VARIANCE PATH DIFFUSION: PER-REGION NOISE SCHEDULES FOR INPAINTING AND MOTION IN-BETWEENING

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models corrupt each coordinate on a single, predefined noise schedule, yet most conditional tasks, such as inpainting, local editing, and motion in-betweening, require noise levels that differ across space and time. Existing methods design such region-wise schedules by hand, each rediscovering empirically the constraints a schedule must satisfy. We derive these constraints instead, by modeling the variance as a random variable: Variance Path Diffusion (VPD) models the cumulative noise as a Gamma process pinned in expectation to a nominal schedule, making variance allocation a free design variable. We prove that the natural shared-rate construction cannot meet a sufficient condition for well-behaved training at any finite stochasticity, and repair it with a Time-Varying-Rate (TVR) construction based on Beta thinning, which restores admissible VPD properties while leaving the marginals intact, and show that the randomness of the path is bounded and empirically does not affect quality. What remains is structure. A single allocation parameter γ, applied at sampling time to one model trained once, spans unconditional generation, editing, and inpainting. VPD covers temporal allocation for motion. VPD achieves the lowest or tied all-joint keyframe error on HumanML3D in-betweening, and sets a new state of the art for inpainting on FFHQ, LSUN Bedroom, and ImageNet.

1 INTRODUCTION

Diffusion models corrupt every coordinate according to a single predefined schedule Ho et al. (2020); Song et al. (2021); Karras et al. (2022). While this symmetry across coordinates is appropriate for unconditional generation, it is unnatural for most conditional tasks. In image inpainting, one region must be synthesized from pure noise while another must be preserved; in local editing, two regions should be assigned different noise levels; in motion in-betweening, keyframes must be preserved while the trajectory between them is generated. These tasks share one requirement: the noise levels vary across regions of the data.

A growing body of empirical work addresses this exact requirement by assigning different noise levels to different regions such as region-aware diffusion Kim et al. (2025), per-pixel time indices Mayet et al. (2025), patch-wise generation Papa et al. (2026), and per-token noise levels along the sequence axis Chen et al. (2024). In each of these methods the region-wise schedule is constructed *by hand*. What is missing is a unifying **principle**. No existing method derives the region-wise formulation from the diffusion process itself and states the conditions such a schedule must satisfy for the resulting process to remain a valid diffusion model with exact reverse transitions, or prescribes how the variance should be allocated.

This paper aims to supply such a principle by elevating the variance itself to a random variable. Rather than empirically assigning different time steps across regions, we model the variance as a stochastic process, specifically a Gamma process, whose mean is pinned to the nominal schedule. We derive the constraint this choice imposes and refer to the resulting framework as Variance Path Diffusion (VPD). The construction of the Gamma process keeps the entire model closed-form, where the forward marginals are Gaussian conditioned on the realized variance, the reverse transition is a Brownian bridge, the ELBO decomposes exactly, and the variance posterior is a generalized inverse

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.

Gaussian distribution. Moreover, a Gamma-Dirichlet duality separates the path into a total budget and its allocation over time, and shows that the data constrain only the total, while the allocation is a design variable. Section 5 takes this literally across space, specifying the allocation with a single ratio under a constraint that fixes the total.

The naive construction, however, fails. Gamma process additivity requires a rate parameter (b) shared across time, while matching the nominal schedule ($\bar{\sigma}_t^2$) in expectation ties the shape parameter (A_t) to the schedule, $A_t = b\bar{\sigma}_t^2$. Taking the ratio of the two endpoints eliminates the rate and pins the dynamic range of the shape to $\sigma_{\min}^2/\sigma_{\max}^2$, a quantity that no choice of schedule profile, number of steps, or stochasticity scale can alter. We identify a sufficient condition, $A_1 > 1$, under which the construction is well behaved. Freeing the rate parameter removes the failure while preserving every closed-form property (§ 4).

Region-wise variance allocation follows naturally in this framework (Fig. 1). Scaling the variance of each region by a deterministic multiplier, subject to a single pixel-weighted constraint that keeps the average on the nominal schedule, yields a one-parameter dial γ that moves the model continuously across unconditional generation ($\gamma = 1$), local editing (moderate γ), and inpainting ($\gamma \gg 1$). Critically, VPD is not constrained to the spatial axis. The same framework in image inpainting transfers to the temporal axis in motion in-betweening, where keyframes play the role of the preserved region. Empirically, VPD achieves state-of-the-art performance on FFHQ, LSUN Bedroom, and ImageNet inpainting and keyframe error metrics on HumanML3D motion in-betweening. Furthermore, VPD provides the dial γ which controls how much of the unmasked region to preserve, showing gradual transitions from inpainting to pure generation. We therefore argue that VPD addresses a broad family of variance-allocation problems.

Our contributions are: (1) **A framework.** We construct Variance Path Diffusion, a diffusion model whose variance path is a Gamma process pinned in expectation to a nominal schedule with every quantity in closed-form formulation (§ 2). We further generalize VPD via a multiplicative Beta-thinning construction with every quantity in closed form, while alleviating problems caused by the shared-rate construction of the Gamma process (§ 4). (2) **A negative result.** Under a shared rate, the condition that guarantees well-behaved training and the stochasticity that motivated the construction are mutually exclusive (§ 3). (3) **Allocation across domains.** We instantiate per-region variance allocation with a single control parameter γ and show that identical mechanisms cover spatial allocation in image domain, and temporal allocation in the motion domain (§ 5).

2 Variance Path Diffusion

Throughout this section, we treat the variance as a single scalar process shared by all coordinates. The extension to per-region variance, which is what makes inpainting and motion in-betweening possible, is introduced in § 5.

2.1 Notation and setup

Let $x_0 \in \mathbb{R}^d$ denote the data and let $t = 1, \cdots, T$ index discrete diffusion steps. We adopt the variance-exploding (VE) parameterization throughout[1], in which the forward process adds noise without rescaling the signal:

$$x_t = x_{t-1} + \sqrt{\beta_t}\epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, I), \tag{1}$$

such that $x_t = x_0 + \sqrt{B_t}\epsilon$ with $B_t = \sum_{s \le t} \beta_s$ the cumulative variance at step t. The terminal distribution is $p(x_T) = \mathcal{N}(0, \sigma_\star^2 I)$ with $\sigma_\star^2 \gg \mathrm{Var}[x_0]$. In standard diffusion, $\{\beta_t\}$ is a fixed schedule. VPD instead treats it as a random variable.

2.2 The VPD construction

Definition 2.1 (VPD forward process). Let the variance increments be independent Gamma variables $\beta_t \sim \mathrm{Ga}(a_t, b)$, where a_t is the shape at step t and $b > 0$ is the rate, shared across all steps.

[1] A variance-preserving parameterization follows by reparameterizing the VE path, see Appendix E.

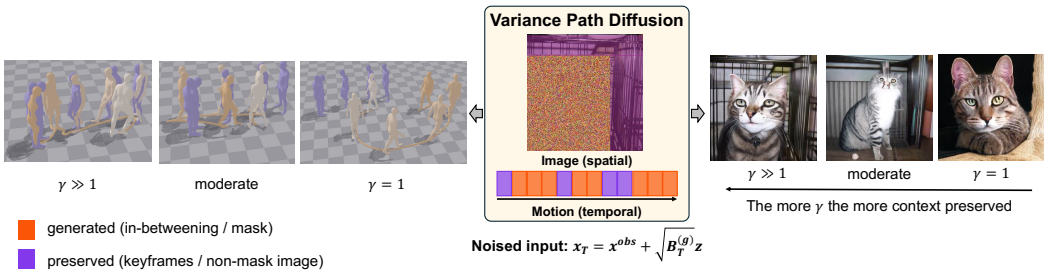

Figure 1: **Overview.** Variance Path Diffusion assigns each region its own noise level through a single ratio γ (center: masked regions start fully noised, the rest near the observation). The same model and the same γ serve both axes: across time in motion in-betweening (left), and across space in image (right). Larger γ preserves more of the observation, $\gamma = 1$ is unconditional generation.

By the additivity of the Gamma distribution with a shared rate parameter, the cumulative variance is again a Gamma distribution. Specifically, $B_t := \sum_{s \le t} \beta_s \sim \mathrm{Ga}(A_t, b), A_t := \sum_{s \le t} a_s$, and the forward process conditioned on the data and the realized variance is the Gaussian, $q(x_t|x_0, B_t) = \mathcal{N}(x_t; x_0, B_t I)$.

The variance is positive almost surely, since Gamma is supported on $\mathbb{R}_{>0}$. Furthermore, the path $B_1 \le \cdots \le B_T$ is monotone almost surely, since the increments are positive. Monotonicity can be viewed as monotone decay of the SNR (Appendix E.1).

Coupling to a noise schedule. The shapes $\{a_t\}$ are determined by requiring the resulting process to match a target noise schedule $\bar{\sigma}_t^2$ in expectation, $\mathbb{E}[B_t] = \frac{A_t}{b} = \bar{\sigma}_t^2$ which yields $A_t = b\bar{\sigma}_t^2$. It is convenient to separate the scale of the shapes from their profile. Let $\phi_t := \bar{\sigma}_t^2/\bar{\sigma}_T^2 \in (0, 1]$ denote the shape profile and $\kappa := A_T$ its scale, such that $A_t = \kappa\phi_t$ and $b = \kappa/\bar{\sigma}_T^2$. Since $B_t \sim \mathrm{Ga}(A_t, b)$, which implies $\mathbb{E}[B_t] = A_t/b$ and $\mathrm{Var}[B_t] = A_t/b^2$, the relative standard deviation (RSD) is $\mathrm{RSD}[B_t] = A_t^{-1/2} = (\kappa\phi_t)^{-1/2}$, so the shape alone governs how far the realized variance deviates from its mean while the mean stays pinned to the schedule for every κ and as $\kappa \to \infty$ the path concentrates and VPD recovers the deterministic VE sampler. The shared rate, however, leaves no freedom in ϕ. The pin $A_t = b\bar{\sigma}_t^2$ forces $\phi_t = \bar{\sigma}_t^2/\bar{\sigma}_T^2$ where the shape profile *is* the schedule profile, and κ is the only remaining degree of freedom. This identity is the source of the degeneracy analyzed in §3, and freeing ϕ from the schedule is precisely what §4 accomplishes.

2.3 Reverse transitions

Given the variance path, the reverse transition is a Gaussian conjugate update. Using $q(x_{t-1}|x_0, B_{t-1}) = \mathcal{N}(x_0, B_{t-1}I)$ and $q(x_t|x_{t-1}, \beta_t) = \mathcal{N}(x_{t-1}, \beta_t I)$, we obtain $q(x_{t-1}|x_t, x_0, B_t) = \mathcal{N}\left(\frac{\beta_t\, x_0 + B_{t-1}\, x_t}{B_t}, \frac{B_{t-1}\beta_t}{B_t} I\right)$. Introducing the *normalized fraction*, $w := \frac{\beta_t}{B_t} \sim \mathrm{Beta}(a_t, A_{t-1})$ (Gamma-Beta relation Appendix B), the transition becomes

$$x_{t-1} = (1 - w)x_t + wx_0 + \sqrt{w(1 - w)B_t}z, \quad z \sim \mathcal{N}(0, I). \tag{2}$$

The mean is an affine interpolation between the current state and the data, and with w the interpolation weight, the injected noise has variance $w(1 - w)B_t$. The variance $w(1 - w)B_t$ is non-negative only when $B_{t-1} \le B_t$, and the transition would be undefined otherwise. We return to this condition in § 4, where the rate is allowed to vary and monotonicity must be imposed explicitly.

Committed paths Conditioning on the variance path therefore makes the entire reverse chain Gaussian, so the path can be drawn once before sampling (Appendix C.8)

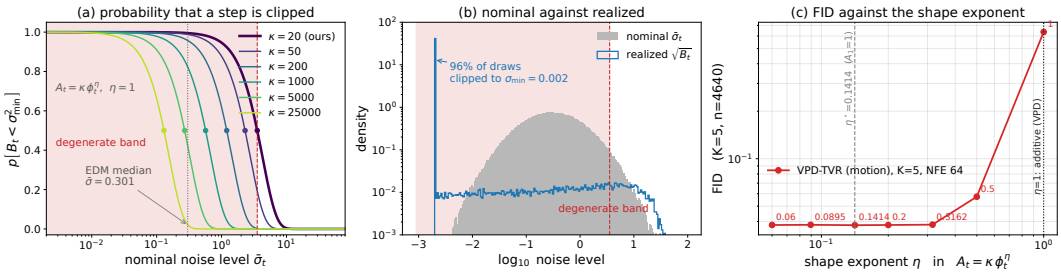

Figure 2: $\sigma \in [0.002, 80]$, $A_t = \kappa\,\phi_t^\eta$. **(a)** Clipping probability against nominal level, $\eta = 1$; pink marks where the median draw is clipped at $\kappa = 20$. **(b)** Nominal against realized training levels, $\eta = 1, \kappa = 20$: 96% collapse onto $\sigma_{\min}$. **(c)** Motion in-betweening FID against η, where $\eta = 1$ is the shared-rate construction.

3 Where The Shared Rate Formulation Fails

The construction of § 2 is exact and its objective is in closed form, yet training on the sampled values collapses. This section shows that the collapse is a structural consequence of the shared rate that Gamma additivity requires. Proposition 3.1 pins A_1/A_T to the endpoint ratio, and Corollary 3.2 turns this into a dichotomy: either $A_1 \leq 1$ and the low-noise regime degrades, or $A_1 > 1$ and the path is nearly deterministic.

3.1 The Endpoint Bound

Matching the noise schedule in expectation fixes the shapes at $A_t = b\bar{\sigma}_t^2$ (§ 2) and the ratio of the two endpoints eliminates the rate parameter.

Proposition 3.1 (Endpoint bound)**.** *For the shared rate construction of § 2, and for any noise schedule $\bar{\sigma}_{1:T}$ with any number of steps T, and any stochasticity scale κ, $\frac{A_1}{A_T} = \phi_1 = \frac{\sigma_{min}^2}{\sigma_{max}^2}$.*

The dynamic range of the shape parameter is therefore fixed by the two endpoints alone. Increasing κ scales A_1 and A_T together and leaves the ratio unchanged and changing the shape of the schedule moves only the intermediate values whereas changing T moves nothing at all. The following Corollary tells what it would take to keep $A_1 > 1$.

Corollary 3.2 (**Antinomy of the shared rate**)**.** *If $A_1 > 1$ then $b > \sigma_{min}^{-2}$ and hence $\kappa = A_T = b\sigma_{max}^2 > \frac{1}{\phi_1}$,* $\mathrm{RSD}[B_T] = \kappa^{-1/2} < \sqrt{\phi_1} = \frac{\sigma_{\min}}{\sigma_{\max}}$.

For the same range this requires $\kappa > 1.6 \times 10^9$ and forces $\mathrm{RSD}[B_T] < 2.5 \times 10^{-5}$. Under a shared rate parameter one cannot have both a well behaved low noise regime and a path that is stochastic in any meaningful sense. The parameter κ trades one against the other but cannot supply both.

3.2 The Degeneracy Is a Band, Not a Tail

The realized variance is clipped at $\sigma_{\min}^2$ for numerical reasons, and Proposition 3.1 makes clipping common rather than exceptional. Writing $z = b\sigma_{\min}^2$, we can determine the clipping probability in closed form, $P(B_t < \sigma_{\min}^2) = \frac{\gamma(A_t, z)}{\Gamma(A_t)} \approx \frac{z^{A_t}}{\Gamma(A_t+1)}$, with $(z \ll 1)$ where γ is the lower incomplete Gamma function, Γ is Gamma function and $A_t = b\bar{\sigma}_t^2$. Since A_t enters as an exponent, the probability tends to one whenever $A_t \to 0$. Figure 2 shows the consequence. Clipping is not confined to a neighborhood of $\sigma_{\min}$ but covers the entire low noise half of the schedule, and marginalizing over the diffusion time as a training batch does, 96% of draws collapse onto a single atom as shown in Fig. 2(b). What the denoiser sees is an atom, a nearly empty band extending to $\bar{\sigma} \approx 3.5$, and a surviving region above it. All these phenomena share a single cause, $A_t = b\bar{\sigma}_t^2$, where the shape is the schedule in different units with its range fixed.

4 GENERALIZED VPD: TIME-VARYING RATE

The trade-off of § 3 stems from a single choice, fixing the rate. Freeing it breaks Gamma additivity, since a sum of Gamma increments with different rates is no longer Gamma. Therefore, we replace the sum by a product: Gamma is closed under multiplication by an independent Beta with matching parameters. The resulting construction contains § 2 as the special case $\eta = 1$ (§ 4.4).

4.1 THE MULTIPLICATIVE CONSTRUCTION

Definition 4.1 (Beta thinning path). Let $\{A_t\}_{t=0}^T$ be strictly increasing and $\{b_t\}_{t=0}^T$ positive, and set $\zeta_t = b_t/b_{t-1}$. Draw $B_T \sim \mathrm{Ga}(A_T, b_T), V_t \sim \mathrm{Beta}(A_{t-1}, A_t - A_{t-1})$ with $\{B_T\} \cup \{V_t\}$ mutually independent, and define $B_{t-1} = \zeta_t V_t B_t$.

We refer to this as the time varying rate (TVR) construction. Note that the first Beta parameter is the cumulative shape and the second is the increment, the reverse of the convention for w in § 2.3.

Proposition 4.2 (Exact marginals). *Under Definition 4.1, $B_t \sim \mathrm{Ga}(A_t, b_t)$ for every t.*

Setting $b_t = A_t/\bar{\sigma}_t^2$ therefore gives $\mathbb{E}[B_t] = \bar{\sigma}_t^2$ exactly, while the relative fluctuation $\mathrm{RSD}[B_t] = A_t^{-1/2}$ is now determined by A_t alone and is decoupled from the schedule. The endpoint bound of Proposition 3.1 does not apply, because it was a statement about $A_t = b\bar{\sigma}_t^2$ under a constant b.

Proposition 4.3 (Monotonicity). *$B_{t-1} \le B_t$ almost surely iff $\zeta_t \le 1$, equivalently $\frac{A_t}{A_{t-1}} \le \frac{\bar{\sigma}_t^2}{\bar{\sigma}_{t-1}^2}$.*

The shape may grow with t, but no faster in relative terms than the variance itself. Monotonicity is required for the Brownian bridge representation of § 2.3 and for the reverse kernel.

4.2 EXACT COARSENING

Composing k steps of Definition 4.1 gives the following proposition.

Proposition 4.4 (Exact coarsening). *For any $k \ge 1$, $B_{t-k} = \left(\prod_{s=t-k+1}^t \zeta_s\right)\left(\prod_{s=t-k+1}^t V_s\right) B_t$ where $\prod_{s=t-k+1}^t \zeta_s = \frac{b_t}{b_{t-k}}$, $\quad \prod_{s=t-k+1}^t V_s \sim \mathrm{Beta}(A_{t-k}, A_t - A_{t-k})$.*

This is the multiplicative counterpart of the role played by Gamma additivity in § 2 and unlike additivity it imposes no constraint on rate parameter.

4.3 REVERSE KERNEL

Define the effective removed fraction $\tilde{w}_t = 1 - \zeta_t V_t$ so that $B_{t-1} = (1 - \tilde{w}_t)B_t$ as in § 2.3. Writing $\mu = \mathbb{E}[V_t] = A_{t-1}/A_t$ and substituting $\zeta_t = (A_t\bar{\sigma}_{t-1}^2)/(A_{t-1}\bar{\sigma}_t^2)$, the shape ratios cancel between the two factors and $\mathbb{E}[\tilde{w}_t] = 1 - \zeta_t\mu = 1 - \frac{\bar{\sigma}_{t-1}^2}{\bar{\sigma}_t^2}$. The mean fraction of variance removed at each step is therefore the nominal schedule itself, independent of the shape profile. Equivalently, this is forced by $\mathbb{E}[B_t] = \bar{\sigma}_t^2$ together with $\tilde{w}_t \perp B_t$. The deterministic factor ζ_t carries the schedule while the Beta factor contributes only multiplicative jitter about it, in contrast to § 2 where $\zeta_t = 1$ and the Beta factor had to produce the entire reduction on its own.

4.4 A GENERALIZED FORM OF SHARED RATE CONSTRUCTION

A convenient parameterization takes the shape profile to be a power of the schedule $A_t = \kappa\phi_t^\eta$ where $\eta \in [0, 1]$ and $\kappa := A_T$ with $\phi_t = \bar{\sigma}_t^2/\bar{\sigma}_T^2$ as in § 2.2. First, *monotonicity is automatic.* Proposition 4.3 requires $A_t/A_{t-1} \le \bar{\sigma}_t^2/\bar{\sigma}_{t-1}^2$, and substituting gives $(\bar{\sigma}_t^2/\bar{\sigma}_{t-1}^2)^\eta \le \bar{\sigma}_t^2/\bar{\sigma}_{t-1}^2$, which holds for every t whenever $\eta \le 1$. Second, *the shared rate construction is the endpoint* $\eta = 1$. There $A_t = \kappa\phi_t = b\bar{\sigma}_t^2$ with $b = \kappa/\bar{\sigma}_T^2$ constant, so $\zeta_t = 1$, and $1 - V_t$ has the law of w from § 2.3. Definition 4.1 then reduces exactly to the construction of § 2.2, and Proposition 3.1 applies. The VPD training collapse is shown in Fig. 2(c). Third, *the dynamic range is under direct control.* Since $A_1/\kappa = \phi_1^\eta$, the ratio that Proposition 3.1 pinned to ϕ_1 becomes ϕ_1^η, and requiring $A_1 > 1$ gives the design inequality $\eta < \frac{\log\kappa}{L}$, where $L := \log(1/\phi_1)$. The analysis of § 2-4 establishes

Table 1: Inpainting performance compared with SoTAs on the FFHQ, LSUN Bedroom, and ImageNet. All baseline values are from RAD Kim et al. (2025). We mark **Bold** for the best, and underline for the second best performance.

Method	FFHQ						LSUN Bedroom						ImageNet					
	Box		Extreme		Wide		Box		Extreme		Wide		Box		Extreme		Wide	
	FID↓	LPIPS↓	FID↓	LPIPS↓	FID↓	LPIPS↓	FID↓	LPIPS↓	FID↓	LPIPS↓	FID↓	LPIPS↓	FID↓	LPIPS↓	FID↓	LPIPS↓	FID↓	LPIPS↓
LaMa Suvorov et al. (2022)	27.7	0.086	61.7	0.492	23.2	0.096	-	-	-	-	-	-	-	-	-	-	-	-
Score-SDE Song et al. (2021)	30.3	0.135	48.6	0.488	29.8	0.132	23.7	0.648	24.1	0.648	23.2	0.644	57.2	0.200	86.6	0.495	62.0	0.183
DDRM Kawar et al. (2022)	28.4	0.109	48.1	0.532	27.5	0.113	20.5	0.166	33.1	0.450	26.4	0.190	74.3	0.224	106.9	0.492	75.1	0.211
RePAINT Lugmayr et al. (2022)	25.7	0.093	35.9	0.398	24.2	0.108	20.5	0.176	23.5	0.461	21.4	0.161	54.0	0.177	84.0	0.479	59.0	0.166
MCG Chung et al. (2022)	23.7	0.089	30.6	0.366	22.1	0.099	19.9	0.131	22.0	0.395	20.9	0.108	48.1	0.132	58.4	0.448	56.9	0.124
DDNM Wang et al. (2022)	30.4	0.089	87.7	0.353	30.4	0.089	22.7	0.150	53.3	0.431	23.2	0.126	63.8	0.187	80.6	0.476	64.5	0.167
DeqIR Cao et al. (2024)	24.2	0.093	64.2	0.368	27.4	0.099	22.2	0.176	43.9	0.461	22.0	0.153	66.5	0.195	99.5	0.505	68.4	0.182
RAD Kim et al. (2025)	22.1	0.074	33.4	0.317	**21.5**	0.078	19.2	0.131	21.6	0.399	20.8	0.107	47.0	**0.118**	57.8	**0.374**	56.7	0.104
VPD (ours)	**21.6**	**0.066**	**26.7**	**0.295**	21.7	**0.073**	**16.0**	**0.114**	**17.1**	**0.371**	**15.8**	**0.099**	**41.4**	**0.118**	**51.4**	0.380	**38.5**	**0.097**

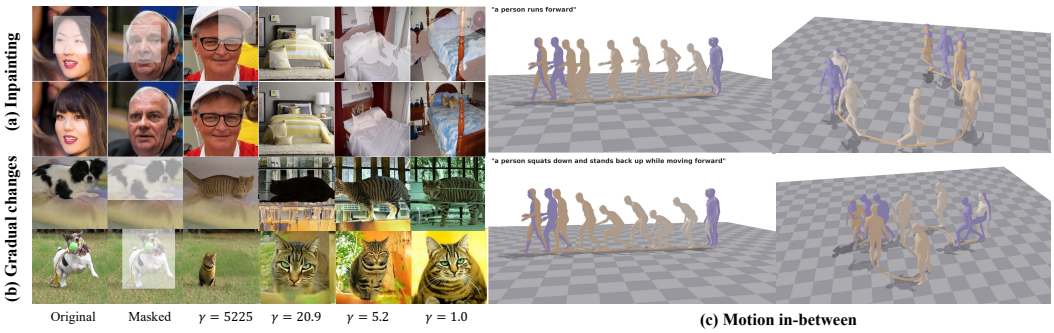

Figure 3: **(a)** Inpainting on FFHQ and LSUN Bedroom; **top**: masked input, **bottom**: result. **(b)** The γ dial on one model: large γ preserves the unmasked region, smaller γ regenerates it progressively. **(c)** Motion in-betweening. Purple: keyframes, orange: generated. **Left**: same keyframes, different prompts (run, squat); **right**: long trajectories. More in Appendix H.

what a per-region variance schedule must satisfy. We therefore use the law for its structure, not its randomness: the construction below takes the mean path and places the spatial contrasts in the rate.

5 Per-Region Variance and Inpainting

Everything so far has treated the variance as a single scalar shared by all coordinates. Once the variance may differ across regions, with a per-pixel level as the limiting case, one formulation covers tasks usually handled separately: inpainting, local editing, and generation. Several recent methods parameterize a pixel-dependent noise level for this reason Kim et al. (2025); Mayet et al. (2025); Papa et al. (2026). This section places that allocation in the framework of § 2-4, fixing the total and specifying how it is split between regions. There are two ways to allocate. One gives each region its own shape under a shared rate, which yields a spatial analogue of the duality of § C.3 but it encounters the difficulty of § 3 while training (Appendix G). We take the other, which leaves the shape untouched. This section considers two regions and leaves general partitions to future work.

5.1 The construction

Let M be a binary mask representing the region to be edited, define $\mathcal{G}_1 = M$ and $\mathcal{G}_2 = M^c$, and write $f := |M|/d$ for the area fraction. Nothing depends on the shape, connectivity, or size of M, so an arbitrary free-form region is admissible. The partition is chosen at sampling time, since the denoiser receives a per-pixel level and has no representation of regions.

Draw a single scalar TVR path $\tilde{B}_{1:T}$ from Definition 4.1, fix a deterministic pair (γ_1, γ_2), and set $B_t^{(g)} = \gamma_g^2 \tilde{B}_t$ with $g = 1, 2$, subject to the pixel weighted pin

$$f\gamma_1^2 + (1-f)\gamma_2^2 = 1. \tag{3}$$

which states that the pixel averaged variance still follows the nominal schedule, $\sum_g (d_g/d)\mathbb{E}[B_t^{(g)}] = \bar{\sigma}_t^2$ where d_g is the size of region g. As in § C.3, the total is fixed and the allocation is free. This is the

Table 2: Quantitative results of motion in-betweening under the CondMDI evaluation protocol on HumanML3D with N=4640 samples, all methods are re-evaluated. K is the number of keyframes randomly placed along the motion; keyframe error is the mean 3D position error over all 22 joints at the keyframes. OmniControl receives the root-joint of the K keyframes instead of full joints.

Method	Conditioning	NFE	FID↓	R-precision (Top-3)↑	Diversity →	Foot Skate Ratio ↓	kf-err (root) ↓	kf-err (all)
Ground Truth	–	–	–	0.771	9.36	0.072	–	–
MDM + imputation Tevet et al. (2022)	$K = 1$	1000	0.865	0.700	8.47	0.087	0.800	0.808
OmniControl (root-joint condition) Xie et al. (2024)	$K = 1$	1000	1.191	0.721	8.86	**0.067**	**0.016**	0.166
CondMDI (Cohan et al., 2024)	$K = 1$	1000	0.231	0.741	9.66	0.089	0.159	0.161
ACMDM + imputation Meng et al. (2025)	$K = 1$	64	0.483	0.744	9.46	0.093	0.116	0.134
VPD (ours)	$K = 1$	64	**0.106**	**0.773**	9.37	0.080	0.042	**0.057**
MDM + imputation Tevet et al. (2022)	$K = 5$	1000	1.733	0.730	8.26	0.102	0.621	0.627
OmniControl (root-joint condition) Xie et al. (2024)	$K = 5$	1000	1.402	0.747	9.23	**0.070**	0.051	0.165
CondMDI (Cohan et al., 2024)	$K = 5$	1000	0.083	0.752	9.56	0.091	0.070	0.072
ACMDM + imputation Meng et al. (2025)	$K = 5$	64	1.338	0.741	8.91	0.129	0.060	0.074
VPD (ours)	$K = 5$	64	**0.038**	**0.779**	9.51	0.085	**0.022**	**0.032**
MDM + imputation Tevet et al. (2022)	$K = 20$	1000	0.479	0.773	8.93	0.091	0.255	0.260
OmniControl (root-joint condition) Xie et al. (2024)	$K = 20$	1000	0.933	0.758	8.89	**0.072**	0.059	0.152
CondMDI (Cohan et al., 2024)	$K = 20$	1000	**0.022**	0.765	8.80	0.090	0.035	0.036
ACMDM + imputation Meng et al. (2025)	$K = 20$	64	0.073	**0.781**	9.00	0.091	**0.011**	**0.017**
VPD (ours)	$K = 20$	64	0.037	0.773	9.24	0.082	**0.011**	**0.017**

conceptual analogy of Gamma-Dirichlet duality. Unlike the duality, the allocation is deterministic, whereas the counterpart is in Appendix G. Here γ_g^2 is a multiplier on the nominal level, so the pin constrains only its pixel weighted average and an individual region may exceed one. Parameterizing by the level ratio between the two regions $\gamma := \gamma_1/\gamma_2$, the pin has the unique solution $\gamma_2^2 = \frac{1}{f\gamma^2+1-f}$ and $\gamma_1^2 = \gamma^2\gamma_2^2$. A single parameter therefore moves the construction across three downstream tasks: generation at $\gamma = 1$, where two regions are assigned the same variance budget; local editing at moderate γ, where the surroundings are partially regenerated; and inpainting with $\gamma \gg 1$, where $\gamma_2 \to 0$. Multiplying a Gamma variate by a constant leaves the shape and divides the rate, $B_t^{(g)} \sim \mathrm{Ga}(A_t, b_t/\gamma_g^2)$, so the level ratio never enters the shape profile and the condition of § 3 is imposed once on the scalar path. Because γ_g doesn't depend on t, it cancels in $B_{t-1}^{(g)}/B_t^{(g)}$, so the removed fraction $\tilde{w}_t$ is common to both regions and monotonicity, exact coarsening, and Theorem C.2 carry over unchanged.

5.2 Training

At every training step we draw a fresh mask and level ratio. The ratio is $\gamma = 1$ with probability p_{uni}, $\gamma = \gamma_{\text{cap}}$ with probability p_{cor}, and otherwise log-uniform on $[1, \gamma_{\text{cap}}]$, where $\gamma_{\text{cap}}(\bar{f}, \bar{\sigma}_t)$ is the largest ratio at which the low region stays above $\sigma_{\min}$ and the high region below $\tau\sigma_{\max}$ (Proposition F.2). The model still trains to learn unconditional generation with $\gamma = 1$, while $\gamma = \gamma_{\text{cap}}$ is the configuration specifically for the inpainting task Kim et al. (2025); Mayet et al. (2025). For images, M is drawn from rectangles, free-form blobs, and strokes following LaMa, with area log-uniform on $[f_{\text{lo}}, f_{\text{hi}}]$ and complemented with probability p_{inv}. For motion, M is a temporal mask drawn with equal probability from four masks: a contiguous span covering 20-80% of frames, one to five random frames, a prefix of 10-50%, or a regular grid of period $\mathrm{Unif}\{2, \cdots, 10\}$ at random phases. For sampling, no modification of sampler is required. Note that there is no replacement step, the known coordinates are never pasted back, and *preservation is carried by the path* (Appendix F.4).

6 Related Works

Per-region noise schedule A rapidly growing body of work relaxes the homogeneous noise schedule of standard diffusion process. In images, per-pixel or per-patch noise levels have been introduced independently for spatially variant denoising Pearl et al. (2023), mask-based inpainting Kim et al. (2025); Mayet et al. (2025), inference-time editing control Levin & Fried (2025), training-efficient patch-level noise Schusterbauer et al. (2026), and fully asynchronous per-token noise as a joint diffusion Papa et al. (2026). Along the temporal axis, TEDi Zhang et al. (2024) maps motion sequence time to diffusion time, and Diffusion Forcing Chen et al. (2024) generalizes this to independent per-frame levels. Each of these designs its schedule by hand and each reports, empirically, a constraint it had to impose: a normalization on the total corruption Kim et al. (2025); Papa et al. (2026), or a cap on the disparity between regions Hu et al. (2026). None derives these from a law of the schedule itself. Section 5 recovers each of them as a consequence of the construction of § 2-§ 4.

Image inpainting and motion in-betweening Both generate from a partial observation, spatial or temporal, that must be preserved. Training-free inpainting imposes the observation at every step by replacement, projection, or guidance (Lugmayr et al., 2022; Wang et al., 2022; Kawar et al., 2022; Chung et al., 2022; 2024), while training-based methods hold the observed region at timestep zero and finetune (Kim et al., 2025; Mayet et al., 2025). Motion in-betweening follows the same split, by imputation (Tevet et al., 2022) or by keyframe conditioning (Cohan et al., 2024; Xie et al., 2024) (extended discussion in Appendix A).

7 EXPERIMENTS

We evaluate VPD on FFHQ Karras et al. (2019), LSUN Bedroom Yu et al. (2016), and ImageNet Deng et al. (2009) for inpainting task with 256×256 size images and HumanML3D Guo et al. (2022) for motion in-betweening task.

7.1 IMPLEMENTATION DETAILS

We finetune the public guided-diffusion ADM models at 256×256 (FFHQ, LSUN Bedroom, ImageNet) Dhariwal & Nichol (2021). Two sets of parameters are trained: (i) a per-pixel conditioning adapter that mirrors the model's own scalar noise embedding and is initialized so that a spatially uniform noise map reproduces the frozen model exactly, and (ii) LoRA factors of rank 16 Hu et al. (2021) on every convolutional and linear weight of the model. We use Adam with learning rate 10^{-4}, $\beta = (0.9, 0.999)$, no weight decay, 1K warmup steps, gradient clipping at 1.0, and an EMA of 0.9999. The batch size is 32 for FFHQ and 16 for LSUN Bedroom and ImageNet. Following the RAD protocol, we train for 300K steps on FFHQ and LSUN Bedroom and 800K steps on ImageNet.

For motion in-betweening, we finetune the pretrained ACMDM Meng et al. (2025) on HumanML3D with LoRA of rank 16 for 100K steps (Adam, learning rate 10^{-4}, batch size 64). Keyframe regions are drawn from four masks (segments, random keyframes, prefixes, regular sparse pins) and the same γ sampler is used with $\kappa = 20$ and $\eta = 0.06$.

Evaluation metrics. For image inpainting, we follow RAD protocol with Fréchet Inception Distance (FID) and LPIPS, over 1000 held-out images per mask type. We use three masks: box, wide strokes, and extreme. For motion in-betweening, we follow the CondMDI protocol on the HumanML3D test split, and report FID, R-precision (top-3), Diversity, foot skate ratio, and keyframe error in meters.

7.2 RESULTS

Image Inpainting. We follow the protocol of RAD, finetuning the same pretrained ADM checkpoint with LoRA. Table 1 reports FID and LPIPS on FFHQ, LSUN Bedroom, and ImageNet. VPD is best in 15 of 18 settings. Fig. 3 (b) sweeps γ: the region outside the mask is progressively regenerated as γ decreases, until at $\gamma = 1$ the image is fully generated.

Motion In-betweening. We follow the CondMDI protocol, conditioning on K randomly placed keyframes. Since ACMDM diffuses in a latent that spans four frames, one pinned latent fixes four consecutive frames. All baselines receive the same four-frame spans (single-frame results in Tab. 8).[2] Tab. 2 shows that VPD reaches the lowest keyframe error at every K, by 2-4 times over CondMDI, while other metrics remain comparable. Fig. 3 (c) shows that the in-between motion follows the text condition while satisfying both end keyframes. For more results, please check App. H.

Ablation study. We ablate three choices: the training distribution of § 5.2, the shape exponent η of § 4.4, and the stochasticity scale κ. *Training distribution.* Table 3 varies the training distribution. Removing the endpoint probability ($p_{\text{cor}} = 0$), or training with the reimplemented AsyncPatch sampler, which likewise has no endpoint, costs 6.9 and 3.6 FID on FFHQ extreme masks, while

[2] ACMDM diffuses in a latent with a temporal stride of four, so two keyframes that fall in the same four-frame window pin the same latent. With $K = 20$, the keyframes occupy 15.6 distinct latents on average. All methods receive the same keyframes, but our model sees them at the resolution of its latent.

Table 3: Training-distribution ablation, FID. Gen.: 50K unconditional samples at NFE 128. †Our reimplementation on the two-region map.

Data	Model	Gen.	Box	Extreme	Wide
FFHQ	ADM baseline	13.9	–	–	–
	VPD ($p_{\text{cor}} = 0.1$)	15.6	21.6	26.7	21.7
	$p_{\text{cor}} = 0$	12.0	23.7	33.6	22.4
	$p_{\text{cor}} = 1$	13.5	21.5	28.2	21.7
	Perlin masks	13.0	21.5	26.9	21.6
	AsyncPatch†	11.9	22.2	30.3	22.0
LSUN	ADM baseline	5.7	–	–	–
	VPD ($p_{\text{cor}} = 0.1$)	3.0	16.0	17.1	15.8
	$p_{\text{cor}} = 1$	8.9	16.0	16.6	15.9
ImageNet	ADM baseline	26.4	–	–	–
	VPD ($p_{\text{cor}} = 0.1$)	33.2	41.4	51.4	38.5
	$p_{\text{cor}} = 1$	48.6	41.6	54.3	39.7

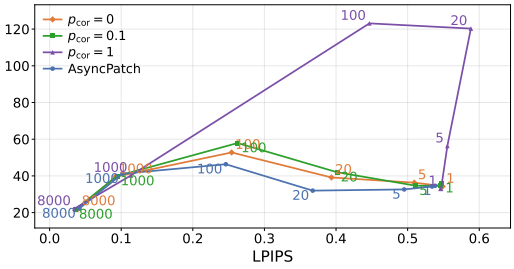

Figure 4: FID vs. LPIPS to source as γ decreases from 8000 to 1 on FFHQ.

RAD's Perlin masks change no inpainting number. Therefore the gain does not come from the mask family. Training only at the endpoint ($p_{\text{cor}} = 1$) matches the full mixture on box and wide masks, but its FID triples at intermediate γ (Fig. 4) and, on LSUN and ImageNet, it generates worse by 5.9 and 15.4 FID. The continuum is therefore what makes editing possible and, on two of three datasets, what preserves generation. *Shape exponent.* Fig. 2(c) sweeps η on motion in-betweening. The sufficient condition $A_1 > 1$ corresponds to $\eta \leq 0.141$, yet FID stays flat to $\eta = 0.316$ ($A_1 \approx 0.025$) and breaks at $\eta = 0.5$. What tracks the break is the fraction of realized levels clamped at σ_{min}:3% at $\eta = 0.316$, 21% at 0.5 and 75% at 1. The denoiser tolerates a small share of floored levels. $A_1 > 1$ is a condition under which the construction is guaranteed to be well behaved. *Stochasticity.* Appendix Table 4 varies κ at sampling time on one checkpoint trained at $\kappa = 5$. Across κ values, FID varies by less than 0.1 at NFE 64 and 0.2 at NFE 4. The randomness of the path does not affect quality, as Proposition C.5 anticipates. We further discuss what the stochastic formulation contributes at § 8.

8 Conclusion and Discussion

Why a random variance The construction of § 5 is deterministic, and the randomness of the path is inconsequential. What the probabilistic framing supplied is a set of statements that a deterministic schedule cannot make: that the total budget and its allocation separate (§ C.3), that the allocation is pinned (§ 5.1), that the contrast belongs in the rate (Appendix G), and that the training distribution must cover both the endpoint inference occupies and the continuum between it and homogeneous training (§ 7.2). RAD and TD-Paint train only the former and AsyncPatch only the latter. Table 3 reproduces each choice within our model and shows what it loses. AsyncPatch and RAD reached the first two empirically, by observing that unconstrained per-token timesteps fail and fixing the mean; AsynDM shows the last by its absence where training-free schedule cannot reach the endpoint and must limit the timestep difference between regions. The third has no empirical counterpart, since it concerns a quantity, the shape of the variance law, that a deterministic schedule does not possess.

Limitations The allocation of § 5 is specified rather than learned by given masks. The framework separates the total budget from its allocation, but the temporal allocation carries no information the data can use (§ C.3) and the magnitude of the path's own randomness is bounded (Prop. C.5). The axis that remains open is the spatial direction, where the allocation is informative but must be supplied by conditions (Theorem C.4). Learning it from a conditioning signal as $p_\phi(\gamma_g|c)$ for a sketch, or a prompt, requires no change to the construction, since γ_g may depend on c as long as it does not depend on t, which is the natural next step. A second limitation is that preservation is approximate where replacement methods are exact.

Conclusion We promoted the variance of a diffusion process to a random variable and asked what follows. The total budget separates from its allocation, the natural shared-rate construction fails, a time-varying rate removes the failure while keeping every closed-form property, and the randomness itself is bounded and, empirically, does not affect quality. What remains is a single per-region construction, a scalar path rescaled under a pin, that one model uses for inpainting, local editing, and motion in-betweening. Prior per-region schedules found pieces of this structure one failure at a time, while here it is derived.

9 AI DISCLOSURE

In this work, we used generative AI tools to help check mathematical claims, provide critical ingredients for proving, assist in writing and verifying proofs, provide feedback on the design of experiments and ablations, implement methods, and finding related works. We did not use generative AI tools to generate synthetic datasets, develop the conceptual framework, or clean or reformat datasets; translation and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to edit parts of the paper for readability and grammar. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-DPM: An Analytic Estimate of the Optimal Reverse Variance in Diffusion Probabilistic Models, May 2022. URL http://arxiv.org/abs/2201.06503.

Jiezhang Cao, Yue Shi, Kai Zhang, Yulun Zhang, Radu Timofte, and Luc Van Gool. Deep Equilibrium Diffusion Restoration with Parallel Sampling. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2824–2834, June 2024. doi: 10.1109/CVPR52733.2024.00273. URL https://ieeexplore.ieee.org/abstract/document/10658596.

Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion Forcing: Next-token Prediction Meets Full-Sequence Diffusion, December 2024. URL http://arxiv.org/abs/2407.01392.

Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving Diffusion Models for Inverse Problems using Manifold Constraints. In *Advances in Neural Information Processing Systems*, volume 35, pp. 25683–25696. Curran Associates, Inc., 2022. doi: 10.52202/068431-1862. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/a48e5877c7bf86a513950ab23b360498-Abstract-Conference.html.

Hyungjin Chung, Jeongsol Kim, Michael T. Mccann, Marc L. Klasky, and Jong Chul Ye. Diffusion Posterior Sampling for General Noisy Inverse Problems, May 2024. URL http://arxiv.org/abs/2209.14687. arXiv:2209.14687 [stat].

Setareh Cohan, Guy Tevet, Daniele Reda, Xue Bin Peng, and Michiel van de Panne. Flexible Motion In-betweening with Diffusion Models, May 2024. URL http://arxiv.org/abs/2405.11126.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848. URL https://ieeexplore.ieee.org/abstract/document/5206848.

Prafulla Dhariwal and Alexander Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating Diverse and Natural 3D Human Motions From Text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5152–5161, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Guo_Generating_Diverse_and_Natural_3D_Human_Motions_From_Text_CVPR_2022_paper.html.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models, October 2021. URL http://arxiv.org/abs/2106.09685.

Zijing Hu, Yunze Tong, Fengda Zhang, Junkun Yuan, Jun Xiao, and Kun Kuang. Asynchronous Denoising Diffusion Models for Aligning Text-to-Image Generation, February 2026. URL http://arxiv.org/abs/2510.04504.

Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4396–4405, June 2019. doi: 10.1109/CVPR.2019.00453. URL https://ieeexplore.ieee.org/abstract/document/8953766.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the Design Space of Diffusion-Based Generative Models. *Advances in Neural Information Processing Systems*, 35:26565–26577, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/a98846e9d9cc01cfb87eb694d946ce6b-Abstract-Conference.html.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising Diffusion Restoration Models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23593–23606. Curran Associates, Inc., 2022. doi: 10.52202/068431-1714. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/95504595b6169131b6ed6cd72eb05616-Abstract-Conference.html.

Sora Kim, Sungho Suh, and Minsik Lee. RAD: Region-Aware Diffusion Models for Image Inpainting. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2439–2448, June 2025. doi: 10.1109/CVPR52734.2025.00233. URL https://ieeexplore.ieee.org/abstract/document/11095134.

Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational Diffusion Models, April 2023. URL http://arxiv.org/abs/2107.00630.

Eran Levin and Ohad Fried. Differential Diffusion: Giving Each Pixel Its Strength. *Computer Graphics Forum*, 44(2):e70040, 2025. ISSN 1467-8659. doi: 10.1111/cgf.70040. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.70040.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using Denoising Diffusion Probabilistic Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11451–11461, June 2022. doi: 10.1109/CVPR52688.2022.01117. URL https://ieeexplore.ieee.org/abstract/document/9880056.

Eugene Lukacs. A Characterization of the Gamma Distribution. *The Annals of Mathematical Statistics*, 26:6, June 1955. URL https://www.jstor.org/stable/2236887?casa_token=OLxiRfy0XvcAAAAA%3AX7iLmg9_oZKtwxFMTrkEHneXBSnMDfQZ6GsqE_9fJEwTSV-LdVFs1GuXRKhJltWFBEDX6axJ95afD3SuA3i-egazDXzzhVgmnzMxRv6EQKi-ldIj7-6I&seq=1.

Tsiry Mayet, Pourya Shamsolmoali, Simon Bernard, Eric Granger, Romain Hérault, and Clement Chatelain. TD-Paint: Faster Diffusion Inpainting Through Time-Aware Pixel Conditioning. *International Conference on Learning Representations*, 2025:102537–102565, May 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/fe489a28a54583ee802b8e2955c024c2-Abstract-Conference.html.

Zichong Meng, Zeyu Han, Xiaogang Peng, Yiming Xie, and Huaizu Jiang. Absolute Coordinates Make Motion Generation Easy, May 2025. URL http://arxiv.org/abs/2505.19377.

Kushagra Pandey, Jaideep Pathak, Yilun Xu, Stephan Mandt, Michael Pritchard, Arash Vahdat, and Morteza Mardani. Heavy-Tailed Diffusion Models, October 2024. URL http://arxiv.org/abs/2410.14171.

Samuele Papa, Valentin De Bortoli, Guillaume Couairon, Daniel Sýkora, Romuald Elie, and Klaus Greff. AsyncPatch Diffusion: Spatially-flexible image generation, June 2026. URL http://arxiv.org/abs/2606.07079.

Naama Pearl, Yaron Brodsky, Dana Berman, Assaf Zomet, Alex Rav Acha, Daniel Cohen-Or, and Dani Lischinski. SVNR: Spatially-variant Noise Removal with Denoising Diffusion, June 2023. URL http://arxiv.org/abs/2306.16052.

Johannes Schusterbauer, Ming Gui, Yusong Li, Pingchuan Ma, Felix Krause, and Björn Ommer. Denoising, Fast and Slow: Difficulty-Aware Adaptive Sampling for Image Generation, April 2026. URL http://arxiv.org/abs/2604.19141.

Dario Shariatian, Umut Simsekli, and Alain Oliviero Durmus. Denoising Levy Probabilistic Models. *International Conference on Learning Representations*, 2025:96991–97024, May 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f12c9da400fec2594a05e37b0507fd00-Abstract-Conference.html.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations, February 2021. URL http://arxiv.org/abs/2011.13456.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust Large Mask Inpainting with Fourier Convolutions. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3172–3182, January 2022. doi: 10.1109/WACV51458.2022.00323. URL https://ieeexplore.ieee.org/document/9707077.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human Motion Diffusion Model, October 2022. URL http://arxiv.org/abs/2209.14916.

Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-Shot Image Restoration Using Denoising Diffusion Null-Space Model, December 2022. URL http://arxiv.org/abs/2212.00490.

Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. OmniControl: Control Any Joint at Any Time for Human Motion Generation, April 2024. URL http://arxiv.org/abs/2310.08580.

EUN BI YOON, Keehun Park, Sungwoong Kim, and Sungbin Lim. Score-based Generative Models with Lévy Processes. In *Advances in Neural Information Processing Systems*, volume 36, pp. 40694–40707. Curran Associates, Inc., 2023. doi: 10.52202/075280-1772. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8011b23e1dc3f57e1b6211ccad498919-Abstract-Conference.html.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop, June 2016. URL http://arxiv.org/abs/1506.03365.

Zihan Zhang, Richard Liu, Rana Hanocka, and Kfir Aberman. TEDi: Temporally-Entangled Diffusion for Long-Term Motion Synthesis. In *ACM SIGGRAPH 2024 Conference Papers*, SIGGRAPH '24, pp. 1–11, New York, NY, USA, 2024. Association for Computing Machinery. doi: 10.1145/3641519.3657515. URL https://dl.acm.org/doi/10.1145/3641519.3657515.

Appendix Table of Contents

CONTENTS

A Extended Related Works **15**

B The Gamma–Beta Relation **15**

C Further discussion of Variance Path Diffusion construction (§ 2) **16**

C.1 Detailed derivation of reverse kernel. 16

C.2 Variance posterior 17

C.3 Path posterior: Gamma-Dirichlet duality 17

C.4 The reverse covariance 18

C.5 Score, denoiser, and the training objective 19

C.6 Exact ELBO 20

C.7 What follows with $A_1 > 1$ 22

C.8 Committed paths 23

D Proofs of theorem, proposition, and corollary **23**

D.1 Proof of Proposition 3.1 23

D.2 Proof of Corollary 3.2 23

D.3 Proof of Proposition 4.2 23

D.4 Proof of Proposition 4.3 23

D.5 Proof of Proposition 4.4 24

E The Variance-Preserving formulation **24**

E.1 The state map 25

E.2 The increment structure transfers 25

E.3 The reverse kernel 26

F Further discussion of per-region extension §5 **26**

F.1 Details of Eq. 3 26

F.2 Invariance of the path properties under region-wise rescaling 26

F.3 Saturation and the start level 27

F.4 Sampling details 27

G Shape allocation instead of the rate **27**

G.1 Construction 28

G.2 Why it is not adopted 28

H More results **28**

H.1 Hyperparameters 28
H.2 Motion 29
H.3 Image 29

A Extended Related Works

Per-region noise schedule A rapidly growing body of work relaxes the homogeneous noise schedule of the standard diffusion process. In images, per-pixel or per-patch levels have been introduced independently for spatially variant denoising, where SVNR Pearl et al. (2023) gives each pixel its own time embedding inferred from the input noise. For inpainting, RAD Kim et al. (2025) and TD-Paint Mayet et al. (2025) hold the observed region at timestep zero and train on Perlin or inpainting masks. For inference-time editing control, Differential Diffusion Levin & Fried (2025) assigns each pixel its own edit strength via a user provided change map, so that regions re-entere the generative process at different noise levels, while AsynDM Hu et al. (2026) denoises prompt-relevant regions, identified from cross-attention at each step, on a slower timestep schedule thand the rest of the image to improve text-image alignment. Similarly, AsyncPatch Papa et al. (2026) assigns every token an independent diffusion process and trains as a joint diffusion.

Random noise levels Several works replace Gaussian noise by a heavy tail distribution Pandey et al. (2024); Shariatian et al. (2025); YOON et al. (2023). LIM YOON et al. (2023) and DLPM Shariatian et al. (2025) drive the forward process by α-stable Lévy increments. Unlike Gamma family, which is closed under addition only at a common rate, stable laws are closed under addition at any scale, so nothing couples the rate to the schedule and the degeneracy of § 3 does not arise. The price is infinite variance, under which a pin of the form $\mathbb{E}[B_t] = \bar{\sigma}_t^2$ is undefined. t-EDM Pandey et al. (2024) places a Gamma law on the precision rather than the variance (ours). Its Student-t kernel is a Gaussian whose variance is divided by one Gamma draw shared across the whole trajectory. Precision-Gamma is conjugate to the Gaussian, which yields closed-form Student-t marginals and heavy tails, while Variance-Gamma is not, which is why our posterior is GIG. What the variance side buys is additivity in time, and a path with the monotonicity and coarsening of § 2-§ 4, whereas a single shared precision has no temporal structure. In the decomposition of Corollary B.3, t-EDM randomizes the total and keeps the temporal allocation at the nominal schedule. § 5 does in the opposite way, pinning the total and moving the allocation across spatial direction.

Image inpainting and motion in-betweening Both tasks generate from a partial observation that must be preserved. They differ only in the axis along which the observation is given, spatial for inpainting and temporal for in-betweening.

Diffusion inpainting divides into two approaches. Training-free methods treat the mask as an inverse problem and impose the observation at every reverse step with the schedule unchanged, by replacement Lugmayr et al. (2022), by projection onto the range-null space Wang et al. (2022); Kawar et al. (2022), or by guidance through a likelihood gradient Chung et al. (2022; 2024); Song et al. (2021). Preservation is exact by construction, and the masked region is generated at the same noise level as the context. Training-based methods instead change what the model sees. RAD Kim et al. (2025) and TD-Paint Mayet et al. (2025) hold the observed region at timestep zero and fine-tune on that configuration, and LaMa Suvorov et al. (2022) is the non-diffusion reference whose mask families we adopt. Our construction belongs to the second group, but trains on the continuum between homogeneous generation and the inpainting endpoint.

Motion in-betweening has followed the same division. MDM Tevet et al. (2022) in-betweens by imputation, replacing keyframe features at every step as RePaint does. CondMDI Cohan et al. (2024) trains a keyframe-conditioned model with randomly placed constraints, OmniControl Xie et al. (2024) conditions on joint trajectories through a control branch, and TEDi Zhang et al. (2024) and Diffusion Forcing Chen et al. (2024) assign per-frame noise levels for autoregressive generation. All receive the constraint through the denoiser's input. The construction of § 5 carries it in the noise level of the pinned frames instead, with the same pin and ratio as the spatial case.

B The Gamma–Beta Relation

Lemma B.1 (Gamma–Beta decomposition; the sufficiency direction of Lukacs (1955))**.** *Let $A_1, A_2 > 0$ and $b > 0$, and let $G_1 \sim \mathrm{Ga}(A_1, b)$ and $G_2 \sim \mathrm{Ga}(A_2, b)$ be independent with the same rate b, where their density are*

$$f_{G_i}(g) = \frac{b^{A_i}}{\Gamma(A_i)}\, g^{A_i - 1} e^{-bg} \quad \textit{for } g > 0, \quad f_{G_i}(g) = 0 \quad \textit{for } g \le 0.$$

Then

$$S := G_1 + G_2 \sim \mathrm{Ga}(A_1 + A_2,\, b), \qquad R := \frac{G_1}{G_1 + G_2} \sim \mathrm{Beta}(A_1, A_2),$$

and S and R are independent.

Proof. The map $\varphi(g_1, g_2) = \big(g_1 + g_2,\ g_1/(g_1 + g_2)\big)$ is a continuously differentiable bijection from $(0,\infty)^2$ onto the rectangle $(0,\infty) \times (0,1)$, with inverse $\psi(s, r) = \big(sr,\ s(1-r)\big)$, and $(S, R) = \varphi(G_1, G_2)$. Its Jacobian is

$$\left|\det \frac{\partial(g_1, g_2)}{\partial(s, r)}\right| = \left|\det \begin{pmatrix} r & s \\ 1-r & -s \end{pmatrix}\right| = \big|-sr - s(1-r)\big| = s.$$

By independence, the joint density of (G_1, G_2) is the product of the marginals,

$$f_{G_1,G_2}(g_1, g_2) = \frac{b^{A_1}}{\Gamma(A_1)} g_1^{A_1-1} e^{-bg_1} \cdot \frac{b^{A_2}}{\Gamma(A_2)} g_2^{A_2-1} e^{-bg_2} = \frac{b^{A_1+A_2}}{\Gamma(A_1)\Gamma(A_2)}\, g_1^{A_1-1} g_2^{A_2-1} e^{-b(g_1+g_2)}$$

on $(0,\infty)^2$. Substituting $(g_1, g_2) = \psi(s, r)$ and multiplying by the Jacobian,

$$\begin{aligned} f_{S,R}(s, r) &= \frac{b^{A_1+A_2}}{\Gamma(A_1)\Gamma(A_2)} (sr)^{A_1-1} \big(s(1-r)\big)^{A_2-1} e^{-bs} \cdot s \\ &= \underbrace{\frac{b^{A_1+A_2}}{\Gamma(A_1+A_2)}\, s^{A_1+A_2-1} e^{-bs}}_{=:\,p(s)} \cdot \underbrace{\frac{r^{A_1-1}(1-r)^{A_2-1}}{\mathrm{B}(A_1, A_2)}}_{=:\,q(r)} \end{aligned}$$

for $(s, r) \in (0,\infty) \times (0,1)$, using $\mathrm{B}(A_1, A_2) = \Gamma(A_1)\Gamma(A_2)/\Gamma(A_1 + A_2)$. Because the domain is a rectangle, extending p and q by zero makes $f_{S,R} = p\, q$ hold on all of $\mathbb{R}^2$. Since p and q are probability densities, they are the marginals of S and R, and $f_{S,R} = f_S f_R$ gives $S \perp R$. Specially, $p(s) = \mathrm{Ga}(A_1 + A_2, b)$ and $q(r) = \mathrm{Beta}(A_1, A_2)$. □

Corollary B.2 (Additivity)**.** *If $\beta_s \sim \mathrm{Ga}(a_s, b)$ are independent with a common rate b, then*

$$B_t = \sum_{s \le t} \beta_s \sim \mathrm{Ga}(A_t, b) \text{ with } A_t = \sum_{s \le t} a_s.$$

Proof. Apply the first part of the Lemma B.1 repeatedly. □

Corollary B.3 (The removed fraction is Beta)**.** *With β_t and $B_{t-1} = \sum_{s<t} \beta_s$ as above,*

$$w := \frac{\beta_t}{B_t} = \frac{\beta_t}{\beta_t + B_{t-1}} \sim \mathrm{Beta}(a_t, A_{t-1}), \quad w \perp B_t.$$

Proof. Set $G_1 = \beta_t \sim \mathrm{Ga}(a_t, b)$ and $G_2 = B_{t-1} \sim \mathrm{Ga}(A_{t-1}, b)$, which are independent because the B_{t-1} does not involve β_t. Lemma B.1 gives the $w = G_1/(G_1 + G_2) \sim \mathrm{Beta}(a_t, A_{t-1})$ and $B_t = G_1 + G_2 \sim \mathrm{Ga}(A_t, b)$ are independent. □

C Further discussion of Variance Path Diffusion construction (§ 2)

C.1 Detailed derivation of reverse kernel.

$$\begin{aligned} q(x_{t-1} | x_t, x_0, B_t) &= \frac{q(x_t | x_{t-1}, x_0, B_t)\, q(x_{t-1} | x_0, B_t)}{q(x_t | x_0, B_t)} \\ &\propto q(x_t | x_{t-1}, x_0, \beta_t)\, q(x_{t-1} | x_0, B_{t-1}) \\ &= \mathcal{N}(x_t; x_{t-1}, \beta_t I)\, \mathcal{N}(x_{t-1}; x_0, B_{t-1} I) \\ &= \mathcal{N}(\tilde{\mu}_t, \tilde{\sigma}_t^2 I), \end{aligned}$$

with

$$\tilde{\mu}_t = \frac{\beta_t\, x_0 + B_{t-1}\, x_t}{B_t}, \quad \tilde{\sigma}_t^2 = \frac{B_{t-1}\, \beta_t}{B_t}.$$

C.2 VARIANCE POSTERIOR

Because both the likelihood and the prior are conjugate in B_t, the posterior over the realized variance is available in closed form. Writing $r = \|x_t - x_0\|$,

$$q(B_t \mid x_t, x_0) \;\propto\; \underbrace{B_t^{-d/2}\, e^{-r^2/(2B_t)}}_{\text{Gaussian likelihood}} \cdot \underbrace{B_t^{A_t-1}\, e^{-bB_t}}_{\text{Gamma prior}} = \mathrm{GIG}\big(\lambda = A_t - \tfrac{d}{2},\ \chi = r^2,\ \psi = 2b\big)\,,$$

where GIG denotes the generalized inverse Gaussian distribution. All moments of B_t given (x_t, x_0) are therefore ratios of modified Bessel functions,

$$\mathbb{E}[B_t^s | x_t, x_0] = \left(\frac{r^2}{2b}\right)^{s/2} \frac{K_{\lambda+s}(r\sqrt{2b})}{K_\lambda(r\sqrt{2b})}.$$

This is a property that only stochastic variance diffusion has. Given a clean image and a noisy observation, the noise level that produced it is a well-defined random variable with a tractable posterior. One application is the setting of SVNR Pearl et al. (2023), where the noise level of a real image varies across pixels and must be estimated before denoising. The GIG posterior gives that estimate per region from the residual against the denoiser's prediction, without a separate estimator. We leave this to future work.

C.3 PATH POSTERIOR: GAMMA-DIRICHLET DUALITY

Let $w_{1:t} := \left(\frac{\beta_1}{B_t}, \ldots, \frac{\beta_t}{B_t}\right)$ denote the normalized allocation of the total budget across time. By the Gamma-Dirichlet duality,

$$w_{1:t} \sim \mathrm{Dir}(a_1, \ldots, a_t) \quad \perp \quad B_t \sim \mathrm{Ga}(A_t, b).$$

That is, the *total* variance accumulated by time t and its *allocation* across the preceding steps are independent, and the change of variables $\beta_{1:t} \leftrightarrow (B_t, w_{1:t})$ separates the path into these two factors.

This separation propagates to the posterior. Since the forward likelihood depends on the path only through B_t, that is $x_t \perp w_{1:t} \mid (x_0, B_t)$, we obtain

$$\begin{aligned} q(B_t, w_{1:t}|x_0, x_t) &= \frac{q(x_t|B_t, w_{1:t}, x_0) q(B_t, w|x_0)}{q(x_t|x_0)} \\ &\propto q(x_t|x_0, B_t) \cdot q(B_t, w_{1:t}) \\ &= \underbrace{q(x_t|x_0, B_t)\mathrm{Ga}(B_t; A_t, b)}_{\text{depend on } B_t} \times \underbrace{\mathrm{Dir}(w_{1:t}; a_1, \cdots, a_t)}_{\text{depend on } w_{1:t}}, \end{aligned}$$

using in addition that the path prior is independent of the data, $q(B_t, w_{1:t} \mid x_0) = q(B_t, w_{1:t})$, and that $q(x_t \mid x_0)$ is constant in the variables of interest. Normalizing each factor gives

$$q(B_t \mid x_0, x_t) = \mathrm{GIG}\left(A_t - \tfrac{d}{2}, r^2, 2b\right), \qquad q(w_{1:t} \mid x_0, x_t) = \mathrm{Dir}(a_1, \ldots, a_t).$$

The two factors behave differently under observation. The first differs from its prior $\mathrm{Ga}(A_t, b)$, so (x_0, x_t) is informative about the total budget. The second, however, is identical to its prior $\mathrm{Dir}(a_1, \cdots, a_t)$, so the observation carries no information at all about how that budget was allocated across steps.

Two consequences matter later. First, the entire variance path can be sampled given (x_0, x_t) by drawing B_t from the GIG posterior and the allocation from unchanged Dirichlet prior, which is what makes the ELBO of § C.6 exact. Second, the allocation is a design variable rather than something the data determines, and § 5 exploits this by allocating variance across spatial regions instead of across time.

This statement concerns a single observation. Observing the state at several times would constrain the ratios of the corresponding cumulative variances, and these ratios are functions of the allocation, so the allocation would no longer be conditionally independent of the data. This does not arise in our setting, since both training and sampling condition on one step at a time.

C.4 THE REVERSE COVARIANCE

Equation 2 is conditioned on both the path and the clean data. In generation neither is available. The sampler knows B_t but must marginalize over x_0, and over the split w that produced B_{t-1} from B_t. Doing both yields the following.

Lemma C.1 (Law of total covariance)**.** *For a random vector X and a random variable Y,* $\mathrm{Cov}[X] = \mathbb{E}_Y[\mathrm{Cov}[X \mid Y]] + \mathrm{Cov}_Y[\mathbb{E}[X \mid Y]]$.

Proof. Let's define $M := \mathbb{E}[X \mid Y]$. The tower property gives $\mathbb{E}[XX^\top] = \mathbb{E}_Y[\mathrm{Cov}[X \mid Y]] + \mathbb{E}_Y[MM^\top]$ and $\mathbb{E}[X]\mathbb{E}[X]^\top = \mathbb{E}_Y[M]\mathbb{E}_Y[M]^\top$. Then $\mathrm{Cov}[X] = \mathbb{E}[XX^\top] - \mathbb{E}[X]\mathbb{E}[X]^\top = \mathbb{E}_Y[\mathrm{Cov}[X|Y]] + \mathrm{Cov}_Y[\mathbb{E}[X|Y]]$ □

Theorem C.2 (Stochastic-variance reverse moments). *Let $\hat{\mu}_0$ and Σ_0 denote the mean and covariance of $p(x_0 \mid x_t, B_t)$, and let the scalar $w \sim \mathrm{Beta}(a_t, A_{t-1})$ be independent of x_0 and of B_t given x_t. Then the exact one-step reverse moments are*

$$\mathbb{E}\big[x_{t-1} \mid x_t, B_t\big] = \bar{w}\,\hat{\mu}_0 + (1-\bar{w})\,x_t, \qquad \bar{w} = \mathbb{E}[w] = \frac{a_t}{A_t},$$

$$\mathrm{Cov}\big[x_{t-1} \mid x_t, B_t\big] = \underbrace{c_1\,B_t I}_{\textit{bridge noise}} + \underbrace{c_2\,\Sigma_0}_{\textit{denoiser uncertainty}} + \underbrace{c_3\,(\hat{\mu}_0 - x_t)(\hat{\mu}_0 - x_t)^\top}_{\textit{path uncertainty (rank one)}},$$

with

$$c_1 = \mathbb{E}[w(1-w)] = \frac{a_t A_{t-1}}{A_t(A_t+1)}, \quad c_2 = \mathbb{E}[w^2] = \frac{a_t(a_t+1)}{A_t(A_t+1)}, \quad c_3 = \mathrm{Var}[w] = \frac{a_t A_{t-1}}{A_t^2(A_t+1)}.$$

Proof. Note that condition on (x_t, B_t) two quantities remain random. $x_0 \sim p(x_0 \mid x_t, B_t)$ with mean $\hat{\mu}_0$ and covariance Σ_0. $w \sim \mathrm{Beta}(a_t, A_{t-1})$ is independent of x_0 and of B_t. The noise $z \sim \mathcal{N}(0, I)$ is independent of both. Let's rearrange the Eq. 2 by writing $u := x_0 - x_t$, where $\mathbb{E}[u] = \hat{\mu}_0 - x_t =: \hat{u}$ and $\mathrm{Cov}[u] = \Sigma_0$ gives

$$x_{t-1} = x_t + w\,u + \sqrt{w(1-w)B_t}\,z. \tag{4}$$

Mean. Since $w \perp u$ and $\mathbb{E}[z] = 0$, $\mathbb{E}[x_{t-1}] = x_t + \mathbb{E}[w]\,\mathbb{E}[u] = \bar{w}\hat{\mu}_0 + (1-\bar{w})x_t$.

Covariance. Note that

$$\mathrm{Cov}[x_{t-1}|x_t, B_t] = \mathrm{Cov}\left[x_t + wu + \sqrt{w(1-w)B_t}z\right] = \mathrm{Cov}\left[wu + \sqrt{w(1-w)B_t}z\right].$$

Define $P := wu$ and $Q := \sqrt{w(1-w)B_t}z$ then $\mathrm{Cov}[P+Q] = \mathrm{Cov}[P] + \mathrm{Cov}[Q] + \mathrm{Cov}[P,Q] + \mathrm{Cov}[P,Q]^\top$. Note that $\mathrm{Cov}[P,Q] = \mathbb{E}[PQ^\top] - \mathbb{E}[P]\mathbb{E}[Q]^\top = 0$ because $\mathbb{E}[Q] = \mathbb{E}[\sqrt{w(1-w)B_t}z] = \mathbb{E}[\sqrt{w(1-w)B_t}]\mathbb{E}[z] = 0$ and $\mathbb{E}[PQ^\top] = \mathbb{E}[w\sqrt{w(1-w)B_t}uz^\top] = 0$. Therefore, we only need to compute $\mathrm{Cov}[P]$ and $\mathrm{Cov}[Q]$. Let's compute $\mathrm{Cov}[Q]$ first as follows:

$$\mathrm{Cov}[Q] = \mathrm{Cov}\left[\sqrt{w(1-w)B_t}z\right] = \mathbb{E}\left[w(1-w)B_t zz^\top\right] = \mathbb{E}[w(1-w)]B_t\mathbb{E}[zz^\top] = c_1 B_t I.$$

For $\mathrm{Cov}[wu]$, we need to apply Lemma C.1 with $Y = w$. Then

$$\mathrm{Cov}[wu] = \underbrace{\mathbb{E}_w[\mathrm{Cov}[wu|w]]}_{\text{(A)}} + \underbrace{\mathrm{Cov}_w[\mathbb{E}[wu|w]]}_{\text{(B)}}.$$

(A). As w is given, $\mathrm{Cov}[wu|w] = w^2\mathrm{Cov}[u] = w^2\Sigma_0$ and $\mathbb{E}_w[w^2\Sigma_0] = \mathbb{E}[w^2]\Sigma_0 = c_2\Sigma_0$.

(B). Similarly, $\mathbb{E}[wu|w] = w\mathbb{E}[u] = w\hat{u}$, then $\mathrm{Cov}_w[w\hat{u}] = \mathrm{Var}[w]\hat{u}\hat{u}^\top = c_3\hat{u}\hat{u}^\top$.

□

Each term has a distinct origin. The first is the bridge noise of Eq. 2, the randomness that remains even when x_0 is known exactly. The second propagates the denoiser's uncertainty about x_0 through the interpolation, scaled by w^2 because the mean depends on x_0 with slope w. The third arises because w itself is random. The reverse step knows the *direction* it should move, namely toward $\hat{\mu}_0$,

but not *how far*, and that one-dimensional uncertainty produces a rank-one contribution along the residual.

The coefficient ratios make the structure explicit:

$$\frac{c_3}{c_1} = \frac{1}{A_t}, \quad \frac{c_2}{c_1} = \frac{a_t + 1}{A_{t-1}}$$

The relative size of the path-uncertainty term depends only on the accumulated shape A_t, and in particular is independent of the data dimension d. The relative size of the denoiser-uncertainty term grows as the step becomes coarse (large a_t relative to A_{t-1}), so the correction matters most in the few-step regime. Both vanish in the deterministic limit $\kappa \to \infty$, where w concentrates on $\bar{w}$ and Theorem C.2 reduces to the familiar VE ancestral kernel.

The derivation of Theorem C.2 used only three facts about w, namely that it is a scalar, that it is independent of x_0 given (x_t, B_t), and that $B_{t-1} \leq B_t$. Therefore, under TVR construction (§ 4), all three hold for $\tilde{w}_t$, so Theorem C.2 applies verbatim with $c_1 = \mathbb{E}[\tilde{w}_t(1-\tilde{w}_t)], \quad c_2 = \mathbb{E}[\tilde{w}_t^2], \quad c_3 = \mathrm{Var}[\tilde{w}_t]$, each available in closed form from the Beta moments of V_t.

Corollary C.3 (Containment of the deterministic sampler). *Fix a schedule $\bar{\sigma}_{1:T}$ and set $A_t = \kappa\phi_t, b = \kappa/\bar{\sigma}_T^2$ as in § 2.2. Then for every $\kappa > 0$,*

$$\mathbb{E}[B_t] = \bar{\sigma}_t^2, \qquad \bar{w} = \frac{a_t}{A_t} = 1 - \frac{\bar{\sigma}_{t-1}^2}{\bar{\sigma}_t^2},$$

both independent of κ, while

$$\mathrm{Var}[B_t] = \frac{\bar{\sigma}_t^4}{\kappa\phi_t} = O(\kappa^{-1}), \qquad c_3 = \mathrm{Var}[w] = \frac{a_t A_{t-1}}{A_t^2(A_t+1)} = O(\kappa^{-1}).$$

Consequently $B_t \to \bar{\sigma}_t^2$ and $w \to \bar{w}$ as $\kappa \to \infty$, the rank-one term vanishes, and Theorem C.2 reduces to the deterministic VE ancestral kernel $c_1 B_t I + c_2 \Sigma_0$ with constant $c_1 = \bar{w}(1-\bar{w}), c_2 = \bar{w}^2$.

Proof. The two identities are immediate from $A_t = \kappa\phi_t$ and $b = \kappa/\bar{\sigma}_T^2$, in which κ cancels. For the variances, $\mathrm{Var}[B_t] = A_t/b^2$ and a_t, A_{t-1}, A_t are all proportional to κ, so the numerator of c_3 scales as κ^2 and the denominator as κ^3. Convergence in probability follows from Chebyshev's inequality, the means being already equal to the limits. □

Two points are worth stating explicitly. First, the mean path and the mean interpolation weight are exactly the deterministic schedule at every κ, not only in the limit; κ modulates the fluctuation around the schedule and nothing else. This is what allows κ to be varied at sampling time while keeping results comparable. Second, because the deterministic sampler is recovered exactly rather than approximately, VPD cannot underperform its deterministic counterpart by construction. The baseline is a point of the family.

Note that κ acts as a single dial for the stochasticity of the path, leaving the mean path untouched. Corollary C.3 in § C.4 makes precise the sense in which $\kappa \to \infty$ recovers the deterministic VE sampler.

Relation to prior work The first two terms are the VE analogue of the optimal reverse variance characterized by Analytic-DPM Bao et al. (2022); as in that line of work, Σ_0 is a property of the data distribution and must be estimated, and we do so with a learned diagonal head. What is closed-form here is the *mixing*. The coefficients c_1, c_2, c_3 follow from the path law alone. The third term has no counterpart in fixed-variance diffusion. It exists only because the noise level is itself a random variable, and it vanishes identically when the path is deterministic.

C.5 Score, denoiser, and the training objective

Marginalizing the variance rather than conditioning on it yields a conditional score in closed form. From $q(x_t|x_0) = \int \mathcal{N}(x_t|x_0, BI)p(B)dB$,

$$\begin{aligned}\nabla_{x_t} \log q(x_t|x_0) &= \frac{\nabla_{x_t} q(x_t|x_0)}{q(x_t|x_0)} \\ &= \frac{\int \nabla_{x_t} \mathcal{N}(x_t; x_0, BI) p(B) dB}{q(x_t|x_0)} \\ &= -(x_t - x_0) \int \frac{1}{B} \frac{\mathcal{N}(x_t; x_0, BI) p(B)}{q(x_t|x_0)} dB \\ &= -(x_t - x_0) \mathbb{E}_{q(B_t|x_t,x_0)}[B_t^{-1}|x_t, x_0] \\ &= -(x_t - x_0) \frac{\sqrt{2b}}{r} \frac{K_{\lambda-1}(\sqrt{2b}r)}{K_\lambda(\sqrt{2b}r)},\end{aligned}$$

with $\lambda = A_t - \frac{d}{2}$, using the GIG moment identity of § C.2. Because the mixing variable is a scalar, the direction of the score is unchanged by marginalization; only its magnitude is modified.

This expression is exact, but it is not what we train. For image-scale problems $\lambda = A_t - d/2$ is large and negative, and the Bessel ratio is numerically delicate in that regime. More importantly, marginalizing is unnecessary under a committed path, since B_t is drawn by the forward process and therefore known. Conditioning on it restores an ordinary Gaussian smoothing problem, for which Tweedie's formula gives

$$\mathbb{E}[x_0|x_t, B_t] = x_t + B_t \nabla_{x_t} \log q(x_t|B_t),$$

where q_B, is the marginal at fixed B_t. Two standard facts then apply: the minimizer of the squared error is the conditional mean, and by the identity above the conditional mean determines the score. Hence a single regression suffices, and we train a denoiser $\hat{x}_\theta$ with

$$\mathcal{L}_{\text{denoiser}} = \mathbb{E}_{x_0, B_t, \epsilon}\left[\|x_0 - \hat{x}_\theta(x_0 + \sqrt{B_t}\epsilon, B_t)\|_2^2\right].$$

Conditioning on B_t removes every special function from the training objective regardless of the law generating B_t. The argument above uses only that $q(x_t|x_0, B_t)$ is Gaussian. The path law and the denoiser are therefore separable, which is why a denoiser trained under one path law can be sampled under another without retraining, a property we use in § 4.

C.6 Exact ELBO

VPD admits a variational bound in which the path is an explicit latent variable. Let

$$\text{generative model:} \quad p_\theta(x_{0:T}, \beta_{1:T}) = p(x_T) p_\theta(\beta_{1:T} \mid x_T) \prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t, \beta_t),$$

$$\text{inference model:} \quad q(x_{1:T}, \beta_{1:T} \mid x_0) = q(\beta_{1:T}) \prod_{t=1}^{T} q(x_t \mid x_{t-1}, \beta_t),$$

with $q(\beta_{1:T}) = \prod_t \mathrm{Ga}(\beta_t; a_t, b)$ the fixed path prior. Because every conditional above is available in closed form, the bound decomposes exactly.

Theorem C.4 (Exact ELBO).

$$\log p_\theta(x_0) \geq \mathcal{L}_{rec} - \mathcal{L}_{diff} - \mathcal{L}_{prior} + \mathcal{L}_{path},$$

where

$$
\begin{aligned}
\mathcal{L}_{\text{rec}} &= \mathbb{E}_q\big[\log p_\theta(x_0 \mid x_1, \beta)\big] \\
\mathcal{L}_{\text{diff}} &= \sum_{t=2}^{T} \mathbb{E}_{q(\beta)\, q(x_t|x_0,B_t)}\Big[\mathrm{KL}\big(q(x_{t-1} \mid x_t, x_0, \beta) \,\big\|\, p_\theta(x_{t-1} \mid x_t, \beta)\big)\Big] \\
\mathcal{L}_{\text{prior}} &= \mathbb{E}_{q(\beta)}\Big[\mathrm{KL}\big(q(x_T \mid x_0, \beta) \,\big\|\, p(x_T)\big)\Big] \\
\mathcal{L}_{\text{path}} &= \mathbb{E}_{q(\beta)\, q(x_T|x_0,\beta)}\big[\log p_\theta(\beta_{1:T} \mid x_T)\big] \;+\; H\big[q(\beta_{1:T})\big].
\end{aligned}
$$

Proof. We can introduce inference model q and derive inequality using Jensen inequality:

$$
\log p_\theta(x_0) = \log \mathbb{E}_q\left[\frac{p_\theta(x_{0:T}, \beta_{1:T})}{q(x_{1:T}, \beta_{1:T}|x_0)}\right] \geq \mathbb{E}_q\left[\log \frac{p_\theta(x_{0:T}, \beta_{1:T})}{q(x_{1:T}, \beta_{1:T}|x_0)}\right].
$$

Factorizing the inside of $\log$ gives

$$
\begin{aligned}
\frac{p_\theta(x_{0:T}, \beta_{1:T})}{q(x_{1:T}, \beta_{1:T}|x_0)} &= \frac{p(x_T)p_\theta(\beta_{1:T}|x_T)\prod_t p_\theta(x_{t-1}|x_t, \beta_t)}{q(\beta_{1:T})\prod_t q(x_t|x_{t-1}, \beta_t)} \\
&= \underbrace{\frac{p(x_T)\prod_t p_\theta(x_{t-1}|x_t, \beta_t)}{\prod_t q(x_t|x_{t-1}, \beta_t)}}_{\text{mean chain given } \beta} \cdot \underbrace{\frac{p_\theta(\beta_{1:T}|x_T)}{q(\beta_{1:T})}}_{\text{path}}
\end{aligned} \tag{5}
$$

Note that mean chain given β term is exactly same as DDPM ELBO conditioned on β. Given the Markov property, we can derive $q(x_t|x_{t-1}, \beta_t) = q(x_t|x_{t-1}, x_0, \beta_t) = \frac{q(x_{t-1}|x_t, x_0, \beta_t)q(x_t|x_0, \beta_t)}{q(x_{t-1}|x_0, \beta_{t-1})}$, then the numerator becomes

$$
\prod_{t=1}^{T} q(x_t|x_{t-1}, \beta_t) = q(x_T|x_{t-1}, x_0, \beta_T)\prod_{t=2}^{T} q(x_{t-1}|x_t, x_0, \beta_t)
$$

Now let's place it back to Eq. 5 then

$$
\frac{p(x_T)\prod_t p_\theta(x_{t-1}|x_t, \beta_t)}{\prod_t q(x_t|x_{t-1}, \beta_t)} = \frac{p(x_T)\prod_t p_\theta(x_{t-1}|x_t, \beta_t)}{q(x_T|x_0, \beta_{1:T})\prod_{t=2}^{T} q(x_{t-1}|x_t, x_0, \beta_t)}.
$$

With the $\log$ the ELBO is derived as follows:

$$
\mathcal{L} = \mathbb{E}_q\left[\underbrace{\log p_\theta(x_0|x_1, \beta)}_{\mathcal{L}_{\text{rec}}} + \underbrace{\log \frac{p(x_T)}{q(x_T|x_0, \beta)}}_{-\mathcal{L}_{\text{prior}}} + \underbrace{\sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t, \beta)}{q(x_{t-1}|x_t, x_0, \beta)}}_{-\mathcal{L}_{\text{diff}}} + \underbrace{\log \frac{p_\theta(\beta_{1:T}|x_T)}{q(\beta_{1:T})}}_{\mathcal{L}_{\text{path}}}\right]
$$

Each term can be further derived as

$$
\mathcal{L}_{\text{diff}} = -\sum_{t=2}^{T} \mathbb{E}_q\left[\log \frac{q(x_{t-1}|x_t, x_0, \beta)}{p_\theta(x_{t-1}|x_t, \beta)}\right] = -\sum_{t=2}^{T} \mathbb{E}_q\Big[\mathrm{KL}\big(q(x_{t-1}|x_t, x_0, \beta)\|p_\theta(x_{t-1}|x_t, \beta)\big)\Big],
$$

$$
\begin{aligned}
\mathcal{L}_{\text{prior}} = -\mathbb{E}_q\left[\log \frac{p_T}{q(x_T|x_0, \beta)}\right] &= \mathbb{E}_{q(\beta)}\left[\mathbb{E}_{q(x_T|x_0,\beta)}\left[\log \frac{q(x_T|x_0, \beta)}{p(x_T)}\right]\right] \\
&= \mathbb{E}_{q(\beta)}\Big[\mathrm{KL}\big(q(x_T|x_0, \beta)\|p(x_T)\big)\Big],
\end{aligned}
$$

$$
\mathcal{L}_{\text{path}} = \mathbb{E}_q\left[\log \frac{p_\theta(\beta_{1:T}|x_T)}{q(\beta_{1:T})}\right] = \mathbb{E}_q[\log p_\theta(\beta_{1:T}|x_T)] + H[q(\beta_{1:T})]
$$

□

Diffusion term. Since q and p_θ share the same covariance in Eq. 2 and differ only in the mean, the per-step KL is

$$\mathrm{KL}_t = \frac{\|\tilde{\mu}_t - \mu_\theta\|^2}{2\tilde{\sigma}_t^2} = \frac{\beta_t}{2B_{t-1}B_t}\|x_0 - \hat{x}_\theta(x_t, B_t)\|_2^2 = \frac{1}{2B_t}\cdot\frac{w}{1-w}\|x_0 - \hat{x}_\theta\|_2^2.$$

The Gamma-Dirichlet duality of § C.3 lets us take the expectation over the split separately from the denoising error, because the denoiser's input depends on the path only through B_t:

$$\mathbb{E}_q[\text{weight}\cdot\text{error}] = \mathbb{E}_{B_t,\epsilon}\left[\mathbb{E}_{w|B_t}[\text{weight}]\cdot\underbrace{\text{error}(B_t,\epsilon)}_{\text{independent to split}}\right], \quad \mathbb{E}\left[\frac{w}{1-w}\right] = \frac{a_t}{A_{t-1}-1}$$

The weight is finite if and only if $A_{t-1} > 1$. We show in § 3 that under the shared-rate construction this condition fails on the majority of steps, proving that objective is not merely inconvenient but vacuous. In practice, therefore, we use the standard EDM weighting.

Prior term. With $q(x_T|x_0,\beta_{1:T}) = \mathcal{N}(x_0, B_T I)$ and $p(x_T) = \mathcal{N}(0,\sigma_\star^2 I)$,

$$\mathrm{KL} = \frac{1}{2}\left[d\frac{B_T}{\sigma_\star^2} + \frac{\|x_0\|_2^2}{\sigma_\star^2} - d + d\log\frac{\sigma_\star^2}{B_T}\right],$$

and since $B_T \sim \mathrm{Ga}(A_T, b)$ has $\mathbb{E}[B_T] = A_T/b$ and $\mathbb{E}[\log B_t] = \psi^{(0)}(A_T) - \log b$, the expectation is closed form:

$$\mathcal{L}_{\text{prior}} = \frac{1}{2}\left[\frac{d\,A_T}{b\,\sigma_\star^2} + \frac{\|x_0\|^2}{\sigma_\star^2} - d + d\log\sigma_\star^2 - d\big(\psi^{(0)}(A_T) - \log b\big)\right].$$

Path term. Two choices of path model are available: a fixed Gamma process, in which case $p_\theta(\beta|x_T) = q(\beta)$ and $\mathcal{L}_{\text{path}} = 0$; or an amortized generator $p_\theta(\beta_{1:T}|x_T)$, optionally conditioned on a class label. In the amortized case,

$$\mathcal{L}_{\text{path}} = \underbrace{\mathbb{E}_{q(\beta),\, x_T = x_0 + \sqrt{B_T}\epsilon}\big[\log p_\theta(\beta_{1:T} \mid x_T)\big]}_{\text{cross-entropy}} + \underbrace{\textstyle\sum_t \big[a_t - \log b + \log\Gamma(a_t) + (1-a_t)\,\psi^{(0)}(a_t)\big]}_{\text{Gamma entropy (constant)}}.$$

An amortized generator must receive a conditioning signal. Without one it degenerates to a copy of the path prior. We derive this term for completeness, since it shows that the path can in principle be learned as a latent variable. We leave this direction as future works.

C.7 What follows with $A_1 > 1$

The coefficient of the path uncertainty term of Theorem C.2 is $c_3 = \zeta_t^2\mathrm{Var}[V_t]$, so its total magnitude along the chain is $\mathcal{M} := \sum_t \mathrm{Var}[V_t]$.

Proposition C.5 (Total magnitude of the rank one term). *On a fine grid,*

$$\mathcal{M} \to \log\frac{A_T(A_1+1)}{A_1(A_T+1)} \xrightarrow[A_T\to\infty]{} \log\Big(1+\frac{1}{A_1}\Big),$$

which is decreasing in A_1; hence $\mathcal{M} < \log 2$ whenever $A_1 > 1$.

Proof. Since a_t is the increment in A and $A_{t-1}/A_t \to 1$ on a fine grid, the summand is $\big(A_t(A_t+1)\big)^{-1}\Delta A$, so $\mathcal{M}$ is a Riemann sum in A. Partial fractions give the anti-derivative $\log\frac{A}{A+1}$, which is bounded above by zero and tends to zero as $A\to\infty$, so the integral converges at the upper endpoint for every κ including $\kappa = \infty$, and diverges at the lower endpoint as $A_1 \to 0$. Therefore,

$$\mathcal{M} \longrightarrow \int_{A_1}^{\kappa}\frac{dA}{A(A+1)} = \log\frac{\kappa(A_1+1)}{A_1(\kappa+1)} \xrightarrow[\kappa\to\infty]{} \log\Big(1+\frac{1}{A_1}\Big).$$

For $A_1 > 1$. total magnitude $\mathcal{M}$ has upper bound $\mathcal{M} < \log 2$. □

The total depends only on the endpoints and not on the interpolating profile, so a learned time profile can redistribute the path uncertainty but cannot increase it. And the construction with a large third therm of Theorem C.2 is one that violates $A_1 > 1$.

C.8 COMMITTED PATHS

Equation 2 shows that *conditioning on the variance path makes the entire reverse chain Gaussian.* This suggests the sampling scheme of the framework: draw the whole path $\beta_{1:T}$ once, and then run

$$p_\theta(x_{0:T-1} \mid x_T, \beta_{1:T}) = \prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t, \beta_t),$$

a chain in which every step is a Gaussian transition with known coefficients. We call this a *committed path*. The stochasticity of the variance is resolved before generation begins, not interleaved with it. Committing has three benefits. The denoiser sees B_t as a known conditioning input, so its training problem stays standard (§ C.4); the reverse coefficients are available in closed form instead of requiring marginalization over an unknown B_t; and the forward marginals are matched exactly at every step by construction. The mean path used in all experiments of § 7 is the committed path with β_t set to its expectation.

D PROOFS OF THEOREM, PROPOSITION, AND COROLLARY

D.1 PROOF OF PROPOSITION 3.1

Proposition 3.1 (Endpoint bound)**.** *For the shared rate construction of § 2, and for any noise schedule $\bar\sigma_{1:T}$ with any number of steps T, and any stochasticity scale κ, $\frac{A_1}{A_T} = \phi_1 = \frac{\sigma^2_{min}}{\sigma^2_{max}}$.*

Proof. Matching the schedule in expectation gives $A_t = b\bar\sigma_t^2$ with b common to every t (§ 2). Taking the ratio at the two endpoints cancels b, and only $\bar\sigma_1 = \sigma_{\min}$ and $\bar\sigma_T = \sigma_{\max}$ remain. Neither κ, which enters through b, nor the interior values $\bar\sigma_{2:T-1}$, nor T appears. □

D.2 PROOF OF COROLLARY 3.2

Corollary 3.2 (**Antinomy of the shared rate**)**.** *If $A_1 > 1$ then $b > \sigma_{min}^{-2}$ and hence $\kappa = A_T = b\sigma^2_{max} > \frac{1}{\phi_1}$,* $\mathrm{RSD}[B_T] = \kappa^{-1/2} < \sqrt{\phi_1} = \frac{\sigma_{\min}}{\sigma_{\max}}$.

Proof. From $A_1 = b\sigma^2_{\min}$, the hypothesis gives $b > \sigma^{-2}_{\min}$. Multiplying by $\sigma^2_{\max}$ and using $\kappa = A_T = b\sigma^2_{\max}$ gives $\kappa > \sigma^2_{\max}/\sigma^2_{\min} = 1/\phi_1$, and $\mathrm{RSD}[B_T] = \kappa^{-1/2}$ then gives the second bound. □

D.3 PROOF OF PROPOSITION 4.2

Proposition 4.2 (**Exact marginals**)**.** *Under Definition 4.1, $B_t \sim \mathrm{Ga}(A_t, b_t)$ for every t.*

Proof. By backward induction. If $X \sim \mathrm{Ga}(A, b)$ and $V \sim \mathrm{Beta}(A', A - A')$ are independent then $VX \sim \mathrm{Ga}(A', b)$, which follows from the Gamma representation of the Beta together with Lukacs' theorem. Scaling gives $c\mathrm{Ga}(A', b) = \mathrm{Ga}(A', b/c)$, and $\zeta_t = \frac{b_t}{b_{t-1}}$ by definition. □

D.4 PROOF OF PROPOSITION 4.3

Proposition 4.3 (**Monotonicity**)**.** *$B_{t-1} \le B_t$ almost surely iff $\zeta_t \le 1$, equivalently $\frac{A_t}{A_{t-1}} \le \frac{\bar\sigma_t^2}{\bar\sigma_{t-1}^2}$.*

Proof. Since $B_t > 0$ almost surely, $B_{t-1} \le B_t$ is equivalent to $\zeta_t V_t \le 1$. Sufficiency is trivial since $V_t \in (0,1)$. For necessity, suppose $\zeta_t > 1$ then $\frac{1}{\zeta_t} \in (0,1)$, and since the Beta density is positive on all of $(0,1)$ we have $P(V_t > 1/\zeta_t) > 0$, so $B_{t-1} > B_t$ with positive probability. □

D.5 Proof of Proposition 4.4

Proposition 4.4 (Exact coarsening). *For any $k \ge 1$, $B_{t-k} = \left(\prod_{s=t-k+1}^{t} \zeta_s\right)\left(\prod_{s=t-k+1}^{t} V_s\right) B_t$ where $\prod_{s=t-k+1}^{t} \zeta_s = \frac{b_t}{b_{t-k}}$, $\quad \prod_{s=t-k+1}^{t} V_s \sim \mathrm{Beta}(A_{t-k}, A_t - A_{t-k})$.*

Proof. The first is a telescoping product. For the second, the Mellin transform of a Beta variable in our parameterization factors as a ratio,

$$\mathbb{E}[V_s^\nu] = \frac{\Gamma(A_{s-1}+\nu)\Gamma(A_s)}{\Gamma(A_{s-1})\Gamma(A_s+\nu)} = \frac{g_\nu(A_{s-1})}{g_\nu(A_s)}, \quad g_\nu = \frac{\Gamma(A+\nu)}{\Gamma(A)},$$

so independence gives a telescoping product,

$$\mathbb{E}\left[\prod_s V_s^\nu\right] = \prod_{s=t-k+1}^{t} \frac{g_\nu(A_{s-1})}{g_\nu(A_s)} = \frac{g_\nu(A_{t-k})}{g_\nu(A_t)},$$

which is the ν-th moment of $\mathrm{Beta}(A_{t-k}, A_t - A_{t-k})$. Since both laws are supported on $[0,1]$, equality of all integer moments determines the distribution. □

Proposition 4.4 concerns a single transition, and equality of transitions does not by itself determine a joint law, since correlated blocks would reproduce the same marginals while giving different trajectory statistics. Let $\mathcal{T} = \{t_0 < \cdots < t_K = T\}$ be a subgrid and let $P^{(j)}$ denote the product of the V_s within block j. First, unrolling Definition 4.1 writes the whole trajectory as a deterministic function $\Phi(B_T, P^{(K)}, \cdots, P^{(1)})$, and the path constructed directly on $\mathcal{T}$ is the same function of the corresponding inputs, so it suffices to compare the two input tuples. Second, each $P^{(j)}$ is a function of the V_s in a single block, the blocks are disjoint, and $\{B_T\} \cup \{V_s\}$ are mutually independent by construction, so the tuple is a product measure whose marginals are given by Proposition 4.4. The two joint laws therefore coincide, and the model is projectively consistent under arbitrary, possibly non uniform, subsampling of the time grid.

E The Variance-Preserving formulation

The main formulation is written in Variance-Exploding (VE) convention of § 2.1, where the state is $x_t = x_0 + \sqrt{B_t}\epsilon$ and the accumulated variance budget B_t is the random variable. However, most of the diffusion literature and pretrained models that we use for finetuning are built on the Variance Preserving (VP) convention $x_t = \sqrt{\bar\alpha_t}x_0 + \sqrt{1-\bar\alpha_t}\epsilon$ Ho et al. (2020); Kim et al. (2025); Dhariwal & Nichol (2021).

In this appendix, we provide the change of variables for both constructions: VPD in § 2, and TVR § 4. The two conventions are related by a rescaling that holds sample by sample, for every draw of the random path and every draw of the noise, the VE state times a scalar is the VP state, with the same x_0 and the same ϵ. This allows every later statement to be transported by substitution instead of re-derived. The rescaling itself is not new. It is the $x = s(t)(x_0 + \sigma(t)\epsilon)$ factorization of Karras et al. (2022), and the underlying fact that a diffusion process is determined by its signal-to-noise ratio, with the overall scale carrying no information, is due to Kingma et al. (2023), and Song et al. (2021) established VE and VP families and their relation. We restate it only because those treatments take the noise level to be a deterministic schedule, whereas B_t is random variable here, so the rescaling factor is not a known results in advance but a function of budget that particular trajectory drew.

E.1 THE STATE MAP

Both conventions describe the same family of Gaussian up to a scale, so the map is fixed by matching the signal-to-noise ratio, following Karras et al. (2022).

Lemma E.1 (Change of variables). *Fix any realization of the path $B_{1:T}$ and any noise draw ϵ, and set*

$$\bar{\alpha}_t = \frac{1}{1+B_t}, \quad s_t = \sqrt{\bar{\alpha}_t},$$

then $x_t^{VP} := s_t x_t^{VE}$ satisfies $x_t^{VP} = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$ with the same ϵ.

Proof. $s_t\sqrt{B_t} = \sqrt{B_t/(1+B_t)} = \sqrt{1-\bar{\alpha}_t}$ and $s_t = \sqrt{\bar{\alpha}_t}$ by definition. □

The lemma is an algebraic identity between two expressions for the same vector, not an equality of distributions, and it involves no expectations, no limit, and no approximation. Therefore, a trajectory simulated in one convention can be converted to the other at any step and converted back with no loss, allowing VPD samplers to run in either way. The map is a decreasing bijection of $(0,\infty)$ onto $(0,1)$, whose SNR is

$$\mathrm{SNR}_t := \frac{\bar{\alpha}_t}{1-\bar{\alpha}_t} = \frac{1}{B_t},$$

so the accumulated budget is exactly the inverse signal-to-noise ratio. Every statement in the main text can be read as a statement about a random inverse-SNR process with a Gamma law. The monotonicity of B_t established in § 2.2 becomes monotone decay of SNR_t, and the terminal condition $B_T = \sigma_\star^2$ becomes $\bar{\alpha}_T = (1+\sigma_\star^2)^-1$.

E.2 THE INCREMENT STRUCTURE TRANSFERS

In the VP convention the forward step is $x_t = \sqrt{1-\beta_t^{\mathrm{VP}}}x_{t-1} + \sqrt{\beta_t^{\mathrm{VP}}}\epsilon_t$ with $\beta_t^{\mathrm{VP}} = 1 - \bar{\alpha}_t/\bar{\alpha}_{t-1}$. Substituting Lemma E.1,

$$\beta_t^{\mathrm{VP}} = 1 - \frac{1+B_{t-1}}{1+B_t} = \frac{B_t - B_{t-1}}{1+B_t} = \frac{\beta_t}{1+B_t},$$

so the VP increment is the VE increment discounted by the current total. This gives the duality preserved.

Proposition E.2 (The increment structure transfers). *Let $\tilde{w}_t = 1 - B_{t-1}/B_t$ be the removed fraction of § 4.3, which is $w \sim \mathrm{Beta}(a_t, A_{t-1})$ under the shared rate and $1-\zeta_t V_t$ under the time-varying rate. In the two conventions the forward increments are*

$$\beta_t^{\mathrm{VE}} = \tilde{w}_t\, B_t, \qquad \beta_t^{\mathrm{VP}} = \tilde{w}_t\,(1-\bar{\alpha}_t),$$

the same random fraction ($\tilde{w}$) applied to the current noise variance of each convention. In particular $\tilde{w}_t \perp \bar{\alpha}_t$ whenever $\tilde{w}_t \perp B_t$, which holds in both constructions.

Proof. The VE identity is the definition of $\tilde{w}_t$. For VP, $\beta_t^{\mathrm{VP}} = \beta_t^{\mathrm{VE}}/(1+B_t)$ by the increment identity above, and $B_t/(1+B_t) = 1-\bar{\alpha}_t$ by Lemma E.1. Therefore

$$\beta_t^{\mathrm{VP}} = \frac{\beta_t}{1+B_t} = \frac{\beta_t}{B_t}(1-\bar{\alpha}_t) = \frac{B_t - B_{t-1}}{B_t}(1-\bar{\alpha}_t) = \tilde{w}_t(1-\bar{\alpha}_t).$$

Independence transfers because $\bar{\alpha}_t = (1+B_t)^{-1}$ is a function of B_t alone and $\tilde{w}_t \perp B_t$ is Lemma B.1 under the shared rate and is imposed by Definition 4.1 under the time-varying rate. □

Every statement of the main text about $\tilde{w}_t$, the coefficient of Theorem C.2 and the coarsening of Proposition 4.4 holds in the VP convention with B_t replaced by $1-\bar{\alpha}_t$.

E.3 THE REVERSE KERNEL

Because the VP state is a deterministic rescaling of the VE state given the path, the reverse transition transport without a new derivation.

Proposition E.3 (The reverse kernel of VP)**.** *Conditioned on the path, the reverse kernel of § 2.3 in VP coordinates is*

$$q(x_{t-1}^{VP}|x_t^{VP}, x_0) = \mathcal{N}\left(\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t^{VP}}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t^{VP}}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t^{VP}, \tilde{\beta}_t I\right),$$

where $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t^{VP} = w(1-\bar{\alpha}_{t-1})$.

Proof. Let's define $\alpha_t^{\mathrm{VP}} := \bar{\alpha}_t/\bar{\alpha}_{t-1} = 1-\beta_t^{\mathrm{VP}}$. Multiplying the VE kernel of Eq. equation 2 by s_{t-1} and substituting $x_t^{\mathrm{VE}} = x_t^{\mathrm{VP}}/s_t$ gives mean $s_{t-1}w\,x_0 + (1-w)(s_{t-1}/s_t)\,x_t^{\mathrm{VP}}$ and covariance $s_{t-1}^2\,w(1-w)B_t\,I$. Now, let's match the coefficients.

For the coefficient of x_0, $\beta_t^{\mathrm{VP}} = w(1-\bar{\alpha}_t)$ by Proposition E.2, so $\sqrt{\bar{\alpha}_{t-1}}\beta_t^{\mathrm{VP}}/(1-\bar{\alpha}_t) = s_{t-1}w$.

For the coefficient of x_t^{VP}, $s_{t-1}/s_t = (\alpha_t^{\mathrm{VP}})^{-1/2}$, and $(1-\bar{\alpha}_{t-1})/(1-\bar{\alpha}_t) = (1-w)\,\bar{\alpha}_{t-1}/\bar{\alpha}_t = (1-w)/\alpha_t^{\mathrm{VP}}$ using $B_{t-1} = (1-w)B_t$. Hence, $\sqrt{\alpha_t^{\mathrm{VP}}}(1-\bar{\alpha}_{t-1})/(1-\bar{\alpha}_t) = \sqrt{\alpha_t^{\mathrm{VP}}}\cdot(1-w)/\alpha_t^{\mathrm{VP}} = (1-w)/\sqrt{\alpha_t^{\mathrm{VP}}} = (1-w)\,s_{t-1}/s_t$.

For the covariance: $s_{t-1}^2(1-w)B_t = \bar{\alpha}_{t-1}B_{t-1} = 1-\bar{\alpha}_{t-1}$, so covariance is same as $w(1-\bar{\alpha}_{t-1})I$. □

F FURTHER DISCUSSION OF PER-REGION EXTENSION §5

This appendix supports § 5.

F.1 DETAILS OF EQ. 3

By setting Eq. 3, the pixel averaged variance follows the nominal schedule. Specifically,

$$\sum_g \frac{d_g}{d}\mathbb{E}[B_t^{(g)}] = \sum_g \frac{d_g}{d}\gamma_g^2\mathbb{E}[\bar{B}_t] = \sum_g \frac{d_g}{d}\gamma_g^2\bar{\sigma}_t^2 = \underbrace{(f\gamma_1^2 + (1-f)\gamma_2^2)}_{=1\text{by Eq. 3}}\bar{\sigma}_t^2$$

This pin ensures that the total mean of variance budget follows nominal variance.

F.2 INVARIANCE OF THE PATH PROPERTIES UNDER REGION-WISE RESCALING

Proposition F.1 (Invariance). *Let $B_t^{(g)} = \gamma_g^2\tilde{B}_t$ with γ_g constant in t. Then monotonicity holds for all regions under the original condition $\zeta_t \le 1$ (Proposition 4.3), exact coarsening holds per region with the Beta law of Proposition 4.4 unchanged, and the effective removed fraction*

$$\tilde{w}_t^{(g)} = 1 - \frac{B_{t-1}^{(g)}}{B_t^{(g)}} = 1 - \zeta_t V_t = \tilde{w}_t,$$

is common to both regions. Therefore Theorem C.2 applies to each with the same coefficients c_1, c_2, and c_3.

Proof. γ_g^2 is a positive constant and it is factored out of the equations. □

Conditioned on x_0 and the removed fraction, the reverse kernel is

$$x_{t-1}^{(g)} = (1-\tilde{w}_t)\,x_t^{(g)} + \tilde{w}_t\,x_0^{(g)} + \sqrt{\tilde{w}_t(1-\tilde{w}_t)B_t^{(g)}}\;z^{(g)},$$

in which the interpolation weight is a single scalar shared by the whole image and the ratio enters only through the injected noise scale.

F.3 SATURATION AND THE START LEVEL

Proposition F.2 (**Admissible level ratio**). *Let $\sigma_t^{(g)} := \gamma_g \bar{\sigma}_t$ with (γ_1, γ_2) given by the pin, and write $S_t := (\tau\sigma_{\max}/\bar{\sigma}_t)^2$. The preserved region stays above the floor, $\sigma_t^{(2)} \geq \sigma_{\min}$, if and only if*

$$\gamma^2 \leq \frac{(\bar{\sigma}_t/\sigma_{\min})^2 - 1 + f}{f},$$

and the edited region stays within the trained range, $\sigma_t^{(1)} \leq \tau\sigma_{\max}$, if and only if

$$\gamma^2 \leq \frac{S_t(1-f)}{1 - S_t f} \quad \textit{when } S_t f < 1,$$

the latter holding for every γ when $S_t f \geq 1$. We write $\gamma_{\mathrm{cap}}(f, \bar{\sigma}_t)$ for the largest γ satisfying both.

Proof. Substitute $\gamma_2^2 = 1/(f\gamma^2 + 1 - f)$ and $\gamma_1^2 = \gamma^2\gamma_2^2$ into each inequality and clear the positive denominator $f\gamma^2 + 1 - f$; both become linear in γ^2. □

Proposition F.3 (**Saturation**). *Under the pin, $\lim_{\gamma\to\infty} \gamma_1 = 1/\sqrt{f}$ and $\lim_{\gamma\to\infty} \gamma_2 = 0$.*

Proof. From equation 3, $\gamma_1^2 = \gamma^2/(f\gamma^2 + 1 - f) \to 1/f$ and $\gamma_2^2 = \gamma_1^2/\gamma^2 \to 0$. □

The ratio is effectively a knob on the preserved region alone, since $\sigma^{(1)}$ approaches its ceiling quickly while $\sigma^{(2)}$ continues toward zero. The ceiling depends on the mask area, and it is larger for smaller masks. Since the denoiser is trained on a bounded range of levels, the operative out-of-distribution constraint is depending on the mask area.

Let $\tau\sigma_{\max}$ be the largest level seen in training. By saturation described above, the region variance reaches $\bar{\sigma}_t/\sqrt{f}$, which exceeds $\tau\sigma_{\max}$ whenever $\bar{\sigma}_t > \tau\sigma_{\max}\sqrt{f}$, which can occur only when $f < 1/\tau^2$. Capping γ to stay in range does not help, because the pin then forces $\gamma_2^2 \geq (1-\tau^2 f)/(1-f)$ and the preserved region is barely preserved. The remedy is to start the chain lower.

Corollary F.4 (**Start level**). *The second condition of Proposition F.2 is automatically satisfied at every step with $\bar{\sigma}_t \leq \tau\sigma_{\max}\sqrt{f}$. A chain started at any $\bar{\sigma}'_T \leq \tau\sigma_{\max}\sqrt{f}$ therefore admits every ratio $\gamma \geq 1$.*

Proof. By Proposition F.3, $\gamma_1\bar{\sigma}_t \leq \bar{\sigma}_t/\sqrt{f}$ for every γ, and this is at most $\tau\sigma_{\max}$ when $\bar{\sigma}_t \leq \tau\sigma_{\max}\sqrt{f}$. Every level of the chain is at or below its start, so the condition holds throughout if it holds at $\bar{\sigma}'_T$. □

Nothing is lost by starting lower. At $\bar{\sigma}'_T$ the masked region sits at $\gamma_1\bar{\sigma}'_T \to \tau\sigma_{\max}$, the top of the trained range, so it is regenerated as completely as from $\sigma_{\max}$.

F.4 SAMPLING DETAILS

Given an observation and a mask, we take the mean path $\tilde{B}_t = \bar{\sigma}_t^2$, choose the start level follows Corollary F.4 which is $\bar{\sigma}'_T \leq \tau\sigma_{\max}\sqrt{f}$, fix a single γ for the whole chain, form $B_t^{(g)}$ by the pin, initialize $x_T = x^{\mathrm{obs}} + \sqrt{B_T^{(g)}}z$, and run the reverse chain of § 4.3. There is no replacement step, the known coordinates are never pasted back, and *preservation is carried by the path.*

G SHAPE ALLOCATION INSTEAD OF THE RATE

Section 5 places the spatial allocation in the rate, where $B_t^{(g)} \sim \mathrm{Ga}\left(A_t, \frac{b_t}{\gamma_g^2}\right)$. The alternative way is to place it in the shape, which is the direct spatial analogue of the shared rate construction of § 2. We record it because it is the natural first attempt, and because it meets the same obstacle of § 3 again.

Table 4: Sampling-time κ on FFHQ box inpainting, one checkpoint trained at $\kappa = 5$.

κ	5	10	20	40	80	320	∞
RSD$[B_T]$	0.447	0.316	0.224	0.158	0.112	0.056	0
FID, NFE 4	27.77	27.62	27.58	27.57	27.56	27.59	27.63
FID, NFE 64	21.62	21.60	21.57	21.55	21.55	21.55	21.54

G.1 Construction

Give each region its own shape $A_t^{(g)}$ under the rate b_t, which is shared across the regions, and draw the group processes independently. Writing $S_t = \sum_g B_t^{(g)}$, the shared rate gives the spatial Gamma-Dirichlet duality:

$$S_t \sim \mathrm{Ga}\left(\sum_g A_t^{(g)}, b_t\right), \quad \left(\frac{B_t^{(g)}}{S_t}\right)_g \sim \mathrm{Dir}\left(A_t^{(1)}, \cdots, A_t^{(G)}\right), \quad S_t \perp \left(\frac{B_t^{(g)}}{S_t}\right)_g,$$

so the total variance budget and its spatial allocation are independent, exactly as the total and the temporal allocation were in § C.3. In this formulation, two properties follow that the rate route does not provide. First, the allocation becomes a latent variable. Being a Dirichlet factor independent of the total, it can be sampled, inferred, or amortized on a conditioning signal. Second, the rank-one term of Theorem C.2 becomes rank G. The removed fractions are now independent across regions, so the third term of Theorem C.2 is $\sum_g c_3^{(g)} \rho^{(g)} (\rho^{(g)})^\top$, with $\rho^{(g)}$ the residual restricted to region g and extended by zero. These are supported on disjoint coordinates and orthogonal.

Writing the allocation as $A_t^{(g)} = A_t G s_g$ with $\sum_g s_g = 1$ and matching the schedule in the pixel-weighted sense gives

$$b_t = \frac{1}{\bar{\sigma}_t^2} \sum_g \frac{d_g}{d} A_t^{(g)},$$

reduces to the scalar pin and matters when they are unequal, as in editing with a small mask. Without the weighting, changing the allocation would silently change the total level and confound the effect of allocation with a departure from the schedule.

G.2 Why it is not adopted

The two allocation methods are equivalent on the mean path. Under the rate allocation, the level of region g is $\gamma_g^2 \bar{\sigma}_t^2$ and under the shape allocation, it is $\mathbb{E}[B_t^{(g)}] = A_t G s_g / b_t = G s_g \bar{\sigma}_t^2$. Because the mean path sets $\mathbb{E}[B_t] = A_t / b_t = \bar{\sigma}_t^2$. Setting $\gamma_g^2 = G s_g$ makes the two identical, and the pin $\sum_g (d_g/d) \gamma_g^2 = 1$ (Eq. 3) becomes $\sum_g (d_g/d) G s_g = 1$. The difference is when the path is actually drawn.

Under the shape allocation method, a region with a small share has a small shape, $A_t^{(g)} = A_t G s_g$, and the condition $A_1 > 1$ of § 3 now applies separately to every region. The region receiving less variance is pushed toward the collapse of Fig. 2. (b). The allocation is bounded below by the same inequality that governed the scalar path. Under the rate route the shape is A_t for every region and every γ, so the condition is imposed once and the training draws are well behaved, including the inpainting endpoint at which the preserved region sits at $\sigma_{\min}$ by design. This is the reason § 5 places the allocation in the rate.

H More results

H.1 Hyperparameters

We report hyperparamters for our experiments at Table. 5, 6, and 7.

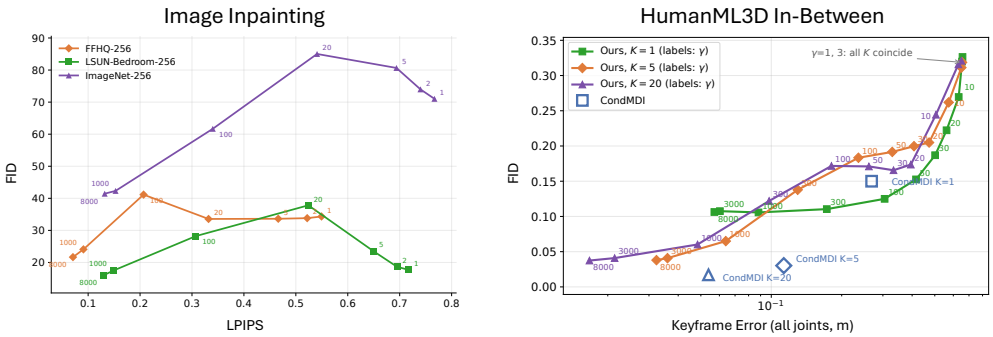

Figure 5: γ **sweep for image inpainting (left) and motion in-between (right).** $\gamma = 8000$ preserves outside contents while $\gamma = 1$ is pure generation. As γ decreases, both LPIPS and Keyframe error constantly increase. Note that at $\gamma = 1$ all K conditions merge into same unconditional generation points for motion in-between.

Table 5: Hyperparameters. The two settings share the construction and differ only where marked.

		Image (ADM, 256^2)	Motion (ACMDM)
Noise-level			
level distribution		$\ln \sigma \sim \mathcal{N}(P_{\text{mean}}, P_{\text{std}}^2)$	$t \sim \mathcal{U}(0, 1)$
P_{mean}, P_{std}		-1.2, 1.2	–
$\sigma_{\min}$, $\sigma_{\max}$		0.01, 80	0.002, 80
Variance path (TVR)			
shape at $\sigma_{\max}$	$\kappa = A_T$	5	20
shape exponent	η	0.06	0.06
Region law			
homogeneous probability	p_{uni}	0.3	0.3
endpoint probability	p_{cor}	0.1	0.1
otherwise	γ	log-uniform on $[1, \gamma_{\text{cap}}(f, \bar{\sigma}_t)]$	
soft ceiling inside γ_{cap} γ	γ_{hi}	10^4	10^3
trained tolerance	τ	2	–
area fraction	f	log-uniform on $[0.02, 0.85]$	–
complement probability	p_{inv}	0.15	–
mask families		rect / blob / stroke	4 temporal families
Optimization			
optimizer		Adam, $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, no weight decay	Adam, $\beta = (0.9, 0.99)$, $\epsilon = 10^{-8}$, weight decay 10^{-5}
learning rate		10^{-4}	10^{-4}
batch size		32 (FFHQ), 16 (LSUN, ImageNet)	64
steps		300K (FFHQ, LSUN), 800K (ImageNet)	100K
warmup / grad. clip		1000 / 1.0	– / –
EMA		0.9999	0.9999
LoRA rank		16	16

H.2 Motion

Table 8 compares single-keyframe conditioning with the four-consecutive-frame conditioning (†) that our latent model requires. The additional frames lower every baseline's keyframe error and, for MDM and OmniControl at $K \geq 5$, improve FID as well. We report the † rows in the main text so that all methods receive the same information. On keyframe error, this is the more favorable for the baselines. Fig 6 conditions on the prefix and suffix keyframes and varies the text prompt. VPD follows the text in the in-betweening segment while satisfying the keyframes at both ends. Fig 8 shows what happens when the two conditions conflict. With keyframes pinned at both ends and the prompt "turns around", the keyframes win and the text is ignored, which is expected as we assign low variance budget at keyframes making hard to change. With a compatible prompt such as "jumps" or "zigzag", both are satisfied. When only a prefix is pinned, the future is unconstrained and the model follows "turns around" as well.

H.3 Image

Figure 10 shows the qualitative results compared to DDNM and DPS with VPD. As DDNM and DPS are training-free methods, we use our finetuned ADM model with 100 NFE. Figures 12-14 give further inpainting results on FFHQ, LSUN Bedroom, and ImageNet, and Figures 15-17 un-

Table 6: Per-dataset settings. Everything not listed here is shared (Table 5).

	FFHQ-256	LSUN Bed.	ImageNet-256	IN-256 (CFG)
Baseline model	ADM/P2 94M	ADM 552M	ADM 552M	ADM 552M (cond.)
Batch size	32	16	16	16
Steps	300K	300K	800K	800K
Label dim	0	0	0	1000
Label dropout	0	0	0	0.15

Table 7: Inference. The start level is the total noise put inside the region; outside stays at $\sigma_{\min}$ throughout.

		Image	Motion
Start level (inside)	σ_{in}	$80 = \sigma_{\max}$	γ-dial
Start level (outside)		$\sigma_{\min} = 0.01$	$\sigma_{\min} = 0.01$
Function evaluations	NFE	100	64
Path at inference		deterministic (mean)	deterministic (mean)
Classifier-free guidance		–	2.0
γ (editing / γ-plane)		$1 \to 8000$	$1 \to 8000$

conditional samples from the same models. Unconditional generation is the case $\gamma = 1$, which the training mixture of § 5.2 samples with probability p_{uni}. Figure 11 sweeps γ on ImageNet with a class condition. At large γ the object inside the mask is replaced and the background is untouched; as γ decreases the background is progressively regenerated as well, until at $\gamma = 1$ the image is fully regenerated from the class condition alone.

Table 8: Quantitative results of motion in-betweening under the CondMDI evaluation protocol on HumanML3D with N=4640 samples, all methods are re-evaluated. † represents that model receives four consecutive frames of keyframe.

Method	Conditioning	NFE	FID↓	R-precision (Top-3)↑	Diversity →	Foot Skate Ratio ↓	kf-err (root) ↓	kf-err (all)
Ground Truth	–	–	–	0.771	9.36	0.072	–	–
MDM + imputation	$K = 1$	1000	0.809	0.680	8.81	0.089	0.827	0.838
MDM + imputation †	$K = 1$	1000	0.865	0.700	8.47	0.087	0.800	0.808
OmniControl (root-joint condition)	$K = 1$	1000	0.906	0.724	9.05	0.064	0.011	0.168
OmniControl (root-joint condition) †	$K = 1$	1000	1.191	0.721	8.86	0.067	0.016	0.166
CondMDI	$K = 1$	1000	0.150	0.752	9.68	0.099	0.267	0.269
CondMDI†	$K = 1$	1000	0.231	0.741	9.66	0.089	0.159	0.161
VPD (ours)†	$K = 1$	64	0.106	0.773	9.37	0.080	0.042	0.057
MDM + imputation	$K = 5$	1000	3.971	0.659	8.18	0.132	0.746	0.756
MDM + imputation†	$K = 5$	1000	1.733	0.730	8.26	0.102	0.621	0.627
OmniControl (root-joint condition)	$K = 5$	1000	1.446	0.736	9.05	0.067	0.048	0.174
OmniControl (root-joint condition)†	$K = 5$	1000	1.402	0.747	9.23	0.070	0.051	0.165
CondMDI	$K = 5$	1000	0.030	0.766	9.26	0.095	0.112	0.113
CondMDI†	$K = 5$	1000	0.083	0.752	9.56	0.091	0.070	0.072
VPD (ours)†	$K = 5$	64	0.038	0.779	9.51	0.085	0.022	0.032
MDM + imputation	$K = 20$	1000	5.321	0.686	8.24	0.228	0.519	0.526
MDM + imputation†	$K = 20$	1000	0.479	0.773	8.93	0.091	0.255	0.260
OmniControl (root-joint condition)	$K = 20$	1000	2.159	0.730	8.97	0.070	0.059	0.160
OmniControl (root-joint condition)†	$K = 20$	1000	0.933	0.758	8.89	0.072	0.059	0.152
CondMDI	$K = 20$	1000	0.017	0.767	9.47	0.090	0.053	0.054
CondMDI†	$K = 20$	1000	0.022	0.765	8.80	0.090	0.035	0.036
VPD (ours)†	$K = 20$	64	0.037	0.773	9.24	0.082	0.011	0.017

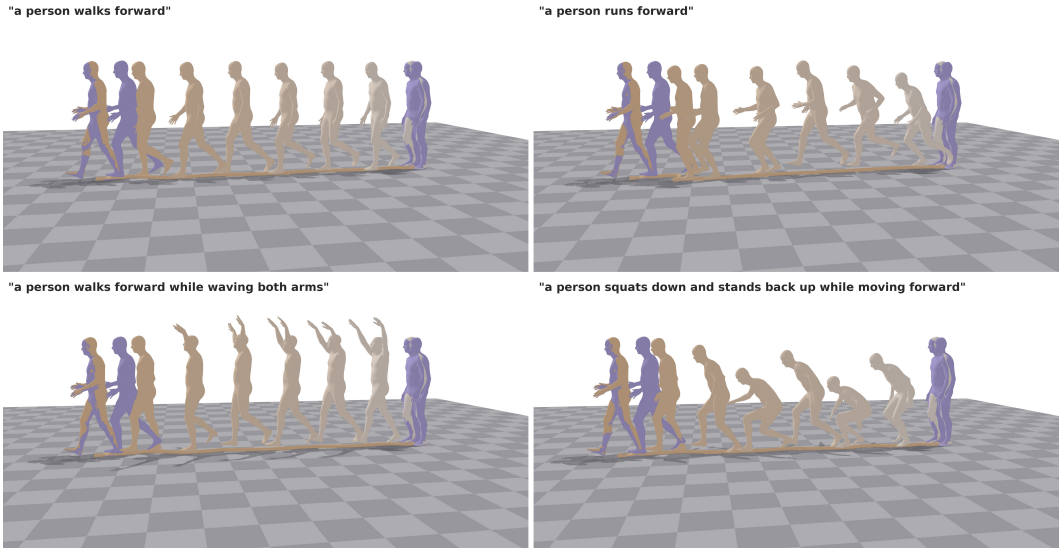

Figure 6: Different text-conditioned in-between generation with same keyframes (purple).

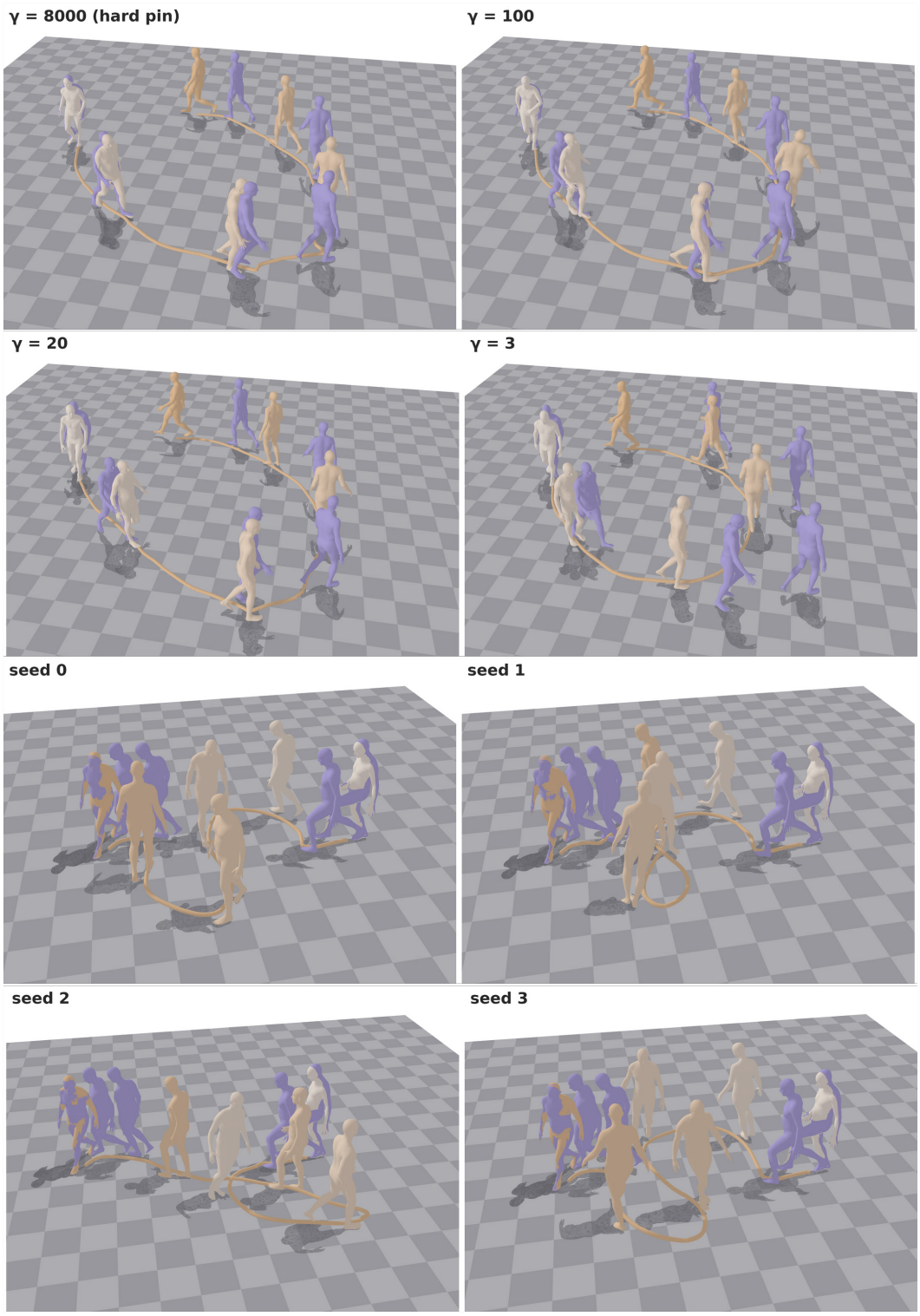

Figure 7: **Top:** the effect of different γ. Generated motion doesn't fulfill the keyframes when $\gamma = 3$. **Bottom:** diverse generation with different seed.

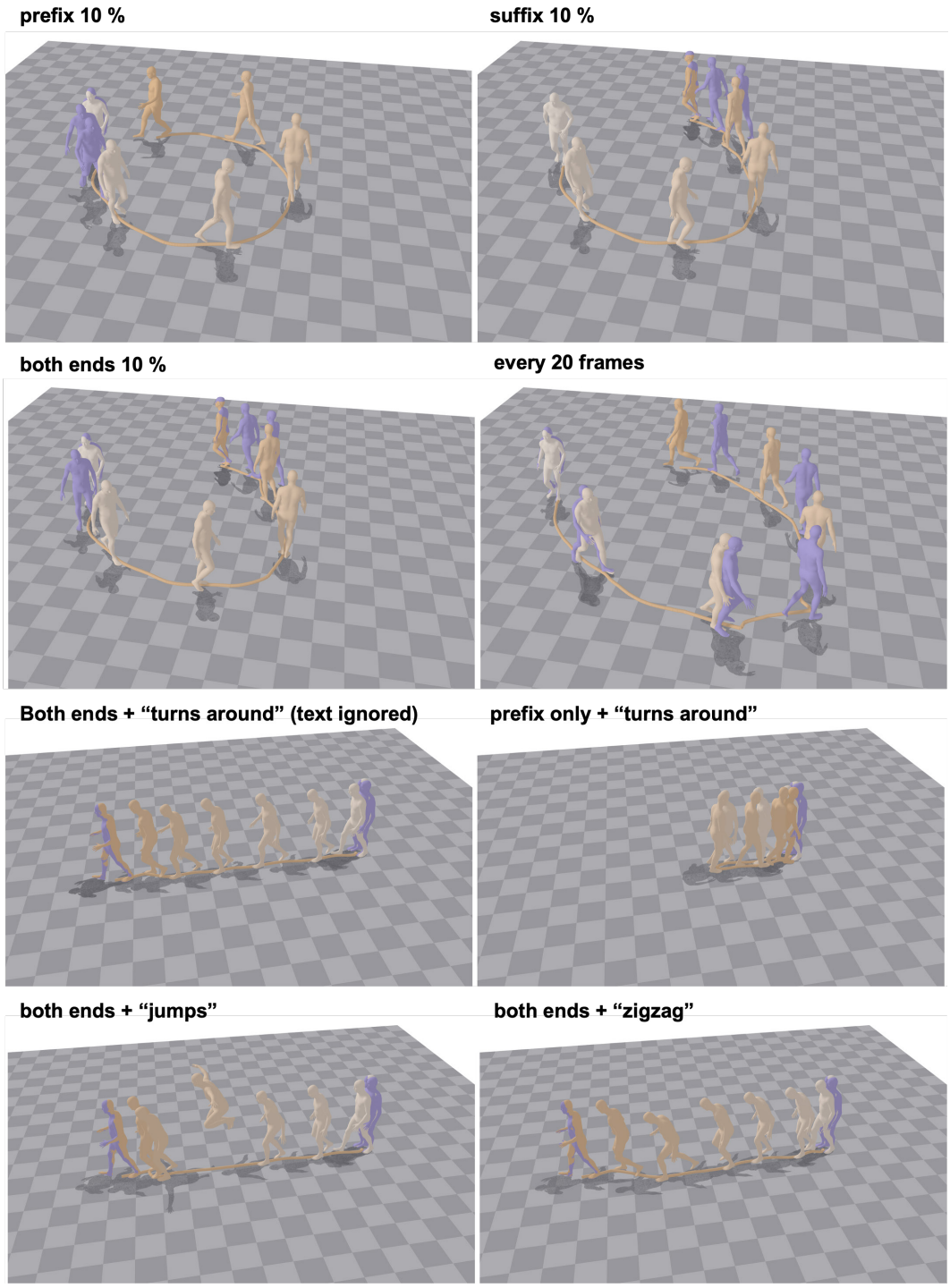

Figure 8: Purple: given keyframe poses (ground truth); orange: generated poses at the same frames, with the generated root trajectory on the floor. **Top:** Prefix, suffix, both-end, and every-20-frame pins (purple) on one clip with the same text. The pin layout alone selects prediction, lead-in, in-betweening, or trajectory following; the model and weights are identical. **Bottom:** Text vs. keyframes. Conflicting text is overridden by end pins (left), followed when the future is free (second), and realized inside the pins when compatible (e.g., jump, zigzag).

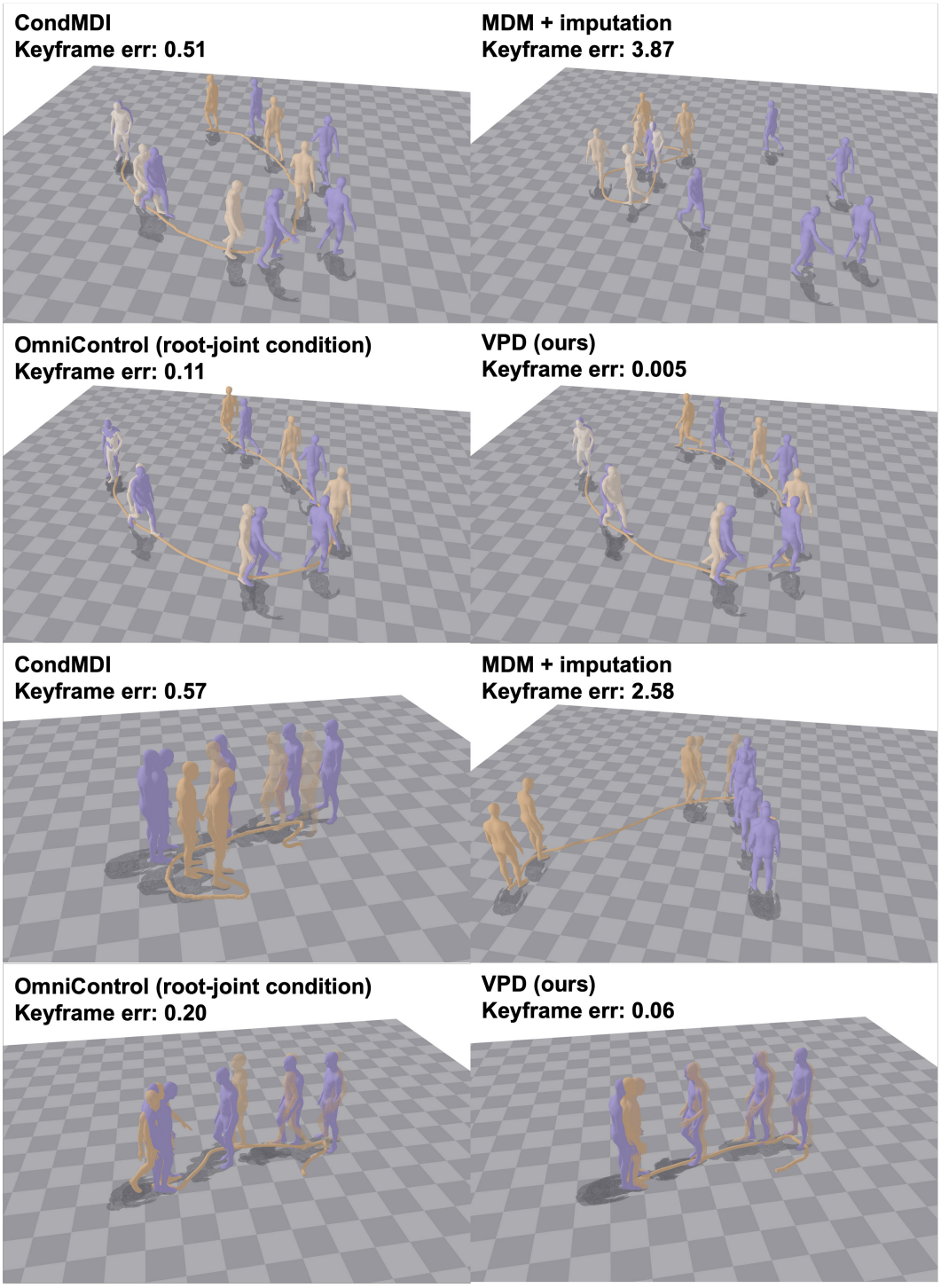

Figure 9: Qualitative in-betweening comparison. Purple: given keyframe poses (ground truth); orange: generated poses at the same frames, with the generated root trajectory on the floor. All methods receive identical keyframes; OmniControl is conditioned on their root positions only. **Top**: prompt "a person walking by making a large circle", one keyframe every 20 frames (10 keyframes). **Bottom**: prompt "a person walks in a counter clockwise circle and turns around…" with K=5 random keyframes.



Figure 10: Qualitative comparison with DDNM Wang et al. (2022), DPS Chung et al. (2024), and VPD (ours). For DDNM and DPS, we use our finetuned ADM model with 100 NFE same as VPD.



Figure 11: Gradual editing on ImageNet 256 with classifier-free guidance: horse to zebra (rows 1-2), dog to cat (rows 3-4), and bird to macaw (rows 5-6). Column 1 shows the input with the mask; columns 2-10 sweep γ from 8000 to 1. Higher γ preserves more of the region outside the mask, and $\gamma = 1$ is full regeneration. The two levels follow from γ and the mask area through the pin (Eq 3).



Figure 12: FFHQ inpaint grid with NFE=100.

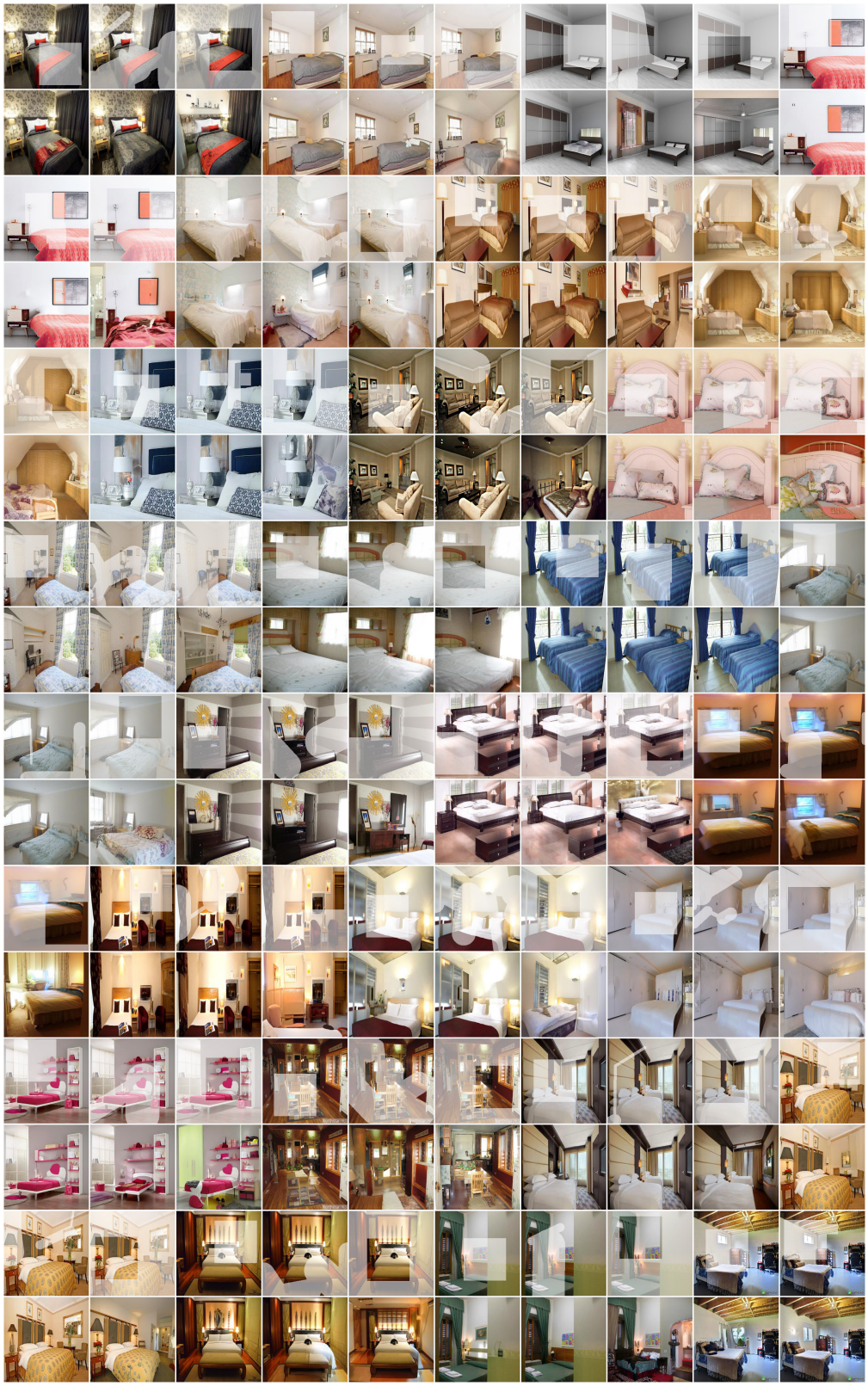

Figure 13: LSUN Bedroom inpaint grid with NFE=100.

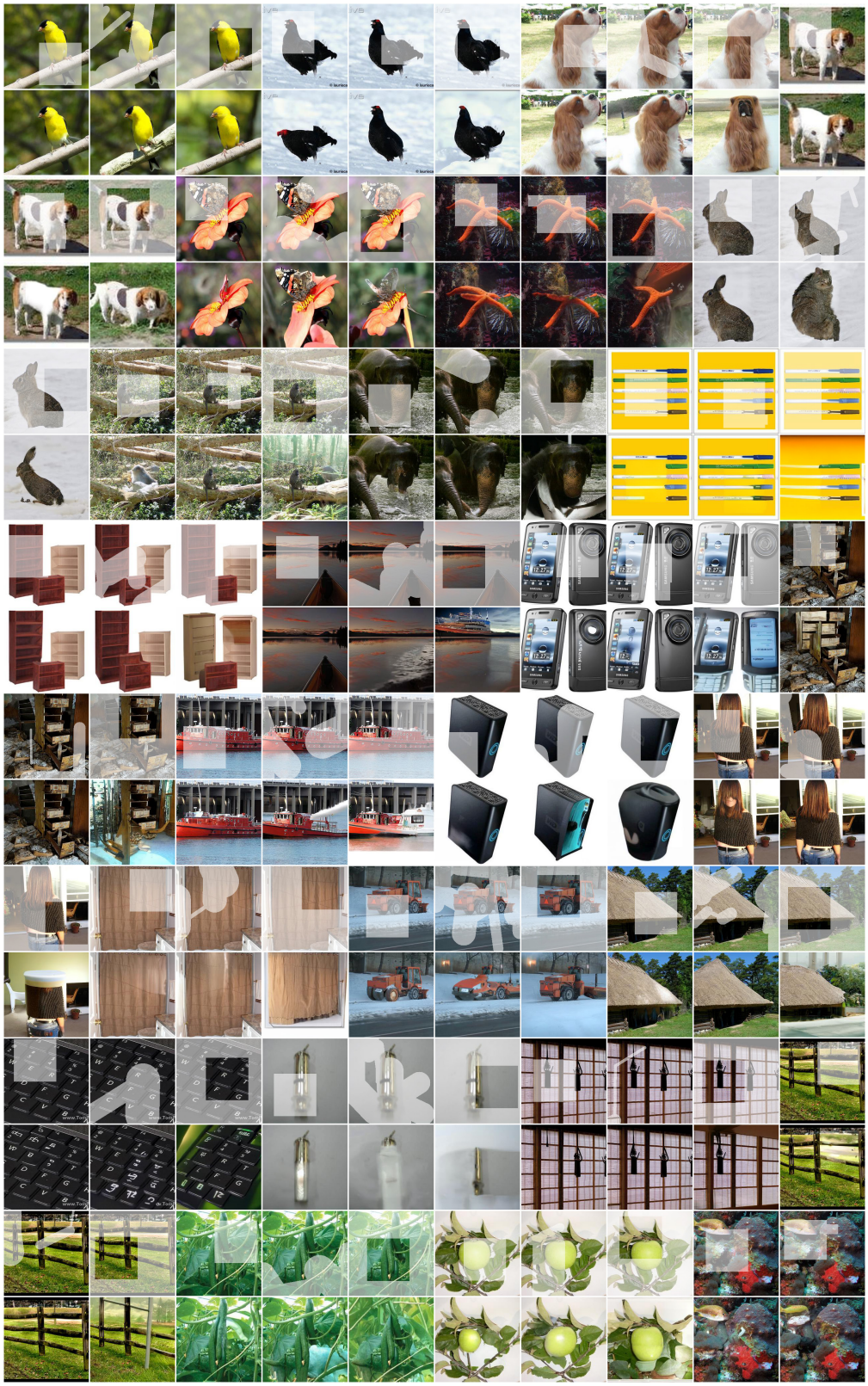

Figure 14: ImageNet inpaint grid with NFE=100.



Figure 15: FFHQ unconditional generation grid with NFE=128. Our FID: 15.55 and baseline (ADM) FID: 13.93 (NFE=128, DDPM).



Figure 16: LSUN Bedroom unconditional generation grid with NFE=128. Our FID: 3.04 and baseline (ADM) FID: 5.71 (NFE=128, DDPM).

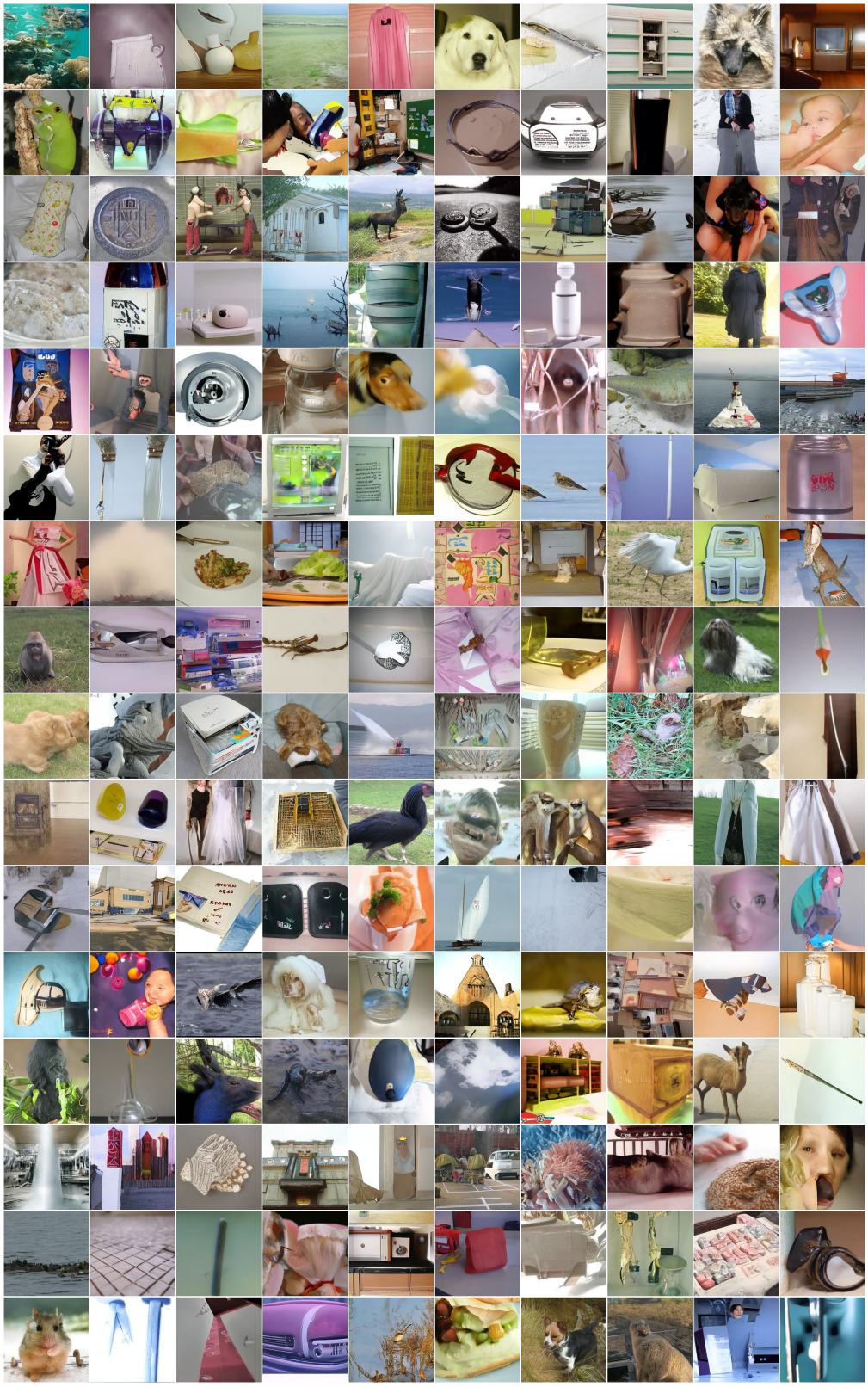

Figure 17: ImageNet unconditional generation grid with NFE=128. Our FID: 33.17 and baseline (ADM) FID: 26.42 (NFE=128, DDPM).