On the Limits of Latent Reuse in Diffusion Models

Yifeng Yu
Department of Mathematical Science
Tsinghua University
yyf22@mails.tsinghua.edu.cn

Lu Yu *
Data Science Institute
Shandong University
lu.yu@sdu.edu.cn

Abstract

Diffusion models are often trained in low-dimensional latent spaces, which are then reused for related but shifted datasets. In this work, we study when such latent reuse remains reliable under distribution shift. We consider a source-target setting in which both datasets are approximately low-dimensional but may lie near different subspaces. We show that freezing and reusing a source latent space induces a target-domain score error governed by two quantities: the principal-angle misalignment between the source and target subspaces, and the target ambient noise amplified by the diffusion time scale. We further establish an endpoint KL lower bound for continuous-time sampling under exact terminal initialization, translating this structural obstruction into a distribution-level limitation. To explore alternatives to frozen reuse, we analyze shared source-target representations and derive both mixed-distribution oracle bounds and target-domain score-estimation guarantees. Our results provide theoretical guidance on when latent reuse is reliable and when learning a shared representation may be necessary.

1 Introduction

Diffusion models have recently achieved remarkable breakthroughs across multiple domains [27, 26, 36]. Their success is driven by a denoising principle: a forward process gradually corrupts data with Gaussian noise, while a learned score function is used to reverse this process and generate new samples [29]. In many modern applications, however, diffusion models are not trained directly in the original ambient space. Instead, high-dimensional data such as high-resolution images are first compressed by an encoder, often the encoder of a variational autoencoder (VAE), and the diffusion model is trained in the resulting lower-dimensional latent space [28, 24, 7, 8, 34]. This strategy is computationally attractive and is consistent with the empirical observation that many high-dimensional datasets exhibit strong low-dimensional structure [25]. After pretraining, the representation map is often kept fixed and reused for related datasets, while a new score model is trained in the induced latent coordinates [7, 8, 17, 34].

While convenient, latent reuse raises a fundamental question: *when does a latent space learned from a source distribution remain reliable under target shift?* The answer is not obvious, since a representation that is effective for the source distribution need not preserve all directions relevant to the target score.

Existing theory provides important insight into why diffusion models can exploit low-dimensional structure. A growing line of work establishes score approximation, score estimation, and distribution recovery guarantees when the data lie on, or near, an unknown low-dimensional subspace or manifold, with rates depending primarily on the intrinsic dimension rather than the ambient dimension [9, 18, 33, 14, 22, 3, 31, 10, 37, 19, 15, 1]. These results, however, are largely single-distribution in nature: they explain the benefits of low-dimensional modeling, while leaving open how a learned representation behaves when reused under distribution shift.

*Corresponding author

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

A separate line of work studies transfer and adaptation of diffusion models across domains [11, 23, 30, 16, 21, 35, 13]. For example, recent work represents the target-domain score through a pretrained source-domain score together with additional guidance [23, 16], while other work establishes sample-efficient transfer guarantees for conditional diffusion models through the lens of representation learning [11]. These works show how pretrained diffusion models or learned representations can be adapted across related tasks. Complementary to this perspective, we study a different widely used practice: freezing a data latent space learned from a source distribution and reusing it as the coordinate system for score estimation on a shifted target distribution. This distinction is important because latent reuse can fail for reasons that are invisible in existing single-distribution low-dimensional theory. Even when both source and target distributions are approximately low-dimensional, the source latent space may be poorly aligned with the target signal subspace, and the target distribution may contain ambient noise outside the reused representation. In this case, no downstream score model trained only on the reused latent coordinates can fully recover the missing target directions.

Motivated by this gap, we take a first step toward a mathematical understanding of latent reuse under source-target distribution shift. We consider a model where both distributions are approximately low-dimensional but may concentrate near different subspaces, allowing us to isolate how representation mismatch affects score estimation. Our analysis shows that frozen reuse can fail for geometric reasons that cannot be resolved solely by more target samples or a more expressive downstream score model. This motivates mixed source–target training as an alternative strategy for learning a shared representation adapted to both distributions. Towards this, our contributions can be summarized as follows.

- **Identifying the limits of frozen latent reuse.** Theorem 2 establishes an irreducible target-domain score error for the projected score class, even with an arbitrarily expressive downstream score model. Proposition 3 translates this obstruction into an endpoint KL lower bound for continuous-time sampling under exact terminal initialization
- **Quantifying when frozen reuse is reliable.** Proposition 4 constructs an explicit frozen coordinate comparator. Theorem 6 gives a finite-sample guarantee for a general core class, and Corollary 7 specializes it to ReLU networks. These results separate statistical, approximation, and geometric errors, thereby clarifying when the cost of latent reuse is small and when it remains intrinsic.
- **Mixed training beyond frozen reuse.** In Proposition 8, Theorem 10, and Corollary 11, we analyze mixed source–target training through a shared-projector oracle. Corollary 9 gives the corresponding target-domain guarantee after freezing the shared representation. Controlled Gaussian-subspace experiments illustrating our theoretical findings are provided in Appendix D.

Overall, our results suggest that latent reuse should be viewed not only as a downstream score-estimation problem, but also as a representation-level constraint governed by the relative geometry of the source and target low-dimensional structures. In particular, the bounds show that low-dimensionality alone does not guarantee reliable reuse: alignment, latent dimension, and ambient-noise effects all influence the resulting score error. These observations provide a step toward a more precise theoretical understanding of when frozen latent reuse is appropriate and when learning a shared representation may be beneficial.

Notation. For a matrix A, let $\|A\|_{\mathrm{op}}$ and $\|A\|_F$ denote its operator and Frobenius norms. For symmetric M, let $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ denote its extreme eigenvalues. For orthonormal U, write $P_U := UU^\top$ and $P_U^\perp := I - P_U$. For a metric class $(\mathcal{F}, d)$, let $\mathcal{N}(\epsilon, \mathcal{F}, d)$ be its ϵ-covering number. Denote $a \lesssim b$ if $a \leqslant Cb$ for an absolute constant C, and use $\widetilde{\mathcal{O}}(\cdot)$ to hide logarithmic factors.

2 Preliminaries

We begin by introducing the diffusion model used throughout the paper, followed by the source-target setting in which latent representations are learned, frozen, and reused under distribution shift.

2.1 Diffusion Models

Framework. We consider the Ornstein–Uhlenbeck forward process

$$\mathrm{d}X_t = -\frac{1}{2}X_t\,\mathrm{d}t + \mathrm{d}W_t, \qquad t \in [0,T],$$

where the initial point $X_0 \sim p_0$ follows an unknown data distribution p_0, and W_t denotes a standard D-dimensional Brownian motion. This process admits the explicit transition kernel

$$X_t \mid X_0 \sim \mathcal{N}\big(\alpha(t)X_0, h(t)I_D\big),$$

with $\alpha(t) = e^{-t/2}, h(t) = 1 - e^{-t}$. Thus, the marginal density of X_t is

$$p_t(x) = \int_{\mathbb{R}^D} \mathcal{N}\big(x; \alpha(t)x_0, h(t)I_D\big)p_0(x_0)\,\mathrm{d}x_0.$$

Sampling from p_0 can be performed by reversing the noising process [2, 5]. The corresponding reverse-time SDE is

$$\mathrm{d}X_t^{\leftarrow} = \left[\frac{1}{2}X_t^{\leftarrow} + \nabla_x \log p_{T-t}\big(X_t^{\leftarrow}\big)\right]\mathrm{d}t + \mathrm{d}\bar{W}_t, \tag{1}$$

where $\bar{W}_t$ is a standard Brownian motion, independent of W_t. The vector field $\nabla_x \log p_t(x)$ is called the *score function* for p_t. It is typically unknown, since p_t is not available in closed form.

Score Matching. In practice, one trains a parametrized score network $s_\theta(x,t)$ to approximate $\nabla_x \log p_t(x)$, typically through an L^2 score-matching objective of the form

$$\underset{\theta\in\Theta}{\text{minimize}} \quad \mathbb{E}\left[\big\|s_\theta(X_t,t) - \nabla_x \log p_t(X_t)\big\|_2^2\right].$$

Substituting the learned score estimate into the reverse process (1) yields the practical continuous-time backward SDE, which is typically approximated by a discretization method (e.g. [39, 6, 12]).

2.2 Source-target Setting

In this work, we study latent reuse in a source-target setting. The source distribution p_1 represents the distribution used to learn the latent representation, while the target distribution p_2 represents the shifted distribution on which this representation is reused. We assume that both distributions are approximately low-dimensional, with each lying near a possibly different linear subspace and perturbed by isotropic ambient noise.

Assumption 1. *For each domain $i \in \{1,2\}$, a data point $x \sim p_i$ is generated via*

$$x = A_i z_i + \varepsilon_i,$$

where $A_i \in \mathbb{R}^{D\times d_i}$ has orthonormal columns, i.e., $A_i^\top A_i = I_{d_i}$, $z_i \sim p_{i,z}$ is a latent variable in $\mathbb{R}^{d_i}$, and the ambient noise $\varepsilon_i \sim \mathcal{N}(0,\sigma_i^2 I_D)$ is independent of z_i, with $\sigma_i > 0$.

Assumption 1 extends the exact linear-subspace model commonly used in low-dimensional diffusion theory [9, 14, 38] by allowing isotropic ambient noise. Here, $A_i z_i$ represents the intrinsic low-dimensional signal, while ε_i captures deviations from $\mathrm{col}(A_i)$. The source and target subspaces are unknown and may differ, allowing us to quantify the effect of latent reuse through subspace misalignment and target ambient noise.

Under this model, Gaussian noising yields an explicit score decomposition.

Lemma 1. *Let $X_0 \sim p_i$ satisfy Assumption 1. Under the forward diffusion process $X_t|X_0 \sim \mathcal{N}(\alpha(t)X_0, h(t)I_D)$, define $\tilde{h}_i(t) := \alpha^2(t)\sigma_i^2 + h(t)$, then the exact score function $\nabla \log p_{i,t}(x)$ rigorously decomposes into mutually orthogonal components*

$$\nabla \log p_{i,t}(x) = \underbrace{A_i \nabla \log p_{i,t}^{\mathrm{LD}}(A_i^\top x)}_{s_\parallel(A_i^\top x,t)\text{: on-support score}} \underbrace{-\frac{1}{\tilde{h}_i(t)}(I_D - A_iA_i^\top)x}_{s_\perp(x,t)\text{: ortho. score}},$$

where $p_{i,t}^{\mathrm{LD}}(z') = \int_{\mathbb{R}^{d_i}} \phi_{i,t}(z'|z)p_{i,z}(z)\,\mathrm{d}z$ with $\phi_{i,t}(\cdot|z)$ being the Gaussian density function of $\mathcal{N}(\alpha(t)z, \tilde{h}_i(t)I_{d_i})$.

Lemma 1 decomposes the score into a latent component along $\mathrm{col}(A_i)$ and an explicit orthogonal component. The latent component depends only on $A_i^\top x$, while the orthogonal component has magnitude governed by the effective variance $\widetilde{h}_i(t) = \alpha^2(t)\sigma_i^2 + h(t)$. This decomposition is the basis for our analysis of latent reuse: when a source subspace is frozen and reused on a target distribution, any mismatch between $\mathrm{col}(A_1)$ and $\mathrm{col}(A_2)$ directly affects the target score.

For later use, we define $f_i^*(z,t) := \tilde{h}_i(t)\nabla \log p_{i,t}^{\mathrm{LD}}(z) + z$. Then, the decomposition in Lemma 1 can be written equivalently as

$$\nabla \log p_{i,t}(x) = \frac{1}{\widetilde{h}_i(t)} A_i f_i^*(A_i^\top x, t) - \frac{1}{\widetilde{h}_i(t)} x.$$

3 Main Results

We now present our main theoretical results. Section 3.1 analyzes frozen latent reuse on the target distribution and identifies the structural bias induced by subspace mismatch and target ambient noise. Section 3.2 complements this with an explicit comparator and a finite-sample bound for target-domain training with a frozen projector. Finally, Section 3.3 studies mixed source-target training as an alternative to frozen reuse.

3.1 Irreducible Error under Frozen Latent Reuse

We first identify a structural limitation of frozen latent reuse: even if the downstream score model is arbitrarily expressive, a fixed source latent space may leave irreducible target-domain error.

Let $V_1 \in \mathbb{R}^{D\times d_1}$ be the source projector, with $V_1^\top V_1 = I_{d_1}$. Under frozen latent reuse, the target score model takes the form

$$s_{V_1,\theta}(x,t) = \frac{1}{h(t)} V_1 f_\theta(V_1^\top x, t) - \frac{x}{h(t)}, \qquad f_\theta \in \mathcal{F},$$

where $\mathcal{F}$ is a prescribed class of measurable functions from $\mathbb{R}^{d_1} \times [t_0, T]$ to $\mathbb{R}^{d_1}$. Define the corresponding score class and its population oracle risk as

$$\mathcal{S}_{\mathrm{freeze}}(V_1) := \{s_{V_1,\theta} : f_\theta \in \mathcal{F}\}, \qquad \mathcal{B}_{\mathrm{freeze}}(V_1) := \inf_{s\in\mathcal{S}_{\mathrm{freeze}}(V_1)} \mathcal{L}_2(s),$$

where

$$\mathcal{L}_2(s) := \frac{1}{T-t_0}\int_{t_0}^{T} \mathbb{E}_{X_t\sim p_{2,t}} \|s(X_t,t) - \nabla \log p_{2,t}(X_t)\|_2^2 \,\mathrm{d}t.$$

Here, θ indexes elements of the general function class $\mathcal{F}$, with ReLU networks considered as a special case. To isolate the representation-level obstruction, we define the enlarged structural class

$$\mathcal{S}_{\mathrm{str}}(V_1) := \left\{ s_{V_1,f}(x,t) = \frac{1}{h(t)} V_1 f(V_1^\top x, t) - \frac{1}{h(t)} x : f \text{ measurable} \right\},$$

and define

$$\mathcal{B}_{\mathrm{str}}(V_1) := \inf_{s\in\mathcal{S}_{\mathrm{str}}(V_1)} \mathcal{L}_2(s).$$

Since $\mathcal{S}_{\mathrm{freeze}}(V_1) \subseteq \mathcal{S}_{\mathrm{str}}(V_1)$, we have $\mathcal{B}_{\mathrm{freeze}}(V_1) \geqslant \mathcal{B}_{\mathrm{str}}(V_1)$. Hence, the following lower bound holds regardless of the core architecture. Our target-domain estimation guarantees are conditional on V_1 and remain valid when V_1 is learned from source data independently of the target training sample, by conditioning on the learned projector.

Let $r = \min\{d_1, d_2\}$, and let $0 \leqslant \theta_1 \leqslant \cdots \leqslant \theta_r \leqslant \pi/2$ be the principal angles between $\mathrm{col}(V_1)$ and $\mathrm{col}(A_2)$, defined by

$$\sigma_j(V_1^\top A_2) = \cos\theta_j, \qquad j = 1, \ldots, r.$$

We now establish a lower bound on the structural bias induced by a frozen projection.

Theorem 2. *Suppose Assumption 1 holds for p_2. Let*

$$g_2(y,t) := h(t)\nabla \log p_{2,t}^{\mathrm{LD}}(y) + y.$$

For $Y_t := A_2^\top X_t$, define

$$M_{g,2}(t) := \mathbb{E}[g_2(Y_t,t)g_2(Y_t,t)^\top], \qquad \mu_2(t) := \lambda_{\min}(M_{g,2}(t)).$$

Then,

$$\mathcal{B}_{\mathrm{str}}(V_1) \geqslant \frac{1}{T-t_0}\int_{t_0}^{T}\left[\frac{\mu_2(t)}{h^2(t)}\left(d_2 - \sum_{j=1}^{r}\cos^2\theta_j\right) + \frac{\sigma_2^4\alpha^4(t)}{h^2(t)\widetilde{h}_2(t)}\left(D - d_1 - d_2 + \sum_{j=1}^{r}\cos^2\theta_j\right)\right]\mathrm{d}t.$$

Consequently, the same lower bound holds for $\mathcal{B}_{\mathrm{freeze}}(V_1)$.

The first term captures target signal directions missed by the frozen subspace and vanishes when $\mathrm{col}(A_2) \subseteq \mathrm{col}(V_1)$. The second term captures the residual error arising from ambient noise, reflecting the discrepancy between the fixed orthogonal score in the projected model and the true target score with effective variance $\widetilde{h}_2(t)$. This error may persist even under perfect subspace alignment. Thus, the lower bound identifies an irreducible limitation of the specified projected score class that cannot be eliminated by increasing the expressivity of the core model.

We next show that this structural obstruction also leads to a distribution-level limitation, by exploiting the fixed orthogonal sampling dynamics of the projected score class.

Proposition 3. *Let $P = V_1V_1^\top$, and let $q_{V_1,0}$ be the law of PX_0 for $X_0 \sim p_2$. Write $q_{V_1,t}$ for its forward OU marginal. Assume the target has a finite second moment and that the score risks and relative entropies below are finite. Consider an admissible core f for which the reverse SDE driven by $s_{V_1,f}$ is well posed, initialized from $q_{V_1,T}$, and denote its output at forward time $t_0 > 0$ by $\widehat{p}_{V_1,f,t_0}$. Then*

$$D_{\mathrm{KL}}(p_{2,t_0}\|\widehat{p}_{V_1,f,t_0}) \geqslant D_{\mathrm{KL}}(p_{2,t_0}\|q_{V_1,t_0}) = D_{\mathrm{KL}}(p_{2,T}\|q_{V_1,T}) + \frac{T-t_0}{2}\mathcal{B}_{\mathrm{str}}(V_1) \geqslant \frac{T-t_0}{2}\mathcal{B}_{\mathrm{str}}(V_1).$$

Combined with Theorem 2, Proposition 3 shows that the geometric obstruction to frozen latent reuse also induces an endpoint KL lower bound. This result applies to continuous-time sampling with exact terminal initialization and measures the discrepancy from the smoothed target p_{2,t_0}. Extensions to standard Gaussian initialization and time-discretized sampling require separate analysis. The argument relies on the structure of the projected score class and does not constitute a general converse to the Girsanov upper bound.

3.2 Achievability and Statistical Error under Frozen Reuse

The preceding lower bound identifies regimes in which frozen reuse incurs intrinsic error. We next complement this result by studying what can be achieved when the frozen subspace is sufficiently aligned with the target distribution. To this end, we introduce an explicit frozen-coordinate comparator and derive the corresponding statistical guarantee.

To establish the bounds, we require mild regularity of the target latent distribution and its induced score map.

Assumption 2. *The target latent density $p_{2,z} > 0$ is twice continuously differentiable. Moreover, there exist constants $B_2, C_{1,2}, C_{2,2} > 0$ such that, for all $\|z\|_2 \geqslant B_2$, it holds that*

$$p_{2,z}(z) \leqslant (2\pi)^{-d_2/2}C_{1,2}\exp\{-C_{2,2}\|z\|_2^2/2\}.$$

Assumption 3. *The target core score map $f_2^*(z,t) := \widetilde{h}_2(t)\nabla \log p_{2,t}^{\mathrm{LD}}(z) + z$ is uniformly β_2-Lipschitz in $z \in \mathbb{R}^{d_2}$ for all $t \in [0,T]$.*

These assumptions are standard in theoretical analyses of diffusion models [9, 4]. They ensure that the target latent score is regular enough to be approximated by the frozen-coordinate comparator introduced below. Let

$$B := V_1^\top A_2, \qquad g_2(y,t) := h(t)\nabla \log p_{2,t}^{\mathrm{LD}}(y) + y.$$

We use the explicit frozen-coordinate comparator

$$f_{\mathrm{comp}}(z,t) := Bg_2(B^\dagger z, t),$$

where $B^\dagger$ is the Moore–Penrose pseudoinverse. This comparator reconstructs a target latent coordinate from the frozen coordinate $z = V_1^\top x$ through $B^\dagger z$, applies the target latent score map g_2, and maps it back to the frozen coordinates through B. The induced frozen score model is

$$s_{\mathrm{comp}}(x,t) := \frac{1}{h(t)} V_1 f_{\mathrm{comp}}(V_1^\top x, t) - \frac{1}{h(t)} x,$$

and we denote $\mathcal{U}_{\mathrm{comp}}(V_1) := \mathcal{L}_2(s_{\mathrm{comp}})$. Since $s_{\mathrm{comp}} \in \mathcal{S}_{\mathrm{str}}(V_1)$, we have $\mathcal{B}_{\mathrm{str}}(V_1) \leqslant \mathcal{U}_{\mathrm{comp}}(V_1)$. The following result bounds the comparator risk $\mathcal{U}_{\mathrm{comp}}(V_1)$ in terms of the principal-angle misalignment between the frozen and target subspaces and the target ambient noise.

Proposition 4. *Suppose Assumptions 1, 2, and 3 hold for p_2. Let $\theta_1, \dots, \theta_r$ be the principal angles between $\mathrm{col}(V_1)$ and $\mathrm{col}(A_2)$, where $r = \min\{d_1, d_2\}$. Let $L_g(t)$ be a Lipschitz constant of $g_2(\cdot, t)$, and define*

$$\lambda_{g,2}^{\max}(t) := \lambda_{\max}\big(\mathbb{E}[g_2(Y_t,t)g_2(Y_t,t)^\top]\big), \qquad M_{Y,2}(t) := \mathbb{E}[Y_t Y_t^\top], \qquad Y_t := A_2^\top X_t.$$

If $d_1 \geqslant d_2$ *and* $B = V_1^\top A_2$ *has full column rank, it holds that*

$$\mathcal{U}_{\mathrm{comp}}(V_1) \leqslant \frac{1}{T-t_0}\int_{t_0}^{T} \left[\frac{\lambda_{g,2}^{\max}(t)}{h^2(t)} \sum_{j=1}^{d_2} \sin^2\theta_j + \frac{2\|B\|_{\mathrm{op}}^2 \widetilde{h}_2(t) L_g^2(t)}{h^2(t)} \sum_{j=1}^{d_2} \tan^2\theta_j + \frac{\sigma_2^4 \alpha^4(t)}{h^2(t)\widetilde{h}_2(t)} \left(D + d_1 - d_2 - \sum_{j=1}^{d_2} \cos^2\theta_j \right) \right] \mathrm{d}t.$$

If $d_1 < d_2$ and $B = V_1^\top A_2$ has full row rank, *it holds that*

$$\mathcal{U}_{\mathrm{comp}}(V_1) \leqslant \frac{1}{T-t_0}\int_{t_0}^{T} \left[\frac{\lambda_{g,2}^{\max}(t)}{h^2(t)} \left(d_2 - \sum_{j=1}^{d_1} \cos^2\theta_j \right) + \frac{2\|B\|_{\mathrm{op}}^2 L_g^2(t)}{h^2(t)} \mathrm{Tr}\big((I_{d_2} - B^\dagger B) M_{Y,2}(t)\big) + \frac{2\|B\|_{\mathrm{op}}^2 \widetilde{h}_2(t) L_g^2(t)}{h^2(t)} \sum_{j=1}^{d_1} \tan^2\theta_j + \frac{\sigma_2^4 \alpha^4(t)}{h^2(t)\widetilde{h}_2(t)} \left(D + d_1 - d_2 - \sum_{j=1}^{d_1} \cos^2\theta_j \right) \right] \mathrm{d}t.$$

When $d_1 \geqslant d_2$, the frozen space has sufficient dimension to represent the target subspace, and the main difficulty is stability, as reflected by $\tan^2\theta_j$. When $d_1 < d_2$, the frozen space is dimensionally insufficient, leading to the additional information-loss term $\mathrm{Tr}\left((I_{d_2} - B^\dagger B) M_{Y,2}(t)\right)$ in addition to the output-side residual $d_2 - \sum_{j=1}^{d_1} \cos^2\theta_j$. Thus, frozen reuse can fail either because the reused space is poorly aligned with the target subspace, or because it is too small to contain the target-relevant directions.

We next extend the comparator analysis to a general core function class $\mathcal{F}$ by controlling the approximation error and the statistical error arising from empirical denoising-risk minimization. Define the weighted approximation error

$$\mathcal{A}_{\mathrm{comp}}(f_\theta; V_1) := \frac{1}{T-t_0}\int_{t_0}^{T} \frac{1}{h^2(t)} \mathbb{E}_{X_t \sim p_{2,t}} \|f_\theta(V_1^\top X_t, t) - f_{\mathrm{comp}}(V_1^\top X_t, t)\|_2^2 \, \mathrm{d}t,$$

and let $\mathcal{A}_{\mathcal{F}}(V_1) := \inf_{f_\theta \in \mathcal{F}} \mathcal{A}_{\mathrm{comp}}(f_\theta; V_1)$.

Lemma 5. *For any $\eta > 0$,*

$$\mathcal{B}_{\mathrm{freeze}}(V_1) \leqslant (1+\eta)\mathcal{U}_{\mathrm{comp}}(V_1) + \left(1 + \frac{1}{\eta}\right) \mathcal{A}_{\mathcal{F}}(V_1). \tag{2}$$

Lemma 5 separates the explicit comparator risk from its approximation error within $\mathcal{F}$. We then pass from the frozen oracle risk to finite-sample training. Let $x_1, \dots, x_{N_2}$ be i.i.d. samples from p_2. For any score network s, define the target-domain denoising loss

$$\ell_2(x; s) := \frac{1}{T - t_0} \int_{t_0}^{T} \mathbb{E}_{X_t \sim p_{2,t}(\cdot \mid X_0 = x)} \left[\|\nabla_{X_t} \log \phi_t(X_t \mid x) - s(X_t, t)\|_2^2 \right] \mathrm{d}t,$$

where $\phi_t(\cdot \mid x)$ denotes the Gaussian transition density of X_t conditional on $X_0 = x$. Define

$$\mathcal{R}_2(s) := \mathbb{E}_{X_0 \sim p_2}[\ell_2(X_0; s)], \qquad \widehat{\mathcal{R}}_2(s) := \frac{1}{N_2} \sum_{i=1}^{N_2} \ell_2(x_i; s),$$

and let

$$\widehat{s}_{\mathrm{freeze}} \in \operatorname*{argmin}_{s \in \mathcal{S}_{\mathrm{freeze}}(V_1)} \widehat{\mathcal{R}}_2(s).$$

Although $\widehat{s}_{\mathrm{freeze}}$ is trained by minimizing the empirical denoising risk, our target quantity is the population score error $\mathcal{L}_2$. The denoising score-matching identity implies that there exists a constant E_2, independent of s, such that

$$\mathcal{R}_2(s) = \mathcal{L}_2(s) + E_2.$$

The identity between the denoising objective and score matching follows from the denoising score matching argument of Vincent [32]. Therefore, population minimization of $\mathcal{R}_2$ is equivalent to population minimization of $\mathcal{L}_2$. Moreover, if

$$\mathrm{StatErr}_{N_2} := 2 \sup_{s \in \mathcal{S}_{\mathrm{freeze}}(V_1)} \left| \widehat{\mathcal{R}}_2(s) - \mathcal{R}_2(s) \right|,$$

then the empirical minimizer satisfies the oracle inequality

$$\mathcal{L}_2(\widehat{s}_{\mathrm{freeze}}) \leqslant \mathcal{B}_{\mathrm{freeze}}(V_1) + \mathrm{StatErr}_{N_2}.$$

Combining this inequality with the comparator decomposition in (2) gives the following bound.

Theorem 6. *Suppose Assumptions 1, 2, and 3 hold for p_2. Let V_1 and $\mathcal{F}$ be fixed independently of the target training sample. Assume that $\mathcal{F}$ is a measurable core function class satisfying $\sup_{f \in \mathcal{F}, z, t} \|f(z, t)\|_2 \leqslant K$, with finite covering numbers as defined below and suitable separability conditions ensuring measurability of the empirical suprema. Let $\widehat{s}_{\mathrm{freeze}}$ be an empirical denoising-risk minimizer over $\mathcal{S}_{\mathrm{freeze}}(V_1)$. For $\delta \in (0, 1)$, choose sufficiently large tail constants $C_z, C_\perp > 0$ and set*

$$R_z = C_z(1 + \sigma_2) \left(\sqrt{d_2} + \sqrt{\log \frac{8N_2}{\delta}} \right), \qquad R_\perp = C_\perp \sigma_2 \left(\sqrt{D - d_2} + \sqrt{\log \frac{8N_2}{\delta}} \right).$$

Let $P_2^\perp = I_D - A_2 A_2^\top$, and define

$$\mathcal{T}_2 := \{x : \|A_2^\top x\|_2 \leqslant R_z,\ \|P_2^\perp x\|_2 \leqslant R_\perp\}.$$

For $\iota > 0$, define the covering number

$$\mathcal{N}_2(\iota) := \mathcal{N}(\iota, \{\ell_2(\cdot; s)\mathbf{1}_{\mathcal{T}_2} : s \in \mathcal{S}_{\mathrm{freeze}}(V_1)\}, \|\cdot\|_\infty).$$

Set $C_T := \left(\frac{T}{1 - e^{-T}} \right)^2$ and

$$\mathrm{Est}_{N_2}(V_1; \mathcal{F}, a, \iota, \delta) := C \left[\frac{C_T(K^2 + R_z^2 + R_\perp^2)}{aN_2 t_0 (T - t_0)} \left(\log \mathcal{N}_2(\iota) + \log \frac{8}{\delta} \right) + \iota \right],$$

where $C > 0$ is a sufficiently large constant independent of N_2, a, ι, δ. Then, for any fixed $a \in (0, 1)$, $\eta > 0$, and $\iota > 0$, with probability at least $1 - \delta$,

$$\mathcal{L}_2(\widehat{s}_{\mathrm{freeze}}) \leqslant (1 + a)^2 \left[(1 + \eta) \mathcal{U}_{\mathrm{comp}}(V_1) + \left(1 + \frac{1}{\eta} \right) \mathcal{A}_{\mathcal{F}}(V_1) \right] + \mathrm{Est}_{N_2}(V_1; \mathcal{F}, a, \iota, \delta) + (2a + a^2) E_2,$$

where $E_2 = \mathcal{R}_2(s) - \mathcal{L}_2(s)$ is independent of s and satisfies

$$0 \leqslant E_2 \leqslant \frac{D}{T - t_0} \log \frac{e^T - 1}{e^{t_0} - 1}.$$

If $\widehat{s}_{\mathrm{freeze}}$ is an approximate empirical minimizer with optimization error at most ξ, the bound has the additional term $(1 + a)\xi$.

The bound relies on bounded loss envelopes, tail control, covering entropy, and approximation in the weighted score metric. Regularity assumptions are imposed on the comparator to ensure its approximability, rather than on individual functions in $\mathcal{F}$.

The following corollary specializes Theorem 6 to ReLU networks, yielding an explicit approximation and estimation bound.

Corollary 7. *Let $\mathcal{F} = \mathcal{F}_{\mathrm{ReLU}}^{(d_1)}(L, M, J, K, \kappa, \gamma, \gamma_t)$ be the ReLU network class defined in Appendix A. Under the conditions of Theorem 6, assume additionally that the comparator satisfies the time-regularity condition in Proposition 14, with the network architecture chosen according to that proposition and Lemma 12. For any $\epsilon \in (0,1)$, setting $a = \epsilon^2$ and $\eta = 1$ yields, with probability at least $1-\delta$,*

$$\mathcal{L}_2(\widehat{s}_{\mathrm{freeze}}) \leqslant C \Bigg[\mathcal{U}_{\mathrm{comp}}(V_1) + \frac{C_T(d_1+1)\epsilon^2}{t_0(T-t_0)} + \epsilon^2 E_2 + \frac{C_T(K^2 + R_z^2 + R_\perp^2)}{N_2 t_0 (T-t_0)\epsilon^2} \left(\log \mathcal{N}_2(\iota) + \log \frac{8}{\delta} \right) + \iota \Bigg].$$

An explicit bound on $\log \mathcal{N}_2(\iota)$ is given in Lemma 13.

The bounds separate finite-sample estimation error, function-class approximation error, and geometry-dependent comparator error. Increasing the expressivity of the core model may reduce approximation error at the cost of greater statistical complexity. While the comparator bound characterizes the achievable error, the structural lower bound in Theorem 2, together with its sampling consequence in Proposition 3, establishes an irreducible limitation of frozen latent reuse. In particular, subspace misalignment or insufficient latent dimension can induce errors that persist regardless of the number of target samples or the expressivity of the core model. This motivates learning a shared representation from mixed source–target data to better capture the target-relevant structure.

3.3 Mixed Source-target Training

We study shared representations at the population-oracle level by comparing the frozen source projector V_1 with a shared projector W_k learned from mixed source-target data. We establish mixture-distribution oracle bounds and a target-domain guarantee in Corollary 9, where the shared projector is frozen for target score estimation. Finite-sample estimation of the shared projector is left for future work. To characterize how well a shared representation captures the signal structures of both domains, consider the mixture distribution

$$p_{\mathrm{mix}} := \omega_1 p_1 + \omega_2 p_2,$$

where $\omega_1, \omega_2 > 0$ and $\omega_1 + \omega_2 = 1$. For a k-dimensional projector $W \in \mathbb{R}^{D\times k}$ and weights $c_1, c_2 > 0$, define the weighted signal residual

$$\Gamma_k(W) := \omega_1 c_1 \|P_W^\perp A_1\|_F^2 + \omega_2 c_2 \|P_W^\perp A_2\|_F^2 .$$

Here, c_i scales the contribution of the signal residual from component i. Thus, $\Gamma_k(W)$ measures the weighted amount of source and target signal lost outside $\mathrm{col}(W)$.

Proposition 8. *Let $A_i \in \mathbb{R}^{D\times d_i}$ have orthonormal columns, and constants $c_1, c_2 > 0$. For $W \in \mathbb{R}^{D\times k}$ with $W^\top W = I_k$, let $\Gamma_k(W) = \sum_{i=1}^2 \omega_i c_i \|P_W^\perp A_i\|_F^2$. Set $M_{\mathrm{mix}} := \sum_{i=1}^2 \omega_i c_i A_i A_i^\top$, and let $\lambda_1(M_{\mathrm{mix}}) \geqslant \cdots \geqslant \lambda_D(M_{\mathrm{mix}}) \geqslant 0$. Then, for any $k \in \{\max(d_1, d_2), \ldots, D\}$, it holds that*

$$\inf_{W: W^\top W = I_k} \Gamma_k(W) = \mathrm{Tr}(M_{\mathrm{mix}}) - \sum_{j=1}^k \lambda_j(M_{\mathrm{mix}}) = \sum_{j=k+1}^D \lambda_j(M_{\mathrm{mix}}).$$

The infimum is attained by any W_k whose columns span a top-k eigenspace of M_{mix}. In particular, the optimal residual is nonincreasing in k and vanishes if $\mathrm{span}(A_1) + \mathrm{span}(A_2) \subseteq \mathrm{span}(W_k)$.

Proposition 8 gives a simple geometric reason to retrain a shared latent space. If the frozen source projector V_1 misses target directions, then W_k minimizes the joint source–target signal residual among all k-dimensional projectors. When W_k contains the union of the source and target signal subspaces, this signal-side residual is completely removed.

Corollary 9. *Let $W \in \mathbb{R}^{D\times k}$ have orthonormal columns and be fixed independently of the target score-training sample. Define*

$$B_W := W^\top A_2, \qquad f_{\mathrm{comp},W}(z,t) := B_W g_2(B_W^\dagger z, t),$$
$$s_{\mathrm{comp},W}(x,t) := \frac{W f_{\mathrm{comp},W}(W^\top x, t) - x}{h(t)},$$

and $\mathcal{U}_{\mathrm{comp}}(W) := \mathcal{L}_2(s_{\mathrm{comp},W})$. Under the conditions of Theorem 6, the target-domain estimation bound holds with (V_1, d_1) replaced by (W, k) and a core class on $\mathbb{R}^k$. The comparator bound in Proposition 4 also applies under the corresponding rank condition, with target signal residual

$$\Delta_2(W) := \|P_W^\perp A_2\|_F^2 = d_2 - r + \sum_{j=1}^{r} \sin^2\vartheta_j, \qquad r = \min\{k, d_2\},$$

where ϑ_j are the principal angles between $\mathrm{col}(W)$ and $\mathrm{col}(A_2)$. If $\mathrm{col}(A_2) \subseteq \mathrm{col}(W)$, the target signal-mismatch and pseudoinverse reconstruction terms vanish, leaving only ambient-noise, approximation, and estimation errors.

The result applies in particular to the population oracle W_k. For a shared projector $\widehat{W}_k$ estimated from target data, the guarantee remains valid under independent sample splitting, while reusing the same samples requires further analysis. Moreover, minimizing the joint signal residual does not necessarily improve the target-domain risk at a fixed rank.

We next connect the geometric improvement of the mixed projector to the mixed score-oracle error. For a fixed orthonormal projector $U \in \mathbb{R}^{D\times m}$, with $U = V_1$ or $U = W_k$, define

$$\mathcal{B}_{\mathrm{mix}}(U) := \inf_{s\in\mathcal{S}_{\mathrm{mix}}(U)} \mathcal{L}_{\mathrm{mix}}(s),$$

where

$$\mathcal{S}_{\mathrm{mix}}(U) := \left\{ s_{U,\theta}(x,t) = \frac{1}{h(t)} U f_\theta(U^\top x, t) - \frac{1}{h(t)} x : f_\theta \in \mathcal{F}_{\mathrm{ReLU}}^{(m)} \right\}$$

is the score class using the projected coordinates $U^\top x$, and $\mathcal{L}_{\mathrm{mix}}$ denotes the integrated score risk under p_{mix}:

$$\mathcal{L}_{\mathrm{mix}}(s) := \frac{1}{T - t_0} \int_{t_0}^{T} \mathbb{E}_{X_t\sim p_{\mathrm{mix},t}} \left[\|s(X_t,t) - \nabla \log p_{\mathrm{mix},t}(X_t)\|_2^2 \right] \mathrm{d}t.$$

The leading projector-dependent quantity is

$$\Gamma(U) := \sum_{i=1}^{2} \omega_i \bar{c}_i \|P_U^\perp A_i\|_F^2, \quad \text{with} \quad \bar{c}_i := \frac{1}{T - t_0} \int_{t_0}^{T} \frac{\lambda_{g,i}^{\max}(t)}{h^2(t)} \,\mathrm{d}t,$$

where

$$\lambda_{g,i}^{\max}(t) := \lambda_{\max}(M_{g,i}(t)), \qquad M_{g,i}(t) := \mathbb{E}_{X_t\sim p_{i,t}} \left[g_i(A_i^\top X_t, t) g_i(A_i^\top X_t, t)^\top \right],$$

with

$$g_i(y,t) := h(t)\nabla \log p_{i,t}^{\mathrm{LD}}(y) + y.$$

Thus, $\Gamma(U)$ measures the weighted source–target signal residual outside the shared latent space. With the natural weights $c_i = \bar{c}_i$, the next result bounds the mixed oracle risk and uses W_k to minimize the leading signal-residual term.

Theorem 10. *Suppose Assumptions 1, 2, and 3 hold for both domains, with possibly different constants, and that all comparator and approximation risks appearing below are finite. The mixed comparator, reconstruction risks, and posterior-compression term are defined in Appendix C.3. Let W_k be the optimizer in Proposition 8 with $c_i = \bar{c}_i$. Then, for any $U \in \{V_1, W_k\}$ and $\eta > 0$,*

$$\mathcal{B}_{\mathrm{mix}}(U) \leqslant (1+\eta) \left[\Gamma(U) + \sum_{i=1}^{2} \omega_i \bar{n}_i \,\mathrm{Tr}(P_U^\perp P_i^\perp) + 2\sum_{i=1}^{2} \omega_i \mathfrak{R}_i(U) + 2\mathfrak{P}(U) \right]$$
$$+ \left(1 + \frac{1}{\eta}\right) \inf_{f_\theta\in\mathcal{F}_{\mathrm{ReLU}}^{(m)}} \mathcal{A}_{\mathrm{mix}}(f_\theta; U).$$

Theorem 10 separates the mixed oracle risk into a leading signal-residual term, an ambient-noise term, reconstruction $\mathfrak{R}_i(U)$ and posterior-compression terms $\mathfrak{P}(U)$. The auxiliary quantities $\mathfrak{R}_i(U)$ and $\mathfrak{P}(U)$ capture, respectively, the componentwise reconstruction error from the projected coordinates $U^\top x$ and the loss of mixture-component information after projection. Their explicit forms, together with the full mixed oracle decomposition, are given in Appendix C.3.

The leading projector-dependent term in the upper bound of Theorem 10 is $\Gamma(U)$. Since $M_{\mathrm{mix}} = \sum_{i=1}^{2} \omega_i \bar{c}_i A_i A_i^\top$, Proposition 8 implies that the optimal k-dimensional mixed projector W_k satisfies $\Gamma(W_k) = \sum_{j=k+1}^{D} \lambda_j(M_{\mathrm{mix}})$. Thus, compared with frozen reuse, the shared projector W_k reduces the leading signal residual by

$$\Gamma(V_1) - \Gamma(W_k) = \Gamma(V_1) - \sum_{j=k+1}^{D} \lambda_j(M_{\mathrm{mix}}).$$

Thus, mixed training improves the oracle upper bound when this reduction dominates the corresponding changes in the ambient-noise, reconstruction, posterior-compression, and ReLU approximation terms. The following corollary illustrates the most favorable case, where the mixed projector contains both source and target signal subspaces.

Corollary 11. *Suppose W_k satisfies $\mathrm{span}(A_1) + \mathrm{span}(A_2) \subseteq \mathrm{span}(W_k)$. Then, for any $\eta > 0$,*

$$\mathcal{B}_{\mathrm{mix}}(W_k) \leqslant (1+\eta)\left[\sum_{i=1}^{2} \omega_i \bar{n}_i (D-k) + 2\sum_{i=1}^{2} \omega_i \bar{n}_i (k-d_i) + 2\mathfrak{P}(W_k)\right] + \left(1 + \frac{1}{\eta}\right) \inf_{f_\theta \in \mathcal{F}_{\mathrm{ReLU}}^{(k)}} \mathcal{A}_{\mathrm{mix}}(f_\theta; W_k)\,,$$

where $\bar{n}_i := \frac{1}{T-t_0}\int_{t_0}^{T} \frac{\alpha^4(t)\sigma_i^4}{h^2(t)\tilde{h}_i(t)}\,\mathrm{d}t$. In particular, if $V_1 = A_1$, then

$$\Gamma(V_1) = \omega_2 \bar{c}_2 \|P_{A_1}^{\perp} A_2\|_F^2, \qquad \Gamma(W_k) = 0.$$

Thus, in the containment regime, mixed training removes the target signal-mismatch term that remains under source-aligned reuse.

The posterior-compression term $\mathfrak{P}(W_k)$ does not necessarily vanish under the containment condition alone. It vanishes under the stronger condition that the projected coordinate $W_k^\top X_t$ is sufficient for identifying the mixture component. For instance, this holds when W_k contains both signal subspaces and the two mixture components have identical ambient-noise distributions. In this case, mixed training eliminates the source–target signal-mismatch term, leaving only ambient-noise, reconstruction, and ReLU approximation contributions.

4 Discussion

Our analysis considers a noisy linear low-dimensional model, in which the source and target distributions concentrate near potentially different subspaces. This setting makes the geometric limitations of latent reuse explicit through principal angles and ambient-noise effects, but does not capture nonlinear latent structures commonly encountered in practice. A natural extension is to consider distributions concentrated near smooth manifolds, as studied in [3, 37, 31]. Another direction is to relax the light-tailed latent and Gaussian ambient-noise assumptions and investigate the robustness of latent reuse under heavy-tailed distributions and more general distribution shifts [40, 20].

Another limitation is that our mixed-training analysis is carried out at the oracle level, treating the shared projector as fixed once chosen. This isolates the representation-level mechanism, but does not account for the finite-sample cost of learning the projector from mixed data. A more complete theory would combine representation learning error, score-estimation error, and sampling error in a single end-to-end guarantee. It would also be interesting to study adaptive mixtures, where the relative amount of source and target data is chosen according to the degree of distribution shift.

Acknowledgments and Disclosure of Funding

This work was partially supported by the City University of Hong Kong Startup Fund 9610707 and the Hong Kong Research Grants Council through ECS Grant No. 21306325 and GRF Grant No. 11309626.

References

[1] Beatrice Achilli, Enrico Ventura, Gianluigi Silvestri, Bao Pham, Gabriel Raya, Dmitry Krotov, Carlo Lucibello, and Luca Ambrogioni. Losing dimensions: Geometric memorization in generative diffusion. *arXiv preprint arXiv:2410.08727*, 2024.

[2] Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

[3] Iskander Azangulov, George Deligiannidis, and Judith Rousseau. Convergence of diffusion models under the manifold hypothesis in high-dimensions. *arXiv preprint arXiv:2409.18804*, 2024.

[4] Adam Block, Youssef Mroueh, and Alexander Rakhlin. Generative modeling with denoising auto-encoders and langevin sampling. *arXiv preprint arXiv:2002.00107*, 2020.

[5] Patrick Cattiaux, Giovanni Conforti, Ivan Gentil, and Christian Léonard. Time reversal of diffusion processes under a finite entropy condition. In *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, volume 59, pages 1844–1881. Institut Henri Poincaré, 2023.

[6] Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pages 4735–4763. PMLR, 2023.

[7] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

[8] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024.

[9] Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. *Proceedings of Machine Learning Research*, 202:5327–5350, 2023. ISSN 2640-3498.

[10] Minshuo Chen, Renyuan Xu, Yumin Xu, and Ruixun Zhang. Diffusion factor models: Generating high-dimensional returns with factor structure. *arXiv preprint arXiv:2504.06566*, 2025.

[11] Ziheng Cheng, Tianyu Xie, Shiyue Zhang, and Cheng Zhang. Provable sample-efficient transfer learning conditional diffusion models via representation learning. *arXiv preprint arXiv:2502.04491*, 2025.

[12] Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2208.05314*, 2022.

[13] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7323–7334, 2023.

[14] Jerry Yao-Chieh Hu, Weimin Wu, Zhuoru Li, Sophia Pi, Zhao Song, and Han Liu. On statistical rates and provably efficient criteria of latent diffusion transformers (dits). *Advances in Neural Information Processing Systems*, 37:31562–31628, 2024.

[15] Zhihan Huang, Yuting Wei, and Yuxin Chen. Denoising diffusion probabilistic models are optimally adaptive to unknown low dimensionality. *Mathematics of Operations Research*, 2026.

[16] Julian Kleutgens, Claudio Battiloro, Lingkai Kong, Benjamin Grewe, Francesca Dominici, and Mauricio Tec. Guided transfer learning for discrete diffusion models. *arXiv preprint arXiv:2512.10877*, 2025.

[17] Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning of latent diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18262–18272, 2025.

[18] Gen Li and Yuling Yan. Adapting to unknown low-dimensional structures in score-based diffusion models. *Advances in Neural Information Processing Systems*, 37:126297–126331, 2024.

[19] Jiadong Liang, Zhihan Huang, and Yuxin Chen. Low-dimensional adaptation of diffusion models: Convergence in total variation. *arXiv preprint arXiv:2501.12982*, 2025.

[20] Wei Luo, Neil K Chada, Shijie Zhang, and Lu Yu. Robustness of diffusion models under distribution shift. *arXiv preprint arXiv:2609.27546*, 2026.

[21] Taehong Moon, Moonseok Choi, Gayoung Lee, Jung-Woo Ha, and Juho Lee. Fine-tuning diffusion models with limited data. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022.

[22] Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, pages 26517–26582. PMLR, 2023.

[23] Yidong Ouyang, Liyan Xie, Hongyuan Zha, and Guang Cheng. Transfer learning for diffusion models. *Advances in Neural Information Processing Systems*, 37:136962–136989, 2024.

[24] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[25] Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. *arXiv preprint arXiv:2104.08894*, 2021.

[26] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pages 8599–8608. PMLR, 2021.

[27] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[28] 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*, pages 10684–10695, 2022.

[29] 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.

[30] Zixing Song, Ziqiao Meng, and José Miguel Hernández-Lobato. Domain-adapted diffusion model for protac linker design through the lens of density ratio in chemical space. In *Forty-second International Conference on Machine Learning*, 2025.

[31] Rong Tang and Yun Yang. Adaptivity of diffusion models to manifold structures. In *International Conference on Artificial Intelligence and Statistics*, pages 1648–1656. PMLR, 2024.

[32] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

[33] Peng Wang, Huijie Zhang, Zekai Zhang, Siyi Chen, Yi Ma, and Qing Qu. Diffusion models learn low-dimensional distributions via subspace clustering. In *2025 IEEE 10th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 211–215. IEEE, 2025.

[34] Pingyu Wu, Kai Zhu, Yu Liu, Liming Zhao, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Improved video vae for latent video diffusion model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18124–18133, 2025.

[35] Enze Xie, Lewei Yao, Han Shi, Zhili Liu, Daquan Zhou, Zhaoqiang Liu, Jiawei Li, and Zhenguo Li. Difffit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4230–4239, 2023.

[36] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923*, 2022.

[37] Konstantin Yakovlev and Nikita Puchkin. Generalization error bound for denoising score matching under relaxed manifold assumption. *arXiv preprint arXiv:2502.13662*, 2025.

[38] Ruofeng Yang, Bo Jiang, Cheng Chen, Ruinan Jin, Baoxiang Wang, and Shuai Li. Few-shot diffusion models escape the curse of dimensionality. *Advances in Neural Information Processing Systems*, 37:68528–68558, 2024.

[39] Yifeng Yu and Lu Yu. Advancing wasserstein convergence analysis of score-based models: Insights from discretization and second-order acceleration. *Advances in Neural Information Processing Systems*, 38:138411–138465, 2026.

[40] Yifeng Yu and Lu Yu. Diffusion models with heavy-tailed targets: Score estimation and sampling guarantees. *arXiv preprint arXiv:2601.06715*, 2026.

A Additional Key Definitions

In this work, we consider the ReLU network $\mathcal{F}_{\mathrm{ReLU}}^{(d_1)}(L, M, J, K, \kappa, \gamma, \gamma_t)$ defined via

$$\mathcal{F}_{\mathrm{ReLU}}^{(d_1)}(L, M, J, K, \kappa, \gamma, \gamma_t) = \Bigg\{ f(z,t) = W_L\sigma(\cdots\sigma(W_1[z^\top, t]^\top + b_1)\cdots) + b_L : \\ \text{network width bounded by } M,\ \sup_{z,t}\|f(z,t)\|_2 \leqslant K, \\ \max\{\|b_i\|_\infty, \|W_i\|_\infty\} \leqslant \kappa \text{ for } i = 1, \cdots, L, \\ \sum_{i=1}^{L}(\|W_i\|_0 + \|b_i\|_0) \leqslant J, \\ \|f(z_1,t) - f(z_2,t)\|_2 \leqslant \gamma\|z_1 - z_2\|_2 \text{ for any } t \in [0,T], \\ \|f(z,t_1) - f(z,t_2)\|_2 \leqslant \gamma_t|t_1 - t_2| \text{ for any } z \Bigg\}.$$

B Proof of Auxiliary Lemmas

Proof of Lemma 1. The proof follows the exact orthogonal decomposition framework as Lemma 1 in [9], with one crucial algebraic modification regarding the transition kernel.

Under our Assumption 1, the initial data incorporates ambient noise $X_0 = A_i z + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2 I_D)$. By the convolution of independent Gaussian distributions, the conditional transition kernel of the forward diffusion process becomes $X_t|z \sim \mathcal{N}(\alpha(t)A_i z, \tilde{h}_i(t)I_D)$, where the effective variance is strictly $\tilde{h}_i(t) = \alpha^2(t)\sigma_i^2 + h(t)$.

Substituting this effective variance $\tilde{h}_i(t)$ for the pure diffusion variance $h(t)$ in the spatial Pythagorean decomposition (i.e., $\|x - \alpha(t)A_i z\|_2^2 = \|A_i^\top x - \alpha(t)z\|_2^2 + \|(I_D - A_i A_i^\top)x\|_2^2$) directly yields the stated factorization of the marginal density and the corresponding score function.

□

Proof of Lemma 5. For any $f_\theta \in \mathcal{F}$,

$$s_{V_1,\theta}(x,t) - \nabla \log p_{2,t}(x) = s_{\mathrm{comp}}(x,t) - \nabla \log p_{2,t}(x) + \frac{V_1}{h(t)}\big[f_\theta(V_1^\top x, t) - f_{\mathrm{comp}}(V_1^\top x, t)\big].$$

Apply $\|u+v\|_2^2 \leqslant (1+\eta)\|u\|_2^2 + (1+1/\eta)\|v\|_2^2$, use $V_1^\top V_1 = I_{d_1}$, and integrate under the target forward marginals. This gives

$$\mathcal{L}_2(s_{V_1,\theta}) \leqslant (1+\eta)\mathcal{U}_{\mathrm{comp}}(V_1) + \left(1 + \frac{1}{\eta}\right)\mathcal{A}_{\mathrm{comp}}(f_\theta; V_1).$$

Taking the infimum over $f_\theta \in \mathcal{F}$ proves the result. □

Lemma 12 (Explicit compact-domain ReLU approximation)**.** *Let $R > 0$, $0 < t_0 < T$, and let $F : \mathbb{R}^m \times [t_0, T] \to \mathbb{R}^m$ satisfy*

$$\|F(z,t) - F(z',t')\|_2 \leqslant L_z\|z - z'\|_2 + L_t|t - t'|, \qquad \sup_{[-R,R]^m \times [t_0,T]}\|F(z,t)\|_2 \leqslant K_0.$$

For $K_0 > 0$ and $\epsilon > 0$, set

$$p = m + 1, \qquad \bar{\epsilon} = \min\{\epsilon, K_0/\sqrt{m}\}, \\ \Lambda = 2R\sqrt{m}L_z + (T - t_0)L_t, \\ q = \max\{1, \lceil 2\Lambda/\bar{\epsilon}\rceil\}, \qquad n_g = (q+1)^p.$$

There exists a ReLU network f_θ such that

$$\sup_{[-R,R]^m \times [t_0,T]}\|f_\theta - F\|_\infty \leqslant \epsilon, \qquad \sup_{z \in \mathbb{R}^m, t \in [0,T]}\|f_\theta(z,t)\|_2 \leqslant 2K_0.$$

It can be included in the class $\mathcal{F}^{(m)}_{\mathrm{ReLU}}(L, M, J, K, \kappa, \gamma, \gamma_t)$ *with*

$$L \leqslant C\log(2pn_g), \quad M \leqslant Cmpn_g, \qquad J \leqslant Cmpn_g\log(2pn_g),$$
$$K = 2K_0, \qquad \kappa \leqslant C\left(1 + K_0 + \Lambda + R^{-1} + \frac{1+t_0}{T-t_0}\right),$$
$$\gamma = \frac{\sqrt{m}\Lambda}{2R}, \qquad \gamma_t = \frac{\sqrt{m}\Lambda}{T-t_0}.$$

Here C *is an absolute constant. If* $K_0 = 0$*, the zero network suffices.*

Proof. Normalize the compact domain to $[0,1]^p$, and write the normalized map as $\widetilde{F}$. Each coordinate of $\widetilde{F}$ is Λ-Lipschitz in the sup norm. On the grid $\mathcal{G} = \{0, 1/q, \ldots, 1\}^p$, define

$$g_j(u) := \max_{v\in\mathcal{G}} \left\{\widetilde{F}_j(v) - \Lambda\|u - v\|_\infty\right\}, \qquad j = 1, \ldots, m.$$

Lipschitz continuity gives $g_j(u) \leqslant \widetilde{F}_j(u)$. Taking a grid point at sup-norm distance at most $1/q$ gives $0 \leqslant \widetilde{F}_j(u) - g_j(u) \leqslant 2\Lambda/q \leqslant \bar{\epsilon}$. Thus $\|g(u)\|_2 \leqslant K_0 + \sqrt{m}\bar{\epsilon} \leqslant 2K_0$. Before evaluating g, clamp each normalized input coordinate to $[0,1]$. This gives the global output bound without changing the approximation on the compact domain. The identities $|a| = \mathrm{ReLU}(a) + \mathrm{ReLU}(-a)$, $\max(a,b) = \mathrm{ReLU}(a-b) + b$, and $\mathrm{clip}_{[0,1]}(a) = \mathrm{ReLU}(a) - \mathrm{ReLU}(a-1)$ give an exact ReLU implementation. Balanced max trees, with signed intermediate quantities represented by positive and negative channels, give the displayed conservative depth, width, sparsity, and parameter bounds. Each g_j is Λ-Lipschitz in the normalized sup norm; undoing the input scaling gives the stated spatial and time Lipschitz bounds. □

Lemma 13 (Covering entropy of the truncated denoising loss)**.** *Take a bounded ReLU class with latent dimension* m*, depth* L*, width at most* M*, sparsity at most* $J \geqslant 1$*, parameter bound* κ*, and global output bound* K*. Put*

$$R_* := \sqrt{R_z^2 + R_\perp^2}, \qquad I_h := \frac{1}{T-t_0}\int_{t_0}^{T}\frac{\mathrm{d}t}{h(t)^2}, \qquad P_{\mathrm{par}} := L(M+m+2)^2,$$

$$C_{\mathrm{par}} := (L+1)\big[2(M+m+2)(\kappa+1)\big]^{L+1}, \qquad H_{\mathrm{loss}} := 2I_h(K+R_*)C_{\mathrm{par}}(1+R_*+\sqrt{m}+T).$$

For the truncated denoising loss class in Theorem 6,

$$\log\mathcal{N}_2(\iota) \leqslant CJ\log\left(2 + \frac{C(P_{\mathrm{par}}+1)(\kappa+1)H_{\mathrm{loss}}}{\iota}\right).$$

Proof. For two networks with a fixed parameter support and parameter sup-norm distance u, layerwise perturbation estimates give

$$\|f_\theta(z,t) - f_{\theta'}(z,t)\|_2 \leqslant C_{\mathrm{par}}(1 + \|z\|_2 + T)u.$$

The denoising residual equals $(\alpha(t)x - Vf_\theta(V^\top X_t, t))/h(t)$. On the truncation region, $\|x\|_2 \leqslant R_*$, and $\mathbb{E}[\|V^\top X_t\|_2 \mid X_0 = x] \leqslant R_* + \sqrt{m}$. Since both cores are bounded by K, the difference between their integrated denoising losses is at most $H_{\mathrm{loss}}u$, uniformly on the truncation region. There are at most $(P_{\mathrm{par}}+1)^{CJ}$ supports with at most J nonzero parameters. A parameter grid of resolution proportional to ι/H_{loss}, taking a representative from each nonempty cell of the constrained class, gives the claimed covering bound.

□

C Proof of Main Results

C.1 Proof of Section 3.1

Proof of Theorem 2. Fix $t \in [t_0, T]$, and define

$$P_2 := A_2A_2^\top, \quad P_2^\perp := I_D - P_2, \quad \rho_2(t) := \frac{\alpha^2(t)\sigma_2^2}{\widetilde{h}_2(t)}.$$

$$G_{2,t}(x) := h(t)\nabla \log p_{2,t}(x) + x = A_2 g_2(A_2^\top x, t) + \rho_2(t) P_2^\perp x.$$

By the definition of the frozen-projection class, for any measurable f,

$$s_{V_1,f}(X_t, t) - \nabla \log p_{2,t}(X_t) = \frac{1}{h(t)} \left[V_1 f(V_1^\top X_t, t) - G_{2,t}(X_t) \right].$$

Therefore,

$$b_{\mathrm{str}}(t; V_1) = \frac{1}{h^2(t)} \inf_{f_t \text{ measurable}} \mathbb{E}\left[\left\| V_1 f_t(V_1^\top X_t) - G_{2,t}(X_t) \right\|_2^2 \right].$$

Since $V_1 f_t(V_1^\top X_t) \in \mathrm{col}(V_1)$, we decompose

$$G_{2,t}(X_t) = P_{V_1} G_{2,t}(X_t) + P_{V_1}^\perp G_{2,t}(X_t).$$

Then $V_1 f_t(V_1^\top X_t) - P_{V_1} G_{2,t}(X_t) \in \mathrm{col}(V_1)$, whereas $P_{V_1}^\perp G_{2,t}(X_t) \in \mathrm{col}(V_1)^\perp$. Hence the two terms are orthogonal and

$$\left\| V_1 f_t(V_1^\top X_t) - G_{2,t}(X_t) \right\|_2^2 = \left\| V_1 f_t(V_1^\top X_t) - P_{V_1} G_{2,t}(X_t) \right\|_2^2 + \left\| P_{V_1}^\perp G_{2,t}(X_t) \right\|_2^2.$$

Taking the infimum over f_t, the first term is nonnegative. Therefore,

$$b_{\mathrm{str}}(t; V_1) \geqslant \frac{1}{h^2(t)} \mathbb{E}\left[\left\| P_{V_1}^\perp G_{2,t}(X_t) \right\|_2^2 \right]. \tag{3}$$

Recall that

$$G_{2,t}(X_t) = A_2 g_2(A_2^\top X_t, t) + \rho_2(t) P_2^\perp X_t.$$

Let

$$Y_t := A_2^\top X_t, \qquad U_t := P_2^\perp X_t.$$

Then

$$P_{V_1}^\perp G_{2,t}(X_t) = P_{V_1}^\perp A_2 g_2(Y_t, t) + \rho_2(t) P_{V_1}^\perp U_t.$$

By the same orthogonal factorization argument used in the noisy score decomposition lemma, $Y_t = A_2^\top X_t$ and $U_t = P_2^\perp X_t$ are independent under $p_{2,t}$. Since $g_2(Y_t, t)$ is measurable with respect to Y_t, it is independent of U_t. Note that $U_t \sim \mathcal{N}(0, \tilde{h}_2(t) P_2^\perp)$, and $\mathbb{E}[U_t] = 0$. Therefore, the cross term vanishes:

$$\mathbb{E}\left[g_2(Y_t, t)^\top A_2^\top P_{V_1}^\perp U_t \right] = \mathbb{E}[g_2(Y_t, t)]^\top A_2^\top P_{V_1}^\perp \mathbb{E}[U_t] = 0.$$

Consequently,

$$\mathbb{E}\left[\left\| P_{V_1}^\perp G_{2,t}(X_t) \right\|_2^2 \right] = \mathbb{E}\left[\left\| P_{V_1}^\perp A_2 g_2(Y_t, t) \right\|_2^2 \right] + \rho_2^2(t) \mathbb{E}\left[\left\| P_{V_1}^\perp U_t \right\|_2^2 \right]. \tag{4}$$

We first lower bound the first term. Since

$$M_{g,2}(t) = \mathbb{E}\left[g_2(Y_t, t) g_2(Y_t, t)^\top \right],$$

we have

$$\mathbb{E}\left[\left\| P_{V_1}^\perp A_2 g_2(Y_t, t) \right\|_2^2 \right] = \mathrm{Tr}\left(A_2^\top P_{V_1}^\perp A_2 M_{g,2}(t) \right).$$

The matrix $A_2^\top P_{V_1}^\perp A_2$ is positive semidefinite. Since $M_{g,2}(t) \succeq \mu_2(t) I_{d_2}$, we obtain

$$\mathrm{Tr}\left(A_2^\top P_{V_1}^\perp A_2 M_{g,2}(t) \right) \geqslant \mu_2(t) \,\mathrm{Tr}\left(A_2^\top P_{V_1}^\perp A_2 \right) = \mu_2(t) \| P_{V_1}^\perp A_2 \|_F^2.$$

Thus,

$$\mathbb{E}\left[\left\| P_{V_1}^\perp A_2 g_2(Y_t, t) \right\|_2^2 \right] \geqslant \mu_2(t) \| P_{V_1}^\perp A_2 \|_F^2. \tag{5}$$

Next, since $\mathbb{E}[U_t U_t^\top] = \tilde{h}_2(t) P_2^\perp$,

$$\mathbb{E}\left[\left\| P_{V_1}^\perp U_t \right\|_2^2 \right] = \mathrm{Tr}\left(P_{V_1}^\perp \mathbb{E}[U_t U_t^\top] P_{V_1}^\perp \right) = \tilde{h}_2(t) \,\mathrm{Tr}\left(P_{V_1}^\perp P_2^\perp \right).$$

Using

$$\rho_2(t) = \frac{\alpha^2(t)\sigma_2^2}{\tilde{h}_2(t)},$$

we get

$$\rho_2^2(t)\mathbb{E}\left[\left\|P_{V_1}^{\perp}U_t\right\|_2^2\right] = \frac{\alpha^4(t)\sigma_2^4}{\tilde{h}_2(t)}\operatorname{Tr}\left(P_{V_1}^{\perp}P_2^{\perp}\right). \tag{6}$$

Combining (3), (4), (5), and (6), we obtain

$$b_{\mathrm{str}}(t;V_1) \geqslant \frac{\mu_2(t)}{h^2(t)}\|P_{V_1}^{\perp}A_2\|_F^2 + \frac{\sigma_2^4\alpha^4(t)}{h^2(t)\tilde{h}_2(t)}\operatorname{Tr}\left(P_{V_1}^{\perp}P_2^{\perp}\right).$$

For any $s_{V_1,f} \in \mathcal{S}_{\mathrm{str}}(V_1)$, we have

$$r_{2,t}(s_{V_1,f}) \geqslant b_{\mathrm{str}}(t;V_1)\,.$$

Therefore,

$$\begin{aligned}
\mathcal{B}_{\mathrm{str}}(V_1) &= \inf_{f \text{ measurable}} \mathcal{L}_2(s_{V_1,f}) \\
&= \inf_{s_{V_1,f}\in\mathcal{S}_{\mathrm{str}}(V_1)} \frac{1}{T-t_0}\int_{t_0}^{T} r_{2,t}(s_{V_1,f})\,\mathrm{d}t \\
&\geqslant \frac{1}{T-t_0}\int_{t_0}^{T} b_{\mathrm{str}}(t;V_1)\,\mathrm{d}t\,.
\end{aligned}$$

It remains to express the two geometric quantities using principal angles. By the definition of principal angles,

$$\|V_1^{\top}A_2\|_F^2 = \sum_{j=1}^{r}\cos^2\theta_j.$$

For the first term,

$$\|P_{V_1}^{\perp}A_2\|_F^2 = \operatorname{Tr}\left(A_2^{\top}P_{V_1}^{\perp}A_2\right) = \operatorname{Tr}(A_2^{\top}A_2) - \operatorname{Tr}(A_2^{\top}P_{V_1}A_2) = d_2 - \|V_1^{\top}A_2\|_F^2 = d_2 - \sum_{j=1}^{r}\cos^2\theta_j.$$

For the second term,

$$\operatorname{Tr}\left(P_{V_1}^{\perp}P_2^{\perp}\right) = \operatorname{Tr}\left((I_D - P_{V_1})(I_D - P_2)\right) = D - d_1 - d_2 + \operatorname{Tr}(P_{V_1}P_2) = D - d_1 - d_2 + \|V_1^{\top}A_2\|_F^2.$$

Therefore,

$$\operatorname{Tr}\left(P_{V_1}^{\perp}P_2^{\perp}\right) = D - d_1 - d_2 + \sum_{j=1}^{r}\cos^2\theta_j.$$

Substituting these two identities into the integrated lower bound proves the principal-angle form.

□

Proof of Proposition 3. Let $V_{\perp}$ complete V_1 to an orthogonal basis, and write $z = V_1^{\top}x$, $u = V_{\perp}^{\top}x$. Let r_t denote the density of $V_1^{\top}X_t$ under the target OU process. By construction,

$$q_{V_1,t}(x) = r_t(z)\varphi_{h(t)I_{D-d_1}}(u), \qquad \nabla\log q_{V_1,t}(x) = V_1\nabla\log r_t(z) - \frac{V_{\perp}u}{h(t)}.$$

The conditional-score identity gives

$$\mathbb{E}[V_1^{\top}\nabla\log p_{2,t}(X_t) \mid V_1^{\top}X_t = z] = \nabla\log r_t(z).$$

Consequently, the pointwise population minimizer in the projected class is $\nabla \log q_{V_1,t}$, corresponding to the measurable core $f^*(z,t) = h(t)\nabla \log r_t(z) + z$. Conditional least squares and orthogonal decomposition therefore yield

$$\mathcal{B}_{\mathrm{str}}(V_1) = \frac{1}{T-t_0}\int_{t_0}^{T} \mathbb{E}_{p_{2,t}} \|\nabla \log p_{2,t} - \nabla \log q_{V_1,t}\|_2^2 \,\mathrm{d}t.$$

Both $p_{2,t}$ and $q_{V_1,t}$ follow the same forward OU equation. The relative entropy dissipation identity is

$$\frac{\mathrm{d}}{\mathrm{d}t} D_{\mathrm{KL}}(p_{2,t}\|q_{V_1,t}) = -\frac{1}{2}\mathbb{E}_{p_{2,t}} \|\nabla \log p_{2,t} - \nabla \log q_{V_1,t}\|_2^2.$$

It follows by differentiating the relative entropy, substituting the two Fokker–Planck equations, and integrating by parts. At strictly positive times, Gaussian smoothing and the finite-second-moment assumption permit cutoff approximation of this calculation. Integrating from t_0 to T proves the stated equality. It remains to compare an arbitrary admissible sampler with q_{V_1,t_0}. The reverse drift in the orthogonal coordinates is $u/2 - u/h(t)$, independent of the projected coordinates and of the core. Starting from the product law $q_{V_1,T}$, the two coordinate processes are independent and the orthogonal output remains $\mathcal{N}(0, h(t_0)I_{D-d_1})$. Hence the sampler output has the form $\widehat{r}_{t_0}(z)\varphi_{h(t_0)I_{D-d_1}}(u)$. The KL chain rule gives

$$D_{\mathrm{KL}}(p_{2,t_0}\|\widehat{p}_{V_1,f,t_0}) = D_{\mathrm{KL}}(p_{2,t_0}\|q_{V_1,t_0}) + D_{\mathrm{KL}}(r_{t_0}\|\widehat{r}_{t_0}) \geqslant D_{\mathrm{KL}}(p_{2,t_0}\|q_{V_1,t_0}).$$

Combining the two identities proves the proposition.

□

C.2 Proof of Section 3.2

Proof of Proposition 4. Fix $t \in [t_0, T]$. By the definition of $G_{2,t}$, for any measurable $f : \mathbb{R}^{d_1} \times [t_0, T] \to \mathbb{R}^{d_1}$,

$$s_{V_1,f}(X_t,t) - \nabla \log p_{2,t}(X_t) = \frac{1}{h(t)}\left[V_1 f(V_1^\top X_t, t) - G_{2,t}(X_t)\right].$$

Therefore,

$$r_{2,t}(s_{V_1,f}) = \frac{1}{h^2(t)}\mathbb{E}\left[\left\|V_1 f(V_1^\top X_t,t) - G_{2,t}(X_t)\right\|_2^2\right].$$

Decompose $G_{2,t}(X_t)$ into its components in $\mathrm{col}(V_1)$ and $\mathrm{col}(V_1)^\perp$:

$$G_{2,t}(X_t) = P_{V_1}G_{2,t}(X_t) + P_{V_1}^\perp G_{2,t}(X_t).$$

Since $V_1 f(V_1^\top X_t,t) - P_{V_1}G_{2,t}(X_t) \in \mathrm{col}(V_1)$ and $P_{V_1}^\perp G_{2,t}(X_t) \in \mathrm{col}(V_1)^\perp$, the two terms are orthogonal. Hence

$$r_{2,t}(s_{V_1,f}) = \frac{1}{h^2(t)}\mathbb{E}\left[\left\|f(V_1^\top X_t,t) - V_1^\top G_{2,t}(X_t)\right\|_2^2\right] + \frac{1}{h^2(t)}\mathbb{E}\left[\left\|P_{V_1}^\perp G_{2,t}(X_t)\right\|_2^2\right].$$

Consider the measurable predictor

$$f_{\mathrm{comp}}(z,t) := Bg_2(B^\dagger z, t).$$

Define

$$s_{\mathrm{comp}}(x,t) = \frac{1}{h(t)}V_1 f_{\mathrm{comp}}(V_1^\top x,t) - \frac{1}{h(t)}x, \qquad \mathcal{U}_{\mathrm{comp}}(V_1) := \mathcal{L}_2(s_{\mathrm{comp}}).$$

Since f_{comp} is measurable, $s_{\mathrm{comp}} \in \mathcal{S}_{\mathrm{str}}(V_1)$. Hence

$$\mathcal{B}_{\mathrm{str}}(V_1) \leqslant \mathcal{U}_{\mathrm{comp}}(V_1).$$

To bound $\mathcal{U}_{\mathrm{comp}}(V_1)$, we have

$$r_{2,t}(s_{\mathrm{comp}}) = \frac{1}{h^2(t)}\underbrace{\mathbb{E}\left[\left\|f_{\mathrm{comp}}(V_1^\top X_t,t) - V_1^\top G_{2,t}(X_t)\right\|_2^2\right]}_{\mathcal{I}_t} + \frac{1}{h^2(t)}\underbrace{\mathbb{E}\left[\left\|P_{V_1}^\perp G_{2,t}(X_t)\right\|_2^2\right]}_{\mathcal{O}_t}.$$

We first bound $\mathcal{O}_t$. Recall that

$$G_{2,t}(X_t) = A_2 g_2(Y_t, t) + \rho_2(t) U_t, \qquad Y_t := A_2^\top X_t, \quad U_t := P_2^\perp X_t.$$

By the same orthogonal factorization argument used in the noisy score decomposition lemma, Y_t and U_t are independent, $U_t \sim \mathcal{N}(0, \tilde{h}_2(t) P_2^\perp)$, and $\mathbb{E}[U_t] = 0$. Therefore,

$$\mathcal{O}_t = \mathbb{E}\left[\left\|P_{V_1}^\perp A_2 g_2(Y_t, t)\right\|_2^2\right] + \rho_2^2(t)\mathbb{E}\left[\left\|P_{V_1}^\perp U_t\right\|_2^2\right].$$

Using $M_{g,2}(t) \preceq \lambda_{g,2}^{\max}(t) I_{d_2}$, we have

$$\mathbb{E}\left[\left\|P_{V_1}^\perp A_2 g_2(Y_t, t)\right\|_2^2\right] = \operatorname{Tr}(A_2^\top P_{V_1}^\perp A_2 M_{g,2}(t)) \leqslant \lambda_{g,2}^{\max}(t) \operatorname{Tr}(A_2^\top P_{V_1}^\perp A_2) = \lambda_{g,2}^{\max}(t) \|P_{V_1}^\perp A_2\|_F^2.$$

Also,

$$\mathbb{E}\left[\left\|P_{V_1}^\perp U_t\right\|_2^2\right] = \operatorname{Tr}\left(P_{V_1}^\perp \mathbb{E}[U_t U_t^\top] P_{V_1}^\perp\right) = \tilde{h}_2(t) \operatorname{Tr}\left(P_{V_1}^\perp P_2^\perp\right).$$

Thus

$$\mathcal{O}_t \leqslant \lambda_{g,2}^{\max}(t) \|P_{V_1}^\perp A_2\|_F^2 + \rho_2^2(t) \tilde{h}_2(t) \operatorname{Tr}\left(P_{V_1}^\perp P_2^\perp\right). \tag{7}$$

Next, we bound $\mathcal{I}_t$. Recall

$$Z_t := V_1^\top X_t, \qquad Q_t := V_1^\top G_{2,t}(X_t).$$

Since $X_t = A_2 Y_t + U_t$, we have

$$Z_t = V_1^\top X_t = V_1^\top A_2 Y_t + V_1^\top U_t = B Y_t + V_1^\top U_t.$$

Similarly,

$$Q_t = V_1^\top G_{2,t}(X_t) = B g_2(Y_t, t) + \rho_2(t) V_1^\top U_t.$$

From the expression of Z_t, we have

$$B^\dagger Z_t = B^\dagger B Y_t + B^\dagger V_1^\top U_t.$$

Therefore,

$$Y_t - B^\dagger Z_t = (I_{d_2} - B^\dagger B) Y_t - B^\dagger V_1^\top U_t.$$

Moreover,

$$Q_t - f_{\mathrm{comp}}(Z_t, t) = B\left[g_2(Y_t, t) - g_2(B^\dagger Z_t, t)\right] + \rho_2(t) V_1^\top U_t.$$

Using the $L_g(t)$-Lipschitz continuity of $g_2(\cdot, t)$, we get

$$\left\|B\left[g_2(Y_t, t) - g_2(B^\dagger Z_t, t)\right]\right\|_2 \leqslant \|B\|_{\mathrm{op}} L_g(t) \left\|Y_t - B^\dagger Z_t\right\|_2.$$

Hence

$$\|Q_t - f_{\mathrm{comp}}(Z_t, t)\|_2 \leqslant \|B\|_{\mathrm{op}} L_g(t) \left\|Y_t - B^\dagger Z_t\right\|_2 + \rho_2(t) \|V_1^\top U_t\|_2.$$

Using $(a+b)^2 \leqslant 2a^2 + 2b^2$, we obtain

$$\mathcal{I}_t \leqslant 2\|B\|_{\mathrm{op}}^2 L_g^2(t) \mathbb{E}\left[\left\|Y_t - B^\dagger Z_t\right\|_2^2\right] + 2\rho_2^2(t) \mathbb{E}\left[\left\|V_1^\top U_t\right\|_2^2\right]. \tag{8}$$

We now compute the two expectations. Since Y_t and U_t are independent and $\mathbb{E}[U_t] = 0$, the cross term in $\mathbb{E}\|Y_t - B^\dagger Z_t\|_2^2$ vanishes. Therefore,

$$\mathbb{E}\left[\left\|Y_t - B^\dagger Z_t\right\|_2^2\right] = \mathbb{E}\left[\left\|(I_{d_2} - B^\dagger B) Y_t\right\|_2^2\right] + \mathbb{E}\left[\left\|B^\dagger V_1^\top U_t\right\|_2^2\right].$$

Since $B^\dagger B$ is the orthogonal projector onto the row space of B, $I_{d_2} - B^\dagger B$ is also an orthogonal projector. Thus

$$\mathbb{E}\left[\left\|(I_{d_2} - B^\dagger B) Y_t\right\|_2^2\right] = \operatorname{Tr}\left((I_{d_2} - B^\dagger B) M_{Y,2}(t)\right).$$

Also, since $\mathbb{E}[U_t U_t^\top] = \tilde{h}_2(t) P_2^\perp$,

$$\mathbb{E}\left[\left\|B^\dagger V_1^\top U_t\right\|_2^2\right] = \tilde{h}_2(t)\,\mathrm{Tr}\left(B^\dagger V_1^\top P_2^\perp V_1 (B^\dagger)^\top\right).$$

Similarly,

$$\mathbb{E}\left[\left\|V_1^\top U_t\right\|_2^2\right] = \tilde{h}_2(t)\,\mathrm{Tr}\left(V_1^\top P_2^\perp V_1\right).$$

Substituting these identities into (8) yields

$$\begin{aligned}\mathcal{I}_t \leqslant\ & 2\|B\|_{\mathrm{op}}^2 L_g^2(t)\,\mathrm{Tr}\left((I_{d_2} - B^\dagger B) M_{Y,2}(t)\right) + 2\|B\|_{\mathrm{op}}^2 L_g^2(t)\tilde{h}_2(t)\,\mathrm{Tr}\left(B^\dagger V_1^\top P_2^\perp V_1 (B^\dagger)^\top\right) \\ & + 2\rho_2^2(t)\tilde{h}_2(t)\,\mathrm{Tr}\left(V_1^\top P_2^\perp V_1\right).\end{aligned} \tag{9}$$

Combining (7), and (9), we obtain the general time-t bound. Integrating over $t \in [t_0, T]$ gives the integrated bound.

It remains to derive the two principal-angle forms.

First suppose $d_1 \geqslant d_2$ and B has full column rank. Then $B^\dagger B = I_{d_2}$, so

$$\mathrm{Tr}\left((I_{d_2} - B^\dagger B) M_{Y,2}(t)\right) = 0.$$

The singular values of $B = V_1^\top A_2$ are

$$\sigma_j(B) = \cos\theta_j, \qquad j = 1, \ldots, d_2.$$

Hence

$$\|P_{V_1}^\perp A_2\|_F^2 = \sum_{j=1}^{d_2} \sin^2\theta_j.$$

Moreover,

$$\mathrm{Tr}(P_{V_1}^\perp P_2^\perp) = D - d_1 - d_2 + \sum_{j=1}^{d_2} \cos^2\theta_j,$$

and

$$\mathrm{Tr}(V_1^\top P_2^\perp V_1) = d_1 - \sum_{j=1}^{d_2} \cos^2\theta_j.$$

A direct calculation in principal-vector coordinates gives

$$\mathrm{Tr}\left(B^\dagger V_1^\top P_2^\perp V_1 (B^\dagger)^\top\right) = \sum_{j=1}^{d_2} \tan^2\theta_j.$$

Using $\rho_2^2(t)\tilde{h}_2(t) = \alpha^4(t)\sigma_2^4/\tilde{h}_2(t)$ gives the stated bound for $d_1 \geqslant d_2$.

Now suppose $d_1 < d_2$ and B has full row rank. Then $B^\dagger B$ is the orthogonal projector onto a d_1-dimensional subspace of $\mathbb{R}^{d_2}$, and the term

$$\mathrm{Tr}\left((I_{d_2} - B^\dagger B) M_{Y,2}(t)\right)$$

remains as an irreducible information-loss term. In this case $r = d_1$, and

$$\|P_{V_1}^\perp A_2\|_F^2 = d_2 - \sum_{j=1}^{d_1} \cos^2\theta_j,$$

because A_2 has $d_2 - d_1$ target directions that cannot be contained in $\mathrm{col}(V_1)$. Furthermore,

$$\mathrm{Tr}(P_{V_1}^\perp P_2^\perp) = D - d_1 - d_2 + \sum_{j=1}^{d_1} \cos^2\theta_j,$$

and

$$\mathrm{Tr}(V_1^\top P_2^\perp V_1) = d_1 - \sum_{j=1}^{d_1} \cos^2\theta_j.$$

The same principal-coordinate calculation gives

$$\mathrm{Tr}\left(B^\dagger V_1^\top P_2^\perp V_1 (B^\dagger)^\top\right) = \sum_{j=1}^{d_1} \tan^2\theta_j.$$

Using again $\rho_2^2(t)\tilde{h}_2(t) = \alpha^4(t)\sigma_2^4/\tilde{h}_2(t)$ yields the stated bound for $d_1 < d_2$. This proves the proposition. □

Proposition 14 (ReLU approximation of the frozen-coordinate comparator)**.** *Suppose Assumptions 1, 2, and 3 hold for the target distribution p_2. Assume additionally that for every $R > 0$,*

$$\tau_{\mathrm{comp}}(R) := \sup_{t\in[t_0,T],\, \|z\|_\infty \leqslant R} \|\partial_t f_{\mathrm{comp}}(z,t)\|_2 < \infty.$$

Let $V_1 \in \mathbb{R}^{D\times d_1}$ be the frozen projector with $V_1^\top V_1 = I_{d_1}$, and let

$$B := V_1^\top A_2.$$

Define

$$L_{\mathrm{comp}} := \|B\|_{\mathrm{op}}\|B^\dagger\|_{\mathrm{op}}(\beta_2 + 1).$$

Let

$$G_{\mathrm{comp},0} := \sup_{t\in[t_0,T]} \|f_{\mathrm{comp}}(0,t)\|_2.$$

Recall the frozen-coordinate comparator

$$f_{\mathrm{comp}}(z,t) := Bg_2(B^\dagger z, t), \qquad z \in \mathbb{R}^{d_1}, \quad t \in [t_0,T].$$

Let

$$Z_t := V_1^\top X_t, \qquad P_{Z,t} := (V_1^\top)_{\#} p_{2,t}.$$

For any approximation level $\epsilon > 0$, choose

$$R_Z = C_{Z,1}\left(\sqrt{d_1} + \sqrt{d_2} + \sqrt{\log\frac{C_{Z,2}(1 + G_{\mathrm{comp},0}^2 + L_{\mathrm{comp}}^2)}{\epsilon^2}}\right).$$

Define

$$K_{\mathrm{comp}}(R_Z) := G_{\mathrm{comp},0} + L_{\mathrm{comp}}\sqrt{d_1}R_Z.$$

Then there exists a ReLU network

$$f_\theta \in \mathcal{F}_{\mathrm{ReLU}}^{(d_1)}(L, M, J, K, \kappa, \gamma, \gamma_t)$$

such that

$$\sup_{t\in[t_0,T]} \|f_\theta(\cdot,t) - f_{\mathrm{comp}}(\cdot,t)\|_{L^2(P_{Z,t})} \leqslant \sqrt{d_1+1}\epsilon.$$

The architecture parameters are chosen as in Lemma 12 with $m = d_1, F = f_{\mathrm{comp}}, R = R_Z, L_z = L_{\mathrm{comp}}, L_t = \tau_{\mathrm{comp}}(R_Z), K_0 = K_{\mathrm{comp}}(R_Z)$. Consequently,

$$\inf_{f_\theta\in\mathcal{F}_{\mathrm{ReLU}}^{(d_1)}} \mathcal{A}_{\mathrm{comp}}(f_\theta; V_1) \leqslant (d_1+1)\epsilon^2 \cdot \frac{1}{T-t_0}\int_{t_0}^T \frac{1}{h^2(t)}\,\mathrm{d}t.$$

In particular, with $C_T = (T/(1-e^{-T}))^2$,

$$\inf_{f_\theta\in\mathcal{F}_{\mathrm{ReLU}}^{(d_1)}} \mathcal{A}_{\mathrm{comp}}(f_\theta; V_1) \leqslant \frac{C_T(d_1+1)\epsilon^2}{t_0(T-t_0)}.$$

Here the architecture is chosen according to Lemma 12. For fixed T and $t_0 \leqslant T/2$, this has order $(d_1+1)\epsilon^2/t_0$.

Proof of Proposition 14. For each $t \in [t_0, T]$, define the frozen-coordinate distribution

$$P_{Z,t} := (V_1^\top)_{\#} p_{2,t}.$$

Equivalently, if $X_t \sim p_{2,t}$, then

$$Z_t := V_1^\top X_t \sim P_{Z,t}.$$

By the definition of $\mathcal{A}_{\mathrm{comp}}$, for any $f_\theta \in \mathcal{F}^{(d_1)}_{\mathrm{ReLU}}$,

$$\mathcal{A}_{\mathrm{comp}}(f_\theta; V_1) = \frac{1}{T - t_0} \int_{t_0}^{T} \frac{1}{h^2(t)} \, \|f_\theta(\cdot, t) - f_{\mathrm{comp}}(\cdot, t)\|^2_{L^2(P_{Z,t})} \, \mathrm{d}t.$$

Therefore, it suffices to construct $f_\theta \in \mathcal{F}^{(d_1)}_{\mathrm{ReLU}}(L, M, J, K, \kappa, \gamma, \gamma_t)$ such that

$$\sup_{t \in [t_0, T]} \|f_\theta(\cdot, t) - f_{\mathrm{comp}}(\cdot, t)\|_{L^2(P_{Z,t})} \leqslant \sqrt{d_1 + 1}\epsilon.$$

Indeed, such a bound immediately implies

$$\mathcal{A}_{\mathrm{comp}}(f_\theta; V_1) \leqslant (d_1 + 1)\epsilon^2 \cdot \frac{1}{T - t_0} \int_{t_0}^{T} \frac{1}{h^2(t)} \, \mathrm{d}t.$$

Since $h(t) = 1 - e^{-t}$, the following bound holds with $C = C_T = (T/(1 - e^{-T}))^2$:

$$\frac{1}{T - t_0} \int_{t_0}^{T} \frac{1}{h^2(t)} \, \mathrm{d}t \leqslant \frac{C}{t_0(T - t_0)}.$$

Hence

$$\mathcal{A}_{\mathrm{comp}}(f_\theta; V_1) \lesssim \frac{C(d_1 + 1)\epsilon^2}{t_0(T - t_0)}.$$

Thus, the remaining task is to prove the uniform-in-time $L^2(P_{Z,t})$ approximation bound for f_{comp}.

Step 1: Regularity transfer from f_2^* to f_{comp}. Recall that

$$f_2^*(y, t) = \tilde{h}_2(t) \nabla \log p^{\mathrm{LD}}_{2,t}(y) + y, \qquad g_2(y, t) = h(t) \nabla \log p^{\mathrm{LD}}_{2,t}(y) + y.$$

Therefore,

$$g_2(y, t) = \frac{h(t)}{\tilde{h}_2(t)} f_2^*(y, t) + \left(1 - \frac{h(t)}{\tilde{h}_2(t)}\right) y.$$

By Assumption 3, $f_2^*(\cdot, t)$ is β_2-Lipschitz uniformly over $t \in [0, T]$. Since $0 \leqslant h(t)/\tilde{h}_2(t) \leqslant 1$, for any $y_1, y_2 \in \mathbb{R}^{d_2}$ and $t \in [t_0, T]$,

$$\|g_2(y_1, t) - g_2(y_2, t)\|_2 \leqslant (\beta_2 + 1)\|y_1 - y_2\|_2.$$

Now recall

$$f_{\mathrm{comp}}(z, t) = B g_2(B^\dagger z, t), \qquad B = V_1^\top A_2.$$

Define

$$L_{\mathrm{comp}} := \|B\|_{\mathrm{op}} \|B^\dagger\|_{\mathrm{op}} (\beta_2 + 1).$$

Then, for any $z_1, z_2 \in \mathbb{R}^{d_1}$ and $t \in [t_0, T]$,

$$\|f_{\mathrm{comp}}(z_1, t) - f_{\mathrm{comp}}(z_2, t)\|_2 \leqslant L_{\mathrm{comp}} \|z_1 - z_2\|_2.$$

Indeed,

$$\|f_{\mathrm{comp}}(z_1, t) - f_{\mathrm{comp}}(z_2, t)\|_2 = \|B[g_2(B^\dagger z_1, t) - g_2(B^\dagger z_2, t)]\|_2.$$

Using the Lipschitz continuity of $g_2(\cdot, t)$, we obtain

$$\|B[g_2(B^\dagger z_1, t) - g_2(B^\dagger z_2, t)]\|_2 \leqslant \|B\|_{\mathrm{op}} (\beta_2 + 1) \|B^\dagger (z_1 - z_2)\|_2.$$

Thus,

$$\|f_{\mathrm{comp}}(z_1,t)-f_{\mathrm{comp}}(z_2,t)\|_2 \leqslant \|B\|_{\mathrm{op}}\|B^\dagger\|_{\mathrm{op}}(\beta_2+1)\|z_1-z_2\|_2 = L_{\mathrm{comp}}\|z_1-z_2\|_2.$$

Step 2: Growth and time regularity of f_{comp}. We next record two regularity properties that will be used in the truncation argument and in the ReLU construction on the compact domain.

First, f_{comp} has at most linear growth in z. Let

$$G_{2,0} := \sup_{t\in[t_0,T]} \|g_2(0,t)\|_2.$$

Under the assumed regularity of the target-domain core map on the compact interval $[t_0,T]$, $G_{2,0}<\infty$. Define

$$G_{\mathrm{comp},0} := \sup_{t\in[t_0,T]} \|f_{\mathrm{comp}}(0,t)\|_2 \leqslant \|B\|_{\mathrm{op}} G_{2,0}.$$

Then, for any $z\in\mathbb{R}^{d_1}$ and $t\in[t_0,T]$,

$$\|f_{\mathrm{comp}}(z,t)\|_2 \leqslant G_{\mathrm{comp},0} + L_{\mathrm{comp}}\|z\|_2.$$

Indeed, by the definition of f_{comp},

$$\|f_{\mathrm{comp}}(z,t)\|_2 \leqslant \|B\|_{\mathrm{op}}\|g_2(B^\dagger z,t)\|_2.$$

Using the Lipschitz bound for g_2, we have

$$\|g_2(B^\dagger z,t)\|_2 \leqslant \|g_2(0,t)\|_2 + (\beta_2+1)\|B^\dagger z\|_2.$$

Therefore,

$$\|f_{\mathrm{comp}}(z,t)\|_2 \leqslant \|B\|_{\mathrm{op}}G_{2,0} + \|B\|_{\mathrm{op}}\|B^\dagger\|_{\mathrm{op}}(\beta_2+1)\|z\|_2 = G_{\mathrm{comp},0} + L_{\mathrm{comp}}\|z\|_2.$$

Second, define the time-regularity modulus of the comparator on the truncated domain by

$$\tau_{\mathrm{comp}}(R) := \sup_{t\in[t_0,T],\,\|z\|_\infty\leqslant R} \|\partial_t f_{\mathrm{comp}}(z,t)\|_2.$$

This quantity will play the role of the time-Lipschitz constant in the construction of the ReLU approximation. Since B and $B^\dagger$ are fixed in time, whenever g_2 is differentiable in t, we have

$$\partial_t f_{\mathrm{comp}}(z,t) = B\partial_t g_2(B^\dagger z,t).$$

In particular, if

$$\tau_{g,2}(R) := \sup_{t\in[t_0,T],\,\|y\|_2\leqslant R} \|\partial_t g_2(y,t)\|_2$$

is finite, then

$$\tau_{\mathrm{comp}}(R) \leqslant \|B\|_{\mathrm{op}}\tau_{g,2}(\|B^\dagger\|_{\mathrm{op}}\sqrt{d_1}R).$$

Step 3: Tail control of the frozen-coordinate distribution. We now show that the frozen-coordinate distribution $P_{Z,t}$ has a uniformly controlled tail over $t\in[t_0,T]$. Under Assumption 1, we may write

$$X_t = \alpha(t)A_2 z + \alpha(t)\epsilon_2 + \sqrt{h(t)}\xi,$$

where $z\sim p_{2,z}$, $\epsilon_2\sim\mathcal{N}(0,\sigma_2^2 I_D)$, $\xi\sim\mathcal{N}(0,I_D)$, and these random variables are independent. Hence

$$Z_t = V_1^\top X_t = \alpha(t)Bz + \alpha(t)V_1^\top\epsilon_2 + \sqrt{h(t)}V_1^\top\xi.$$

Since $\|B\|_{\mathrm{op}}\leqslant 1$, $\alpha(t)\leqslant 1$, and $h(t)\leqslant 1$, we have

$$\|Z_t\|_2 \leqslant \|z\|_2 + \|G_t\|_2,$$

where

$$G_t := \alpha(t)V_1^\top\epsilon_2 + \sqrt{h(t)}V_1^\top\xi \sim \mathcal{N}(0,(\alpha^2(t)\sigma_2^2+h(t))I_{d_1}).$$

In particular,

$$\alpha^2(t)\sigma_2^2 + h(t) \leqslant \sigma_2^2 + 1.$$

By Assumption 2, z has a sub-Gaussian-type norm tail. Therefore, there exist dimension-free constants $c_Z, C_Z, C_{Z,0} > 0$, depending only on the tail constants of $p_{2,z}$ and on σ_2, such that for all $u \geqslant 0$,

$$\sup_{t\in[t_0,T]} \mathbb{P}\left(\|Z_t\|_2 > C_{Z,0}(\sqrt{d_2} + \sqrt{d_1}) + u\right) \leqslant C_Z \exp(-c_Z u^2).$$

Consequently, there exists a constant $C_Z' > 0$ such that for all $R \geqslant 2C_{Z,0}(\sqrt{d_2} + \sqrt{d_1})$,

$$\sup_{t\in[t_0,T]} \mathbb{E}\left[(1 + \|Z_t\|_2^2)\mathbf{1}_{\{\|Z_t\|_2 > R\}}\right] \leqslant C_Z'(1 + R^2)\exp(-c_Z R^2/C_Z'),$$

which follows from

$$\mathbb{E}[\|Z_t\|_2^2 \mathbf{1}_{\{\|Z_t\|_2 > R\}}] = R^2\mathbb{P}(\|Z_t\|_2 > R) + 2\int_R^\infty u\mathbb{P}(\|Z_t\|_2 > u)\,\mathrm{d}u$$

together with the preceding tail bound. In particular, choosing

$$R_Z = C_{Z,1}\left(\sqrt{d_1} + \sqrt{d_2} + \sqrt{\log\frac{C_{Z,2}(1 + G_{\mathrm{comp},0}^2 + L_{\mathrm{comp}}^2)}{\epsilon^2}}\right)$$

with sufficiently large constants $C_{Z,1}, C_{Z,2} > 0$, and enlarging $C_{Z,1}$ to absorb the polynomial factor generated by $K_{\mathrm{comp}}(R_Z)$, gives

$$\sup_{t\in[t_0,T]} \mathbb{E}\left[(1 + \|Z_t\|_2^2)\mathbf{1}_{\{\|Z_t\|_2 > R_Z\}}\right] \leqslant \frac{\epsilon^2}{C\left(1 + K_{\mathrm{comp}}^2(R_Z) + G_{\mathrm{comp},0}^2 + L_{\mathrm{comp}}^2\right)}.$$

This is possible because $K_{\mathrm{comp}}^2(R_Z) \leqslant C(1 + G_{\mathrm{comp},0}^2 + L_{\mathrm{comp}}^2 d_1 R_Z^2)$, while the tail bound decays exponentially in R_Z^2.

Step 4: ReLU approximation on the truncated domain. Let

$$\mathcal{D}_Z(R_Z) := [-R_Z, R_Z]^{d_1} \times [t_0, T].$$

By Steps 1 and 2, f_{comp} satisfies the assumptions of Lemma 12 on $\mathcal{D}_Z(R_Z)$ with

$$m = d_1, \qquad L_z = L_{\mathrm{comp}}, \qquad L_t = \tau_{\mathrm{comp}}(R_Z), \qquad K_0 = K_{\mathrm{comp}}(R_Z).$$

Applying Lemma 12, we obtain a network $f_\theta \in \mathcal{F}_{\mathrm{ReLU}}^{(d_1)}(L, M, J, K, \kappa, \gamma, \gamma_t)$ such that

$$\sup_{(z,t)\in\mathcal{D}_Z(R_Z)} \|f_\theta(z,t) - f_{\mathrm{comp}}(z,t)\|_\infty \leqslant \epsilon.$$

Therefore,

$$\sup_{(z,t)\in\mathcal{D}_Z(R_Z)} \|f_\theta(z,t) - f_{\mathrm{comp}}(z,t)\|_2 \leqslant \sqrt{d_1}\epsilon.$$

Moreover,

$$\sup_{z\in\mathbb{R}^{d_1}, t\in[t_0,T]} \|f_\theta(z,t)\|_2 \leqslant CK_{\mathrm{comp}}(R_Z).$$

Step 5: Compact-tail decomposition of the $L^2(P_{Z,t})$ error. For each $t \in [t_0, T]$, define the truncation event

$$\mathcal{E}_t := \{\|Z_t\|_\infty \leqslant R_Z\}.$$

Then

$$\begin{aligned}
&\mathbb{E}\left[\|f_\theta(Z_t,t) - f_{\mathrm{comp}}(Z_t,t)\|_2^2\right] \\
&= \mathbb{E}\left[\|f_\theta(Z_t,t) - f_{\mathrm{comp}}(Z_t,t)\|_2^2 \mathbf{1}_{\mathcal{E}_t}\right] + \mathbb{E}\left[\|f_\theta(Z_t,t) - f_{\mathrm{comp}}(Z_t,t)\|_2^2 \mathbf{1}_{\mathcal{E}_t^c}\right].
\end{aligned}$$

On $\mathcal{E}_t$, we have $(Z_t, t) \in [-R_Z, R_Z]^{d_1} \times [t_0, T]$. Therefore, by the compact-domain approximation constructed in Step 4,

$$\mathbb{E}\left[\|f_\theta(Z_t,t) - f_{\text{comp}}(Z_t,t)\|_2^2 \mathbf{1}_{\mathcal{E}_t}\right] \leqslant d_1\epsilon^2.$$

It remains to control the tail contribution. By the global boundedness of the constructed network and the linear growth bound of f_{comp}, we have

$$\|f_\theta(Z_t,t) - f_{\text{comp}}(Z_t,t)\|_2^2 \leqslant 2C^2K_{\text{comp}}^2(R_Z) + 2(G_{\text{comp},0} + L_{\text{comp}}\|Z_t\|_2)^2.$$

Hence, for another absolute constant $C > 0$,

$$\|f_\theta(Z_t,t) - f_{\text{comp}}(Z_t,t)\|_2^2 \leqslant C\left(1 + K_{\text{comp}}^2(R_Z) + G_{\text{comp},0}^2 + L_{\text{comp}}^2\right)(1 + \|Z_t\|_2^2).$$

Since $\mathcal{E}_t^c = \{\|Z_t\|_\infty > R_Z\} \subseteq \{\|Z_t\|_2 > R_Z\}$, the tail estimate from Step 3 gives

$$\begin{aligned}&\sup_{t\in[t_0,T]} \mathbb{E}\left[\|f_\theta(Z_t,t) - f_{\text{comp}}(Z_t,t)\|_2^2 \mathbf{1}_{\mathcal{E}_t^c}\right] \\ &\leqslant C\left(1 + K_{\text{comp}}^2(R_Z) + G_{\text{comp},0}^2 + L_{\text{comp}}^2\right) \sup_{t\in[t_0,T]} \mathbb{E}\left[(1 + \|Z_t\|_2^2)\mathbf{1}_{\{\|Z_t\|_2 > R_Z\}}\right].\end{aligned}$$

By the choice of R_Z in Step 3, we have

$$\sup_{t\in[t_0,T]} \mathbb{E}\left[(1 + \|Z_t\|_2^2)\mathbf{1}_{\{\|Z_t\|_2 > R_Z\}}\right] \leqslant \frac{\epsilon^2}{C\left(1 + K_{\text{comp}}^2(R_Z) + G_{\text{comp},0}^2 + L_{\text{comp}}^2\right)},$$

Therefore,

$$\sup_{t\in[t_0,T]} \mathbb{E}\left[\|f_\theta(Z_t,t) - f_{\text{comp}}(Z_t,t)\|_2^2 \mathbf{1}_{\mathcal{E}_t^c}\right] \leqslant \epsilon^2.$$

Such a choice of R_Z is possible because the frozen-coordinate tail is sub-Gaussian by Step 3, while $K_{\text{comp}}(R_Z)$ grows at most linearly in R_Z. Thus the polynomial factor in R_Z can be absorbed into the exponential tail by increasing the logarithmic constant in the choice of R_Z. Combining the compact and tail parts yields

$$\sup_{t\in[t_0,T]} \mathbb{E}\left[\|f_\theta(Z_t,t) - f_{\text{comp}}(Z_t,t)\|_2^2\right] \leqslant (d_1+1)\epsilon^2.$$

Equivalently,

$$\sup_{t\in[t_0,T]} \|f_\theta(\cdot,t) - f_{\text{comp}}(\cdot,t)\|_{L^2(P_{Z,t})} \leqslant \sqrt{d_1+1}\epsilon.$$

Step 6: From $L^2(P_{Z,t})$ approximation to $\mathcal{A}_{\text{comp}}$. By Step 5, the constructed ReLU network $f_\theta \in \mathcal{F}_{\text{ReLU}}^{(d_1)}$ satisfies

$$\sup_{t\in[t_0,T]} \|f_\theta(\cdot,t) - f_{\text{comp}}(\cdot,t)\|_{L^2(P_{Z,t})} \leqslant \sqrt{d_1+1}\epsilon.$$

Substituting this bound into the definition of $\mathcal{A}_{\text{comp}}$, we obtain

$$\mathcal{A}_{\text{comp}}(f_\theta; V_1) \leqslant (d_1+1)\epsilon^2 \cdot \frac{1}{T-t_0}\int_{t_0}^T \frac{1}{h^2(t)}\,\mathrm{d}t.$$

Since $h(t) = 1 - e^{-t}$, for $t \in [t_0, T]$ we have

$$h(t) \geqslant \frac{1-e^{-T}}{T}t.$$

Therefore,

$$\int_{t_0}^T \frac{1}{h^2(t)}\,\mathrm{d}t \leqslant \left(\frac{T}{1-e^{-T}}\right)^2 \int_{t_0}^T \frac{1}{t^2}\,\mathrm{d}t \leqslant \left(\frac{T}{1-e^{-T}}\right)^2 \frac{1}{t_0}.$$

It follows that

$$\mathcal{A}_{\text{comp}}(f_\theta; V_1) \leqslant (d_1+1)\epsilon^2 \cdot \frac{1}{T-t_0}\left(\frac{T}{1-e^{-T}}\right)^2 \frac{1}{t_0}.$$

Consequently,

$$\inf_{f_\theta \in \mathcal{F}_{\mathrm{ReLU}}^{(d_1)}} \mathcal{A}_{\mathrm{comp}}(f_\theta; V_1) \leqslant C_T(d_1+1)\frac{\epsilon^2}{t_0(T-t_0)},$$

where

$$C_T := \left(\frac{T}{1-e^{-T}}\right)^2.$$

If T is fixed and $t_0 \leqslant T/2$, this is of order

$$\frac{(d_1+1)\epsilon^2}{t_0}.$$

This proves the proposition.

□

Proof of Theorem 6. Write $R_*^2 = R_z^2 + R_\perp^2$ and $\Omega = \bigcap_{j=1}^{N_2}\{x_j \in \mathcal{T}_2\}$. The target latent tail assumption and Gaussian norm concentration give $\mathbb{P}(\Omega^c) \leqslant \delta/4$ for sufficiently large $C_z, C_\perp$. For every core in $\mathcal{F}$, the conditional denoising residual is

$$\nabla_{X_t} \log \phi_t(X_t \mid x) - s_{V_1,\theta}(X_t, t) = \frac{\alpha(t)x - V_1 f_\theta(V_1^\top X_t, t)}{h(t)}.$$

Thus

$$0 \leqslant \ell_2(x; s) \leqslant \frac{2C_T(K^2 + \|x\|_2^2)}{t_0(T-t_0)}.$$

Let $g_s = \ell_2(\cdot; s)\mathbf{1}_{\mathcal{T}_2}$, $\mathcal{R}_2^{\mathrm{tr}}(s) = \mathbb{E}g_s(X_0)$, and $\widehat{\mathcal{R}}_2^{\mathrm{tr}}(s) = N_2^{-1}\sum_j g_s(x_j)$. The same tail estimates, integrated against $K^2 + \|X_0\|_2^2$, give

$$0 \leqslant \sup_s[\mathcal{R}_2(s) - \mathcal{R}_2^{\mathrm{tr}}(s)] \leqslant C\frac{C_T(K^2+R_*^2)}{t_0(T-t_0)}\frac{\delta}{N_2}.$$

On Ω, the full empirical loss equals the truncated empirical loss for every candidate. Each g_s is bounded by $B_u = 2C_T(K^2+R_*^2)/(t_0(T-t_0))$. Bernstein's inequality, $\mathrm{Var}(g_s) \leqslant B_u\mathbb{E}g_s$, and an ι-net give, with probability at least $1-\delta/2$, simultaneously for all candidates,

$$\mathcal{R}_2^{\mathrm{tr}}(s) \leqslant (1+a)\widehat{\mathcal{R}}_2^{\mathrm{tr}}(s) + D_N,$$
$$\widehat{\mathcal{R}}_2^{\mathrm{tr}}(s) \leqslant (1+a)\mathcal{R}_2^{\mathrm{tr}}(s) + D_N,$$

where

$$D_N = C\left[\frac{B_u}{aN_2}\left(\log \mathcal{N}_2(\iota) + \log\frac{8}{\delta}\right) + \iota\right].$$

For completeness, the multiplicative form follows by absorbing the square-root Bernstein term using Young's inequality with parameter proportional to a; replacing a function by an ι-net representative adds at most $(2+a)\iota$. The measurability conditions permit applying the bound to the class suprema. Choose any population comparator $\bar{s} \in \mathcal{S}_{\mathrm{freeze}}(V_1)$. On the intersection of these events, empirical minimization and the two displayed bounds yield

$$\mathcal{R}_2(\widehat{s}_{\mathrm{freeze}}) \leqslant (1+a)^2\mathcal{R}_2(\bar{s}) + \mathrm{Est}_{N_2}(V_1; \mathcal{F}, a, \iota, \delta),$$

after increasing the constant in Est_{N_2} to absorb truncation. An optimization error ξ adds $(1+a)\xi$. Since $\mathcal{R}_2 = \mathcal{L}_2 + E_2$,

$$\mathcal{L}_2(\widehat{s}_{\mathrm{freeze}}) \leqslant (1+a)^2\mathcal{L}_2(\bar{s}) + \mathrm{Est}_{N_2}(V_1; \mathcal{F}, a, \iota, \delta) + (2a+a^2)E_2.$$

Take comparators approaching the class infimum and apply Lemma 5. Finally, the denoising-score-matching identity implies

$$0 \leqslant E_2 \leqslant \frac{D}{T-t_0}\int_{t_0}^{T}\frac{\mathrm{d}t}{h(t)} = \frac{D}{T-t_0}\log\frac{e^T-1}{e^{t_0}-1}.$$

The combined failure probability is at most δ. □

Proof of Corollary 7. Apply Theorem 6 with the ReLU class. Proposition 14 and Lemma 12 give $\mathcal{A}_{\mathcal{F}}(V_1) \leqslant C_T(d_1+1)\epsilon^2/[t_0(T-t_0)]$. Set $a = \epsilon^2$, $\eta = 1$, and use $\epsilon \in (0,1)$ to bound the remaining multiplicative constants. Lemma 13 supplies the stated covering entropy. The term $(2a+a^2)E_2$ is bounded by $3\epsilon^2 E_2$ and is retained in the corollary. □

C.3 Proof of Section 3.3

Proof of Proposition 8. Using $P_W^\perp = I_D - P_W$, we have

$$\|P_W^\perp A_i\|_F^2 = \mathrm{Tr}(A_i^\top P_W^\perp A_i) = d_i - \mathrm{Tr}(W^\top A_i A_i^\top W).$$

Therefore

$$\Gamma_k(W) = \omega_1 c_1 d_1 + \omega_2 c_2 d_2 - \mathrm{Tr}\left(W^\top M_{\mathrm{mix}} W\right).$$

Since

$$\mathrm{Tr}(M_{\mathrm{mix}}) = \omega_1 c_1 d_1 + \omega_2 c_2 d_2,$$

minimizing $\Gamma_k(W)$ is equivalent to maximizing

$$\mathrm{Tr}(W^\top M_{\mathrm{mix}} W)$$

over all $D \times k$ matrices with orthonormal columns. By the Rayleigh–Ritz variational principle, the maximum equals

$$\sum_{j=1}^{k} \lambda_j(M_{\mathrm{mix}}),$$

and it is achieved by taking W to be the top k eigenvectors of M_{mix}. The monotonicity in k follows immediately from the eigenvalue formula. Finally, the residual is zero if and only if $P_W^\perp A_1 = 0$ and $P_W^\perp A_2 = 0$, which is equivalent to both component subspaces being contained in $\mathrm{span}(W)$. □

Corollary 15 (Principal-angle interpretation)**.** *Let $A_i \in \mathbb{R}^{D \times d_i}$ have orthonormal columns for $i = 1, 2$. Set*

$$a := \omega_1 c_1, \qquad b := \omega_2 c_2, \qquad M_{\mathrm{mix}} := a A_1 A_1^\top + b A_2 A_2^\top.$$

Let $r := \min\{d_1, d_2\}$, and let

$$0 \leqslant \varphi_1 \leqslant \cdots \leqslant \varphi_r \leqslant \pi/2$$

be the principal angles between $\mathrm{span}(A_1)$ and $\mathrm{span}(A_2)$. Then M_{mix} is zero on

$$(\mathrm{span}(A_1) + \mathrm{span}(A_2))^\perp.$$

On the two-dimensional principal-angle block associated with φ_j, the eigenvalues of M_{mix} are

$$\lambda_j^\pm = \frac{a+b}{2} \pm \frac{1}{2}\sqrt{(a-b)^2 + 4ab\cos^2\varphi_j}, \qquad j = 1, \ldots, r. \tag{10}$$

If $d_1 > d_2$, then M_{mix} has additionally $d_1 - d_2$ eigenvalues equal to a. If $d_2 > d_1$, then M_{mix} has additionally $d_2 - d_1$ eigenvalues equal to b. All remaining eigenvalues are zero.

Proof of Corollary 15. By the cosine-sine decomposition for two subspaces, there exist orthonormal principal vectors $u_j \in \mathrm{span}(A_1)$ and $v_j \in \mathrm{span}(A_2)$ such that $u_j^\top v_j = \cos\varphi_j$ for $j = 1, \ldots, r$. The action of P_{A_1} and P_{A_2} decomposes into independent principal-angle blocks, together with the extra directions belonging only to the larger subspace.

On the block spanned by u_j and the normalized component of v_j orthogonal to u_j, the matrix $aP_{A_1} + bP_{A_2}$ has trace $a + b$ and determinant $ab\sin^2\varphi_j$. Hence its two eigenvalues are

$$\lambda_j^\pm = \frac{a+b}{2} \pm \frac{1}{2}\sqrt{(a+b)^2 - 4ab\sin^2\varphi_j} = \frac{a+b}{2} \pm \frac{1}{2}\sqrt{(a-b)^2 + 4ab\cos^2\varphi_j}.$$

If $d_1 > d_2$, the $d_1 - d_2$ directions in $\mathrm{span}(A_1)$ orthogonal to $\mathrm{span}(A_2)$ are eigenvectors with eigenvalue a. If $d_2 > d_1$, the additional directions in $\mathrm{span}(A_2)$ have eigenvalue b. On the orthogonal complement of $\mathrm{span}(A_1) + \mathrm{span}(A_2)$, both projectors vanish, so the eigenvalue is zero. □

In the special case $d_1 = d_2 = d$, $c_1 = c_2 = 1$, and $\omega_1 = \pi, \omega_2 = 1 - \pi$, we have $a = \pi$ and $b = 1 - \pi$, so (10) becomes

$$\lambda_j^{\pm} = \frac{1}{2}\left[1 \pm \sqrt{1 - 4\pi(1-\pi)\sin^2\varphi_j}\right], \qquad j = 1, \ldots, d.$$

In the balanced case $\pi = 1/2$, this further reduces to

$$\lambda_j^{\pm} = \frac{1 \pm \cos\varphi_j}{2}.$$

Consequently, the optimal k-dimensional shared projector in Proposition 8 selects the directions corresponding to the largest eigenvalues of M_{mix}. Principal directions with small angles concentrate most of their mixed subspace energy in one direction, whereas directions with large angles require additional latent dimensions to be well represented. If $\varphi_j = 0$, then $\lambda_j^- = 0$, reflecting that the corresponding principal-angle block degenerates into a shared one-dimensional direction.

Proof of Corollary 9. The target-domain comparator and estimation arguments require only a fixed orthonormal representation, not that it be a source-population subspace. Substituting W, k for V_1, d_1 proves the claim. Under containment, $B_W^\top B_W = I_{d_2}$, so all d_2 principal angles are zero and $B_W^\dagger B_W = I_{d_2}$. □

Proof of Theorem 10. **Score structure under mixed training.** Let

$$p_{\mathrm{mix}} := \omega_1 p_1 + \omega_2 p_2, \qquad \omega_1 + \omega_2 = 1,$$

and denote by $p_{\mathrm{mix},t}$ the corresponding forward-diffused density. Since the forward diffusion is linear,

$$p_{\mathrm{mix},t}(x) = \omega_1 p_{1,t}(x) + \omega_2 p_{2,t}(x).$$

Define the posterior mixing weights

$$\pi_i(x,t) := \frac{\omega_i p_{i,t}(x)}{\omega_1 p_{1,t}(x) + \omega_2 p_{2,t}(x)}, \qquad i = 1, 2.$$

Then

$$\nabla \log p_{\mathrm{mix},t}(x) = \pi_1(x,t)\nabla \log p_{1,t}(x) + \pi_2(x,t)\nabla \log p_{2,t}(x).$$

For each component, define

$$g_i(y,t) := h(t)\nabla \log p_{i,t}^{\mathrm{LD}}(y) + y, \qquad \rho_i(t) := 1 - \frac{h(t)}{\tilde{h}_i(t)} = \frac{\alpha^2(t)\sigma_i^2}{\tilde{h}_i(t)}.$$

For $i = 1, 2$, define

$$M_{g,i}(t) := \mathbb{E}_{X_t \sim p_{i,t}}\left[g_i(A_i^\top X_t, t) g_i(A_i^\top X_t, t)^\top\right],$$

and

$$\lambda_{g,i}^{\max}(t) := \lambda_{\max}(M_{g,i}(t)).$$

Let

$$P_i := A_i A_i^\top, \qquad P_i^\perp := I_D - P_i,$$

and

$$G_{i,t}(x) := A_i g_i(A_i^\top x, t) + \rho_i(t) P_i^\perp x.$$

Then the mixed score can be written as

$$\nabla \log p_{\mathrm{mix},t}(x) = \frac{1}{h(t)} G_{\mathrm{mix},t}(x) - \frac{1}{h(t)} x,$$

where

$$G_{\mathrm{mix},t}(x) := \pi_1(x,t) G_{1,t}(x) + \pi_2(x,t) G_{2,t}(x).$$

Let $U \in \mathbb{R}^{D\times m}$ be an ideal latent projector with $U^\top U = I_m$. In the comparison below, U will be instantiated either as the source-aligned projector V_1 with $m = d_1$, or as the shared mixed-data projector W_k with $m = k$. The word "ideal" is used only at the oracle-analysis level: $\mathcal{B}_{\mathrm{mix}}(U)$ measures the best approximation power of the score network class after a projector has been chosen, and does not describe the finite-sample cost of learning the projector itself. Define

$$P_U := UU^\top, \qquad P_U^\perp := I_D - P_U, \qquad H_i(U) := U^\top A_i, \quad i = 1, 2.$$

Consider the mixed-data score class induced by U:

$$\mathcal{S}_{\mathrm{mix}}(U) := \left\{ s_{U,\theta}(x,t) = \frac{1}{h(t)} U f_\theta(U^\top x, t) - \frac{1}{h(t)} x : f_\theta \in \mathcal{F}^{(m)}_{\mathrm{ReLU}} \right\}.$$

For any score network s, define the mixed score-matching risk

$$\mathcal{L}_{\mathrm{mix}}(s) := \frac{1}{T - t_0} \int_{t_0}^{T} \mathbb{E}_{X_t \sim p_{\mathrm{mix},t}} \left[\|s(X_t, t) - \nabla \log p_{\mathrm{mix},t}(X_t)\|_2^2 \right] \mathrm{d}t.$$

The mixed ReLU oracle risk under the projector U is

$$\mathcal{B}_{\mathrm{mix}}(U) := \inf_{s \in \mathcal{S}_{\mathrm{mix}}(U)} \mathcal{L}_{\mathrm{mix}}(s).$$

For $i = 1, 2$, let $q_{i,U,t}$ denote the density of $U^\top X_t$ when $X_t \sim p_{i,t}$. Define the projected posterior mixing weights by

$$\pi_i^U(z,t) := \frac{\omega_i q_{i,U,t}(z)}{\omega_1 q_{1,U,t}(z) + \omega_2 q_{2,U,t}(z)}, \qquad i = 1, 2.$$

Let $C \in \{1, 2\}$ denote the component label, with $\mathbb{P}(C = i) = \omega_i$ and $X_t \mid C = i \sim p_{i,t}$. Then

$$\pi_i^U(z,t) = \mathbb{P}(C = i \mid U^\top X_t = z).$$

Thus, π_i^U is the Bayes posterior component weight after observing only the compressed coordinate $U^\top X_t$. If the exact component conditional predictors

$$\mu_{i,U,t}(z) := \mathbb{E}\left[U^\top G_{i,t}(X_t) \mid U^\top X_t = z,\ C = i \right]$$

were available, then the optimal measurable predictor of $U^\top G_{\mathrm{mix},t}(X_t)$ given $U^\top X_t = z$ would be

$$\sum_{i=1}^{2} \pi_i^U(z,t) \mu_{i,U,t}(z).$$

As in the frozen-projector analysis, directly controlling these conditional predictors is difficult. We therefore use an explicit componentwise comparator:

$$\psi_{i,U}(z,t) := H_i(U) g_i(H_i(U)^\dagger z, t), \qquad i = 1, 2.$$

The mixed explicit comparator is then defined as

$$f^{\mathrm{comp}}_{\mathrm{mix},U}(z,t) := \pi_1^U(z,t)\psi_{1,U}(z,t) + \pi_2^U(z,t)\psi_{2,U}(z,t).$$

Finally, define

$$s^{\mathrm{comp}}_{\mathrm{mix},U}(x,t) := \frac{1}{h(t)} U f^{\mathrm{comp}}_{\mathrm{mix},U}(U^\top x, t) - \frac{1}{h(t)} x,$$

and

$$\mathcal{U}_{\mathrm{mix}}(U) := \mathcal{L}_{\mathrm{mix}}(s^{\mathrm{comp}}_{\mathrm{mix},U}).$$

Define the ReLU approximation error of the mixed comparator under U by

$$\mathcal{A}_{\mathrm{mix}}(f_\theta; U) := \frac{1}{T - t_0} \int_{t_0}^{T} \frac{1}{h^2(t)} \mathbb{E}_{X_t \sim p_{\mathrm{mix},t}} \left[\left\| f_\theta(U^\top X_t, t) - f^{\mathrm{comp}}_{\mathrm{mix},U}(U^\top X_t, t) \right\|_2^2 \right] \mathrm{d}t.$$

Then, for any $\eta > 0$,

$$\mathcal{B}_{\mathrm{mix}}(U) \leqslant (1+\eta)\mathcal{U}_{\mathrm{mix}}(U) + \left(1 + \frac{1}{\eta}\right) \inf_{f_\theta \in \mathcal{F}^{(m)}_{\mathrm{ReLU}}} \mathcal{A}_{\mathrm{mix}}(f_\theta; U).$$

Indeed, for any $f_\theta \in \mathcal{F}^{(m)}_{\mathrm{ReLU}}$, one decomposes

$$\begin{aligned} U f_\theta(U^\top X_t, t) - G_{\mathrm{mix},t}(X_t) = &\left[U f^{\mathrm{comp}}_{\mathrm{mix},U}(U^\top X_t, t) - G_{\mathrm{mix},t}(X_t)\right] \\ &+ U\left[f_\theta(U^\top X_t, t) - f^{\mathrm{comp}}_{\mathrm{mix},U}(U^\top X_t, t)\right], \end{aligned}$$

and applies $\|a+b\|_2^2 \leqslant (1+\eta)\|a\|_2^2 + (1+1/\eta)\|b\|_2^2$, using $U^\top U = I_m$.

The measurable-comparator risk decomposes into an output-side structural term and an in-space comparator term:

$$\mathcal{U}_{\mathrm{mix}}(U) = \mathcal{O}_{\mathrm{mix}}(U) + \mathcal{I}_{\mathrm{mix}}(U),$$

where

$$\mathcal{O}_{\mathrm{mix}}(U) := \frac{1}{T-t_0}\int_{t_0}^{T} \frac{1}{h^2(t)} \mathbb{E}_{X_t \sim p_{\mathrm{mix},t}}\left[\left\|P_U^\perp G_{\mathrm{mix},t}(X_t)\right\|_2^2\right] \mathrm{d}t,$$

and

$$\mathcal{I}_{\mathrm{mix}}(U) := \frac{1}{T-t_0}\int_{t_0}^{T} \frac{1}{h^2(t)} \mathbb{E}_{X_t \sim p_{\mathrm{mix},t}}\left[\left\|f^{\mathrm{comp}}_{\mathrm{mix},U}(U^\top X_t, t) - U^\top G_{\mathrm{mix},t}(X_t)\right\|_2^2\right] \mathrm{d}t.$$

This follows by decomposing $G_{\mathrm{mix},t}(X_t)$ into its P_U- and $P_U^\perp$-components and using the orthogonality of $\mathrm{span}(U)$ and its complement.

The output-side term is controlled by the joint subspace residual. Since

$$G_{\mathrm{mix},t}(x) = \pi_1(x,t)G_{1,t}(x) + \pi_2(x,t)G_{2,t}(x),$$

convexity gives

$$\left\|P_U^\perp G_{\mathrm{mix},t}(x)\right\|_2^2 \leqslant \sum_{i=1}^{2} \pi_i(x,t)\left\|P_U^\perp G_{i,t}(x)\right\|_2^2.$$

Using $\pi_i(x,t)p_{\mathrm{mix},t}(x) = \omega_i p_{i,t}(x)$, we obtain

$$\mathbb{E}_{p_{\mathrm{mix},t}}\left[\left\|P_U^\perp G_{\mathrm{mix},t}(X_t)\right\|_2^2\right] \leqslant \sum_{i=1}^{2} \omega_i \mathbb{E}_{p_{i,t}}\left[\left\|P_U^\perp G_{i,t}(X_t)\right\|_2^2\right].$$

For each component, the same calculation as in the frozen-projector analysis yields

$$\mathbb{E}_{p_{i,t}}\left[\left\|P_U^\perp G_{i,t}(X_t)\right\|_2^2\right] \leqslant \lambda_{g,i}^{\max}(t)\|P_U^\perp A_i\|_F^2 + \frac{\alpha^4(t)\sigma_i^4}{\tilde{h}_i(t)}\,\mathrm{Tr}(P_U^\perp P_i^\perp).$$

Therefore,

$$\mathcal{O}_{\mathrm{mix}}(U) \leqslant \sum_{i=1}^{2} \omega_i \bar{c}_i \|P_U^\perp A_i\|_F^2 + \sum_{i=1}^{2} \omega_i \bar{n}_i\,\mathrm{Tr}(P_U^\perp P_i^\perp),$$

where

$$\bar{c}_i := \frac{1}{T-t_0}\int_{t_0}^{T} \frac{\lambda_{g,i}^{\max}(t)}{h^2(t)}\,\mathrm{d}t, \qquad \bar{n}_i := \frac{1}{T-t_0}\int_{t_0}^{T} \frac{\alpha^4(t)\sigma_i^4}{h^2(t)\tilde{h}_i(t)}\,\mathrm{d}t.$$

If $c_i = \bar{c}_i$ in $\Gamma_k(W)$, the first term is exactly the weighted joint residual

$$\sum_{i=1}^{2} \omega_i \bar{c}_i \|P_U^\perp A_i\|_F^2.$$

In particular, for $U = W_k$, where W_k is the top-k eigenspace of $M_{\mathrm{mix}} = \omega_1 \bar{c}_1 A_1 A_1^\top + \omega_2 \bar{c}_2 A_2 A_2^\top$, we have

$$\mathcal{O}_{\mathrm{mix}}(W_k) \leqslant \sum_{j=k+1}^{D} \lambda_j(M_{\mathrm{mix}}) + \sum_{i=1}^{2} \omega_i \bar{n}_i \operatorname{Tr}(P_{W_k}^{\perp} P_i^{\perp}).$$

For $U = V_1$, the same bound becomes

$$\mathcal{O}_{\mathrm{mix}}(V_1) \leqslant \omega_1 \bar{c}_1 \|P_{V_1}^{\perp} A_1\|_F^2 + \omega_2 \bar{c}_2 \|P_{V_1}^{\perp} A_2\|_F^2 + \sum_{i=1}^{2} \omega_i \bar{n}_i \operatorname{Tr}(P_{V_1}^{\perp} P_i^{\perp}).$$

Thus, W_k improves the output-side structural term whenever the reduction in the joint residual offsets the additional dimension-dependent costs introduced later through approximation and statistical complexity.

It remains to interpret the in-space term $\mathcal{I}_{\mathrm{mix}}(U)$. Define

$$Z := U^\top X_t, \qquad e_{i,U}(X_t, t) := \psi_{i,U}(Z, t) - U^\top G_{i,t}(X_t).$$

Then

$$\begin{aligned}
&f_{\mathrm{mix},U}^{\mathrm{comp}}(Z, t) - U^\top G_{\mathrm{mix},t}(X_t) \\
&= \sum_{i=1}^{2} \pi_i(X_t, t) e_{i,U}(X_t, t) + \sum_{i=1}^{2} (\pi_i^U(Z, t) - \pi_i(X_t, t)) \psi_{i,U}(Z, t) \\
&= \sum_{i=1}^{2} \pi_i(X_t, t) e_{i,U}(X_t, t) + (\pi_1^U(Z, t) - \pi_1(X_t, t))(\psi_{1,U}(Z, t) - \psi_{2,U}(Z, t)).
\end{aligned}$$

Consequently, by $\|a + b\|_2^2 \leqslant 2\|a\|_2^2 + 2\|b\|_2^2$ and $\|\sum_i \pi_i e_i\|_2^2 \leqslant \sum_i \pi_i \|e_i\|_2^2$, we have

$$\begin{aligned}
&\mathcal{I}_{\mathrm{mix}}(U) \\
&\leqslant \frac{2}{T - t_0} \int_{t_0}^{T} \frac{1}{h^2(t)} \sum_{i=1}^{2} \omega_i \mathbb{E}_{X_t \sim p_{i,t}} \left[\left\| \psi_{i,U}(U^\top X_t, t) - U^\top G_{i,t}(X_t) \right\|_2^2 \right] \mathrm{d}t \\
&\quad + \frac{2}{T - t_0} \int_{t_0}^{T} \frac{1}{h^2(t)} \mathbb{E}_{X_t \sim p_{\mathrm{mix},t}} \left[\left(\pi_1^U(U^\top X_t, t) - \pi_1(X_t, t) \right)^2 \left\| \psi_{1,U}(U^\top X_t, t) - \psi_{2,U}(U^\top X_t, t) \right\|_2^2 \right] \mathrm{d}t.
\end{aligned}$$

The first term is a component reconstruction error: it measures how well the compressed coordinate $U^\top X_t$ can recover the componentwise score target. The second term is a posterior-compression error: it measures the loss caused by replacing the full posterior weight $\pi_i(X_t, t)$ with the projected posterior weight $\pi_i^U(U^\top X_t, t)$. This posterior-compression term is specific to mixed training and has no analogue in the single-domain frozen-transfer bound.

Define the exact component reconstruction risk and posterior-compression term by

$$\mathfrak{R}_i(U) := \frac{1}{T - t_0} \int_{t_0}^{T} \frac{1}{h(t)^2} \mathbb{E}_{X_t \sim p_{i,t}} \|\psi_{i,U}(Z, t) - U^\top G_{i,t}(X_t)\|_2^2 \, \mathrm{d}t,$$

$$\mathfrak{P}(U) := \frac{1}{T - t_0} \int_{t_0}^{T} \frac{1}{h(t)^2} \mathbb{E}_{X_t \sim p_{\mathrm{mix},t}} \left[\left(\pi_1^U(Z, t) - \pi_1(X_t, t) \right)^2 \|\psi_{1,U}(Z, t) - \psi_{2,U}(Z, t)\|_2^2 \right] \mathrm{d}t.$$

The preceding calculation states $\mathcal{I}_{\mathrm{mix}}(U) \leqslant 2 \sum_i \omega_i \mathfrak{R}_i(U) + 2\mathfrak{P}(U)$. Combining it with the bound on $\mathcal{O}_{\mathrm{mix}}(U)$ and the comparator-to-ReLU inequality proves Theorem 10. The following regularity discussion explains how a ReLU approximation of the comparator can be constructed when the stated compact-domain time regularity and tail conditions hold.

Regularity of the componentwise comparators. We first record how the choice of U affects the regularity of the componentwise comparators. Recall that

$$\psi_{i,U}(z, t) := H_i(U) g_i(H_i(U)^\dagger z, t), \qquad H_i(U) := U^\top A_i, \quad i = 1, 2.$$

By Assumption 3, the map $f_i^*(\cdot, t)$ is β_i-Lipschitz uniformly in t. Since

$$g_i(y,t) = \frac{h(t)}{\tilde{h}_i(t)} f_i^*(y,t) + \left(1 - \frac{h(t)}{\tilde{h}_i(t)}\right) y,$$

and $0 \leqslant h(t)/\tilde{h}_i(t) \leqslant 1$, it follows that $g_i(\cdot, t)$ is $(\beta_i + 1)$-Lipschitz uniformly in t. Hence, for any $z, z' \in \mathbb{R}^m$,

$$\|\psi_{i,U}(z,t) - \psi_{i,U}(z',t)\|_2 \leqslant L_{\psi,i}(U)\|z - z'\|_2,$$

where

$$L_{\psi,i}(U) := \|H_i(U)\|_{\mathrm{op}} \|H_i(U)^\dagger\|_{\mathrm{op}} (\beta_i + 1).$$

Moreover, if

$$G_{i,0} := \sup_{t \in [t_0,T]} \|g_i(0,t)\|_2 < \infty,$$

then on the box $\|z\|_\infty \leqslant R$,

$$\|\psi_{i,U}(z,t)\|_2 \leqslant K_{\psi,i,U}(R) := \|H_i(U)\|_{\mathrm{op}} G_{i,0} + L_{\psi,i}(U)\sqrt{m}R.$$

Finally, define the compact-domain time regularity modulus

$$\tau_{\psi,i,U}(R) := \sup_{t \in [t_0,T],\, \|z\|_\infty \leqslant R} \|\partial_t \psi_{i,U}(z,t)\|_2.$$

Whenever g_i is differentiable in t, we have

$$\partial_t \psi_{i,U}(z,t) = H_i(U) \partial_t g_i(H_i(U)^\dagger z, t).$$

Thus, if

$$\tau_{g,i}(R) := \sup_{t \in [t_0,T],\, \|y\|_2 \leqslant R} \|\partial_t g_i(y,t)\|_2 < \infty,$$

then

$$\tau_{\psi,i,U}(R) \leqslant \|H_i(U)\|_{\mathrm{op}} \tau_{g,i}\left(\|H_i(U)^\dagger\|_{\mathrm{op}} \sqrt{m} R\right).$$

Regularity of the projected posterior weights. For $i = 1, 2$, the projected density $q_{i,U,t}$ admits the representation

$$q_{i,U,t}(z) = \int_{\mathbb{R}^{d_i}} \varphi_{m,\tilde{h}_i(t)}(z - \alpha(t) H_i(U) y) p_{i,z}(y)\, \mathrm{d}y,$$

where $\varphi_{m,\tilde{h}_i(t)}$ is the m-dimensional Gaussian density with covariance $\tilde{h}_i(t) I_m$. Since $\tilde{h}_i(t) = \alpha^2(t)\sigma_i^2 + h(t)$ is bounded away from zero on $[t_0, T]$, the Gaussian convolution makes $q_{i,U,t}$ positive and smooth in z. Under the positivity, smoothness, and tail assumptions on $p_{i,z}$, differentiation under the integral sign gives finite constants on each compact box:

$$m_{q,i,U}(R) := \inf_{t \in [t_0,T],\, \|z\|_\infty \leqslant R} q_{i,U,t}(z) > 0,$$

$$L_{q,i,U}^{(z)}(R) := \sup_{t \in [t_0,T],\, \|z\|_\infty \leqslant R} \|\nabla_z q_{i,U,t}(z)\|_2 < \infty,$$

$$L_{q,i,U}^{(t)}(R) := \sup_{t \in [t_0,T],\, \|z\|_\infty \leqslant R} |\partial_t q_{i,U,t}(z)| < \infty.$$

No equality between σ_1 and σ_2 is required; the constants above may depend on σ_1, σ_2 separately.

Since

$$\pi_i^U(z,t) = \frac{\omega_i q_{i,U,t}(z)}{\omega_1 q_{1,U,t}(z) + \omega_2 q_{2,U,t}(z)},$$

we have, for the two-component mixture,

$$\nabla_z \pi_1^U(z,t) = \pi_1^U(z,t)\pi_2^U(z,t)\left(\nabla_z \log q_{1,U,t}(z) - \nabla_z \log q_{2,U,t}(z)\right).$$

Therefore, on $\|z\|_\infty \leqslant R$,

$$\|\nabla_z \pi_1^U(z,t)\|_2 \leqslant L_{\pi,U}(R),$$

where one may take

$$L_{\pi,U}(R) := \frac{1}{4}\left(\frac{L_{q,1,U}^{(z)}(R)}{m_{q,1,U}(R)} + \frac{L_{q,2,U}^{(z)}(R)}{m_{q,2,U}(R)}\right).$$

Similarly,

$$|\partial_t \pi_1^U(z,t)| \leqslant \tau_{\pi,U}(R),$$

where

$$\tau_{\pi,U}(R) := \frac{1}{4}\left(\frac{L_{q,1,U}^{(t)}(R)}{m_{q,1,U}(R)} + \frac{L_{q,2,U}^{(t)}(R)}{m_{q,2,U}(R)}\right).$$

The same bounds hold for π_2^U, since $\pi_2^U = 1 - \pi_1^U$.

Regularity of the projected mixed comparator. We now combine the componentwise comparator regularity and the projected posterior regularity. Recall

$$f_{\mathrm{mix},U}^{\mathrm{comp}}(z,t) = \pi_1^U(z,t)\psi_{1,U}(z,t) + \pi_2^U(z,t)\psi_{2,U}(z,t).$$

Equivalently,

$$f_{\mathrm{mix},U}^{\mathrm{comp}}(z,t) = \psi_{2,U}(z,t) + \pi_1^U(z,t)\left(\psi_{1,U}(z,t) - \psi_{2,U}(z,t)\right).$$

For $\|z\|_\infty, \|z'\|_\infty \leqslant R$, using the bounds above gives

$$\|f_{\mathrm{mix},U}^{\mathrm{comp}}(z,t) - f_{\mathrm{mix},U}^{\mathrm{comp}}(z',t)\|_2 \leqslant L_{\mathrm{mix},U}(R)\|z - z'\|_2,$$

where one may take

$$L_{\mathrm{mix},U}(R) := L_{\psi,1}(U) + L_{\psi,2}(U) + L_{\pi,U}(R)\left(K_{\psi,1,U}(R) + K_{\psi,2,U}(R)\right).$$

Similarly, for $\|z\|_\infty \leqslant R$ and $t, t' \in [t_0, T]$,

$$\|f_{\mathrm{mix},U}^{\mathrm{comp}}(z,t) - f_{\mathrm{mix},U}^{\mathrm{comp}}(z,t')\|_2 \leqslant \tau_{\mathrm{mix},U}(R)|t - t'|,$$

where one may take

$$\tau_{\mathrm{mix},U}(R) := \tau_{\psi,1,U}(R) + \tau_{\psi,2,U}(R) + \tau_{\pi,U}(R)\left(K_{\psi,1,U}(R) + K_{\psi,2,U}(R)\right).$$

Thus the regularity constants entering the ReLU approximation and covering-number arguments depend on U through the component conditioning factors $H_i(U)^\dagger$, the projected posterior regularity constants $L_{\pi,U}(R), \tau_{\pi,U}(R)$, and the latent dimension m.

Comparison between $U = V_1$ and $U = W_k$. The preceding bounds make explicit how the projector choice affects the regularity of the projected mixed comparator. If $U = V_1$ and V_1 is well aligned with A_1, then $H_1(V_1) = V_1^\top A_1$ is well conditioned, and $L_{\psi,1}(V_1)$ is moderate. However, if A_2 is far from the source subspace, then $H_2(V_1) = V_1^\top A_2$ may be rank deficient or ill conditioned. In that case, the component-2 comparator can suffer from large reconstruction error, and the regularity constant $L_{\psi,2}(V_1)$ may deteriorate through $\|H_2(V_1)^\dagger\|_{\mathrm{op}}$.

By contrast, W_k is chosen to reduce the joint subspace residual

$$\Gamma_k(W) = \omega_1 c_1 \|P_W^\perp A_1\|_F^2 + \omega_2 c_2 \|P_W^\perp A_2\|_F^2.$$

Reducing the joint residual can improve component reconstruction when the relevant singular values of $H_i(W_k) = W_k^\top A_i$ improve, but this is not a separate monotonicity guarantee. In the containment case, $H_i(W_k)$ has orthonormal columns, all its singular values equal one, and the spatial regularity bound reduces to the intrinsic constant $\beta_i + 1$.

However, the posterior-weight regularity constants $L_{\pi,W_k}(R)$ and $\tau_{\pi,W_k}(R)$ are controlled by the projected densities $q_{i,W_k,t}$. These constants need not improve monotonically with k: a higher-dimensional projection may preserve more information for distinguishing the two components, but it may also create sharper posterior transitions or smaller projected-density lower bounds on a fixed box. Therefore, the main guaranteed benefit of W_k is the reduction of the joint structural residual and the potential improvement of component reconstruction conditioning, while the posterior-compression and dimension-dependent statistical costs must be tracked separately.

□

Proof of Corollary 11. Under containment, $H_i = W_k^\top A_i$ satisfies $H_i^\top H_i = I_{d_i}$, so $H_i^\dagger = H_i^\top$ and $H_i^\dagger W_k^\top X_t = A_i^\top X_t$. Consequently,

$$\psi_{i,W_k}(W_k^\top X_t, t) - W_k^\top G_{i,t}(X_t) = -\rho_i(t) W_k^\top P_i^\perp X_t.$$

Since $P_i^\perp X_t$ is centered Gaussian with covariance $\widetilde{h}_i(t) P_i^\perp$,

$$\mathfrak{R}_i(W_k) = \bar{n}_i \operatorname{Tr}(W_k^\top P_i^\perp W_k) = \bar{n}_i(k - d_i), \qquad \operatorname{Tr}(P_{W_k}^\perp P_i^\perp) = D - k.$$

Also $\Gamma(W_k) = 0$. Substitute these identities into Theorem 10. If $V_1 = A_1$, then $P_{V_1}^\perp A_1 = 0$, giving the stated expression for $\Gamma(V_1)$.

□

D Synthetic Experiments

D.1 Geometric Checks and Results

We illustrate the geometric mechanisms using Gaussian-subspace models, without neural-network optimization. The target structural-oracle expectations are analytic Gaussian moments; the explicit-comparator mixture expectations are estimated by stratified Monte Carlo. All reported risks are averaged over OU time. The complete settings and numerical checks are specified below. These evaluations isolate representation geometry rather than finite-sample learning rates or sampling quality.

Target-domain alignment and missing directions. We generate $X_0 = A_2 Z + 0.08\varepsilon$, with $D = 64$, $Z \sim \mathcal{N}(0, \Lambda)$, $\varepsilon \sim \mathcal{N}(0, I_{64})$, and eight latent variances evenly spaced from 0.7 to 1.3. Table 1 fixes $d_1 = d_2 = 8$ and varies all eight principal angles together. Over $[t_0, T] = [0.02, 4]$, the structural-oracle risk grows from 0.585 at alignment to 92.775 near orthogonality. A rank-four frozen subspace contained in the same target subspace instead incurs risk 47.241, despite all four defined principal angles being zero. This illustrates the missing-direction term $d_2 - r$, which is not captured by the observed principal angles alone.

Table 1: Time-averaged target-domain structural-oracle risk and the theoretical lower bound, using 961 time nodes. All eight principal angles equal the reported angle.

Common angle	$\sum_j \sin^2 \theta_j$	Lower bound	Structural oracle
$0°$	0.000	0.585	0.585
$30°$	2.000	16.336	24.900
$60°$	6.000	47.838	72.581
$89°$	7.998	63.569	92.775

Shared representations and ambient noise. We also consider $X_i = A_i Z_i + \sigma_i \varepsilon_i$, where $Z_i \sim \mathcal{N}(0, I_4)$, $\varepsilon_i \sim \mathcal{N}(0, I_{32})$, and all four source-target principal angles are $60°$. We use $[t_0, T] = [0.03, 4]$ and $p_{\mathrm{mix}} = (p_1 + p_2)/2$. Table 2 reports $\mathcal{L}_{\mathrm{mix}}(s^{\mathrm{comp}}_{\mathrm{mix},U})$, the risk of the explicit comparator defined in Appendix C.3. It is neither target-domain risk nor the structural-oracle risk obtained by optimizing all cores. The shared rank-k projector is a top-k eigenspace of $M_{\mathrm{mix}} = \frac{1}{2}\bar{c}_1 A_1 A_1^\top + \frac{1}{2}\bar{c}_2 A_2 A_2^\top$, with the coefficients prescribed by Theorem 10.

At rank four, the shared projector reduces this comparator risk by 34.3% under equal noise and 30.0% under unequal noise. Increasing the rank further reduces the evaluated risk in these settings. The union representation has zero signal residual, but the total comparator risk remains positive. Ambient noise contributes residual score error; with unequal component noise, projection can additionally lose information about the mixture posterior. These observations illustrate the separate terms in the mixed-risk bound rather than a guarantee of universal target-domain improvement.

D.2 Experimental Settings and Numerical Evaluation

Within each data-generating model, the latent Gaussian variables and ambient Gaussian noise are independent.

Table 2: Explicit-comparator mixture score risk: mean $\pm$ Monte Carlo standard error over 10 repetitions, using 385 time nodes. Equal noise: $(\sigma_1, \sigma_2) = (0.05, 0.05)$; unequal noise: $(0.02, 0.20)$.

Representation	Rank	Equal noise	Unequal noise
Source only	4	11.7218 ± 0.0043	13.3775 ± 0.0042
Shared	4	7.7045 ± 0.0028	9.3638 ± 0.0029
Shared	6	3.9955 ± 0.0008	5.6185 ± 0.0013
Union	8	0.022438 ± 0.000002	1.6568 ± 0.0003

Target-domain principal-angle check. Set $D = 64$, $d_1 = d_2 = 8$, $\sigma_2 = 0.08$, and

$$X_0 = A_2 Z + \sigma_2 \varepsilon, \qquad Z \sim \mathcal{N}(0, \Lambda), \quad \varepsilon \sim \mathcal{N}(0, I_{64}), \qquad \Lambda = \operatorname{diag}\left(0.7 + \frac{0.6(j-1)}{7}\right)_{j=1}^{8}.$$

For standard coordinate vectors e_j, take $V_1 = [e_1, \ldots, e_8]$ and $A_2 = [\cos\theta\, e_j + \sin\theta\, e_{8+j}]_{j=1}^{8}$. Thus all eight principal angles equal θ, not just the largest angle. The original sweep is $\theta \in \{0, 15, 30, 45, 60, 75, 89\}^\circ$; Table 1 displays four settings re-evaluated on a refined grid. The OU interval is $[t_0, T] = [0.02, 4]$.

Missing-direction check. Keep $D = 64$, $d_2 = 8$, Λ, σ_2, and the time interval unchanged. The implementation examines $d_1 \in \{4, 8, 12\}$ and maximum angles $\vartheta \in \{0, 30, 60, 75, 85, 89\}^\circ$. Writing $r = \min(d_1, 8)$, use the angle vector $\theta_j = (j-1)\vartheta/(r-1)$, $j = 1, \ldots, r$. First set $A_{\mathrm{can}} = [e_1, \ldots, e_8]$, and let the first r columns of V_1 be $\cos\theta_j e_j + \sin\theta_j e_{8+j}$. If $d_1 = 12$, append $e_{17}, \ldots, e_{20}$. To avoid aligning the covariance eigenvectors with the principal-vector basis, set $A_2 = A_{\mathrm{can}} Q$, where Q is the orthogonal factor from the unpivoted NumPy QR decomposition of an 8×8 standard-normal matrix drawn with `default_rng(20260719)`. This rotation leaves the target subspace and principal angles unchanged, but rotates the anisotropic covariance within it. The reported missing-direction comparison uses $\vartheta = 0$, $d_1 = 4$ versus $d_1 = 8$. For $d_1 = 4$, the four defined principal angles are zero, while four target directions are missing; this is subspace containment, not equality of the two subspaces.

Forward scores and structural-oracle evaluation. We use $\alpha(t) = e^{-t/2}$, $h(t) = 1 - e^{-t}$, and $\widetilde{h}_i(t) = h(t) + \alpha(t)^2 \sigma_i^2$. For a centered Gaussian component with latent covariance Λ_i,

$$\Sigma_{i,t} = \alpha(t)^2 A_i \Lambda_i A_i^\top + \widetilde{h}_i(t) I_D, \qquad \nabla \log p_{i,t}(x) = -\Sigma_{i,t}^{-1} x.$$

For a target covariance Σ_t and an orthonormal basis V, conditional expectation gives the structural minimizer

$$s_V^*(x,t) = -V(V^\top \Sigma_t V)^{-1} V^\top x - \frac{(I_D - VV^\top)x}{h(t)}.$$

Defining $H_t = \Sigma_t^{-1} - V(V^\top \Sigma_t V)^{-1} V^\top - (I_D - VV^\top)/h(t)$, its instantaneous score risk is $\operatorname{Tr}(H_t \Sigma_t H_t^\top)$. The code equivalently evaluates the orthogonal-output residual and the conditional-variance residual of $G_t(X_t) = X_t + h(t)\nabla \log p_{2,t}(X_t)$, and adds them. The lower bound is computed from the smallest eigenvalue of $M_{g,2}(t)$ in Theorem 2, not fitted to the oracle curve. All target expectations are analytic Gaussian moments; only the time integral is numerical. The nonzero aligned baseline is the ambient-noise mismatch of the fixed shortcut $-P_V^\perp x/h(t)$, not Monte Carlo error.

Shared-representation model and strategies. Set $D = 32$, $d_1 = d_2 = 4$, $\Lambda_1 = \Lambda_2 = I_4$, $A_1 = [e_1, \ldots, e_4]$, and $A_2 = [\cos(60^\circ) e_j + \sin(60^\circ) e_{4+j}]_{j=1}^{4}$. Each domain is $X_0^{(i)} = A_i Z_i + \sigma_i \varepsilon_i$, with $Z_i \sim \mathcal{N}(0, I_4)$, $\varepsilon_i \sim \mathcal{N}(0, I_{32})$. The mixture weights are always $\omega_1 = \omega_2 = 1/2$. The two noise scenarios are $(\sigma_1, \sigma_2) = (0.05, 0.05)$ and $(0.02, 0.20)$; the OU interval is $[0.03, 4]$. The source-only representation is $U = A_1$. The shared rank-four and rank-six representations use the leading eigenvectors of

$$M_{\mathrm{mix}} = \tfrac{1}{2}\bar{c}_1 A_1 A_1^\top + \tfrac{1}{2}\bar{c}_2 A_2 A_2^\top,$$

as returned by `numpy.linalg.eigh`, sorted by decreasing eigenvalue. The rank-eight representation is the same construction at $k = 8$; its column space is exactly $\operatorname{col}(A_1) + \operatorname{col}(A_2)$. These four

representations are population constructions, not estimated projectors. The rank-six cutoff lies in a repeated-eigenvalue block; no unique rank-six subspace or eigengap is assumed.

Spectral coefficients and explicit comparator. With $q_i(t) = \alpha(t)^2 + \widetilde{h}_i(t)$, the isotropic latent core and its second moment are

$$g_i(y,t) = \left(1 - \frac{h(t)}{q_i(t)}\right) y, \qquad \lambda_{g,i}^{\max}(t) = \frac{\alpha(t)^4(1+\sigma_i^2)^2}{q_i(t)}.$$

We integrate $\lambda_{g,i}^{\max}(t)/h(t)^2$ to obtain $\bar{c}_i$. The 25-node coefficients used to construct the representations are $(\bar{c}_1, \bar{c}_2) = (7.569859, 7.569859)$ for equal noise and $(7.552641, 7.877504)$ for unequal noise (rounded). We hold these representations fixed when refining the risk integral, thereby separating evaluation accuracy from representation selection. They are prescribed by the model and time interval, not tuned. Mixture weights remain $(1/2, 1/2)$; the normalized spectral weights $\omega_i \bar{c}_i / \sum_j \omega_j \bar{c}_j$ are approximately $(0.5, 0.5)$ and $(0.4895, 0.5105)$, respectively. For $H_i = U^\top A_i$, the implementation uses

$$\psi_{i,U}(z,t) = H_i g_i(H_i^\dagger z, t), \qquad s_{\mathrm{mix},U}^{\mathrm{comp}}(x,t) = \frac{U \sum_i \pi_i^U(U^\top x, t)\psi_{i,U}(U^\top x, t) - x}{h(t)}.$$

The full and projected posterior weights are evaluated from Gaussian densities with covariances $\Sigma_{i,t}$ and $U^\top \Sigma_{i,t} U$, respectively, using stable log-density normalization. The reference mixture score is exactly $\sum_i \pi_i(x,t)(-\Sigma_{i,t}^{-1}x)$. Thus the score functions are analytic, but the mixture risk expectation is estimated by Monte Carlo. Specifically, the reported mixture risk is

$$\mathcal{L}_{\mathrm{mix}}(s_{\mathrm{mix},U}^{\mathrm{comp}}) = \frac{1}{T - t_0} \int_{t_0}^{T} \mathbb{E}_{X_t \sim p_{\mathrm{mix},t}} \| s_{\mathrm{mix},U}^{\mathrm{comp}}(X_t, t) - \nabla \log p_{\mathrm{mix},t}(X_t) \|_2^2 \, \mathrm{d}t.$$

Time quadrature and Monte Carlo. For n nodes we set

$$t_\ell = t_0 (T/t_0)^{\ell/(n-1)}, \quad \ell = 0, \ldots, n-1, \qquad \frac{1}{T - t_0} \sum_{\ell=0}^{n-2} \frac{t_{\ell+1} - t_\ell}{2} \{ r(t_\ell) + r(t_{\ell+1}) \}.$$

The original target checks use $n = 121$; the original mixture check uses $n = 25$. No adaptive quadrature tolerance is specified: these are fixed-grid evaluations. For every time node in the mixture check, draw 2,500 independent noisy samples from each component, compute the sample-average squared score error separately, and combine the two averages with weights $1/2$. Each original 25-node mixture entry therefore uses 125,000 noisy vectors in total, with new draws at successive time nodes. The original source, shared rank-four, shared rank-six, and union evaluations use `default_rng` seeds 1234, 1235, 1238, and 1241, respectively, reset for each noise scenario. The nonconsecutive seeds reflect intervening empirical-projector configurations in the original script; those configurations are not included in the result tables. The final 385-node evaluation uses 1,925,000 noisy vectors per repetition and configuration; the grids and repetitions are specified below.

Refined final evaluation and uncertainty. The final target table and missing-direction comparison use 961 time nodes. Repeating these evaluations with 481 nodes changes every displayed target risk and lower bound by less than 0.009%; the original 121-node values differ from the 961-node values by less than 0.18%. For each mixture configuration, we evaluate 10 independent repetitions on each of the 25-, 49-, 97-, 193-, and 385-node grids, using repetition seeds 50000–50009 and 2,500 samples per component per node. The final mixture table uses the 385-node mean over these 10 repetitions, with error bars equal to the sample standard deviation across repetitions divided by $\sqrt{10}$. These are Monte Carlo standard errors of the reported mean, not standard deviations of individual runs, confidence intervals, or errors due to learning a representation. They do not include time-quadrature bias, which is checked separately by grid refinement. Coarser 25-node integration changes the mixture estimates by a few percent, motivating the refinement. The representations and risk definitions are unchanged, and the ordering of the four strategies is preserved on every evaluated grid. All grids are fixed in advance; the grid comparison is an empirical numerical check, not a certified quadrature-error bound. Table 3 reports representative coarse and refined evaluations. The largest relative difference between the 193- and 385-node mixture means is 0.051%, including Monte Carlo variability.

Table 3: Mixture-risk grid sensitivity, holding each representation fixed. Entries are means over 10 Monte Carlo repetitions. The reported risks use 385 nodes; uncertainty estimates are in Table 2.

Noise	Representation	25 nodes	193 nodes	385 nodes
Equal	Source, rank 4	11.9737	11.7243	11.7218
Equal	Shared, rank 4	7.8674	7.7055	7.7045
Equal	Shared, rank 6	4.0773	3.9957	3.9955
Equal	Union, rank 8	0.023361	0.022449	0.022438
Unequal	Source, rank 4	13.6820	13.3806	13.3775
Unequal	Shared, rank 4	9.5804	9.3655	9.3638
Unequal	Shared, rank 6	5.7615	5.6198	5.6185
Unequal	Union, rank 8	1.7105	1.6575	1.6568

Compute and software. The final re-evaluation used an AMD Ryzen 9 7940HX CPU, one BLAS thread, and a machine with 15.2 GiB usable RAM, running Windows 11, Python 3.12.7, NumPy 1.26.4, pandas 2.2.2, and threadpoolctl 3.5.0. The audit entry point uses `numpy.trapz` as the equivalent fallback for `numpy.trapezoid` on NumPy 1.26. Running the target grid checks took about 4 seconds. The 193- and 385-node mixture checks together took 235 seconds for equal noise and 232 seconds for unequal noise; the full refined command took 471 seconds and peaked at approximately 97 MiB process working memory. The additional 25-, 49-, and 97-node audit took 118 seconds, for approximately 589 seconds total numerical re-evaluation time. These are measured CPU runtimes for the reported synthetic evaluations and their numerical checks, not historical training times or the compute of other project experiments. The original exact wall times and CPU model were not retained; the available original metadata and the new measured audit are supplied separately. The evaluation uses CPU NumPy linear algebra; no GPU, external dataset, fitted network, optimizer, or train/test split is required. The original run metadata records Python 3.12.3 and NumPy 2.4.2 on Linux x86-64; the final re-evaluation environment is specified above.

The accompanying supplementary code archive contains the synthetic-data generators, numerical evaluation scripts, original and refined result files, and a README specifying dependencies and reproduction commands.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction are substantiated by the detailed theoretical analyses are presented in the paper. These sections clearly outline our contributions and scope, which are consistently reflected throughout the findings.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: In the last section, we discuss the limitations and propose several promising new directions.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: For all the theoretical results, we provide the full set of assumptions and complete proofs for them.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes].

   Justification: The appendix provides the experimental settings, risk evaluation procedures, and numerical implementation details needed to reproduce the reported results.

   Guidelines:

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

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

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

Answer: [Yes].

Justification: The supplementary code archive provides the synthetic-data generators, evaluation scripts, original and refined results, and installation and execution instructions. No external dataset or pretrained model is required.

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes].

Justification: Appendix D specifies the synthetic data distributions, subspace configurations, evaluation metrics, and numerical procedures.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes].

Justification: Table 2 reports Monte Carlo standard errors. Appendix D describes their computation and separately assesses the sensitivity of the results to time discretization.

Guidelines:

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

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

8. **Experiments compute resources**

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

   Answer: [Yes].

   Justification: The computational environment and approximate execution times are reported in Appendix D.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: This work fully complies with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: This is a theoretical work without addressing or having any societal impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: This work poses no such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [N/A]

    Justification: This work does not use existing assets.

    Guidelines:

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

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

13. **New assets**

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

    Answer: [Yes].

    Justification: The supplementary archive documents the synthetic-evaluation code, numerical outputs, dependencies, reproduction commands, and limitations. No new dataset or pretrained model is introduced.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing or research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: This work does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

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