(y) \in \mathcal{A}_U \qquad \forall y \in \mathbb{R}^n$. Write $f := f^\star_{t_0}, \qquad Df(y) = UD(y)U^\top$, with $D(y)$ diagonal.

We first show that the posterior mean itself separates in the U-frame.

If $v \in \mathrm{range}(U)^\perp$, then $Df(y)v = 0$. Therefore $\frac{d}{ds} f(y + sv) = 0$, so f is constant along every affine fiber orthogonal to $\mathrm{range}(U)$.

Moreover, $U_\perp^\top Df(y) = 0$, so $U_\perp^\top f(y)$ is constant over $\mathbb{R}^n$. Denote this constant by $b_\perp$ and set $b := U_\perp b_\perp \in \mathrm{range}(U)^\perp$. It remains to examine the tangent coordinates. For $z \in \mathbb{R}^k$, define

$h_i(z) := u_i^\top f(Uz)$. For $i \neq j$, $\frac{\partial h_i}{\partial z_j}(z) = u_i^\top Df(Uz)u_j = 0$. Thus h_i depends only on z_i. Hence there exist scalar functions $m_i : \mathbb{R} \to \mathbb{R}$ such that

$$f(y) = b + \sum_{i=1}^{k} u_i\, m_i(u_i^\top y). \tag{12}$$

Factorization of the noisy law. Let p_{t_0} denote the density of Y_{t_0}. In the orthogonal coordinates $y = Uz + U_\perp w$, define $\widetilde{p}(z, w) := p_{t_0}(Uz + U_\perp w)$. By Tweedie's formula, $\nabla \log p_{t_0}(y) = \frac{f(y)-y}{t_0^2}$. Combining this with equation 12 yields, for every $i = 1, \dots, k$,

$$\frac{\partial}{\partial z_i} \log \widetilde{p}(z, w) = \frac{m_i(z_i) - z_i}{t_0^2}, \tag{13}$$

while in the orthogonal complement,

$$\nabla_w \log \widetilde{p}(z, w) = \frac{b_\perp - w}{t_0^2}. \tag{14}$$

For each i, choose an antiderivative A_i satisfying $A_i'(z_i) = \frac{m_i(z_i)-z_i}{t_0^2}$. Consider $F(z, w) := \log \widetilde{p}(z, w) - \sum_{i=1}^{k} A_i(z_i) + \frac{|w-b_\perp|^2}{2t_0^2}$. The two score identities above show that $\nabla F(z, w) = 0$. Since the domain is connected, F is constant. After normalization,

$$\widetilde{p}(z, w) = \left(\prod_{i=1}^{k} p_{i,t_0}(z_i) \right) \varphi_{t_0^2 I}(w - b_\perp), \tag{15}$$

for suitable one-dimensional densities p_{i,t_0}. This score-integration step is the classical log-Hessian factorization mechanism (Lin, 1997; Theis, 2004). The subsequent deconvolution steps establish the additional conclusion about the clean law and its affine support.

Thus the tangent coordinates $U^\top Y_{t_0}$ are mutually independent, while $U_\perp^\top Y_{t_0} \sim \mathcal{N}(b_\perp, t_0^2 I)$ is independent of them.

Recovering the clean support. Let $B := U_\perp^\top X$. Since $U_\perp^\top Y_{t_0} = B + t_0 \varepsilon_\perp \sim \mathcal{N}(b_\perp, t_0^2 I)$, their characteristic functions satisfy $\phi_B(v) e^{-t_0^2|v|^2/2} = e^{i\langle b_\perp, v\rangle} e^{-t_0^2|v|^2/2}$. The Gaussian factor is nonzero for every v, so it can be canceled: $\phi_B(v) = e^{i\langle b_\perp, v\rangle}$. Therefore $B = b_\perp$ a.s. and the clean distribution is supported on the affine subspace $b + \mathrm{range}(U)$.

Recovering clean independence. Define $\Psi := U^\top X$, $Z := U^\top Y_{t_0} = \Psi + t_0 \varepsilon$. By equation 15, the coordinates of Z are mutually independent, hence

$$\phi_Z(s) = \prod_{i=1}^{k} \phi_{Z_i}(s_i). \tag{16}$$

Gaussian convolution gives $\phi_Z(s) = \phi_\Psi(s) e^{-t_0^2\|s\|^2/2}$ and, coordinatewise, $\phi_{Z_i}(s_i) = \phi_{\Psi_i}(s_i) e^{-t_0^2 s_i^2/2}$. Substituting these expressions into equation 16 and canceling the nonvanishing Gaussian factors yields $\phi_\Psi(s) = \prod_{i=1}^{k} \phi_{\Psi_i}(s_i)$. Therefore $\Psi_1, \dots, \Psi_k$ are mutually independent.

Combining this with the deterministic normal component gives $X = b + U\Psi$ a.s., which proves (i).

Finally, suppose the scalar denoisers at t_0 are tied, $m_1 = \cdots = m_k$. Then equation 13 shows that all one-dimensional noisy marginals have the same score. Their log-densities therefore differ only by constants, which normalization removes. Hence their noisy distributions coincide. Since Gaussian convolution has a nonvanishing characteristic function, deconvolution implies that the clean marginals of $\Psi_1, \dots, \Psi_k$ are also identical. □

A.2 Proof of corollary 3.2

Proof. By theorem 3.1, $X = b + U\Psi$, where the coordinates of Ψ are mutually independent. At any noise level $t > 0$, $U^\top Y_t = \Psi + t\varepsilon, \qquad U_\perp^\top Y_t = b_\perp + t\varepsilon_\perp$. The posterior over Ψ therefore factorizes coordinatewise, exactly as in the forward implication above. Consequently, $f_t^\star(y) = b + Um_t(U^\top y)$, with $[m_t(z)]_i = \mathbb{E}[\Psi_i \mid \Psi_i + t\varepsilon_i = z_i]$. The same frame U thus represents the Bayes denoiser for every $t > 0$. □

The proof relies on the injectivity of Gaussian convolution. Since the Gaussian characteristic function never vanishes, independence of the noisy coordinates implies independence of the clean coordinates. However, this implication is not uniformly stable under approximation, as shown in proposition 4.2.

A.3 Variance-preserving Gaussian observations

Consider a variance-preserving Gaussian observation $Z_s = \alpha_s X + \sigma_s G$, where $\alpha_s > 0$, $\sigma_s > 0$, and $G \sim \mathcal{N}(0, I_n)$ is independent of X. Rescaling gives $Z_s/\alpha_s = X + t_s G$, with $t_s = \sigma_s/\alpha_s$. Consequently, the posterior-mean denoiser $h_s(z) := \mathbb{E}[X \mid Z_s = z]$ satisfies

$$h_s(z) = f_{t_s}^\star(z/\alpha_s), \qquad Dh_s(z) = \frac{1}{\alpha_s} Df_{t_s}^\star(z/\alpha_s).$$

The scalar multiplication and invertible input rescaling preserve the fixed diagonalizing frame. Thus the exact one-noise characterization applies to variance-preserving Gaussian observations whenever $\alpha_s, \sigma_s > 0$. This observation-level equivalence does not by itself transfer training objectives or reverse-process KL bounds between different noise parameterizations.

B Simultaneous Diagonalization and the Commutation Certificate

This section supplies the linear-algebra argument behind the frame-free formulation in section 3. The only fact we need is that a pairwise commuting family of real symmetric operators admits a common orthonormal eigenbasis.

B.1 Simultaneous diagonalization

Lemma B.1 (Simultaneous diagonalization)**.** *Let $\mathcal{M}$ be a pairwise commuting family of real symmetric operators on a finite-dimensional real inner-product space E: $AB = BA \qquad \forall A, B \in \mathcal{M}$. Then there exists an orthonormal basis of E consisting of common eigenvectors of every member of $\mathcal{M}$. Equivalently, there exists an orthogonal matrix Q such that $Q^\top AQ$ is diagonal for every $A \in \mathcal{M}$.*

Proof. We argue recursively on $\dim(E)$.

If every $A \in \mathcal{M}$ is a scalar multiple of the identity, then every orthonormal basis is simultaneously an eigenbasis and there is nothing to prove.

Otherwise, choose some $A \in \mathcal{M}$ having at least two distinct eigenvalues. Since A is real symmetric, the spectral theorem gives an orthogonal decomposition $E = \bigoplus_{\lambda \in \sigma(A)} E_\lambda$, where $E_\lambda = \{v \in E : Av = \lambda v\}$ is the eigenspace corresponding to λ.

We claim that every E_λ is invariant under every $B \in \mathcal{M}$. Indeed, if $v \in E_\lambda$, then $A(Bv) = B(Av) = B(\lambda v) = \lambda Bv$, where we used $AB = BA$. Hence $Bv \in E_\lambda$. Therefore each operator $B \in \mathcal{M}$ restricts to a symmetric operator $B|_{E_\lambda} : E_\lambda \to E_\lambda$. The restricted family $\mathcal{M}_\lambda := \{B|_{E_\lambda} : B \in \mathcal{M}\}$ remains pairwise commuting.

Because A has at least two distinct eigenvalues, every eigenspace E_λ has dimension strictly smaller than $\dim(E)$. By the induction hypothesis, on each E_λ there exists an orthonormal basis consisting of common eigenvectors of all operators in $\mathcal{M}_\lambda$.

The eigenspaces of a real symmetric operator corresponding to distinct eigenvalues are mutually orthogonal. Concatenating the common orthonormal eigenbases obtained on all E_λ therefore gives an orthonormal basis of E that simultaneously diagonalizes every member of $\mathcal{M}$. □

B.2 Proof of the commutation certificate

We now apply lemma B.1 to the Bayes Jacobian family.

Fix $t_0 > 0$ and write $J(y) := Df^\star_{t_0}(y)$. Recall from equation 2 that

$$J(y) = \frac{1}{t_0^2}\,\mathrm{Cov}(X \mid Y_{t_0} = y), \tag{17}$$

and hence every $J(y)$ is real symmetric and positive semidefinite.

Define the active subspace $S := \mathrm{span}\big\{\mathrm{range}(J(y)) : y \in \mathbb{R}^n\big\}$. Assume $\dim S = k$ and that the Jacobians commute pairwise:

$$[J(y), J(y')] := J(y)J(y') - J(y')J(y) = 0 \qquad \forall y, y' \in \mathbb{R}^n. \tag{18}$$

Proposition B.2 (Commutation certificate)**.** *Under the assumptions above, there exists $U \in \mathrm{St}(n,k)$ with $\mathrm{range}(U) = S$ and some $b \perp S$ such that $X = b + U\Psi$ a.s., where the coordinates of Ψ are mutually independent. Moreover, the same frame U gives the coordinatewise Bayes representation of $f^\star_t$ for every $t > 0$.*

Proof. By equation 17, every $J(y)$ is real symmetric. By definition of S, $\mathrm{range}(J(y)) \subseteq S \quad \forall y$. We first show that every $J(y)$ vanishes on $S^\perp$. Fix $v \in S^\perp$. For any $w \in \mathbb{R}^n$, symmetry gives $\langle J(y)v, w\rangle = \langle v, J(y)w\rangle$. But $J(y)w \in S$, while $v \perp S$. Therefore $\langle v, J(y)w\rangle = 0 \quad \forall w$, which implies

$$J(y)v = 0 \qquad \forall v \in S^\perp. \tag{19}$$

Thus each $J(y)$ leaves S invariant and acts trivially on its orthogonal complement. Consider now the restricted family $\mathcal{J}_S = \{J(y)|_S : y \in \mathbb{R}^n\}$. Every member of $\mathcal{J}_S$ is symmetric on S, and by equation 18 the family is pairwise commuting. By lemma B.1, there exists an orthonormal basis $u_1, \dots, u_k$ of S that simultaneously diagonalizes all $J(y)|_S$.

Let $U = [u_1 \ \cdots \ u_k] \in \mathrm{St}(n,k)$. For every y, there is therefore a diagonal matrix $D(y) = \mathrm{diag}\big(d_1(y), \dots, d_k(y)\big)$ such that the restriction of $J(y)$ to S is represented by $D(y)$ in this basis. Together with equation 19, this gives $J(y) = UD(y)U^\top \quad \forall y \in \mathbb{R}^n$. Equivalently, $Df^\star_{t_0}(y) \in \mathcal{A}_U \quad \forall y$. The hypothesis of theorem 3.1 is now satisfied. Consequently, there exist $b \in \mathrm{range}(U)^\perp = S^\perp$ and a random vector $\Psi = (\Psi_1, \dots, \Psi_k)$ with mutually independent coordinates such that

$$X = b + U\Psi \qquad \text{a.s.} \tag{20}$$

Finally, corollary 3.2 applies to equation 20, giving $f^\star_t(y) = b + Um_t(U^\top y) \quad \forall t > 0$, where each m_t acts coordinatewise. Thus the common eigenframe identified from the Jacobian family at the single level t_0 is also the Bayes-optimal frame along the entire Gaussian denoising path. □

Why symmetry matters. Pairwise commutation alone is not sufficient for arbitrary real matrices to guarantee orthogonal simultaneous diagonalization. Here symmetry comes automatically from the Bayes identity $Df^\star_t(y) = t^{-2}\,\mathrm{Cov}(X \mid Y_t = y)$, which is why the commutation condition can be converted directly into a fixed orthonormal frame.

Interpretation. The fixed-frame condition in theorem 3.1 appears to require that a candidate frame U be known in advance. Proposition B.2 removes that requirement. At the exact Bayes level, the condition can instead be expressed intrinsically through the family of Jacobians itself: $[J(y), J(y')] = 0 \quad \forall y, y'$. The shared frame is then simply their common eigenbasis on the active subspace.

For learned denoisers, however, approximate commutation is used only as a descriptive diagnostic. The impossibility result in proposition 4.2 shows that even stronger approximate diagonality at positive noise cannot, without additional assumptions, provide a distribution-free quantitative certificate of clean independence.

C Proofs for Approximate Sharing

This section gives the complete proofs of theorem 4.1 and proposition 4.2. We first establish the exact relation between total correlation and coordinatewise denoising regret, and then prove the absence of a distribution-free approximate converse.

C.1 Proof of theorem 4.1

Let $W = V^\top Z$ and let P denote the law of W. Define $Q := \prod_{i=1}^k P_{W_i}$, the product of the one-dimensional marginals of P.

Because V is orthogonal and the Gaussian noise is isotropic, we may work entirely in the W coordinates. For $t > 0$, define the Gaussian convolutions $P_t := P * \mathcal{N}(0, t^2 I_k), \qquad Q_t := Q * \mathcal{N}(0, t^2 I_k)$.

Since both Q and the Gaussian channel factorize coordinatewise, $Q_t = \prod_{i=1}^k \left(P_{W_i} * \mathcal{N}(0, t^2)\right)$. But the ith factor on the right-hand side is precisely the ith marginal of P_t. Hence

$$D_{\mathrm{KL}}(P_t \| Q_t) = \mathrm{TC}_V(t). \tag{21}$$

Let $f_{P,t}$ and $f_{Q,t}$ denote the posterior means under the priors P and Q, respectively:

$$f_{P,t}(y) := \mathbb{E}_P[W \mid W + t\varepsilon = y], \tag{22}$$
$$f_{Q,t}(y) := \mathbb{E}_Q[W \mid W + t\varepsilon = y]. \tag{23}$$

Because Q is a product distribution, $f_{Q,t}$ acts coordinatewise. After rotation back by V, it is exactly the constrained denoiser $g_{V,t}$ in equation 4.

Under the true law P, conditional expectation is the L^2 projection of W onto the sigma-algebra generated by $W + t\varepsilon$. Therefore, for any measurable estimator of the noisy observation, the Pythagorean identity applies. In particular,

$$\begin{aligned} &\mathbb{E}_P \left\| W - f_{Q,t}(W + t\varepsilon) \right\|^2 - \mathbb{E}_P \left\| W - f_{P,t}(W + t\varepsilon) \right\|^2 \\ &\qquad = \mathbb{E}_{Y \sim P_t} \left\| f_{P,t}(Y) - f_{Q,t}(Y) \right\|^2 . \end{aligned} \tag{24}$$

By orthogonal invariance, the left-hand side is precisely $\mathcal{R}_V(t)$. Thus

$$\mathcal{R}_V(t) = \mathbb{E}_{Y \sim P_t} \left\| f_{P,t}(Y) - f_{Q,t}(Y) \right\|^2 . \tag{25}$$

For $t > 0$, Gaussian smoothing makes both P_t and Q_t have strictly positive smooth densities, which we denote by p_t and q_t. The relative de Bruijn identity gives

$$\frac{d}{d(t^2)} D_{\mathrm{KL}}(P_t \| Q_t) = -\frac{1}{2} \mathbb{E}_{Y \sim P_t} \left\| \nabla \log p_t(Y) - \nabla \log q_t(Y) \right\|^2 . \tag{26}$$

Tweedie's formula applied to the two priors gives

$$f_{P,t}(y) = y + t^2 \nabla \log p_t(y), \tag{27}$$
$$f_{Q,t}(y) = y + t^2 \nabla \log q_t(y). \tag{28}$$

Subtracting,

$$f_{P,t}(y) - f_{Q,t}(y) = t^2 \left(\nabla \log p_t(y) - \nabla \log q_t(y) \right). \tag{29}$$

Using $\frac{d(t^2)}{dt} = 2t$, together with equation 26, equation 29, and equation 25, yields

$$\begin{aligned} \frac{d}{dt} D_{\mathrm{KL}}(P_t \| Q_t) &= -t\, \mathbb{E}_{P_t} \left\| \nabla \log p_t - \nabla \log q_t \right\|^2 \\ &= -\frac{1}{t^3} \mathbb{E}_{P_t} \left\| f_{P,t} - f_{Q,t} \right\|^2 \\ &= -\frac{\mathcal{R}_V(t)}{t^3}. \end{aligned} \tag{30}$$

Combining equation 21 and equation 30, $\frac{d}{dt}\mathrm{TC}_V(t) = -\frac{\mathcal{R}_V(t)}{t^3}$. Integrating from t_0 to T gives $\mathrm{TC}_V(t_0) - \mathrm{TC}_V(T) = \int_{t_0}^{T} \frac{\mathcal{R}_V(t)}{t^3}\, dt$, which proves equation 6.

We now justify explicit sufficient endpoint conditions for equation 7. Suppose $\mathbb{E}\|W\|^2 < \infty$ and $\mathrm{TC}(W) = D_{\mathrm{KL}}(P\|Q) < \infty$. Since the same Gaussian channel sends P and Q to P_t and Q_t, the data-processing inequality gives $\mathrm{TC}_V(t) = D_{\mathrm{KL}}(P_t\|Q_t) \le D_{\mathrm{KL}}(P\|Q)$. As $t \downarrow 0$, $P_t \Rightarrow P$ and $Q_t \Rightarrow Q$. Joint lower semicontinuity of relative entropy under weak convergence, combined with this upper bound, yields $\lim_{t\downarrow 0} \mathrm{TC}_V(t) = \mathrm{TC}(W)$.

For the large-noise endpoint let $\mu = \mathbb{E}W$. Because Q_t is the product of the marginals of P_t, the relative-entropy chain rule and convexity in the first argument give

$$\begin{aligned} 0 \le \mathrm{TC}_V(t) &\le D_{\mathrm{KL}}\big(P_t \,\big\|\, \mathcal{N}(\mu, t^2 I_k)\big) \\ &\le \mathbb{E}\left[D_{\mathrm{KL}}\big(\mathcal{N}(W, t^2 I_k) \,\big\|\, \mathcal{N}(\mu, t^2 I_k)\big)\right] \\ &= \frac{\mathbb{E}\|W-\mu\|^2}{2t^2} \longrightarrow 0. \end{aligned} \tag{31}$$

Taking $t_0 \downarrow 0$ and $T \uparrow \infty$ in equation 6, using nonnegativity of regret, proves equation 7. This is also the specialization of the Gaussian mismatched-estimation identity to $Q = \prod_i P_{W_i}$. □

Noise-schedule consequence. Let π be a density supported on $[t_0, T]$. Then

$$\begin{aligned} \mathbb{E}_{t\sim\pi}\mathcal{R}_V(t) &= \int_{t_0}^{T} t^3 \pi(t)\, \frac{\mathcal{R}_V(t)}{t^3}\, dt \\ &\le \left(\sup_{t\in[t_0,T]} t^3\pi(t)\right) \mathrm{TC}(V^\top Z). \end{aligned} \tag{32}$$

Representation-learning consequence. Whenever equation 7 is finite for the candidate frames, it holds pointwise in V, and hence

$$\arg\min_{V\in O(k)} \mathrm{TC}(V^\top Z) = \arg\min_{V\in O(k)} \int_0^\infty \frac{\mathcal{R}_V(t)}{t^3}\, dt. \tag{33}$$

C.2 Generative consequence

For completeness, we justify the distribution-level statement equation 8.

Let $P_W := \mathcal{L}(V^\top Z), \qquad Q_W := \prod_{i=1}^k P_{W_i}$, and let $Q_V := V_\# Q_W$ be the product-of-marginals law rotated back to the original coordinates.

By construction, the Bayes posterior mean of Q_W under isotropic Gaussian corruption acts coordinatewise. Rotating back by V, its Bayes denoiser is exactly $g_{V,t}$. Hence the full path $\{g_{V,t}\}_{t>0}$ is the exact Bayes denoising path associated with Q_V. Under the ideal reverse-diffusion construction, this path therefore generates Q_V.

Moreover, relative entropy is invariant under invertible orthogonal changes of coordinates, so

$$\begin{aligned} D_{\mathrm{KL}}(P_Z\|Q_V) &= D_{\mathrm{KL}}\left(P_{V^\top Z} \,\|\, Q_W\right) \\ &= D_{\mathrm{KL}}\left(P_{V^\top Z} \,\middle\|\, \prod_i P_{(V^\top Z)_i}\right) \\ &= \mathrm{TC}(V^\top Z). \end{aligned} \tag{34}$$

This proves equation 8.

C.3 Connection to likelihood-weighted diffusion KL bounds

Theorem 4.1 has a direct connection to likelihood-weighted score matching, but its main purpose is to quantify *representation-induced* denoising regret. We spell out the relation to prior work to distinguish a general generative bound from the exact identity obtained here. The Gaussian mismatched-estimation identity (Verdú, 2010) provides the information–estimation foundation. In diffusion terminology, Theorem 1 of Song et al. (2021) bounds the generative KL of an arbitrary learned score

using a likelihood-weighted score error and a terminal mismatch; their Theorem 2 establishes equality when the learned score is the exact noisy-score path of a single alternative distribution and the reverse process has the matching terminal law. Equation (5) of Kadkhodaie et al. (2024) expresses the corresponding general bound in terms of excess denoising MSE. Thus an exact KL–denoising equality is not, by itself, the novelty claimed here. What our choice of alternative law contributes is the *total-correlation interpretation* of both sides and its use in the subsequent analysis of fixed versus noise-dependent frames.

Matching the diffusion weighting. Retain the notation from Appendix C.1: $P = \mathcal{L}(W)$ and $Q = \prod_i P_{W_i}$, with $P_t = P * \mathcal{N}(0, t^2 I_k)$ and $Q_t = Q * \mathcal{N}(0, t^2 I_k)$. For the variance-exploding Gaussian channel $Y_t = W + t\varepsilon$, the forward diffusion coefficient satisfies $g(t)^2 = 2t$ when t parameterizes the noise standard deviation. Hence the likelihood-weighted score discrepancy is

$$\frac{1}{2}\int_{t_0}^{T} g(t)^2\, \mathbb{E}_{Y\sim P_t}\|\nabla \log p_t(Y) - \nabla \log q_t(Y)\|^2\, dt = \int_{t_0}^{T} t\, \mathbb{E}_{Y\sim P_t}\|\nabla \log p_t(Y) - \nabla \log q_t(Y)\|^2\, dt \tag{35}$$

$$= \int_{t_0}^{T} \frac{\mathbb{E}_{Y\sim P_t}\|f_{P,t}(Y) - f_{Q,t}(Y)\|^2}{t^3}\, dt, \tag{36}$$

where the last step is Tweedie's identity. The t^{-3} factor therefore matches the likelihood weighting after converting scores to posterior means; it is not an independently chosen weighting. Here $t_0 > 0$ avoids a potentially degenerate zero-noise coefficient, and endpoint limits are taken only under the assumptions of Theorem 4.1.

Why equality holds here. For any two suitable laws whose Gaussian convolutions admit the relative de Bruijn identity, the derivation in Appendix C.1 gives the *two-law* identity

$$D_{\mathrm{KL}}(P_{t_0}\|Q_{t_0}) = \int_{t_0}^{T} \frac{\mathbb{E}_{Y\sim P_t}\|f_{P,t}(Y) - f_{Q,t}(Y)\|^2}{t^3}\, dt + D_{\mathrm{KL}}(P_T\|Q_T), \qquad 0 < t_0 < T. \tag{37}$$

The terminal KL term is present even for *exact* score paths at finite T; in particular, it is not appropriate to silently identify P_T and Q_T. If $t_0 \downarrow 0$ and $T \uparrow \infty$ under the endpoint conditions proved in Appendix C.1, this becomes

$$D_{\mathrm{KL}}(P\|Q) = \int_{0}^{\infty} \frac{\mathbb{E}_{Y\sim P_t}\|f_{P,t}(Y) - f_{Q,t}(Y)\|^2}{t^3}\, dt. \tag{38}$$

The infinite-noise endpoint is an analytical limit, not a finite noise prior used in an implementation. Equation 38 concerns the Bayes posterior means of two fixed laws; it does not assert equality for an arbitrary approximate neural denoiser.

In our case $Q = \prod_i P_{W_i}$ has two additional properties. First, Gaussian convolution preserves its product structure, so Q_t is the product of the marginals of P_t and

$$D_{\mathrm{KL}}(P_t\|Q_t) = \mathrm{TC}_V(t). \tag{39}$$

Second, its Bayes posterior mean is coordinatewise. The Pythagorean identity in equation 25 therefore identifies the numerator of equation 37 as precisely $\mathcal{R}_V(t)$, the excess Bayes risk of discarding cross-coordinate information. Combining the two properties recovers the finite-interval and all-noise statements of Theorem 4.1, without replacing its existing proof.

Finally, the *generative* identification uses the exact Bayes path of Q, not merely a small score error. An ideal reverse process initialized from Q_T and driven by its exact score produces Q (in W coordinates), or $Q_V = V_{\#}Q$ in the original coordinates. Orthogonal invariance of KL then gives $D_{\mathrm{KL}}(P_Z\|Q_V) = \mathrm{TC}(V^\top Z)$, as established in Appendix C.2. A general learned score initialized from some other finite-noise prior has instead a KL *bound* with its terminal mismatch, not this automatic identification. Moreover, the equality concerns the factorization cost for a *specified* frame; the additional cost of enforcing the *same* frame across noise levels is separately defined by Δ_{tie} in equation 9.

C.4 PROOF OF PROPOSITION 4.2

Recall the family $p_m(x_1, x_2) = \varphi(x_1)\varphi(x_2)\left[1 + \epsilon \sin(mx_1)\sin(mx_2)\right], \qquad 0 < \epsilon < 1$, where φ is the standard Gaussian density.

Define $h_m(x_1, x_2) := \sin(mx_1)\sin(mx_2), \qquad q(x_1, x_2) := \varphi(x_1)\varphi(x_2)$. Then $p_m = q(1+\epsilon h_m)$. Since $|h_m| \le 1$ and $\epsilon < 1$, p_m is everywhere positive. Furthermore, $\int_{\mathbb{R}} \varphi(x)\sin(mx)\,dx = 0$. It follows immediately that p_m is normalized and that both of its one-dimensional marginals are exactly standard Gaussian. Therefore the total correlation of P_m is simply $\mathrm{TC}(P_m) = D_{\mathrm{KL}}(P_m \| Q)$, where Q is the independent standard Gaussian law with density q.

The clean dependence stays bounded away from zero. Set $F(u) := (1+u)\log(1+u) - u$. Since $F(0) = F'(0) = 0, \qquad F''(u) = \frac{1}{1+u}$, for $|u| \le \epsilon$ we have $F''(u) \ge \frac{1}{1+\epsilon}$. Hence $F(u) \ge \frac{u^2}{2(1+\epsilon)}$.

Because $\mathbb{E}_Q h_m = 0$,

$$\begin{aligned} \mathrm{TC}(P_m) &= \mathbb{E}_Q\left[(1+\epsilon h_m)\log(1+\epsilon h_m)\right] \\ &= \mathbb{E}_Q F(\epsilon h_m) \\ &\ge \frac{\epsilon^2}{2(1+\epsilon)}\mathbb{E}_Q h_m^2. \end{aligned} \tag{40}$$

For $G \sim \mathcal{N}(0,1)$, $\mathbb{E}\sin^2(mG) = \frac{1-e^{-2m^2}}{2}$. Independence under Q therefore gives

$$\mathbb{E}_Q h_m^2 = \frac{(1-e^{-2m^2})^2}{4} \ge \frac{(1-e^{-2})^2}{4}. \tag{41}$$

Combining equation 40 and equation 41, $\mathrm{TC}(P_m) \ge \frac{\epsilon^2(1-e^{-2})^2}{8(1+\epsilon)} > 0 \qquad \forall m \ge 1$. This proves the uniform lower bound in proposition 4.2.

Gaussian smoothing exponentially suppresses the dependence. Fix $t > 0$ and write $q_t := 1 + t^2$. Under the one-dimensional Gaussian reference model $X \sim \mathcal{N}(0,1), \qquad Y = X + tG$, the conditional distribution is $X \mid Y = y \sim \mathcal{N}\left(\frac{y}{q_t}, \frac{t^2}{q_t}\right)$. Consequently,

$$\mathbb{E}[\sin(mX) \mid Y = y] = \exp\left(-\frac{m^2t^2}{2q_t}\right)\sin\left(\frac{my}{q_t}\right). \tag{42}$$

Applying equation 42 independently in the two coordinates gives the exact smoothed density

$$\begin{aligned} p_{m,t}(y_1, y_2) &= \varphi_{q_t}(y_1)\varphi_{q_t}(y_2)r_{m,t}(y), \\ r_{m,t}(y) &:= 1 + \delta_{m,t}\sin(\lambda_{m,t}y_1)\sin(\lambda_{m,t}y_2), \end{aligned} \tag{43}$$

where φ_{q_t} denotes the $\mathcal{N}(0, q_t)$ density and $\delta_{m,t} := \epsilon\exp\left(-\frac{m^2t^2}{q_t}\right), \qquad \lambda_{m,t} := \frac{m}{q_t}$.

In particular, Gaussian smoothing leaves both marginals exactly $\mathcal{N}(0, q_t)$ while attenuating the dependence amplitude by the factor $\exp\left(-\frac{m^2t^2}{1+t^2}\right)$. The same convexity argument used above also yields the quantitative bound

$$\mathrm{TC}(P_{m,t}) \le \frac{\delta_{m,t}^2}{8(1-\epsilon)} \le \frac{\epsilon^2}{8(1-\epsilon)}\exp\left(-\frac{2m^2t^2}{1+t^2}\right). \tag{44}$$

Thus the dependence visible after any fixed positive amount of Gaussian smoothing decays exponentially in m^2.

Uniform convergence of the posterior mean. From equation 43, $\nabla \log p_{m,t}(y) = -\frac{y}{q_t} + \nabla \log r_{m,t}(y)$. Tweedie's formula therefore gives

$$f^\star_{m,t}(y) = \frac{y}{q_t} + t^2\nabla \log r_{m,t}(y). \tag{45}$$

Because $r_{m,t}(y) \ge 1 - \epsilon$ and $\|\nabla r_{m,t}(y)\| \le \sqrt{2}\,\delta_{m,t}\lambda_{m,t}$, we obtain

$$\sup_y \left\| f^\star_{m,t}(y) - \frac{y}{q_t} \right\| \le \frac{\sqrt{2}\,t^2\delta_{m,t}\lambda_{m,t}}{1-\epsilon}. \tag{46}$$

Uniform convergence of the Jacobian. Using $\nabla^2 \log r = \frac{\nabla^2 r}{r} - \frac{\nabla r \nabla r^\top}{r^2}$, we bound the two terms separately.

Every entry of $\nabla^2 r_{m,t}$ has absolute value at most $\delta_{m,t}\lambda_{m,t}^2$, hence $\|\nabla^2 r_{m,t}(y)\|_{\mathrm{op}} \le 2\delta_{m,t}\lambda_{m,t}^2$. Also, $\|\nabla r_{m,t}(y)\|^2 \le 2\delta_{m,t}^2\lambda_{m,t}^2$. Since $\delta_{m,t} \le \epsilon$ and $r_{m,t} \ge 1-\epsilon$, $\sup_y \|\nabla^2 \log r_{m,t}(y)\|_{\mathrm{op}} \le \frac{2\delta_{m,t}\lambda_{m,t}^2}{(1-\epsilon)^2}$.

Differentiating equation 45, $Df^\star_{m,t}(y) = \frac{1}{q_t} I_2 + t^2 \nabla^2 \log r_{m,t}(y)$, and therefore

$$\sup_y \left\| Df^\star_{m,t}(y) - \frac{1}{q_t} I_2 \right\|_{\mathrm{op}} \le \frac{2t^2 \delta_{m,t}\lambda_{m,t}^2}{(1-\epsilon)^2}. \tag{47}$$

Now fix any compact positive-noise interval $t \in [\tau, T], \qquad 0 < \tau < T < \infty$. Then $\frac{t^2}{1+t^2} \ge \frac{\tau^2}{1+\tau^2} =: c_\tau > 0$. Consequently, $\delta_{m,t} \le \epsilon e^{-c_\tau m^2}$ uniformly over $t \in [\tau, T]$.

The right-hand side of equation 46 is therefore bounded uniformly by a constant times $me^{-c_\tau m^2}$, while the right-hand side of equation 47 is bounded by a constant times $m^2 e^{-c_\tau m^2}$. Both converge to zero as $m \to \infty$. Hence

$$\sup_{t\in[\tau,T]} \sup_y \left\| f^\star_{m,t}(y) - \frac{y}{1+t^2} \right\| \longrightarrow 0, \tag{48}$$

$$\sup_{t\in[\tau,T]} \sup_y \left\| Df^\star_{m,t}(y) - \frac{1}{1+t^2} I_2 \right\|_{\mathrm{op}} \longrightarrow 0. \tag{49}$$

This proves the two uniform convergences in proposition 4.2. □

C.5 No one-noise stability modulus

Corollary C.1 (No one-noise stability modulus)**.** *Fix $t_0 > 0$ and let $\mathcal{A}_I := \{D \in \mathbb{R}^{2\times 2} : D \text{ is diagonal}\}$. There is no distribution-free function $\omega\colon [0,\infty) \to [0,\infty)$ with $\omega(r) \to 0$ as $r \downarrow 0$ such that every smooth positive law P on $\mathbb{R}^2$ satisfies*

$$\mathrm{TC}(P) \le \omega\left(\sup_y \mathrm{dist}_{\mathrm{op}}\big(Df^\star_{t_0}(y), \mathcal{A}_I\big) \right), \tag{50}$$

where $\mathrm{dist}_{\mathrm{op}}(A, \mathcal{A}_I) := \inf_{D\in\mathcal{A}_I} \|A - D\|_{\mathrm{op}}$.

Proof. Apply proposition 4.2 to P_m. The Jacobian convergence in that proposition makes the displayed distance tend to zero, while the clean total correlation stays uniformly bounded away from zero, contradicting any modulus $\omega(r) \to 0$. □

C.6 The same construction also hides noncommutation

The counterexample is stronger than pointwise near-diagonality in the displayed frame. It also makes pairwise Bayes-Jacobian commutators vanish over every compact positive-noise interval.

Write $J_{m,t}(y) = Df^\star_{m,t}(y) = \frac{1}{q_t} I_2 + E_{m,t}(y)$, where, by equation 47, $\sup_{t\in[\tau,T]} \sup_y \|E_{m,t}(y)\|_{\mathrm{op}} \longrightarrow 0$.

Since scalar multiples of the identity commute with every matrix, $[J_{m,t}(y), J_{m,s}(y')] = [E_{m,t}(y), E_{m,s}(y')]$. For any matrices E, F, $\|[E,F]\|_{\mathrm{op}} \le 2\|E\|_{\mathrm{op}}\|F\|_{\mathrm{op}}$. Therefore

$$\sup_{\substack{t,s\in[\tau,T]\\ y,y'\in\mathbb{R}^2}} \|[J_{m,t}(y), J_{m,s}(y')]\|_{\mathrm{op}} \longrightarrow 0. \tag{51}$$

Thus vanishing cross-input and cross-noise commutators can coexist with clean total correlation bounded away from zero in the displayed coordinates. This construction alone does not establish a lower bound on $\inf_{V\in O(2)} \mathrm{TC}(V^\top X_m)$; it does not resolve frame-optimized quantitative stability.

D Additional Derivations and Identifiability for the Rademacher–Gaussian Model

This section collects additional calculations for the solvable model introduced in section 5. We derive the scalar posterior mean, record the score and Jacobian of the full denoising path, describe several limiting regimes, and prove that a single positive-noise denoiser identifies the model parameters up to signed permutation.

Recall the model

$$X = U\Psi, \qquad \Psi = aS + \xi, \tag{52}$$

where $U \in \mathrm{St}(n,k)$, $\quad S_i \overset{\text{iid}}{\sim} \mathrm{Unif}\{-1,+1\}$, $\quad \xi \sim \mathcal{N}(0, \beta I_k)$, with $a, \beta > 0$.

For convenience, throughout this section write $q_t := \beta + t^2$, $\quad r_t := q_t^{-1}$.

D.1 Scalar posterior mean

Consider one latent coordinate $V = aS + \xi$, observed through $Z = V + t\varepsilon$, where $S \sim \mathrm{Unif}\{-1,+1\}$, $\quad \xi \sim \mathcal{N}(0,\beta)$, $\quad \varepsilon \sim \mathcal{N}(0,1)$, independently.

Conditioned on $S = s$, $Z \mid S = s \sim \mathcal{N}(as, q_t)$. Equal sign priors give the log-posterior odds

$$\begin{aligned} \log \frac{\mathbb{P}(S=1 \mid Z=z)}{\mathbb{P}(S=-1 \mid Z=z)} &= -\frac{(z-a)^2}{2q_t} + \frac{(z+a)^2}{2q_t} \\ &= \frac{2az}{q_t}. \end{aligned} \tag{53}$$

Hence $\mathbb{E}[S \mid Z=z] = \tanh\left(\frac{az}{q_t}\right)$.

Conditional on both $Z = z$ and $S = s$, Gaussian regression gives $\mathbb{E}[\xi \mid Z=z, S=s] = \frac{\beta}{q_t}(z - as)$. Therefore

$$\begin{aligned} m_t(z) &:= \mathbb{E}[V \mid Z=z] \\ &= a\,\mathbb{E}[S \mid Z=z] + \mathbb{E}[\xi \mid Z=z] \\ &= \frac{\beta}{q_t} z + \frac{at^2}{q_t} \tanh\left(\frac{az}{q_t}\right). \end{aligned} \tag{54}$$

This scalar map underlies the closed-form denoising path used in the finite-sample analysis.

D.2 Full Bayes denoiser, score, and Jacobian

Because the latent coordinates are independent and identically distributed, applying equation 54 coordinatewise gives

$$f_{U,a,\beta}(y,t) = U\left[\beta r_t U^\top y + a(1-\beta r_t)\tanh\big(a r_t U^\top y\big)\right], \tag{55}$$

where $\tanh$ acts coordinatewise.

Since $1 - \beta r_t = \frac{t^2}{q_t}$, this is equivalent to $f_{U,a,\beta}(y,t) = U\left[\frac{\beta}{q_t}U^\top y + \frac{at^2}{q_t}\tanh\left(\frac{aU^\top y}{q_t}\right)\right]$.

Thus the full continuum of Bayes denoisers has one common linear encoder $U^\top$, one common decoder U, and the same coordinatewise nonlinearity. The noise scale enters only through the scalar coefficients.

Potential representation. Define $\Phi_t(y) := \frac{\beta}{2q_t}\|U^\top y\|^2 + t^2 \sum_{i=1}^k \log\cosh\left(\frac{au_i^\top y}{q_t}\right)$. Direct differentiation gives $\nabla\Phi_t(y) = f_{U,a,\beta}(y,t)$. The denoiser is therefore a gradient map, consistently with the generic Bayes identity.

Let $P_U := UU^\top$. Using Tweedie's formula, $\nabla \log p_t(y) = \frac{f_{U,a,\beta}(y,t)-y}{t^2}$, we obtain

$$\nabla \log p_t(y) = -\frac{(I - P_U)y}{t^2} + \frac{1}{q_t} U \left[-U^\top y + a \tanh\left(\frac{aU^\top y}{q_t} \right) \right]. \tag{56}$$

Differentiating the scalar posterior mean gives $g_t(z) := m_t'(z) = \frac{\beta}{q_t} + \frac{a^2 t^2}{q_t^2} \operatorname{sech}^2\left(\frac{az}{q_t}\right) > 0$. Consequently, $Df_{U,a,\beta}(y,t) = U \operatorname{diag}\left(g_t(u_1^\top y), \ldots, g_t(u_k^\top y)\right) U^\top$.

Every Jacobian therefore has range contained in $\operatorname{range}(U)$ and is diagonal in the same frame for every input and every noise level. In particular, $[Df_{U,a,\beta}(y,t), Df_{U,a,\beta}(y',s)] = 0 \qquad \forall y, y', t, s$. The model is thus an exact equality case of the commutation certificate in proposition B.2.

D.3 Limiting regimes

The scalar map equation 54 has several useful limits.

Vanishing observation noise. As $t \downarrow 0$, $m_t(z) \longrightarrow z$. Thus the Bayes denoiser approaches the identity on the active subspace.

Vanishing mixture amplitude. As $a \downarrow 0$, $m_t(z) \longrightarrow \frac{\beta}{\beta+t^2} z$, the posterior mean for a centered Gaussian prior of variance β.

Vanishing within-component variance. For fixed $t > 0$, as $\beta \downarrow 0$, $m_t(z) \longrightarrow a \tanh\left(\frac{az}{t^2}\right)$. This is the pure Rademacher sign-detection regime.

Large-noise limit. As $t \to \infty$, expanding $\tanh$ around zero gives $m_t(z) = \frac{a^2+\beta}{t^2} z + O\left(t^{-4}|z| + t^{-6}|z|^3\right)$. Thus the denoiser becomes small and asymptotically linear at high noise.

D.4 Identifiability from one positive-noise denoiser

The main text uses the Rademacher–Gaussian model primarily to study finite-sample sharing. The model also admits a stronger exact identifiability result: one positive-noise Bayes denoiser determines its latent axes and scalar parameters.

Some ambiguity is unavoidable. In the Gaussian limit $a = 0$, equation 55 becomes $f_{U,0,\beta}(y,t) = \frac{\beta}{\beta+t^2} P_U y$, which depends on U only through the projector P_U. Hence U and UR, $R \in O(k)$, cannot be distinguished. For $a > 0$, the non-Gaussian component breaks this continuous rotational symmetry down to signed permutations.

Theorem D.1 (Identifiability from one positive-noise denoiser)**.** *Fix $t_0 > 0$. Within the model equation 52 with $a, \beta > 0$, the function $y \longmapsto f_{U,a,\beta}(y, t_0)$ determines*

1. *the latent dimension k;*
2. *the unordered signed set of axes $\{\pm u_1, \ldots, \pm u_k\}$;*
3. *the scalar parameters a and β.*

The only remaining ambiguity in U is signed permutation of its columns.

Proof. Set $q := \beta + t_0^2$. The scalar posterior mean at t_0 is $m(z) = \frac{\beta}{q} z + \frac{a t_0^2}{q} \tanh\left(\frac{az}{q}\right)$. Differentiating at the origin gives

$$c_1 := m'(0) = \frac{\beta}{q} + \frac{a^2 t_0^2}{q^2} > 0 \tag{57}$$

and

$$c_3 := m'''(0) = -\frac{2a^4 t_0^2}{q^4} < 0. \tag{58}$$

Recover the active subspace and k. At the origin, $Df_{U,a,\beta}(0,t_0) = c_1 P_U$. Since $c_1 > 0$, $\mathrm{range}\big(Df_{U,a,\beta}(0,t_0)\big) = \mathrm{range}(U)$. Thus $k = \mathrm{rank}\big(Df_{U,a,\beta}(0,t_0)\big)$, and the active subspace itself is identified.

The Jacobian alone does not identify the individual axes, because $c_1 P_U$ is rotationally invariant within the active subspace.

Recover the individual axes. We therefore inspect the third derivative at the same point. For any $v \in \mathbb{R}^n$, $D^3 f_{U,a,\beta}(0,t_0)[v,v,v] = c_3 \sum_{j=1}^k u_j \langle u_j, v\rangle^3$. Define the observable scalar quartic form $H(v) := \big\langle v, D^3 f_{U,a,\beta}(0,t_0)[v,v,v]\big\rangle$. Then $H(v) = c_3 \sum_{j=1}^k \langle u_j, v\rangle^4$.

Restrict to $\|v\| = 1$ and write $\alpha_j := \langle u_j, v\rangle$. Since the u_j are orthonormal, $\sum_{j=1}^k \alpha_j^2 \le 1$. Moreover,

$$\sum_{j=1}^k \alpha_j^4 \le \left(\sum_{j=1}^k \alpha_j^2\right)^2 \le 1. \tag{59}$$

Equality throughout holds if and only if $v \in \mathrm{range}(U)$ and exactly one α_j is nonzero, which is equivalent to $v = \pm u_j$ for some j.

Because $c_3 < 0$, minimizing $H(v)$ over the unit sphere is equivalent to maximizing the quartic sum in equation 59. Hence $\arg\min_{\|v\|=1} H(v) = \{\pm u_1, \dots, \pm u_k\}$. The unordered signed set of latent axes is therefore determined by the denoiser. Evaluating H at any recovered axis also identifies c_3.

Recover a and β. From equation 58, define $h := \left(-\frac{c_3}{2t_0^2}\right)^{1/4} = \frac{a}{q}$. Since $\beta = q - t_0^2$, substituting into equation 57 gives

$$\begin{aligned} c_1 &= \frac{q - t_0^2}{q} + \frac{a^2 t_0^2}{q^2} \\ &= 1 - \frac{t_0^2}{q} + h^2 t_0^2. \end{aligned} \tag{60}$$

Therefore $\frac{1}{q} = h^2 + \frac{1-c_1}{t_0^2}$. Thus q is identified, after which $a = hq, \qquad \beta = q - t_0^2$. This recovers all model parameters. □

Interpretation. The argument has a simple structure. The Jacobian at one point identifies the active subspace but is rotationally invariant within that subspace. A higher-order derivative resolves the remaining rotation by exposing the non-Gaussian coordinate axes. Thus, in this solvable model, the intrinsic frame guaranteed abstractly by the commutation certificate can be extracted explicitly from the Bayes-denoiser function at a single positive noise level.

E Finite-Sample Learning of the Rademacher–Gaussian Frame

This section proves the finite-sample result stated in theorem 5.1. We first give a sharper finite-net version, then derive the simplified rate used in the main text and the corresponding comparison with L untied frames.

Recall the model $X = U_\star(aS + \xi), \qquad U_\star \in \mathrm{St}(n,k)$, with $S_i \overset{\text{iid}}{\sim} \mathrm{Unif}\{-1,+1\}, \qquad \xi \sim \mathcal{N}(0, \beta I_k)$, and observations $Y = X + \mathsf{T}G, \qquad G \sim \mathcal{N}(0, I_n), \qquad \mathsf{T} \sim \nu$, where T is independent of (X, G) and ν is an arbitrary distribution on $(0,\infty)$.

For a candidate frame $U \in \mathrm{St}(n,k)$, let $f_U = f_{U,a,\beta}$ denote the closed-form denoising path from equation 55, and define $\ell_U(X,Y,\mathsf{T}) := \|X - f_U(Y,\mathsf{T})\|^2, \qquad \mathcal{L}_\nu(U) := \mathbb{E}\,\ell_U$. The true frame is well specified:

$$f_{U_\star}(Y,\mathsf{T}) = \mathbb{E}[X \mid Y, \mathsf{T}]. \tag{61}$$

Let $d_{n,k} := nk - \frac{k(k+1)}{2}$ be the dimension of $\mathrm{St}(n,k)$. For $0 < \eta \le 1$ and $(n,k) \neq (1,1)$, let $\mathcal{U}_\eta$ be an operator-norm η-net satisfying

$$\log |\mathcal{U}_\eta| \le d_{n,k} \log\left(\frac{C_0}{\eta}\right) \tag{62}$$

for a universal constant C_0.

When $n = k = 1$, set $\mathcal{U}_\eta = \{-1, +1\}$. In this case, $\mathrm{St}(1,1) = \{-1, +1\}$, and both frames induce the same denoiser. Indeed, for $U \in \{-1, +1\}$, the oddness of $\tanh$ and $U^2 = 1$ give, from equation 55,

$$f_U(y,t) = \frac{\beta}{\beta + t^2} y + \frac{at^2}{\beta + t^2} \tanh\left(\frac{ay}{\beta + t^2}\right).$$

Hence $f_U = f_{U_\star}$ for every candidate frame, so $\mathcal{L}_\nu(U) - \mathcal{L}_\nu(U_\star) = 0$ for all ν. The exceptional case therefore requires no covering-number bound.

Given N i.i.d. triples $(X_i, Y_i, \mathsf{T}_i)_{i=1}^N$, define $\widehat{\mathcal{L}}_\nu(U) := \frac{1}{N} \sum_{i=1}^N \ell_U(X_i, Y_i, \mathsf{T}_i)$ and let $\widehat{U}_\eta \in \arg\min_{U \in \mathcal{U}_\eta} \widehat{\mathcal{L}}_\nu(U)$. The finite net is used only as a proof device; it is not intended as a practical optimization algorithm.

E.1 A UNIFORM SUB-EXPONENTIAL LOSS BOUND

For notational convenience, write $q_t := \beta + t^2, \qquad \alpha_t := \frac{\beta}{q_t}, \qquad b_t := \frac{at^2}{q_t}, \qquad P_U := UU^\top$. Then

$$f_U(Y,t) = \alpha_t P_U Y + b_t U \tanh\left(\frac{aU^\top Y}{q_t}\right). \tag{63}$$

Lemma E.1 (Uniform sub-exponential loss)**.** *There is a universal constant $c_0 > 0$ such that, for every $U \in \mathrm{St}(n,k)$ and every distribution ν of T,*

$$\|\ell_U(X, Y, \mathsf{T})\|_{\psi_1} \le B_{a,\beta,k} := c_0 k (a^2 + \beta). \tag{64}$$

The bound is uniform in the ambient dimension n and in the noise scale.

Proof. From equation 63, $f_U(Y,t) = \alpha_t P_U Y + b_t U \tanh\left(\frac{aU^\top Y}{q_t}\right)$. We use $0 \le \alpha_t \le 1, \qquad |b_t| \le a, \qquad \|\tanh(\cdot)\| \le \sqrt{k}$, together with $\alpha_t t = \frac{\beta t}{\beta + t^2} \le \frac{\sqrt{\beta}}{2}$.

Since $X = U_\star(aS + \xi), \qquad Y = X + tG$, the triangle inequality yields $\|X - f_U(Y,t)\| \le 3a\sqrt{k} + 2\|\xi\| + \frac{\sqrt{\beta}}{2}\|P_U G\|$. Using $(x+y+z)^2 \le 3x^2 + 3y^2 + 3z^2$, we obtain $\ell_U \le 27a^2 k + 12\|\xi\|^2 + \frac{3\beta}{4}\|P_U G\|^2$.

Now $\frac{\|\xi\|^2}{\beta} \sim \chi^2_k$, and, for every fixed U, $\|P_U G\|^2 \sim \chi^2_k$. The sub-exponential Orlicz norm of a χ^2_k variable is $O(k)$. Hence $\|\ell_U\|_{\psi_1} \le c_0 k(a^2 + \beta)$ for a universal c_0, uniformly over U, t, and ν. No boundedness or truncation event is required. □

E.2 APPROXIMATION ERROR OF A NEARBY FRAME

Lemma E.2 (Quadratic approximation of the Bayes frame)**.** *For every $U \in \mathrm{St}(n,k)$,*

$$\mathcal{L}_\nu(U) - \mathcal{L}_\nu(U_\star) \le K_{a,\beta,k} \|U - U_\star\|_{\mathrm{op}}^2, \tag{65}$$

where

$$K_{a,\beta,k} := 4k \left(3a^2 + 2\beta + \frac{a^4(a^2 + \beta)}{\beta^2}\right). \tag{66}$$

The bound is uniform over the distribution ν of the noise level.

Proof. Because $f_{U_\star}$ is the true conditional expectation, equation 61 and the orthogonality principle imply $\mathcal{L}_\nu(U) - \mathcal{L}_\nu(U_\star) = \mathbb{E}\,\|f_U(Y, \mathsf{T}) - f_{U_\star}(Y, \mathsf{T})\|^2$.

Fix a noise level t, set $V := U_\star, \qquad \delta := \|U - V\|_{\mathrm{op}}$. We separate the linear and nonlinear pieces of equation 63.

For the linear term,

$$\mathbb{E}\|\alpha_t(P_U - P_V)Y\|^2 \le 4k(a^2+\beta)\delta^2. \tag{67}$$

Here we use $\|P_U - P_V\|_F \le 2\|U-V\|_F \le 2\sqrt{k}\,\delta$ and $\alpha_t^2(a^2+\beta+t^2) \le a^2+\beta$. For the nonlinear term, the map $\tanh$ is 1-Lipschitz. Using this fact together with the bounds on b_t and the second moments of Y gives

$$\begin{aligned}\mathbb{E}\left\| b_t U \tanh\left(\frac{aU^\top Y}{q_t}\right) - b_t V \tanh\left(\frac{aV^\top Y}{q_t}\right)\right\|^2 \\ \le 2k\left(a^2 + \frac{a^4(a^2+\beta)}{\beta^2}\right)\delta^2.\end{aligned} \tag{68}$$

Finally, using $\|x+y\|^2 \le 2\|x\|^2 + 2\|y\|^2$ to combine equation 67 and equation 68 yields $\mathbb{E}\|f_U(Y,t) - f_V(Y,t)\|^2 \le K_{a,\beta,k}\delta^2$, with $K_{a,\beta,k}$ as in equation 66. The estimate is uniform in t, so averaging over $\mathsf{T} \sim \nu$ proves equation 65. □

E.3 Finite-net ERM bound

We can now state the sharper form of the result used in the main text.

Theorem E.3 (Finite-net shared-frame learning)**.** *There exist universal constants $c_0, c_1, c_2 > 0$ such that, for every $0 < \delta < 1$, with probability at least $1-\delta$,*

$$\begin{aligned}\mathcal{L}_\nu(\widehat{U}_\eta) - \mathcal{L}_\nu(U_\star) \le K_{a,\beta,k}\eta^2 \\ + c_1 B_{a,\beta,k}\sqrt{\frac{\Gamma_\eta(\delta)}{N}} + c_2 B_{a,\beta,k}\frac{\Gamma_\eta(\delta)}{N},\end{aligned} \tag{69}$$

where $\Gamma_\eta(\delta) := d_{n,k}\log\left(\frac{C_0}{\eta}\right) + \log\left(\frac{2}{\delta}\right)$, and $B_{a,\beta,k}$ and $K_{a,\beta,k}$ are defined in equation 64 and equation 66, respectively.

The bound is uniform over the noise-level distribution ν.

Proof. For $n = k = 1$, the excess risk is identically zero by the observation following equation 62, so the bound holds directly. In the remainder, assume $(n,k) \ne (1,1)$. By lemma E.1, for each fixed $U \in \mathcal{U}_\eta$, Bernstein's inequality for sub-exponential random variables gives

$$\left|\widehat{\mathcal{L}}_\nu(U) - \mathcal{L}_\nu(U)\right| \le cB_{a,\beta,k}\left(\sqrt{\frac{u}{N}} + \frac{u}{N}\right) \tag{70}$$

with probability at least $1 - 2e^{-u}$, for a universal constant c.

Choose $u = \log\left(\frac{2|\mathcal{U}_\eta|}{\delta}\right)$. A union bound then gives, with probability at least $1-\delta$,

$$\sup_{U\in\mathcal{U}_\eta}\left|\widehat{\mathcal{L}}_\nu(U) - \mathcal{L}_\nu(U)\right| \le cB_{a,\beta,k}\left(\sqrt{\frac{\Gamma_\eta(\delta)}{N}} + \frac{\Gamma_\eta(\delta)}{N}\right), \tag{71}$$

where we used the entropy estimate equation 62.

Let $\overline{U} \in \mathcal{U}_\eta$ satisfy $\|\overline{U} - U_\star\|_{\mathrm{op}} \le \eta$. By empirical optimality of $\widehat{U}_\eta$, $\widehat{\mathcal{L}}_\nu(\widehat{U}_\eta) \le \widehat{\mathcal{L}}_\nu(\overline{U})$. Therefore

$$\begin{aligned}&\mathcal{L}_\nu(\widehat{U}_\eta) - \mathcal{L}_\nu(U_\star) \\ &\le 2\sup_{U\in\mathcal{U}_\eta}\left|\widehat{\mathcal{L}}_\nu(U) - \mathcal{L}_\nu(U)\right| + \mathcal{L}_\nu(\overline{U}) - \mathcal{L}_\nu(U_\star).\end{aligned} \tag{72}$$

Apply lemma E.2 to the last term: $\mathcal{L}_\nu(\overline{U}) - \mathcal{L}_\nu(U_\star) \le K_{a,\beta,k}\eta^2$. Combining this with equation 71 and absorbing universal numerical constants into c_1, c_2 proves equation 69. □

E.4 Rate used in the main text

Choose $\eta = N^{-1/2}$. Then, up to logarithmic factors and at fixed confidence, equation 69 gives

$$\mathcal{L}_\nu(\widehat{U}_\eta) - \mathcal{L}_\nu(U_\star) = \widetilde{O}\left(B_{a,\beta,k}\sqrt{\frac{d_{n,k}}{N}} + \frac{K_{a,\beta,k} + B_{a,\beta,k}d_{n,k}}{N}\right). \tag{73}$$

This is the form stated in theorem 5.1.

A sufficient total-sample scaling for excess denoising risk at most ε is therefore

$$N_{\mathrm{share}}(\varepsilon) = \widetilde{O}\left(\frac{B^2_{a,\beta,k}d_{n,k}}{\varepsilon^2} + \frac{K_{a,\beta,k} + B_{a,\beta,k}d_{n,k}}{\varepsilon}\right). \tag{74}$$

The key point is that this rate depends on the complexity of *one* Stiefel frame even if the training distribution ν places mass on many different noise levels.

E.5 Comparison with L untied frames

Now fix L noise levels and keep the total budget of training triples equal to N. Suppose instead that the learner may fit unrelated frames $U_1, \dots, U_L \in \mathrm{St}(n,k)$. When $n = k = 1$, all frames induce the same Bayes denoiser, so both shared and untied estimators have zero excess risk. Thus assume $(n,k) \neq (1,1)$ in the following comparison. The parameter space is then the product $\mathrm{St}(n,k)^L$. Taking the product of L operator-norm η-nets gives $\log|\mathcal{U}_\eta^L| \le Ld_{n,k}\log\left(\frac{C_0}{\eta}\right)$. Repeating exactly the same covering-and-concentration argument therefore replaces $d_{n,k}$ by $Ld_{n,k}$. At the level of the leading terms,

$$\text{shared:} \quad \widetilde{O}\left(\sqrt{\frac{d_{n,k}}{N}}\right), \qquad \text{untied:} \quad \widetilde{O}\left(\sqrt{\frac{Ld_{n,k}}{N}}\right). \tag{75}$$

More explicitly, the corresponding sufficient sample scaling is

$$N^{(L)}_{\mathrm{untied}}(\varepsilon) = \widetilde{O}\left(\frac{B^2_{a,\beta,k}Ld_{n,k}}{\varepsilon^2} + \frac{K_{a,\beta,k} + B_{a,\beta,k}Ld_{n,k}}{\varepsilon}\right). \tag{76}$$

This comparison isolates the representation-pooling effect: when a single frame is correctly specified, observations from all noise levels contribute to estimating the same geometric parameter, whereas the untied class replicates that parameter across levels.

The comparison should not be interpreted as a minimax separation. The number L is a property of the chosen discretized untied baseline, not an intrinsic dimension of continuous-time diffusion; equation 74–equation 76 are sufficient upper bounds; and N counts independent noisy training triples rather than distinct clean observations. Their purpose is to isolate the estimation side of the tradeoff discussed in section 4: sharing reduces representation-estimation complexity, while under misspecification it may also incur the nonzero approximation penalty Δ_{tie}.

F Controlled Synthetic Experiments

This section gives the implementation details for the controlled experiments used to test the exact commutation certificate and cross-noise frame transfer. The pooling and bias–complexity experiments are described separately in sections G and H.

F.1 Exact-Bayes commutation under controlled dependence

We first describe the exact-Bayes experiment used to visualize the commutation characterization.

Consider the two-dimensional latent model $X = aS + \xi, \quad \xi \sim \mathcal{N}(0, \beta I_2)$, where ξ is independent of S, with $a = 1, \quad \beta = 0.25$, and sign distribution $\mathbb{P}(S = s) = \frac{1}{4}(1 + \rho s_1 s_2), \quad s \in \{-1,+1\}^2, \quad \rho \in [0,1]$.

At $\rho = 0$, S_1 and S_2 are independent in the standard basis. At $\rho = 1$, $S_1 = S_2$ almost surely, and the model becomes factorial again after the 45° rotation

$$R_{45} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}. \tag{77}$$

Nonfactorization at intermediate dependence. For every $0 < \rho < 1$, the law of X does not factorize in any orthogonal frame.

Proof. Since S is centered with $\mathrm{Cov}(S_1, S_2) = \rho$,

$$\mathrm{Cov}(X) = \begin{pmatrix} a^2 + \beta & a^2\rho \\ a^2\rho & a^2 + \beta \end{pmatrix}.$$

Its eigenvalues are $\beta + a^2(1 + \rho)$ and $\beta + a^2(1 - \rho)$, with respective unit eigenvectors $e_+ = (1, 1)^\top/\sqrt{2}$ and $e_- = (1, -1)^\top/\sqrt{2}$. They are distinct when $0 < \rho < 1$. If an orthogonal frame made the coordinates of X independent, it would diagonalize this covariance matrix; its directions would therefore have to be e_+, e_-, up to signs and permutation.

To rule out this unique candidate, set $T_\pm = (S_1 \pm S_2)/\sqrt{2}$ and $Y_\pm = e_\pm^\top X = aT_\pm + \eta_\pm$. Isotropy and independence of ξ imply that η_+, η_- are independent $\mathcal{N}(0, \beta)$ variables, jointly independent of S. Since $S_1^2 = S_2^2 = 1$, $T_\pm^2 = 1 \pm S_1S_2$, so $T_+^2T_-^2 = 0$ almost surely and $\mathbb{E}[T_\pm^2] = 1 \pm \rho$. Conditional independence of Y_+ and Y_- given S yields

$$\begin{aligned} \mathrm{Cov}(Y_+^2, Y_-^2) &= a^4 \,\mathrm{Cov}(T_+^2, T_-^2) \\ &= -a^4(1 - \rho^2) < 0. \end{aligned}$$

Independent coordinates would have independent squares and hence zero covariance between their squares, a contradiction. Signs and permutations do not affect this conclusion. □

For each ρ, we enumerate all four sign states exactly. Conditioned on $S = s$, the noisy observation $Y_t = X + tG$ satisfies $Y_t \mid S = s \sim \mathcal{N}\left(as, (\beta + t^2)I_2\right)$. Posterior sign probabilities are therefore available exactly by Bayes' rule and are evaluated numerically using stable log-sum-exp.

Gaussian regression gives $Df_t^\star(y) = \frac{\beta}{\beta+t^2} I_2 + \frac{a^2t^2}{(\beta+t^2)^2} \,\mathrm{Cov}\,(S \mid Y_t = y)$. Thus no learned approximation is involved in this experiment.

We sample independently $(Y_t, t), \qquad (Y'_s, s)$, with $t, s \sim \mathrm{Unif}\{0.2, 0.5, 0.8, 1.1\}$, and evaluate the normalized squared Frobenius commutator

$$\mathcal{C} := \frac{\|J_t(Y_t)J_s(Y'_s) - J_s(Y'_s)J_t(Y_t)\|_F^2}{\left(\|J_t(Y_t)\|_F \|J_s(Y'_s)\|_F + \varepsilon\right)^2}. \tag{78}$$

For each value of ρ, the reported commutator is averaged over independently sampled input/noise-level pairs. The full sweep uses 100 independent seeds and 3000 Jacobian pairs per seed.

At the two factorial endpoints, commutation holds up to numerical precision: in the standard basis at $\rho = 0$, and in the rotated basis equation 77 at $\rho = 1$. At intermediate dependence, the commutator becomes nonzero and peaks near $\rho = 0.7$.

As an additional diagnostic, for a candidate rotation V we evaluate the normalized off-diagonal energy

$$\mathcal{E}_{\mathrm{off}}(V) := \frac{\sum_j \|\,\mathrm{offdiag}(V^\top J_j V)\|_F^2}{\sum_j \|J_j\|_F^2 + \varepsilon}. \tag{79}$$

We compare the standard frame, the fixed 45° frame, and the best global two-dimensional rotation. At intermediate ρ, no single fixed rotation makes the Bayes-Jacobian family diagonal.

Finally, we visualize the local eigenframe at $\rho = 0.7, \qquad t = 0.5$. Eigenvectors are displayed only at locations where the normalized eigenvalue gap exceeds 0.03. Their spatially varying orientation illustrates directly the loss of a global fixed frame.

Figure 4 summarizes the commutator, fixed-frame, and local-eigenframe diagnostics.

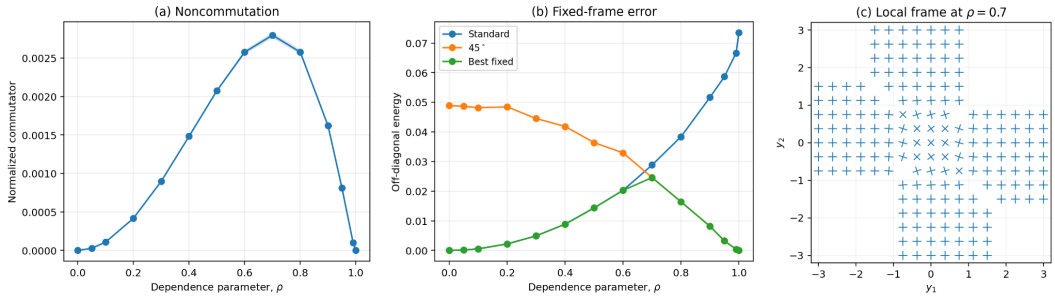


Figure 4: **Bayes-Jacobian commutation tracks orthogonal factorization.** **(a)** Mean normalized cross-input, cross-noise commutator versus the dependence parameter ρ; shaded bands are 95% normal-approximation intervals over 100 independent seeds. Commutation holds at $\rho = 0$, fails at intermediate dependence, and returns at $\rho = 1$ in a rotated factorial frame. **(b)** Normalized off-diagonal energy in the standard frame, the 45° frame, and the best fixed rotation. No fixed frame diagonalizes the family well at intermediate ρ, while the appropriate endpoint frame is exact. **(c)** Local eigenframe at $\rho = 0.7$ and $t = 0.5$. Crosses are shown only where the normalized eigengap exceeds 0.03; the spatially varying orientation illustrates the loss of a global eigenframe.

F.2 Independently trained denoisers

The main controlled neural experiment asks whether the same cross-noise geometry can emerge without shared weights or a shared time-conditioned architecture.

We use $X = U(aS + \xi)$, with $n = 12, \quad k = 3, \quad a = 1, \quad \beta = 0.25$, where $U \in \mathrm{St}(12, 3)$ is a dense random orthonormal frame and $\xi \sim \mathcal{N}(0, 0.25 I_3)$ is independent of S. Dependence is introduced only between S_1 and S_2 according to the sign-dependence law above, while S_3 is an independent Rademacher variable. The same nonfactorization conclusion holds for this embedded model when $0 < \rho < 1$. To see this, write $U = [u_1 \ u_2 \ u_3]$. The covariance of $aS + \xi$ has the three pairwise distinct, positive eigenvalues $\beta + a^2(1 + \rho)$, $\beta + a^2(1 - \rho)$, and $\beta + a^2$, with eigenvectors $(1, 1, 0)^\top/\sqrt{2}$, $(1, -1, 0)^\top/\sqrt{2}$, and $(0, 0, 1)^\top$, respectively. The ambient covariance of X has these same three simple positive eigenvalues and nine additional zero eigenvalues. Thus an ambient orthogonal frame with independent coordinates would have to contain $u_+ = (u_1 + u_2)/\sqrt{2}$ and $u_- = (u_1 - u_2)/\sqrt{2}$, up to signs and permutation. Their projected coordinates satisfy exactly the preceding squared-moment calculation:

$$\mathrm{Cov}\big((u_+^\top X)^2, (u_-^\top X)^2\big) = -a^4(1 - \rho^2) < 0.$$

They cannot be independent, regardless of the choice of basis in the nine-dimensional zero-eigenvalue subspace.

We use $\rho \in \{0, 0.25, 0.50, 0.70, 0.85, 1\}$, and train separate denoisers at $t \in \{0.2, 0.5, 0.8, 1.1\}$. The reference level used for one-noise frame recovery is $t_0 = 0.5$.

For each of ten outer seeds, one random U is drawn and reused across the entire ρ sweep. Within a fixed noise level, initialization and base random-number streams are paired across ρ to reduce nuisance variation. Different noise levels, however, use independent initializations and independent training streams. Consequently, denoisers at different t share neither weights nor optimization history.

Across 6 ρ-values $\times$ 4 noise levels $\times$ 10 outer seeds, the experiment contains 240 independently trained denoisers.

F.3 Network architecture and optimization

Each denoiser is a fully mixing residual MLP with

- hidden width 128;
- three residual blocks;

- SiLU activations;
- no noise-level or time input.

Thus every network is trained for exactly one value of t.

Training uses on-the-fly synthetic samples with

- batch size 2048;
- AdamW optimization;
- initial learning rate 10^{-3};
- cosine decay to 10^{-5};
- weight decay 10^{-6};
- gradient-norm clipping at 5.

Checkpoint selection uses only held-out denoising error relative to the exact Bayes posterior mean. In particular, neither the frame-fit objective nor any cross-noise transfer quantity is used for early stopping or model selection.

The low-noise condition $t = 0.2$ is the most difficult approximation problem. After the initial sweep, a uniform continuation pass is therefore applied to all 60 models trained at $t = 0.2$, using a new independent Bayes-validation split. Checkpoint selection during this continuation again uses only Bayes-approximation error. All three configured approximation-quality checks passed after this uniform cleanup. Across the 60 networks at $t = 0.2$, relative excess denoising risk averaged 0.8415%, with a maximum of 1.1725% (threshold 2%). Across all 240 networks, the maximum normalized squared symmetric-Jacobian error was 1.6089% (threshold 10%), and the maximum normalized antisymmetric Jacobian energy was 0.2600% (threshold 2%).

To define the Jacobian metrics, write $J_\theta = Df_\theta$, $S_\theta = (J_\theta + J_\theta^\top)/2$, $A_\theta = (J_\theta - J_\theta^\top)/2$, and $J^\star = Df_t^\star$. The metrics are, respectively,

$$\frac{\sum_j \|S_\theta(y_j) - J^\star(y_j)\|_F^2}{\sum_j \|J^\star(y_j)\|_F^2 + \varepsilon} \qquad \text{and} \qquad \frac{\sum_j \|A_\theta(y_j)\|_F^2}{\sum_j \|J_\theta(y_j)\|_F^2 + \varepsilon},$$

where $\varepsilon > 0$ is a numerical stabilizer. These metrics use 1,024 held-out Jacobians per condition; function-risk evaluations use 8,192 held-out samples per condition.

F.4 Jacobian computation and frame recovery

For every trained denoiser, exact network Jacobians are computed by automatic differentiation and symmetrized: $J_{\mathrm{sym}}(y) := \frac{1}{2}\left(Df_\theta(y) + Df_\theta(y)^\top\right)$. Symmetrization isolates the symmetric geometry relevant to the Bayes Jacobian and joint diagonalization. It excludes any antisymmetric component of the learned network Jacobian, which need not be purely numerical noise. The resulting algebra residual is therefore a diagnostic of symmetric geometry, not the complete learned-Jacobian defect.

The active three-dimensional subspace is estimated from the leading eigenspace of $\frac{1}{M}\sum_{j=1}^M J_{\mathrm{sym}}(y_j)^2$. Within the recovered subspace, we then perform approximate joint diagonalization of the held-out Jacobian family.

The primary frame $\widehat{U}_{t_0}$ is fitted using only held-out Jacobians from $t_0 = 0.5$. After fitting, this frame is frozen and evaluated on independently trained denoisers at all other noise levels.

As an oracle comparison, a separate target-specific frame is fitted independently at each evaluation noise level. The gap between the frozen-source-frame error and this target-specific fit isolates the extra cost of reusing the one-noise frame.

The main frame fit uses 512 held-out Jacobians at t_0, and neural geometry is evaluated using 1024 held-out Jacobians per condition.

For the exact-Bayes reference, posterior moments are evaluated in closed form. In the final numerical calculation we use 32,768 samples per condition and exploit the known two-dimensional rotation

structure of the (S_1, S_2) block. This produces a globally solved exact reference frame rather than relying on a generic approximate joint-diagonalization local optimum.

F.5 Geometric and functional metrics

For symmetric Jacobians $J_1, \dots, J_M$ and a candidate $W \in \mathrm{St}(n, k)$, the fixed-algebra residual used in the main text is

$$\mathcal{E}_{\mathrm{alg}}(W) = \frac{\sum_{j=1}^{M} \left\| J_j - W \,\mathrm{diag}\left(\mathrm{diag}(W^\top J_j W)\right) W^\top \right\|_F^2}{\sum_{j=1}^{M} \|J_j\|_F^2 + \varepsilon}. \tag{80}$$

For cross-noise transfer, the frame fitted at t_0 is frozen and evaluated at each target level. We additionally report the *excess algebra error*

$$\Delta\mathcal{E}_{\mathrm{alg}}(t) := \mathcal{E}_{\mathrm{alg}}(\widehat{U}_{t_0}; t) - \mathcal{E}_{\mathrm{alg}}(\widehat{U}_t; t), \tag{81}$$

where $\widehat{U}_t$ is fitted directly at the target level.

To verify that the geometric effect also matters functionally, we construct the exact optimal scalar posterior mean along the recovered axes while freezing the source-level frame. We then measure

$$\mathcal{E}_{\mathrm{func}}(t) := \frac{\mathrm{MSE}_{\mathrm{frozen}}(t) - \mathrm{MSE}_{\mathrm{Bayes}}(t)}{\mathrm{MSE}_{\mathrm{Bayes}}(t)}. \tag{82}$$

This separates failure of the frame itself from ordinary neural approximation error.

F.6 Uncertainty and paired comparisons

For the main transfer curves, non-reference noise levels are first averaged within each outer seed and ρ value. Uncertainty is then computed across the ten independent outer seeds.

The primary contrast compares the intermediate-dependence regime $\rho = 0.70$ with the factorial endpoints $\rho = 0$ and $\rho = 1$. These contrasts are paired within seed and use Student-t 95% confidence intervals.

This paired design is important because the same random latent frame and, where appropriate, matched base random streams are reused across the ρ sweep, while different noise levels remain independently trained. The resulting comparison therefore isolates the change in the latent dependence structure rather than variation caused by different embeddings or unrelated initializations.

G Experimental Details and Robustness for Cross-Noise Pooling

This section gives the implementation details and robustness controls for the fixed-total-budget pooling experiment in section 6.2. The purpose of the experiment is to isolate the statistical advantage of estimating one shared frame from data collected across several noise levels.

The data follow the correctly specified Rademacher–Gaussian model from section 5, with $a = 1$, $\beta = 0.25$. The scalar parameters are treated as known, so the only learned representation parameter is the Stiefel frame.

For a free matrix $W \in \mathbb{R}^{n \times k}$, we enforce the orthonormality constraint by reduced QR factorization at every optimization step: $U = \mathrm{qf}(W) \in \mathrm{St}(n, k)$. All risks use the exact denoiser equation 55; there is therefore no neural approximation error in this experiment.

G.1 Optimization protocol

Both shared and untied estimators are optimized using full-batch Adam with

- learning rate 5×10^{-2};
- 600 optimization steps;

- five independent random restarts.

The shared estimator chooses the restart with the smallest total empirical training loss. The untied estimator is given the stronger advantage of choosing the best restart separately for every level-specific frame.

Within every paired comparison, shared and untied estimators use exactly the same training examples and exactly the same test examples.

In the main balanced experiment, datasets are nested as the total sample budget N increases. For example, the $N = 512$ training set is a subset of the $N = 1024$ training set for the same seed and number of noise levels L. This reduces unnecessary Monte Carlo variation in the sample-efficiency curves.

G.2 Fixed-total-budget experiment

We use $n = 20, \quad k = 5$, and vary $L \in \{1, 2, 4, 8, 16\}$ noise levels over the interval $[0.5, 1.5]$. For each pair (N, L), the total number of independent noisy training triples is fixed at N.

The shared estimator fits one frame $U \in \mathrm{St}(20, 5)$ using all N triples.

The untied estimator fits $U_1, \ldots, U_L \in \mathrm{St}(20, 5)$, with the same total budget divided across the L levels.

This experiment therefore holds total data fixed while changing only how many representation parameters must be estimated.

The full sweep uses $N \in \{512, 1024, 2048, 4096\}$ in the principal $N \times L$ visualization, together with additional sample-size evaluations used in the sample-efficiency plots.

The qualitative behavior is consistent across all tested budgets: the shared estimator is nearly insensitive to increasing L, whereas the untied estimator deteriorates as the same total number of training triples must support more separately estimated frames.

At $L = 1$, the shared and untied parameterizations coincide up to numerical optimization tolerance. This serves as a direct implementation check.

G.3 Control: fixed number of examples per noise level

The fixed-total-budget experiment intentionally makes each untied frame see fewer examples as L grows. To distinguish the pooling effect from the simple fact that more tasks are present, we perform a second control in which every noise level receives exactly 128 training triples.

The total training budget is therefore $N = 128L$.

Under this design, every untied frame sees the same number of training examples regardless of L. Accordingly, its mean excess denoising risk remains approximately constant:

L	2	4	8	16
untied excess risk	0.126	0.136	0.132	0.138

(83)

The shared estimator, however, benefits from the growing total number of examples because every observation informs the same frame. Its mean excess risk decreases from 0.0624 at $L = 2$ to 0.00859 at $L = 16$.

This control uses five paired random seeds.

G.4 Control: i.i.d. sampling of the noise level

The theoretical result in theorem 5.1 is formulated for i.i.d. triples whose noise levels are random. The main experiment, in contrast, uses a balanced allocation across levels in order to remove random occupancy variation and thereby make the comparison more favorable to the untied baseline.

We therefore repeat the experiment with $N = 1024$ and draw the noise level of every training triple independently and uniformly from the same L-point grid.

Across ten paired seeds, the shared excess risk remains approximately 0.015–0.017 over $L = 1, \dots, 16$, while the untied excess risk increases from approximately 0.0172 at $L = 1$ to 0.255 at $L = 16$.

Thus the fixed-budget pooling effect is not an artifact of balanced or stratified allocation across noise levels.

G.5 Optimization and representation-recovery diagnostics

A possible alternative explanation for the shared estimator's lower test error is that the larger untied model is simply harder to optimize. The training losses argue against this explanation.

Across the 600 paired comparisons in the main experiment, the untied model obtains training loss no larger than that of the shared model in $591/600 = 98.5\%$ of cases.

Nevertheless, for every tested (L, N) pair with $L > 1$, the untied model has larger test excess risk in all 30 paired random seeds.

Thus the richer untied class generally fits the training objective at least as well, while generalizing worse.

We also monitor recovery of the representation itself. Let $P_\star = U_\star U_\star^\top, \qquad \widehat{P} = \widehat{U}\widehat{U}^\top$. The active-subspace error is measured by $\|\widehat{P} - P_\star\|_F$. To measure recovery of the individual non-Gaussian axes, the columns of $\widehat{U}$ are aligned to $U_\star$ up to signed permutation, and the resulting frame discrepancy is reported separately.

Both representation diagnostics exhibit the same qualitative shared-versus-untied behavior as the denoising excess risk.

H Experimental Details for the Bias–Complexity Crossover

This section gives the exact population formulas and finite-sample protocol for the deliberately misspecified experiment in section 6.2. Unlike the previous experiment, the clean distribution here does not admit one frame that is optimal over the entire noise range.

The clean random vector is

$$X = S^{(1)} + bR_\theta S^{(2)} + \xi, \tag{84}$$

where $S^{(1)}, S^{(2)} \overset{\text{ind}}{\sim} \mathrm{Unif}\{-1, +1\}^2$, $b = 0.5, \qquad \theta = \frac{\pi}{4}, \qquad \xi \sim \mathcal{N}(0, \beta I_2), \qquad \beta = 0.02.$

Equivalently, P_X is the equal-weight 16-component Gaussian mixture $P_X = \frac{1}{16}\sum_{c=1}^{16} \mathcal{N}(\mu_c, \beta I_2)$, with component means $\mu_{s,r} = s + bR_\theta r, \qquad s, r \in \{-1, +1\}^2$.

The two Rademacher structures therefore live in different orientations: the stronger component is axis-aligned while the weaker component is rotated by 45°.

H.1 Exact unrestricted Bayes denoiser

Write $q_t := \beta + t^2, \qquad \alpha_t := \frac{\beta}{q_t}$. For noisy observation $Y_t = X + tG$, the posterior mixture weights are $w_c(y, t) = \frac{\exp\left(-\|y - \mu_c\|^2/(2q_t)\right)}{\sum_d \exp(-\|y - \mu_d\|^2/(2q_t))}$.

Gaussian regression inside each mixture component gives the exact Bayes denoiser $f_t^\star(y) = \alpha_t y + (1 - \alpha_t)\sum_{c=1}^{16} w_c(y, t)\mu_c$.

H.2 Exact coordinatewise denoiser for a candidate frame

Let $V_\phi = [v_1(\phi), v_2(\phi)] \in O(2)$ be a rotation by angle ϕ, and define the projected mixture means $m_{c,j} := v_j^\top \mu_c$.

The exact scalar marginal Bayes map in coordinate j is

$$h_{\phi,j,t}(z) = \alpha_t z + (1-\alpha_t)\frac{\sum_{c=1}^{16} m_{c,j}\exp\big(-(z-m_{c,j})^2/(2q_t)\big)}{\sum_{c=1}^{16}\exp(-(z-m_{c,j})^2/(2q_t))}. \tag{85}$$

The corresponding coordinatewise denoiser is therefore

$$g_{\phi,t}(y) = V_\phi \begin{pmatrix} h_{\phi,1,t}(v_1^\top y) \\ h_{\phi,2,t}(v_2^\top y) \end{pmatrix}. \tag{86}$$

Thus the coordinatewise scalar estimators are exact. The only restricted or learned quantity is the frame angle ϕ.

H.3 Likelihood-weighted population calculation

The primary noise interval is $[t_0, T] = [0.35, 0.70]$. To match the weighting in the total-correlation/regret identity, we use the *normalized* noise density

$$C_{t_0,T} := \int_{t_0}^{T} t^{-3}\,dt = \frac{t_0^{-2} - T^{-2}}{2}, \qquad \pi_{t_0,T}(t) := \frac{t^{-3}}{C_{t_0,T}} \quad (t \in [t_0, T]). \tag{87}$$

All population risks reported below are expectations under $\pi_{t_0,T}$, whereas equation 9 uses the unnormalized measure $t^{-3}\,dt$.

We sample this distribution by drawing $u \sim \mathrm{Unif}[0,1]$ and setting

$$t(u) = \left[t_0^{-2} - u\left(t_0^{-2} - T^{-2}\right)\right]^{-1/2}. \tag{88}$$

Candidate frames are restricted to the grid

$$\Phi := \{0^\circ, 0.5^\circ, 1^\circ, \ldots, 89.5^\circ\}. \tag{89}$$

For $r_\phi(t) := \mathcal{R}_{V_\phi}(t)$ and the eight equal-mass noise bins $I_1, \ldots, I_8$ induced by equation 88, define the normalized, grid-restricted structural risks

$$\begin{aligned} \overline{\mathcal{B}}^{\Phi}_{\mathrm{share}} &:= \min_{\phi\in\Phi} \mathbb{E}_{t\sim\pi_{t_0,T}} r_\phi(t), \\ \overline{\mathcal{B}}^{\Phi}_{\mathrm{point}} &:= \mathbb{E}_{t\sim\pi_{t_0,T}} \min_{\phi\in\Phi} r_\phi(t), \\ \overline{\mathcal{B}}^{\Phi}_{\mathrm{untied},8} &:= \frac{1}{8}\sum_{\ell=1}^{8} \min_{\phi\in\Phi} \mathbb{E}_{t\sim\pi_{t_0,T}(\cdot|I_\ell)} r_\phi(t). \end{aligned} \tag{90}$$

Here $r_\phi(t)$ is the full-vector excess Bayes MSE from equation 5, as computed in the experiment; the shared and bin-specific empirical frames are selected using full-vector squared loss, whose frame-independent Bayes-risk term cancels in expectation.

Population expectations under the Gaussian mixture are evaluated using tensor-product Gauss–Hermite quadrature of order 14 in the two Gaussian coordinates.

The t^{-3}-weighted noise integral is divided into eight equal-mass bins in u, with six-node Gauss–Legendre quadrature within each bin. Increasing either quadrature order does not change the reported population values at the displayed precision.

The resulting normalized, grid-restricted structural risk floors are

$$\overline{\mathcal{B}}^{\Phi}_{\mathrm{point}} = 0.01705875, \tag{91}$$

$$\overline{\mathcal{B}}^{\Phi}_{\mathrm{untied},8} = 0.01707878, \tag{92}$$

$$\overline{\mathcal{B}}^{\Phi}_{\mathrm{share}} = 0.01851098. \tag{93}$$

Thus the normalized pointwise (continuous-in-noise) and eight-bin tying gaps, both restricted to Φ, are respectively

$$\overline{\Delta}^{\Phi}_{\mathrm{tie}} := \overline{\mathcal{B}}^{\Phi}_{\mathrm{share}} - \overline{\mathcal{B}}^{\Phi}_{\mathrm{point}} = 0.00145223,$$
$$\overline{\Delta}^{\Phi}_{\mathrm{tie},8} := \overline{\mathcal{B}}^{\Phi}_{\mathrm{share}} - \overline{\mathcal{B}}^{\Phi}_{\mathrm{untied},8} = 0.00143220. \tag{94}$$

For the same angle grid, the unnormalized pointwise gap is $\Delta^{\Phi}_{\mathrm{tie}} = C_{t_0,T}\overline{\Delta}^{\Phi}_{\mathrm{tie}}$. It approximates the unrestricted Δ_{tie} of equation 9; equality is not asserted unless the unrestricted shared and pointwise optima are attained on the grid. The displayed values also use numerical quadrature over the continuous noise interval.

The best shared frame is $\phi^{\star}_{\mathrm{share}} = 0^{\circ}$; the first two bins prefer 45°, and the remaining six prefer 0°. Thus the preferred coordinatewise representation genuinely changes inside the chosen noise interval.

H.4 Finite-sample crossover protocol

We use total budgets $N \in \{64, 128, 256, 512, 1024, 2048, 4096, 8192\}$.

The weighted noise interval is partitioned into the same eight equal-mass bins. Every bin receives exactly $N/8$ independent training triples.

Within each bin, u is sampled uniformly over that bin's interval and mapped to $t(u)$ using equation 88. Thus the noise scale remains continuous inside each bin and follows the appropriate conditional t^{-3}-weighted law.

The shared estimator minimizes empirical squared denoising loss over the angle grid equation 89 using all N training triples.

The untied estimator minimizes the same objective separately in each of the eight bins, so each level-specific frame receives only $N/8$ training triples.

Datasets are nested across N within each random seed, and shared and untied estimators always use exactly the same training triples.

The experiment uses 100 paired random seeds.

After a frame has been selected, its performance is evaluated using the precomputed population structural risk rather than a finite test set. This removes test-set Monte Carlo noise and isolates representation-estimation error.

H.5 Noise-window controls

To verify that the asymptotic disadvantage of sharing comes from a change of preferred representation rather than from some unrelated property of the clean distribution, we keep the clean law equation 84 fixed and change only the positive-noise window.

Low-noise window. For $t \in [0.20, 0.33]$, the 45° frame is optimal throughout the interval. The measured normalized, grid-restricted difference $\overline{\Delta}^{\Phi}_{\mathrm{tie}}$ is below numerical precision: 1.7×10^{-18}.

High-noise window. For $t \in [0.45, 0.75]$, the 0° frame is optimal throughout, and the measured tying penalty is numerically zero.

Transition window. For the primary interval $t \in [0.35, 0.70]$, the pointwise-optimal grid frame changes with t, producing the nonzero normalized, grid-restricted penalty $\overline{\Delta}^{\Phi}_{\mathrm{tie}} \approx 1.45 \times 10^{-3}$.

Thus the large-sample disadvantage of sharing appears specifically when the evaluated noise range crosses a region in which the optimal representation changes.

Table 1: Complete finite-sample bias–complexity crossover. Shared and untied estimators use the same total training budget and the same samples within each paired seed.

N	Shared risk	Eight-bin untied risk	Shared − untied
64	0.02347 ± 0.00123	0.02592 ± 0.00060	-0.00246 ± 0.00115
128	0.02163 ± 0.00099	0.02447 ± 0.00054	-0.00284 ± 0.00087
256	0.02013 ± 0.00068	0.02293 ± 0.00046	-0.00280 ± 0.00066
512	0.01920 ± 0.00040	0.02137 ± 0.00042	-0.00217 ± 0.00049
1024	0.01871 ± 0.00005	0.01956 ± 0.00028	-0.00084 ± 0.00028
2048	0.01860 ± 0.00002	0.01847 ± 0.00018	0.00012 ± 0.00018
4096	0.01856 ± 0.00001	0.01773 ± 0.00009	0.00083 ± 0.00010
8192	0.01854 ± 0.00001	0.01747 ± 0.00006	0.00107 ± 0.00006

H.6 Complete crossover results

Table 1 reports the complete finite-sample sweep. Entries are means with approximate 95% confidence intervals over 100 paired seeds. Negative values of the final column favor the shared estimator.

The finite-sample pattern is therefore the one predicted by the bias–complexity decomposition. At small N, the shared estimator benefits from pooling all observations into one frame estimate. As N grows, estimation error shrinks for both classes, exposing the nonzero structural floor of the shared parameterization. The ordering therefore reverses once the persistent tying penalty dominates the finite-sample estimation benefit.

I Learned-Representation Experiments

This section gives the complete implementation and evaluation details for the learned-representation experiments summarized in section 6.3. We first study VAE-family representations of dSprites, where the empirical aggregate posterior admits exact finite-mixture Bayes calculations. We then study learned CIFAR-10 autoencoder latents with learned time-conditioned denoisers and controlled interventions on cross-coordinate dependence.

I.1 VAE-family representations on dSprites

We train matched convolutional implementations of four representation objectives: VAE, β-VAE, FactorVAE, β-TCVAE. All models use the same encoder/decoder architecture and latent dimension $k = 10$.

Each model is trained for $12{,}000$ optimization steps with batch size $= 256$, learning rate $= 10^{-4}$, using $500{,}000$ dSprites examples.

The objective coefficients are

$$\beta = 4 \qquad \text{for } \beta\text{-VAE}, \tag{95}$$

$$\gamma = 10 \qquad \text{for FactorVAE}, \tag{96}$$

$$\beta_{\mathrm{TC}} = 6 \qquad \text{for } \beta\text{-TCVAE}. \tag{97}$$

For FactorVAE, the total-correlation penalty uses a discriminator with five fully connected hidden layers of width 512, each followed by a LeakyReLU activation (negative slope 0.2), and a scalar-logit output. The discriminator is trained with Adam (learning rate 10^{-4}, $\beta_1 = 0.5$, $\beta_2 = 0.9$). At each training step, we first freeze the discriminator and update the VAE using the reconstruction loss, the KL term, and the weighted discriminator-logit estimate of total correlation ($\gamma = 10$). We then perform one discriminator update on a separate minibatch: samples from its joint aggregate posterior are contrasted with samples obtained by independently permuting each latent coordinate across that minibatch, using binary cross-entropy with logits. Thus each VAE update is followed by one discriminator update.

For each objective, we train five independent representation seeds, $0, 1, 2, 3, 4$, giving a total of 20 independently trained learned representations.

I.2 Empirical aggregate posterior

For every held-out dSprites example x_i, retain the encoder mean $\mu_i \in \mathbb{R}^k$ and diagonal posterior variance $v_i \in \mathbb{R}^k_+$.

The empirical aggregate posterior is represented using all $M = 237{,}280$ held-out examples:

$$\widehat{q}_\phi(z) = \frac{1}{M}\sum_{i=1}^{M} \mathcal{N}\left(z; \mu_i, \mathrm{diag}(v_i)\right). \tag{98}$$

Before evaluation, the representation is centered and divided by a single global scalar scale estimated from $45{,}000$ held-out posterior draws. Coordinates are *not* whitened individually, since coordinate-wise whitening would alter the relative geometry that the experiment is intended to measure.

We treat equation 98 as an empirical plug-in distribution. We do not claim that this finite mixture is exactly equal to the population aggregate posterior.

I.3 Exact Bayes calculations for the empirical mixture

Let $Z_0 \sim \widehat{q}_\phi$ and corrupt it by isotropic Gaussian noise: $Z_t = Z_0 + t\varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I_k)$.

For mixture component i, the corrupted observation distribution is $Z_t \mid i \sim \mathcal{N}\left(\mu_i, \mathrm{diag}(v_i + t^2\mathbf{1})\right)$.

Conditioned on component i and observation $Z_t = y$, Gaussian regression gives $m_{i,t}(y) = \mu_i + \frac{v_i}{v_i+t^2} \odot (y - \mu_i)$, where division is coordinatewise.

Let $\omega_i(y,t) := \mathbb{P}(i \mid Z_t = y)$ denote the full k-dimensional mixture posterior weight. The exact plug-in Bayes posterior mean for the empirical mixture is $\widehat{f}^\star_t(y) = \sum_{i=1}^M \omega_i(y,t) m_{i,t}(y)$.

All mixture probabilities are evaluated using stable chunked log-sum-exp calculations.

To construct the optimal coordinatewise rule, for coordinate j we instead compute posterior mixture weights using only the one-dimensional marginal observation y_j. Applying the scalar posterior mean separately in every coordinate gives $\widehat{g}^{\mathrm{coord}}_t(y)$.

The corresponding exact plug-in coordinatewise regret is $\widehat{\mathcal{R}}(t) = \mathbb{E}\left\| \widehat{f}^\star_t(Z_t) - \widehat{g}^{\mathrm{coord}}_t(Z_t) \right\|^2$.

Thus the theory-facing denoisers in this experiment contain no learned denoiser approximation error.

Likewise, the posterior covariance of the empirical Gaussian mixture can be evaluated exactly. The Bayes Jacobian is therefore $\widehat{J}_t(y) = \frac{1}{t^2}\,\mathrm{Cov}_{\widehat{q}_\phi}\left(Z_0 \mid Z_t = y\right)$.

All denoisers and Jacobians used in the dSprites theory-facing calculations are exact for the finite empirical mixture, up to Monte Carlo integration and numerical arithmetic.

I.4 Finite-interval total-correlation/regret test

For every one of the 20 learned representations, we directly test the finite-interval identity from theorem 4.1 over $t \in [0.2, 1.1]$.

In the independent higher-precision validation, the regret integral is evaluated on a 33-point geometric grid using Simpson integration with the actual, nonuniform noise-level coordinates:

$$\int_{0.2}^{1.1} \frac{\widehat{\mathcal{R}}(t)}{t^3}\, dt. \tag{99}$$

For the paired endpoint total-correlation difference,

$$\widehat{\mathrm{TC}}(0.2) - \widehat{\mathrm{TC}}(1.1), \tag{100}$$

we use $4{,}000$ paired common-random-number samples per representation. The regret curve uses $1{,}024$ posterior samples per noise level, with common random draws reused across grid nodes.

Table 2: Learned dSprites representations. All theory-facing quantities use the complete $M = 237{,}280$-component empirical aggregate-posterior mixture. "TC drop" denotes $\widehat{\mathrm{TC}}(0.2) - \widehat{\mathrm{TC}}(1.1)$, and "regret" denotes $\int_{0.2}^{1.1} t^{-3}\widehat{\mathcal{R}}(t)\,dt$.

Representation	Clean TC	TC drop	Regret	Commutator	Basis transfer
VAE	14.124 ± 0.185	4.711 ± 0.045	4.699 ± 0.069	$(2.982 \pm 0.089) \times 10^{-2}$	0.00785 ± 0.00070
β-VAE	3.940 ± 0.032	2.057 ± 0.024	2.038 ± 0.026	$(5.169 \pm 0.265) \times 10^{-3}$	0.00181 ± 0.00052
FactorVAE	11.076 ± 0.289	2.964 ± 0.097	2.935 ± 0.114	$(1.622 \pm 0.058) \times 10^{-2}$	0.01964 ± 0.00094
β-TCVAE	1.722 ± 0.063	0.931 ± 0.050	0.933 ± 0.048	$(2.674 \pm 0.148) \times 10^{-3}$	0.00172 ± 0.00056

Integrating each sample's regret curve yields a Monte Carlo standard error that accounts for dependence across noise levels; the two endpoint estimates are similarly paired when computing the total-correlation drop.

For a grid-resolution check, we recompute each integral on the nested 17-point subgrid using the same samples. Across the 20 representations, the largest relative change between the 17- and 33-point integrals is 0.016%. The largest absolute seed-level relative discrepancy between equation 100 and equation 99 is 5.48%; however, the largest discrepancy divided by its estimated combined Monte Carlo standard error is 1.67. Here the combined standard error uses the endpoint and integral Monte Carlo errors in quadrature; these are distinct from the across-representation standard errors reported in Table 2.

I.5 dSprites geometric diagnostics

Bayes-Jacobian geometry is evaluated at $t \in \{0.20, 0.35, 0.50, 0.75, 1.10\}$.

For every noise level we use independent validation and test sets of 64 exact plug-in Bayes Jacobians.

The normalized cross-input, cross-noise commutator is evaluated using 2000 independently sampled Jacobian pairs.

For one-noise frame transfer, approximate joint diagonalizers are fitted at the three reference levels $t_0 \in \{0.35, 0.50, 0.75\}$. Each fit uses three random restarts. The resulting fixed frame is then evaluated on independent test Jacobians at all other geometry levels.

The transfer score reported in the main paper averages the excess off-diagonal error relative to a separately fitted target-level frame, over all reference levels and all non-reference evaluation levels.

The exact Jacobians satisfy the expected Bayes structural checks: the largest observed asymmetry is numerically zero, and the smallest observed Jacobian eigenvalue in the evaluated set is 5.75×10^{-3}.

I.6 Complete dSprites results

Table 2 reports the final method-level results. Uncertainty is the standard error across five independently trained representation seeds.

Defining the per-seed relative gap as integrated regret minus TC drop, divided by TC drop, the seed-averaged relative gaps are

$$\text{VAE:} \quad -0.26\% \pm 0.71\%, \tag{101}$$

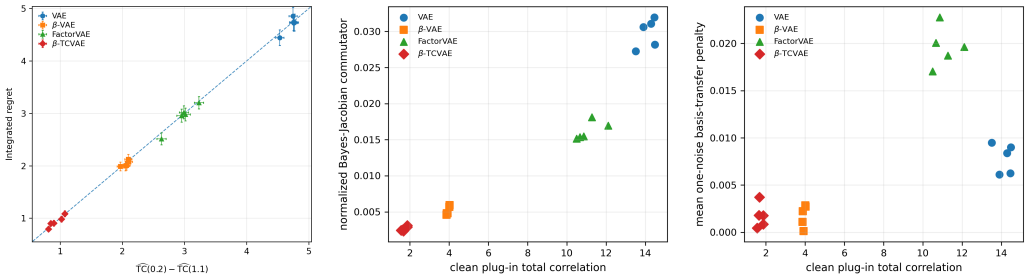
$$\beta\text{-VAE:} \quad -0.90\% \pm 0.88\%, \tag{102}$$

$$\text{FactorVAE:} \quad -1.05\% \pm 0.92\%, \tag{103}$$

$$\beta\text{-TCVAE:} \quad +0.23\% \pm 1.59\%. \tag{104}$$

Thus the finite empirical-mixture experiment numerically recovers the theorem-backed TC–regret relation across independently learned representations rather than only in a single example.

Figure 5 visualizes the identity test and the two geometry diagnostics across all 20 learned representations.



(a) **(b)** **(c)**

Figure 5: **Multi-seed dSprites experiment on learned VAE representations.** Each point is one independently trained representation; there are five seeds for each of four objectives, and all quantities use the full 237,280-component held-out aggregate-posterior mixture. **(a)** The higher-precision finite-interval total-correlation drop closely follows the exact plug-in Bayes regret integral predicted by equation 6; error bars denote approximate 95% Monte Carlo intervals conditional on each learned representation. **(b)** Lower clean plug-in total correlation is accompanied by more nearly commuting exact Bayes Jacobians; the pooled descriptive Spearman correlation is 0.983. **(c)** One-noise basis transfer has a weaker pooled association with clean TC ($\rho = 0.687$) and is non-monotone across objectives, with FactorVAE providing the clearest exception. Panels (b)–(c) are geometric diagnostics, not distribution-free approximate converses.

I.7 Commutation and basis-transfer trends

The mean normalized Bayes-Jacobian commutators have the same method-level ordering as clean plug-in total correlation: β-TCVAE $<$ β-VAE $<$ FactorVAE $<$ VAE.

Across all 20 independently trained representations, the pooled descriptive Spearman correlation between clean plug-in total correlation and the commutator diagnostic is $\rho_{\mathrm{Spearman}} = 0.983$.

Because the representations are grouped by training objective, this pooled correlation is descriptive. Between-objective differences may dominate the association, so it does not establish an equally strong relationship within each objective or support an independent-sample significance test. The same qualification applies to the pooled frame-transfer correlation reported below.

The one-noise frame-transfer statistic is substantially less monotone. Its pooled descriptive Spearman correlation with clean total correlation is $\rho_{\mathrm{Spearman}} = 0.687$.

FactorVAE provides the clearest example of this distinction: it has lower clean plug-in total correlation than the standard VAE but larger one-noise transfer error.

This behavior is consistent with the asymmetric approximate theory: theorem 4.1 gives an exact forward relation between total correlation and integrated coordinatewise denoising regret, whereas the paper does not predict a monotone quantitative relation from clean dependence to one-noise frame-transfer error.

I.8 Empirical-mixture-size validation

A finite aggregate-posterior mixture can itself induce dependence through the shared mixture index. Before fixing the final protocol, we therefore perform a nested mixture-size validation using $M \in \{80{,}000, 160{,}000, 237{,}280\}$ with common random numbers.

From $M = 160{,}000$ to the complete $M = 237{,}280$ held-out mixture, the changes in the finite-interval endpoint TC drop are

	VAE	β-VAE	FactorVAE	β-TCVAE
relative change	2.08%	3.40%	6.00%	0.50%

(105)

corresponding respectively to 0.90, 0.74, 1.28, 0.06 paired Monte Carlo standard errors.

This validation motivated using the complete held-out empirical posterior for every final run rather than an arbitrary smaller component set.

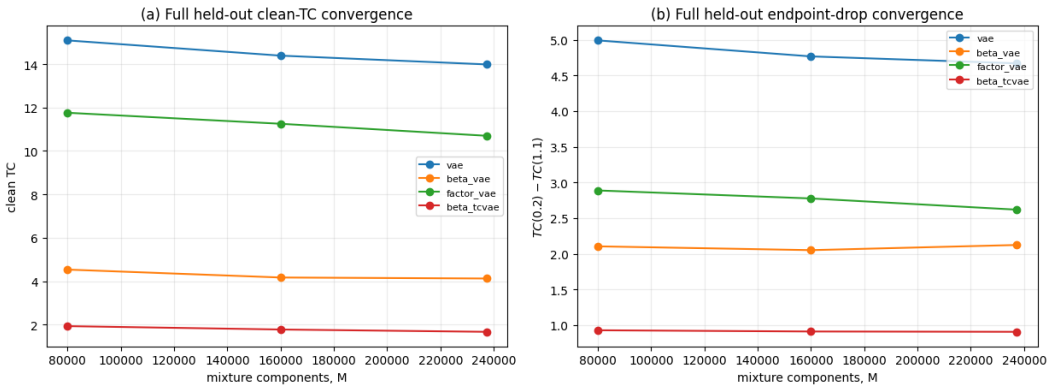


Figure 6: **Empirical-mixture-size validation used to select the final learned-latent protocol.** Nested aggregate-posterior mixtures use $M \in \{80{,}000, 160{,}000, 237{,}280\}$ components, with 237,280 equal to the entire dSprites set held out from representation training. **(a)** Clean plug-in total correlation. **(b)** The theorem-relevant finite-interval drop $\widehat{\mathrm{TC}}(0.2) - \widehat{\mathrm{TC}}(1.1)$. The final $160{,}000 \to 237{,}280$ changes are small relative to paired Monte Carlo uncertainty, motivating the use of the complete held-out mixture in the final 20-run study.

Figure 6 shows the nested-mixture validation.

I.9 CIFAR-10 AUTOENCODER REPRESENTATIONS

We next study a setting in which both the latent representation and the denoiser are learned from real images.

CIFAR-10 images are scaled to $[-1, 1]$.

The autoencoder uses three stride-2 convolutional encoder layers with $64,\ 128,\ 256$ channels, followed by a 32-dimensional linear bottleneck. The decoder mirrors the encoder using transposed-convolution layers.

The autoencoder is trained on a fixed $30{,}000$ image subset of the CIFAR-10 training set for 8 epochs using AdamW with learning rate $= 2 \times 10^{-3}$, weight decay $= 10^{-4}$.

Latent coordinates are centered and standardized using moments computed from the autoencoder training set.

I.10 MARGINAL-PRESERVING FACTORIAL CONTROLS

Before constructing the factorial controls, we apply a fixed smooth invertible coordinatewise contrast map

$$z_j \longmapsto z_j + \alpha_j z_j^3, \tag{106}$$

where the coefficients α_j are deterministically permuted over the interval $[0, 0.3]$. The coordinates are then recentered and rescaled.

Because equation 106 is coordinatewise and invertible, it changes marginal shape and improves axis identifiability without changing the copula.

From these transformed real latents, we construct three conditions.

Original latents. The transformed empirical latent distribution is used without further modification.

Factorial control. Every coordinate is independently permuted across training examples. This preserves each empirical one-dimensional marginal exactly and is intended to suppress cross-coordinate dependence, but the fixed finite empirical distribution need not factorize exactly.

Rotated factorial control. A single fixed random matrix $Q \in O(32)$ is applied to the shuffled samples. The planted rotation supplies known reference directions for frame recovery, but does not imply exact independence of the fixed empirical distribution.

These are latent-space distributional interventions; the shuffled samples are not decoded and treated as image datasets.

I.11 Latent denoiser architecture and training

For each condition we train a fully mixing time-conditioned residual MLP with

- hidden width 384;
- four SiLU residual blocks;
- a 64-dimensional sinusoidal embedding of $\log \sigma$.

The network is trained to predict the clean latent Z from $Z_\sigma = Z + \sigma G$, with $\sigma \sim$ log-uniform$[0.08, 3]$.

Training uses 25 epochs, batch size $= 512$, with AdamW, learning rate $= 2 \times 10^{-3}$, weight decay $= 10^{-4}$.

The confirmatory endpoint experiment uses five matched training seeds $101,\ 202,\ 303,\ 404,\ 505$.

I.12 Frame fitting and evaluation

Geometry is evaluated at $\sigma \in \{0.15, 0.30, 0.60, 1.20\}$, with the predeclared source level $\sigma_0 = 0.60$.

For each model, we use 256 held-out examples to fit the source frame and 512 disjoint held-out examples for evaluation.

Exact 32×32 network Jacobians are obtained by automatic differentiation and symmetrized: $S_i = \frac{1}{2}\left(J_i + J_i^\top\right)$.

For an orthogonal frame U, define the normalized fixed-algebra residual

$$\mathcal{E}_{\mathrm{img}}(U;\sigma) = \frac{1}{m}\sum_{i=1}^{m} \frac{\|\operatorname{offdiag}(U^\top S_i U)\|_F^2}{\|S_i\|_F^2 + \varepsilon}. \tag{107}$$

This residual measures joint diagonalizability of symmetrized network Jacobians; it does not include their antisymmetric components or measure the full learned-Jacobian approximation error.

Frames are fitted by Jacobi/Givens approximate joint diagonalization using only source-level Jacobians. The fitted source frame is then frozen.

For the target-level oracle baseline, the same fitter is run independently at each evaluation level and includes the frozen source-level frame among its initializations.

I.13 Marginal-matched endpoint results

At the source level $\sigma_0 = 0.6$, the target-fitted held-out fixed-algebra residual is

$$\text{original latents:} \quad 0.05356 \pm 0.00427, \tag{108}$$

$$\text{factorial control:} \quad 0.002372 \pm 0.000119, \tag{109}$$

$$\text{rotated factorial:} \quad 0.002258 \pm 0.000101. \tag{110}$$

Uncertainties are standard errors across the five independently trained denoisers.

Thus marginal-preserving coordinate shuffling, intended to suppress dependence, reduces the observed symmetric-algebra defect by more than an order of magnitude.

This reduction is not explained by isotropy. For random orthogonal frames, the corresponding residuals are

$$\text{factorial control:} \quad 0.02694 \pm 0.00160, \tag{111}$$

$$\text{rotated factorial:} \quad 0.02601 \pm 0.00152. \tag{112}$$

The source-level fitted frame also aligns with the planted reference directions up to sign and permutation. The mean matched cosine is

$$\text{factorial control:} \quad 0.9639 \pm 0.0049, \tag{113}$$

$$\text{rotated factorial:} \quad 0.9544 \pm 0.0048. \tag{114}$$

The planted dense rotation is therefore recovered almost as well as the original coordinate reference frame.

I.14 One-noise frame transfer

Using only the frame fitted at $\sigma_0 = 0.6$, we evaluate excess algebra error at $\sigma \in \{0.15, 0.30, 1.20\}$.

Averaged across these three non-reference levels, the excess residual is

$$\text{factorial:} \quad (6.1 \pm 7.9) \times 10^{-6}, \tag{115}$$

$$\text{rotated factorial:} \quad (3.97 \pm 0.30) \times 10^{-5}, \tag{116}$$

$$\text{original latents:} \quad 0.001336 \pm 0.000102. \tag{117}$$

The shuffled controls therefore show very small observed source-frame transfer error, including after the planted rotation.

Unlike the synthetic experiment in section F.2, these CIFAR denoisers are time-conditioned and share parameters across noise levels. This experiment therefore tests whether the predicted input–output geometry remains visible in a learned image representation; it does not establish that the sharing itself arose without architectural weight sharing.

I.15 Controlled Gaussian-copula dependence path

To move beyond the shuffled-control and real-latent endpoints, we introduce a continuous dependence intervention starting from the shuffled empirical control (which may retain residual dependence).

For each coordinate, an empirical marginal CDF fitted to the factorial training latents is used to Gaussianize the samples: $z_j \mapsto g_j$. We then group the 32 coordinates into the fixed pairs $(1,2), (3,4), \ldots, (31,32)$. For every pair, define

$$g'_{2j-1} = g_{2j-1}, \tag{118}$$

$$g'_{2j} = \rho g_{2j-1} + \sqrt{1-\rho^2}\, g_{2j}. \tag{119}$$

Each coordinate is then mapped back through its original empirical quantile function. This approximately preserves all one-dimensional marginal distributions while introducing dependence independently inside the 16 coordinate pairs.

For the ideal Gaussian copula, the nominal total correlation is $\mathrm{TC}(\rho) = -8\log(1-\rho^2)$.

The dependence levels are fixed in advance as $\rho \in \{0, 0.2, 0.4, 0.6, 0.75\}$.

The achieved pair correlations closely match their nominal targets. The maximum unintended cross-pair correlation is 0.0167, and the maximum normalized marginal-quantile distortion is 0.0053.

I.16 Dependence-path training protocol

The dependence-path experiment uses the same denoiser architecture and five training seeds as the endpoint experiment.

Within every replicate, all five values of ρ use matched

- network initialization;
- minibatch-shuffle random streams;
- noise-level random streams;

- Gaussian-corruption random streams.

This produces a paired experimental design across dependence levels.

For each trained model, 192 held-out examples are used to fit the source frame, and 384 disjoint examples are used for geometric evaluation.

All reported uncertainties are computed across the five independently trained denoiser seeds, not across evaluation examples.

I.17 Dependence-path results

The source-level fixed-algebra residual increases from $0.002363 \pm 0.000116 \qquad (\rho = 0)$ to

$$0.004773 \pm 0.000133 \qquad (\rho = 0.2), \tag{120}$$
$$0.012002 \pm 0.000451 \qquad (\rho = 0.4), \tag{121}$$
$$0.016537 \pm 0.000950 \qquad (\rho = 0.6), \tag{122}$$

before decreasing slightly to $0.015602 \pm 0.000837 \qquad (\rho = 0.75)$.

Thus the effect rises strongly with introduced dependence and then saturates at the strongest intervention.

The per-seed Spearman correlations between nominal controlled total correlation and fixed-algebra residual are $0.9, \quad 0.9, \quad 1.0, \quad 0.9, \quad 0.9$. We do not interpret these values as evidence for a universal monotone map from total correlation to geometric defect. The empirically supported statement is weaker: controlled dependence produces a large and reproducible departure from the common diagonal algebra, followed by saturation.

The one-noise frame-transfer penalty exhibits the same qualitative rise-and-saturate behavior.

I.18 Additional training diagnostics and limitations

We report two additional diagnostics for the dependence-path experiment. Neither alone establishes unchanged approximation quality across the different data distributions.

First, final denoising MSE actually *decreases* across the dependence sweep, from approximately $0.316 \qquad (\rho = 0)$ to $0.281 \qquad (\rho = 0.75)$.

Second, at $\sigma_0 = 0.6$, the symmetrized-Jacobian positive-semidefiniteness check has zero negative-eigenvalue fraction at every dependence level. The mean Jacobian symmetry defect increases from 0.0955 to 0.1299.

Because the data distribution changes with ρ, the Bayes risk may also change: decreasing raw MSE does not rule out increased error relative to the corresponding Bayes denoiser. The symmetrized-Jacobian checks likewise do not establish proximity to the Bayes Jacobian. Distribution-dependent approximation error may therefore contribute to the observed geometric trends.

These experiments should nevertheless be interpreted in the direction supported by the theory. The shuffled controls exhibit small symmetric-frame residuals, and controlled dependence changes them systematically. However, by proposition 4.2, a small positive-noise algebra residual is not a distribution-free quantitative certificate that the clean latent law has small total correlation.

J Dependence Hidden Below the Gaussian Smoothing Scale

This experiment visualizes the mechanism behind proposition 4.2. The proposition shows analytically that dependence can remain substantial in the clean distribution while becoming arbitrarily difficult to detect from the Bayes denoiser at any fixed positive amount of Gaussian smoothing. Here we evaluate the same construction numerically.

Recall the family $p_m(x_1, x_2) = \varphi(x_1)\varphi(x_2)\,[1 + \epsilon \sin(mx_1)\sin(mx_2)]$, with $\epsilon = 0.8$ and frequencies $m \in \{1, 2, 3, 4, 5, 6, 7, 8, 10, 12, 15, 20\}$.

Both one-dimensional marginals are exactly standard Gaussian for every m, while the two coordinates remain dependent.

All positive-noise quantities are evaluated over $t \in [0.2, 1.1]$.

J.1 Clean and noisy total correlation

Because the marginals of P_m are exactly standard Gaussian, the clean total correlation is $\mathrm{TC}(P_m) = D_{\mathrm{KL}}\left(P_m \,\|\, \mathcal{N}(0, I_2)\right)$.

We estimate the relevant expectations using scrambled Sobol quasi-Monte Carlo under the independent Gaussian product reference.

The clean dependence remains substantial throughout the entire frequency sweep. Numerically, $\mathrm{TC}(P_1) \approx 0.0638$, and for large m the clean total correlation approaches approximately 0.086.

For $\epsilon = 0.8$, the uniform lower bound from proposition 4.2, $\mathrm{TC}(P_m) \geq \frac{\epsilon^2(1-e^{-2})^2}{8(1+\epsilon)}$, equals approximately 0.0332. Thus the clean dependence stays well above zero across the sweep.

By contrast, after Gaussian corruption the dependence is rapidly suppressed. At the smallest positive noise level used in the experiment, $t = 0.2$, the noisy total correlation decreases from $\mathrm{TC}(P_{1,0.2}) \approx 5.72 \times 10^{-2}$ to $\mathrm{TC}(P_{20,0.2}) \approx 3.47 \times 10^{-15}$. Larger values of t suppress the oscillatory dependence even more strongly.

J.2 Exact Bayes-Jacobian deviation

From section C.4, the smoothed density can be written $p_{m,t}(y) = \varphi_{1+t^2}(y_1)\varphi_{1+t^2}(y_2) r_{m,t}(y)$, where $r_{m,t}(y) = 1 + \delta_{m,t} \sin(\lambda_{m,t} y_1) \sin(\lambda_{m,t} y_2)$, with $\delta_{m,t} = \epsilon \exp\left(-\frac{m^2 t^2}{1+t^2}\right), \qquad \lambda_{m,t} = \frac{m}{1+t^2}$.

The exact Bayes denoiser satisfies $f^\star_{m,t}(y) = \frac{y}{1+t^2} + t^2 \nabla \log r_{m,t}(y)$, and therefore $Df^\star_{m,t}(y) - \frac{1}{1+t^2} I_2 = t^2 \nabla^2 \log r_{m,t}(y)$.

We evaluate the uniform deviation

$$\mathcal{D}_m := \sup_{t \in [0.2, 1.1]} \sup_{y \in \mathbb{R}^2} \left\| Df^\star_{m,t}(y) - \frac{1}{1+t^2} I_2 \right\|_{\mathrm{op}}. \tag{123}$$

Because $r_{m,t}$ depends on y only through the periodic phases $\lambda_{m,t} y_1, \qquad \lambda_{m,t} y_2$, the supremum over y is approximated numerically on a dense phase grid. The resulting value is then maximized over $t \in [0.2, 1.1]$.

The deviation is not monotone for the smallest frequencies, which is consistent with the asymptotic nature of proposition 4.2. It peaks near $m = 4$ at approximately $\mathcal{D}_4 \approx 0.451$. Afterward it rapidly collapses:

$$\mathcal{D}_{10} \approx 6.43 \times 10^{-2}, \tag{124}$$

$$\mathcal{D}_{15} \approx 1.16 \times 10^{-3}, \tag{125}$$

$$\mathcal{D}_{20} \approx 2.46 \times 10^{-6}. \tag{126}$$

Thus, even though the clean total correlation remains substantial, the entire positive-noise Bayes Jacobian family becomes nearly indistinguishable from the isotropic Gaussian Jacobian $(1 + t^2)^{-1} I_2$.

J.3 Cross-input, cross-noise commutators

We also evaluate the basis-free commutator diagnostic.

For every frequency m, draw $t, s \overset{\mathrm{iid}}{\sim} \mathrm{Unif}[0.2, 1.1]$ and sample inputs exactly from P_m using rejection sampling.

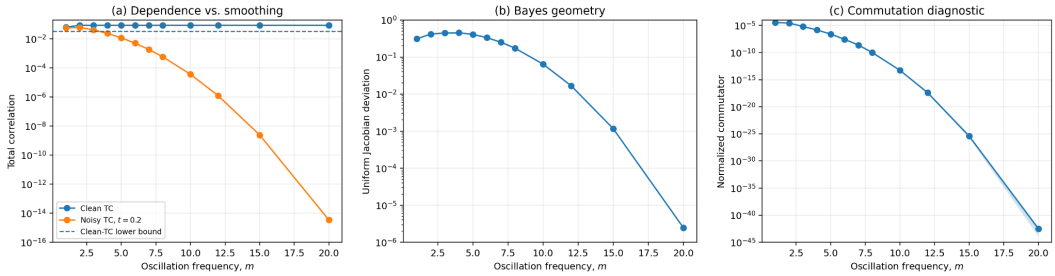


Figure 7: **Dependence can hide below the Gaussian smoothing scale.** **(a)** Clean total correlation remains bounded away from zero as the oscillation frequency m increases, while total correlation after the least positive noise level in the evaluation interval, $t = 0.2$, collapses. **(b)** The uniform exact Bayes-Jacobian deviation $\sup_{t\in[0.2,1.1],y} \|Df^\star_{m,t}(y) - (1+t^2)^{-1} I_2\|_{\mathrm{op}}$ tends rapidly to zero. **(c)** Mean normalized Bayes-Jacobian commutators for $t, s \overset{\mathrm{iid}}{\sim} \mathrm{Unif}[0.2, 1.1]$ collapse over the same frequency sweep; shaded bands are 95% normal-approximation intervals over 50 independent seeds.

For independently sampled noisy inputs, define $J_{m,t} := Df^\star_{m,t}(Y_t)$, $J'_{m,s} := Df^\star_{m,s}(Y'_s)$. We report the normalized squared Frobenius commutator

$$\mathcal{C}_m := \mathbb{E}\left[\frac{\|J_{m,t}J'_{m,s} - J'_{m,s}J_{m,t}\|_F^2}{\left(\|J_{m,t}\|_F\|J'_{m,s}\|_F + \varepsilon\right)^2}\right]. \tag{127}$$

The mean commutator decreases extremely rapidly with m:

$$\mathcal{C}_1 \approx 3.35 \times 10^{-5}, \tag{128}$$

$$\mathcal{C}_{10} \approx 5.20 \times 10^{-14}, \tag{129}$$

$$\mathcal{C}_{20} \approx 3.30 \times 10^{-43}. \tag{130}$$

The reported uncertainty bands are 95% normal-approximation intervals over 50 independent seeds.

Figure 7 summarizes the numerical scale obstruction across clean dependence, positive-noise geometry, and cross-noise commutation.

These results illustrate the instability established in proposition 4.2. As m increases, the clean total correlation remains bounded away from zero, while the dependence becomes increasingly difficult to detect after Gaussian smoothing. Both the Bayes-Jacobian deviation and the cross-noise commutators approach zero because the oscillatory interaction is suppressed at every fixed positive noise level.

The experiment illustrates the analytical result rather than providing an independent proof of it. Moreover, the lower bound on clean total correlation concerns the original coordinate frame. The observed decay of the commutators does not establish the corresponding impossibility result for the frame-optimized quantity $\inf_{V\in O(2)} \mathrm{TC}(V^\top X_m)$.