UNDERSTANDING PREDICTION TARGET, TIME SAMPLING, AND MODEL CAPACITY IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Denoising diffusion models require several coupled design choices: what the network predicts, how timesteps are sampled, and how much capacity the architecture provides. Three recent empirical findings shed light on how to make these choices: (i) clean data prediction can have higher generation quality than noise or velocity prediction; (ii) sampling timesteps from a logit-normal distribution can outperform uniform sampling; and (iii) generation quality can peak at an intermediate model capacity. However, a theoretical understanding of these observations, and particularly of how the three choices interact, remains limited. In this paper, we show that theoretical counterparts of all three observations arise in a tractable linear model with Gaussian data and a time-gated predictor. Specifically, our spectral analysis reveals how prediction targets, timestep sampling, and model capacity jointly determine which spectral components of the data are represented by the model and, in turn, shape the generated distribution, a principle we refer to as spectral selection. This analysis further shows that the optimal design choices depend on the spectral structure of the data. We provide empirical support for our theoretical findings on synthetic and real data.

1 INTRODUCTION

Denoising diffusion models generate data by learning a transformation from noise to data, and have become central to generative modeling (Ho et al., 2020; Lipman et al., 2023; Liu et al., 2023). In the flow-matching formulation adopted throughout this paper, a clean sample x and Gaussian noise ϵ are interpolated at timestep $t \in [0, 1]$ as $z_t = tx + (1 - t)\epsilon$ and a network takes (z_t, t) as input. The resulting training procedure leaves several design choices.

We focus on three such choices: what the network predicts, how timesteps are sampled, and how much capacity the architecture has. The first is the *prediction target*: the network can be trained to predict the injected noise ϵ (ϵ-prediction) as popularized by diffusion models (Ho et al., 2020); the velocity $v = x - \epsilon$ that transports noise toward data (v-prediction) as standard in flow matching (Lipman et al., 2023); or the clean data x itself (x-prediction) (Li & He, 2026). The second is *timestep sampling*: the distribution from which t is drawn during training. Uniform sampling is the common default, while nonuniform alternatives such as logit-normal sampling can be used to concentrate training on particular ranges of timesteps. The third is *model capacity*: how much input–output structure the architecture can represent, often controlled by network width and depth.

Recent empirical studies have revealed several nontrivial effects involving subsets of these design choices. **Observation 1: x-prediction can improve generation quality over standard prediction targets.** Specifically, directly predicting clean data can remain effective for data in high-dimensional ambient spaces when model capacity is limited (Li & He, 2026; Lu et al., 2026). **Observation 2: Nonuniform timestep sampling can outperform uniform sampling.** In particular, distributions with an interior mode, such as logit-normal distributions, can improve generation quality over uniform sampling across different prediction targets (Esser et al., 2024; Yao et al., 2024). **Observation 3: Generation quality can peak at a moderate model capacity.** Specifically, introducing bottlenecks or choosing moderate network capacity can improve generation quality (Bao et al., 2023; Li & He, 2026). Our experiment in Figure 1 also reproduces all three qualitative trends in an image generation setting distinct from those of the original studies.

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

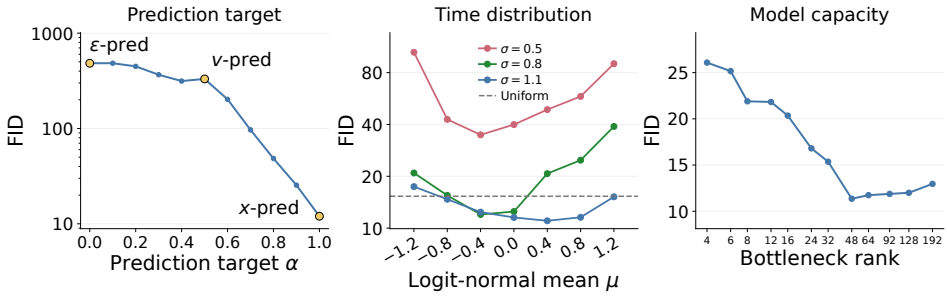


Figure 1: **Prediction target, timestep sampling, and model capacity sweeps.** FID-50K on CelebA-HQ across different design choices. In these experiments, x-prediction, nonuniform timestep sampling, and intermediate model capacity yield better generation quality.

Although these findings offer useful guidance, a theoretical understanding of these phenomena remains limited. Moreover, existing studies typically vary only subsets of the design choices (Bao et al., 2023; Esser et al., 2024; Li & He, 2026) and therefore do not reveal how all three interact. This raises a fundamental question: *How do prediction target, timestep sampling, and model capacity jointly affect generation?*

A key to understanding this interaction is finite model capacity. Under the manifold hypothesis, clean data may concentrate near low-dimensional structure (Carlsson, 2009; Pope et al., 2021), whereas noise and velocity contain variation throughout the ambient space. Different prediction targets may therefore require different amounts of capacity to represent well. When the model lacks sufficient capacity to represent all of this structure, minimizing the training loss favors components whose representation yields a larger loss reduction. Because these loss reductions can vary across timesteps as the input z_t changes with t, timestep sampling can change which components are prioritized.

Building on this intuition, we study a tractable model with Gaussian data and a linear time-gated predictor. In this setting, the data structure is characterized by its covariance spectrum. This allows us to formalize the preceding intuition as a principle we refer to as *spectral selection*: three design choices jointly determine which eigendirections are selected by the training loss minimizer and, in turn, shape the generated distribution. Using this spectral characterization, we derive theoretical counterparts to Observations 1–3 and show how preferred design choices vary with the spectral structure, particularly with data dimensionality.

Our contributions are summarized as follows.

- **Spectral characterization of training and generation.** In Section 4, we define a *spectral score* that measures the reduction in training loss achieved by optimally representing the target along each eigendirection of the data covariance. We show that these scores determine which directions the loss minimizer represents. We then characterize how its generated distribution depends on the three design choices.
- **Design choices in terms of data dimensionality.** In Section 5, we analyze how optimal design choices depend on data dimensionality and derive theoretical counterparts to Observations 1–3. We further show that preference can shift from x-prediction to v-prediction as dimensionality increases.
- **Beyond the training loss minimizer.** In Section 6, we extend the analysis beyond the training loss minimizer by quantifying how training loss translates into generation error and how timestep sampling can control this amplification.
- **Empirical validation.** In Section 7, we provide empirical support for the theoretical findings across synthetic and real data.

2 Related Work

Prediction target, model capacity, and data dimensionality. Li & He (2026) advocate x-prediction for high-dimensional pixel-space generation, motivated by the manifold hypothesis discussed above. They also observe the U-shaped dependence of generation error on bottleneck dimension. Gagneux et al. (2026) further find empirically that the preferred prediction target depends on data dimensionality: smaller intrinsic dimension can favor x-prediction, whereas larger intrinsic dimension can

favor v-prediction. Our analysis is consistent with these findings and shows how data dimensionality affects target preference and why additional capacity can worsen generation.

Jin & Wang (2026) study a continuous family of prediction targets whose parameter interpolates between ϵ-, v-, and x-prediction, corresponding to the target family analyzed in our work. Their main analysis considers whitened low-dimensional data embedded in a high-dimensional ambient space and a time-independent predictor, and compares prediction targets through regression loss. Our analysis differs in three main respects: we introduce bottleneck width as an explicit model-capacity constraint with a time-dependent predictor; we jointly analyze prediction target and timestep sampling; and we compare prediction targets through the Wasserstein error of the generated endpoint distribution, since equal regression errors can correspond to different generation errors; see Section 6.

Time sampling and loss parametrization. Nonuniform timestep sampling is widely used to place greater emphasis on certain timestep ranges during training (Kingma & Gao, 2023; Hang et al., 2023), and logit-normal sampling in particular has been shown to improve generation over uniform sampling (Esser et al., 2024; Yao et al., 2024). Loss parametrization provides another way to control the relative importance of timesteps during training (Li & He, 2026; Gagneux et al., 2026). With a fixed prediction target, evaluating the loss in another prediction parameterization induces a time-dependent reweighting of the loss. Our analysis shows that the resulting timestep weighting can affect generation in two ways: when the training loss remains above its minimum, it controls how residual errors along represented eigendirections are penalized across timesteps; even under exact loss minimization, it can change which directions the learned predictor represents.

3 Problem Setup

We introduce a tractable model in which prediction target, timestep sampling, and model capacity can be studied explicitly. Following the flow-based formulation, the setup uses a Gaussian source, a linear interpolation path, a regression training loss, and ODE-based generation through the induced velocity field (Lipman et al., 2023; Albergo & Vanden-Eijnden, 2023; Liu et al., 2023). Similar Gaussian and linearized settings have been used to isolate the effects of data dimensionality, loss weighting, and prediction parameterization in diffusion and flow models (Gagneux et al., 2026; Jin & Wang, 2026).

Data. Let $x \in \mathbb{R}^D$ denote a data sample. We assume a centered Gaussian data distribution

$$x \sim \mathcal{N}(0, H), \qquad H = \textstyle\sum_{i=1}^{D} \lambda_i e_i e_i^\top, \qquad 1 = \lambda_1 > \lambda_2 > \cdots > \lambda_D > 0, \tag{1}$$

where $\{e_i\}_{i=1}^D$ is an orthonormal eigenbasis of H and λ_i are the covariance eigenvalues. The normalization $\lambda_1 = 1$ fixes the overall scale. Gaussian data models are a standard probe for studying diffusion and flow matching because the Bayes predictor, spectral decomposition, and generated covariance can be computed in closed form (Gagneux et al., 2026; Jin & Wang, 2026).

Predictor class and bottleneck. In diffusion and flow models, the timestep typically enters the network through feature-wise modulation: a time embedding is used to compute scaling factors and additive biases for hidden representation of the network (Perez et al., 2018). Adaptive normalization layers, standard in U-Net backbones (Dhariwal & Nichol, 2021) and diffusion transformers (Peebles & Xie, 2023), are a representative instance. We study a linear gated predictor that retains the multiplicative part of this conditioning while discarding the additive shift and the normalization:

$$f(z; t) = U\big(s(t) \odot V^\top z\big) = U\,\mathrm{diag}(s(t))\,V^\top z, \qquad s(t) = E\phi_K(t), \tag{2}$$

where $U, V \in \mathbb{R}^{D\times m}$, $E \in \mathbb{R}^{m\times K}$ are trainable matrices, with U and V shared across timesteps, and $\phi_K(t) \in \mathbb{R}^K$ is a time-embedding vector. Since our focus is the bottleneck m rather than the expressivity of the time embedding, we first allow $s(t)$ to be any continuous function mapping $(0, 1) \to \mathbb{R}^m$ and analyze the predictor class

$$\mathcal{F}_m = \big\{(z, t) \mapsto U\big(s(t) \odot V^\top z\big) \,\big|\, U, V \in \mathbb{R}^{D\times m},\ s \in C((0, 1); \mathbb{R}^m)\big\}. \tag{3}$$

We control model capacity through the bottleneck dimension m: the predictor $U\mathrm{diag}(s(t))V^\top$ has rank at most m at each timestep. This setting allows us to characterize a training loss minimizer in closed form. Later, in Section 6, we also study predictors with positive excess loss.

Prediction target and induced velocity. Let $\epsilon \sim \mathcal{N}(0, I_D)$ be independent Gaussian noise. For $t \in [0, 1]$, the network takes (z_t, t) as input and predicts a target y^α for $\alpha \in [0, 1]$ where

$$z_t = tx + (1 - t)\epsilon, \qquad y^\alpha = \alpha x - (1 - \alpha)\epsilon. \tag{4}$$

The choices $\alpha = 1$, $\alpha = 1/2$, and $\alpha = 0$ correspond to x-prediction, v-prediction up to a factor $1/2$, and ϵ-prediction up to sign, respectively. This target family is also considered by Jin & Wang (2026). For $t \in (0, 1)$, define $d_\alpha(t) := \alpha(1 - t) + (1 - \alpha)t$. Then the pair (z_t, y^α) determines the velocity, so any predictor f of y^α induces a velocity field by time-dependent linear transformation:

$$v = x - \epsilon = \frac{y^\alpha - (2\alpha - 1)z_t}{d_\alpha(t)}, \qquad v_f^\alpha(z; t) = \frac{f(z; t) - (2\alpha - 1)z}{d_\alpha(t)}. \tag{5}$$

Training loss and timestep sampling. The target used to compute training loss can be chosen differently of the model's prediction target (Salimans & Ho, 2022; Li & He, 2026; Gagneux et al., 2026). We show below that the effect of this choice of training loss can be incorporated into timestep sampling. Let parameters $\alpha, \beta \in [0, 1]$ specify the prediction target and the target used in the loss, respectively. To compute the training loss against y^β, we convert to the target representation:

$$f^\beta(z_t; t) = (2\beta - 1)z_t + d_\beta(t)v_f^\alpha(z_t; t), \qquad y^\beta = (2\beta - 1)z_t + d_\beta(t)v. \tag{6}$$

Under a timestep sampling density $w : (0, 1) \to (0, \infty)$ satisfying $\int_0^1 w(t)\, dt = 1$, β- and v-loss are

$$\mathcal{L}_{w,\beta}(f) := \int_0^1 w(t)\mathbb{E}\|f^\beta(z_t; t) - y^\beta\|^2\, dt = \int_0^1 \left(w(t)d_\beta(t)^2\right)\mathbb{E}\|v_f^\alpha(z_t; t) - v\|^2\, dt, \tag{7}$$

$$\mathcal{L}_{w,v}(f) := \int_0^1 w(t)\mathbb{E}\|v_f^\alpha(z_t; t) - v\|^2\, dt, \tag{8}$$

where the expectation is over independent $x \sim \mathcal{N}(0, H)$ and $\epsilon \sim \mathcal{N}(0, I_D)$; we study the population objective, isolating the effects of the three choices from finite-sample estimation error.

Since $w(t)d_\beta(t)^2$ is always normalizable, any β-loss can be converted to v-loss by reweighting timestep sampling density, up to a normalization constant.[1] We therefore study both the choice of training loss and timestep sampling by varying the sampling density under v-loss. To simplify notation, we omit the subscript v and write $\mathcal{L}_w$ for $\mathcal{L}_{w,v}$ unless otherwise stated.

Generated distribution. For any predictor $f \in \mathcal{F}_m$, generation starts from Gaussian noise and follows the induced velocity field. Let $(Z_t)_{t\in[0,1]}$ be the generated trajectory, continuous on $[0, 1]$, differentiable on $(0, 1)$, and satisfying

$$Z_0 \sim \mathcal{N}(0, I_D), \qquad \frac{dZ_t}{dt} = v_f^\alpha(Z_t; t) = \frac{f(Z_t; t) - (2\alpha - 1)Z_t}{d_\alpha(t)}, \qquad t \in (0, 1). \tag{9}$$

We denote the law of Z_t by $p_t(f)$. Then $p_0(f) = \mathcal{N}(0, I_D)$, and $p_1(f)$ is the generated endpoint distribution. Because the induced velocity field is linear, $p_t(f)$ remains centered Gaussian, so the generated distribution is fully determined by its covariance.

Evaluation distribution. To allow for nuisance variation in the training distribution, we evaluate generation against the clean distribution $p_{\text{clean}} := \mathcal{N}(0, H_{\text{clean}})$ where $H_{\text{clean}} := \sum_{i=1}^d \lambda_i e_i e_i^\top$ for $1 \le d \le D$. The first d covariance directions are treated as data signal, while the remaining directions represent nuisance variation. Setting $d = D$ recovers evaluation against the training distribution.

4 Spectral Selection: From Training Loss Minimum to Generation

In this section, we characterize the effects of the three design choices on a training loss minimizer through the principle of *spectral selection*. We first show that a width-m minimizer selects the m covariance eigendirections that yield the largest reductions in training loss. These direction-wise reductions are quantified by what we call the *spectral score*, which depends on the data spectrum, prediction target, and timestep sampling. We then show that selected directions recover the corresponding data variances at the end of generation, whereas omitted directions have a default variance determined by the prediction target. All proofs in this section are deferred to Appendix B.

[1] The converse need not hold. For $\beta = 1$, *i.e.,* x-loss, uniform sampling under v-loss would require a density proportional to $(1 - t)^{-2}$, which is not normalizable for $t \in (0, 1)$.

4.1 Spectral Score: Loss Reduction for Representing Each Eigendirection

To quantify the reduction in training loss from representing each eigendirection, we first characterize the Bayes-optimal velocity predictor. The Bayes-optimal predictors for other targets can be obtained directly from it by a time-dependent linear transformation.

Proposition 1 (Bayes-optimal velocity). *Fix $t \in (0,1)$. Among all measurable velocity fields $g : \mathbb{R}^D \to \mathbb{R}^D$, the minimizer of $\mathbb{E}\big[\|g(z_t) - v\|^2\big]$ is the conditional mean $v^\star(z;t) = \mathbb{E}[v \mid z_t = z]$, given by*

$$v^\star(z;t) = \big(tH - (1-t)I_D\big)\big(t^2 H + (1-t)^2 I_D\big)^{-1} z = \sum_{i=1}^{D} \frac{t\lambda_i - (1-t)}{t^2\lambda_i + (1-t)^2}(z^\top e_i)e_i. \tag{10}$$

The corresponding Bayes-optimal loss is $\mathcal{L}_w^{\text{Bayes}} := \int_0^1 w(t)\,\mathbb{E}\big[\|v^\star(z_t;t) - v\|^2\big]dt$.

The conditional mean $v^\star(z;t)$ is $\mathrm{Cov}(v,z)\mathrm{Var}(z)^{-1}z$, since (z_t, v) is jointly Gaussian, and therefore decomposes along the covariance eigendirections. Although $\mathcal{F}_m$ is not constrained to align with this basis, we show below that a training loss minimizer in $\mathcal{F}_m$ selects m of these eigendirections. This motivates quantifying the reduction in the loss obtained by representing each eigendirection.

Definition 1 (Spectral score). For a prediction target y^α with $\alpha \in [0,1]$ and a time-weighting density w, define the spectral score by

$$S_w^\alpha(\lambda) := \int_0^1 w(t)\frac{\big[\alpha t\lambda - (1-\alpha)(1-t)\big]^2}{d_\alpha(t)^2\big[t^2\lambda + (1-t)^2\big]}\,dt. \tag{11}$$

For each covariance eigendirection, the spectral score measures the reduction in v-loss obtained by using the Bayes-optimal prediction rather than omitting that direction. Note that this value depends on the covariance eigenvalue λ_i, the prediction target α, and the timestep sampling density w.

4.2 Training: Selecting Eigendirections with the Largest Spectral Scores

We next show that a training loss minimizer selects the m eigendirections with the largest scores.

Theorem 1 (Width-m training loss minimizer). *Fix $\alpha \in [0,1]$, a timestep sampling density w, and a bottleneck width $m \le D$. Assume $S_w^\alpha(\lambda_i) < \infty$ for all i. Let $I^\star$ index the m largest spectral scores of $\{S_w^\alpha(\lambda_i)\}_{i=1}^D$. Then a minimizer of $\mathcal{L}_w$ over $\mathcal{F}_m$ is*

$$f^\star(z;t) = \sum_{i\in I^\star} \frac{\alpha t\lambda_i - (1-\alpha)(1-t)}{t^2\lambda_i + (1-t)^2}(z^\top e_i)e_i, \tag{12}$$

realized in $\mathcal{F}_m$ by $U^\star = V^\star = [e_i]_{i\in I^\star}$ and the time-gates $s_i^\star(t) = \frac{\alpha t\lambda_i - (1-\alpha)(1-t)}{t^2\lambda_i + (1-t)^2}$, for $i \in I^\star$.

$$\min_{f\in\mathcal{F}_m} \mathcal{L}_w(f) = \mathcal{L}_w^{\text{Bayes}} + \textstyle\sum_{i\notin I^\star} S_w^\alpha(\lambda_i), \tag{13}$$

where $\mathcal{L}_w^{\text{Bayes}}$ is defined in Proposition 1.

Thus, the spectral scores determine which directions are selected, while the bottleneck width determines how many are selected. Along these directions, the minimizer matches the Bayes-optimal prediction, while each omitted direction adds its spectral score to the training loss.

4.3 Generation: Endpoint Distribution along Selected and Omitted Directions

Theorem 1 identifies the eigendirections selected by the training loss minimizer. We now solve the corresponding generation ODE induced by this predictor to characterize the generated distribution in closed form along both selected and omitted directions.

Theorem 2 (Generated covariance of the training loss minimizer). *Let $\alpha \in (0,1]$,[2] and let $f^\star$ training loss minimizer specified in Theorem 1 with selected index set $I^\star$. Then the ODE induced by $f^\star$*

[2] For $\alpha = 0$, an omitted direction satisfies $dZ_i(t)/dt = Z_i(t)/t$, which is singular at $t = 0$ and incompatible with $Z_i(0) \sim \mathcal{N}(0,1)$.

preserves Gaussianity, with covariance diagonal in the eigenbasis of H. Specifically, for $t \in [0,1]$, $p_t(f^\star) = \mathcal{N}\big(0, \sum_{i=1}^{D} \lambda_i^{\mathrm{gen}}(t) e_i e_i^\top\big)$, where

$$\lambda_i^{\mathrm{gen}}(t) = \begin{cases} t^2\lambda_i + (1-t)^2, & i \in I^\star, \\ \big((\alpha(1-t) + (1-\alpha)t)/\alpha\big)^2, & i \notin I^\star. \end{cases} \tag{14}$$

In particular, $\lambda_i^{\mathrm{gen}}(1) = \lambda_i$ for $i \in I^\star$, while $\lambda_i^{\mathrm{gen}}(1) = \sigma_\alpha^2$ otherwise, where $\sigma_\alpha := (1-\alpha)/\alpha$.

Thus, selected directions recover the corresponding data variances at the end of generation, whereas all omitted directions have the same target-dependent variance σ_α^2. Therefore, at the loss minimizer, the prediction target and timestep sampling affect generation through spectral selection, and the prediction target further affects generation by determining the variance along omitted directions.

For any prediction target $\alpha \in (0,1)$, the ordering of eigendirections by their direction-wise loss reductions can change across timesteps. For example, near $t = 1$, the loss reduction increases with λ, whereas at $t = 1-\alpha$ it decreases with λ. Therefore, changing the timestep sampling distribution can change the spectral score ranking and, consequently, which eigendirections are selected.

Remark 1 (Time invariance of spectral selection for x-prediction)**.** For x-prediction with finite spectral scores, *i.e.,* $\alpha = 1$, the spectral score is increasing in λ for any timestep sampling density w. Thus, a training loss minimizer in $\mathcal{F}_m$ selects the m leading covariance eigendirections independently of w. In practice, however, timestep sampling affects generation quality under x-prediction, as observed in Figure 1 and prior work (Li & He, 2026), suggesting that timestep sampling can affect generation beyond spectral selection. Since practical training need not attain the exact minimizer, later in Section 6 we quantify how the effects of suboptimality on generation vary across timesteps; see also Oko et al. (2023) for a related analysis of how errors at different timesteps propagate to the generated distribution.

5 Optimal Design Choices in Terms of Data Dimensionality

In this section, we apply the preceding spectral characterization to analyze how data dimensionality affects the optimal design choices, recovering theoretical counterparts to Observations 1–3. We first show that suitable logit-normal timestep sampling achieves the smallest generation error for each fixed combination of prediction target and bottleneck width. We then compare these attainable errors to characterize the optimal target and width. All proofs are deferred to Appendix C.

5.1 Optimal Timestep Sampling and Endpoint-Optimal Selection

Motivated by the empirical benefits of logit-normal timestep sampling, we ask whether a suitable sampling density can make a training loss minimizer select directions that minimize generation error. For a given prediction target and bottleneck width, let $f^\star$ be a training loss minimizer from Theorem 1 with selected index set $I^\star$. By Theorem 2, selected directions recover their observed variances, whereas omitted directions have the target-dependent variance σ_α^2. Thus, endpoint generation error relative to the clean distribution p_{clean} is

$$W_2^2\big(p_1(f^\star), p_{\mathrm{clean}}\big) = \mathcal{E}_d(I^\star, \alpha) := \sum_{\substack{i \le d, \\ i \notin I^\star}} \Big(\sqrt{\lambda_i} - \sigma_\alpha\Big)^2 + \sum_{\substack{i > d, \\ i \in I^\star}} \lambda_i + \sum_{\substack{i > d, \\ i \notin I^\star}} \sigma_\alpha^2. \tag{15}$$

For fixed α and m, we call an index set *endpoint-optimal* if it minimizes $\mathcal{E}_d(I, \alpha)$ over all size-m index sets I. Although a training loss minimizer need not select such a set, the following theorem shows that suitable logit-normal timestep sampling can make the minimizer select an endpoint-optimal set.

Theorem 3 (Endpoint-optimal selection by timestep sampling)**.** *Let $\alpha \in (0,1)$,*[3] *and let $m < D$. Then there exists a timestep sampling density $w^{(\alpha,m)}$ and a training loss minimizer $f^\star \in \arg\min_{f \in \mathcal{F}_m} \mathcal{L}_{w^{(\alpha,m)}}(f)$ whose selected index set is endpoint-optimal. Consequently,*

$$W_2^2\big(p_1(f^\star), p_{\mathrm{clean}}\big) = \min_{|I|=m} \mathcal{E}_d(I, \alpha). \tag{16}$$

Moreover, $w^{(\alpha,m)}$ can be chosen as a sufficiently concentrated logit-normal density.

[3] For $\alpha = 1$, every timestep density with finite spectral scores selects $\{1, \dots, m\}$; this set is endpoint-optimal exactly when $m \le d$ or $m = D$.

The construction concentrates sampling near a timestep whose instantaneous scores favor an endpoint-optimal set. We would like to note that this logit-normal density is a sufficient construction for realizing endpoint optimal set, rather than a characterization of all timestep distributions that attain it.

5.2 Optimal Prediction Target

We next compare the minimum achievable generation errors across prediction targets α for each fixed bottleneck width, with timestep sampling optimized separately for each target.

Theorem 4 (Optimal prediction target). *For $m \le d$, there exist a prediction target $\alpha_m^\star$, a timestep sampling density $w^{(\alpha_m^\star, m)}$, and a corresponding training loss minimizer $f^\star \in \arg\min_{f \in \mathcal{F}_m} \mathcal{L}_{w^{(\alpha_m^\star, m)}}(f)$ with selected index set $I^\star$ attaining the minimum generation error:*

$$W_2^2\big(p_1(f^\star), p_{\text{clean}}\big) = \min_{\alpha \in (0,1]} \min_{|I|=m} \mathcal{E}_d(I, \alpha). \tag{17}$$

Moreover, the optimal prediction target $\alpha_m^\star$ is given by

$$\alpha_m^\star = \begin{cases} \dfrac{D-m}{D-m+\sum_{i \le d, i \notin I^\star} \sqrt{\lambda_i}}, & m < d, \\ 1, & m = d. \end{cases} \tag{18}$$

On the other hand, for $d < m < D$, the infimum generation error over training loss minimizers equals $\min_{\alpha \in (0,1]} \min_{|I|=m} \mathcal{E}_d(I, \alpha)$. This infimum is not attained, but can be approached as $\alpha \uparrow 1$.

The preferred target depends on the data spectrum and on model capacity relative to the clean signal dimension d. When $m < d$, some signal directions must remain omitted, so targets with $\alpha < 1$ help recover part of the missing signal variance. When $m = d$, x-prediction ($\alpha = 1$) exactly recovers the clean distribution. For $d < m < D$, targets close to x-prediction keep the default variance small while allowing suitable timestep sampling to select nuisance directions with small variances.

To see more clearly how data dimensionality affects the optimal target, consider the power-law spectrum $\lambda_i = i^{-\gamma}$ and evaluate generation against the training distribution, *i.e.,* $d = D$. The power-law spectra have been observed in real data and used in theoretical analyses (Maloney et al., 2022; Worschech & Rosenow, 2025). Here, the decay exponent γ controls the effective data dimensionality: faster decay concentrates variance in fewer directions. For fixed $1 \le m \le D - 2$,[4] the optimal target approaches x-prediction as $\gamma \to \infty$, when variance concentrates in the leading direction, and v-prediction as $\gamma \downarrow 0$, when variance becomes spread across all directions in the ambient space.

5.3 Optimal Model Capacity

We next compare the minimum attainable generation errors across bottleneck widths m for each fixed prediction target α, with timestep sampling optimized.

Theorem 5 (Optimal model capacity). *Let $\alpha \in (0, 1]$. There exist a bottleneck width $m_\alpha^\star$, a timestep sampling density $w(\alpha, m_\alpha^\star)$, and a corresponding training loss minimizer $f^\star \in \arg\min_{f \in \mathcal{F}_{m_\alpha^\star}} \mathcal{L}_{w(\alpha, m_\alpha^\star)}(f)$ attaining the minimum generation error:*

$$W_2^2\big(p_1(f^\star), p_{\text{clean}}\big) = \min_{1 \le m \le D} \min_{|I|=m} \mathcal{E}_d(I, \alpha). \tag{19}$$

Moreover, the optimal width can be chosen as

$$m_\alpha^\star = d + \#\{i > d : \lambda_i < \sigma_\alpha^2\}, \tag{20}$$

where $\#$ denotes cardinality. The minimum generation error equals $\sum_{i=d+1}^{D} \min\{\lambda_i, \sigma_\alpha^2\}$.

The optimal width reflects the direction-wise benefit of representation. Representing a signal direction removes its contribution to generation error, whereas representing a nuisance direction is beneficial only when $\lambda_i < \sigma_\alpha^2$. For x-prediction, $\sigma_1 = 0$: generation error decreases to zero as the width increases to d, then increases beyond d as nuisance directions are represented. In contrast, for v-prediction, $\sigma_{1/2}^2 = 1$, so the optimized error is nonincreasing in width and full capacity is optimal, under our spectral normalization.

[4] The restriction $m \le D - 2$ excludes cases in which the optimal target need not be unique: every target is optimal when $m = D$, while for $m = D - 1$, the generated variance σ_α^2 along the single omitted direction can exactly match the corresponding covariance eigenvalue for a suitable α.

6 Generation Beyond Exact Minimization

The preceding analysis shows how timestep sampling affects generation through spectral selection at exact minimizers of the training loss. For x-prediction, however, this spectral selection is independent of timestep sampling (Remark 1), even though generation quality varies with sampling in practice (Li & He, 2026, Table 3). In this section, we study how time sampling affects generation beyond spectral selection by considering predictors with positive excess training loss, as can arise from approximation error due to finite-dimensional time embeddings or optimization error due to incomplete training, including early stopping. We focus on small excess loss and prediction targets near x-prediction.

For clarity, we consider $d = D$, so generation is evaluated against the training distribution $\mathcal{N}(0, H)$. To study how prediction errors propagate during generation, we analyze predictors that retain the spectral structure of the exact minimizers in Theorem 1. Both the choice of directions and the predictions along them may be suboptimal. This spectral structure allows us to trace how errors along each direction affect its endpoint variance. Specifically, we consider *spectral predictors*

$$f(z;t) = \sum_{i \in I} s_i^\alpha(t)(z^\top e_i)e_i, \qquad |I| \le m.$$

Here, I indexes the directions used by the predictor and is fixed over time. Each continuous function $s_i^\alpha : (0,1) \to \mathbb{R}$ scales the input component $z^\top e_i$ at timestep t.

At an exact loss minimizer, generation recovers the data variances λ_i along each direction $i \in I$. When $s_i^\alpha(t)$ differs from its Bayes-optimal value, the resulting velocity error accumulates along the generation trajectory and can change the endpoint variance. Timestep sampling determines how strongly these errors are penalized during training. We therefore compare sampling densities by the worst-case generation error under a common excess-loss budget η, measured relative to each density's minimum $\mathcal{L}_w^\star := \min_{g \in \mathcal{F}_m} \mathcal{L}_w(g)$ at the fixed target α.

Theorem 6 (Optimal timestep sampling near x-prediction)**.** *Let $1 \le m < D$. There exist constants $\alpha_0 \in (0,1)$ and $\eta_0 > 0$, depending only on $\{\lambda_i\}_{i=1}^D$ and m, such that, for every $\alpha \in (\alpha_0, 1)$ and $0 < \eta < \eta_0$, the density*

$$w^\star(t) = \frac{[t^2 + (1-t)^2]^{-1/2}}{\int_0^1 [s^2 + (1-s)^2]^{-1/2}\, ds} \tag{21}$$

uniquely minimizes the worst-case generation error

$$\sup_{\substack{f \in \mathcal{F}_m,\ f \text{ spectral} \\ \mathcal{L}_w(f) - \mathcal{L}_w^\star \le \eta}} W_2^2\big(p_1(f), \mathcal{N}(0,H)\big) \tag{22}$$

over all positive continuous timestep sampling densities w satisfying $\int_0^1 w(t)^{-1}\, dt < \infty$. Under $w^\star$, every spectral predictor with excess loss at most η has represented set $I = \{1, \dots, m\}$.

When excess training loss remains, $w^\star$ assigns greater weight to timesteps where errors can have a larger impact on generation, thereby minimizing worst-case generation error. The optimal density $w^\star$ peaks at $t = 1/2$, and this interior peak is qualitatively consistent with the empirical preference of logit-normal sampling (Esser et al., 2024; Yao et al., 2024). Matching this optimal density under other training loss choices such as the x-loss $\mathcal{L}_{w,1}$ or the ϵ-loss $\mathcal{L}_{w,0}$ would require non-normalizable sampling densities,[5] which may help explain the empirical preference for v-loss (Li & He, 2026). Proofs are deferred to Appendix D.

7 Experiments

In this section, we empirically test the theoretical findings on synthetic data and examine whether qualitative trends persist on real data. Experimental setup and details are provided in Appendix E.

Empirical training and generation. We examine whether numerical training recovers the spectral selection and generated covariance characterized in Section 4. The learned subspace aligns with the eigendirections having the largest spectral scores $S_w^\alpha(\lambda_i)$ as discussed in Theorem 1, across the tested configurations. The generated covariance is also consistent with Theorem 2, recovering the

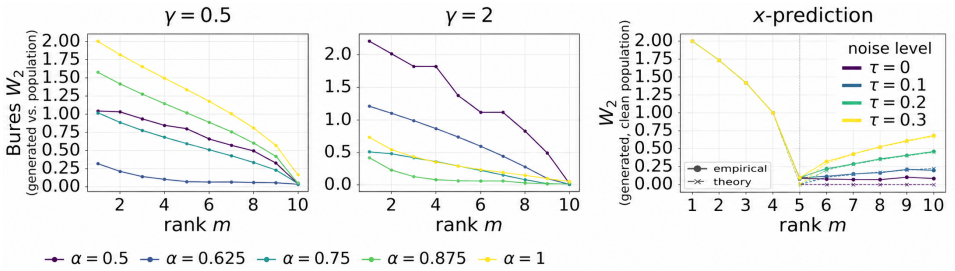


Figure 2: **Design choices across data dimensionality.** The two left panels show that faster spectral decay favors more x-like targets, while slower decay favors more v-like targets. The right panel shows that nuisance variation can make an intermediate bottleneck width optimal. These trends are consistent with Section 5.

data variance along selected directions and the target-dependent variance σ_α^2 along omitted directions. Detailed comparisons are provided in Figure 3.

Prediction targets under power-law spectra. We examine power-law spectra with $\gamma \in \{0.5, 2\}$, varying the bottleneck width $m \in \{1, \ldots, 10\}$ and prediction target $\alpha \in \{0.5, 0.625, 0.75, 0.875, 1\}$. A single fixed logit-normal timestep distribution is used throughout the sweep. The slowly decaying spectrum favors more v-like targets, whereas the rapidly decaying spectrum favors more x-like targets, consistent with the theoretical dependence on spectral decay Figure 2.

Model capacity under nuisance variation. We test whether increasing capacity initially reduces generation error but eventually increases it. We set $d = 5$ and $D = 10$, adding isotropic nuisance variation outside the signal subspace. Under x-prediction, we vary the bottleneck width $m \in \{1, \ldots, 10\}$ and nuisance variance $\lambda_{\text{noise}} \in \{0.1, 0.2, 0.3, 0.4\}$, evaluating generation against the clean distribution. The generation error exhibits a U-shaped dependence on capacity, with its minimum near $m = d$. Analogous results for power-law spectra are also presented in Figure 4.

Validation on Real Data. We also examine whether these qualitative trends persist on the CelebA-HQ dataset (Karras et al., 2018) using the JiT-B/8 model (Li & He, 2026). We generate images directly in pixel space, where natural images are expected to concentrate near a low-dimensional manifold (Li & He, 2026). We separately vary prediction targets, timestep sampling distributions, and bottleneck widths in the patch embedding, evaluating generation quality with FID-50K. The results show benefits from x-prediction, suitably chosen logit-normal sampling, and an intermediate bottleneck width, consistent with prior empirical findings (Li & He, 2026). Experimental details are provided in Section E.2.

8 Conclusion and Discussion

Our analysis jointly studies three design choices that are often considered separately within a tractable model. Within this model, the spectral scores determine which covariance eigendirections are selected by a training-loss minimizer. Along selected directions, generation recovers the corresponding data variances, whereas along omitted directions, the prediction target determines the generated variance. This characterization yields theoretical counterparts of the three motivating observations and shows that preferred design choices depend on the spectral structure of the data. Beyond the exact minimizer, our analysis also provides additional insight into the empirical preference such as logit-normal timestep sampling and v-loss.

Limitations and future work. Our theory deliberately uses Gaussian data, a linear time-gated predictor, and population training objectives to isolate the effects of the three design choices and obtain closed-form characterizations. These simplifications leave several important questions open. Finite-sample effects and optimization dynamics may alter the selected directions and residual errors relative to our population analysis. In our linear model, depth collapses into an equivalent linear map and therefore does not provide an additional capacity axis. For nonlinear models, however, depth can change both expressivity and feature learning, and its interaction with the three design choices remains open. Although our real-data experiments show that the qualitative phenomena persist beyond the Gaussian setting, a quantitative theory connecting data geometry to prediction target, timestep sampling, and model capacity remains an important direction for future work.

[5]The required densities are proportional to $w^\star(t)/(1-t)^2$ and $w^\star(t)/t^2$, respectively.

AI USE STATEMENT

We used generative AI tools to improve the clarity and readability of the manuscript and assist with modifications to experimental code. We reviewed the revised text for mathematical accuracy and logical coherence, and checked the code changes for consistency with the intended implementation. We take full responsibility for the final manuscript, experimental code, and reported results.

REPRODUCIBILITY STATEMENT

The modeling assumptions are stated in Section 3 and in the theorem statements in Sections 4 to 6. Proofs and supplementary theoretical details are provided in Appendices B to D. Appendix E describes the experimental setups, including the model configurations, parameter sweeps, and evaluation procedures.

REFERENCES

Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations*, 2023.

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22669–22679. IEEE, 2023.

Gunnar Carlsson. Topology and data. *Bulletin of the American mathematical society*, 46(2):255–308, 2009.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024.

Anne Gagneux, Ségolène Tiffany Martin, Rémi Gribonval, and Mathurin Massias. Training flow matching: The role of weighting and parameterization. In *ICLR 2026 2nd Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy*, 2026.

Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-SNR weighting strategy. In *International Conference on Computer Vision*, 2023.

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

Qing Jin and Chaoyang Wang. Revisiting diffusion model predictions through dimensionality. *arXiv preprint arXiv:2601.21419*, 2026.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Hk99zCeAb.

Diederik P Kingma and Ruiqi Gao. Understanding diffusion objectives as the ELBO with simple data augmentation. In *Advances in Neural Information Processing Systems*, 2023.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Conference on Computer Vision and Pattern Recognition*, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Yiyang Lu, Susie Lu, Qiao Sun, Hanhong Zhao, Zhicheng Jiang, Xianbang Wang, Tianhong Li, Zhengyang Geng, and Kaiming He. One-step latent-free image generation with pixel mean flows. *arXiv preprint arXiv:2601.22158*, 2026.

Alexander Maloney, Daniel A Roberts, and James Sully. A solvable model of neural scaling laws. *arXiv preprint arXiv:2210.16859*, 2022.

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

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *International Conference on Computer Vision*, 2023.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, 2018.

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

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

Marshall H. Stone. The generalized weierstrass approximation theorem. *Mathematics Magazine*, 21 (5):237–254, 1948.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, 2020.

Roman Worschech and Bernd Rosenow. Analyzing neural scaling laws in two-layer networks with power-law data spectra. In *International Conference on Learning Representations*, volume 2025, pp. 89098–89145, 2025.

Jingfeng Yao, Cheng Wang, Wenyu Liu, and Xinggang Wang. Fasterdit: Towards faster diffusion transformers training without architecture modification. *Advances in neural information processing systems*, 37:56166–56189, 2024.

A Missing Details for Section 3

The proposition states a standard approximation condition on the time features, analogous to the universal approximation assumptions used for rich basis expansions and Fourier-style embeddings (Stone, 1948; Tancik et al., 2020).

Proposition 2 (Approximation of time gates). *Let $\{\varphi_k\}_{k\geq 1} \subset C([0,1])$ be scalar time features whose linear span is dense in $C([0,1])$ under the uniform norm, and let $\phi_K(t) = (\varphi_1(t), \ldots, \varphi_K(t)) \in \mathbb{R}^K$. For any compact interval $J \subset (0,1)$, any $s \in C(J;\mathbb{R}^m)$, and any $\varepsilon > 0$, there exist K and $E \in \mathbb{R}^{m\times K}$ such that $\sup_{t\in J} \|E\phi_K(t) - s(t)\|_2 \leq \varepsilon$.*

Proof of Proposition 2. Write $s(t) = (s_1(t), \ldots, s_m(t))$. Since J is a compact interval, each coordinate $s_j \in C(J)$ admits a continuous extension $\widetilde{s}_j \in C([0,1])$. By density of the span of $\{\varphi_k\}_{k\geq 1}$ in $C([0,1])$, for any $\eta > 0$ there exist K_j and coefficients $a_{j1}, \ldots, a_{jK_j}$ such that

$$\sup_{t\in[0,1]} \left| \widetilde{s}_j(t) - \sum_{k=1}^{K_j} a_{jk}\varphi_k(t) \right| \leq \eta.$$

Let $K = \max_j K_j$ and pad shorter coefficient vectors with zeros. Define $E \in \mathbb{R}^{m\times K}$ by $E_{jk} = a_{jk}$. Then, for every $t \in J$,

$$\|E\phi_K(t) - s(t)\|_2 \leq \sqrt{m}\,\eta.$$

Taking $\eta = \varepsilon/\sqrt{m}$ proves the claim. □

By Proposition 2, increasing K makes the finite-dimensional gate in $\mathcal{F}_m^{(K)}$ expressive enough to approximate arbitrary continuous time gates on compact subintervals of $(0,1)$.

B Proofs for Section 4

B.1 Proofs for Section 4.1

Proof of Proposition 1. For fixed t, the squared-loss minimizer over all measurable functions of z_t is the conditional expectation $\mathbb{E}[v \mid z_t]$. It remains to compute this conditional expectation.

Since $x \sim \mathcal{N}(0, H)$ and $\epsilon \sim \mathcal{N}(0, I_D)$ are independent, the pair (v, z_t) is jointly Gaussian and centered. Thus, the covariance between v and z_t is

$$\mathrm{Cov}(v, z_t) = \mathbb{E}[(x-\epsilon)(tx + (1-t)\epsilon)^\top] = tH - (1-t)I_D. \tag{23}$$

The covariance of z_t is $\mathrm{Cov}(z_t) = t^2H + (1-t)^2 I_D$. The Gaussian conditioning formula gives

$$\mathbb{E}[v \mid z_t = z] = \mathrm{Cov}(v, z_t)\mathrm{Cov}(z_t)^{-1} z.$$

Substituting the covariance expressions yields the matrix form in Equation (10). Since $H = \sum_i \lambda_i e_i e_i^\top$, both matrices are diagonal in the basis $\{e_i\}_{i=1}^D$, which gives the coordinate form in Equation (10). □

B.2 Proofs for Section 4.2

Proof of Theorem 1. We first rewrite the velocity loss as a weighted regression loss for y^α. By the first identity in Equation (5),

$$v_f^\alpha(z_t; t) - v = \frac{f(z_t; t) - y^\alpha}{\alpha(1-t) + (1-\alpha)t}.$$

Let $g^\alpha(z; t) := \mathbb{E}[y^\alpha \mid z_t = z]$. Because the map from y^α to v in Equation (5) is affine in y^α, the velocity induced by g^α is the Bayes velocity $v^\star$ from Proposition 1. The Pythagorean identity for conditional expectation gives

$$\begin{aligned}
\mathcal{L}_w(f) &= \mathcal{L}_w^{\mathrm{Bayes}} + \int_0^1 w(t)\mathbb{E}\left[\left\| \frac{f(z_t;t) - g^\alpha(z_t;t)}{\alpha(1-t) + (1-\alpha)t} \right\|^2 \right] dt \\
&= \mathcal{L}_w^{\mathrm{Bayes}} + \int_0^1 \frac{w(t)}{\left[\alpha(1-t) + (1-\alpha)t\right]^2} \mathbb{E}\left[\|f(z_t;t) - g^\alpha(z_t;t)\|^2 \right] dt.
\end{aligned} \tag{24}$$

We next compute g^α. In the eigenbasis of H, write $x_i = x^\top e_i$, $\epsilon_i = \epsilon^\top e_i$, and $z_i = z_t^\top e_i$. Then

$$z_i = tx_i + (1-t)\epsilon_i, \qquad y_i^\alpha = \alpha x_i - (1-\alpha)\epsilon_i.$$

The pair (y_i^α, z_i) is centered Gaussian with $\mathrm{Var}(z_i) = t^2\lambda_i + (1-t)^2$ and $\mathrm{Cov}(y_i^\alpha, z_i) = \alpha t\lambda_i - (1-\alpha)(1-t)$. Therefore,

$$\mathbb{E}[y_i^\alpha \mid z_i] = \frac{\alpha t\lambda_i - (1-\alpha)(1-t)}{t^2\lambda_i + (1-t)^2} z_i.$$

Thus the unconstrained Bayes predictor for y^α is

$$g^\alpha(z;t) = \sum_{i=1}^{D} \frac{\alpha t\lambda_i - (1-\alpha)(1-t)}{t^2\lambda_i + (1-t)^2} (z^\top e_i) e_i. \tag{25}$$

It remains to impose the bottleneck. For any $f \in \mathcal{F}_m$, the output of $f(\cdot;t)$ lies in the column space of U; denote this subspace by $\mathcal{U}$. Then $\dim(\mathcal{U}) \le m$. Let $P_\mathcal{U}$ be the orthogonal projection onto $\mathcal{U}$. Among all predictors whose outputs lie in $\mathcal{U}$, the pointwise optimal approximation to g^α is $P_\mathcal{U} g^\alpha$. Hence

$$\begin{aligned}
&\int_0^1 \frac{w(t)}{\big[\alpha(1-t) + (1-\alpha)t\big]^2} \mathbb{E}\left[\|f(z_t;t) - g^\alpha(z_t;t)\|^2\right] dt \\
&\quad \ge \int_0^1 \frac{w(t)}{\big[\alpha(1-t) + (1-\alpha)t\big]^2} \mathbb{E}\left[\|g^\alpha(z_t;t) - P_\mathcal{U} g^\alpha(z_t;t)\|^2\right] dt.
\end{aligned} \tag{26}$$

The weighted squared norm of the component of g^α captured by $\mathcal{U}$ is

$$\int_0^1 \frac{w(t)}{\big[\alpha(1-t) + (1-\alpha)t\big]^2} \mathbb{E}\left[\|P_\mathcal{U} g^\alpha(z_t;t)\|^2\right] dt = \sum_{i=1}^{D} S_w^\alpha(\lambda_i)\, \|P_\mathcal{U} e_i\|^2. \tag{27}$$

Indeed, by Section B.1, $z_i = z_t^\top e_i$ are centered and uncorrelated, so cross terms vanish. The contribution of direction e_i is

$$\begin{aligned}
&\int_0^1 \frac{w(t)}{\big[\alpha(1-t) + (1-\alpha)t\big]^2} \left[\frac{\alpha t\lambda_i - (1-\alpha)(1-t)}{t^2\lambda_i + (1-t)^2}\right]^2 \mathbb{E}[z_i^2]\, dt \\
&\quad = \int_0^1 w(t) \frac{\big[\alpha t\lambda_i - (1-\alpha)(1-t)\big]^2}{\big[\alpha(1-t) + (1-\alpha)t\big]^2 \big[t^2\lambda_i + (1-t)^2\big]}\, dt = S_w^\alpha(\lambda_i).
\end{aligned} \tag{28}$$

Let $r_i = \|P_\mathcal{U} e_i\|^2$. Then $0 \le r_i \le 1$ and $\sum_{i=1}^D r_i = \mathrm{tr}(P_\mathcal{U}) \le m$. Therefore,

$$\sum_{i=1}^{D} S_w^\alpha(\lambda_i) r_i \le \sum_{i \in I^\star} S_w^\alpha(\lambda_i),$$

with equality when $\mathcal{U} = \mathrm{span}\{e_i : i \in I^\star\}$. Thus the optimal m-dimensional output subspace keeps the m directions with largest spectral scores.

This bound is achieved within $\mathcal{F}_m$ by choosing $U^\star = V^\star = [e_i]_{i \in I^\star}$ and the gates $s_i^\star(t) = \frac{\alpha t\lambda_i - (1-\alpha)(1-t)}{t^2\lambda_i + (1-t)^2}$ for $i \in I^\star$. The omitted directions contribute exactly $\sum_{i \notin I^\star} S_w^\alpha(\lambda_i)$ above the Bayes velocity loss. This proves $\min_{f \in \mathcal{F}_m} \mathcal{L}_w(f) = \mathcal{L}_w^{\mathrm{Bayes}} + \sum_{i \notin I^\star} S_w^\alpha(\lambda_i)$ and the form of $f^\star$ in Equation (12). □

B.3 Proofs for Section 4.3

Proof of Theorem 2. Since $f^\star$ in Equation (12) is diagonal in the eigenbasis $\{e_i\}_{i=1}^D$, the induced velocity field is also diagonal in this basis, as the transformation inEquation (5) preserves diagonal structure. Write $Z_i(t) = Z_t^\top e_i$. The coordinates evolve independently by scalar linear ODEs.

First consider $i \in I^\star$. In this direction,

$$f_i^\star(z;t) = \frac{\alpha t\lambda_i - (1-\alpha)(1-t)}{t^2\lambda_i + (1-t)^2} z_i.$$

Substituting this into Equation (5) gives

$$\frac{\frac{\alpha t\lambda_i - (1-\alpha)(1-t)}{t^2\lambda_i + (1-t)^2} - (2\alpha - 1)}{\alpha(1-t) + (1-\alpha)t} = \frac{t\lambda_i - (1-t)}{t^2\lambda_i + (1-t)^2}. \tag{29}$$

Hence

$$\frac{dZ_i(t)}{dt} = \frac{t\lambda_i - (1-t)}{t^2\lambda_i + (1-t)^2} Z_i(t).$$

Since

$$\frac{d}{dt}\log\left(t^2\lambda_i + (1-t)^2\right) = \frac{2(t\lambda_i - (1-t))}{t^2\lambda_i + (1-t)^2},$$

the solution is $Z_i(t) = \sqrt{t^2\lambda_i + (1-t)^2}\, Z_i(0)$. Because $Z_i(0) \sim \mathcal{N}(0,1)$, we obtain

$$\lambda_i^{\text{gen}}(t) = t^2\lambda_i + (1-t)^2, \qquad i \in I^\star.$$

Now consider $i \notin I^\star$. The predictor has no component in direction e_i, so $f_i^\star(z;t) = 0$. Therefore

$$\frac{dZ_i(t)}{dt} = \frac{1-2\alpha}{\alpha(1-t) + (1-\alpha)t} Z_i(t).$$

For $\alpha \in (0,1]$, this ODE has solution

$$Z_i(t) = \frac{\alpha(1-t) + (1-\alpha)t}{\alpha} Z_i(0).$$

Thus

$$\lambda_i^{\text{gen}}(t) = \left(\frac{\alpha(1-t) + (1-\alpha)t}{\alpha}\right)^2, \qquad i \notin I^\star.$$

Since $Z_0 \sim \mathcal{N}(0, I_D)$ and each coordinate evolves by a deterministic scalar multiplier, Z_t remains centered Gaussian with diagonal covariance in the basis $\{e_i\}_{i=1}^D$. This proves Equation (14); evaluating at $t = 1$ then gives the endpoint eigenvalues. □

B.3.1 Missing Details for Section 4.3

In this section, we show that, for x-prediction, the spectral score is increasing in λ for every timestep, so the training-loss minimizer selects the directions with the largest eigenvalues regardless of the timestep distribution.

Proof. When $\alpha = 1$, the spectral score becomes

$$S_w^1(\lambda) = \int_0^1 w(t) \frac{t^2\lambda^2}{(1-t)^2\left[t^2\lambda + (1-t)^2\right]}\, dt. \tag{30}$$

For every fixed $t \in (0,1)$, the integrand is strictly increasing for $\lambda > 0$, since its derivative with respect to λ is

$$\frac{t^2\lambda\left[t^2\lambda + 2(1-t)^2\right]}{(1-t)^2\left[t^2\lambda + (1-t)^2\right]^2} > 0.$$

for $\lambda > 0$. Hence the ordering of the scores agrees with the ordering of the eigenvalues, independently of w, up to ties. Theorem 1 therefore implies that the training loss minimizer selects the eigendirections corresponding to the m largest eigenvalues. □

C Missing Details and Proofs for Section 5

We first characterize all training loss minimizers and derive their endpoint generation errors (Appendix C.1). We then prove the results on timestep sampling (Appendix C.2), prediction target (Appendix C.3), and model capacity (Appendix C.4).

Let $\mathcal{W}_\alpha$ denote the class of sampling densities with finite spectral scores, as required in Theorem 1. Note that if $0 < \alpha < 1$, then the spectral scores are always finite. We write $[r] = \{1, \dots, r\}$, with $[0] = \varnothing$, and interpret sums over empty sets as zero. For notational convenience, we write $\sigma := \sigma_\alpha = (1-\alpha)/\alpha$ throughout this appendix.

C.1 Missing Details: Training Loss Minimizers and Endpoint Generation Error

Theorem 1 constructs a spectral training loss minimizer. To optimize generation error over all training loss minimizers, we need the converse: every minimizer must have this form. The fixed input and output spaces of $\mathcal{F}_m$ are essential to this conclusion.

Lemma 1 (Spectral form of every training loss minimizer)**.** *Fix $\alpha \in (0, 1]$, $w \in \mathcal{W}_\alpha$, and $1 \le m \le D$. A predictor $f \in \mathcal{F}_m$ minimizes $\mathcal{L}_w$ if and only if*

$$f(z;t) = \sum_{i \in I} a_i^\alpha(t)(z^\top e_i)e_i, \qquad a_i^\alpha(t) = \frac{\alpha t \lambda_i - (1-\alpha)(1-t)}{t^2\lambda_i + (1-t)^2}, \tag{31}$$

where I indexes m largest spectral scores $S_w^\alpha(\lambda_i)$. Thus, ties in the spectral scores can allow different selected index sets, but do not allow rotations between covariance eigendirections.

Proof. Write $f(z;t) = M(t)z$, where $M(t) = U\,\mathrm{diag}(s(t))V^\top$, and set

$$q_i(t) = t^2\lambda_i + (1-t)^2, \qquad G_\alpha(t) = \sum_{i=1}^D a_i^\alpha(t) e_i e_i^\top.$$

The Bayes target predictor is $G_\alpha(t)z$. Let P be the orthogonal projection onto $\mathrm{col}(U)$. Since $PM(t) = M(t)$, the two components of

$$(M(t) - G_\alpha(t))z_t = (M(t) - PG_\alpha(t))z_t - (I_D - P)G_\alpha(t)z_t$$

are orthogonal. The Bayes loss is finite, since it is at most $\mathbb{E}\|v\|^2 = \mathrm{tr}(H) + D$; finite spectral scores then imply a finite minimum loss by Theorem 1. The regression decomposition gives

$$\begin{aligned}\mathcal{L}_w(f) - \mathcal{L}_w^{\mathrm{Bayes}} = &\int_0^1 \frac{w(t)}{d_\alpha(t)^2}\mathbb{E}\big[\|(M(t) - PG_\alpha(t))z_t\|^2\big]\,dt \\ &+ \sum_{i=1}^D S_w^\alpha(\lambda_i)\big(1 - \|Pe_i\|^2\big).\end{aligned} \tag{32}$$

All scores are strictly positive: their integrands vanish at most at one interior timestep. Moreover, $0 \le \|Pe_i\|^2 \le 1$ and $\sum_i \|Pe_i\|^2 = \mathrm{rank}\,P \le m$. If I_0 indexes m largest scores, it follows that

$$\sum_{i=1}^D S_w^\alpha(\lambda_i)\|Pe_i\|^2 \le \sum_{i \in I_0} S_w^\alpha(\lambda_i).$$

The minimizer in Theorem 1 attains this bound and makes the integral in (32) zero. Every minimizer must therefore satisfy both equalities. Strict positivity of the scores gives $\mathrm{rank}\,P = m$.

The function $\mathbb{E}\|(M(t) - PG_\alpha(t))z_t\|^2/d_\alpha(t)^2$ is continuous and nonnegative on $(0,1)$. Every open subinterval has positive sampling mass, so a zero integral forces this function to vanish at every interior timestep. Since $\mathrm{Cov}(z_t) \succ 0$, we obtain

$$M(t) = PG_\alpha(t), \qquad 0 < t < 1.$$

The row space of $M(t)$ lies in the fixed space $\mathrm{col}(V)$. On the other hand, $G_\alpha(t) \to \alpha I_D$ as $t \uparrow 1$, and hence $M(t) \to \alpha P$. Because $\alpha > 0$, this implies $\mathrm{range}(P) \subseteq \mathrm{col}(V)$. Both spaces have dimension m, so they are equal. Consequently,

$$PG_\alpha(t)(I_D - P) = M(t)(I_D - P) = 0.$$

Taking transposes shows that $(I_D - P)G_\alpha(t)P = 0$ as well. Thus P commutes with $G_\alpha(t)$.

For each fixed $t \in (0,1)$, the eigenvalue of $G_\alpha(t)$ associated with λ is strictly increasing in λ, since

$$\frac{\partial}{\partial \lambda}\frac{\alpha t\lambda - (1-\alpha)(1-t)}{t^2\lambda + (1-t)^2} = \frac{t(1-t)d_\alpha(t)}{[t^2\lambda + (1-t)^2]^2} > 0.$$

It follows that $G_\alpha(t)$ has distinct eigenvalues. An orthogonal projection commuting with this matrix must project onto a subset of its eigenvectors. Hence $P = \sum_{i\in I} e_i e_i^\top$ for a size-m set I. Equality in the captured-score bound requires I to index m largest scores, which proves the necessary form in (31). The converse follows from Theorem 1. □

Endpoint generation error. For any size-m index set I, the spectral predictor in (31) has the explicit generated trajectory

$$e_i^\top Z_t = \begin{cases} \sqrt{q_i(t)}\, e_i^\top Z_0, & i \in I, \\ \dfrac{d_\alpha(t)}{\alpha}\, e_i^\top Z_0, & i \notin I. \end{cases}$$

These expressions solve the generation ODE and extend continuously to both endpoints. In particular, the endpoint exists for every training loss minimizer covered by Lemma 1, including at $\alpha = 1$. Selected directions have variance λ_i, and omitted directions have variance σ^2. The generated and clean covariances share an eigenbasis, so their squared Wasserstein distance is

$$\mathcal{E}_d(I,\alpha) = \sum_{\substack{i\le d \\ i\notin I}} (\sqrt{\lambda_i} - \sigma)^2 + \sum_{\substack{i>d \\ i\in I}} \lambda_i + \sum_{\substack{i>d \\ i\notin I}} \sigma^2. \tag{33}$$

This proves (15) for every training loss minimizer.

C.2 Proofs for Section 5.1

Proof of Theorem 3. Fix $0 < \alpha < 1$ and $m < D$, and let $n = D - m$. The proof has two parts. First, an endpoint-optimal set can be chosen to omit a consecutive block of eigendirections. Second, a suitable timestep strictly separates that block from the selected directions in instantaneous spectral score. Concentrated logit-normal sampling preserves this separation.

The omitted directions form a consecutive block. For $i \notin I$, the reduction in endpoint error obtained by selecting direction i is

$$g_i := \mathcal{E}_d(I,\alpha) - \mathcal{E}_d(I\cup\{i\},\alpha) = \begin{cases} (\sqrt{\lambda_i} - \sigma)^2, & i \le d, \\ \sigma^2 - \lambda_i, & i > d. \end{cases} \tag{34}$$

These gains do not depend on I. An endpoint-optimal size-m set therefore omits n directions with the smallest gains.

The sequence $(g_i)_{i=1}^D$ is nonincreasing and then nondecreasing, allowing either part to be empty. To verify this, first suppose $\sigma \le \sqrt{\lambda_d}$. The signal gains are then nonincreasing, whereas the nuisance gains are increasing. Either sign of the jump at d is consistent with a single minimum. If $\sigma > \sqrt{\lambda_d}$, the signal gains themselves first decrease and then increase. When $d < D$, the boundary jump satisfies

$$g_{d+1} - g_d = 2\sigma\sqrt{\lambda_d} - \lambda_d - \lambda_{d+1} > \lambda_d - \lambda_{d+1} > 0,$$

so the increasing part continues into the nuisance directions. When $d = D$, there is no boundary to check.

For such a sequence, the indices of its n smallest entries can be chosen consecutively: start at a minimum and extend to the smaller of the two adjacent entries, resolving ties in either direction. Thus an endpoint-optimal set has the form

$$I^\star = [k] \cup \{k+n+1, \dots, D\} \qquad \text{for some } 0 \le k \le m.$$

A timestep separates the selected and omitted directions. Let $L = \lambda_{k+n}$ and $U = \lambda_{k+1}$. The omitted eigenvalues lie in $[L, U]$, and every selected eigenvalue lies outside this interval. With $h = t/(1-t)$, the instantaneous spectral score can be written as

$$\frac{(1+h)^2}{(1+\sigma h)^2} Q_h(\lambda), \qquad Q_h(\lambda) = \frac{(h\lambda - \sigma)^2}{1 + h^2\lambda}. \tag{35}$$

The prefactor is positive and independent of λ, so only Q_h matters for the ranking. Direct calculation gives

$$Q_h'(\lambda) = 1 - \frac{(1+h\sigma)^2}{(1+h^2\lambda)^2}, \quad Q_h''(\lambda) = \frac{2h^2(1+h\sigma)^2}{(1+h^2\lambda)^3} > 0, \tag{36}$$

$$Q_h(U) - Q_h(L) = \frac{h(U-L)\{LUh^3 + (L+U-\sigma^2)h - 2\sigma\}}{(1+h^2U)(1+h^2L)}. \tag{37}$$

For $n \geq 2$, choose $h_0 > 0$ satisfying

$$LUh_0^3 + (L+U-\sigma^2)h_0 - 2\sigma = 0, \qquad t_0 = \frac{h_0}{1+h_0}. \tag{38}$$

This root exists and is unique. Indeed, the cubic is negative at zero and tends to infinity. Its derivative is strictly increasing on the positive axis, so the cubic either increases throughout or first decreases while negative and then increases. In either case, it crosses zero exactly once. The root gives $Q_{h_0}(L) = Q_{h_0}(U)$. Strict convexity implies that values inside $[L, U]$ are at most this common value and values outside are strictly larger. Thus every selected direction has a strictly larger score than every omitted direction. For $n = 1$, set $h_0 = \sigma/L$ instead. The unique omitted direction then has score zero, while every selected direction has positive score.

Logit-normal sampling preserves the strict ranking. Let T_τ satisfy

$$\operatorname{logit}(T_\tau) \sim \mathcal{N}(\log h_0, \tau^2), \qquad \tau > 0.$$

As $\tau \downarrow 0$, T_τ converges in probability to t_0. Each instantaneous score extends to a bounded continuous function on $[0, 1]$ for the fixed interior target α. Its expectation therefore converges to its value at t_0. There are finitely many selected–omitted pairs, so all their strict score gaps persist for sufficiently small $\tau > 0$. The resulting logit-normal density belongs to $\mathcal{W}_\alpha$ and makes $I^\star$ the unique set of m largest scores. Theorem 1 then supplies the required training loss minimizer. □

The boundary target and full width. At $\alpha = 1$, the instantaneous score is strictly increasing in λ for any timestep $t \in (0, 1)$, so every $w \in \mathcal{W}_1$ selects $[m]$. The gains in (34) become λ_i on signal directions and $-\lambda_i$ on nuisance directions. Hence $[m]$ is endpoint-optimal exactly when $m \leq d$ or $m = D$. For $d < m < D$, endpoint optimality instead requires all signal directions and the $m - d$ smallest-variance nuisance directions. This proves the qualification in the theorem's footnote. At full width, all directions are represented and the generated covariance is H for every $\alpha \in (0, 1]$.

The attainable error at a fixed target. For the remaining proofs, define

$$R_m(\alpha) := \inf_{w \in \mathcal{W}_\alpha} \min_{f \in \arg\min_{g \in \mathcal{F}_m} \mathcal{L}_w(g)} W_2^2\big(p_1(f), p_{\text{clean}}\big). \tag{39}$$

The inner minimum exists by Lemma 1, because only finitely many selected index sets are possible. That lemma and the sampling theorem give

$$R_m(\alpha) = \begin{cases} \min_{|I|=m} \mathcal{E}_d(I, \alpha), & 0 < \alpha < 1, \\ \mathcal{E}_d([m], 1), & \alpha = 1. \end{cases} \tag{40}$$

The infimum over densities is attained in both cases. At full width, the same formula holds by the preceding observation.

C.3 Proofs for Section 5.2

Proof of Theorem 4. Fix $m < D$. The change of variables $\sigma = (1-\alpha)/\alpha$ maps $\alpha \in (0, 1]$ bijectively onto $\sigma \in [0, \infty)$. For a fixed size-m set I, write

$$A_I = \sum_{\substack{i \le d \\ i \notin I}} \sqrt{\lambda_i}, \qquad B_I = \sum_{\substack{i \le d \\ i \notin I}} \lambda_i + \sum_{\substack{i > d \\ i \in I}} \lambda_i.$$

Expanding the omitted-direction terms gives

$$\mathcal{E}_d(I, \alpha) = (D - m)\sigma^2 - 2A_I\sigma + B_I. \tag{41}$$

The coefficient $D - m$ counts all omitted directions, including nuisance directions.

Widths below the signal dimension: $m < d$. Every size-m set omits at least one signal direction, so $A_I > 0$. The quadratic in (41) has its unique minimum at

$$\sigma_I^\star = \frac{A_I}{D - m} > 0, \qquad \alpha_I^\star = \frac{D - m}{D - m + A_I} \in (0, 1).$$

There are finitely many size-m sets, so minimizing their individual minima produces a globally optimal pair $(I_0, \alpha^\star)$, with $\alpha^\star \in (0, 1)$. By Theorem 3, some sampling density realizes an endpoint-optimal set $I^\star$ at this target. The pair $(I^\star, \alpha^\star)$ is therefore globally optimal as well. In particular, $\alpha^\star$ must minimize the quadratic for this realized set $I^\star$. Thus

$$\alpha_m^\star = \frac{D - m}{D - m + \sum_{i \le d,\, i \notin I^\star} \sqrt{\lambda_i}},$$

and the associated training loss minimizer attains the stated joint minimum.

Width equal to the signal dimension: $m = d$. The set $I = [d]$ and target $\alpha = 1$ give zero endpoint error. Every density in $\mathcal{W}_1$ realizes this set, so the lower bound of zero is attained by a training loss minimizer. Furthermore, zero error cannot select any nuisance direction. Since the set has size d, it must equal $[d]$; because $m < D$, at least one nuisance direction remains omitted and forces $\sigma = 0$. Hence $\alpha = 1$ is the unique optimal target in this case.

Widths above the signal dimension: $d < m < D$. Define

$$I^\dagger = [d] \cup \{D - m + d + 1, \ldots, D\}, \qquad B_m = \sum_{i=D-m+d+1}^{D} \lambda_i.$$

Every size-m set selects at least $m - d$ nuisance directions. Their variance contribution alone gives

$$\mathcal{E}_d(I, \alpha) \ge \sum_{\substack{i > d \\ i \in I}} \lambda_i \ge B_m. \tag{42}$$

Strict ordering of the spectrum implies that equality in the second inequality requires exactly the $m - d$ smallest nuisance directions. The other d selected directions must then be all signal directions, giving $I = I^\dagger$. Since $D - m > 0$ nuisance directions remain omitted, equality in the first inequality also requires $\sigma = 0$. Thus the joint minimum over index sets and targets is B_m, attained only by $(I^\dagger, 1)$.

This pair cannot be realized by training loss minimization: at $\alpha = 1$, every admissible density selects $[m]$, and

$$R_m(1) = \sum_{i=d+1}^{m} \lambda_i > B_m.$$

For $0 < \alpha < 1$, however, (40) gives

$$B_m \le R_m(\alpha) \le \mathcal{E}_d(I^\dagger, \alpha) = B_m + (D - m)\sigma^2.$$

Therefore $R_m(\alpha) \to B_m$ as $\alpha \uparrow 1$. The infimum over trained predictors equals the joint minimum over sets and targets, but no trained predictor attains it. □

Power-law spectra. We next prove the target limits stated in Section 5.2. Let $d = D$, $\lambda_i = i^{-\gamma}$ with $\gamma > 0$, and $1 \le m \le D - 2$. In this case the optimum has an explicit form:

$$\alpha_m^\star(\gamma) = \frac{D - m}{D - m + \sum_{i=m+1}^{D} i^{-\gamma/2}}. \tag{43}$$

To prove this, set $n = D - m \ge 2$ and $a_i = i^{-\gamma/2}$. For an omitted set $J = \{j_1 < \cdots < j_n\}$, optimizing its endpoint error over $\sigma \ge 0$ yields

$$\min_{\sigma \ge 0} \sum_{r=1}^{n} (a_{j_r} - \sigma)^2 = \frac{1}{n} \sum_{r<s} (a_{j_r} - a_{j_s})^2, \qquad \sigma_J^\star = \frac{1}{n} \sum_{r=1}^{n} a_{j_r}. \tag{44}$$

The constraint $\sigma \ge 0$ is inactive because all a_i are positive.

The sequence a_i is strictly decreasing and strictly convex, so the gaps $a_i - a_{i+1}$ strictly decrease with i. For $r < s$, the ordered indices satisfy $j_r \le j_s - (s - r)$ and $j_s \le m + s$. Consequently,

$$a_{j_r} - a_{j_s} \ge a_{j_s-(s-r)} - a_{j_s} \ge a_{m+r} - a_{m+s}.$$

The second inequality compares sums of the same number of consecutive gaps; shifting those gaps to larger indices decreases their sum. Thus the tail set $J^\star = \{m + 1, \ldots, D\}$ minimizes every term in (44). It is the unique minimizer. Indeed, if $J \ne J^\star$, then $j_1 < m + 1$. For the pair $(r, s) = (1, 2)$, either $j_2 < m + 2$, making the second inequality strict, or $j_2 = m + 2$ and $j_1 < j_2 - 1$, making the first strict.

The unique jointly optimal selected set is therefore $[m]$, and the optimal default standard deviation is the average of $a_{m+1}, \ldots, a_D$. Converting back to α proves (43). This target is interior and is attainable by Theorem 3. As $\gamma \downarrow 0$, every term in the average tends to one, so $\alpha_m^\star(\gamma) \to 1/2$. As $\gamma \to \infty$, every tail index is at least two, so the average tends to zero and $\alpha_m^\star(\gamma) \to 1$. If $m = D - 1$, any single omitted variance can be matched exactly by an appropriate target, giving multiple optima. If $m = D$, every target in $(0, 1]$ recovers the training distribution.

C.4 Proofs for Section 5.3

Proof of Theorem 5. Fix $\alpha \in (0, 1]$. A selected signal direction contributes zero endpoint error, whereas an omitted signal direction contributes a nonnegative amount. A nuisance direction contributes either λ_i if selected or σ^2 if omitted. Hence, for every width and every selected set,

$$\mathcal{E}_d(I, \alpha) \ge \sum_{i=d+1}^{D} \min\{\lambda_i, \sigma^2\}. \tag{45}$$

The set

$$I_\alpha^\star = [d] \cup \{i > d : \lambda_i < \sigma^2\}$$

achieves this bound and has cardinality $m_\alpha^\star = d + \#\{i > d : \lambda_i < \sigma^2\}$. It remains to show that the bound is attainable by a training loss minimizer at this width.

If $m_\alpha^\star = D$, full-width training recovers H and attains the bound directly. If $\alpha = 1$ and $m_\alpha^\star < D$, then $m_\alpha^\star = d$ and every density in $\mathcal{W}_1$ selects $[d]$, again attaining the bound. In the remaining case, $0 < \alpha < 1$ and $m_\alpha^\star < D$. Theorem 3 realizes an endpoint-optimal set at this width. Its error equals the lower bound because $I_\alpha^\star$ already attains that bound among size-$m_\alpha^\star$ sets. Finally, Lemma 1 ensures that every training loss minimizer at every width is subject to (45). This proves global optimality and the claimed minimum error. □

The theorem gives an optimal width, which need not be unique. A nuisance direction with $\lambda_i = \sigma^2$ can be selected or omitted without changing its error. The same is true of a signal direction with $\sqrt{\lambda_i} = \sigma$. For an interior target, the sampling theorem and the full-width case above establish attainability at each resulting width allowed by the model.

Capacity under x-prediction. At $\alpha = 1$, every admissible density selects $[m]$, so

$$R_m(1) = \begin{cases} \sum_{i=m+1}^{d} \lambda_i, & m \le d, \\ \sum_{i=d+1}^{m} \lambda_i, & m > d. \end{cases} \tag{46}$$

In particular, for $1 \le m < D$,

$$R_{m+1}(1) - R_m(1) = \begin{cases} -\lambda_{m+1}, & m < d, \\ \lambda_{m+1}, & m \ge d. \end{cases}$$

The error therefore strictly decreases up to width d and strictly increases thereafter, on whichever of these ranges are nonempty. Its unique minimum is zero at $m = d$. Substituting $\lambda_i = i^{-\gamma}$ gives the power-law specialization.

Capacity under v-prediction. At $\alpha = 1/2$, $\sigma = 1$. All selection gains are nonnegative: they are $(\sqrt{\lambda_i} - 1)^2$ on signal directions and $1 - \lambda_i$ on nuisance directions. Adding any omitted direction can therefore only reduce the endpoint error or leave it unchanged. Since (40) identifies $R_m(1/2)$ with the minimum over size-m sets, $R_m(1/2)$ is nonincreasing in m. Full width is optimal, although zero selection gains can make smaller widths optimal as well.

D Proofs for Section 6

Proof of Theorem 6. We first derive the loss and endpoint-error decompositions for an arbitrary spectral predictor. We then show that $w^\star$ uniquely minimizes the worst-case endpoint error in the stated regime.

Step 1: Excess loss and endpoint error. Fix $\alpha \in (0, 1)$ and a positive continuous timestep sampling density w satisfying $\int_0^1 w(t)^{-1}\, dt < \infty$. Write $q_i(t) := t^2\lambda_i + (1-t)^2$ for the variance of z_t along e_i. Since $\lambda_D > 0$, each q_i is bounded above and away from zero on $[0, 1]$. Consequently,

$$A_i(w) := \int_0^1 \frac{dt}{w(t)q_i(t)} < \infty.$$

The spectral scores are finite because $\alpha(1-t) + (1-\alpha)t \ge \min\{\alpha, 1-\alpha\} > 0$. The Bayes velocity loss, $\int_0^1 w(t) \sum_i \lambda_i / q_i(t)\, dt$, is also finite.

Let I be the represented set of a spectral predictor, with $|I| \le m$. For each $i \in I$, define

$$b_i^{\alpha,\star}(t) = \frac{\alpha t\lambda_i - (1-\alpha)(1-t)}{q_i(t)}, \qquad \rho_i^\alpha(t) = \frac{s_i^\alpha(t) - b_i^{\alpha,\star}(t)}{\alpha(1-t) + (1-\alpha)t}.$$

Thus $\rho_i^\alpha(t)$ is the coefficient of the velocity residual along e_i. Write

$$E_i := \int_0^1 w(t)q_i(t)(\rho_i^\alpha(t))^2\, dt$$

for its contribution to the loss, and define

$$P_{\alpha,w}(I) := \max_{|J|=m} \sum_{j\in J} S_w^\alpha(\lambda_j) - \sum_{i\in I} S_w^\alpha(\lambda_i).$$

Because the spectral scores are nonnegative, $P_{\alpha,w}(I) \ge 0$ for every $|I| \le m$.

Orthogonality to the conditional mean decomposes the population loss into the Bayes loss and the squared deviation from the Bayes predictor. A represented direction contributes E_i to this deviation, while an omitted direction contributes $S_w^\alpha(\lambda_i)$. Subtracting the minimum loss from Theorem 1 therefore gives

$$\mathcal{L}_w(f) - \mathcal{L}_w^\star = P_{\alpha,w}(I) + \sum_{i\in I} E_i. \tag{47}$$

In particular, finite excess loss implies $E_i < \infty$ for every $i \in I$. Cauchy–Schwarz then yields

$$\int_0^1 |\rho_i^\alpha(t)|\, dt \le \sqrt{E_i A_i(w)} < \infty. \tag{48}$$

Let $Z_i(t) := e_i^\top Z_t$. Along a represented direction, the generation ODE is

$$\frac{dZ_i(t)}{dt} = \left[\frac{t\lambda_i - (1-t)}{q_i(t)} + \rho_i^\alpha(t)\right] Z_i(t).$$

The first term in brackets equals $\frac{1}{2}\frac{d}{dt}\log q_i(t)$. Since $q_i(0) = 1$, the unique solution continuous at $t = 0$ is

$$Z_i(t) = \sqrt{q_i(t)} \exp\left(\int_0^t \rho_i^\alpha(s)\, ds\right) Z_i(0).$$

By (48), this solution extends continuously to $t = 1$. Its endpoint variance is $\lambda_i e^{2\delta_i}$, where

$$\delta_i = \int_0^1 \rho_i^\alpha(t)\, dt.$$

For an omitted direction, the zero predictor coefficient instead gives

$$Z_i(t) = \frac{\alpha(1-t) + (1-\alpha)t}{\alpha} Z_i(0), \qquad i \notin I.$$

Its endpoint standard deviation is therefore $\sigma = \sigma_\alpha = (1-\alpha)/\alpha$.

The generated distribution remains centered Gaussian and diagonal in the covariance eigenbasis. Define

$$B_\alpha(I) := \sum_{i \notin I} (\sqrt{\lambda_i} - \sigma)^2.$$

The Wasserstein formula for diagonal Gaussian covariances gives

$$\begin{aligned} W_2^2\big(p_1(f), \mathcal{N}(0, H)\big) &= \sum_{i \in I} \lambda_i (e^{\delta_i} - 1)^2 + \sum_{i \notin I} \left(\sqrt{\lambda_i} - \frac{1-\alpha}{\alpha}\right)^2 \\ &= B_\alpha(I) + \sum_{i \in I} \lambda_i (e^{\delta_i} - 1)^2. \end{aligned} \tag{49}$$

Step 2: Exact worst-case allocation of residual loss. For $u \ge 0$, define

$$F_i(u; w) := \lambda_i (e^{\sqrt{u A_i(w)}} - 1)^2.$$

By (48) and $(e^{-x} - 1)^2 \le (e^x - 1)^2$ for $x \ge 0$, the endpoint-error contribution of direction i is at most $F_i(E_i; w)$. Moreover,

$$F_i(u; w) = \lambda_i \sum_{k=2}^{\infty} \frac{2^k - 2}{k!} A_i(w)^{k/2} u^{k/2}.$$

Every term is nonnegative and convex in u, and the series converges for each $u \ge 0$. Hence $F_i(\,\cdot\,; w)$ is convex and $F_i(0; w) = 0$.

Fix a nonempty represented set I and a residual-loss budget $E > 0$. Whenever $\sum_{i \in I} E_i \le E$, convexity gives

$$\sum_{i \in I} F_i(E_i; w) \le \sum_{i \in I} \frac{E_i}{E} F_i(E; w) \le \max_{i \in I} F_i(E; w).$$

This bound is attained by assigning the entire budget to a maximizing direction i_0 and setting

$$\rho_{i_0}^\alpha(t) = \sqrt{\frac{E}{A_{i_0}(w)}} \frac{1}{w(t) q_{i_0}(t)}, \qquad \rho_i^\alpha(t) = 0 \quad (i \in I \setminus \{i_0\}). \tag{50}$$

Indeed, the chosen residual has energy E and integral $\sqrt{EA_{i_0}(w)}$. The resulting predictor coefficients are continuous on $(0,1)$ and nonzero near $t=1$, so the predictor belongs to $\mathcal{F}_m$ and has represented set I. Its endpoint is well-defined by (48). For $E=0$, the bound is attained by setting all residuals to zero.

Thus, for a fixed nonempty set I and residual-loss budget $E \geq 0$, the exact worst-case contribution from represented directions is $\max_{i\in I} F_i(E;w)$. For $I=\varnothing$, there is no residual contribution. Since there are finitely many represented sets and every $A_i(w)$ is finite, Equations (47) and (49) also imply finite worst-case generation error for every finite excess-loss budget.

Step 3: Uniform separation of represented sets near x-prediction. Let

$$I_0 := \{1,\dots,m\}, \qquad \Delta := \lambda_m - \lambda_{m+1} > 0,$$

and let

$$C := \int_0^1 q_1(t)^{-1/2}\,dt, \qquad \text{so that} \qquad w^\star(t) = q_1(t)^{-1/2}/C.$$

The density $w^\star$ extends to a positive continuous function on $[0,1]$ and satisfies the required reciprocal-integrability condition.

For every $I \neq I_0$ with $|I| \leq m$, the ordering of the eigenvalues gives

$$\sum_{i\in I_0} \lambda_i - \sum_{i\in I} \lambda_i \geq \Delta.$$

The continuous extension $B_1(I) = \sum_{i\notin I} \lambda_i$ therefore satisfies $B_1(I) - B_1(I_0) \geq \Delta$. Since there are finitely many represented sets, for all $\alpha < 1$ sufficiently close to 1,

$$B_\alpha(I) \geq B_\alpha(I_0) + \Delta/2 \qquad (I \neq I_0,\ |I| \leq m). \tag{51}$$

We next establish a corresponding training-loss gap. Put $\sigma = \sigma_\alpha$. The fraction in the spectral score satisfies

$$\frac{[t\lambda - \sigma(1-t)]^2}{[1-t+\sigma t]^2[t^2\lambda + (1-t)^2]} = \frac{\lambda+\sigma^2}{(1-t+\sigma t)^2} - \frac{\lambda}{t^2\lambda + (1-t)^2}.$$

The nonnegative kernel $\sigma/(1-t+\sigma t)^2$ integrates to one on $[0,1]$, while its integral over $[0,1-\varepsilon]$ tends to zero as $\sigma \downarrow 0$ for every $\varepsilon \in (0,1)$. Since $w^\star$ is continuous at 1, and $\lambda/[t^2\lambda + (1-t)^2]$ is bounded for $\lambda > 0$, it follows that

$$\sigma S^\alpha_{w^\star}(\lambda) \longrightarrow w^\star(1)\lambda \qquad (\alpha \uparrow 1). \tag{52}$$

The maximum defining $P_{\alpha,w^\star}(I)$ ranges over finitely many sets. Therefore, for every $I \neq I_0$ with $|I| \leq m$,

$$\sigma P_{\alpha,w^\star}(I) \longrightarrow w^\star(1)\left(\sum_{i\in I_0}\lambda_i - \sum_{i\in I}\lambda_i\right) \geq w^\star(1)\Delta > 0.$$

Hence $P_{\alpha,w^\star}(I) \to \infty$ for every such I. We may consequently choose $\alpha_0 \in (0,1)$, depending only on H, m, such that, for every $\alpha \in (\alpha_0, 1)$, (51) holds and

$$P_{\alpha,w^\star}(I_0) = 0, \qquad P_{\alpha,w^\star}(I) \geq \Delta \quad (I \neq I_0,\ |I| \leq m). \tag{53}$$

The first equality follows because some size-m set attains zero selection penalty, and the second inequality excludes every set other than I_0. Thus I_0 is the unique training-optimal represented set under $w^\star$.

Step 4: A uniform positive excess-loss budget. For any $\lambda_i < \lambda_1$, direct subtraction gives

$$\lambda_1 A_1(w^\star) - \lambda_i A_i(w^\star) = (\lambda_1 - \lambda_i)\int_0^1 \frac{(1-t)^2}{w^\star(t)q_1(t)q_i(t)}\,dt > 0.$$

Also,

$$\frac{F_i(u;w^\star)}{u} \longrightarrow \lambda_i A_i(w^\star) \qquad (u \downarrow 0).$$

It follows that direction 1 maximizes $F_i(u; w^\star)$ over $i \in I_0$ for all sufficiently small $u > 0$. This conclusion holds simultaneously over the finite set I_0; equal eigenvalues give identical functions.

Choose $\eta_0 > 0$ small enough that this dominance holds for $0 < u \le \eta_0$, and that

$$\eta_0 < \Delta, \qquad F_1(\eta_0; w^\star) < \Delta/2.$$

Such a choice exists because $F_1(u; w^\star) \to 0$ as $u \downarrow 0$. It depends only on H, m and is independent of α.

Fix $\alpha \in (\alpha_0, 1)$ and $0 < \eta < \eta_0$. By Equations (47) and (53), every spectral predictor with excess loss at most η under $w^\star$ must have represented set I_0. In particular, no predictor using fewer than m directions satisfies the excess-loss bound.

Let $R_{\alpha,\eta}(w)$ denote the supremum in (22). For $w^\star$, the selection penalty is zero and the entire budget η is available for coefficient residuals. Step 2 and the dominance of direction 1 therefore give

$$\begin{aligned} R_{\alpha,\eta}(w^\star) &= B_\alpha(I_0) + F_1(\eta; w^\star) \\ &= B_\alpha(I_0) + \lambda_1(e^{C\sqrt{\eta}} - 1)^2 < B_\alpha(I_0) + \Delta/2, \end{aligned} \tag{54}$$

where we used $A_1(w^\star) = C^2$.

Step 5: Global optimality and uniqueness. Consider any other positive continuous timestep sampling density w satisfying $\int_0^1 w(t)^{-1}\, dt < \infty$.

If I_0 is training-optimal under w, then $P_{\alpha,w}(I_0) = 0$. Applying the residual construction in (50) with budget η on direction 1 yields

$$R_{\alpha,\eta}(w) \ge B_\alpha(I_0) + \lambda_1(e^{\sqrt{\eta A_1(w)}} - 1)^2.$$

Cauchy–Schwarz and $\int_0^1 w(t)\, dt = 1$ imply

$$C^2 = \left(\int_0^1 \sqrt{w(t)}\, \frac{dt}{\sqrt{w(t) q_1(t)}} \right)^2 \le A_1(w).$$

Equality holds if and only if $w = q_1^{-1/2}/C$ almost everywhere, and hence everywhere on $(0, 1)$ by continuity. Thus $A_1(w) > C^2$ whenever $w \ne w^\star$. Since $\eta > 0$ and $a \mapsto (e^{\sqrt{\eta a}} - 1)^2$ is strictly increasing for $a > 0$, (54) gives

$$R_{\alpha,\eta}(w) > R_{\alpha,\eta}(w^\star) \qquad (w \ne w^\star).$$

If I_0 is not training-optimal under w, choose a training-optimal size-m set $I \ne I_0$. Its predictor with Bayes coefficients on I has zero excess loss and endpoint error $B_\alpha(I)$. Consequently,

$$R_{\alpha,\eta}(w) \ge B_\alpha(I) \ge B_\alpha(I_0) + \Delta/2 > R_{\alpha,\eta}(w^\star),$$

by Equations (51) and (54).

These two cases cover every density under consideration, including those with tied training-optimal sets. They prove that $w^\star$ is the unique minimizer and complete the proof. □

Why reciprocal integrability is required. The preceding proof shows that $\int_0^1 w(t)^{-1}\, dt < \infty$ is sufficient for finite worst-case generation error. In the spectral predictor class considered here, it is also necessary for any positive excess-loss budget.

To see this, retain $\alpha \in (0, 1)$ and $\lambda_D > 0$, but suppose w is a positive continuous density with $\int_0^1 w(t)^{-1}\, dt = \infty$. Then $A_i(w) = \infty$ for every i. The spectral scores remain finite for $\alpha \in (0, 1)$, so a training-optimal spectral predictor still exists. Choose one of its represented directions i and nonzero continuous, compactly supported functions χ_n on $(0, 1)$ such that $0 \le \chi_n \le 1$ and $\chi_n \uparrow 1$ pointwise. Set

$$K_n := \int_0^1 \frac{\chi_n(t)}{w(t) q_i(t)}\, dt, \qquad \rho_{i,n}^\alpha(t) := \sqrt{\frac{\eta}{K_n}}\, \frac{\chi_n(t)}{w(t) q_i(t)},$$

and leave all other velocity residuals zero. Each K_n is positive and finite, and monotone convergence gives $K_n \to \infty$. The excess loss of this perturbation satisfies

$$\int_0^1 w(t)q_i(t)(\rho_{i,n}^{\alpha}(t))^2\,dt = \frac{\eta}{K_n}\int_0^1 \frac{\chi_n(t)^2}{w(t)q_i(t)}\,dt \le \eta,$$

whereas

$$\int_0^1 \rho_{i,n}^{\alpha}(t)\,dt = \sqrt{\eta K_n} \longrightarrow \infty.$$

Each residual is continuous and compactly supported, so each perturbed predictor has a well-defined, finite endpoint covariance. Nevertheless, its endpoint-error contribution $\lambda_i(e^{\sqrt{\eta K_n}}-1)^2$ diverges. Thus the worst-case generation error is infinite for every $\eta > 0$.

Center bias and logit-normal sampling. Under the normalization $\lambda_1 = 1$,

$$q_1(t) = 2(t-\tfrac{1}{2})^2 + \tfrac{1}{2}.$$

Hence $w^\star$ is symmetric about $t = 1/2$ and uniquely maximized there.

The comparison with logit-normal sampling concerns this center bias, not satisfaction of the theorem's density restriction. Indeed, a logit-normal density with location μ and scale $s > 0$ is

$$w_{\mathrm{LN}}(t) = \frac{1}{s\sqrt{2\pi}\,t(1-t)}\exp\left(-\frac{(\log(t/(1-t))-\mu)^2}{2s^2}\right).$$

As $t \downarrow 0$,

$$\log w_{\mathrm{LN}}(t) = -\frac{(\log t)^2}{2s^2} + O(|\log t|),$$

so $w_{\mathrm{LN}}(t) = o(t)$. Therefore $\int_0^1 w_{\mathrm{LN}}(t)^{-1}\,dt = \infty$. Its empirical usefulness is compatible with this observation: the theorem studies worst-case residuals without additional constraints on their concentration in time.

Loss space and timestep sampling. For a fixed induced velocity, define

$$x_f(z;t) := z + (1-t)v_f^{\alpha}(z;t), \qquad \epsilon_f(z;t) := z - tv_f^{\alpha}(z;t).$$

Along the training interpolation,

$$x_f(z_t;t) - x = (1-t)(v_f^{\alpha}(z_t;t) - v), \qquad \epsilon_f(z_t;t) - \epsilon = -t(v_f^{\alpha}(z_t;t) - v).$$

Thus reproducing the $w^\star$-weighted velocity loss through timestep sampling alone under x- or ϵ-loss requires densities proportional to

$$\frac{w^\star(t)}{(1-t)^2} \qquad \text{or} \qquad \frac{w^\star(t)}{t^2},$$

respectively. Neither is integrable because $w^\star$ has strictly positive limits at both endpoints. This comparison concerns matching the loss weighting for the same induced velocity; it does not establish a universal performance ordering between different loss spaces.

Why the theorem excludes $\alpha = 1$. On the full time interval, reciprocal integrability is incompatible with finite spectral scores at $\alpha = 1$. For any $\lambda > 0$, finiteness of $S_w^1(\lambda)$ requires

$$\int_{1/2}^1 \frac{w(t)}{(1-t)^2}\,dt < \infty.$$

If $\int_0^1 w(t)^{-1}\,dt < \infty$ also held, Cauchy–Schwarz would imply

$$\int_{1/2}^{1-\varepsilon}\frac{dt}{1-t} \le \left(\int_{1/2}^1 \frac{w(t)}{(1-t)^2}\,dt\right)^{1/2}\left(\int_{1/2}^1 \frac{dt}{w(t)}\right)^{1/2}.$$

The right-hand side is finite and independent of ε, while the left-hand side diverges as $\varepsilon \downarrow 0$. Thus finite spectral scores and the residual control used above are incompatible at $\alpha = 1$ on the full time interval. Since $m < D$, every spectral predictor omits a positive-variance direction; under a density satisfying reciprocal integrability, its population loss at $\alpha = 1$ is therefore infinite.

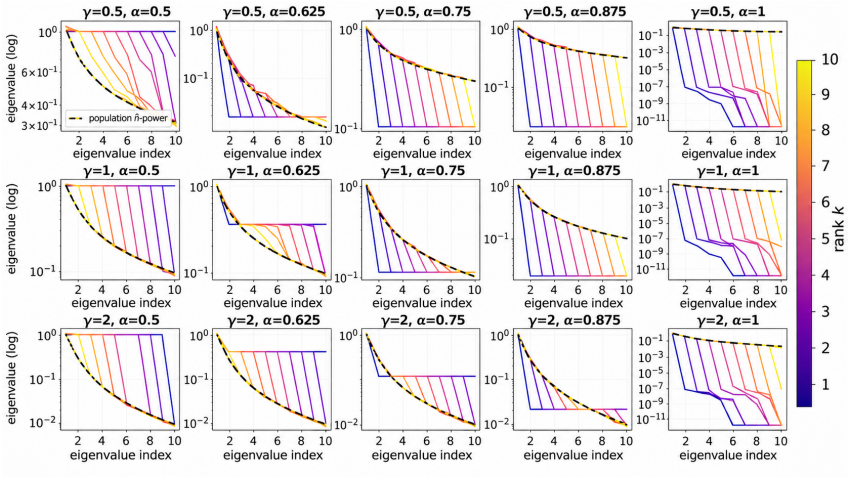


Figure 3: **Generated spectra and spectral selection.** Colored curves show the generated covariance spectra for model ranks $m = 1, \ldots, 10$, while black dashed curves show the power-law data spectra $\lambda_i = i^{-\gamma}$. Rows and columns vary the spectral decay γ and prediction target α, respectively. As m increases, the model reproduces the data variance along more eigendirections, while omitted directions remain near the target-dependent variance σ_α^2, consistent with Theorem 2.

E Further Details for Section 7

E.1 Synthetic Data Experiments

Synthetic experiments setup. All synthetic experiments use centered Gaussian data $x \sim \mathcal{N}(0, H)$ with ambient dimension $D = 10$ unless stated otherwise. Independent noise is sampled as $\epsilon \sim \mathcal{N}(0, I_D)$, and training uses the interpolation $z_t = tx + (1-t)\epsilon$ from the main text. We use a finite-dimensional time-embedding implementation of the gated bottleneck predictor in Equation (2) and train it by numerical optimization. Generation follows the induced velocity field, which we integrate numerically from Gaussian noise. We evaluate the generated endpoint through its covariance and the Wasserstein-2 distance to the corresponding evaluation distribution.

Spectral selection and generated covariance. To test Theorems 1 and 2, we compare the subspace learned by numerical optimization with the covariance eigendirections ranked by the spectral scores $S_w^\alpha(\lambda_i)$. We then compare the generated covariance with the population characterization in Theorem 2. Across the tested configurations, numerical training preferentially represents the directions with the largest spectral scores. Along represented directions, the generated eigenvalues approach the corresponding data eigenvalues, whereas omitted directions approach the target-dependent variance $((1-\alpha)/\alpha)^2$.

For example, when $\gamma = 2$ and $\alpha = 0.75$, the omitted-direction variance is

$$\left(\frac{1-0.75}{0.75}\right)^2 = \frac{1}{9}.$$

Together with the learned-subspace comparison, these experiments test both stages of the spectral characterization: which eigendirections are selected and how this selection determines the generated covariance.

Prediction targets under power-law spectra. To examine the effect of dimensional structure of data, we use power-law covariance eigenvalues $\lambda_i = i^{-\gamma}$ and compare $\gamma \in \{0.5, 2\}$ while varying the bottleneck width m and prediction target α. A single logit-normal timestep distribution is used throughout the sweep, rather than retuning timestep sampling for each target. Thus, the comparison tests whether the qualitative shift in target preference with spectral decay persists under fixed timestep sampling. As reported in Figure 2, the slowly decaying spectrum favors more v-like targets, whereas the rapidly decaying spectrum favors more x-like targets.

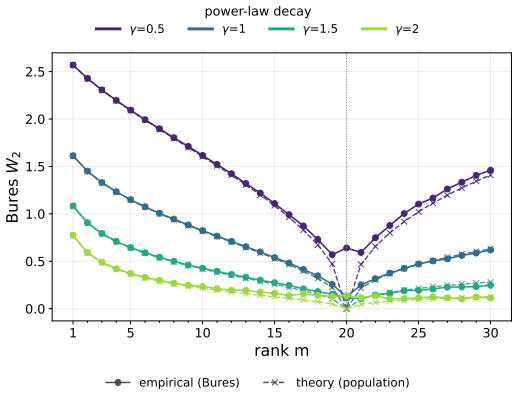


Figure 4: Under power-law covariance spectra $\lambda_i = i^{-\gamma}$, we measure W_2 distance to the clean signal distribution versus bottleneck width m. Solid and dashed lines show empirical results and theoretical results, respectively. nuisance variation can make an intermediate bottleneck width optimal, consistent with the Section 5.

Model capacity under nuisance variation. For the capacity experiments, we use a rank-d clean signal together with isotropic nuisance variation outside the signal subspace:

$$H_{\text{clean}} = \sum_{i=1}^{d} \lambda_i e_i e_i^\top, \qquad H_{\text{obs}} = H_{\text{clean}} + \lambda_{\text{noise}} \sum_{i=d+1}^{D} e_i e_i^\top.$$

We fix $d = 5$, vary the bottleneck width m, and consider several values of λ_{noise}. This setting differs slightly from the power-law nuisance spectrum discussed in Section 5.3: all nuisance directions have the same variance. It nevertheless isolates the same capacity tradeoff. Under x-prediction, widths $m < d$ omit clean signal directions, whereas widths $m > d$ begin to represent nuisance variation. The resulting clean-generation error therefore exhibits the same qualitative U-shaped dependence on model capacity described in Section 5.3. Additional deviations in the zero-nuisance setting arise from the finite time embedding used in numerical training.

We also test the power-law setting, taking $H_{\text{obs}} = \sum_{i=1}^{D} i^{-\gamma} e_i e_i^\top$ and retaining its first d eigendirections as the clean signal. Here, nuisance variances follow the power-law decay instead of remaining constant. Figure 4 compares empirical W_2 distances to the clean distribution with theoretical predictions under x-prediction, showing the same U-shaped dependence on capacity with a minimum near $m = d$.

E.2 Real-Data Experiments

Experimental setup. We study unconditional image generation on CelebA-HQ (Karras et al., 2018), using all 30,000 images at 256×256 resolution. We adopt JiT-B/8, a pixel-space Transformer adapted from Just image Transformers (Li & He, 2026), with 8×8 RGB patches, 12 Transformer blocks, and hidden width 768. All models are trained from scratch for 1,000 epochs: 500 epochs of block-masked training followed by 500 epochs without masking. We report final-checkpoint FID using 50,000 generated images and the 30,000 real images as the reference, with EMA weights and a 50-step Heun sampler without classifier-free guidance.

Controlled sweeps. We separately investigate prediction targets, timestep distributions, and patch-embedding bottleneck widths, holding the remaining settings fixed within each sweep. For prediction targets, we sweep

$$u_\alpha = \alpha x - (1 - \alpha)\epsilon, \qquad \alpha \in \{0, 0.1, \ldots, 1\},$$

covering noise, velocity, and clean-data prediction at $\alpha = 0$, 0.5, and 1, respectively, up to fixed sign and scale factors. All parameterizations use a common stabilized velocity-space training objective.

For timestep sampling, we use clean-data prediction and vary the location and dispersion of a logit-normal distribution:

$$\mathrm{logit}(t) \sim \mathcal{N}(\mu, \sigma^2), \quad \mu \in \{-1.2, -0.8, -0.4, 0, 0.4, 0.8, 1.2\}, \quad \sigma \in \{0.5, 0.8, 1.1\}.$$

Here, σ denotes the standard deviation in logit space. Uniform timestep sampling is included as a baseline. For the bottleneck sweep, we factorize the linear patch embedding as $\mathbb{R}^{192} \to \mathbb{R}^r \to \mathbb{R}^{768}$ and evaluate

$$r \in \{4, 6, 8, 12, 16, 24, 32, 48, 64, 92, 128, 192\}.$$

An 8×8 RGB patch has 192 dimensions, so the composed projection has rank at most $\min(r, 192)$. Thus, $r = 192$ represents the input-dimensional boundary, rather than an arbitrary upper limit on model width. The Transformer depth and hidden width remain fixed. Prediction-target and timestep sweeps use $r = 128$; prediction-target and bottleneck sweeps use $(\mu, \sigma) = (-0.4, 0.8)$. The bottleneck sweep uses x-prediction.

Empirical observations. The results are qualitatively consistent with the three non-trivial empirical observations discussed above, extending these trends to CelebA-HQ. First, clean-data prediction substantially outperforms noise-containing target parameterizations, achieving an FID of 12.00. Second, generation quality depends strongly on timestep sampling: the best evaluated logit-normal setting, $(\mu, \sigma) = (0.4, 1.1)$, achieves 11.03, compared with 15.31 for uniform sampling. The preferred mean depends on the distribution width, highlighting the importance of considering both parameters. Third, increasing bottleneck width does not monotonically improve performance. FID decreases from 26.09 at $r = 4$ to 11.36 at $r = 48$, before increasing to 12.96 at $r = 192$. An intermediate bottleneck therefore outperforms the input-dimensional boundary in this setting. These single-seed results support qualitative consistency, rather than a claim of statistical significance.