ℓ_p-POWER HEAVY-TAILED DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We develop a variational framework for heavy-tailed diffusion based on an ℓ_p-power family, extending Student-t-based models beyond $p = 2$. Using the family's Gaussian scale-mixture structure, we construct diffusion paths for $p \in [1, 2]$ via random-precision Gaussians and derive an adjacent-state p-th power objective through γ-power-divergence projection. At $p = 2$, the objectives are equivalent up to a constant, but differ for $p \in [1, 2)$. Theoretical analysis identifies regimes where the adjacent-state objective is more sensitive to rare components than denoising, and conditions under which $p \in (1, 2)$ yields lower reconstruction risk than either endpoint. Experiments on synthetic data, language, and image generation reveal complementary strengths over Gaussian and Student-t baselines: the adjacent-state objective better preserves rare modes under heavy-tailed noise, while denoising achieves better reconstruction on contaminated data.

1 INTRODUCTION

Recent work in generative modeling and representation learning has explored heavy-tailed distributions as alternatives to Gaussian assumptions. Applications include improving the diversity of images generated from class-imbalanced datasets (Yoon et al., 2023; Kim et al., 2024), modeling environmental extremes (Lafon et al., 2026; Zhang et al., 2026), enhancing prosodic diversity in speech synthesis (Lian et al., 2025), and learning text representation (Jalalzai et al., 2020). These efforts reflect two complementary motivations: capturing extreme observations in the tails of data distributions and learning flexible representations that preserve underrepresented patterns in heterogeneous data.

In this work, we approach heavy-tailed modeling through variational inference, drawing on variational autoencoders (VAEs) that minimize γ-power divergence between data and model joint distributions (Kim et al., 2024; Cho et al., 2026). The hierarchical variational interpretation of diffusion motivates extending this formulation to latent trajectories (Ho et al., 2020b; Kingma et al., 2021). A closely related model is Student-t diffusion (Pandey et al., 2025), which derives a quadratic objective from a sum of per-step conditional γ-power divergences. Following the divergence minimization view, we provide a complementary variational derivation over complete diffusion trajectories, recovering an equivalent quadratic denoising objective in the Student-t case. This reduction requires an explicit derivation because γ-power divergence lacks the additive chain rule of the Kullback-Leibler (KL) divergence.

We further extend this variational formulation to the ℓ_p-power family ($1 \leq p \leq 2$), which includes the Student-t and symmetric Pareto distributions used in VAEs (Kim et al., 2024; Cho et al., 2026), and derive the corresponding ℓ_p^p training losses. At the population level, these losses target conditional Fréchet p-centers, connecting medians at $p = 1$ with conditional means at $p = 2$. Intermediate exponents retain sensitivity to error magnitude while reducing the relative influence of large errors compared with squared loss, a classical motivation in robust estimation (Forsythe, 1972). The family's Gaussian scale-mixture (GSM) representation allows us to construct diffusion paths driven by Gaussian noise with random precisions.

The derived objective raises a further question: does replacing adjacent-state prediction with clean-data denoising preserve its population minimizers beyond the quadratic setting? We examine this relationship and how the exponent and prediction target affect two distinct goals: retaining rare components and suppressing corruption. Further related work is discussed in Appendix A.

Contributions. Our main contributions are as follows:

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.

- **GSM characterization and variational construction.** We characterize the ℓ_p-power family and its Gaussian scale-mixture (GSM) structure by establishing that such representations exist when $0 < p \leq 2$. Building on this characterization, we construct heavy-tailed diffusion paths for $1 \leq p \leq 2$ using Gaussian paths with random precisions and derive an adjacent-state ℓ_p^p variational training objective through the γ-power divergence.
- **Comparison with denoising objective.** We show that the adjacent-state objective and its clean-data denoising surrogate differ only by a parameter-independent constant at $p = 2$, whereas their population minimizers can differ for $1 \leq p < 2$.
- **Theoretical analysis and empirical evaluation.** We analyze the statistical properties of the two objectives and empirically demonstrate their respective strengths in rare-mode preservation and denoising under contamination.

2 Theoretical Background

In this section, we review diffusion models and information geometry, and introduce the ℓ_p-power family and its Gaussian scale-mixture (GSM) structure. Further details and proofs are provided in Appendix B. We first introduce two conventions used throughout the paper.

Definition 2.1 (θ-equivalent objectives). *Objectives $\mathcal{J}$ and $\mathcal{L}$ are θ-equivalent, written $\mathcal{J} \equiv_\theta \mathcal{L}$, if $\mathcal{J}(\theta) = a\mathcal{L}(\theta) + b$ for all θ, where $a > 0$ and $b \in \mathbb{R}$ are independent of θ.*

Unless otherwise stated, all probabilistic measures are assumed to be absolutely continuous with respect to their natural Lebesgue reference measures, and all expectations and losses appearing below are assumed finite For a forward path distribution F, $\mathbb{E}_{t,F}$ denotes expectation over independent draws of the time index and the forward path and expectations over t use $t \sim \mathrm{Unif}\{1, \ldots, T\}$.

2.1 Divergence Projection in Variational Diffusion Models

Diffusion models associate an observation x_0 with a latent trajectory $x_{1:T}$. Let $F(x_{0:T})$ be a fixed inference joint density and $G_\theta(x_{0:T})$ a learned generative joint density. For a chosen divergence $\mathcal{D}$, consider the joint path projection problem

$$\theta^\star \in \arg\min_\theta \mathcal{D}(F \| G_\theta). \tag{1}$$

The standard variational formulation uses the KL divergence for $\mathcal{D}$. When the data entropy and the relevant objectives are finite, the resulting path objective is θ-equivalent to the data-averaged negative ELBO (Kingma & Welling, 2013; Ho et al., 2020b; Kingma et al., 2021).

For Markov forward and reverse paths, write data and model joint densities as

$$F(x_{0:T}) = f_{\text{data}}(x_0) \prod_{t=1}^{T} f_t(x_t \mid x_{t-1}), \qquad G_\theta(x_{0:T}) = g_T(x_T) \prod_{t=1}^{T} g_{\theta,t}(x_{t-1} \mid x_t).$$

Let $f_{t-1|t}$ denote the reverse conditional kernel induced by F. With g_T fixed, the KL chain rule gives

$$\mathcal{D}_{\mathrm{KL}}(F \| G_\theta) \equiv_\theta \mathbb{E}_{t,F} \left[\mathcal{D}_{\mathrm{KL}} \left(f_{t-1|t}(\cdot \mid X_t) \| g_{\theta,t}(\cdot \mid X_t) \right) \right]. \tag{2}$$

Thus, path KL projection and the time-averaged local KL objective are θ-equivalent. A standard distributional choice is Gaussian reverse kernels $g_{\theta,t}(\cdot \mid x_t) = \mathcal{N}_n\big(\mu_\theta(t, x_t), s_t^2 I_n\big)$, with fixed $s_t^2 > 0$ for $t = 1, \ldots, T$. Then, (2) is θ-equivalent to the quadratic adjacent loss

$$\mathcal{L}^{\mathrm{G}}_{\mathrm{adj},2}(\theta) := \mathbb{E}_{t,F} \left[a_t \| X_{t-1} - \mu_\theta(t, X_t) \|_2^2 \right], \quad a_t := 1/(2s_t^2). \tag{3}$$

For the standard Gaussian forward process, $X_t = \sqrt{\alpha_t} X_{t-1} + \sqrt{\beta_t}\varepsilon_t$, where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, $0 < \beta_t < 1$, and $\varepsilon_t \sim \mathcal{N}_n(0, I_n)$ are independent across time and of X_0. The parameterization $\mu_\theta(t, x) = \frac{1}{\sqrt{\alpha_t}} \left(x - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x, t) \right)$ leads to the weighted denoising loss

$$\mathcal{L}^{\mathrm{G}}_{\mathrm{denoise},2}(\theta) := \mathbb{E}_{t,X_0,\varepsilon} \left[w_t \| \varepsilon - \epsilon_\theta(X_t, t) \|_2^2 \right], \qquad w_t := 2(\beta_t / s_t)^2 (\alpha_t (1 - \bar{\alpha}_t))^{-1}. \tag{4}$$

Conditioning on (X_0, X_t) decomposes the adjacent squared error into a posterior-mean error and a θ-independent conditional variance. Under the parameterization above, the posterior-mean error is precisely a weighted noise-prediction error (Ho et al., 2020b). Consequently,

$$\mathcal{D}_{\mathrm{KL}}(F\|G_\theta) \equiv_\theta \mathcal{L}^{\mathrm{G}}_{\mathrm{adj},2} \equiv_\theta \mathcal{L}^{\mathrm{G}}_{\mathrm{denoise},2}. \tag{5}$$

The equivalences above rely on distinct ingredients: the KL chain rule connects path projection to the time-averaged local KL objective, while the quadratic conditional-mean decomposition connects adjacent prediction to denoising. For divergences beyond KL, an analogous additive chain rule need not hold. Likewise, replacing squared error by a nonquadratic loss does not generally preserve these equivalences. The following sections examine these two relationships for heavy-tailed models.

2.2 The γ-Power Divergence and Power Families

In KL divergence-based formulations, the model-dependent term involves $\log(\cdot)$, which is affine in the sufficient statistics for exponential families. By contrast, certain heavy-tailed distributions with polynomially decaying densities admit an analogous affine representation through a power transformation of the density. This motivates the use of the γ-power divergence for heavy-tailed generative models (Kim et al., 2024; Pandey et al., 2025; Cho et al., 2026).

For $\gamma \in (-1, \infty) \setminus \{0\}$, define the γ-power divergence (Eguchi, 2021; Kim et al., 2024) as

$$\mathcal{D}_\gamma(f\|g) := -\frac{1}{\gamma}\left\{\int f(x)\left(\frac{g(x)}{\|g\|_{1+\gamma}}\right)^\gamma dx - \|f\|_{1+\gamma}\right\}. \tag{6}$$

Here $\|h\|_q := (\int h(x)^q dx)^{1/q}$. As $\gamma \to 0$, $\mathcal{D}_\gamma$ converges to the KL divergence, while $(1+\gamma a)^{1/\gamma} \to e^a$. Thus, the same limit recovers both the KL divergence and the exponential-family form. For a fixed statistic T, a power family has density $g_\theta(x) \propto (1 + \gamma\theta^\top T(x))_+^{1/\gamma}$. On its support, $g_\theta(x)^\gamma$ is affine in $T(x)$, up to a multiplicative factor determined by the family parameters. The corresponding information-geometric interpretation is also discussed in Appendix B.2.

Student t and symmetric Pareto distributions provide examples of such power-form densities (Kim et al., 2024; Cho et al., 2026). This algebraic compatibility yields tractable prediction objectives for matched power-family models: with fixed scales and suitable integrability conditions, the model-dependent term reduces to an expectation of the defining residual statistic.

In contrast, KL divergence between power-family distributions generally does not admit a comparable simplification. A Gaussian scale-mixture representation offers an alternative through conditional Gaussian calculations, but these generally lead to a different objective. Specifically, let $\bar{f}(x) = \int f(x \mid \lambda)\,\Pi(d\lambda)$ and $\bar{g}_\theta(x) = \int g_\theta(x \mid \lambda)\,\Pi(d\lambda)$ share a common latent-precision law Π. Then

$$\mathcal{D}_{\mathrm{KL}}(\bar{f}\|\bar{g}_\theta) \le \mathbb{E}_{\Lambda\sim\Pi}\left[\mathcal{D}_{\mathrm{KL}}\big(f(\cdot \mid \Lambda)\|g_\theta(\cdot \mid \Lambda)\big)\right]. \tag{7}$$

When the conditional densities are Gaussian, the KL divergence inside the expectation has a closed-form expression. Its average over the latent precision, however, is generally an upper-bound surrogate for the marginal KL divergence, with a gap that can depend on the model parameters. This implies that conditional Gaussian calculations do not automatically recover the marginal KL projection, motivating the use of γ-power divergence objectives directly on the marginal heavy-tailed distributions.

2.3 The ℓ_p-Power Family and Its Gaussian Scale-Mixture Structure

We next define the ℓ_p-power family and characterize its GSM representations as a basis for diffusion modeling, building on existing scale-mixture representations.

Let $D_s := \mathrm{diag}(s)$. For $p > 0$, $\nu > 0$, $m \in \mathbb{R}^d$, and $s \in (0,\infty)^d$, define the ℓ_p-power density by

$$f^{(d)}_{p,\nu}(x \mid m, s) := \frac{C_{p,\nu,d}}{|D_s|}\left(1 + \frac{\|D_s^{-1}(x-m)\|_p^p}{\nu}\right)^{-\frac{\nu+d}{p}}, C_{p,\nu,d} := \frac{\Gamma((\nu+d)/p)}{\Gamma(\nu/p)\nu^{d/p}[2\Gamma(1+1/p)]^d}. \tag{8}$$

We denote the corresponding distribution by $F^{(d)}_{p,\nu}(m, D_s)$. This family provides a common generalization of the Student t and symmetric Pareto families, recovered at $p = 2$ and $p = 1$, respectively. The parameter p controls the shape of the equal-density contours, s sets the scale along each coordinate, and ν controls tail heaviness, with smaller ν giving heavier tails. For $\nu > p$, we choose $\gamma = -p/(\nu + d)$ in γ-power divergence, so that $(f^{(d)}_{p,\nu})^\gamma$ is affine in $\|D_s^{-1}(x-m)\|_p^p$.

Table 1: Closure properties of the ℓ_p power family under different latent precision specifications.

Precision specification	Markov path	ℓ_p-power family ? $f(x_t \mid x_{t-1})$	$f(x_t \mid x_0)$	$f(x_{t-1} \mid x_t, x_0)$
Temporally dependent ($\mathbf{\Lambda}$)	No	✗	✓	✓
Temporally independent ($\mathbf{\Lambda}_t$)	Yes	✓	✗	✗

Gaussian scale mixture representation. Gaussian noise with random precisions connects this family to diffusion paths. Definition 2.2 and Theorem 2.3 specify the range of p for which the power family admits a GSM representation.

Definition 2.2 (Gaussian scale mixture). *A centered $X \in \mathbb{R}^d$ is a GSM if $X \overset{\mathrm{d}}{=} \Lambda^{-1/2} Z$, where $Z \sim \mathcal{N}_d(0, I_d)$, $\Lambda = \mathrm{diag}(\Lambda_1, \ldots, \Lambda_d)$ is a random precision matrix that is positive definite almost surely, and Z is independent of Λ. We call X a common-precision GSM if $\Lambda_1 = \cdots = \Lambda_d$ almost surely (Andrews & Mallows, 1974).*

Theorem 2.3 (GSM Representations of power form densities). *For $c, m, p > 0$ with $mp > d$, the density on $\mathbb{R}^d$ given by $f(x) \propto (1 + c\|x\|_p^p)^{-m}$ is a GSM if and only if $0 < p \leq 2$. When $d \geq 2$, it has a common-precision representation if and only if $p = 2$.*

Although the family also admits a GSM representation for $0 < p < 1$, we focus on $1 \leq p \leq 2$ for diffusion modeling to ensure convexity of the function $x \mapsto \|x\|_p^p$.

Latent Representation and Direct Sampling The GSM representation provides a conditionally Gaussian description of the ℓ_p-power family, but its explicit construction involves tilted positive-stable variables (see Appendix B.3). For sampling, we combine the representations in Arellano-Valle & Richter (2012) to obtain the following sampler using only Gamma and Rademacher random variables.

Proposition 2.4 (Gamma and Generalized Normal Representation). *Let $0 < p \leq 2$ and $\nu > 0$. For i.i.d. Rademacher variables S_i, let $W_i \overset{\mathrm{iid}}{\sim} \mathrm{Gamma}(1/p, 1/p)$ and $U \sim \mathrm{Gamma}(\nu/p, 1/p)$, independently, where the second Gamma parameter is the rate. Then*

$$X := m + D_s(\nu/U)^{1/p}(S_i W_i^{1/p})_{i=1}^d \sim F_{p,\nu}^{(d)}(m, D_s).$$

Here, $S_i W_i^{1/p}$ are independent generalized-normal variables with density $g_p(y) \propto \exp(-|y|^p/p)$, and $(\nu/U)^{1/p}$ is their shared random scale. This direct sampler complements the GSM representation used to describe the diffusion paths below.

3 The ℓ_p-Power Heavy-Tailed Diffusion Models

Building on the GSM representation established in the previous section, we develop an ℓ_p-power diffusion model whose training objective is derived from the γ-power divergence. Proofs are provided in Appendix D. The following assumptions apply throughout this section.

Assumption 3.1 (Parameters, schedules, and moments). *Fix $1 \leq p \leq 2$, $\nu > p$, and $T \geq 2$. Assume $\mathbb{E}\|X_0\|_p^p < \infty$. The forward schedules satisfy $(\mu_0, \sigma_0) = (1, 0)$, $\sigma_t > 0$ for $1 \leq t \leq T$, $\mu_t > 0$ for $1 \leq t < T$, and $\mu_T \geq 0$. The decoder scales $\tau_0, \ldots, \tau_{T-1}$ are fixed and positive.*

3.1 ℓ_p-Power Diffusion Forward and Backward Path

To guide our choice of path distributions, we compare joint and stepwise closure properties in Table 1. The joint construction uses jointly sampled precisions to preserve ℓ_p-power marginals and bridges conditional on X_0, but generally becomes non-Markovian after integrating out the precisions. Integrating out X_0 may also break the closure property of forward transition.

In contrast, independent stepwise precisions yield Markovian ℓ_p-power transitions, but generally lose marginal and bridge closure. Products of transition densities instead admit a Dirichlet-average representation of inverse-power kernels via Feynman parametrization (Cai et al., 2025); see Appendix E.1.

We therefore adopt the construction with temporally coupled latent precisions $\mathbf{\Lambda}$, which preserves the ℓ_p-power forms of $f(x_t \mid x_0)$ and $f(x_{t-1} \mid x_t, x_0)$, enabling direct sampling of adjacent forward states conditional on X_0. Within this joint formulation, we interpret training as a γ-power divergence projection of the forward path distribution onto the reverse model family induced by joint ℓ_p-power densities. This follows the joint-projection view of variational diffusion models.

Forward and backward paths. We now use the GSM representation of joint ℓ_p-power densities to define the forward and reverse paths. In each construction, the Gaussian vector and the latent precision matrix are independent. For the forward path, they are also independent of X_0.

Lemma 3.2 (Forward path distributions). *Draw a GSM precision $\mathbf{\Lambda}^f$ for $F^{(Tn)}_{p,\nu}(0, I_{Tn})$ and $Z^f \sim \mathcal{N}_{Tn}(0, I_{Tn})$. Conditionally on X_0 and $\mathbf{\Lambda}^f$, we have*

$$X_{1:T} \mid \{X_0 = x_0, \mathbf{\Lambda}^f = \boldsymbol{\lambda}\} \sim \mathcal{N}_{Tn}\big(\mu_f(x_0), D_{s_f} \boldsymbol{\lambda}^{-1} D_{s_f}\big)\,.$$

Marginalizing $\mathbf{\Lambda}^f$ gives

$$X_{1:T} \mid \{X_0 = x_0\} \sim F^{(Tn)}_{p,\nu}(\mu_f(x_0), D_{s_f}), \qquad X_t \mid \{X_0 = x_0\} \sim F^{(n)}_{p,\nu}(\mu_t x_0, \sigma_t I_n),\ t = 1, \ldots, T.$$

For $2 \le t \le T$, the conditional distribution of X_{t-1} given (X_t, X_0) is

$$X_{t-1} \mid \{X_t = x_t, X_0 = x_0\} \sim F^{(n)}_{p,\nu+n}\left(\mu_{t-1} x_0,\ \sigma_{t-1}\left[\frac{\nu + \|(x_t - \mu_t x_0)/\sigma_t\|_p^p}{\nu + n}\right]^{1/p} I_n\right). \quad (9)$$

Likewise, we construct the reverse path from joint ℓ_p-power density. Draw a GSM precision $\mathbf{\Lambda}^r$ for $F^{(N)}_{p,\nu}(0, I_N)$ and an independent $Z^r \sim \mathcal{N}_N(0, I_N)$, and set $U_{0:T} := (\mathbf{\Lambda}^r)^{-1/2} Z^r$. Starting from $X_T := U_T$, recursively define $X_{t-1} := \mu_\theta(t, X_t) + \tau_{t-1} U_{t-1},\ t = T, \ldots, 1$. For a path $x_{0:T}$, write the standardized residual vectors $u_T := x_T$ and $u_{t-1} := (x_{t-1} - \mu_\theta(t, x_t))/\tau_{t-1}$.

Let $G_{\theta,p}$ denote the reverse model defined by the preceding recursion with $U_{0:T} \sim F^{(N)}_{p,\nu}(0, I_N)$, where $N = (T+1)n$. Its terminal prior and reverse conditional distributions are given below.

Lemma 3.3 (Reverse path distributions). *The prior of $G_{\theta,p}$ is $G_T = F^{(n)}_{p,\nu}(0, I_n)$. For $t = 1, \ldots, T$, define $k_t := (T - t + 1)n,\ S_{p,t} := \sum_{j=t}^T \|u_j\|_p^p,\ a_{p,t} := \left(\frac{\nu + S_{p,t}}{\nu + k_t}\right)^{1/p}$. The reverse conditional distributions are*

$$X_{t-1} \mid \{X_{t:T} = x_{t:T}\} \sim F^{(n)}_{p,\nu+k_t}\left(\mu_\theta(t, x_t), \tau_{t-1} a_{p,t} I_n\right). \quad (10)$$

At each reverse step, the conditional distribution depends on $x_{t:T}$ only through x_t and the accumulated residual statistic $S_{p,t}$. Thus, reverse sampling is Markovian in the augmented state $(X_t, S_{p,t})$ and the samplers only need to maintain and update this pair. Specifically, starting from $X_T \sim F^{(n)}_{p,\nu}(0, I_n)$, the stochastic sampling recursively draws (10) using the trained predictor μ_θ, with parameters $(\nu + k_t, \tau_{t-1} a_{p,t})$, and updates $S_{p,t}$ after each draw. Appendix E.3 summarizes this process.

3.2 Adjacent-state Objective via γ-Power Divergence and Denoising Surrogate

Joint projection and adjacent-state loss. Following the joint-projection interpretation of variational diffusion in (1), we formulate training as $\arg\min_{\theta \in \Theta} \mathcal{D}_{\gamma_\star, p}(F \| G_{\theta,p})$ with $\gamma_{\star,p} := -\frac{p}{\nu + (T+1)n}$. This choice matches the power exponent of $G_{\theta,p}$: the transformation from $U_{0:T}$ to $X_{0:T}$ has constant Jacobian determinant $\prod_{t=1}^T \tau_{t-1}^n$, so $G_{\theta,p}$ retains the exponent $-\frac{\nu+(T+1)n}{p}$ of the joint noise density.

The joint power form of $G_{\theta,p}$ allows the model-dependent part of this objective to be expressed through the adjacent-state loss.

Theorem 3.4 (Divergence formulations and adjacent loss).

$$\mathcal{D}_{\gamma_\star, p}(F \| G_{\theta,p}) \equiv_\theta \mathbb{E}_{t, X_0}\left[\mathbb{E}_{F_{X_{1:T} \mid X_0}}\left[\frac{1}{\nu}\left\|\frac{X_{t-1} - \mu_\theta(t, X_t)}{\tau_{t-1}}\right\|_p^p\right]\right] =: \mathcal{L}_{\mathrm{adj},p}(\theta). \quad (11)$$

Remark 3.5 (Independent stepwise reverse precisions)**.** *Alternatively, the reverse process can be constructed by sampling reverse precisions independently at each step according to the Markov kernel* $g^{\mathrm{ind}}_{\theta,t} = f^{(n)}_{p,\nu}(\cdot \mid \mu_\theta(t, x_t), \tau_{t-1}\mathbf{1}_n)$*. For* $\gamma_{\mathrm{loc},p} = -p/(\nu + n)$ *and suitable weights* $w_t > 0$*,*

$$\mathcal{L}_{\mathrm{loc},p}(\theta) := \mathbb{E}_t\Big[w_t \mathbb{E}_F\Big[\mathcal{D}_{\gamma_{\mathrm{loc},p}}\Big(\tilde{f}_{t-1|t,0} \| g^{\mathrm{ind}}_{\theta,t}\Big)\Big]\Big] \equiv_\theta \mathcal{L}_{\mathrm{adj},p}(\theta), \tag{12}$$

where conditioning arguments are suppressed and $\tilde{f}_{t-1|t,0} = f_{t-1|t,0}$ *for* $t \geq 2$*, while* $\tilde{f}_{0|1,0} = f_{0|1}(\cdot \mid X_0)$*. This changes the reverse sampling by updating* (ν, τ_{t-1})*, corresponding to independent precision draws across time steps. See Appendices E.2 and E.3 and, for* $p = 2$*, Pandey et al. (2025).*

Denoising loss with conditional mean. Motivated by Gaussian denoising loss (4), we replace the sampled target X_{t-1} in the adjacent loss with its conditional mean Y_{t-1} and examine the θ-equivalence of the adjacent and denoising loss.

Denote the conditional mean of X_t given X_0 by $Y_t := \mathbb{E}_F[X_t \mid X_0] = \mu_t X_0,\ t = 0, \ldots, T$. Replacing the sampled target X_{t-1} in the adjacent loss with Y_{t-1} yields the ℓ_p^p-based denoising loss:

$$\mathcal{L}_{\mathrm{dns},p}(\theta) := \mathbb{E}_{t,X_0}\left[\mathbb{E}_{F_{X_{1:T}|X_0}}\left[\frac{1}{\nu}\left\|\frac{Y_{t-1} - \mu_\theta(t, X_t)}{\tau_{t-1}}\right\|_p^p\right]\right]. \tag{13}$$

As in Gaussian case, a noise prediction parameterization expresses this objective as a weighted ℓ_p^p error $\|\epsilon_t - \epsilon_\theta(X_t, t)\|_p^p$, with ℓ_p-power noise. We next compare the adjacent and denoising losses across different values of p and characterize the relationship between their optimization objectives.

Theorem 3.6 (Adjacent and denoising losses)**.** *Let* $C_p := \frac{n}{T(\nu-p)} \sum_{t=2}^{T} \left(\frac{\sigma_{t-1}}{\tau_{t-1}}\right)^p$*. Then,*

$$\mathcal{L}_{\mathrm{dns},p}(\theta) \leq \mathcal{L}_{\mathrm{adj},p}(\theta) \leq \mathcal{L}_{\mathrm{dns},p}(\theta) + C_p. \tag{14}$$

For $p = 2$*, the upper bound holds with equality for every* θ*, but not in general for* $1 \leq p < 2$*.*

Thus, for $p = 2$, the joint γ-power divergence, adjacent loss, and denoising surrogate are θ-equivalent. For $1 \leq p < 2$, equivalence fails in general, i.e. $\mathcal{D}_{\gamma_{\star,p}}(F \| G_{\theta,p}) \equiv_\theta \mathcal{L}_{\mathrm{adj},p} \not\equiv_\theta \mathcal{L}_{\mathrm{dns},p}$.

3.3 Analysis of Two Training Objectives

Although adjacent prediction and denoising prediction use the same sub-quadratic loss, their different targets can yield different population-optimal predictors. For $1 \leq p \leq 2$ and finite conditional p-th moments, define the conditional Fréchet p-center by

$$\mathsf{m}_p(U \mid V = v) \in \operatorname*{arg\,min}_{m \in \mathbb{R}^n} \mathbb{E}_F\left[\|U - m\|_p^p \mid V = v\right].$$

We write $\mathsf{m}_p(U)$ for the corresponding unconditional center, which is unique for $p > 1$. Let $\epsilon_t := X_{t-1}/\mu_{t-1} - X_0$. The unrestricted optimal predictors can be chosen as

$$m^\star_{\mathrm{adj},t}(x) = \mu_{t-1}\mathsf{m}_p(X_0 + \epsilon_t \mid X_t = x), \qquad m^\star_{\mathrm{dns},t}(x) = \mu_{t-1}\mathsf{m}_p(X_0 \mid X_t = x).$$

Here, $\epsilon_t \mid (X_0 = z, X_t = x)$ follows a centered ℓ_p-power distribution whose scale depends on (z, x). At $p = 2$, the predictors coincide because $\mathbb{E}[\epsilon_t \mid X_0, X_t] = 0$. For $1 \leq p < 2$, however, the perturbation can shift the conditional p-center. This allows the two predictors to respond differently to rare components of the conditional signal distribution.

For the following comparison, fix $1 < p < 2$ and a time–observation pair (t, x) with $t \geq 2$, and suppress conditioning on $X_t = x$. Throughout this comparison, each additive-noise representation $W + s(W)V$ uses an auxiliary noise vector $V \sim F^{(n)}_{p,\nu+n}(0, I_n)$ independent of the W. Here $s : \mathbb{R}^n \to (0, \infty)$ denotes the conditional scale determined by Lemma 3.2.

Sensitivity to rare conditional components. For analytical tractability, we model the concentrated dominant component as a point mass at c and consider

$$Y_\pi \sim (1 - \pi)\delta_c + \pi Q, \quad c \in \mathbb{R}^n, \quad 0 < \pi < 1,$$

keeping Q and s fixed as $\pi \downarrow 0$. Let $U \sim Q$ satisfy $\mathbb{E}\|U\|_2^2 < \infty$ and $\mathbb{E}[s(U)^p] < \infty$. Write $h := \mathbb{E}[U - c]$ and define $d_\pi := \mathsf{m}_p(Y_\pi)$, $a_\pi := \mathsf{m}_p(Y_\pi + s(Y_\pi)V)$, and $\mathcal{E}_\pi(m) := \mathbb{E}\|m - Y_\pi\|_2^2$.

Proposition 3.7 (Rare-component response and reconstruction risk). *There exist coefficients $q, b \in \mathbb{R}^n$, explicitly given in Appendix D.4, such that, as $\pi \downarrow 0$,*

$$d_\pi = c + \pi^{1/(p-1)} q + o(\pi^{1/(p-1)}), \qquad a_\pi = c + \pi b + o(\pi).$$

Moreover, $\mathcal{E}_\pi(a_\pi) - \mathcal{E}_\pi(d_\pi) = \pi^2\big(\|b-h\|_2^2 - \|h\|_2^2\big) + o(\pi^2)$.

This implies that, for sufficiently small π, adjacent prediction improves reconstruction when its greater sensitivity to rare components better tracks the conditional mean shift.

Reconstruction trade-offs across subspaces. To identify when the adjacent predictor a_π has higher reconstruction risk, decompose its first-order response b into components within and orthogonal to a reference subspace. For $1 \le r < n$, let P be a fixed rank-r orthogonal projector satisfying $Ph = h$ and set $P^\perp := I_n - P$.

Corollary 3.8 (Denoising risk comparison). *Suppose $\|h\|_2^2/r \le H^2$ and $\|P^\perp b\|_2^2/(n-r) \ge B^2$ for some $H \ge 0$ and $B > 0$. If $n/r > 1 + H^2/B^2$, then $\mathcal{E}_\pi(d_\pi) < \mathcal{E}_\pi(a_\pi)$ for sufficiently small $\pi > 0$.*

This gives a sufficient condition for denoising to achieve lower reconstruction risk for sufficiently small π. For fixed H and B, the condition holds when the dimension ratio n/r is sufficiently large.

Effect of the loss exponent on denoising. For $\eta \ge 0$, let $\boldsymbol{x}^{(\eta)} \in (0,1)^D$ be a fixed clean signal satisfying $\|\boldsymbol{x}^{(\eta)} - \boldsymbol{x}^{(0)}\|_\infty \le \eta$. Observe $N \ge 3$ i.i.d. corrupted vectors $\boldsymbol{Y}^{(i,\eta)} \in [0,1]^D$, with N odd. Each coordinate marginal is nondegenerate, assigns positive mass to $\{0,1\}$, and has no atoms outside $\{0, x_j^{(\eta)}, 1\}$. For $p \in [1,2]$, let $\widehat{\boldsymbol{x}}_{N,p}^{(\eta)}$ denote their unique empirical p-center and define the reconstruction risk $\mathcal{R}_N^{(\eta)}(p) := D^{-1}\mathbb{E}\|\widehat{\boldsymbol{x}}_{N,p}^{(\eta)} - \boldsymbol{x}^{(\eta)}\|_2^2$, where the expectation is over corruption.

Proposition 3.9 (Intermediate exponents under bounded corruption). *For each fixed $\eta \ge 0$, there exists $\delta_\eta \in (0,1)$ such that $\mathcal{R}_N^{(\eta)}(p) < \mathcal{R}_N^{(\eta)}(1)$ for every $p \in (1, 1+\delta_\eta)$. If additionally $\mathcal{R}_N^{(\eta)}(1) \le \mathcal{R}_N^{(\eta)}(2)$, then $\operatorname{argmin}_{p\in[1,2]} \mathcal{R}_N^{(\eta)}(p) \subset (1,2)$.*

For bounded observations such as images, the proposition motivates studying intermediate loss exponents under clipped additive Gaussian noise or salt-and-pepper corruption. Furthermore, our experiments empirically examine how the reconstruction-optimal exponent varies with the salt-and-pepper corruption rate ρ and clean color variation η (Figures 1 and 2).

4 Experiments

In this section, we conduct two sets of experiments to examine how the choice of training objective affects the performance of the proposed diffusion framework. The first compares the adjacent (Adj.) and denoising (Dns.) objectives in terms of rare-mode coverage. The second evaluates the practical utility of the dns. in high-dimensional applications, comparing it with Gaussian DDPM and Student's t-EDM (Pandey et al., 2025). To isolate the effect of the loss exponent p, we fix the degrees-of-freedom parameter ν. Experimental details are provided in Appendix F.

4.1 Rare-Mode Coverage and Rare-Word Restoration of the Adjacent ℓ_p^p-Loss

Rare-mode coverage on heavy-tailed noise. We evaluate rare-mode coverage on an imbalanced mixture:

$$X = \mu_C + 0.55\, Z, \quad C \sim \mathrm{Cat}(\pi_{1:K}), \quad Z \sim S_{1.7}(0, I_{30}), \quad C \perp Z. \tag{15}$$

where $S_\alpha(0, I_n)$ denotes a centered isotropic α-stable distribution. The mixture weights follow a power law, $\pi_k \propto k^{-\tau}$ for $k = 1, \dots, K$, where τ controls the degree of imbalance. We set $K = 16$ and $\tau = 1.2$. The component centers $\{\mu_k\}_{k=1}^K$ lie on a sphere of radius 7.5 in a randomly chosen six-dimensional linear subspace of $\mathbb{R}^{30}$.

Table 2 reports total variation–based discrepancy (TV) and head- and rare-mode recall; metric definitions are provided in Appendix F.2. The adj. objective improves all three metrics at the tested $p < 2$, with larger gains at smaller p. At $p = 2$, the objectives are equivalent up to a parameter-independent constant (Theorem 3.6). Gaussian DDPM remains competitive, whereas

Table 2: Rare mode coverage comparison.

Setting	Loss	TV↓	Rare rec.	Head rec.
$p = 1.2$	Adj.	$\mathbf{0.012 \pm 0.003}$	$\mathbf{0.967 \pm 0.006}$	$\mathbf{0.991 \pm 0.007}$
	Dns.	0.062 ± 0.001	0.832 ± 0.007	0.930 ± 0.001
$p = 1.6$	Adj.	$\mathbf{0.013 \pm 0.001}$	$\mathbf{0.958 \pm 0.009}$	$\mathbf{0.990 \pm 0.006}$
	Dns.	0.027 ± 0.003	0.928 ± 0.016	0.972 ± 0.006
$p = 2.0$	Adj.	0.029 ± 0.014	0.909 ± 0.045	$\mathbf{0.973 \pm 0.015}$
	Dns.	$\mathbf{0.023 \pm 0.005}$	$\mathbf{0.947 \pm 0.025}$	0.971 ± 0.003
DDPM		0.020 ± 0.005	0.969 ± 0.039	0.983 ± 0.006
DLPM ($\alpha = 1.7$)		0.075 ± 0.007	0.733 ± 0.021	0.912 ± 0.011



(a) $\eta = 0$ (b) $\eta = 0.05$ (c) $\eta = 0.10$

Figure 1: 4-colored MNIST varying η.

Table 3: Rare-recall gap comparison $\Delta R = R_{\text{adj}} - R_{\text{dns}}$.

n	n/r	$\alpha = 1.7$			$\alpha = 2.0$		
		$p = 1.2$	$p = 1.6$	$p = 1.8$	$p = 1.2$	$p = 1.6$	$p = 1.8$
32	1	$+0.135$	$+0.030$	$+0.009$	$+0.128$	$+0.068$	$+0.027$
	4	$+0.140$	$+0.038$	$+0.021$	$+0.144$	$+0.072$	$+0.040$
	16	$+0.072$	$+0.013$	$+0.006$	$+0.093$	$+0.011$	$\mathbf{-0.001}$
64	1	$+0.070$	$+0.026$	$\mathbf{-0.018}$	$+0.084$	$+0.024$	$\mathbf{-0.012}$
	4	$+0.062$	$+0.019$	$\mathbf{-0.036}$	$+0.062$	$+0.016$	$\mathbf{-0.002}$
	16	$+0.088$	$+0.024$	$\mathbf{-0.005}$	$+0.068$	$+0.011$	$\mathbf{-0.013}$
256	1	$+0.054$	$\mathbf{-0.010}$	$+0.019$	$\mathbf{-0.054}$	$\mathbf{-0.111}$	$\mathbf{-0.035}$
	4	$+0.026$	$\mathbf{-0.054}$	$+0.038$	$\mathbf{-0.056}$	$\mathbf{-0.140}$	$\mathbf{-0.076}$
	16	$+0.077$	$+0.015$	$+0.040$	$\mathbf{-0.060}$	$\mathbf{-0.126}$	$\mathbf{-0.051}$

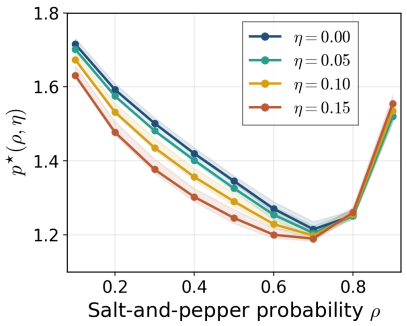


Figure 2: $p^\star$-curve varying (ρ, η).

Table 4: Character-level typo correction on WikiText-103: overall accuracy and head/tail word recall.

Model	**F0.5 (↑)**	**Acc. (↑)**	**Acc. (sentence) (↑)**	**Head rec. (↑)**	**Tail rec. (↑)**
Norvig-10	0.665	0.614	0.117	0.483	0.113
$p = 1.2$, Adj.	$\mathbf{0.836 \pm 0.004}$	$\mathbf{0.864 \pm 0.005}$	$\mathbf{0.306 \pm 0.002}$	$\mathbf{0.991 \pm 0.003}$	$\mathbf{0.345 \pm 0.003}$
$p = 1.2$, Dns.	0.068 ± 0.017	0.477 ± 0.002	0.002 ± 0.000	0.103 ± 0.025	0.002 ± 0.001
$p = 1.8$, Adj.	0.822 ± 0.001	0.854 ± 0.006	0.273 ± 0.004	0.989 ± 0.003	0.316 ± 0.002
$p = 1.8$, Dns.	0.480 ± 0.079	0.630 ± 0.020	0.020 ± 0.015	0.904 ± 0.066	0.078 ± 0.034
t-EDM	0.711 ± 0.002	0.731 ± 0.001	0.176 ± 0.003	0.988 ± 0.005	0.179 ± 0.003
DDPM	0.716 ± 0.003	0.744 ± 0.004	0.251 ± 0.002	0.973 ± 0.002	0.249 ± 0.005

DLPM (Shariatian et al., 2025) performs worse in this setting, suggesting that heavy-tailed diffusion noise alone does not guarantee better rare-mode coverage. Table 3 compares rare-mode recall between Adj. and Dns. across dimension and component-noise settings. Adj. generally achieves higher rare-mode recall at $n = 32$, whereas Dns. is favored across all displayed settings at $n = 256$ with $\alpha = 2.0$. Still, Adj. remains competitive even at $n = 256$ for $\alpha = 1.7$ and the comparison shows no consistent monotonic trend in n/r. These results show that relative rare-mode coverage can vary with dimension, component-noise distribution and loss exponent even within the same task.

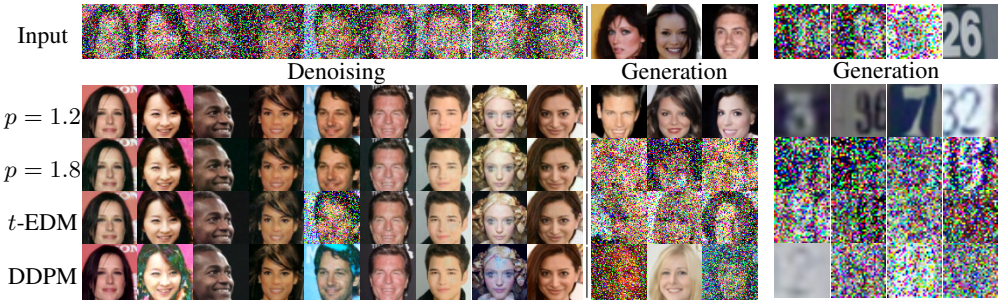
Rare-word restoration in typo correction. We evaluate rare-word restoration on WikiText-103 (Merity et al., 2016), motivated by the power law of word frequencies. In each sentence, we select a fraction f of words with at least four letters and replace a fraction q of each selected word's letters, with $f = q = 0.5$. Models reconstruct the clean sentence. Our frequent-based baseline, Norvig-10 (Norvig, 2007), corrects out-of-vocabulary words using same length training words within 10 substitutions, preferring fewer substitutions and then higher frequency. Excluding proper nouns and words occurring fewer than five times in training, we rank test occurrences by frequency and define the top and bottom 10% as head and tail. Table 4 shows that diffusion models achieve high head accuracy, with differences concentrated in the tail, where ℓ_p diffusion largest gains over DDPM and t-EDM. Norvig-10 ignores sentence context and performs substantially worse in both groups.

4.2 Denoising Applications of Denoising ℓ_p^p-Loss

Optimal loss exponents on four-colored MNIST. We begin with a controlled experiment on four-colored MNIST to illustrate the role of intermediate loss exponents. Figure 1 shows examples with varying η in Proposition 3.9, which controls the magnitude of color variation around the nominal colors. After applying salt-and-pepper noise with probability ρ, we compute empirical p-centers for $p \in [1, 2]$ and select $p^\star(\rho, \eta)$ to minimize reconstruction MSE against the clean image. Figure 2 shows that $p^\star$ lies in $(1, 2)$ across a broad range of settings, consistent with Proposition 3.9. As ρ

Table 5: Quantitative results on denoising and unconditional generation on various datasets.

			Denoising			Generation		
Dataset	**Model**	**loss**	**PSNR (↑)**	**SSIM (↑)**	**LPIPS (↓)**	**FID (↓)**	**Precision (↑)**	**Recall (↑)**
Omniglot	$p = 1.2$	Dns.	$\mathbf{37.010 \pm 0.057}$	$\mathbf{0.994 \pm 0.000}$	$\mathbf{0.001 \pm 0.000}$	$\mathbf{43.676 \pm 1.163}$	$\mathbf{0.603 \pm 0.023}$	0.503 ± 0.010
	$p = 1.8$	Dns.	36.311 ± 0.141	0.992 ± 0.001	0.001 ± 0.000	198.586 ± 31.738	0.270 ± 0.077	0.650 ± 0.017
	t-EDM		36.601 ± 0.256	0.990 ± 0.003	0.001 ± 0.000	283.640 ± 9.288	0.126 ± 0.015	$\mathbf{0.753 \pm 0.030}$
	DDPM		36.070 ± 0.384	0.992 ± 0.001	0.001 ± 0.000	191.062 ± 44.432	0.328 ± 0.106	0.733 ± 0.017
CIFAR10	$p = 1.2$	Dns.	$\mathbf{35.735 \pm 1.421}$	0.970 ± 0.020	0.002 ± 0.001	$\mathbf{126.016 \pm 17.512}$	$\mathbf{0.406 \pm 0.024}$	$\mathbf{0.516 \pm 0.016}$
	$p = 1.8$	Dns.	29.075 ± 9.184	0.806 ± 0.027	0.028 ± 0.005	308.582 ± 8.767	0.100 ± 0.013	0.332 ± 0.017
	t-EDM		26.405 ± 1.620	0.711 ± 0.017	0.071 ± 0.011	308.562 ± 36.376	0.222 ± 0.177	0.322 ± 0.280
	DDPM		35.440 ± 0.603	$\mathbf{0.971 \pm 0.008}$	$\mathbf{0.002 \pm 0.000}$	217.222 ± 15.373	0.273 ± 0.048	0.490 ± 0.053
SVHN	$p = 1.2$	Dns.	$\mathbf{45.944 \pm 0.235}$	$\mathbf{0.995 \pm 0.001}$	$\mathbf{0.001 \pm 0.000}$	$\mathbf{163.157 \pm 13.460}$	0.424 ± 0.027	0.485 ± 0.015
	$p = 1.8$	Dns.	39.568 ± 5.152	0.951 ± 0.007	0.008 ± 0.002	355.903 ± 49.541	0.099 ± 0.066	$\mathbf{0.540 \pm 0.098}$
	$p = 1.8$	Adj.	36.568 ± 4.146	0.941 ± 0.010	0.028 ± 0.004	195.903 ± 12.898	$\mathbf{0.554 \pm 0.090}$	0.080 ± 0.002
	t-EDM		28.423 ± 0.285	0.831 ± 0.002	0.150 ± 0.003	332.691 ± 9.072	0.078 ± 0.014	0.511 ± 0.057
	DDPM		27.024 ± 8.423	0.656 ± 0.015	0.141 ± 0.022	303.928 ± 112.558	0.202 ± 0.177	0.321 ± 0.089
CelebA	$p = 1.2$	Dns.	$\mathbf{39.278 \pm 3.067}$	$\mathbf{0.965 \pm 0.012}$	$\mathbf{0.013 \pm 0.002}$	$\mathbf{37.176 \pm 2.316}$	$\mathbf{0.817 \pm 0.039}$	$\mathbf{0.256 \pm 0.025}$
	$p = 1.8$	Dns.	32.327 ± 4.942	0.894 ± 0.016	0.043 ± 0.010	349.356 ± 76.549	0.039 ± 0.066	0.119 ± 0.019
	$p = 1.8$	Adj.	28.560 ± 5.255	0.673 ± 0.021	0.095 ± 0.025	373.324 ± 50.235	0.004 ± 0.002	0.005 ± 0.002
	t-EDM		17.440 ± 7.932	0.303 ± 0.042	0.680 ± 0.050	402.583 ± 30.543	0.000 ± 0.000	0.000 ± 0.000
	DDPM		18.654 ± 2.250	0.519 ± 0.030	0.122 ± 0.014	267.229 ± 72.487	0.124 ± 0.072	0.235 ± 0.017



(a) CelebA (b) SVHN

Figure 3: Qualitative reconstruction and generation on CelebA and SVHN (ℓ_p with Denoising loss.)

increases, $p^\star$ initially decreases before rising again at high corruption levels. Larger η generally favors smaller exponents at low to moderate corruption levels.

Image denoising and generation. We conducted denoising and unconditional generation experiments with ℓ_p-power diffusion on Omniglot (Lake et al., 2015), CIFAR-10 (Krizhevsky, 2009), SVHN (Netzer et al., 2011), and CelebA (Liu et al., 2015). Restoration was evaluated using PSNR, SSIM, and LPIPS (Zhang et al., 2018), and unconditional generation using FID (Heusel et al., 2017) and improved precision and recall (Sajjadi et al., 2018).

We apply per-channel salt-and-pepper noise ($q = 0.5$) once to a fraction $\rho = 0.75$ of training images and keep this contamination fixed. Training images serve as diffusion targets, with clean counterparts of contaminated images excluded from training; only observation corruption is resampled at each step. A single model learns both $p(X \mid Y)$ and $p(X)$ by zeroing observations for a fraction $\lambda = 0.5$ of samples, without classifier-free guidance at inference. Despite receiving no direct supervision for corruption removal, ℓ_p-power models assign probability mass to cleaner outputs in both modes, consistent with the population optimum induced by the ℓ_p^p loss.

5 Conclusion

We developed an ℓ_p-power heavy-tailed diffusion framework using Gaussian scale mixtures to construct random-precision paths for $1 \leq p \leq 2$. Joint γ-power divergence projection yields an adjacent-state prediction loss equivalent to conventional denoising at $p = 2$, Our analysis characterizes how these objectives differ in their sensitivity to rare components and their robustness in denoising. Empirical results support this distinction: adjacent-state training better preserves rare modes under heavy-tailed noise, whereas denoising training achieves better denoising accuracy. Further discussion and limitations are provided in Appendix G.

AI USE STATEMENT

In this work, we used generative AI tools to polish the writing, verify mathematical statements and proofs, and assist with implementing experimental code and conducting experiments. All authors reviewed the AI-assisted text, code, and experimental results. We take responsibility for the final content of this work, including all text, claims, code, and results produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The appendices include proofs of the theoretical results, together with the assumptions under which they hold, as well as details of the experimental settings. To support reproducibility for experiments, we also provide the Python code in the supplementary materials.

REFERENCES

Shun-ichi Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

D. F. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(1):99–102, 1974.

Reinaldo B. Arellano-Valle and Wolf-Dieter Richter. On skewed continuous $l_{n,p}$-symmetric distributions. *Chilean Journal of Statistics*, 3(2):193–212, 2012.

Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*, 2018.

Aymene Mohammed Bouayed, Samuel Deslauriers-Gauthier, Adrian Iaccovelli, and David Naccache. Conditional-t^3vae: Equitable latent space allocation for fair generation, 2025.

Diana Cai, Robert M. Gower, David Blei, and Lawrence K. Saul. Fisher meets feynman: score-based variational inference with a product of experts. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Mincheol Cho, Yedarm Seong, and Joong-Ho Won. Pareto variational autoencoder. In *The Fourteenth International Conference on Learning Representations*, 2026.

Quan Dao, Binh Ta, Tung Pham, and Anh Tran. A high-quality robust diffusion framework for corrupted dataset. In *Computer Vision – ECCV 2024*, pp. 107–123. Springer, 2024.

Jacob Deasy, Nikola Simidjievski, and Pietro Liò. Heavy-tailed denoising score matching. *arXiv preprint arXiv:2112.09788*, 2021.

Shinto Eguchi. Pythagoras theorem in information geometry and applications to generalized linear models. In *Handbook of Statistics*, volume 45, pp. 15–42. Elsevier, 2021.

Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, 2 edition, 1999.

Alan B. Forsythe. Robust estimation of straight line regression coefficients by minimizing pth power deviations. *Technometrics*, 14(1):159–166, 1972.

Robert Gruhlke, Valentin Resseguier, and Merveille Talla. Multiplicative diffusion models: Beyond gaussian latents. In *The Fourteenth International Conference on Learning Representations*, 2026.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6626–6637, 2017.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020b.

Hamid Jalalzai, Pierre Colombo, Chloé Clavel, Eric Gaussier, Giovanna Varni, Emmanuel Vignon, and Anne Sabourin. Heavy-tailed representations, text polarity classification & data augmentation. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 4295–4307. Curran Associates, Inc., 2020.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

T. Anderson Keller and Max Welling. Topographic VAEs learn equivariant capsules. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Artem Khrapov, Vadim Popov, Tasnima Sadekova, Assel Yermekova, and Mikhail Kudinov. Improving diffusion models's data-corruption resistance using scheduled pseudo-huber loss, 2024.

Juno Kim, Jaehyuk Kwon, Mincheol Cho, Hyunjong Lee, and Joong-Ho Won. t^3-variational autoencoder: Learning heavy-tailed data with student's t and power divergence. In *The Twelfth International Conference on Learning Representations*, 2024.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *The Second International Conference on Learning Representations*, 2013.

Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Advances in Neural Information Processing Systems 34*, pp. 21696–21707, 2021.

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

Nicolas Lafon, Philippe Naveau, and Ronan Fablet. A vae approach to sample multivariate extremes. *Journal of Statistical Computation and Simulation*, 96(12):2927–2963, 2026.

Brenden Lake, Ruslan Salakhutdinov, and Joshua Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350:1332–1338, 12 2015.

Ji Li and Chao Wang. Integrating reweighted least squares with plug-and-play diffusion priors for noisy image restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6163–6171, 2026. doi: 10.1609/aaai.v40i8.37541.

Lei Li and Yuexiao Dong. Robust diffusion models via divergence-induced weighted denoising, 2026.

Weihan Li, Linyun Zhou, Jian Yang, Shengxuming Zhang, Xiangtong Du, Xiuming Zhang, Jing Zhang, Chaoqing Xu, Mingli Song, and Zunlei Feng. L-diffusion: Laplace diffusion for efficient pathology image segmentation. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 36954–36973. PMLR, 13–19 Jul 2025.

Qi Lian, Yu Qi, and Yueming Wang. Cauchy diffusion: A heavy-tailed denoising diffusion probabilistic model for speech synthesis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24549–24557, Apr. 2025.

Haoyu Liu, Tingyu Zhu, Nanshan Jia, Jinghai He, and Zeyu Zheng. Learning to simulate from heavy-tailed distribution via diffusion model. *Operations Research*, 2026.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016.

Mehrdad Moradi and Kamran Paynabar. Rddpm: Robust denoising diffusion probabilistic model for unsupervised anomaly segmentation. In *Proceedings of the ICCV 2025 Workshop on Vision-based Industrial InspectiON (VISION)*, 2025.

Eliya Nachmani, Robin San Roman, and Lior Wolf. Denoising diffusion gamma models. *arXiv preprint arXiv:2110.05948*, 2021.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171. PMLR, 18–24 Jul 2021.

Peter Norvig. How to write a spelling corrector, 2007. URL `https://www.norvig.com/spell-correct.html`.

Kushagra Pandey, Jaideep Pathak, Yilun Xu, Stephan Mandt, Michael Pritchard, Arash Vahdat, and Morteza Mardani. Heavy-tailed diffusion models. In *International Conference on Learning Representations*, 2025.

Mario Pascual-González, Ariadna Jiménez-Partinen, R. M. Luque-Baena, Fátima Nagib-Raya, and Ezequiel López-Rubio. SLIM-Diff: Shared latent image-mask diffusion with L_p loss for data-scarce epilepsy FLAIR MRI, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Vadim Popov, Assel Yermekova, Tasnima Sadekova, Artem Khrapov, and Mikhail Kudinov. Improved sampling algorithms for Lévy-Itô diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018.

René L. Schilling, Renming Song, and Zoran Vondraček. *Bernstein Functions: Theory and Applications*, volume 37 of *De Gruyter Studies in Mathematics*. De Gruyter, 2 edition, 2012.

Dario Shariatian, Umut Simsekli, and Alain Durmus. Heavy-tailed diffusion with denoising lévy probabilistic models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Hiroshi Takahashi, Tomoharu Iwata, Yuki Yamanaka, Masanori Yamada, and Satoshi Yagi. Student-t variational autoencoder for robust density estimation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 2696–2702, 2018.

Lucas Theis and Matthias Bethge. Generative image modeling using spatial LSTMs. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Lucas Theis, Reshad Hosseini, and Matthias Bethge. Mixtures of conditional gaussian scale mixtures applied to multiscale image representations. *PLOS ONE*, 7(7):e39857, 2012.

Martin J. Wainwright and Eero P. Simoncelli. Scale mixtures of gaussians and the statistics of natural images. In *Advances in Neural Information Processing Systems*, volume 12, 1999.

Keito Wakatsuki and Hideaki Shimazaki. Self-regulating annealing in heavy-tailed diffusion models, 2026.

Mike West. On scale mixtures of normal distributions. *Biometrika*, 74(3):646–648, 1987.

Zhouhao Yang, Yezhen Wang, Kenji Kawaguchi, Vladimir Braverman, and Haoyang Cao. Heavy-tailed flow matching via random clocks. *arXiv preprint arXiv:2607.13841*, 2026.

Eunbi Yoon, Keehun Park, Sungwoong Kim, and Sungbin Lim. Score-based generative models with lévy processes. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Heegeon Yoon and Heeyoung Kim. Multimodal deep generative model for semi-supervised learning under class imbalance. *Technometrics*, 68(3):530–542, 2026. ISSN 0040-1706.

Likun Zhang, Xiaoyu Ma, Christopher K. Wikle, and Raphaël Huser. Fast and flexible emulation of spatial extremes processes via variational autoencoders. *Journal of the American Statistical Association*, 0(0):1–14, 2026.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

CONTENTS

1 Introduction **1**

2 Theoretical Background **2**

2.1 Divergence Projection in Variational Diffusion Models 2

2.2 The γ-Power Divergence and Power Families 3

2.3 The ℓ_p-Power Family and Its Gaussian Scale-Mixture Structure 3

3 The ℓ_p-Power Heavy-Tailed Diffusion Models **4**

3.1 ℓ_p-Power Diffusion Forward and Backward Path 4

3.2 Adjacent-state Objective via γ-Power Divergence and Denoising Surrogate 5

3.3 Analysis of Two Training Objectives 6

4 Experiments **7**

4.1 Rare-Mode Coverage and Rare-Word Restoration of the Adjacent ℓ_p^p-Loss 7

4.2 Denoising Applications of Denoising ℓ_p^p-Loss 8

5 Conclusion **9**

A Related Work **15**

B Details on Theoretical Background **17**

B.1 Gaussian DDPM Posterior and Noise-Prediction Details 17

B.2 Information Geometry of the γ-Power Divergence 17

B.3 Explicit Positive-stable Diagonal GSM Representation 18

C Proofs for Section 2 **21**

C.1 Proof of Theorem 2.3 21

C.2 Proof of Proposition 2.4 23

D Proofs for Section 3 **24**

D.1 Proofs of Lemmas 3.2 and 3.3 24

D.2 Proof of Theorem 3.4 26

D.3 Proof of Theorem 3.6 27

D.4 Proof of Proposition 3.7 and Corollary 3.8 28

D.5 Proof of Proposition 3.9 30

E Details on ℓ_p-Power Diffusion Models **34**

E.1 Feynman Parametrization for Independent Latent Precisions 34

E.2 Justification of the Reverse Markov Path via Conditional Divergence Summation . 35

E.3 Sampling Algorithms for ℓ_p-Power Diffusion 36

F Experimental Details **37**

F.1 Common Setup 37

F.2 Rare-mode Coverage under Heavy-tailed Noise 37

F.3 Rare-word Restoration in Typo Correction 39

F.4 Optimal $p^\star$ Curve of Four-color MNIST Dataset 41

F.5 Image Denoising and Unconditional Generation 42

G Discussion and Limitations **44**

A Related Work

Gaussian scale mixtures in machine learning. Gaussian scale mixtures (GSMs) have been used in statistical image modeling. Wainwright & Simoncelli (1999) use GSMs to model heavy-tailed wavelet marginals and dependencies among neighboring coefficients in natural images. Theis et al. (2012) develop directed multiscale image models based on mixtures of conditional GSMs.

GSMs have also found applications in deep learning. Theis & Bethge (2015) combine mixtures of conditional GSMs with spatial recurrent networks for image generation. For learned image compression, Ballé et al. (2018) use a hyperlatent to predict Gaussian scales for latent image representations. In variational autoencoders, Takahashi et al. (2018) obtain a Student-t decoder likelihood by marginalizing a Gamma-distributed precision, while Keller & Welling (2021) construct a topographic Product of Student-t prior through shared latent variables governing local scales. More recently, Yang et al. (2026) introduce random-clock flow matching with conditionally Gaussian sources. Our work uses GSM representations to construct heavy-tailed diffusion paths and analyze the training objectives induced by reverse scale coupling.

Heavy-tailed generative modeling with γ-power divergence. γ-Power divergence (Eguchi, 2021; Kim et al., 2024) provide an alternative to KL-based variational objectives for heavy-tailed generative models. Kim et al. (2024) formulate a Student-t VAE by replacing the KL divergence between joint distributions with the γ-power divergence. Bouayed et al. (2025) extend this framework using class-conditional Student-t joint priors for balanced generation from imbalanced data, while Yoon & Kim (2026) combine Student-t modeling with γ-power divergence for multimodal semi-supervised classification under class imbalance. Cho et al. (2026) use symmetric Pareto distributions to derive objectives with ℓ_1 loss terms. For diffusion models, Pandey et al. (2025) combine Student-t forward and reverse kernels with local γ-power divergence terms, yielding a quadratic denoising objective. This is the closest objective-level precedent for our $p = 2$ case.

Non-Gaussian and heavy-tailed diffusion. Diffusion models have been extended beyond Gaussian perturbations through several noise constructions. Nachmani et al. (2021) construct diffusion models with Gamma noise, while Deasy et al. (2021) extend denoising score matching to generalized Gaussian perturbations. Li et al. (2025) combine Laplace feature diffusion with contrastive learning for pathology segmentation. Beyond additive perturbations, Gruhlke et al. (2026) develop multiplicative diffusion models that preserve the distribution of data norms and yield non-Gaussian latent distributions.

Heavy-tailed diffusion models have also been explored to improve the modeling of tail behavior and rare events. One line of work uses Student-t distributions: Pandey et al. (2025) develop Student-t diffusion models for heavy-tail estimation. Lian et al. (2025) develop Cauchy Diffusion for speech synthesis using the $\nu = 1$ Student-t distribution and KL-based scale loss. Building on Student-t diffusion, Wakatsuki & Shimazaki (2026) develop an SDE sampler and analyze the resulting self-regulating annealing mechanism. In a score-based formulation, Liu et al. (2026) use the conditional scores of Student-t perturbation kernels as denoising score-matching targets and employ annealed Langevin dynamics to simulate heavy-tailed distributions.

Another line of work uses α-stable noise. Yoon et al. (2023) employ isotropic α-stable Lévy processes, while Popov et al. (2025) develop stochastic samplers that preserve the corresponding time-marginal

distributions. Shariatian et al. (2025) formulate discrete-time Lévy diffusion models using conditional Gaussian representations with KL divergence objective.

Robust objectives in diffusion modeling and restoration. Alternative distribution-matching criteria and residual penalties have been explored for robust diffusion training and image restoration. Dao et al. (2024) introduce an unbalanced optimal transport formulation for robust reverse-process learning in a DDGAN-based framework. Khrapov et al. (2024) introduce a scheduled pseudo-Huber loss for diffusion training on contaminated image and audio data, while Moradi & Paynabar (2025) use Huber and least-trimmed-squares objectives for unsupervised anomaly segmentation. Li & Dong (2026) derive denoising objectives from local conditional f-divergences between Gaussian kernels with matched covariances, yielding residual-dependent gradient weights.

ℓ_p^p-based objective also appear in recent applications: Pascual-González et al. (2026) evaluate ℓ_p^p training losses and prediction parameterizations in Gaussian diffusion models for joint medical image–mask generation, while Li & Wang (2026) combine ℓ_q^q measurement-fidelity terms with pretrained Gaussian diffusion priors for restoration under non-Gaussian noise. These applications retain Gaussian diffusion processes and do not derive their power penalties from heavy-tailed diffusion path objectives. On the other hand, our work derives ℓ_p^p training objectives from such path models and characterizes the population-level relationship between adjacent-state prediction and clean-target denoising.

B Details on Theoretical Background

This section provides details on Gaussian DDPM posteriors and noise prediction, the information geometric details on γ-power divergence, and explicit derivation of positive-stable Gaussian scale mixture representation.

B.1 Gaussian DDPM Posterior and Noise-Prediction Details

The main text states the path-projection view of diffusion models and the KL chain-rule decomposition in Section 2. We record here the Gaussian DDPM identities that connect the complete-data ELBO to the usual noise-prediction parameterization.

With the Gaussian forward schedule $(\alpha_t, \beta_t, \bar{\alpha}_t)$ from Section 2, the one-time conditional distribution is

$$f(x_t \mid x_0) = \mathcal{N}_n\left(\sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I_n\right).$$

Conditioning the forward path on x_0 gives the DDPM ELBO decomposition. When the displayed quantities are finite,

$$\begin{aligned}\mathcal{D}_{\mathrm{KL}}(F\|G_\theta) = C_{\mathrm{term}} + \sum_{t=2}^{T} \mathbb{E}_{f(x_0,x_t)}\left[\mathcal{D}_{\mathrm{KL}}\big(f(x_{t-1} \mid x_t, x_0)\|g_\theta(x_{t-1} \mid x_t)\big)\right] \\ - \mathbb{E}_{f(x_0,x_1)}\left[\log g_\theta(x_0 \mid x_1)\right], \end{aligned} \tag{16}$$

where

$$C_{\mathrm{term}} = -\mathcal{H}(f_{\text{data}}) + \mathbb{E}_{X_0}\mathcal{D}_{\mathrm{KL}}\big(f(x_T \mid X_0)\|g_T(x_T)\big),$$

and $\mathcal{H}(f_{\text{data}}) = -\mathbb{E}_{X_0\sim f_{\text{data}}}\big[\log f_{\text{data}}(X_0)\big]$ is the differential entropy of the data distribution. For a fixed data distribution f_{data} and terminal prior g_T, $\mathcal{H}(f_{\text{data}})$ is independent of θ and therefore constitutes a constant term during model training. Thus the $t \geq 2$ terms use posterior KL divergences conditioned on X_0, while the $t = 1$ term is the reconstruction likelihood. This is the DDPM ELBO decomposition specialized from the path projection, and uses more conditioning information than the reverse conditional $f_{t-1|t}(\cdot \mid X_t)$ in (2.)

The Gaussian posterior in the $t \geq 2$ KL terms is

$$f(x_{t-1} \mid x_t, x_0) = \mathcal{N}_n\left(\widetilde{\mu}_t, \widetilde{\beta}_t I_n\right), \qquad \widetilde{\mu}_t = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon\right), \qquad \widetilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t. \tag{17}$$

Suppose that $g_\theta(x_{t-1} \mid x_t) = \mathcal{N}_n\big(\mu_\theta(x_t,t), \sigma_t^2 I_n\big)$, with fixed reverse variance σ_t^2. Up to terms independent of θ, substituting (17) into each posterior KL term gives

$$\frac{1}{2\sigma_t^2}\left\|\widetilde{\mu}_t - \mu_\theta(x_t,t)\right\|_2^2.$$

Under the standard noise-prediction parameterization

$$\mu_\theta(x_t,t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t,t)\right),$$

this becomes

$$w_t\left\|\varepsilon - \epsilon_\theta(x_t,t)\right\|_2^2, \qquad w_t = \frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}.$$

Uniformly averaging these weighted terms over $t \sim \mathrm{Unif}\{1,\ldots,T\}$ gives the denoising objective in (4); for the Gaussian decoder, the $t = 1$ reconstruction term has the same form with $\widetilde{\mu}_1 = x_0$.

B.2 Information Geometry of the γ-Power Divergence

A smooth divergence on a statistical manifold $\mathcal{M} = \{g_\theta : \theta \in \Theta\}$ induces a Riemannian metric and a pair of dual affine connections $(\nabla, \nabla^\star)$ (Amari, 2016). For the γ-power divergence in (6), the ∇-autoparallel curve between densities g and f is the usual mixture, or m-geodesic,

$$g_t^{(m)}(x) = (1-t)g(x) + tf(x), \qquad t \in [0,1]. \tag{18}$$

Thus, the primal geometry retains the linear mixture interpolation of the KL geometry. In contrast, on an admissible domain, the dual $\nabla^\star$-autoparallel curve is characterized by

$$\left(\frac{g_t^{(\gamma)}(x)}{\|g_t^{(\gamma)}\|_{1+\gamma}}\right)^\gamma = (1-t)\left(\frac{g(x)}{\|g\|_{1+\gamma}}\right)^\gamma + t\left(\frac{f(x)}{\|f\|_{1+\gamma}}\right)^\gamma + c_t, \qquad t \in [0,1], \tag{19}$$

where c_t is independent of x and $g_t^{(\gamma)}$ is normalized to integrate to one. As $\gamma \to 0$, this curve converges to the exponential geodesic $g_t(x) \propto g(x)^{1-t} f(x)^t$.

Moreover, a statistical manifold is called γ-flat if it is autoparallel under $\nabla^\star$, or equivalently if the γ-geodesic between any two of its members remains in the family. In particular, the power family

$$\mathcal{S}_\gamma = \left\{ g_\theta(x) \propto \left(1 + \gamma\theta^\top s(x)\right)_+^{1/\gamma} : \theta \in \Theta \right\} \tag{20}$$

is γ-flat because its normalized γ-power coordinates are affine in the sufficient statistic $s(x)$. Hence power families play for D_γ the role that exponential families play for the KL divergence.

For $m \in \mathbb{R}^d$, $s \in (0,\infty)^d$, and $D_s := \mathrm{diag}(s)$, the case $p = 2$ gives the Student-t density with location m and scatter matrix D_s^2:

$$f_{2,\nu}^{(d)}(x \mid m, s) = \frac{\Gamma((\nu+d)/2)}{\Gamma(\nu/2)(\nu\pi)^{d/2}|D_s|}\left(1 + \frac{\|D_s^{-1}(x-m)\|_2^2}{\nu}\right)^{-(\nu+d)/2}. \tag{21}$$

Its matched divergence parameter is $\gamma_{\star,2} = -2/(\nu+d)$ (Kim et al., 2024).

Similarly, $p = 1$ gives the symmetric Pareto density

$$f_{1,\nu}^{(d)}(x \mid m, s) = \frac{\Gamma(\nu+d)}{\Gamma(\nu)(2\nu)^d|D_s|}\left(1 + \frac{\|D_s^{-1}(x-m)\|_1}{\nu}\right)^{-(\nu+d)}, \tag{22}$$

with matched parameter $\gamma_{\star,1} = -1/(\nu+d)$ (Cho et al., 2026).

For fixed m and ν, varying s yields the γ-flat symPareto scale family, whose sufficient statistics are $(|x_1 - m_1|, \ldots, |x_d - m_d|)^\top$.

For the non-quadratic cases $0 < p < 2$, the γ-flatness statement used here is restricted to fixed-location scale subfamilies. Once μ is fixed, these subfamilies have the parameter-independent statistic

$$s_{\mu,p}(x) = \left(|x_1 - \mu_1|^p, \ldots, |x_d - \mu_d|^p\right)^\top.$$

Allowing μ to vary does not generally preserve a common presentation. The quadratic case $p = 2$ is exceptional because translating a quadratic energy produces only constant, linear, and quadratic statistics in x. For the $p = 1$ symPareto family, Cho et al. (2026) address this issue using a tractable upper-bound surrogate consisting of the γ-power divergence between the corresponding centered distributions and a scale-normalized ℓ_1 translation penalty. This construction suggests analogous surrogates for other non-quadratic values of p.

B.3 Explicit Positive-stable Diagonal GSM Representation

For $0 < p < 2$, the positive-stable Gaussian-mixture representation of generalized-Gaussian densities (West, 1987) gives an explicit diagonal precision distribution for the GSM existence result in Theorem 2.3.

Proposition B.1 (Positive-stable diagonal GSM representation)**.** *For $0 < p < 2$, let $R_{p/2}$ be a positive $p/2$-stable random variable with*

$$\mathbb{E}\left[e^{-tR_{p/2}}\right] = e^{-t^{p/2}}, \qquad t \geq 0.$$

Let Q_p denote the distribution of $L_p := 2p^{-2/p}R_{p/2}$, and define a probability measure H_p by

$$H_p(d\lambda) := \frac{\lambda^{-1/2}}{\mathbb{E}[L_p^{-1/2}]}\, Q_p(d\lambda).$$

Let $A_1, \dots, A_d \overset{\text{iid}}{\sim} H_p$, $U \sim \text{Gamma}(\nu/p, \text{rate} = 1/p)$, $\nu > 0$, *and* $Z \sim \mathcal{N}_d(0, I_d)$ *be mutually independent. Set*

$$\Lambda := \left(\frac{U}{\nu}\right)^{2/p} \text{diag}(A_1, \dots, A_d).$$

Then

$$m + D_s \Lambda^{-1/2} Z \sim F^{(d)}_{p,\nu}(m, D_s).$$

The proof will use the following positive-stable identities to normalize H_p and identify the induced generalized-Gaussian density.

Lemma B.2 (Positive-stable identities)**.** *Let* $0 < \alpha < 1$, *and let* R_α *be a positive* α*-stable random variable satisfying*

$$\mathbb{E}[e^{-tR_\alpha}] = e^{-t^\alpha}, \qquad t \geq 0.$$

Then, for every $r > 0$,

$$\mathbb{E}[R_\alpha^{-r}] = \frac{\Gamma(r/\alpha)}{\alpha \Gamma(r)} = \frac{\Gamma(1 + r/\alpha)}{\Gamma(1 + r)}. \tag{23}$$

Moreover, for $0 < p < 2$, *if*

$$\alpha = \frac{p}{2}, \qquad L_p = 2p^{-2/p} R_{p/2},$$

then

$$\mathbb{E}\left[e^{-y^2 L_p/2}\right] = \exp\left(-\frac{|y|^p}{p}\right), \qquad y \in \mathbb{R}. \tag{24}$$

Proof. For $x > 0$ and $r > 0$, the Gamma integral representation

$$x^{-r} = \frac{1}{\Gamma(r)} \int_0^\infty t^{r-1} e^{-tx}\, dt$$

and Tonelli's theorem give

$$\mathbb{E}[R_\alpha^{-r}] = \frac{1}{\Gamma(r)} \mathbb{E}\left[\int_0^\infty t^{r-1} e^{-tR_\alpha}\, dt\right] = \frac{1}{\Gamma(r)} \int_0^\infty t^{r-1} \mathbb{E}[e^{-tR_\alpha}]\, dt. \tag{25}$$

Using the defining Laplace transform and substituting $u = t^\alpha$ yields

$$\mathbb{E}[R_\alpha^{-r}] = \frac{1}{\Gamma(r)} \int_0^\infty t^{r-1} e^{-t^\alpha}\, dt = \frac{\Gamma(r/\alpha)}{\alpha \Gamma(r)}.$$

The equivalent Gamma-function expression follows from $\Gamma(1 + x) = x\Gamma(x)$.

For the second identity,

$$\mathbb{E}[e^{-y^2 L_p/2}] = \mathbb{E}\left[e^{-y^2 p^{-2/p} R_{p/2}}\right] = \exp\left[-\left(y^2 p^{-2/p}\right)^{p/2}\right] = \exp\left(-\frac{|y|^p}{p}\right).$$

□

Now we prove Proposition B.1.

Proof of Proposition B.1. Set $\alpha := p/2$ and $M_p := \mathbb{E}[L_p^{-1/2}]$. By Lemma B.2, taking $r = 1/2$ gives

$$M_p = 2^{-1/2} p^{1/p} \frac{\Gamma(1 + 1/p)}{\Gamma(3/2)} = \sqrt{\frac{2}{\pi}}\, p^{1/p} \Gamma(1 + 1/p). \tag{26}$$

We next identify the univariate Gaussian mixture induced by H_p. For $y \in \mathbb{R}$, the stable Laplace transform (24) gives

$$\int_0^\infty e^{-\lambda y^2/2}\, Q_p(d\lambda) = \exp\left(-\frac{|y|^p}{p}\right). \tag{27}$$

Thus, if $A \sim H_p$ and $Z \sim \mathcal{N}(0,1)$ are independent, then $G := A^{-1/2}Z$ has density

$$\begin{aligned} g_p(y) &= \int_0^\infty \sqrt{\frac{\lambda}{2\pi}} \exp\left(-\frac{\lambda y^2}{2}\right) H_p(d\lambda) \\ &= \frac{1}{\sqrt{2\pi}M_p} \int_0^\infty \exp\left(-\frac{\lambda y^2}{2}\right) Q_p(d\lambda) \\ &= \frac{1}{2p^{1/p}\Gamma(1+1/p)} \exp\left(-\frac{|y|^p}{p}\right). \end{aligned} \tag{28}$$

Thus, for $\boldsymbol{A} = \mathrm{diag}(A_1, \ldots, A_d)$, the coordinates of $G := \boldsymbol{A}^{-1/2}Z$ are independent generalized-Gaussian random variables with density g_p.

Finally, $\Lambda = \left(\frac{U}{\nu}\right)^{2/p} \boldsymbol{A}$ implies

$$\Lambda^{-1/2}Z = \left(\frac{\nu}{U}\right)^{1/p} G.$$

Applying Proposition 2.4 to these independent generalized-normal coordinates gives

$$m + D_s\Lambda^{-1/2}Z \sim F_{p,\nu}^{(d)}(m, D_s).$$

□

C Proofs for Section 2

C.1 Proof of Theorem 2.3

The proof relies on the complete-monotonicity characterization of univariate Gaussian scale mixtures. We first introduce the required definitions and composition rule.

Definition C.1 (Complete monotonicity). *A function $g : (0, \infty) \to \mathbb{R}$ is* completely monotone *if $g \in C^\infty(0, \infty)$ and*

$$(-1)^k g^{(k)}(y) \geq 0 \qquad \text{for all } y > 0 \text{ and integers } k \geq 0.$$

Definition C.2 (Bernstein function). *A function $b : (0, \infty) \to [0, \infty)$ is a* Bernstein function *if $b \in C^\infty(0, \infty)$ and b' is completely monotone.*

Lemma C.3 (Composition rule). *If f is completely monotone and b is a Bernstein function, then $f \circ b$ is completely monotone (Schilling et al., 2012, Theorem 3.8).*

We apply the following characterization of univariate Gaussian scale mixtures.

Theorem C.4 (Andrews & Mallows (1974)). *Let X have a density f_X symmetric about zero. There exist independent $V > 0$ and $Z \sim \mathcal{N}(0, 1)$ such that $X \overset{\mathrm{d}}{=} Z/V$ if and only if*

$$y \longmapsto f_X(\sqrt{y})$$

is completely monotone on $(0, \infty)$

Proof of Theorem 2.3. We first establish GSM representability for $0 < p \leq 2$ and its impossibility for $p > 2$. We then characterize the common-scale case for $d \geq 2$.

Case $0 < p \leq 2$. For $u > 0$, define

$$h_u(x) := \prod_{i=1}^{d} \frac{(cu)^{1/p}}{2\Gamma(1 + 1/p)} \exp(-cu|x_i|^p)\,.$$

For each coordinate,

$$\int_{\mathbb{R}} e^{-cu|x|^p}\, dx = 2\int_0^\infty e^{-cux^p}\, dx = \frac{2}{p}(cu)^{-1/p}\Gamma(1/p) = 2(cu)^{-1/p}\Gamma(1 + 1/p),$$

so every factor in h_u integrates to one. Let $\alpha := p/2 \in (0, 1]$. For every integer $k \geq 0$,

$$(-1)^k \frac{d^k}{dy^k}\left(\alpha y^{\alpha-1}\right) = \alpha \prod_{j=1}^{k}(j - \alpha) y^{\alpha-k-1} \geq 0,$$

where the product is understood as 1 when $k = 0$. Hence $y \mapsto y^\alpha$ is a Bernstein function. Since $t \mapsto e^{-cut}$ is completely monotone, Lemma C.3 implies that

$$y \longmapsto e^{-cuy^\alpha} = \exp\Big(-cuy^{p/2}\Big)$$

is completely monotone. By Theorem C.4, each coordinate density in h_u is a univariate GSM, and therefore h_u is a GSM in the sense of Definition 2.2.

Now let

$$a := m - \frac{d}{p} > 0, \qquad u \sim \mathrm{Gamma}(a, 1).$$

Then

$$\begin{aligned}\int_0^\infty h_u(x)\frac{u^{a-1}e^{-u}}{\Gamma(a)}\, du &= \frac{c^{d/p}}{[2\Gamma(1 + 1/p)]^d\Gamma(a)}\int_0^\infty u^{m-1}e^{-u(1+c\|x\|_p^p)}\, du \\ &= \frac{c^{d/p}\Gamma(m)}{[2\Gamma(1 + 1/p)]^d\Gamma(m - d/p)}\left(1 + c\|x\|_p^p\right)^{-m}.\end{aligned}$$

Since h_u and the Gamma mixing law are normalized densities, Tonelli's theorem (Folland, 1999) gives

$$\int_{\mathbb{R}^d}\int_0^\infty h_u(x)\frac{u^{a-1}e^{-u}}{\Gamma(a)}\,du\,dx = \int_0^\infty \left(\int_{\mathbb{R}^d} h_u(x)\,dx\right)\frac{u^{a-1}e^{-u}}{\Gamma(a)}\,du = 1.$$

Therefore, $\frac{c^{d/p}\Gamma(m)}{[2\Gamma(1+1/p)]^d\Gamma(m-d/p)}$ is the normalizing constant of the target power-form density. Hence the density is a GSM for $0 < p \le 2$.

Case $p > 2$. Suppose, for contradiction, that the power-form density is a GSM. By Definition 2.2,

$$X_1 = \Lambda_1^{-1/2} Z_1, \qquad Z_1 \sim \mathcal{N}(0,1), \qquad \Lambda_1 \perp Z_1,$$

so every coordinate marginal of a GSM is a univariate GSM.

Let $x_{-1} := (x_2, \dots, x_d)$ and

$$A(x_1) := 1 + c|x_1|^p.$$

The first marginal satisfies

$$f_{X_1}(x_1) \propto \int_{\mathbb{R}^{d-1}} \left(A(x_1) + c\|x_{-1}\|_p^p\right)^{-m}\,dx_{-1}.$$

With the change of variables

$$x_{-1} = \left(\frac{A(x_1)}{c}\right)^{1/p} z, \qquad dx_{-1} = \left(\frac{A(x_1)}{c}\right)^{(d-1)/p} dz,$$

we obtain

$$\begin{aligned} f_{X_1}(x_1) &\propto A(x_1)^{-m+(d-1)/p} \int_{\mathbb{R}^{d-1}} \left(1 + \|z\|_p^p\right)^{-m} dz \\ &\propto \left(1 + c|x_1|^p\right)^{-\widetilde{m}}, \qquad \widetilde{m} := m - \frac{d-1}{p} > 0, \end{aligned}$$

where $\widetilde{m} > 0$ follows from $mp > d$.

Set $r := p/2 > 1$. Then, for some $K > 0$,

$$g(y) := f_{X_1}(\sqrt{y}) = K\left(1 + cy^r\right)^{-\widetilde{m}}.$$

Its second derivative is

$$g''(y) = K\widetilde{m}cr\,y^{r-2}\left(1 + cy^r\right)^{-\widetilde{m}-2}\left[-(r-1) + c(\widetilde{m}r+1)y^r\right].$$

Hence

$$g''(y) < 0 \qquad \text{whenever} \qquad 0 < y < \left(\frac{r-1}{c(\widetilde{m}r+1)}\right)^{1/r}.$$

and therefore g is not completely monotone. Thus the power-form density is not a GSM for $p > 2$.

Common-scale GSMs. For $p = 2$, let

$$\Lambda \sim \text{Gamma}\left(m - \frac{d}{2}, \text{ rate} = \frac{1}{2c}\right), \qquad Z \sim \mathcal{N}_d(0, I_d),$$

independently. Then $\Lambda^{-1/2}Z$ has density proportional to $\left(1 + c\|x\|_2^2\right)^{-m}$ and hence yields a common-scale GSM representation.

Conversely, every common-scale GSM is rotationally invariant. Indeed, for any orthogonal matrix R,

$$R\Lambda^{-1/2}Z = \Lambda^{-1/2}RZ \overset{\text{d}}{=} \Lambda^{-1/2}Z.$$

For $d \ge 2$, consider

$$x = \rho e_1, \qquad x' = \frac{\rho}{\sqrt{2}}(e_1 + e_2),$$

which satisfy

$$\|x\|_2 = \|x'\|_2 = \rho,$$

but

$$\|x\|_p^p = \rho^p, \qquad \|x'\|_p^p = 2^{1-p/2}\rho^p.$$

Thus the power-form density is rotationally invariant only when $p = 2$. Therefore, for $d \ge 2$, it admits a common-scale GSM representation if and only if $p = 2$. □

C.2 Proof of Proposition 2.4

Proof. Define

$$G_i := S_i W_i^{1/p}, \qquad G := (G_1, \dots, G_d)^\top.$$

The density of W_i is

$$f_W(w) = \frac{p^{-1/p}}{\Gamma(1/p)} w^{1/p-1} e^{-w/p}, \qquad w > 0.$$

Since S_i takes the values ± 1 with equal probability, the transformation $w = |y|^p$, whose Jacobian is $p|y|^{p-1}$, gives

$$\begin{aligned} g_p(y) &= \frac{1}{2} f_W(|y|^p)\, p|y|^{p-1} \\ &= \frac{p^{1-1/p}}{2\Gamma(1/p)} \exp\left(-\frac{|y|^p}{p}\right) \\ &= \frac{1}{2p^{1/p}\Gamma(1+1/p)} \exp\left(-\frac{|y|^p}{p}\right). \end{aligned} \tag{29}$$

Thus $G_1, \dots, G_d$ are independent generalized-Gaussian variables with common density g_p.

Set

$$Y := \left(\frac{\nu}{U}\right)^{1/p} G.$$

Conditional on $U = u$, the coordinates of Y are independent and have common scale $(\nu/u)^{1/p}$. Hence

$$f_{Y|U=u}(y) = \left[\frac{(u/\nu)^{1/p}}{2p^{1/p}\Gamma(1+1/p)}\right]^d \exp\left(-\frac{u}{p\nu}\|y\|_p^p\right). \tag{30}$$

The density of U is

$$f_U(u) = \frac{1}{\Gamma(\nu/p)p^{\nu/p}} u^{\nu/p-1} e^{-u/p}, \qquad u > 0.$$

Integrating out U therefore gives

$$f_Y(y) = \frac{\nu^{-d/p}}{[2p^{1/p}\Gamma(1+1/p)]^d \Gamma(\nu/p) p^{\nu/p}} \times \int_0^\infty u^{(\nu+d)/p-1} \exp\left[-\frac{u}{p}\left(1 + \frac{\|y\|_p^p}{\nu}\right)\right] du. \tag{31}$$

Using

$$\int_0^\infty u^{a-1} e^{-bu}\, du = \frac{\Gamma(a)}{b^a}, \qquad a, b > 0,$$

with

$$a = \frac{\nu + d}{p}, \qquad b = \frac{1}{p}\left(1 + \frac{\|y\|_p^p}{\nu}\right),$$

we obtain

$$\begin{aligned} f_Y(y) &= \frac{\Gamma((\nu+d)/p)}{\Gamma(\nu/p)\nu^{d/p}[2\Gamma(1+1/p)]^d} \left(1 + \frac{\|y\|_p^p}{\nu}\right)^{-(\nu+d)/p} \\ &= C_{p,\nu,d} \left(1 + \frac{\|y\|_p^p}{\nu}\right)^{-(\nu+d)/p}. \end{aligned} \tag{32}$$

Finally, $X = m + D_s Y$. The change of variables $y = D_s^{-1}(x - m)$ has Jacobian determinant $|D_s|^{-1}$, so

$$\begin{aligned} f_X(x) &= \frac{1}{|D_s|} f_Y\left(D_s^{-1}(x-m)\right) \\ &= \frac{C_{p,\nu,d}}{|D_s|} \left(1 + \frac{\|D_s^{-1}(x-m)\|_p^p}{\nu}\right)^{-(\nu+d)/p} \\ &= f_{p,\nu}^{(d)}(x \mid m, s). \end{aligned} \tag{33}$$

Therefore $X \sim F_{p,\nu}^{(d)}(m, D_s)$. □

D PROOFS FOR SECTION 3

D.1 PROOFS OF LEMMAS 3.2 AND 3.3

Proof. We first record the marginal identity used in both parts. For $a > 0$, $m > r/p$, and $y \in \mathbb{R}^k$, the change of variables $z = (a + \|y\|_p^p)^{1/p} w$ gives

$$\int_{\mathbb{R}^r} \left(a + \|y\|_p^p + \|z\|_p^p\right)^{-m} dz = c_{p,r,m} \left(a + \|y\|_p^p\right)^{-m+r/p}, \tag{34}$$

where $c_{p,r,m} := \int_{\mathbb{R}^r} (1 + \|w\|_p^p)^{-m} dw < \infty$. Thus every coordinate marginal of $F_{p,\nu}^{(d)}(0, \mathbf{1}_d)$ remains in the ℓ_p-power family, with the same ν and the corresponding dimension.

We next derive forward and backward posteriors.

1. Forward path. By construction and the independence of X_0 and $\epsilon_{1:T}$,

$$X_{1:T} \mid X_0 = x_0 = m_f(x_0) + s_f \odot \epsilon_{1:T}.$$

where $\odot$ denote element-wise product. For $x_{1:T} \in \mathbb{R}^{Tn}$, set $e = s_f^{-1} \odot (x_{1:T} - m_f(x_0))$. The corresponding Jacobian is $\prod_{t=1}^T \sigma_t^{-n}$, so the form of ℓ_p-power family density (8) gives

$$X_{1:T} \mid X_0 = x_0 \sim F_{p,\nu}^{(Tn)}(m_f(x_0), s_f).$$

Applying (34) to all blocks except the t-th and then using $X_t = \mu_t x_0 + \sigma_t \epsilon_t$ yields

$$X_t \mid X_0 = x_0 \sim F_{p,\nu}^{(n)}(\mu_t x_0, \sigma_t \mathbf{1}_n).$$

For $2 \le t \le T$, the same marginal identity gives $(\epsilon_{t-1}, \epsilon_t) \sim F_{p,\nu}^{(2n)}(0, I_{2n})$. Conditioning on $X_0 = x_0$ and $X_t = x_t$ gives $\epsilon_t = e_t := (x_t - \mu_t x_0)/\sigma_t$. The conditional density of ϵ_{t-1} at e is then proportional to

$$\left(1 + \frac{\|e\|_p^p}{\nu + \|e_t\|_p^p}\right)^{-(\nu+2n)/p}.$$

To express this kernel in the standard ℓ_p-power form, set $d_t := ((\nu + \|e_t\|_p^p)/(\nu + n))^{1/p}$, so that

$$\frac{\|e\|_p^p}{\nu + \|e_t\|_p^p} = \frac{\|e/d_t\|_p^p}{\nu + n}.$$

The resulting kernel corresponds to $F_{p,\nu+n}^{(n)}(0, d_t I_n)$. The affine transformation $X_{t-1} = \mu_{t-1} x_0 + \sigma_{t-1} \epsilon_{t-1}$ therefore yields

$$X_{t-1} \mid (X_0 = x_0, X_t = x_t) \sim F_{p,\nu+n}^{(n)}(\mu_{t-1} x_0, \sigma_{t-1} d_t I_n),$$

proving (9). Note that the conditional distribution is the point mass δ_{x_0} at $t = 1$.

2. Reverse path. Let $G_{\theta,p}$ denote the joint density induced by the reverse construction. Applying the change-of-variables formula to $U \sim F_{p,\nu}^{(N)}(0, \mathbf{1}_N)$ gives

$$G_{\theta,p}(x_{0:T}) = \frac{C_{p,\nu,N}}{\bar{\tau}_{\text{rev}}} \left(1 + \frac{\sum_{j=0}^T \|u_j\|_p^p}{\nu}\right)^{-(\nu+N)/p}, \tag{35}$$

where $u_{0:T}$ are the standardized residuals associated with $x_{0:T}$. Since $X_T = U_T$, marginal closure also gives $X_T \sim F_{p,\nu}^{(n)}(0, \mathbf{1}_n)$.

Next, fix $t \in \{1, \dots, T\}$ and write $k_t := (T - t + 1)n$, $S_{p,t} := \sum_{j=t}^T \|u_j\|_p^p$. Integrating out $u_0, \dots, u_{t-2}$ using (34) gives

$$(U_{t-1}, U_{t:T}) \sim F_{p,\nu}^{(n+k_t)}(0, \mathbf{1}_{n+k_t}).$$

Thus, for fixed $u_{t:T}$, the conditional density of U_{t-1} at u is proportional to

$$\left(1 + \frac{\|u\|_p^p + S_{p,t}}{\nu}\right)^{-(\nu+n+k_t)/p}.$$

Indeed, dividing the joint density by the marginal density of $U_{t:T}$ introduces a factor independent of u. Moreover,

$$1 + \frac{\|u\|_p^p + S_{p,t}}{\nu} = \left(1 + \frac{S_{p,t}}{\nu}\right)\left(1 + \frac{\|u\|_p^p}{\nu + S_{p,t}}\right).$$

Absorbing the first factor into the normalizing constant therefore leaves the conditional density kernel

$$\left(1 + \frac{\|u\|_p^p}{\nu + S_{p,t}}\right)^{-(\nu+k_t+n)/p}.$$

In particular, this conditional density depends on $u_{t:T}$ only through $S_{p,t}$.

To identify its scale, set $a_{p,t} := ((\nu + S_{p,t})/(\nu + k_t))^{1/p}$. Then

$$\frac{\|u\|_p^p}{\nu + S_{p,t}} = \frac{\|u/a_{p,t}\|_p^p}{\nu + k_t},$$

so

$$U_{t-1} \mid U_{t:T} = u_{t:T} \sim F^{(n)}_{p,\nu+k_t}(0, a_{p,t}\mathbf{1}_n).$$

$u_{t:T}$ is determined by $x_{t:T}$ through the residual definitions. Conversely, $x_{t:T}$ is recovered from $u_{t:T}$ by starting at $x_T = u_T$ and applying the reverse recursion down to time t. Hence conditioning on $X_{t:T} = x_{t:T}$ is equivalent to conditioning on the corresponding $U_{t:T} = u_{t:T}$.

Lastly, applying $X_{t-1} = \mu_\theta(t, x_t) + \tau_{t-1}U_{t-1}$ therefore gives

$$X_{t-1} \mid X_{t:T} = x_{t:T} \sim F^{(n)}_{p,\nu+k_t}\left(\mu_\theta(t, x_t), \tau_{t-1}a_{p,t}\mathbf{1}_n\right),$$

proving (10). □

D.2 Proof of Theorem 3.4

Proof. Let $G_{\theta,p}$ denote the joint density induced by the reverse construction, and write $N := (T+1)n$, $\gamma := \gamma_{\star,p} = -p/(\nu + N)$, and $\bar{\tau}_{\mathrm{rev}} := \prod_{t=1}^{T} \tau_{t-1}^{n}$. Define

$$R_{\theta,p}(x_{0:T}) := \|x_T\|_p^p + \sum_{t=1}^{T} \left\| \frac{x_{t-1} - \mu_\theta(t, x_t)}{\tau_{t-1}} \right\|_p^p.$$

From (35), we know that

$$G_{\theta,p}(x_{0:T}) = K_p \left(1 + \frac{R_{\theta,p}(x_{0:T})}{\nu}\right)^{-(\nu+N)/p}, \qquad K_p := \frac{C_{p,\nu,N}}{\bar{\tau}_{\mathrm{rev}}}.$$

Applying the same change of variables,

$$\|G_{\theta,p}\|_{1+\gamma}^{1+\gamma} = K_p^{1+\gamma} \bar{\tau}_{\mathrm{rev}} \int_{\mathbb{R}^N} \left(1 + \frac{\|u\|_p^p}{\nu}\right)^{-(\nu+N-p)/p} du.$$

The integral is finite because $\nu > p$ implies $(\nu + N - p)/p > N/p$. Thus $\|G_{\theta,p}\|_{1+\gamma}$ is finite and independent of θ. Moreover, the choice of γ gives

$$G_{\theta,p}(x_{0:T})^{\gamma} = K_p^{\gamma} \left(1 + \frac{R_{\theta,p}(x_{0:T})}{\nu}\right).$$

Set $c_p := (K_p / \|G_{\theta,p}\|_{1+\gamma})^{\gamma} > 0$. Substituting it into the joint γ-power divergence formulation, we obtain

$$\mathcal{D}_\gamma(F \| G_{\theta,p}) = -\frac{1}{\gamma} \left\{ c_p \left(1 + \frac{\mathbb{E}_F R_{\theta,p}(X_{0:T})}{\nu}\right) - \|F\|_{1+\gamma} \right\}.$$

By the definition of $\mathcal{L}_{\mathrm{adj},p}$, we have

$$\frac{\mathbb{E}_F R_{\theta,p}(X_{0:T})}{\nu} = \frac{\mathbb{E}_F \|X_T\|_p^p}{\nu} + T\mathcal{L}_{\mathrm{adj},p}(\theta).$$

Therefore,

$$\mathcal{D}_\gamma(F \| G_{\theta,p}) = -\frac{1}{\gamma} \left\{ c_p \left(1 + \frac{\mathbb{E}_F \|X_T\|_p^p}{\nu}\right) - \|F\|_{1+\gamma} \right\} - \frac{T c_p}{\gamma} \mathcal{L}_{\mathrm{adj},p}(\theta).$$

The first term is independent of θ, and $-Tc_p/\gamma > 0$ because $\gamma < 0$. This proves the claimed θ-equivalence. □

D.3 Proof of Theorem 3.6

Proof. For each t, set

$$R_{\theta,t} := \frac{\mu_{t-1}X_0 - \mu_\theta(t, X_t)}{\tau_{t-1}}, \qquad Z_t := \frac{X_{t-1} - \mu_{t-1}X_0}{\tau_{t-1}}.$$

For $t \geq 2$, $Z_t = (\sigma_{t-1}/\tau_{t-1})\epsilon_{t-1}$; for $t = 1$, $Z_1 = 0$. By Lemma 3.2, the corresponding time-t summands in the adjacent and denoising losses are

$$\mathbb{E}_F\left[\frac{\|R_{\theta,t} + Z_t\|_p^p}{\nu}\right], \qquad \mathbb{E}_F\left[\frac{\|R_{\theta,t}\|_p^p}{\nu}\right],$$

respectively.

We first compute the noise moment. For $E \sim F_{p,\nu}^{(n)}(0, I_n)$, the variable $\|E\|_p^p/\nu$ has a beta-prime distribution, whose density is proportional to

$$y^{n/p-1}(1+y)^{-(n+\nu)/p}, \qquad y > 0.$$

Because $\nu > p$, its expectation is $(n/p)/(\nu/p - 1)$, giving $\mathbb{E}\|E\|_p^p = \frac{n\nu}{\nu-p}$. Therefore,

$$\frac{1}{T\nu}\sum_{t=1}^{T}\mathbb{E}_F\|Z_t\|_p^p = \frac{n}{T(\nu-p)}\sum_{t=2}^{T}\left(\frac{\sigma_{t-1}}{\tau_{t-1}}\right)^p = C_p. \tag{36}$$

For $1 \leq p \leq 2$, conditional symmetry and integrability give $\mathbb{E}[Z_t \mid X_0, X_t] = 0$.

Conditional Jensen's inequality yields

$$\mathbb{E}_F\big[\|R_{\theta,t} + Z_t\|_p^p \mid X_0, X_t\big] \geq \|R_{\theta,t}\|_p^p.$$

For the upper bound, the scalar inequality

$$\frac{|a+b|^p + |a-b|^p}{2} \leq |a|^p + |b|^p, \qquad 1 \leq p \leq 2, \tag{37}$$

summed over coordinates, gives

$$\frac{\|R_{\theta,t} + Z_t\|_p^p + \|R_{\theta,t} - Z_t\|_p^p}{2} \leq \|R_{\theta,t}\|_p^p + \|Z_t\|_p^p.$$

The two terms on the left have the same conditional expectation under reflection. Dividing by ν, averaging over time, and using (36) proves

$$\mathcal{L}_{\mathrm{dns},p}(\theta) \leq \mathcal{L}_{\mathrm{adj},p}(\theta) \leq \mathcal{L}_{\mathrm{dns},p}(\theta) + C_p. \tag{38}$$

For $p = 2$, the parallelogram identity yields equality in the upper bound for every θ:

$$\mathcal{L}_{\mathrm{adj},2}(\theta) = \mathcal{L}_{\mathrm{dns},2}(\theta) + C_2.$$

This proves the θ-equivalence between $\mathcal{L}_{\mathrm{adj},2}$ and $\mathcal{L}_{\mathrm{dns},2}$.

For $1 \leq p < 2$, the upper inequality (37) is strict whenever $ab \neq 0$, and is an equality when $ab = 0$. The integrated upper bound is therefore attained if and only if

$$R_{\theta,t,i}Z_{t,i} = 0, \text{a.s.}$$

for all t and i. These are the equality conditions for (38).

To establish failure of θ-equivalence in general for $1 \leq p < 2$, take $T \geq 2$ and the predictor family $\mu_\theta(t, x) = \tau_{t-1}\theta\mathbf{1}_n$, $\theta \in \mathbb{R}$. Then $R_{\theta,t,i} = \mu_{t-1}X_{0,i}/\tau_{t-1} - \theta$. Define

$$h_p(r, z) := \tfrac{1}{2}\big(|r+z|^p + |r-z|^p\big) - |r|^p.$$

The preceding inequalities give $0 \leq h_p(r, z) \leq |z|^p$, and $h_p(r, z) \to 0$ as $|r| \to \infty$ for fixed z. Conditional symmetry and dominated convergence therefore imply

$$\mathcal{L}_{\mathrm{adj},p}(\theta) - \mathcal{L}_{\mathrm{dns},p}(\theta) = \frac{1}{T\nu}\sum_{t,i}\mathbb{E}_F h_p(R_{\theta,t,i}, Z_{t,i}) \longrightarrow 0 \quad \text{as } \theta \to \infty.$$

For $T \geq 2$, full conditional support makes the gap strictly positive for every finite θ. Since $\mathcal{L}_{\mathrm{dns},p}(\theta) \to \infty$ while the gap tends to zero, any affine relation $\mathcal{L}_{\mathrm{adj},p} = a\mathcal{L}_{\mathrm{dns},p} + b$ would require $a = 1$ and $b = 0$, contradicting positivity. Thus the losses are not generally θ-equivalent for $p < 2$. For $T = 1$, they coincide for every p. □

D.4 Proof of Proposition 3.7 and Corollary 3.8

Proof. Work under the stated local conditional model, with fixed $1 < p < 2$, $\kappa = \nu + n > p$, and $s(c) > 0$. Let $U \sim Q$ be independent of V, and define

$$\psi_p(u) := \operatorname{sgn}(u)|u|^{p-1}, \qquad \psi_p^{-1}(v) = \operatorname{sgn}(v)|v|^{1/(p-1)}.$$

The assumptions $\mathbb{E}\|U\|_2^2 < \infty$ and $\mathbb{E}[s(U)^p] < \infty$, together with independence and $\mathbb{E}\|V\|_p^p < \infty$, ensure that both targets have finite p-th moments. Their objectives are therefore finite, continuous, strictly convex, and coercive, so each has a unique minimizer. Their first-order conditions follow by dominated differentiation on compact sets: for either coordinate target Z and $|m| \le K$, $|m - Z|^{p-1} \le K^{p-1} + |Z|^{p-1}$, whose right-hand side is integrable.

Since the loss separates across coordinates, fix j and set

$$M_j := \mathbb{E}[\psi_p(U_j - c_j)], \qquad C_j := \mathbb{E}[\psi_p(U_j - c_j + s(U)V_j)].$$

These constants are finite, but need not be nonzero. Define the response coefficients by

$$q_j := \psi_p^{-1}(M_j), \qquad b_j := \frac{C_j}{D_p}, \qquad D_p := (p-1)s(c)^{p-2}\mathbb{E}|V_1|^{p-2}.$$

The coordinate marginals of V are identical, symmetric, and have bounded densities. Since $p - 2 \in (-1, 0)$, $D_p \in (0, \infty)$.

1. Denoising objective. Write $d_{\pi,j} := [d_\pi]_j$. Its first-order condition is

$$(1-\pi)\psi_p(d_{\pi,j} - c_j) = \pi\mathbb{E}[\psi_p(U_j - d_{\pi,j})].$$

Strict monotonicity of the score implies that $d_{\pi,j} - c_j$ has the same sign as M_j, including $d_{\pi,j} = c_j$ when $M_j = 0$. Monotonicity then gives

$$|d_{\pi,j} - c_j|^{p-1} \le \frac{\pi|M_j|}{1-\pi},$$

so $d_{\pi,j} \to c_j$ as $\pi \downarrow 0$. For sufficiently small π,

$$|\psi_p(U_j - d_{\pi,j})| \le |U_j - c_j|^{p-1} + 1,$$

whose right-hand side is integrable. Dominated convergence therefore yields

$$\frac{\psi_p(d_{\pi,j} - c_j)}{\pi} = \frac{\mathbb{E}[\psi_p(U_j - d_{\pi,j})]}{1-\pi} \longrightarrow M_j.$$

By homogeneity of ψ_p and continuity of its inverse,

$$\frac{d_{\pi,j} - c_j}{\pi^{1/(p-1)}} = \psi_p^{-1}\left(\frac{\psi_p(d_{\pi,j} - c_j)}{\pi}\right) \longrightarrow q_j.$$

This proves $d_\pi = c + \pi^{1/(p-1)}q + o(\pi^{1/(p-1)})$ at fixed dimension, including coordinates with $q_j = 0$.

2. Adjacent-state objective. Define

$$A_j(u) := \mathbb{E}[\psi_p(u - c_j - s(c)V_j)], \qquad B_j(u) := \mathbb{E}[\psi_p(u - U_j - s(U)V_j)].$$

Both functions are continuous and strictly increasing. Writing $a_{\pi,j} := [a_\pi]_j$, its first-order condition is

$$(1-\pi)A_j(a_{\pi,j}) + \pi B_j(a_{\pi,j}) = 0. \tag{39}$$

Symmetry and oddness give $A_j(c_j) = 0$ and $B_j(c_j) = -C_j$. For every fixed $\varepsilon > 0$, the score at $c_j \pm \varepsilon$ converges to $A_j(c_j \pm \varepsilon)$, which is respectively positive and negative. Strict monotonicity thus implies $a_{\pi,j} \to c_j$.

Let $W_j := s(c)V_j$. Absolute continuity of ψ_p and Fubini's theorem give

$$A_j(c_j + u) - A_j(c_j) = (p-1)\int_0^u \mathbb{E}|v - W_j|^{p-2}\,dv.$$

The bounded density of W_j and $p-2>-1$ imply that $v \mapsto \mathbb{E}|v-W_j|^{p-2}$ is finite and continuous at zero. Indeed, the contribution within distance ε of the singularity is uniformly $O(\varepsilon^{p-1})$, and dominated convergence applies away from the singularity. Consequently,

$$A_j(c_j+u) = D_p u + o(|u|), \qquad u \to 0.$$

Set $\delta_{\pi,j} := a_{\pi,j} - c_j$. Since $D_p > 0$, for sufficiently small $|u|$, $|A_j(c_j+u)| \geq (D_p/2)|u|$. Using equation 39 and local boundedness of B_j gives

$$\frac{(1-\pi)D_p}{2}|\delta_{\pi,j}| \leq \pi|B_j(a_{\pi,j})| = O(\pi),$$

so $\delta_{\pi,j} = O(\pi)$. Substituting the linearization and $B_j(a_{\pi,j}) = -C_j + o(1)$ into the first-order condition,

$$0 = (1-\pi)\big(D_p\delta_{\pi,j} + o(\pi)\big) + \pi\big(-C_j + o(1)\big).$$

Hence $\delta_{\pi,j} = \pi C_j/D_p + o(\pi)$, and thus $a_\pi = c + \pi b + o(\pi)$. This argument also covers $b_j = 0$, without dividing by $\delta_{\pi,j}$.

3. Reconstruction risk. The finite second moment of U ensures that $\mathrm{Cov}(Y_\pi)$ exists. Since $\mathbb{E}Y_\pi = c + \pi h$,

$$\mathcal{E}_\pi(m) = \mathrm{tr}\,\mathrm{Cov}(Y_\pi) + \|m - c - \pi h\|_2^2.$$

Because $1/(p-1) > 1$, the denoising expansion implies $d_\pi - c = o(\pi)$. Therefore,

$$a_\pi - c - \pi h = \pi(b-h) + o(\pi), \qquad d_\pi - c - \pi h = -\pi h + o(\pi).$$

The covariance terms cancel, yielding

$$\begin{aligned}\mathcal{E}_\pi(a_\pi) - \mathcal{E}_\pi(d_\pi) &= \|a_\pi - c - \pi h\|_2^2 - \|d_\pi - c - \pi h\|_2^2 \\ &= \pi^2\big(\|b-h\|_2^2 - \|h\|_2^2\big) + o(\pi^2).\end{aligned}$$

□

Proof of Corollary 3.8. Since $Ph = h$,

$$\|b-h\|_2^2 - \|h\|_2^2 = \|Pb-h\|_2^2 + \|P^\perp b\|_2^2 - \|h\|_2^2 \geq (n-r)B^2 - rH^2.$$

The stated ratio condition makes this coefficient strictly positive, proving the comparison for sufficiently small $\pi > 0$ at the fixed dimension. An interpretation across dimensions requires uniform bounds on H and B; a common small-π range would additionally require uniform control of the remainder.

D.5 Proof of Proposition 3.9

Proof. Fix any $\eta \geq 0$. We first reduce the vector risk to scalar risks at this fixed η, prove the scalar expansion, and then sum the expansions over coordinates. Writing $R_{N,j}^{(\eta)}(p)$ for this coordinate's squared-error risk, we have

$$\mathcal{R}_N^{(\eta)}(p) = \frac{1}{D}\sum_{j=1}^{D} R_{N,j}^{(\eta)}(p). \tag{40}$$

Fix a coordinate j. Throughout the scalar calculation below (Steps 1–3), we suppress the dependence on this fixed η and j for notational convenience, writing

$$c := x_j^{(\eta)}, \qquad Y_i := Y_j^{(i,\eta)}, \qquad \epsilon_i := z_j^{(i,\eta)}, \qquad Q_c := Q_{\boldsymbol{x}^{(\eta)},j}.$$

Thus $Y_i = c + \epsilon_i$ with $\epsilon_i \overset{\text{iid}}{\sim} Q_c$; let $Y \overset{d}{=} Y_1 \in [0,1]$. By assumption, $c \in (0,1)$ and Y has atoms only in $\{0, c, 1\}$ and positive total mass at the endpoints. Write $R_N := R_{N,j}^{(\eta)}$ and $N = 2m+1$. Set

$$q_0 := \mathbb{P}(Y = 0), \qquad q_1 := \mathbb{P}(Y = 1),$$

and, for $p > 1$, define

$$b := \frac{m}{m+1}, \qquad a_p := b^{\frac{1}{p-1}}.$$

We'll establish the following asymptotic expansion as $p \downarrow 1$:

$$R_N(p) - R_N(1) = -C_N a_p + o(a_p), \tag{41}$$

where

$$C_N = 2\binom{N}{m+1}\left[cq_0^{m+1}\big(\mathbb{E}[Y^{1/m}]\big)^m + (1-c)q_1^{m+1}\big(\mathbb{E}[(1-Y)^{1/m}]\big)^m\right] > 0.$$

Since $C_N > 0$, this expansion implies $R_N(p) < R_N(1)$ for all $p > 1$ sufficiently close to one.

Step 1. Bounding the displacement of the p-center from the sample median.

Steps 1 and 2 are deterministic statements for an arbitrary scalar sample, so they require no explicit η or j indices. Fix a sample $y_1, \ldots, y_N \in [0,1]$ and denote $y_{\max} = \max_{i=1,\cdots N} y_i$.

Let u_1 be its median and u_p its unique p-center. Since $p > 1$, the unique minimizer u_p of the objective function implies

$$\sum_{y_i < u_p} (u_p - y_i)^{p-1} = \sum_{y_i > u_p} (y_i - u_p)^{p-1} \tag{42}$$

If $u_p > u_1$, at least $m+1$ observations are at most u_1, whereas at most m observations exceed u_p. The first-order condition therefore implies

$$(m+1)(u_p - u_1)^{p-1} \leq m(y_{\max} - u_p)^{p-1}.$$

Taking the $(p-1)$-th root and recalling that $a_p = (m/(m+1))^{1/(p-1)}$, we obtain

$$\begin{aligned} 0 < u_p - u_1 &\leq a_p(y_{\max} - u_p) \\ &= a_p(y_{\max} - u_1) - a_p(u_p - u_1). \end{aligned}$$

Rearranging gives

$$(1 + a_p)(u_p - u_1) \leq a_p(y_{\max} - u_1),$$

and hence

$$0 < u_p - u_1 \leq \frac{a_p}{1+a_p}(y_{\max} - u_1).$$

The reflected argument applies when $u_p < u_1$. Hence,

$$|u_p - u_1| \leq \frac{a_p}{1+a_p}(y_{\max} - y_{\min}) \leq a_p. \tag{43}$$

In particular, $u_p \to u_1$ as $p \downarrow 1$.

Step 2. Asymptotic behavior of the displacement. Let

$$L := \#\{i \in \{1, \dots, N\} : y_i < u_1\}, \quad U := \#\{i \in \{1, \dots, N\} : y_i > u_1\}, \quad K := N - L - U.$$

Since u_1 is the median of $N = 2m + 1$ observations,

$$L \le m, \qquad U \le m, \qquad K \ge 1.$$

Recall that

$$b := \frac{m}{m+1}, \qquad a_p := b^{1/(p-1)}.$$

We first consider a sequence $p \downarrow 1$ along which $d_p := u_p - u_1 > 0$. By Step 1, $d_p \to 0$. Hence, for sufficiently small $p - 1$, every observation strictly above u_1 remains strictly above u_p. The two sides of (42) can therefore be written as

$$\sum_{y_i < u_p} (u_p - y_i)^{p-1} = \sum_{y_i < u_1} (u_1 - y_i + d_p)^{p-1} + K d_p^{p-1},$$
$$\sum_{y_i > u_p} (y_i - u_p)^{p-1} = \sum_{y_i > u_1} (y_i - u_1 - d_p)^{p-1}.$$

Substituting these expressions into (42) and rearranging yields

$$K d_p^{p-1} = \sum_{y_i > u_1} (y_i - u_1 - d_p)^{p-1} - \sum_{y_i < u_1} (u_1 - y_i + d_p)^{p-1}.$$

For each $y_i \neq u_1$, the corresponding residual magnitude converges to the positive value $|y_i - u_1|$. Consequently,

$$(y_i - u_1 - d_p)^{p-1} \longrightarrow 1, \quad y_i > u_1,$$
$$(u_1 - y_i + d_p)^{p-1} \longrightarrow 1, \quad y_i < u_1.$$

It follows that

$$d_p^{p-1} \longrightarrow \frac{U - L}{K}.$$

Since $K d_p^{p-1} > 0$ along this sequence and its limit is $U - L$, we must have $U \ge L$.

Moreover,

$$b - \frac{U - L}{K} = \frac{(2m+1)(m-U) + L}{(m+1)K} \ge 0,$$

and equality holds if and only if $(L, K, U) = (0, m + 1, m)$.

If $(U - L)/K < b$, choose $b' \in (\frac{U-L}{K}, b)$. Then $d_p^{p-1} \le b'$ and

$$0 < \frac{d_p}{a_p} \le \left(\frac{b'}{b}\right)^{1/(p-1)} \longrightarrow 0 \qquad \text{as } p \downarrow 1.$$

In the equality case $(L, K, U) = (0, m + 1, m)$, the median is the sample minimum and occurs exactly $m + 1$ times. The first-order condition reduces to

$$(m+1) d_p^{p-1} = \sum_{y_i > u_1} (y_i - u_1 - d_p)^{p-1}.$$

Since $a_p^{p-1} = m/(m+1)$, this gives

$$\frac{d_p}{a_p} = \left[\frac{1}{m} \sum_{y_i > u_1} (y_i - u_1 - d_p)^{p-1}\right]^{1/(p-1)} \longrightarrow \left[\prod_{y_i > u_1} (y_i - u_1)\right]^{1/m}.$$

The limit follows from continuity of power means at order zero, since all limiting arguments $y_i - u_1$ in this expression are strictly positive.

Applying the same argument to the reflected sample $1 - y_1, \ldots, 1 - y_N$ handles sequences with $d_p < 0$. When $d_p = 0$, the normalized displacement is zero. Combining these cases yields

$$\frac{u_p - u_1}{a_p} \longrightarrow D(y_1, \ldots, y_N), \quad \text{as } p \to 1,$$

where

$$D(y_1, \ldots, y_N) = \begin{cases} \left[\prod_{y_i > u_1} (y_i - u_1) \right]^{1/m}, & (L, K, U) = (0, m+1, m), \\ -\left[\prod_{y_i < u_1} (u_1 - y_i) \right]^{1/m}, & (L, K, U) = (m, m+1, 0), \\ 0, & \text{otherwise.} \end{cases}$$

Step 3. Expected risk expansion. We now return to the random observations $Y_i = c + \epsilon_i$, $i = 1, \ldots, N$. Steps 1 and 2 apply to every fixed realization $(y_1, \ldots, y_N)$ of these observations. Let

$$\widehat{c}_p := \operatorname*{argmin}_{u \in [0,1]} \sum_{i=1}^{N} |Y_i - u|^p, \qquad p > 1,$$

and let $\widehat{c}_1$ denote their sample median. In the coordinate notation above, $\widehat{c}_p = [\widehat{\boldsymbol{x}}_{N,p}^{(\eta)}]_j$ and $R_N(p) = \mathbb{E}[(\widehat{c}_p - c)^2]$. Setting $\Delta_p := \widehat{c}_p - \widehat{c}_1$, the samplewise limit established in Step 2 gives

$$\frac{\Delta_p}{a_p} \longrightarrow D(Y_1, \ldots, Y_N) \qquad \text{almost surely.}$$

Then

$$\frac{R_N(p) - R_N(1)}{a_p} = \mathbb{E}\left[2(\widehat{c}_1 - c)\frac{\Delta_p}{a_p} + \frac{\Delta_p^2}{a_p} \right].$$

By (43), $|\Delta_p / a_p| \le 1$ and $\Delta_p^2 / a_p \le a_p \to 0$. Dominated convergence theorem therefore yields

$$\lim_{p \downarrow 1} \frac{R_N(p) - R_N(1)}{a_p} = 2\mathbb{E}[(\widehat{c}_1 - c) D(Y_1, \ldots, Y_N)].$$

We now evaluate $2\mathbb{E}[(\widehat{c}_1 - c) D(Y_1, \ldots, Y_N)]$ using the casewise expression for D obtained in Step 2. A nonzero value of D requires exactly $m + 1$ observations to coincide at a sample minimum or maximum. By independence and the assumption on the atoms, this repeated value belongs to $\{0, c, 1\}$ almost surely. The contribution from $\widehat{c}_1 = c$ vanishes.

When exactly $m + 1$ observations equal zero, the contribution is

$$-2c \binom{N}{m+1} q_0^{m+1} \left(\mathbb{E}[Y^{1/m}] \right)^m.$$

Indeed, the remaining m observations contribute their geometric mean; any additional zero makes this product vanish. The analogous contribution from exactly $m + 1$ ones is

$$-2(1 - c) \binom{N}{m+1} q_1^{m+1} \left(\mathbb{E}[(1 - Y)^{1/m}] \right)^m.$$

Their sum is $-C_N$, proving (41).

Since $q_0 + q_1 > 0$ and Y is nondegenerate, at least one of these two terms is strictly negative. Hence $C_N > 0$, and $R_N(p) < R_N(1)$ for all $p > 1$ sufficiently close to one.

Step 4. Vector risk expansion and interior minimizers. We now restore the η and j indices. The scalar argument applies to every coordinate at the fixed η. Let $C_{N,j}^{(\eta)}$ denote the coefficient C_N above with $c = x_j^{(\eta)}$ and $Y = Y_j^{(1,\eta)}$; each coefficient is strictly positive by the corresponding marginal assumptions. Because D is finite and a_p depends only on the common sample size N, (40) yields

$$\mathcal{R}_N^{(\eta)}(p) - \mathcal{R}_N^{(\eta)}(1) = -C_N^{(\eta)} a_p + o(a_p), \qquad C_N^{(\eta)} := \frac{1}{D} \sum_{j=1}^{D} C_{N,j}^{(\eta)} > 0.$$

Thus, for some $\delta_\eta \in (0, 1)$, $\mathcal{R}_N^{(\eta)}(p) < \mathcal{R}_N^{(\eta)}(1)$ whenever $1 < p < 1 + \delta_\eta$.

For every sample, each scalar center is continuous in $p > 1$ by continuity of the objective, compactness of $[0, 1]$, and uniqueness of its minimizer. The deterministic bound (43) gives continuity at $p = 1$. All coordinatewise squared errors are bounded by one, so bounded convergence and (40) imply that $\mathcal{R}_N^{(\eta)}$ is continuous on $[1, 2]$ and attains its minimum. If $\mathcal{R}_N^{(\eta)}(1) \leq \mathcal{R}_N^{(\eta)}(2)$, the strict improvement near $p = 1$ excludes both endpoints from the minimizer set. Lastly, since η is arbitrary, this proves the proposition for every fixed $\eta \geq 0$. □

E Details on ℓ_p-Power Diffusion Models

Remark 3.5 introduces a weighted local-divergence objective for reverse models with independent stepwise precisions. This appendix examines the associated dependence and distributional closure properties, presents the conditional projection interpretation of the denoising objective, and summarizes the sampling procedures.

E.1 Feynman Parametrization for Independent Latent Precisions

Table 1 compares the Markov property and distributional closure of forward paths constructed using joint path-level precision or independent stepwise precision. For the reverse model, When the latent precisions Λ_t are sampled independently across steps, integrating out each precision yields an ℓ_p power transition density $f(x_t \mid x_{t-1})$. The resulting conditional path density $f(x_{1:T} \mid x_0) = \prod_{t=1}^{T} f(x_t \mid x_{t-1})$, however, generally does not belong to the joint ℓ_p power family.

The Feynman parametrization (Cai et al., 2025) expresses this product as a Dirichlet average of inverse power kernels over the full path.

For $k \in \mathbb{N}$, let $\Delta^{k-1} := \{w \in [0,1]^k : \sum_{i=1}^k w_i = 1\}$ denote the standard simplex. For positive exponents $\alpha_1, \dots, \alpha_k$ and $\bar{\alpha} := \sum_{i=1}^k \alpha_i$, the identity

$$\frac{1}{A_1^{\alpha_1} \cdots A_k^{\alpha_k}} = \frac{\Gamma(\bar{\alpha})}{\prod_{i=1}^k \Gamma(\alpha_i)} \int_{\Delta^{k-1}} \frac{\prod_{i=1}^k w_i^{\alpha_i - 1}}{\left(\sum_{i=1}^k w_i A_i\right)^{\bar{\alpha}}} \, dw \tag{44}$$

holds for all $A_1, \dots, A_k > 0$. This can be written as a Dirichlet mixture: for $W \sim \mathrm{Dir}_k(\alpha_1, \dots, \alpha_k)$,

$$\frac{1}{A_1^{\alpha_1} \cdots A_k^{\alpha_k}} = \mathbb{E}_W \left[\left(\sum_{i=1}^k W_i A_i \right)^{-\bar{\alpha}} \right]. \tag{45}$$

We now specialize this identity to ℓ_p-power factors.

Lemma E.1 (Feynman parametrization for a product of ℓ_p-power factors)**.** *Let $0 < p \le 2$, $\nu > 0$, and $\alpha_p := (\nu + n)/p$. For $t = 1, \dots, T$, define*

$$q_t(x_t) = Z_t^{-1} \left(1 + \frac{1}{\nu} \left\| \frac{x_t - \mu_t}{\sigma_t} \right\|_p^p \right)^{-\alpha_p},$$

where $Z_t > 0$ is the normalizing constant. Let $\gamma_{T,p} := -\frac{1}{T\alpha_p} = -\frac{p}{T(\nu+n)}$.

If $W = (W_1, \dots, W_T) \sim \mathrm{Dir}_T(\alpha_p \mathbf{1})$, then

$$\prod_{t=1}^T q_t(x_t) = C_F \, \mathbb{E}_W \left[\left(1 + \frac{1}{\nu} \sum_{t=1}^T W_t \left\| \frac{x_t - \mu_t}{\sigma_t} \right\|_p^p \right)^{-T\alpha_p} \right],$$

where $C_F := \prod_{t=1}^T Z_t^{-1}$.

Proof. Applying (44) with $A_t = 1 + \nu^{-1} \|(x_t - \mu_t)/\sigma_t\|_p^p$ and $\alpha_t = \alpha_p$ gives

$$\prod_{t=1}^T \left(1 + \frac{1}{\nu} \left\| \frac{x_t - \mu_t}{\sigma_t} \right\|_p^p \right)^{-\alpha_p} = \frac{\Gamma(T\alpha_p)}{\Gamma(\alpha_p)^T} \int_{\Delta^{T-1}} \frac{\prod_{t=1}^T w_t^{\alpha_p - 1}}{\left(1 + \frac{1}{\nu} \sum_{t=1}^T w_t \left\| \frac{x_t - \mu_t}{\sigma_t} \right\|_p^p \right)^{T\alpha_p}} \, dw.$$

Multiplying both sides by $\prod_{t=1}^T Z_t^{-1}$ and using the density of $\mathrm{Dir}_T(\alpha_p \mathbf{1})$ yields

$$\prod_{t=1}^T q_t(x_t) = C_F \, \mathbb{E}_W \left[\left(1 + \frac{1}{\nu} \sum_{t=1}^T W_t \left\| \frac{x_t - \mu_t}{\sigma_t} \right\|_p^p \right)^{-T\alpha_p} \right].$$

Now raise both sides to $\gamma_{T,p} = -1/(T\alpha_p)$. Since $\gamma_{T,p} < 0$, the map $z \mapsto z^{\gamma_{T,p}}$ is convex on $(0, \infty)$. Jensen's inequality therefore gives

$$\left(\prod_{t=1}^{T} q_t(x_t)\right)^{\gamma_{T,p}} \leq C_F^{\gamma_{T,p}} \mathbb{E}_W \left[\left(1 + \frac{1}{\nu}\sum_{t=1}^{T} W_t \left\|\frac{x_t - \mu_t}{\sigma_t}\right\|_p^p\right)^{-T\alpha_p\gamma_{T,p}}\right].$$

Since $-T\alpha_p\gamma_{T,p} = 1$, we obtain

$$\left(\prod_{t=1}^{T} q_t(x_t)\right)^{\gamma_{T,p}} \leq C_F^{\gamma_{T,p}} \mathbb{E}_W \left[1 + \frac{1}{\nu}\sum_{t=1}^{T} W_t \left\|\frac{x_t - \mu_t}{\sigma_t}\right\|_p^p\right].$$

For $W \sim \mathrm{Dir}_T(\alpha_p \mathbf{1})$, $\mathbb{E}[W_t] = 1/T$. Hence

$$\left(\prod_{t=1}^{T} q_t(x_t)\right)^{\gamma_{T,p}} \leq C_F^{\gamma_{T,p}} \left(1 + \frac{1}{T\nu}\sum_{t=1}^{T} \left\|\frac{x_t - \mu_t}{\sigma_t}\right\|_p^p\right).$$

This gives the point-wise bound. Finally, integrating this bound against $q(x_{1:T})$ proves the desired bound. □

Since the relative scales vary over a nondegenerate Dirichlet distribution, their average generally cannot be reduced to a single exact ℓ_p-power density with fixed parameters. Thus, except in degenerate cases, independent time-wise ℓ_p-power factors lie outside the single joint ℓ_p-power family used in the main construction. Together with the generic failure of convolution closure for independent ℓ_p-power innovations, this justifies the claims in Table 1.

E.2 Justification of the Reverse Markov Path via Conditional Divergence Summation

The Markov reverse model with independent precisions Λ_t admits an objective based on local divergences, as described in Remark 3.5. We prove that the weighted objective $\mathcal{L}_{\mathrm{loc},p}$ in (12) equals $\mathcal{L}_{\mathrm{adj},p}$ up to a constant independent of θ.

Retain $g^{\mathrm{ind}}_{\theta,t}$ and $\gamma_{\mathrm{loc},p} = -p/(\nu + n)$ from the remark, and take $t \sim \mathrm{Unif}\{1, \ldots, T\}$. For $t \geq 2$, the reference $\tilde{f}_{t-1|t,0}(\cdot \mid X_t, X_0) = f_{t-1|t,0}(\cdot \mid X_t, X_0)$ is the forward conditional density in (9). At $t = 1$, use $\tilde{f}_{0|1,0}(\cdot \mid X_1, X_0) = f_{0|1}(\cdot \mid X_1)$.

Proposition E.2 (Local conditional divergences and adjacent loss)**.** *Under Assumption 3.1, suppose $X_0 \mid X_1$ has a density and $\mathbb{E}_F\|\tilde{f}_{t-1|t,0}(\cdot \mid X_t, X_0)\|_{1+\gamma_{\mathrm{loc},p}} < \infty$ for every t. For $w_t = \frac{p}{\nu+n}\left[\frac{C_{p,\nu,n}(\nu-p)}{\tau_{t-1}^n(\nu+n-p)}\right]^{p/(\nu+n-p)}$,*

$$\mathcal{L}_{\mathrm{loc},p} \equiv_\theta \mathcal{L}_{\mathrm{adj},p}. \tag{46}$$

Proof. Translation invariance makes $\|g^{\mathrm{ind}}_{\theta,t}(\cdot \mid x)\|_{1+\gamma_{\mathrm{loc},p}}$ independent of θ and x. Integrating the power density gives $\|g^{\mathrm{ind}}_{\theta,t}(\cdot \mid x)\|^{1+\gamma_{\mathrm{loc},p}}_{1+\gamma_{\mathrm{loc},p}} = (C_{p,\nu,n}/\tau_{t-1}^n)^{\gamma_{\mathrm{loc},p}}(\nu + n - p)/(\nu - p)$.

The chosen weights therefore yield

$$\left(\frac{g^{\mathrm{ind}}_{\theta,t}(y \mid x)}{\|g^{\mathrm{ind}}_{\theta,t}(\cdot \mid x)\|_{1+\gamma_{\mathrm{loc},p}}}\right)^{\gamma_{\mathrm{loc},p}} = -\frac{\gamma_{\mathrm{loc},p}}{w_t}\left(1 + \frac{1}{\nu}\left\|\frac{y - \mu_\theta(t,x)}{\tau_{t-1}}\right\|_p^p\right).$$

For $t \geq 2$, integrating the residual against $\tilde{f}_{t-1|t,0}$ gives its conditional expectation given (X_t, X_0). At $t = 1$, it gives the conditional expectation of the residual of X_0 given X_1. Substituting into (6) and applying iterated expectation therefore gives, for every t,

$$\begin{aligned}
&\mathbb{E}_F \mathcal{D}_{\gamma_{\mathrm{loc},p}}\left(\tilde{f}_{t-1|t,0}(\cdot \mid X_t, X_0)\|g^{\mathrm{ind}}_{\theta,t}(\cdot \mid X_t)\right) \\
&\quad = \frac{1}{w_t}\left(1 + \frac{1}{\nu}\mathbb{E}_F\left\|\frac{X_{t-1} - \mu_\theta(t, X_t)}{\tau_{t-1}}\right\|_p^p\right) + \frac{1}{\gamma_{\mathrm{loc},p}}\mathbb{E}_F\|\tilde{f}_{t-1|t,0}(\cdot \mid X_t, X_0)\|_{1+\gamma_{\mathrm{loc},p}}.
\end{aligned}$$

Multiplying by w_t and averaging over t yields

$$\mathcal{L}_{\mathrm{loc},p}(\theta) = \mathcal{L}_{\mathrm{adj},p}(\theta) + 1 + \mathbb{E}_t\left[\frac{w_t}{\gamma_{\mathrm{loc},p}}\mathbb{E}_F\|\tilde{f}_{t-1|t,0}(\cdot \mid X_t, X_0)\|_{1+\gamma_{\mathrm{loc},p}}\right].$$

The remaining terms are finite and independent of θ, proving (46). □

Thus, minimizing $\mathcal{L}_{\mathrm{adj},p}$ fits the Markov reverse model through the weighted conditional divergence objective in Remark 3.5. Stochastic sampling from this model draws a new precision Λ_t at each reverse step.

E.3 Sampling Algorithms for ℓ_p-Power Diffusion

Algorithm 1 summarizes stochastic sampling from $G_{\theta,p}$ and deterministic sampling with a denoiser-driven ODE. For the deterministic branch, $\mu(s)$ and $\sigma(s)$ smoothly interpolate the forward schedules, with $\mu(0) = 1$, $\sigma(0) = 0$, and $\sigma(s) > 0$ for $s > 0$; D_θ is extended continuously from the training times to any additional evaluation times. The associated denoiser and velocity are

$$D_\theta(t,x) = \frac{\mu_\theta(t,x)}{\mu_{t-1}}, \qquad t = 1,\dots,T,$$
$$v_{\theta,p}(s,x) = \frac{\dot\sigma(s)}{\sigma(s)}x + \left(\dot\mu(s) - \mu(s)\frac{\dot\sigma(s)}{\sigma(s)}\right)D_\theta(s,x), \qquad s > 0.$$

Algorithm 1 Stochastic and deterministic sampling for ℓ_p-power diffusion

Require: Predictor μ_θ, parameters (p,ν), and scales $\{\tau_{t-1}\}_{t=1}^T$
Require: For the ODE: schedules (μ,σ), denoiser D_θ, and grid $T = s_0 > \cdots > s_K = 0$
Ensure: Generated sample x from either branch

Draw $x \sim F^{(n)}_{p,\nu}(0, I_n)$

Stochastic sampler

$S \leftarrow \|x\|_p^p$
for $t = T,\dots,1$ **do**
  $m \leftarrow \mu_\theta(t,x)$
  $\kappa \leftarrow \nu + (T-t+1)n$
  $a \leftarrow \tau_{t-1}((\nu+S)/\kappa)^{1/p}$
  Draw $x \sim F^{(n)}_{p,\kappa}(m, aI_n)$
  if $t > 1$ **then**
    $S \leftarrow S + \|(x-m)/\tau_{t-1}\|_p^p$
  end if
end for

Network evaluations: T.

Deterministic sampler (Heun)

for $i = 0,\dots,K-1$ **do**
  $h \leftarrow s_{i+1} - s_i$
  $d \leftarrow v_{\theta,p}(s_i, x)$
  $\widetilde{x} \leftarrow x + hd$
  if $s_{i+1} > 0$ **then**
    $\widetilde{d} \leftarrow v_{\theta,p}(s_{i+1}, \widetilde{x})$
    $x \leftarrow x + \frac{h}{2}(d + \widetilde{d})$
  else
    $x \leftarrow \widetilde{x}$ ▷ Final Euler step
  end if
end for

Network evaluations: $2K - 1$.

The stochastic branch follows the exact conditionals of $G_{\theta,p}$ in Lemma 3.3, with draws implemented using Proposition 2.4. For the independent-precision model in Remark 3.5, set $\kappa = \nu$, $a = \tau_{t-1}$, and omit the statistic S.

The deterministic branch uses the Heun scheme of Karras et al. (2022, Algorithm 1), with a final Euler step to avoid evaluating the velocity at $\sigma(0) = 0$. With the conditional-mean denoiser of $Y_s = \mu(s)X_0 + \sigma(s)E$, where $E \sim F^{(n)}_{p,\nu}(0, I_n)$ is independent of X_0, this velocity transports the marginals provided the flow is well posed, initialized from the correct terminal law, and integrated exactly. For $1 \le p < 2$, the conditional p-centers targeted by the two objectives generally differ from conditional means (Section 3.3), so marginal preservation is not guaranteed. The initial prior matches the terminal law when $\mu(T) = 0$ and $\sigma(T) = 1$; for a truncated schedule, it is generally an approximation.

F EXPERIMENTAL DETAILS

We study rare-mode coverage and the choice of loss exponent through two simulation experiments: a heavy-tailed mixture (Appendix F.2) and four-color MNIST (Appendix F.4). We also evaluate the ℓ_p-power construction on character-level typo correction (Appendix F.3) and image denoising and unconditional generation (Appendix F.5). In this section, we summarize the basic experimental setup for each task.

F.1 COMMON SETUP

We summarize the shared controls and reporting conventions for each task below. Detailed experimental settings are provided in the corresponding subsections.

Simulation: rare-mode mixture. Within each matched Adj./Dns. pair, the methods share the training sample, standardization, calibration, initialization, and random seed. Results are reported as the mean and sample standard deviation over three seeds, with the mixture geometry fixed across runs (Appendix F.2).

Simulation: four-color MNIST. Both figures use the same intensity bins and RGB palette. Reconstruction risk is averaged over repeated color-perturbation and corruption draws, and the optimal-exponent curves aggregate image-wise minimizers. The color construction, estimator, and averaging procedure are detailed in Appendix F.4.

Application: image denoising and unconditional generation. Within each dataset, all methods share the data partitions, fixed corrupted training images, backbone architecture, and training budget. The same trained model is evaluated on restoration and unconditional generation, with restoration measured against clean test targets. Results are reported as the mean and sample standard deviation over three seeds (Appendix F.5).

Application: character-level typo correction. All diffusion models share the data partitions, precomputed corruptions, backbone architecture, and training budget. Results are reported as the mean and sample standard deviation over three seeds, varying initialization, batch order, and diffusion noise while sharing the precomputed corruptions. Norvig-10 is a deterministic baseline (Appendix F.3).

F.2 RARE-MODE COVERAGE UNDER HEAVY-TAILED NOISE

Data distribution. The primary comparison in Table 2 and the ablations below use the mixture in equation 15 with $n = 30$, $r = 6$, $\alpha = 1.7$, and sixteen components. The weights are $\pi_k = k^{-1.2} / \sum_{j=1}^{16} j^{-1.2}$, the component scale is 0.55, and the centers are constructed as

$$\mu_k = 7.5\, B v_k, \qquad B \in \mathbb{R}^{n \times r}, \qquad B^\top B = I_r, \qquad \|v_k\|_2 = 1.$$

Here r denotes the dimension of the reference subspace used to place the centers. We obtain B by QR factorization of a Gaussian matrix and select the unit vectors $v_k \in \mathbb{R}^r$ using a max-min restart procedure followed by Riesz-energy repulsion. The geometry is fixed across runs; training datasets are generated using seeds 0, 1, and 2.

Objectives and ablations. The main comparison pairs the adjacent (Adj.) and denoising (Dns.) objectives with matched $F_{p,\nu}$ forward noise and ℓ_p^p residuals for $p \in \{1.2, 1.6, 2.0\}$. At $p \in \{1.2, 1.6\}$, the ablations separately vary the noise family and residual exponent according to Table 6. The matched $p = 2$ setting is the Student-t_ν path with a quadratic residual; DDPM is the corresponding Gaussian denoising baseline. DLPM and DLIM use stable diffusion noise with index 1.7.

Training and sampling. All methods use an MLP denoiser with width 512 and depth 4, trained on 10^5 points for $25{,}000$ steps with batch size 512. We use AdamW with learning rate 2×10^{-4}, 500 warmup steps, gradient clipping at 20, and EMA decay 0.999. Coordinates are standardized using coordinatewise medians and MADs estimated from the training data. We use a cosine forward

Table 6: Forward noise and residual penalties for the mixture experiment. The paired Adj./Dns. ablations use $p \in \{1.2, 1.6\}$; the matched configuration additionally includes $p = 2$.

Setting	Forward noise	Residual penalty
Matched p	$F_{p,\nu}$	ℓ_p^p
t_ν path, ℓ_p^p	$F_{2,\nu}$ (Student-t_ν)	ℓ_p^p
F_p path, ℓ_2^2	$F_{p,\nu}$	ℓ_2^2
Gauss. path, ℓ_p^p	Gaussian	ℓ_p^p
DDPM	Gaussian	Denoising ℓ_2^2
DLPM / DLIM	Stable, $\alpha = 1.7$	DLPM objective

schedule over $[10^{-3}, 0.999]$, $T = 100$ reverse steps, and $\nu = 10$ for the ℓ_p-power noise families. For each noise family, residual exponent, and seed, Adj. and Dns. share the training data, standardization, calibration, initialization, and random seed. Ht-VDM and DDPM use ODE sampling. DLPM uses its ancestral reverse sampler; DLIM uses the same checkpoint with the reverse innovation omitted. Each method generates 5×10^4 samples and is evaluated against an independent reference sample of 10^5 points from the mixture. Results are reported as the mean and sample standard deviation over the three seeds.

Mode assignment. Generated and reference samples are assigned to their nearest center in standardized coordinates. The rejection radius is the empirical 0.99 quantile of the reference nearest-center distances. Samples beyond this radius, as well as non-finite samples, are left unassigned. Let $\hat{\pi}_k$ and π_k^{ref} denote the fractions of all generated and reference samples assigned to component k, respectively; these fractions are not renormalized after rejection. Let $\mathcal{K}_{\mathrm{head}}$ contain the eight components with the largest population weights π_k, and let $\mathcal{K}_{\mathrm{rare}}$ contain the remaining eight.

Evaluation metrics. The assigned-mode mass discrepancy, abbreviated as TV in Table 2, is

$$\mathrm{TV}_{\mathrm{assigned}} := \frac{1}{2} \sum_{k=1}^{16} \left| \hat{\pi}_k - \pi_k^{\mathrm{ref}} \right|.$$

This excludes the unassigned category and uses no renormalization. The total variation distance including rejection would additionally contain $\frac{1}{2}|\hat{\pi}_\varnothing - \pi_\varnothing^{\mathrm{ref}}|$, where the two terms are the generated and reference rejection fractions.

For $\pi_k^{\mathrm{ref}} > 0$, define the mass-based mode recall and its head/rare averages by

$$\mathrm{Rec}_k = \min\left\{1, \frac{\hat{\pi}_k}{\pi_k^{\mathrm{ref}}}\right\}, \qquad R_S = \frac{1}{8} \sum_{k \in \mathcal{K}_S} \mathrm{Rec}_k, \quad S \in \{\mathrm{head}, \mathrm{rare}\}.$$

Table 2 reports these recalls. The ablation results in Table 7 report the mass-based F1 score

$$\mathrm{F1}_k := \frac{2 \min\{\hat{\pi}_k, \pi_k^{\mathrm{ref}}\}}{\hat{\pi}_k + \pi_k^{\mathrm{ref}}}, \qquad \hat{\pi}_k + \pi_k^{\mathrm{ref}} > 0.$$

For positive masses, this is the harmonic mean of Rec_k and $\min\{1, \pi_k^{\mathrm{ref}}/\hat{\pi}_k\}$. It is zero when the reference mass is positive and the generated mass is zero. Rare F1 and Head F1 average $\mathrm{F1}_k$ over their respective sets of eight modes.

Rare-recall gap across dimensions. We vary n, r, and the component-noise index α using the same mixture weights, component scale, and center construction. The full factorial grid is

$$n \in \{32, 64, 256\}, \quad n/r \in \{1, 4, 16\},$$
$$\alpha \in \{1.7, 2.0\}, \qquad p \in \{1.0, 1.2, \ldots, 2.0\}.$$

Each Adj./Dns. pair uses matched $F_{p,\nu}$ forward noise and ℓ_p^p residuals with $\nu = 10$, evaluated over three paired seeds. Training, sampling, and mode assignment follow the protocol above. Table 3 shows the subset $p \in \{1.2, 1.6, 1.8\}$ and reports the mean paired difference

$$\Delta R := R_{\mathrm{rare}}^{\mathrm{Adj.}} - R_{\mathrm{rare}}^{\mathrm{Dns.}},$$

so that positive values favor Adj. and negative values favor Dns.

Table 7: Rare-mode coverage and noise–loss ablations at $n = 30$, $r = 6$, and component-noise index $\alpha = 1.7$. Entries are means and sample standard deviations over three seeds. Bold denotes the better mean within each Adj./Dns. pair. $\mathrm{TV}_{\mathrm{assigned}}$ excludes the unassigned category.

Setting	Loss	$\mathrm{TV}_{\mathrm{assigned}} \downarrow$	Rare F1↑	Head F1↑
$p = 1.2$	Adj.	$\mathbf{0.012 \pm 0.003}$	$\mathbf{0.975 \pm 0.006}$	$\mathbf{0.986 \pm 0.005}$
	Dns.	0.062 ± 0.001	0.908 ± 0.004	0.952 ± 0.000
t_ν path, $\ell_{1.2}^{1.2}$	Adj.	$\mathbf{0.013 \pm 0.003}$	$\mathbf{0.974 \pm 0.003}$	$\mathbf{0.986 \pm 0.005}$
	Dns.	0.060 ± 0.001	0.910 ± 0.005	0.955 ± 0.003
$F_{1.2}$ path, ℓ_2^2	Adj.	0.025 ± 0.004	0.955 ± 0.020	0.977 ± 0.003
	Dns.	$\mathbf{0.023 \pm 0.005}$	$\mathbf{0.967 \pm 0.014}$	$\mathbf{0.981 \pm 0.004}$
Gauss. path, $\ell_{1.2}^{1.2}$	Adj.	$\mathbf{0.018 \pm 0.006}$	$\mathbf{0.972 \pm 0.004}$	$\mathbf{0.981 \pm 0.003}$
	Dns.	0.059 ± 0.005	0.902 ± 0.010	0.960 ± 0.003
$p = 1.6$	Adj.	$\mathbf{0.013 \pm 0.001}$	$\mathbf{0.972 \pm 0.003}$	$\mathbf{0.984 \pm 0.003}$
	Dns.	0.027 ± 0.003	0.962 ± 0.009	0.980 ± 0.003
t_ν path, $\ell_{1.6}^{1.6}$	Adj.	$\mathbf{0.013 \pm 0.001}$	$\mathbf{0.972 \pm 0.002}$	$\mathbf{0.987 \pm 0.001}$
	Dns.	0.027 ± 0.001	0.965 ± 0.005	0.979 ± 0.001
$F_{1.6}$ path, ℓ_2^2	Adj.	0.027 ± 0.006	0.951 ± 0.015	0.978 ± 0.004
	Dns.	$\mathbf{0.020 \pm 0.004}$	$\mathbf{0.970 \pm 0.017}$	$\mathbf{0.983 \pm 0.005}$
Gauss. path, $\ell_{1.6}^{1.6}$	Adj.	$\mathbf{0.019 \pm 0.003}$	$\mathbf{0.971 \pm 0.002}$	0.976 ± 0.002
	Dns.	0.023 ± 0.002	0.968 ± 0.011	$\mathbf{0.985 \pm 0.003}$
$p = 2.0$	Adj.	0.029 ± 0.014	0.947 ± 0.022	0.977 ± 0.012
	Dns.	$\mathbf{0.023 \pm 0.005}$	$\mathbf{0.969 \pm 0.011}$	$\mathbf{0.980 \pm 0.001}$
DDPM		0.020 ± 0.005	0.966 ± 0.010	0.980 ± 0.006
DLPM ($\alpha = 1.7$)		0.075 ± 0.007	0.845 ± 0.014	0.941 ± 0.007
DLIM ($\alpha = 1.7$)		0.620 ± 0.002	0.019 ± 0.004	0.098 ± 0.004

F.3 Rare-word Restoration in Typo Correction

We use WikiText-103 (Merity et al., 2016) with its official article-level training, validation, and test partitions. After detokenization and sentence segmentation, we retain sentences of 16–128 characters containing at least one alphabetic word of four or more letters and discard residual tokenization artifacts. The resulting splits contain 1,408,568, 3,267, and 3,876 sentences, respectively. The vocabulary contains 71 character symbols plus padding and unknown symbols, and all sentences are padded to length 128.

For each sentence, we uniformly select approximately half of the eligible words and replace approximately half of their letters with keyboard-adjacent letters, preserving case. Counts are rounded to the nearest integer, with ties rounded upward. All edits are substitutions, preserving word lengths and boundaries. We precompute 16 corrupted copies of each training sentence and sample one uniformly at each visit. Evaluation corruptions are generated once with a separate seed and shared across methods.

All diffusion models use a shared DiT-style architecture (Peebles & Xie, 2023) with 12 blocks, width 768, four attention heads, feed-forward width 3,072, and dropout 0.1. Characters are represented by jointly learned 64-dimensional embeddings normalized to norm 8. In addition to the diffusion loss, training includes an auxiliary character-classification loss to supervise character prediction. The diffusion state and corrupted sentence embeddings are separately projected and added position-wise, together with learned positional embeddings. Padding positions are included in diffusion. We compare ℓ_p models with $p \in \{1.2, 1.8\}$ and $\nu = 10$ against Gaussian DDPM (Ho et al., 2020a) and t-EDM (Pandey et al., 2025), also with $\nu = 10$.

Each model is trained for 50,000 steps with batches of 1,024 sentences sampled uniformly with replacement. AdamW (Loshchilov & Hutter, 2019) uses a learning rate of 5×10^{-4}, 2,000 warm-up steps, and cosine decay to zero. Weight decay 0.01 is applied to weight matrices, excluding embed-

dings, biases, and normalization parameters. We use bf16 autocast and evaluate EMA parameters with decay 0.999.

The dictionary baseline, Norvig-10 (Norvig, 2007), uses training word counts and searches same-length candidates within 10 substitutions. It preserves words already in the dictionary and words shorter than four letters, selecting other candidates by minimum substitutions, then training frequency and alphabetical order. Head–tail analysis includes eligible words with at least five training occurrences. Words capitalized in at least 90% of non-sentence-initial occurrences, or never observed outside sentence-initial positions, are excluded. The head and tail are defined by training-frequency thresholds covering approximately the upper and lower 10% of the remaining test occurrences.

We report word accuracy and word-level $F_{0.5}$ over eligible words, exact-match accuracy over complete sentences, and head and tail recall over corrupted words in each group. Results for diffusion models are reported as means and sample standard deviations over three seeds, varying initialization, batch order, and diffusion noise while sharing the precomputed corruptions. Norvig-10 is deterministic.

The following examples show actual sentence restoration results.

EXAMPLE 01

Clean	She had been converted from a commercial vessel in New Orleans for river and coastal fighting.
Corrupted	She had been **comceftse** from a commercial **vrwsrl** in New **Iroeznx** for **eicef** and coastal **rihhtunb.**
Norvig-10	She had been **concerted** from a commercial vessel in New Orleans for **eider** and coastal **righting.**
ℓ_p $(p = 1.2)$	She had been converted from a commercial vessel in New Orleans for river and coastal fighting.
t-EDM	She had been **comceftse** from a commercial **vrwsrl** in New **Iroeznx** for **rived** and coastal **righting.**
DDPM	She had been converted from a commercial **crasel** in New Orleans for river and coastal **righting.**

EXAMPLE 02

Clean	Turner, although wounded, remained on the tank until he was killed.
Corrupted	**Turjdf,** although **slunree,** remained on the **ramk ybtik** he was killed.
Norvig-10	**Turned,** although **sounded,** remained on the **rank** until he was killed.
ℓ_p $(p = 1.2)$	Turner, although wounded, remained on the tank until he was killed.
t-EDM	Turner, although **sounded,** remained on the **rank** until he was killed.
DDPM	Turner, although **wounder,** remained on the **rank** until he was killed.

EXAMPLE 03

Clean	Ise was briefly fitted with an aircraft flying-off platform for a Mitsubishi 1MF3 fighter on Turret No.
Corrupted	Ise was **grisrlh** fitted with an **wircfagr** flying-off **ooatfirk** for a Mitsubishi 1MF3 fighter on **Furrwf** No.
Norvig-10	Ise was **bristly** fitted with an aircraft flying-off platform for a Mitsubishi 1MF3 fighter on **Furred** No.
ℓ_p $(p = 1.2)$	Ise was briefly fitted with an aircraft flying-off platform for a Mitsubishi 1MF3 fighter on Turret No.
t-EDM	Ise was **grisrlh** fitted with an **aircrate** flying-off **platfirm** for a Mitsubishi 1MF3 fighter on **Furrwf** No.
DDPM	Ise was briefly fitted with an aircraft flying-off platform for a Mitsubishi 1MF3 fighter on **Furrwf** No.

EXAMPLE 04

Clean	Winter estates on the island, however, were better built and little damaged.
Corrupted	**Wungwr eztsyss** on the island, however, **sdre hdrter ghjlt** and little damaged.
Norvig-10	**Winger** estates on the island, however, were **herter guilt** and little damaged.
ℓ_p $(p = 1.2)$	Winter estates on the island, however, were better built and little damaged.
t-EDM	**Singer extates** on the island, however, were **hdrter** built and little damaged.
DDPM	**Singer** estates on the island, however, were **betted** built and little damaged.

EXAMPLE 05

Clean	The last of the tanks were built in the first quarter of 1942 and delivered at the end of the year.
Corrupted	The last of the **gajkw** were **nuuot** in the **dktst wuzrtde** of 1942 and delivered at the end of the year.
Norvig-10	The last of the **gamow** were **huyot** in the **ditsy** quarter of 1942 and delivered at the end of the year.
ℓ_p $(p = 1.2)$	The last of the tanks were built in the first quarter of 1942 and delivered at the end of the year.
t-EDM	The last of the **gajkw** were built in the first **quarted** of 1942 and delivered at the end of the year.
DDPM	The last of the **banks** were built in the first quarter of 1942 and delivered at the end of the year.

EXAMPLE 06

Clean	The German Imperial Navy ordered its first fifteen Type UB I boats on 15 October 1914.
Corrupted	The **Ferjwn Ompedjak** Navy **orrsdex** its first **cuvteeh Thoe** UB I boats on 15 October 1914.
Norvig-10	The **Ferman** Imperial Navy **oreades** its first fifteen **Thoe** UB I boats on 15 October 1914.
ℓ_p $(p = 1.2)$	The German Imperial Navy ordered its first fifteen Type UB I boats on 15 October 1914.
t-EDM	The German Imperial Navy **ordeded** its first **difteen Thie** UB I boats on 15 October 1914.
DDPM	The German Imperial Navy ordered its first **difteen** Type UB I boats on 15 October 1914.

EXAMPLE 07

Clean	Donald is armed with a special gun that can shoot plungers, popcorn or bubble gum.
Corrupted	**Xlnaod** is **srker** with a **zpedkap** gun **tyar** can shoot plungers, popcorn or **nyhble** gum.
Norvig-10	Donald is **drier** with a special gun **thar** can shoot **plumbers,** popcorn or bubble gum.
ℓ_p $(p = 1.2)$	Donald is armed with a special gun that can shoot plungers, popcorn or bubble gum.
t-EDM	**Conale** is **srker** with a special gun that can shoot plungers, **popcorm** or **nyhble** gum.
DDPM	**Conald** is armed with a special gun that can shoot plungers, popcorn or **nyhble** gum.

EXAMPLE 08

Clean	A head from La Venta is decorated with the head of a bird.
Corrupted	A **hwqd** from La Venta is **drdlfater sitn** the **gewd** of a bird.
Norvig-10	A **hwad** from La **Vents** is **decorater sign** the head of a bird.
ℓ_p $(p = 1.2)$	A head from La Venta is decorated with the head of a bird.
t-EDM	A head from La Venta is **decocated** with the head of a bird.
DDPM	A head from La Venta is **decocated** with the head of a bird.

EXAMPLE 09

Clean	The system moved northwestward and later turned to the northeast around the subtropical ridge.
Corrupted	The system **jpvec horfgsdateard** and later turned to the **noeyneszt** around the **eybrtkpival fudgr.**
Norvig-10	The system moved northwestward and later turned to the northeast around the subtropical **fudge.**
ℓ_p $(p = 1.2)$	The system moved northwestward and later turned to the northeast around the subtropical ridge.
t-EDM	The system moved **northeratward** and later turned to the **northeest** around the **suberipical fudgr.**
DDPM	The system moved **northeestward** and later turned to the **noeyneszt** around the subtropical ridge.

EXAMPLE 10

Clean	High winds killed a man in Hong Kong after knocking him off a platform.
Corrupted	**Hugb eknrs kulosd** a man in Hong Kong **acret** knocking him off a platform.
Norvig-10	**Hugh eobrs** killed a man in Hong Kong after knocking him off a platform.
ℓ_p $(p = 1.2)$	High winds killed a man in Hong Kong after knocking him off a platform.
t-EDM	High **wines** killed a man in Hong **King** after knocking him off a platform.
DDPM	**Hugh** winds killed a man in Hong Kong after knocking him off a platform.

F.4 OPTIMAL $p^\star$ CURVE OF FOUR-COLOR MNIST DATASET

The curve is the mean, over the first 50 MNIST test images, of the exponent $p^\star(\rho, \eta)$ that minimizes reconstruction MSE. Each grayscale image is painted by the bins $[0, 64)$, $[64, 128)$, $[128, 192)$, $[192, 256)$, with palette

$$c_0 = (0.20, 0.50, 0.80),\ c_1 = (0.35, 0.50, 0.65),\ c_2 = (0.65, 0.50, 0.35),\ c_3 = (0.80, 0.50, 0.20).$$

The clean image is the clipped perturbation $X = \mathrm{clip}(c_{g(i)} + \delta_i, 0, 1)$, where δ_i is drawn i.i.d. from $\mathrm{Unif}[-\eta, \eta]^3$ and $g(i)$ is the bin of pixel i. The sweep uses $\eta \in \{0, 0.05, 0.10, 0.15\}$. Salt-and-pepper then acts on X: each pixel is replaced, in all three channels at once, by 0 or by 1 with probability $\rho/2$ each. The grid is $\rho \in \{0, 0.05, \ldots, 1\}$.

The estimator is given the true four-bin partition. Inside each bin and each RGB channel it returns one empirical p-center of the observed intensities: the median at $p = 1$, the mean at $p = 2$, and a bisection of $\sum_i \mathrm{sign}(m - Y_i)\,|m - Y_i|^{p-1}$ otherwise. Every pixel in that bin is then repainted with the resulting constant. The risk $R^{(\eta)}(p)$ is the mean squared error of this reconstruction against X,

averaged over pixels, channels, and 64 independent draws of $(\delta, \text{salt-and-pepper})$. For each image the minimizer $p^\star$ is read from the grid

$$\{1, 2\} \cup \text{linspace}(1.01, 1.12, 8) \cup \text{linspace}(1.18, 1.50, 5) \cup \{1.7, 1.9\},$$

with a three-point parabolic refinement about the grid argmin when that quadratic opens upwards. The plotted curve is the mean of these 50 image-wise minimizers, and the band is their interquartile range. The same figure also shows the mean risk at the chosen $p^\star$.

F.5 Image Denoising and Unconditional Generation

A single network is trained on a training set in which a fraction ρ of the training images is corrupted, and the same trained parameters are evaluated on two tasks: restoration of corrupted test images and unconditional generation. No label indicates which training images are corrupted, and the clean originals of the corrupted training images are never used during training.

We use four datasets (Table 8). CelebA follows the official partition; each image is centre-cropped to 140×140 and resized to 64×64 with interpolation. CIFAR-10 and SVHN keep their official test sets, and their official training sets are split into training and validation sets with class stratification (CIFAR-10 $45{,}000/5{,}000$; SVHN $90\%/10\%$, without the `extra` set). For Omniglot, the background and evaluation sets are pooled (50 alphabets, 1,623 characters, 32,460 drawings), each drawing is resized to 32×32 grayscale, and the drawings are split $9 : 1 : 1$ within each character so that every character appears in every split. Pixel values are scaled to $[-1, 1]$. No data augmentation is used. The validation split is used only to monitor training, and all reported numbers are computed on the test split.

Table 8: Datasets. N_{corr} is the number of corrupted training images at $\rho = 0.75$.

Dataset	Resolution	Channels	Train	Valid	Test	N_{corr}
Omniglot	32×32	1	26,559	2,950	2,951	19,920
CIFAR-10	32×32	3	45,000	5,000	10,000	33,750
SVHN	32×32	3	65,931	7,326	26,032	49,449
CelebA	64×64	3	162,770	19,867	19,962	122,078

All corruption is per-channel salt-and-pepper noise with rate $q = 0.5$. Each coordinate (c, h, w) is selected independently with probability q, and a selected coordinate is replaced by -1 or 1 with equal probability. Since the channels of a pixel are corrupted independently, at least one channel is replaced in $1 - (1 - q)^3 = 87.5\%$ of the pixels of an RGB image. The model is not told which coordinates were replaced. The corrupted images are the first $\lceil \rho N \rceil$ images of a fixed random permutation of the training split, with $\rho = 0.75$. Because this permutation depends only on the dataset, all methods and all seeds are trained with the same set of corrupted images. Each selected image $x^\star$ is replaced by a single corrupted copy $\tilde{x}$ drawn once before training, and this copy is kept unchanged throughout training. The remaining $1 - \rho$ of the images are left unchanged. The diffusion target x_0 of a training example is the image contained in the training set, that is, $\tilde{x}$ for a corrupted image and $x^\star$ otherwise.

At every optimization step, a new observation $y = \text{corrupt}(x_0)$ is drawn for every example in the batch. The noise family and the rate q are the same, and the draw is new at every step. The observation of a corrupted training image therefore carries two independent corruptions: the fixed training-set corruption and the observation corruption resampled at each step. Regardless of whether the training image is corrupted, the observation of each example is replaced by the zero image with probability $\lambda = 0.5$, and this choice is also redrawn at every step. The network receives x_t and y concatenated along the channel axis. The zero observation is the input used for unconditional generation.

All methods use the same U-Net. The input has $2C$ channels (x_t and y) and the output has C channels. The base width is 128, with channel multipliers $(1, 2, 2, 2)$ at 32×32 and $(1, 2, 2, 4)$ at 64×64. Each resolution has two residual blocks, and self-attention (head dimension 64) is placed at 16×16 for 32×32 inputs and at 16×16 and 8×8 for 64×64 inputs. The residual blocks use group normalization with 32 groups and dropout of 0.1. The noise level passes through a 128-dimensional sinusoidal embedding and a two-layer MLP of width 512, and enters every residual block through adaptive group normalization whose modulation weights are initialized to zero. The network has 38.3M parameters at 32×32 and 77.5M at 64×64.

Table 9: Selected configurations for the image experiments.

Component	ℓ_p-power model	DDPM	t-EDM
Prediction parameterization	x_0-prediction	ϵ-prediction	EDM preconditioning
Training objective	$\mathcal{L}_{\mathrm{dns},p}$	Unweighted L_{simple}	t-EDM objective
Noise distribution	ℓ_p-power	Gaussian	Student-t
Sampler	Heun	Ancestral	Heun

In the ℓ_p-power model, the network predicts x_0 and is trained with the denoising surrogate objective $\mathcal{L}_{\mathrm{dns},p}$ with $p \in \{1.2, 1.8\}$ and $\nu = 10$. We use the cosine schedule of Nichol & Dhariwal (2021) with $T = 20$ and set the decoder floor to $\tau_{\min} = 0.0894$, which is the value of σ_1 of this schedule at $T = 20$. The timestep is drawn uniformly from $\{1, \ldots, T\}$ for each image. The terminal state is $x_T \sim F_{p,\nu}(0, I)$. DDPM (Ho et al., 2020a) uses ϵ-prediction with the unweighted objective L_{simple}, with the same schedule and $T = 20$ as the ℓ_p-power model. t-EDM (Pandey et al., 2025) uses Student-t noise with $\nu = 10$ and the EDM preconditioning (Karras et al., 2022). During training, the noise level is drawn from $\log \sigma \sim \mathcal{N}(P_{\mathrm{mean}}, P_{\mathrm{std}}^2)$. Since the published constants assume a unit data scale, all noise levels are multiplied by the data standard deviation σ_{data} measured on each training set (0.809, 0.730, 0.667, and 0.771 for Omniglot, CIFAR-10, SVHN, and CelebA, respectively). That is, $P_{\mathrm{mean}} = -1.2 + \log \sigma_{\mathrm{data}}$, $P_{\mathrm{std}} = 1.2$, and $\sigma \in [0.002, 80] \cdot \sigma_{\mathrm{data}}$, with a time-discretisation exponent of 7.

For the experiments, all models are trained for 100,000 optimization steps. All runs use AdamW (Loshchilov & Hutter, 2019) with $(\beta_1, \beta_2) = (0.9, 0.999)$, a learning rate of 2×10^{-4}, weight decay of 0.01, and a batch size of 256. The learning rate is increased linearly over the first 2,000 steps and held constant afterwards, and no gradient clipping is applied. An exponential moving average of the parameters is maintained with decay 0.9995 and bias correction, and all evaluations use these averaged parameters.

For restoration, the network is conditioned on the observation, and $K = 16$ samples are drawn, averaged, and clipped to $[-1, 1]$. For generation, the zero image is given as the observation. The ℓ_p-power model applies the deterministic Heun branch of Algorithm 1 in the time coordinate over $T = 10$ training timesteps, and the final step to $t = 0$ is taken with Euler's method. This costs $2T - 1 = 19$ network evaluations. DDPM uses ancestral sampling with the fixed small posterior variance (20 evaluations), and t-EDM uses Heun's method with 10 steps (19 evaluations). The number of network evaluations is therefore not matched: the ℓ_p-power model and t-EDM use 19, and DDPM uses 20.

Table 9 summarizes the best-performing configurations among those evaluated for each method in the image denoising and unconditional generation experiments.

G Discussion and Limitations

We discuss the scope of our theoretical results and the limitations of their empirical interpretation.

Discrete and continuous constructions. Our exact path-space projection is established for the coupled discrete construction. Matching its single-time kernels through a continuous GSM construction does not establish a continuum limit of the discrete path law. For $1 < p < 2$, the population-optimal predictor is a conditional p-center, which need not equal the conditional mean and therefore does not generally identify a probability-flow velocity. We treat the deterministic denoising procedure as a heuristic refinement: neither preservation of a prescribed marginal curve nor improvement over direct prediction follows from the discrete projection result.

Parameter selection. Prior Student-t-based generative models reports limited sensitivity to ν in high-dimensional experiments. (Kim et al., 2024; Pandey et al., 2025; Cho et al., 2026; Bouayed et al., 2025) However, we have not examined whether the effect of ν on performance depends on (p). In this work, we fix $\nu = 10$ in the high-dimensional experiments, leaving a systematic study of its interaction with p and data-dependent selection of (p, ν) for future work.

In addition, our path-space reduction holds for any fixed positive decoder schedule. However, the schedule is treated as a design choice, and our theory does not specify how to select it for a given task. Developing data-dependent selection rules and characterizing their effects on reconstruction and generation remain directions for future work.

Scope of the theoretical analysis for relative performance of the two objectives. Our local population analysis characterizes the leading-order reconstruction-risk difference between adjacent and denoising prediction under an idealized mixture assumption. However, we have not used the resulting response coefficients to quantitatively predict the performance gaps observed in our experiments. Extending this comparison to continuous dominant components and accounting for finite-sample estimation, model capacity, and optimization also require further analysis.

In the separate bounded-corruption setting, our results hold for each fixed η under the stated assumptions. They do not characterize how the reconstruction-optimal exponent or its improvement over the endpoint choices varies with the clean-signal variation η. A quantitative analysis of these dependencies, together with their interaction with the corruption rate, could provide more practical guidance for selecting the training objective and loss exponent.

Interpretation across applications. The applications differ in the meaning of atypical training targets. Typo correction uses clean sentences as targets, so rare words remain valid structures to recover. In the image experiments, 75% of the training targets contain fixed impulse corruption, and their clean counterparts are excluded from training. Thus, X_0 need not represent a clean image, and reduced sensitivity to atypical targets may benefit reconstruction against clean references. This distinction offers a possible interpretation of the advantage of adjacent training in Table 4 and the better mean restoration scores of denoising training at $p = 1.8$ on SVHN and CelebA in Table 5. However, the denoising models also exhibit reduced head-word recall. Explaining this broader degradation requires separating the effects of embedding learning, auxiliary classification, and optimization from the population-level mechanism.

Dependence on evaluation criteria. The preferred objective also depends on the evaluation criterion. On SVHN at $p = 1.8$, adjacent training achieves lower FID and higher precision than denoising training, but substantially lower recall (0.080 versus 0.540). Similarly, the $p = 1.2$ denoising model does not achieve the highest generation recall on every dataset despite its strong restoration and FID results. A local reconstruction-risk comparison therefore does not establish a common ordering across reconstruction accuracy, generative fidelity, and coverage.