POSTERIOR FLOW MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Flow matching learns how to transport noise toward data, but its mean prediction does not directly report the endpoint uncertainty remaining at an intermediate state. This paper introduces *Posterior Flow Matching* (PFM), a general generative modeling framework that learns conditional velocity means and componentwise variances with a Gaussian posterior model. The same model evaluation supplies both a transport prediction and endpoint uncertainty, making posterior diagnostics available throughout generation. These learned moments support a mean-driven ordinary differential equation (ODE) and a posterior-driven stochastic differential equation (SDE), whose local noise scales adapt to the predicted uncertainty. Across image, point-cloud, molecular, and action generation, PFM improves generation quality and provides endpoint-posterior readouts throughout generation.

1 INTRODUCTION

Diffusion models (Ho et al., 2020; Song et al., 2021; Cai et al., 2025; Ma et al., 2024; Rombach et al., 2022; Hoogeboom et al., 2023; Yu et al., 2025; Zheng et al., 2026; Wang et al., 2025) and flow matching (FM) (Lipman et al., 2023; Albergo & Vanden-Eijnden, 2023; Esser et al., 2024; Liu et al., 2023; Jin et al., 2025; Klein et al., 2023) generate data by progressively transforming noise into structured outputs (Karras et al., 2022). Diffusion learns a score for reversing a noising process (Lai et al., 2025), while FM learns a velocity field along a reference path (Lipman et al., 2024; Lee et al., 2024; Kornilov et al., 2024). *These predictions guide generation but do not directly report the endpoint uncertainty remaining at an intermediate state.*

This uncertainty arises because several source-data pairs can be compatible with the current intermediate state, each prescribing a different velocity toward its data endpoint (Guo & Schwing, 2025; Zhang et al., 2025b;a; Ma et al., 2026). Standard FM regression estimates only the conditional mean of these velocity targets (Lipman et al., 2023), which supplies the transport direction but does not report how widely the compatible endpoints are dispersed. Learning this dispersion alongside the mean would reveal which parts of a developing output remain uncertain and provide local scales for stochastic perturbations, connecting uncertainty estimation directly to generation.

We realize this idea with *Posterior Flow Matching* (PFM), which extends FM regression to learn the conditional velocity mean and componentwise variance with a Gaussian posterior model. The same model evaluation supplies both the transport prediction and endpoint-posterior readouts. The velocity mean defines a mean-driven ordinary differential equation (PFM ODE), while the velocity covariance sets the local noise scales of a posterior-driven stochastic differential equation (PFM SDE). Under either sampler, the endpoint-posterior readouts describe the evolving output and its uncertainty without a separate diagnostic model. We evaluate PFM on image generation, 3D point-cloud generation, molecular conformation generation, and action generation for robotics control. Across these tasks, PFM SDE improves generation quality over PFM ODE and other baselines. Covariance ablations further support using the learned posterior variance to guide sampling. Both samplers use the same trained posterior model and provide endpoint uncertainty estimates throughout generation.

The paper provides three key contributions. *First*, we extend flow matching regression to posterior learning by jointly fitting the conditional velocity mean and componentwise variance with one Gaussian negative log-likelihood (NLL) objective. An affine relation converts these moments into endpoint-posterior readouts at essentially no additional cost. *Second*, we derive the mean-driven ODE and posterior-driven SDE samplers from these moments. Both provably preserve the reference marginals under exact dynamics, and the SDE dissipates the relative entropy of density perturbations, weighting each coordinate by its predicted velocity variance. *Third*, we show that this

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.

single formulation transfers across image, point-cloud, molecular, and action generation, where it improves generation quality while its posterior diagnostics quantitatively describe endpoint uncertainty throughout generation. Through this work, we call for treating uncertainty as an integral part of generation rather than a post-hoc diagnostic, laying a principled foundation for reliable and interpretable generative models.

2 Related Work

Our work connects flow matching and diffusion models with posterior uncertainty estimation and variance-aware generation.

Flow matching and diffusion models. Diffusion probabilistic models learn to reverse gradual data corruption through denoising objectives (Ho et al., 2020). Score-based modeling connects reverse SDEs to probability-flow ODEs, enabling stochastic and deterministic sampling within one model (Song et al., 2021). FM learns velocity fields by regressing conditional targets along reference paths (Lipman et al., 2023; Tong et al., 2024; Pooladian et al., 2023; Albergo et al., 2024; Ma et al., 2025). Stochastic interpolants place transport and diffusion in a common framework, where a density path admits dynamics with adjustable diffusion (Albergo et al., 2025; Ma et al., 2024). PFM builds on this correspondence: its learned posterior mean supplies transport, and its covariance selects local diffusion scales within the compatible family.

Uncertainty quantification in generative models. BayesDiff propagates uncertainty from a last-layer Laplace approximation through diffusion (Kou et al., 2024), while Generative Uncertainty combines Bayesian model uncertainty with a semantic likelihood for sample assessment (Jazbec et al., 2025). Posterior moments can also be obtained from denoiser derivatives (Manor & Michaeli, 2024) or velocity derivatives in flow models (Lv et al., 2026). UA-Flow learns heteroscedastic velocity uncertainty with supervision corrected toward a marginal-velocity objective and propagates it for assessment and guidance (Han et al., 2026); Flowing with Confidence propagates input-dependent network noise along trajectories (de Kruiff et al., 2026). PFM instead directly fits the conditional moments of sampled FM velocity targets. The affine endpoint relation then gives endpoint-posterior readouts at the current state, without propagating uncertainty along the trajectory or computing model derivatives to read out the variance.

Posterior modeling and variance-aware generation. Improved DDPM learns reverse-process variances (Nichol & Dhariwal, 2021), and Analytic-DPM estimates optimal reverse variances from a pretrained score model (Bao et al., 2022). These describe reverse transitions, whereas PFM models conditional velocity and endpoint dispersion. Pixel-wise aleatoric uncertainty has been used to guide diffusion sampling (De Vita & Belagiannis, 2025). Posterior Augmented Flow Matching reduces training-target variance through posterior-weighted candidate averages (Stoica et al., 2026). PFM preserves the sampled targets and fits their conditional moments with one Gaussian NLL. The same covariance supplies endpoint diagnostics and the SDE diffusion tensor, paired with the score and divergence corrections required for compatible transport.

3 Preliminaries

We introduce the reference targets used for learning, the score-based view of generation, and the correspondence used to construct compatible samplers. Throughout, blue highlights velocity targets and posterior means; orange highlights posterior variance parameters and diffusion coefficients.

3.1 Reference paths and flow matching

We consider continuous states in $\mathbb{R}^d$, with generation time t increasing from noise at $t=0$ to data at $t=1$. Uppercase letters denote random variables; Table 6 in Appendix A.1 collects the notation used throughout. Let $X_0 \sim p_0$ and $X_1 \sim p_1 = p_{\text{data}}$ be independent, and define the reference path and its velocity target by

$$X_t = (1-t)X_0 + tX_1, \qquad V = X_1 - X_0, \qquad Z = (X_0, X_1). \tag{1}$$

Each pair Z fixes an interpolation, while the marginal velocity averages compatible targets:

$$v(x,t) = \mathbb{E}[V \mid X_t = x], \qquad \partial_t p_t = -\nabla_x \cdot (v p_t), \tag{2}$$

where p_t is the law of X_t. When this law has a differentiable positive density, its score is $s(x,t) = \nabla_x \log p_t(x)$. A sampler Y_t with $dY_t = v(Y_t,t)dt$ and $Y_0 \sim p_0$ follows these marginals when the flow and density evolution are well posed.

Conditional flow matching (CFM) (Lipman et al., 2023; Albergo & Vanden-Eijnden, 2023) trains a neural velocity v_η using targets conditioned on Z:

$$\mathcal{L}_{\mathrm{CFM}}(\eta) = \frac{1}{2}\mathbb{E}_{t,X_0,X_1}\|V - v_\eta(X_t,t)\|^2, \qquad t \sim r, \quad t < 1. \tag{3}$$

The time law r is fixed independently of the parameters. This sample-based loss differs from regression to v by a parameter-independent conditional-variance term, so both objectives have the same gradient (Lipman et al., 2024) under the differentiation conditions in Appendix A.2. This regression identifies the conditional mean needed for transport, linking the sample-based objective to the marginal velocity field that drives generation; PFM extends this regression by also learning the conditional variance of the same targets. In conditional generation, C specifies a class, observation, or reference structure held fixed within a trajectory, whereas Z specifies the sampled reference interpolation; thus, PFM conditions on C while its learned variance captures variation across paths Z compatible with (X_t, C). We omit C from the main formulation and present the complete conditional treatment in Appendix E.

3.2 Score-based diffusion

Differently from FM, diffusion starts from a prescribed noising process (Ho et al., 2020; Song et al., 2021). Writing noising time as $\tau = 1 - t$, let

$$d\bar{X}_\tau = f(\bar{X}_\tau, \tau)d\tau + g(\tau)dW_\tau, \qquad \text{and} \qquad \bar{X}_0 \sim p_{\mathrm{data}}, \tag{4}$$

where f is the noising drift, g a scalar noise coefficient, and W standard d-dimensional Brownian motion. Let $\bar{p}_\tau$ be the law of $\bar{X}_\tau$ and $\bar{s}(x,\tau) = \nabla_x \log \bar{p}_\tau(x)$ its score. At generation time, a network $s_\phi(x,t)$ estimates $\bar{s}(x, 1-t)$, through denoising score matching with a known perturbation kernel (Song et al., 2021; Lai et al., 2025) (see A.3). In increasing generation time, the reverse SDE and probability-flow ODE are (Song et al., 2021)

$$dY_t = [-f(Y_t, 1-t) + g(1-t)^2 \bar{s}(Y_t, 1-t)]dt + g(1-t)dW_t, \tag{5}$$

$$dY_t = [-f(Y_t, 1-t) + \tfrac{1}{2}g(1-t)^2 \bar{s}(Y_t, 1-t)]dt. \tag{6}$$

They share the noising-path marginals with the exact score and initialization from the terminal noising law, while in practice, both use s_ϕ. This ODE–SDE pairing is a special case of the correspondence below, which also permits state-dependent diffusion. The parameters η, ϕ, and θ index FM, diffusion, and PFM, respectively.

3.3 A shared ODE–SDE correspondence

To compare deterministic and stochastic generation within a common framework, we first determine when an ODE and an SDE induce the same evolving density. This correspondence will allow PFM to use its learned posterior variance to control stochasticity while preserving the reference marginals under exact dynamics. We write the stochastic sampler as an Itô SDE,

$$dY_t = \underbrace{b(Y_t,t)dt}_{\text{drift increment}} + \underbrace{B(Y_t,t)dW_t}_{\text{Brownian noise increment}}.$$

Here, b is the full drift and B maps a Brownian increment into state-space noise, so the Itô form separates directed motion from a zero-mean random perturbation. To quantify this perturbation, define the diffusion tensor $a = BB^\top$, whose matrix divergence is $(\nabla_x \cdot a)_i = \sum_j \partial_{x_j} a_{ij}$. The tensor a is a local covariance rate, so an increment of duration h has conditional covariance $h\,a(x,t) + o(h)$.

The remaining question is how the drift b must change when a is introduced so that the sampler continues to follow a prescribed density path p_t. Writing $s = \nabla_x \log p_t$ for its score and rewriting the Fokker–Planck equation as a continuity equation gives (Song et al., 2021; Albergo et al., 2025)

$$v = b - \frac{1}{2}(\nabla_x \cdot a + a\,s), \qquad b = \underbrace{v}_{\text{ODE velocity}} + \underbrace{\frac{1}{2}(\nabla_x \cdot a + a\,s)}_{\text{drift correction}}. \tag{7}$$

Equation 7 provides the desired correspondence: after choosing a positive-semidefinite a, the correction added to v compensates for the probability flux induced by the stochastic perturbation. Thus, a density path transported by v also admits an SDE with the corresponding drift; with smooth coefficients, suitable boundary behavior, unique density evolution, and matching initialization, the two dynamics have the same one-time marginals. Appendix A.5 gives the conditions and derivation.

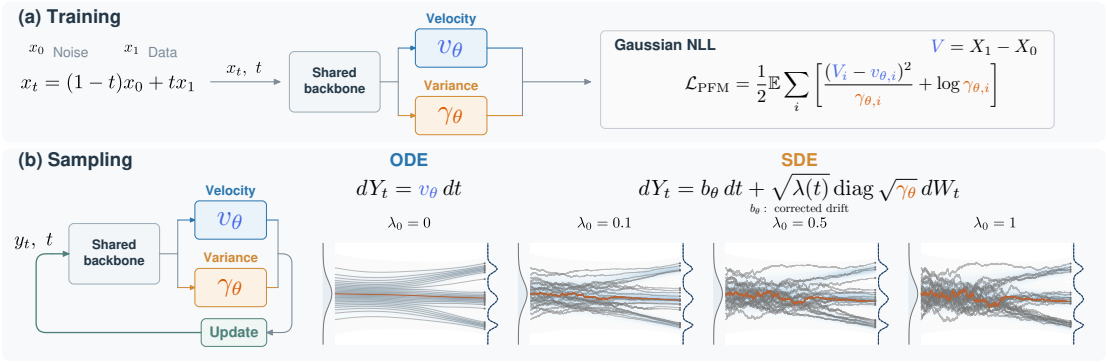


Figure 1: **Posterior Flow Matching. (a) Training.** A shared network learns the velocity mean v_θ and log variance ℓ_θ by Gaussian NLL (Eq. 13). The diagram displays $\gamma_\theta = \exp \ell_\theta$. **(b) Sampling.** These outputs define the PFM ODE and SDE (Eqs. 16 and 19), evaluated at (Y_t, t). Trajectories share a checkpoint and initial particles, with $\lambda(t) = \lambda_0(1-t)$. See Figure 8 for the full overview.

Diffusion provides a familiar instance of this construction. For the reverse SDE in Eq. 5, $a = g(1-t)^2 I$ is independent of the state, so its divergence vanishes and Eq. 7 recovers the probability-flow ODE in Eq. 6. For FM (and PFM), using this correspondence requires a score in addition to the learned velocity. To obtain it without training another model, specialize Eq. 1 to $X_0 \sim \mathcal{N}(0, I)$; the resulting Gaussian kernel gives (Lipman et al., 2024)

$$X_t \mid X_1 \sim \mathcal{N}(tX_1, (1-t)^2 I), \qquad s(x,t) = \frac{t\, v(x,t) - x}{1-t}, \quad t < 1. \tag{8}$$

A velocity prediction therefore supplies the derived score $\widehat{s}_\eta = (t\, v_\eta - x)/(1-t)$ needed by the drift correction. Standard FM can then choose a prescribed isotropic tensor $a = \kappa(t) I$, with $\kappa(t) \geq 0$, which assigns the same covariance rate in every direction (Albergo et al., 2025). Its drift follows from Eq. 7, its noise amplitude is $\sqrt{\kappa(t)}$, and $\kappa \equiv 0$ recovers the ODE. This construction provides a stochastic FM sampler but does not use posterior information to adapt its noise across states and coordinates; PFM will instead choose the diffusion tensor from learned posterior moments in Section 4.2. Appendices A.4 and A.6 derive the score identity and baseline samplers.

4 Method

PFM learns the conditional velocity mean and variance to describe the data endpoints compatible with an intermediate state (Figure 1, with the full overview in Figure 8). We first connect posterior fitting to endpoint information and define a Gaussian learning objective. The learned moments then determine ODE and SDE sampling and provide endpoint-posterior readouts for diagnostics.

4.1 Learning conditional velocity posteriors

The FM mean supplies a transport direction, but leaves the spread of compatible endpoints unspecified. At a fixed time, this remaining uncertainty is described by $p(X_1 \mid X_t)$ and its conditional entropy. To formalize how much endpoint information the state retains, mutual information (MI) relates it to the information available in the state through $I(X_t; X_1) = H(X_1) - H(X_1 \mid X_t)$, where H denotes differential entropy. For an endpoint-posterior model q_θ^x, the expected log-likelihood gives the variational bound (Barber & Agakov, 2003; Poole et al., 2019)

$$I(X_t; X_1) \geq H(X_1) + \mathbb{E} \log q_\theta^x(X_1 \mid X_t, t). \tag{9}$$

Because the reference path fixes the MI, minimizing the posterior NLL tightens this bound by reducing its gap, the expected Kullback–Leibler (KL) divergence, also called relative entropy, from the true posterior to q_θ^x. Appendix B.1 derives this identity under finite-entropy and integrability conditions. For external generation inputs C, the same argument uses conditional mutual information (CMI), developed in Appendix E.2.

To fit this posterior while retaining the FM transport parameterization, use the reference-path identity

$$X_1 = X_t + (1-t)V. \tag{10}$$

At fixed (x, t) with $t < 1$, this invertible affine map converts a velocity posterior into an endpoint posterior. Their NLLs differ only by $d \log(1-t)$, independently of θ (Appendix B.2). We can therefore fit the endpoint posterior in velocity coordinates, where its mean directly supplies transport.

To estimate both moments in one model evaluation, a shared backbone predicts the velocity mean and log variance,

$$(v_\theta, \ell_\theta) = F_\theta(x, t), \qquad \gamma_\theta = \exp \ell_\theta, \qquad \Sigma_\theta = \operatorname{diag} \gamma_\theta, \tag{11}$$

which define a diagonal-Gaussian approximation to the conditional velocity posterior:

$$q_\theta(v \mid x, t) = \mathcal{N}(v; v_\theta(x, t), \Sigma_\theta(x, t)). \tag{12}$$

Both heads depend on the full state. Up to a constant, the training objective is

$$\begin{aligned} \mathcal{L}_{\mathrm{PFM}}(\theta) &= \frac{1}{2}\mathbb{E}\left[(V - v_\theta)^\top {\Sigma_\theta}^{-1}(V - v_\theta) + \log\det \Sigma_\theta\right] \\ &= \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{d}\left(\frac{(V_i - v_{\theta,i})^2}{\gamma_{\theta,i}} + \log \gamma_{\theta,i}\right)\right]. \end{aligned} \tag{13}$$

The expectation uses the same reference pairs and time law as FM, with outputs evaluated at (X_t, t). The weighted residual and log-variance terms jointly fit both moments. The proposition below asks whether this objective preserves the FM transport target while recovering conditional dispersion.

Proposition 4.1 (Conditional moment recovery; proof in Appendix B.4)**.** *Fix (x, t) and suppose $V \mid X_t = x$ has finite second moments and positive componentwise variances. The expected NLL of $q_\theta(\cdot \mid x, t)$ is uniquely minimized over its mean and positive diagonal variances at*

$$v_\theta(x, t) = v(x, t), \qquad \gamma_{\theta,i}(x, t) = \operatorname{Var}(V_i \mid X_t = x). \tag{14}$$

Appendix B.4 proves the result. Fixing $\Sigma_\theta = I$ recovers the FM objective in Eq. 3, while learning it adds conditional variance estimation without changing the population mean target. The shared backbone and both heads are trained jointly by Algorithm 1.

To make the learned uncertainty useful for prediction and diagnostics, the affine map in Eq. 10 converts these velocity moments into endpoint-posterior readouts:

$$\mu_\theta(x, t) = x + (1 - t)v_\theta(x, t), \qquad {\Sigma_\theta}^x(x, t) = (1 - t)^2 \Sigma_\theta(x, t). \tag{15}$$

These endpoint-posterior readouts require no additional head. Appendix B.2 derives their moment transformation, and Appendix F develops the alternative of predicting the endpoint mean directly.

4.2 Posterior-driven ODE and SDE sampling

PFM uses its learned mean and variance to separate transport from adaptive noise. Both samplers start from $Y_0 \sim \mathcal{N}(0, I)$. The learned mean defines the PFM ODE and the derived score of Eq. 8:

$$dY_t = v_\theta(Y_t, t)dt, \qquad \widehat{s}_\theta(x, t) = \frac{t\, v_\theta(x, t) - x}{1 - t}. \tag{16}$$

To scale stochastic perturbations by learned uncertainty, we choose a diffusion covariance rate that is isotropic after normalization by the local velocity posterior standard deviations:

$${\Sigma_\theta}^{-1/2} a_\theta {\Sigma_\theta}^{-1/2} = \lambda(t) I, \qquad \lambda(t) \geq 0. \tag{17}$$

This gives the diffusion tensor and a Brownian noise coefficient

$$a_\theta = \lambda(t)\Sigma_\theta, \qquad B_\theta = \sqrt{\lambda(t)}\, \operatorname{diag} \sqrt{\gamma_\theta}. \tag{18}$$

Thus posterior variances set relative noise strengths across states and coordinates, while the inference-time schedule λ controls the overall rate. Appendix C.1 derives the corresponding increment covariance and its locally normalized form.

Adding diffusion changes density evolution, so it must be paired with a drift correction. The correspondence in Eq. 7 gives the PFM SDE

$$dY_t = \left[v_\theta + \frac{1}{2}\nabla_x \cdot a_\theta + \frac{1}{2} a_\theta \widehat{s}_\theta\right](Y_t, t)dt + B_\theta(Y_t, t)dW_t. \tag{19}$$

The score correction compensates for diffusion across density gradients, and the divergence correction accounts for spatially varying covariance. At the population mean in Proposition 4.1, the

derived score is exact and these dynamics preserve the reference marginals under the conditions in Appendix A.5. Setting $\lambda = 0$ removes diffusion and both corrections, recovering the PFM ODE. Appendix C.2 derives the drift, and Appendix G gives the numerical samplers.

Relative-entropy dissipation. Beyond preserving the reference marginals, the diffusion also determines how density perturbations evolve. Let p_t be the reference density and let $\widetilde{p}_t$ start from another density under the same compatible coefficients, with the score in their common drift fixed by p_t (Fontbona & Jourdain, 2016; Wibisono et al., 2017).

Proposition 4.2 (Relative-entropy dissipation; proof in Appendix C.3)**.** *On $[0, T]$ with $T < 1$, suppose both densities are positive and smooth, the displayed quantities are finite, and boundary terms vanish under integration by parts. For the SDE with $a = \lambda\Sigma_\theta$,*

$$\frac{d}{dt}\,\mathrm{KL}(\widetilde{p}_t\|p_t) = -\frac{\lambda(t)}{2}\int \widetilde{p}_t \sum_i \gamma_{\theta,i}\left(\partial_{x_i}\log\frac{\widetilde{p}_t}{p_t}\right)^2 dx \le 0. \tag{20}$$

When $\lambda \equiv 0$, both densities follow the same ODE, and their KL remains constant.

Posterior variances weight coordinatewise dissipation; KL decreases strictly when $\lambda(t) > 0$ and the integral above is positive. Appendix C.3 proves the general matrix form, and Appendix H.6 illustrates one-dimensional KL evolution and cumulative dissipation.

Relation to the standard generators. To isolate the role of the learned variance, Table 1 compares FM, diffusion, and PFM under the shared ODE–SDE correspondence. FM and PFM share the reference path and population velocity target. PFM additionally learns the variance that supplies local diffusion scales and endpoint diagnostics. Appendix A.6 derives the baseline entries, and Appendix H.6 visualizes the dynamics.

Table 1: Predicted fields, generation dynamics, and posterior moments. Fields use (Y_t, t), with f, g at noising time $1-t$. The tensor a is the Brownian covariance rate. Semicolons separate velocity and endpoint readouts. Baseline endpoint means follow Eq. 33, with † denoting a Gaussian perturbation kernel. PFM readouts follow Eq. 15. Appendix A.6 gives the reference paths and derivations.

Model / predicted fields	Sampler	Generation dynamics	a	Moment readouts
FM (v_η)	ODE	$dY_t = v_\eta\,dt$	0	v_η; μ_η
FM (v_η)	SDE	$dY_t = [v_\eta + \frac{1}{2}\kappa\widehat{s}_\eta]dt + \sqrt{\kappa}\,dW_t$	κI	v_η; μ_η
Diffusion (s_ϕ)	ODE	$dY_t = [-f + \frac{1}{2}g^2 s_\phi]dt$	0	${\mu_\phi}^\dagger$
Diffusion (s_ϕ)	SDE	$dY_t = [-f + g^2 s_\phi]dt + g\,dW_t$	$g^2 I$	${\mu_\phi}^\dagger$
PFM (v_θ, ℓ_θ)	ODE	$dY_t = v_\theta\,dt$	0	$(v_\theta, \Sigma_\theta)$; $(\mu_\theta, {\Sigma_\theta}^x)$
PFM (v_θ, ℓ_θ)	SDE	$dY_t = [v_\theta + \frac{1}{2}\nabla_x\cdot a_\theta + \frac{1}{2}a_\theta\widehat{s}_\theta]dt + B_\theta\,dW_t$	$\lambda\Sigma_\theta$	$(v_\theta, \Sigma_\theta)$; $(\mu_\theta, {\Sigma_\theta}^x)$

4.3 Posterior Diagnostics

The endpoint-posterior readouts in Eq. 15 make the evolving predictions interpretable under either sampler. At a generated state $Y_t = y$, they approximate the moments of the reference endpoint posterior $p(X_1 \mid X_t = y)$. The mean describes a predicted endpoint, and the covariance describes the dispersion of compatible data endpoints, as illustrated across domains in Figures 2–5.

To assess these readouts, we construct reference states by interpolating independent noise and known data endpoints using Eq. 1. At matched times, we compare predicted variances with mean squared prediction residuals. We also compare Gaussian NLLs at a shared predicted mean to assess the covariance fit. Appendix D explains these moment and likelihood checks, connecting posterior learning to the quantitative diagnostics in Section 5.2.

5 Experiments

Experimental Setup. We evaluate PFM on image generation, point-cloud generation, molecular generation, and action generation. Across tasks, PFM learns Gaussian velocity posteriors, with mean and covariance parameterizations adapted to each task. ImageNet (Deng et al., 2009) predicts the endpoint mean directly (Appendix F), while molecular generation (Ramakrishnan et al., 2014) uses a covariance projected onto the centered coordinate subspace (Appendix H.4). PFM ODE and PFM SDE sampling share the trained model within each task and category. Appendix H provides the core implementation settings, task-specific adaptations, and supplementary results.

Table 2: PushT control scores ↑.

Method	20 steps	50 steps
FM	0.825	0.872
Diffusion Policy	0.823	0.867
PFM ODE	0.827	0.878
PFM SDE	**0.848**	**0.914**

Table 3: Molecule generation. R/P: recall/precision.

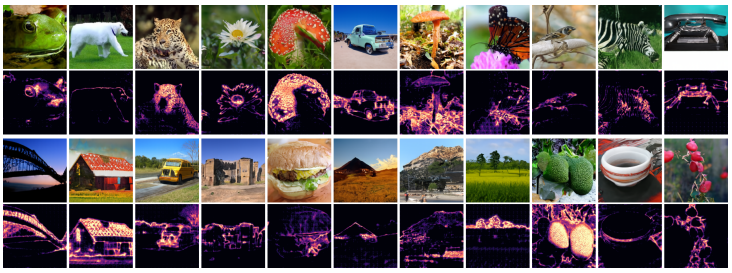
Method	Sampler	COV-R (%) ↑	COV-P (%) ↑	MAT-R (Å) ↓	MAT-P (Å) ↓
FM (ET-Flow)	ODE	96.47	94.05	0.073	0.098
FM (ET-Flow–SO(3))	ODE	95.98	92.10	0.076	0.110
PFM	ODE	96.44	93.99	0.075	0.108
PFM	SDE	**98.37**	**96.42**	**0.063**	**0.081**

Table 4: ImageNet-256.

Method	FID ↓	IS ↑
FM	25.42	63.2
Diffusion	27.38	62.1
PFM ODE	24.45	65.44
PFM SDE	**22.28**	**68.06**

Table 5: ShapeNet (CD). Ideal 1-NNA: 50%.

	Chair			Airplane		
Method	MMD×10^3 ↓	COV (%) ↑	1-NNA (%)	MMD×10^3 ↓	COV (%) ↑	1-NNA (%)
Diffusion	12.28	48.94	60.11	3.28	48.71	64.83
FM	12.04	49.23	61.74	3.26	49.34	63.19
PFM ODE	11.57	50.98	59.95	3.23	50.13	62.70
PFM SDE	**10.50**	**51.58**	**58.23**	**3.06**	**51.80**	**60.23**



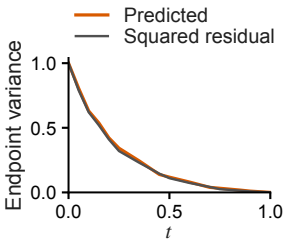


Figure 2: Left: ImageNet-256 endpoint means μ_θ, each paired with an earlier-time variance map from $\Sigma_\theta{}^x$ on the same trajectory. Maps average over channels and brighten with increasing variance. Right: comparison of predicted endpoint variance and mean squared prediction residuals over time.

5.1 Generation across Domains

Image generation. PFM generates ImageNet-256 (Deng et al., 2009) images spanning objects and scenes. In Table 4, PFM SDE improves Fréchet inception distance (FID) (Heusel et al., 2017) from 24.45 to 22.28 and inception score (IS) (Salimans et al., 2016) from 65.44 to 68.06 relative to PFM ODE, also outperforming the listed FM and diffusion baselines. Figure 2 shows representative samples; Appendix H.2 adds CIFAR-10 (Krizhevsky, 2009) results.

3D point-cloud generation. PFM generates Chair and Airplane shapes on ShapeNet (Chang et al., 2015). Using Chamfer distance (CD), we evaluate minimum matching distance (MMD), coverage (COV) (Achlioptas et al., 2018), and 1-nearest-neighbor accuracy (1-NNA) (Yang et al., 2019). MMD averages the distance from each reference shape to its nearest generated shape. In both categories, PFM SDE lowers MMD-CD, increases coverage, and moves 1-NNA-CD closer to its ideal value of 50% relative to PFM ODE and the listed baselines (Table 5). These gains span both geometric fidelity and distributional coverage. Figure 3 follows the emerging endpoint geometry.

Molecular conformation generation. PFM generates molecular conformations on GEOM-QM9 (Axelrod & Gómez-Bombarelli, 2022). PFM SDE improves all four metrics (Hassan et al., 2024) over PFM ODE and the listed reference methods (Table 3): recall and precision coverage (COV-R/P) rise to 98.37% and 96.42%, while recall and precision matching distances (MAT-R/P) fall to 0.063 and 0.081 Å. Thus the improvement combines broader conformation coverage with more accurate generated conformations. Figure 4 follows the evolving conformations.

Action generation on PushT. On PushT (Florence et al., 2022; Chi et al., 2025), PFM generates action sequences conditioned on observations for use in a control loop. PFM SDE improves the control score from 0.827 to 0.848 at 20 steps and from 0.878 to 0.914 at 50 steps, exceeding the listed FM and Diffusion Policy (Chi et al., 2025) scores at both budgets (Table 2). Figure 5 links generated actions to execution across planning calls.

5.2 Posterior Diagnostics

Visualizing endpoint predictions and variances. Alongside endpoint means μ_θ, PFM supplies covariances $\Sigma_\theta{}^x$ through Eq. 15. Image maps highlight dispersion around object boundaries, while molecular maps show atomwise variance (Figures 2 and 4). Point-cloud ellipsoids and action ellipses show endpoint standard deviations around predicted means and contract over time (Figures 3 and 5).

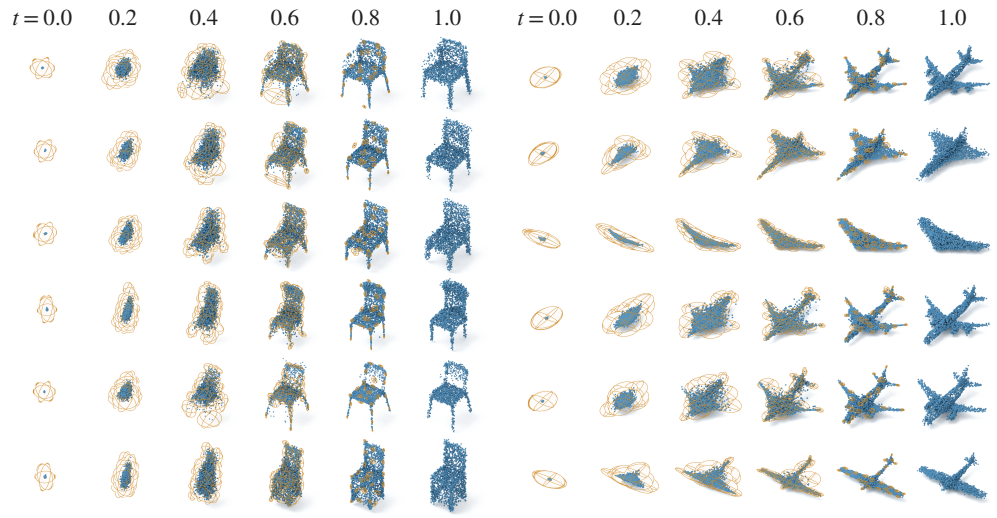


Figure 3: Chair (left) and Airplane (right) endpoint predictions over generation time. Each row follows one example. Blue points show μ_θ, and orange ellipsoids show one-standard-deviation contours from ${\Sigma_\theta}^x$ at a fixed subset of points tracked across time.

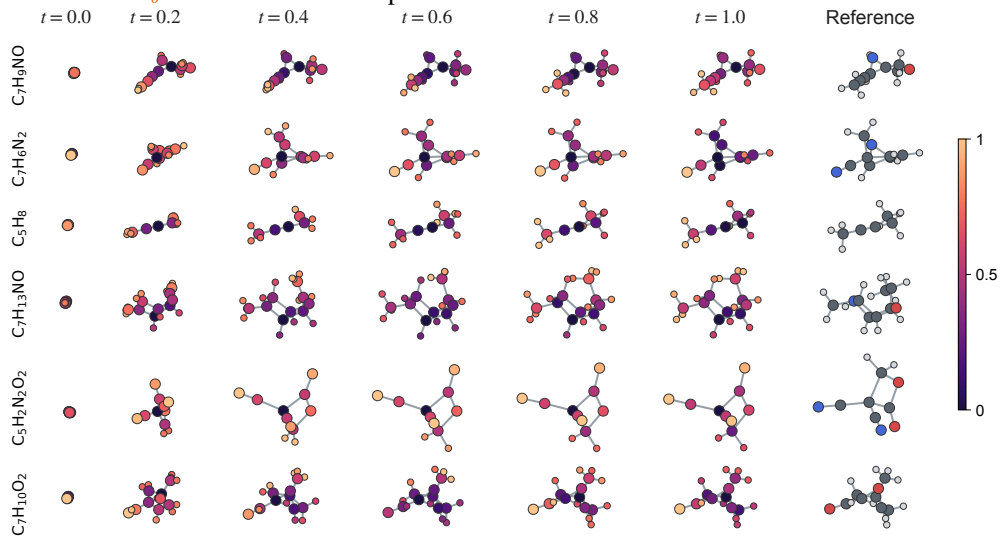


Figure 4: Molecular endpoint means μ_θ over generation time, followed by reference conformations. Colors show atomwise variance from ${\Sigma_\theta}^x$. Reference colors denote elements.

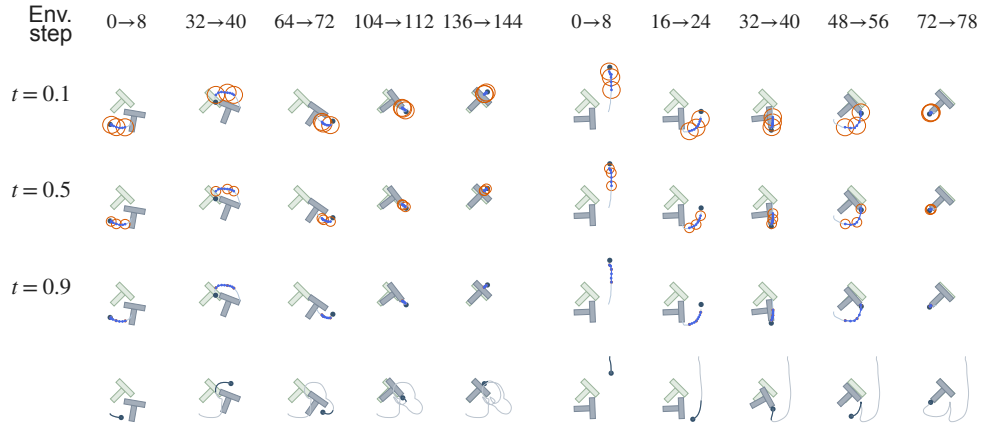


Figure 5: PushT posterior readouts for two episodes. Columns indicate the environment steps from planning to execution, and t denotes generation time. Upper rows show endpoint action means μ_θ (blue) and standard-deviation ellipses from ${\Sigma_\theta}^x$ (orange). The bottom row shows execution.

Quantitative posterior checks. The right panel of Figure 2 compares predicted endpoint variance with mean squared prediction residuals over time. The curves closely track each other as both decrease, linking the predicted uncertainty to the evolving prediction error.

At each evaluated time, the comparisons in Section 4.3 show broad agreement between mean predicted variances and mean squared residuals on reference states (Figure 6(a–c)), with departures in low-variance and sparse point-cloud bins. With the molecular mean fixed, learned covariance yields lower Gaussian NLL at every evaluated time than a covariance that depends only on time and is fitted on validation data (d). Appendix H.4 defines the covariance comparison and explains its implications for posterior fitting.

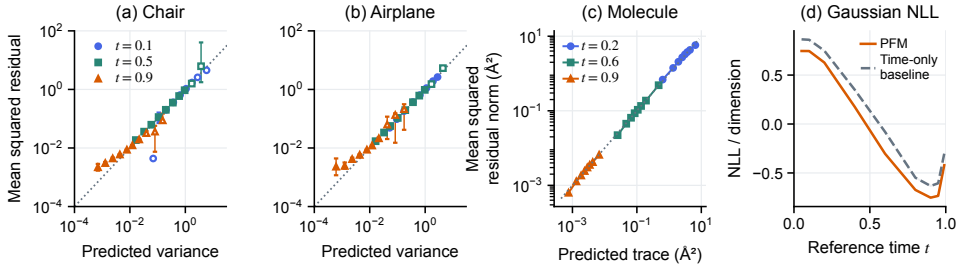


Figure 6: Posterior diagnostics grouped at fixed reference times. (a,b) Chair/Airplane endpoint variances versus mean squared residuals. (c) Molecular covariance traces versus mean squared residual norms. (d) Molecular velocity-posterior NLL per subspace dimension, with a shared predicted mean (lower is better). Dotted lines in (a–c) indicate equality. Bars/bands show 95% bootstrap intervals, and hollow markers indicate sparse bins.

5.3 Ablation Studies and Sampling Cost

Role of posterior covariance. In $a_\theta = \lambda(t)\Sigma_\theta$ (Eq. 18), the posterior covariance determines local noise allocation. The comparison uses a shared mean model and the same scalar noise schedule for the two SDEs. PFM SDE outperforms PFM ODE and PFM + isotropic SDE across all four tasks. Appendix H.7.1 details the comparison.

Inference-time stochasticity. With this covariance field fixed, λ controls the overall stochasticity. Figure 7 illustrates gains at intermediate λ_0 in $\lambda(t) = \lambda_0(1-t)$ relative to PFM ODE ($\lambda_0 = 0$), followed by declining quality at larger scales. Thus the same posterior model supports inference-time adjustment: covariance determines local noise allocation, while its overall strength affects sampling quality. Appendix H.7.2 connects this adjustment to the one-dimensional generation trajectories.

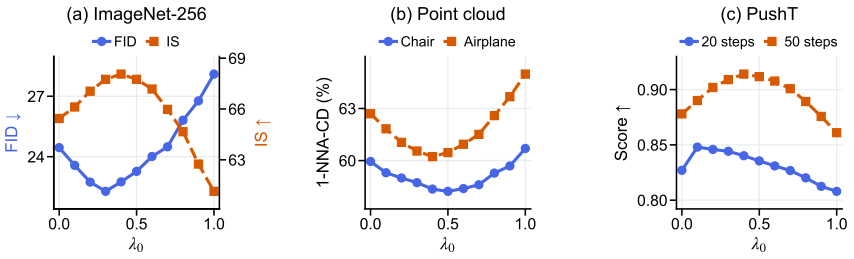


Figure 7: Stochasticity sweeps with $\lambda(t) = \lambda_0(1-t)$. (a) ImageNet-256 FID and IS. (b) Chair and Airplane 1-NNA-CD (ideal: 50%). (c) PushT control scores at two integration budgets.

Sampling time. These sampling choices also differ in cost. PFM ODE has sampling times close to FM ODE, while PFM SDE takes 1.16–1.33× the PFM ODE time. Appendix H.7.3 reports the per-task times and discusses the additional variance-field derivatives.

6 Conclusion

We introduced Posterior Flow Matching (PFM), a general generative framework that connects data generation with endpoint uncertainty estimation. Across image, point-cloud, molecular, and action generation, PFM improves generation quality while providing uncertainty readouts throughout the generation process.

AI USE DISCLOSURE

We used large language models to polish the writing and improve the readability of the manuscript. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content.

REFERENCES

Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3D point clouds. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 40–49. PMLR, 2018. URL https://proceedings.mlr.press/v80/achlioptas18a.html.

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic Interpolants: A Unifying Framework for Flows and Diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025. URL https://jmlr.org/papers/v26/23-1605.html.

Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=li7qeBbCR1t.

Michael Samuel Albergo, Mark Goldstein, Nicholas Matthew Boffi, Rajesh Ranganath, and Eric Vanden-Eijnden. Stochastic interpolants with data-dependent couplings. In *International Conference on Machine Learning*, pp. 921–937. PMLR, 2024.

Simon Axelrod and Rafael Gómez-Bombarelli. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9:185, 2022. doi: 10.1038/s41597-022-01288-4. URL https://www.nature.com/articles/s41597-022-01288-4.

Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-DPM: an Analytic Estimate of the Optimal Reverse Variance in Diffusion Probabilistic Models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=0xiJLKH-ufZ.

David Barber and Felix V. Agakov. The IM algorithm: A variational approach to information maximization. In *Advances in Neural Information Processing Systems*, volume 16, 2003. URL https://papers.nips.cc/paper_files/paper/2003/file/a6ea8471c120fe8cc35a2954c9b9c595-Paper.pdf.

Shengqu Cai, Eric Ryan Chan, Yunzhi Zhang, Leonidas Guibas, Jiajun Wu, and Gordon Wetzstein. Diffusion self-distillation for zero-shot customized image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18434–18443, 2025.

Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An information-rich 3D model repository, 2015. URL https://arxiv.org/abs/1512.03012.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

Friso de Kruiff, Dario Coscia, Max Welling, and Erik Bekkers. Flowing with Confidence. arXiv:2605.18472, 2026. URL https://arxiv.org/abs/2605.18472.

Michele De Vita and Vasileios Belagiannis. Diffusion Model Guided Sampling with Pixel-Wise Aleatoric Uncertainty Estimation. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pp. 3844–3854, 2025. URL https://openaccess.thecvf.com/content/WACV2025/html/De_Vita_Diffusion_Model_Guided_Sampling_with_Pixel-Wise_Aleatoric_Uncertainty_Estimation_WACV_2025_paper.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://www.image-net.org/static_files/papers/imagenet_cvpr09.pdf.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=FPnUhsQJ5B.

Pete Florence, Corey Lynch, Andy Zeng, Oscar A Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. In *Conference on robot learning*, pp. 158–168. PMLR, 2022.

Joaquin Fontbona and Benjamin Jourdain. A trajectorial interpretation of the dissipations of entropy and Fisher information for stochastic differential equations. *The Annals of Probability*, 44(1): 131–170, 2016. doi: 10.1214/14-AOP969. URL https://arxiv.org/abs/1107.3300.

Pengsheng Guo and Alex Schwing. Variational rectified flow matching. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Rk18ZikrFI.

Juyeop Han, Lukas Lao Beyer, and Sertac Karaman. Flow Matching with Uncertainty Quantification and Guidance. arXiv:2602.10326, 2026. URL https://arxiv.org/abs/2602.10326.

Majdi Hassan, Nikhil Shenoy, Jungyoon Lee, Hannes Stärk, Stephan Thaler, and Dominique Beaini. ET-Flow: Equivariant Flow-Matching for Molecular Conformer Generation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 128798–128824. Curran Associates, Inc., 2024. doi: 10.52202/079017-4091. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e8bd617e7dd0394ceadf37b4a7773179-Abstract-Conference.html.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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.

Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning*, pp. 13213–13232. PMLR, 2023.

Metod Jazbec, Eliot Wong-Toi, Guoxuan Xia, Dan Zhang, Eric Nalisnick, and Stephan Mandt. Generative uncertainty in diffusion models. In Silvia Chiappa and Sara Magliacane (eds.), *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 1837–1858. PMLR, 21–25 Jul 2025. URL https://proceedings.mlr.press/v286/jazbec25a.html.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong MU, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=66NzcRQuOq.

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, 2022.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6114.

Leon Klein, Andreas Krämer, and Frank Noé. Equivariant flow matching. *Advances in Neural Information Processing Systems*, 36:59886–59910, 2023.

Nikita Kornilov, Petr Mokrov, Alexander Gasnikov, and Aleksandr Korotin. Optimal flow matching: Learning straight trajectories in just one step. *Advances in Neural Information Processing Systems*, 37:104180–104204, 2024.

Siqi Kou, Lei Gan, Dequan Wang, Chongxuan Li, and Zhijie Deng. BayesDiff: Estimating Pixel-wise Uncertainty in Diffusion via Bayesian Inference. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.11142.

Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Chieh-Hsin Lai, Yang Song, Dongjun Kim, Yuki Mitsufuji, and Stefano Ermon. The principles of diffusion models, 2025. URL https://arxiv.org/abs/2510.21890.

Sangyun Lee, Zinan Lin, and Giulia Fanti. Improving the training of rectified flows. *Advances in neural information processing systems*, 37:63082–63109, 2024.

Tianhong Li and Kaiming He. Back to Basics: Let Denoising Generative Models Denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36115–36125, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_Back_to_Basics_Let_Denoising_Generative_Models_Denoise_CVPR_2026_paper.html.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky T. Q. Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code, 2024. URL https://arxiv.org/abs/2412.06264.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Shitong Luo and Wei Hu. Diffusion probabilistic models for 3D point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2837–2845, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Luo_Diffusion_Probabilistic_Models_for_3D_Point_Cloud_Generation_CVPR_2021_paper.html.

Li Lv, Jiarui Xing, Xinyan Liang, and Jian Wang. Uncertainty in a Single Pass: A Closed-Form Identity for One-Step Flow Matching. arXiv:2605.00941v5, 2026. URL https://arxiv.org/pdf/2605.00941v5.

Chenrui Ma, Xi Xiao, Tianyang Wang, Xiao Wang, and Yanning Shen. Stochastic interpolants via conditional dependent coupling. *arXiv preprint arXiv:2509.23122*, 2025.

Chenrui Ma, Xi Xiao, Tianyang Wang, Xiao Wang, and Yanning Shen. Learning Straight Flows: Variational Flow Matching for Efficient Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38154–38164, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Ma_Learning_Straight_Flows_Variational_Flow_Matching_for_Efficient_Generation_CVPR_2026_paper.html.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

Hila Manor and Tomer Michaeli. On the Posterior Distribution in Denoising: Application to Uncertainty Quantification. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=adSGeugiuj.

Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://arxiv.org/abs/1509.08731.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/nichol21a.html.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky TQ Chen. Multisample flow matching: Straightening flows with minibatch couplings. In *International Conference on Machine Learning*, pp. 28100–28127. PMLR, 2023.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 5171–5180. PMLR, 2019. URL https://proceedings.mlr.press/v97/poole19a.html.

Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 652–660, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Qi_PointNet_Deep_Learning_CVPR_2017_paper.html.

Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1: 140022, 2014. doi: 10.1038/sdata.2014.22. URL https://www.nature.com/articles/sdata201422.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. URL https://arxiv.org/abs/1409.0575.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2011.13456.

George Stoica, Sayak Paul, Matthew Wallingford, Vivek Ramanujan, Abhay Nori, Winson Han, Ali Farhadi, Ranjay Krishna, and Judy Hoffman. Posterior Augmented Flow Matching. arXiv:2605.00825, 2026. URL https://arxiv.org/abs/2605.00825.

Philipp Thölke and Gianni De Fabritiis. TorchMD-NET: Equivariant transformers for neural network based molecular potentials. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2202.02541.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and Generalizing Flow-Based Generative Models with Minibatch Optimal Transport. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=CD9Snc73AW.

Fu-Yun Wang, Ling Yang, Zhaoyang Huang, Mengdi Wang, and Hongsheng Li. Rectified Diffusion: Straightness Is Not Your Need in Rectified Flow. In *International Conference on Learning Representations*, pp. 31420–31445, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/4df9a5e6bad9e64ebcea453e031142bb-Abstract-Conference.html.

Andre Wibisono, Varun Jog, and Po-Ling Loh. Information and estimation in Fokker–Planck channels. In *IEEE International Symposium on Information Theory*, 2017. URL https://arxiv.org/abs/1702.03656.

Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. PointFlow: 3D point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. URL https://arxiv.org/abs/1906.12320.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation Alignment for Generation: Training Diffusion Transformers Is Easier Than You Think. In *International Conference on Learning Representations*, pp. 87400–87442, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d9e42b4d7163931f3689d6d6fbaa11d0-Abstract-Conference.html.

Yichi Zhang, Yici Yan, Alex Schwing, and Zhizhen Zhao. Hierarchical rectified flow matching with mini-batch couplings. *arXiv preprint arXiv:2507.13350*, 2025a.

Yichi Zhang, Yici Yan, Alex Schwing, and Zhizhen Zhao. Towards hierarchical rectified flow. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=6F6qwdycgJ.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion Transformers with Representation Autoencoders. In *International Conference on Learning Representations*, pp. 35791–35820, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3c4141c12660ad3625eb4ae845e0a6f9-Abstract-Conference.html.

APPENDIX CONTENTS

A Notation and Shared Foundations . . . 16
A.1 Notation reference . . . 16
A.2 Conditional-to-marginal flow matching . . . 16
A.3 Diffusion training and conditional scores . . . 17
A.4 The Gaussian velocity–score identity . . . 18
A.5 Fokker–Planck correspondence and marginal compatibility . . . 18
A.6 Diffusion and FM as sampling special cases . . . 19

B Posterior Learning and Coordinate Transformations . . . 19
B.1 Endpoint posterior fitting and information . . . 20
B.2 Velocity and endpoint coordinates . . . 20
B.3 Gaussian model and likelihood . . . 20
B.4 Proof of conditional moment recovery . . . 21

C Posterior-Driven Dynamics . . . 22
C.1 From posterior normalization to the diffusion tensor . . . 22
C.2 Compatible drift and its coordinate form . . . 22
C.3 Proof of relative-entropy dissipation . . . 22

D Posterior Diagnostics . . . 23
D.1 Posterior readouts and residual transformations . . . 23
D.2 Moment and likelihood checks . . . 24
D.3 Readouts along generated trajectories . . . 24

E Conditional Generation . . . 24
E.1 Conditional reference construction . . . 24
E.2 Learning conditional velocity posteriors . . . 25
E.3 Posterior-driven ODE and SDE sampling . . . 26
E.4 Posterior diagnostics . . . 26

F Endpoint-Mean Parameterization . . . 26
F.1 Direct endpoint-mean model . . . 26
F.2 Likelihood and conditional moments . . . 27
F.3 Generation dynamics . . . 27
F.4 Readouts and parameterization correspondence . . . 27

G Training and Numerical Sampling . . . 28
G.1 Training update . . . 28
G.2 Variance derivatives for the SDE drift . . . 28
G.3 Numerical sampling . . . 29
G.4 Effect of derivative estimation . . . 30
G.5 Terminal time and endpoint output . . . 30

H Supplementary Experiments . . . 32
H.1 Implementation details . . . 32
H.2 Image generation . . . 32
H.3 Point-cloud generation . . . 33
H.4 Molecular conformation generation . . . 33
H.5 Action generation on PushT . . . 34
H.6 One-dimensional generation dynamics . . . 36
H.7 Ablation Studies and Sampling Cost . . . 40

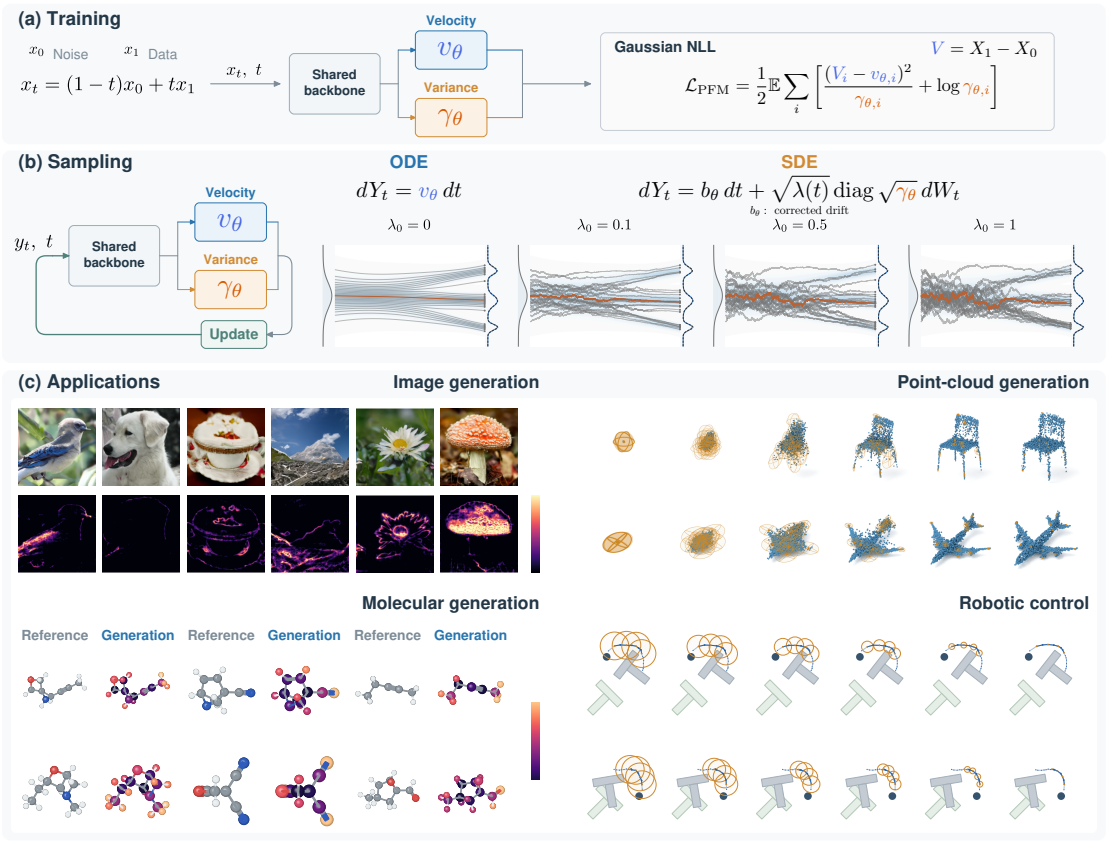


Figure 8: **Posterior Flow Matching. (a) Training.** A shared network learns the velocity mean v_θ and log variance ℓ_θ by Gaussian NLL (Eq. 13). The diagram displays $\gamma_\theta = \exp \ell_\theta$. **(b) Sampling.** These outputs define the PFM ODE and SDE (Eqs. 16 and 19), evaluated at (Y_t, t). Trajectories share a checkpoint and initial particles, with $\lambda(t) = \lambda_0(1-t)$. **(c) Applications.** Endpoint readouts (Eq. 15) pair means with transformed variance maps for images and molecules, or standard-deviation contours for point clouds and actions. Point-cloud rows show Chair (top) and Airplane (bottom).

A Notation and Shared Foundations

This appendix establishes the foundations used in Section 3 and the generator comparison in Section 4.2. For FM and PFM, independent endpoints $X_0 \sim \mathcal{N}(0, I)$ and $X_1 \sim p_{\text{data}}$ define $X_t = (1-t)X_0 + tX_1$ and the velocity target $V = X_1 - X_0$. We first identify its conditional mean as the transport field. We then derive the diffusion score and endpoint readouts, establish the shared ODE–SDE correspondence, and recover the baseline entries of Table 1.

A.1 Notation reference

Table 6 collects the shared symbols. Parameter subscripts identify the model, and μ denotes an endpoint-mean readout. Barred processes, densities, and scores refer to diffusion noising time. Appendix E introduces the external generation condition, and Appendix F changes the directly predicted mean coordinate.

A.2 Conditional-to-marginal flow matching

We prove the continuity identity in Eq. 2 and the gradient equivalence stated after Eq. 3 (Lipman et al., 2023; Albergo & Vanden-Eijnden, 2023; Lipman et al., 2024). For a smooth compactly supported test function ψ, the reference interpolation gives

$$\frac{d}{dt}\mathbb{E}[\psi(X_t)] = \mathbb{E}[\nabla\psi(X_t)^\top V] = \mathbb{E}[\nabla\psi(X_t)^\top v(X_t, t)].$$

Table 6: Core notation used in the main text. Local notation is defined in the relevant appendix. Here d is the state dimension and I the identity matrix.

Symbol	Meaning
$X_0, X_1;\ Z$	Noise and data endpoints; their pair defining a reference path.
$X_t;\ Y_t;\ \bar{X}_\tau$	FM/PFM reference state; generation process; diffusion noising process.
$t;\ \tau = 1 - t;\ C$	Generation time, noising time, and external generation condition.
$p_0;\ p_1 = p_{\text{data}}$	Source and data distributions.
$p_t;\ \bar{p}_\tau$	FM/PFM reference density; diffusion noising density.
$V;\ v$	Sampled velocity target; its conditional mean.
$\eta; \phi; \theta$	Parameters of FM, diffusion, and PFM.
$v_\eta;\ v_\theta$	FM velocity predictor; PFM velocity mean.
$q_\theta;\ q_\theta^x$	Velocity and endpoint posterior models.
F_θ	Network predicting velocity mean and log variance.
$\ell_\theta; \gamma_\theta; \Sigma_\theta$	Velocity log-variance vector, componentwise variance vector, and diagonal covariance.
$\mathcal{L}_{\text{PFM}};\ \mathcal{L}_{\text{CFM}}$	PFM Gaussian negative log-likelihood (NLL) objective; CFM regression objective.
$H;\ I(\cdot;\cdot)$	Differential entropy; mutual information (conditional when given C).
r	Training time law.
$s;\ \bar{s};\ s_\phi$	FM/PFM score; diffusion noising score; predictor of $\bar{s}(x, 1-t)$ in generation time.
$\widehat{s}_\eta; \widehat{s}_\theta$	Scores derived from the corresponding velocity predictions.
$b;\ b_\theta$	General SDE drift; PFM drift computed from posterior outputs.
$B;\ a = BB^\top$	Brownian noise coefficient; diffusion tensor (covariance rate).
$B_\theta;\ a_\theta = \lambda\Sigma_\theta$	PFM Brownian noise coefficient; posterior-driven diffusion tensor.
$f;\ g;\ W$	Prescribed noising drift, scalar noise coefficient, and standard Brownian motion.
$\kappa(t); \lambda(t)$	FM isotropic covariance rate; PFM posterior-covariance scale.
$\widetilde{p}_t$	Density evolved from a different initial law under the same exact compatible dynamics as p_t.
T	Terminal integration time.
$\mu_\eta;\ \mu_\phi;\ \mu_\theta$	Endpoint-mean readouts of FM, diffusion, and PFM.
Σ_θ^x	PFM endpoint-posterior diagonal covariance.
$\alpha_t; \sigma_t$	Diffusion kernel signal scale and cumulative noise standard deviation.

Finite endpoint second moments imply integrability of V. Together with the bounded gradient of ψ, this justifies differentiating along the path. This is the weak form of the continuity equation. A well-defined ODE flow with unique density evolution and initialization p_0 follows these marginals.

The marginal regression objective corresponding to CFM is

$$\mathcal{L}_{\text{FM}}(\eta) = \frac{1}{2}\mathbb{E}_{t,X_t}\|v_\eta(X_t,t) - v(X_t,t)\|^2. \tag{21}$$

Both losses use the same parameter-independent time law r and reference interpolation. Write $D_\eta v_\eta$ for the parameter Jacobian, with output coordinates indexing rows. Under finite risks and conditions allowing differentiation under expectation,

$$\begin{aligned}\nabla_\eta \mathcal{L}_{\text{CFM}} &= \mathbb{E}[(D_\eta v_\eta)^\top (v_\eta - V)] \\ &= \mathbb{E}[(D_\eta v_\eta)^\top (v_\eta - \mathbb{E}[V \mid X_t, t])] \\ &= \mathbb{E}[(D_\eta v_\eta)^\top (v_\eta - v)] = \nabla_\eta \mathcal{L}_{\text{FM}}.\end{aligned} \tag{22}$$

All network quantities are evaluated at (X_t, t). Conditioning on these inputs leaves the prediction and its parameter Jacobian fixed, giving the second equality.

Expanding $V - v_\eta = (V - v) + (v - v_\eta)$ gives a second view of the same result:

$$\mathbb{E}\|V - v_\eta(X_t,t)\|^2 = \mathbb{E}\|V - v(X_t,t)\|^2 + \mathbb{E}\|v(X_t,t) - v_\eta(X_t,t)\|^2. \tag{23}$$

The cross term vanishes because $\mathbb{E}[V - v \mid X_t, t] = 0$. The first term is independent of the predictor, and the second is minimized by the conditional mean. Equations 22 and 23 justify the sample-based training in Section 3.1 and identify the velocity target recovered by Eq. 45.

A.3 Diffusion training and conditional scores

For the noising process in Section 3.2, pair the data endpoint X_1 with $\bar{X}_0 = X_1$ and draw $\bar{X}_{1-t}$ from its perturbation kernel $\bar{p}_{1-t}(\cdot \mid X_1)$. Denoising score matching uses (Song et al., 2021)

$$\mathcal{L}_{\text{score}}(\phi) = \frac{1}{2}\mathbb{E}_{t,X_1,\bar{X}_{1-t}}\left[w(t)\left\|s_\phi(\bar{X}_{1-t},t) - \nabla_x \log \bar{p}_{1-t}(x \mid X_1)|_{x=\bar{X}_{1-t}}\right\|^2\right], \tag{24}$$

where $X_1 \sim p_{\text{data}}$, the time law is fixed, and $w(t) > 0$ is a prescribed weight. A Gaussian kernel with signal scale α_t and cumulative noise standard deviation $\sigma_t > 0$ gives

$$\bar{p}_{1-t}(x \mid x_1) = \mathcal{N}(x; \alpha_t x_1, \sigma_t^2 I), \qquad \nabla_x \log \bar{p}_{1-t}(x \mid x_1) = -\frac{x - \alpha_t x_1}{\sigma_t^2}. \tag{25}$$

The cumulative perturbation scale σ_t and instantaneous noise coefficient $g(1-t)$ have distinct roles.

For a positive marginal density, differentiation under the mixture integral and Bayes' rule give

$$\begin{aligned}\bar{s}(x,1-t) &= \frac{\int \nabla_x \bar{p}_{1-t}(x \mid x_1)\, p_1(dx_1)}{\bar{p}_{1-t}(x)} \\ &= \mathbb{E}[\nabla_x \log \bar{p}_{1-t}(x \mid X_1) \mid \bar{X}_{1-t} = x].\end{aligned} \tag{26}$$

Assuming these derivatives and expectations exist and the squared risks are finite, the conditional-mean decomposition in Eq. 23 identifies the unrestricted minimizer as $s_\phi(x,t) = \bar{s}(x,1-t)$ wherever the training time law has support.

For the Gaussian kernel, substituting Eq. 25 into Eq. 26 yields

$$\bar{s}(x,1-t) = \frac{\alpha_t \mathbb{E}[X_1 \mid \bar{X}_{1-t} = x] - x}{\sigma_t^2}, \qquad \mathbb{E}[X_1 \mid \bar{X}_{1-t} = x] = \frac{x + \sigma_t^2 \bar{s}(x,1-t)}{\alpha_t}, \tag{27}$$

with the second identity requiring $\alpha_t \neq 0$. This relation supplies the diffusion endpoint readout used in Table 1. The baseline comparison in Appendix A.6 collects it together with the FM readout after constructing both samplers.

A.4 The Gaussian velocity–score identity

We derive Eq. 8, which supplies the scores used by FM and PFM. Return to the independent Gaussian linear interpolation of Section 3.1 (Lipman et al., 2024). Its density is

$$p_t(x) = \int \mathcal{N}(x; tx_1, (1-t)^2 I)\, p_1(dx_1).$$

Applying the mixture calculation in Eq. 26 gives

$$s(x,t) = \mathbb{E}\left[-\frac{x - tX_1}{(1-t)^2} \,\middle|\, X_t = x \right] = \frac{t\mathbb{E}[X_1 \mid X_t = x] - x}{(1-t)^2}. \tag{28}$$

Taking conditional expectations in Eq. 10 yields $\mathbb{E}[X_1 \mid X_t = x] = x + (1-t)v(x,t)$. Substitution into Eq. 28 gives

$$s(x,t) = \frac{t[x + (1-t)v(x,t)] - x}{(1-t)^2} = \frac{tv(x,t) - x}{1-t},$$

proving Eq. 8. The corresponding learned score obeys

$$\widehat{s}_\theta - s = \frac{t}{1-t}(v_\theta - v), \tag{29}$$

and the same relation holds with η for FM. These derived scores enter the FM SDE in Eq. 32 and the PFM SDE in Eq. 19.

The same calculation at a fixed external condition gives the score identity in Appendix E.3.

A.5 Fokker–Planck correspondence and marginal compatibility

We expand Eq. 7 in Section 3.3 (Song et al., 2021; Albergo et al., 2025). For $dY_t = b\,dt + B\,dW_t$, the tensor $a = BB^\top$ is symmetric positive semidefinite. Brownian increments have covariance hI, so, under local regularity,

$$\mathbb{E}[Y_{t+h} - Y_t \mid Y_t = x] = h\,b(x,t) + o(h), \qquad \mathrm{Cov}(Y_{t+h} - Y_t \mid Y_t = x) = h\,a(x,t) + o(h).$$

The density equation is

$$\partial_t p_t = -\nabla_x \cdot (b p_t) + \frac{1}{2} \sum_{i,j} \partial_{x_i} \partial_{x_j} (a_{ij} p_t). \tag{30}$$

For $s = \nabla_x \log p_t$, the product rule gives

$$\sum_j \partial_{x_j}(a_{ij} p_t) = \left[\sum_j \partial_{x_j} a_{ij} + \sum_j a_{ij} s_j \right] p_t = [(\nabla_x \cdot a) + a\, s]_i p_t. \tag{31}$$

Consequently, Eq. 30 becomes

$$\partial_t p_t = -\nabla_x \cdot \left[\left(b - \tfrac{1}{2} \nabla_x \cdot a - \tfrac{1}{2} a\, s \right) p_t \right].$$

The bracket is the probability-current velocity v in Eq. 7. Solving for b gives the converse construction from an ODE to an SDE.

For the compatibility claim, let a positive density path solve the continuity equation with velocity v on $[0, T]$, $T < 1$. Choose $a = BB^\top$ and $b = v + (\nabla_x \cdot a + a\nabla_x \log p_t)/2$. Assume sufficient coefficient and density regularity for the displayed derivatives, well-posed ODE and SDE dynamics, and unique density evolution with the specified initial law. Decay at infinity or boundary conditions eliminating boundary fluxes complete the density calculation. Substituting Eq. 31 yields

$$-\nabla_x \cdot (bp_t) + \frac{1}{2}\sum_{i,j} \partial_{x_i}\partial_{x_j}(a_{ij}p_t) = -\nabla_x \cdot (vp_t).$$

Thus the reference density solves both equations. Matching initialization and uniqueness then identify the SDE marginals with p_t. This proves the correspondence used in Section 3.3. It concerns one-time marginals, while the processes can have different joint laws across time.

A.6 Diffusion and FM as sampling special cases

We now apply the shared density construction to the diffusion and FM baselines. The resulting dynamics and endpoint means justify the comparison with PFM in Section 4.2 and Table 1. Diffusion uses the prescribed noising fields f, g and score predictor s_ϕ. FM uses the velocity predictor v_η on the independent Gaussian reference path.

The noising density $\bar{p}_\tau$ satisfies

$$\partial_\tau \bar{p}_\tau = -\nabla_x \cdot (f\bar{p}_\tau) + \tfrac{1}{2}g(\tau)^2 \Delta_x \bar{p}_\tau.$$

Reversing time and using $\Delta_x \bar{p}_{1-t} = \nabla_x \cdot (\bar{s}(x, 1-t)\bar{p}_{1-t})$ gives

$$\partial_t \bar{p}_{1-t} = \nabla_x \cdot (f\bar{p}_{1-t}) - \tfrac{1}{2}g^2 \Delta_x \bar{p}_{1-t} = -\nabla_x \cdot [(-f + \tfrac{1}{2}g^2\bar{s})\bar{p}_{1-t}],$$

where $f, g, \bar{s}$ are evaluated at noising time $1 - t$. Applying Appendix A.5 to this density path gives the probability-flow ODE (Song et al., 2021). Choosing $a = g^2 I$ adds $g^2\bar{s}/2$ to its velocity, yielding the reverse SDE. With the exact score and the terminal noising law as initialization, these dynamics reproduce $\bar{p}_{1-t}$. Replacing $\bar{s}(x, 1-t)$ by $s_\phi(x, t)$ gives the diffusion rows of Table 1.

For FM, Appendix A.4 supplies $\widehat{s}_\eta = (tv_\eta - x)/(1-t)$. Choose $a = \kappa(t)I$ with $\kappa(t) \geq 0$ (Albergo et al., 2025; Lipman et al., 2024). Then $\nabla_x \cdot a = 0$ and $B = \sqrt{\kappa}I$, so

$$\begin{aligned} dY_t &= v_\eta(Y_t, t)dt, \\ dY_t &= [v_\eta + \tfrac{1}{2}\kappa\widehat{s}_\eta](Y_t, t)dt + \sqrt{\kappa(t)}\, dW_t. \end{aligned} \tag{32}$$

Both start from $\mathcal{N}(0, I)$. At $v_\eta = v$, the score is exact and Appendix A.5 gives the reference marginals. Setting $\kappa = 0$ returns the ODE. Appendix G.3 supplies the discrete SDE step.

For the endpoint means, taking conditional expectations in $X_1 = X_t + (1-t)V$ gives the FM readout. Replacing the exact score in Eq. 27 by s_ϕ gives its diffusion counterpart:

$$\mu_\eta(x, t) = x + (1-t)v_\eta(x, t), \qquad \mu_\phi(x, t) = \frac{x + \sigma_t^2 s_\phi(x, t)}{\alpha_t}, \quad \alpha_t \neq 0. \tag{33}$$

Together with the dynamics above, these readouts give the baseline entries of Table 1. PFM uses the same correspondence with the posterior-dependent tensor derived in Appendix C.1, and its additional variance readouts follow from Appendix B.2.

B Posterior Learning and Coordinate Transformations

This appendix develops the learning argument in Section 4.1, from endpoint information to the moments fitted by the PFM objective. Independent $X_0 \sim \mathcal{N}(0, I)$ and $X_1 \sim p_{\text{data}}$ define $X_t = (1-t)X_0 + tX_1$ and $V = X_1 - X_0$. At fixed $t < 1$, we fit the endpoint posterior $p(X_1 \mid X_t)$ with q_θ^x. We first derive the information bound in Eq. 9, then express the likelihood in velocity coordinates and prove Proposition 4.1. The information identities assume conditional densities, finite displayed entropies and cross-entropies, and a model density positive on the posterior support.

B.1 Endpoint posterior fitting and information

At fixed time, the available endpoint information and remaining conditional uncertainty satisfy

$$I(X_t; X_1) = H(X_1) - H(X_1 \mid X_t). \tag{34}$$

To connect this identity to posterior fitting, write $p(\cdot \mid X_t)$ for the true endpoint posterior, suppressing fixed time. We use $\mathrm{KL}(p\|q) = \mathbb{E}_p[\log(p/q)]$ for the Kullback–Leibler divergence, or relative entropy. Adding and subtracting its log density in the expected NLL gives

$$\begin{aligned}\mathbb{E}[-\log q_\theta^x(X_1 \mid X_t, t)] &= -\mathbb{E}\log p(X_1 \mid X_t) + \mathbb{E}\log \frac{p(X_1 \mid X_t)}{q_\theta^x(X_1 \mid X_t, t)} \\ &= H(X_1 \mid X_t) + \mathbb{E}_{X_t}\,\mathrm{KL}\left(p(\cdot \mid X_t)\|q_\theta^x(\cdot \mid X_t, t)\right).\end{aligned} \tag{35}$$

Subtracting this identity from $H(X_1)$ and using the nonnegativity of KL gives (Barber & Agakov, 2003; Poole et al., 2019)

$$I(X_t; X_1) \geq H(X_1) + \mathbb{E}\log q_\theta^x(X_1 \mid X_t, t),$$

the bound in Eq. 9. Its gap is the expected posterior KL in Eq. 35.

Equality in Eq. 9 holds when the model matches the true posterior almost everywhere. The reference path fixes the MI, while posterior fitting reduces the approximation gap. The affine relation below expresses the same objective in velocity coordinates.

B.2 Velocity and endpoint coordinates

The reference path implies $X_1 = X_t + (1-t)V$, the affine identity in Eq. 10. At fixed (x, t) with $t < 1$, the map $x_1 = x + (1-t)v$ is invertible and has determinant $(1-t)^d$. The induced endpoint model therefore has density

$$q_\theta^x(x_1 \mid x, t) = \frac{q_\theta((x_1 - x)/(1-t) \mid x, t)}{(1-t)^d}. \tag{36}$$

Taking negative logarithms yields

$$-\log q_\theta^x(X_1 \mid X_t, t) = -\log q_\theta(V \mid X_t, t) + d\log(1-t). \tag{37}$$

The last term is independent of θ, which justifies the velocity-coordinate objective in Section 4.1. The same map also determines the endpoint moments in Eq. 15.

For any true conditional law with finite second moments, the same affine map gives

$$\begin{aligned}\mathbb{E}[X_1 \mid X_t = x] &= x + (1-t)v(x, t), \\ \mathrm{Cov}(X_1 \mid X_t = x) &= (1-t)^2\,\mathrm{Cov}(V \mid X_t = x).\end{aligned} \tag{38}$$

Subtracting the conditional mean from the affine identity multiplies the centered velocity by $1 - t$. Taking its outer product therefore proves the full covariance relation. The model retains the diagonal of this covariance at the population optimum in Proposition 4.1.

The same Jacobian determines the entropy transformation. Assuming finite differential entropies, taking the expected negative log density of the transformed true posterior gives

$$H(X_1 \mid X_t) = H(V \mid X_t) + d\log(1-t). \tag{39}$$

The translation by the observed state leaves conditional entropy unchanged, while the scale $1 - t$ contributes its log determinant. This connects the information bound to the velocity-coordinate likelihood.

B.3 Gaussian model and likelihood

The model $q_\theta = \mathcal{N}(v_\theta, \Sigma_\theta)$ uses $\Sigma_\theta = \mathrm{diag}\,\gamma_\theta$ with positive variances. Applying the affine map to this Gaussian gives

$$q_\theta^x = \mathcal{N}(\mu_\theta, {\Sigma_\theta}^x), \qquad \mu_\theta = x + (1-t)v_\theta, \qquad {\Sigma_\theta}^x = (1-t)^2\Sigma_\theta.$$

These are the moments of the model distribution in Eq. 15.

The complete velocity negative log-likelihood is

$$-\log q_\theta(V \mid X_t, t) = \frac{1}{2}\left[(V - v_\theta)^\top {\Sigma_\theta}^{-1}(V - v_\theta) + \log\det\Sigma_\theta\right] + \frac{d}{2}\log(2\pi). \tag{40}$$

Define $\mathcal{L}_{\mathrm{PFM}} = \mathbb{E}[-\log q_\theta(V \mid X_t, t)] - \frac{d}{2}\log(2\pi)$, which is the objective in Eq. 13. Combining Eqs. 36 and 40 and averaging under the fixed training law gives

$$\mathbb{E}[-\log q_\theta^x(X_1 \mid X_t, t)] = \mathcal{L}_{\mathrm{PFM}} + d\,\mathbb{E}_t \log(1-t) + \frac{d}{2}\log(2\pi). \tag{41}$$

The averaged identity assumes finite expected negative log-likelihoods and integrability of $\log(1-t)$. At each fixed $t < 1$, the pointwise density identity already gives the parameter-independent difference. Since the additional terms in Eq. 41 are independent of θ, minimizing $\mathcal{L}_{\mathrm{PFM}}$ maximizes the variational lower bound in Eq. 9 averaged over the training time law r.

Diagonal covariance. At fixed (x, t),

$$\mathrm{Cov}_{q_\theta}(V_i, V_j \mid x, t) = 0 \quad (i \neq j).$$

The approximation uses d variance outputs instead of the $d(d+1)/2$ entries of a full symmetric covariance. Each output still depends on the complete state, so $\partial_{x_j}\gamma_{\theta,i}$ can be nonzero for $i \neq j$. Conditional covariance of the modeled target and the Jacobian of the predicted variance field are different objects: the former is diagonal here, while the latter can be dense. Appendix G.2 computes the diagonal of this input Jacobian for the SDE drift.

B.4 Proof of conditional moment recovery

We prove Proposition 4.1 and obtain the residual identity used by the diagnostics in Section 4.3.

Proof of Proposition 4.1. Fix a state, time, and coordinate i. Let m_i and ${\sigma_i}^2 > 0$ denote the true conditional mean and variance. For a candidate mean u_i,

$$\mathbb{E}[(V_i - u_i)^2 \mid X_t = x] = \mathbb{E}[(V_i - m_i)^2 \mid X_t = x] + (m_i - u_i)^2,$$

since the centered cross term has zero expectation. A candidate positive variance γ_i therefore gives conditional Gaussian risk

$$R_i(u_i, \gamma_i) = \frac{1}{2}\left[\frac{{\sigma_i}^2 + (m_i - u_i)^2}{\gamma_i} + \log\gamma_i\right]. \tag{42}$$

For any fixed positive γ_i, the unique mean minimizer is $u_i = m_i$. At that mean,

$$\partial_{\gamma_i} R_i = \frac{\gamma_i - {\sigma_i}^2}{2{\gamma_i}^2}.$$

It is negative below ${\sigma_i}^2$ and positive above, giving the unique variance minimum. Summing the coordinate risks proves the result. Only the true conditional first and second moments enter this calculation. □

For a fixed biased mean and a positive residual second moment, minimizing Eq. 42 over variance gives

$$\gamma_i = \mathbb{E}[(V_i - u_i)^2 \mid X_t = x] = {\sigma_i}^2 + (m_i - u_i)^2. \tag{43}$$

This supplies the residual interpretation in Appendix D.2.

The implementations in Appendix H.1 bound each log variance between fixed values $\ell_- < \ell_+$. For the diagonal covariance model, the same risk gives the optimum over the closed interval $[e^{\ell_-}, e^{\ell_+}]$. For a fixed mean u_i, its variance derivative has the sign of $\gamma_i - \mathbb{E}[(V_i - u_i)^2 \mid X_t = x]$, so

$$\gamma_i^\star(u_i) = \mathrm{clip}\big(\mathbb{E}[(V_i - u_i)^2 \mid X_t = x], e^{\ell_-}, e^{\ell_+}\big). \tag{44}$$

Here $\mathrm{clip}(r, a, b) = \min\{b, \max\{a, r\}\}$. With a freely optimized mean at each (x, t), the mean optimum remains $u_i^\star = m_i$ and Eq. 44 becomes the clipped conditional variance $\mathrm{clip}({\sigma_i}^2, e^{\ell_-}, e^{\ell_+})$. Thus the bounds retain conditional moment recovery within the variance interval and set its limiting scales outside it. Appendix G.1 gives the smooth parameterization used to enforce these bounds.

Fixing all variances to one in Eq. 13 removes likelihood weighting and the log determinant, leaving

$$\mathcal{L}_{\mathrm{PFM}} = \frac{1}{2}\mathbb{E}\|V - v_\theta(X_t, t)\|^2. \tag{45}$$

This is the FM specialization stated in Section 4.1. Together, the likelihood transformation and moment recovery show how one PFM objective fits the velocity moments used for generation and the endpoint moments used for diagnostics.

C Posterior-Driven Dynamics

This appendix derives the PFM SDE in Section 4.2 and proves its relative-entropy result. Let p_t be a reference density transported by v, with score $s = \nabla_x \log p_t$. PFM predicts v_θ and the positive diagonal covariance $\Sigma_\theta = \operatorname{diag} \gamma_\theta$. We use the covariance to set local diffusion scales, pair the resulting tensor with a compatible drift, and then derive the dissipation identity for two densities evolving under those same coefficients.

C.1 From posterior normalization to the diffusion tensor

The design principle in Eq. 17 makes the noise isotropic after dividing each noise coordinate by the predicted velocity posterior standard deviation at the current state: $\Sigma_\theta{}^{-1/2} a_\theta \Sigma_\theta{}^{-1/2} = \lambda(t) I$ with $\lambda(t) \geq 0$. At fixed (x, t), the positive diagonal covariance is invertible. Multiplying both sides by $\Sigma_\theta{}^{1/2}$ yields $a_\theta = \lambda \Sigma_\theta$. A square root is $B_\theta = \sqrt{\lambda} \operatorname{diag} \sqrt{\gamma_\theta}$, for which $B_\theta B_\theta{}^\top = a_\theta$.

The normalization in Eq. 17 is a local matrix calculation at the current state. For a Brownian increment $\Delta W \sim \mathcal{N}(0, hI)$, freezing the coefficients at this state gives

$$\operatorname{Cov}(B_\theta \Delta W \mid x, t) = h\lambda \Sigma_\theta, \qquad \operatorname{Cov}(\Sigma_\theta{}^{-1/2} B_\theta \Delta W \mid x, t) = h\lambda I.$$

Thus coordinate i receives noise variance $h\lambda\gamma_{\theta,i}$. The posterior controls relative strengths, and λ converts velocity covariance into a state-increment covariance rate. This explains the component-wise noise allocation in Section 4.2.

C.2 Compatible drift and its coordinate form

For the independent Gaussian reference path, the mean defines $\widehat{s}_\theta = (t v_\theta - x)/(1 - t)$, as derived in Appendix A.4. Pairing the chosen diffusion tensor with this score and the mean gives the drift in Eq. 19:

$$b_\theta = v_\theta + \tfrac{1}{2} \nabla_x \cdot a_\theta + \tfrac{1}{2} a_\theta \widehat{s}_\theta. \tag{46}$$

The positive correction follows from solving the shared correspondence for the SDE drift. Since λ depends only on time,

$$(\nabla_x \cdot a_\theta)_i = \sum_j \partial_{x_j} \left[\lambda(t) \gamma_{\theta,i} \delta_{ij}\right] = \lambda(t) \partial_{x_i} \gamma_{\theta,i},$$

where δ_{ij} is the Kronecker delta. Defining $D_i = \partial_{x_i} \gamma_{\theta,i}$ gives

$$b_{\theta,i} = v_{\theta,i} + \frac{\lambda(t)}{2} \left[D_i + \gamma_{\theta,i} \frac{t v_{\theta,i} - x_i}{1 - t} \right]. \tag{47}$$

The three contributions are transport, correction for spatially varying covariance, and correction along the density score. Equations 46 and 47 are the vector and coordinate forms of the sampler in Section 4.2.

At the population mean $v_\theta = v$, Eq. 29 gives $\widehat{s}_\theta = s$. The conditions in Appendix A.5 then establish the reference marginals for the chosen a_θ. Posterior fitting determines which diffusion tensor is used within this compatible family. Setting $\lambda = 0$ eliminates both corrections and the Brownian term, recovering Eq. 16. Appendix G.3 turns these coefficients into numerical updates.

C.3 Proof of relative-entropy dissipation

Proposition 4.2 describes an initial density perturbation under exact compatible dynamics. To prove it, take any diffusion tensor $a = BB^\top$ and fix the drift $b = v + (\nabla_x \cdot a + a \nabla_x \log p_t)/2$. The reference density p_t and a second density $\widetilde{p}_t$ evolve under these same coefficients from their respective initial laws. Under the smoothness, finiteness, and vanishing-boundary conditions of the proposition, we first establish the general identity (Fontbona & Jourdain, 2016; Wibisono et al., 2017)

$$\frac{d}{dt} \operatorname{KL}(\widetilde{p}_t \| p_t) = -\frac{1}{2} \int \widetilde{p}_t \left(\nabla_x \log \frac{\widetilde{p}_t}{p_t} \right)^\top a \left(\nabla_x \log \frac{\widetilde{p}_t}{p_t} \right) dx \leq 0. \tag{48}$$

Proof of Proposition 4.2. With the common coefficients fixed as above, applying Eq. 31 to the diffusion term for $\widetilde{p}_t$ gives

$$\sum_j \partial_{x_j}(a_{ij}\widetilde{p}_t) = [(\nabla_x \cdot a) + a\nabla_x \log \widetilde{p}_t]_i \widetilde{p}_t.$$

Substitution of the common drift into the Fokker–Planck equation therefore yields

$$\partial_t \widetilde{p}_t = -\nabla_x \cdot (v\widetilde{p}_t) + \frac{1}{2}\nabla_x \cdot \left[a\widetilde{p}_t \nabla_x \log \frac{\widetilde{p}_t}{p_t}\right]. \tag{49}$$

The reference satisfies $\partial_t p_t = -\nabla_x \cdot (v p_t)$. Mass conservation removes the integral of $\partial_t \widetilde{p}_t$ when differentiating KL, leaving

$$\frac{d}{dt}\mathrm{KL}(\widetilde{p}_t \| p_t) = \int \log \frac{\widetilde{p}_t}{p_t}\, \partial_t \widetilde{p}_t \, dx - \int \frac{\widetilde{p}_t}{p_t}\, \partial_t p_t \, dx. \tag{50}$$

All integrals below are over $\mathbb{R}^d$. Under the regularity and vanishing-boundary assumptions in Proposition 4.2, integration by parts turns the two transport contributions into

$$\int v^\top \widetilde{p}_t \nabla_x \log \frac{\widetilde{p}_t}{p_t}\, dx - \int v^\top p_t \nabla_x \frac{\widetilde{p}_t}{p_t}\, dx = 0,$$

because $\widetilde{p}_t \nabla_x \log(\widetilde{p}_t/p_t) = p_t \nabla_x (\widetilde{p}_t/p_t)$. The remaining term is

$$\begin{aligned}
&\frac{1}{2}\int \log \frac{\widetilde{p}_t}{p_t}\, \nabla_x \cdot \left[a\widetilde{p}_t \nabla_x \log \frac{\widetilde{p}_t}{p_t}\right] dx \\
&\qquad = -\frac{1}{2}\int \widetilde{p}_t \left(\nabla_x \log \frac{\widetilde{p}_t}{p_t}\right)^\top a \left(\nabla_x \log \frac{\widetilde{p}_t}{p_t}\right) dx.
\end{aligned}$$

This proves Eq. 48, with the sign following from positive semidefiniteness. Substituting $a = \lambda \,\mathrm{diag}\, \gamma_\theta$ gives the coordinate sum in Eq. 20. Taking $a = 0$ leaves only the canceling transport terms and proves the ODE statement in Proposition 4.2. □

For the isotropic FM choice $a = \kappa I$, all coordinate weights are $\kappa(t)$. PFM uses $\lambda(t)\gamma_{\theta,i}(x,t)$, so the posterior covariance in Eq. 18 determines the coordinatewise weights in relative-entropy dissipation.

When the dissipation is integrable in time, Eq. 48 gives the cumulative identity

$$\begin{aligned}
&\mathrm{KL}(\widetilde{p}_T \| p_T) \\
&\quad + \frac{1}{2}\int_0^T \int \widetilde{p}_t \left(\nabla_x \log \frac{\widetilde{p}_t}{p_t}\right)^\top a \left(\nabla_x \log \frac{\widetilde{p}_t}{p_t}\right) dx\, dt = \mathrm{KL}(\widetilde{p}_0 \| p_0).
\end{aligned} \tag{51}$$

D Posterior Diagnostics

This appendix gives the statistical basis for Section 4.3. Independent noise and data endpoints define $X_t = (1-t)X_0 + tX_1$ and $V = X_1 - X_0$. A PFM model predicts velocity mean v_θ and covariance Σ_θ. Their endpoint readouts are $\mu_\theta = x + (1-t)v_\theta$ and ${\Sigma_\theta}^x = (1-t)^2\Sigma_\theta$. We derive how prediction residuals transform between these coordinates, explain moment checks and Gaussian NLL comparisons, and finally interpret the readouts at states visited by a sampler.

D.1 Posterior readouts and residual transformations

We develop Eq. 15 for reference pairs. The model distribution in Eq. 12 has exact moments

$$\mathbb{E}_{q_\theta}[V \mid x,t] = v_\theta(x,t), \qquad \mathrm{Cov}_{q_\theta}(V \mid x,t) = \Sigma_\theta(x,t).$$

Its induced endpoint moments follow from Appendix B.2. Proposition 4.1 identifies their true population targets.

For evaluation, pair a data endpoint X_1 not used in training with independent noise X_0, and construct the reference state X_t and target velocity V by Eq. 1. The known endpoint then supplies coordinate residuals

$$e_i^v = V_i - v_{\theta,i}(X_t,t), \qquad e_i^x = X_{1,i} - \mu_{\theta,i}(X_t,t).$$

Substituting $X_1 = X_t + (1-t)V$ and $\mu_\theta = X_t + (1-t)v_\theta$ gives $e_i^x = (1-t)e_i^v$, and hence

$$(X_{1,i} - \mu_{\theta,i}(X_t,t))^2 = (1-t)^2(V_i - v_{\theta,i}(X_t,t))^2. \tag{52}$$

This exact relation holds for every reference pair and predicted mean. This relation expresses the same residual comparison in velocity or endpoint coordinates.

D.2 Moment and likelihood checks

We explain how the residuals in Eq. 52 assess the moments predicted in Section 4.3. At fixed (x,t), let m_i and ${\sigma_i}^2$ be the true velocity mean and variance. Equation 43 gives

$$\mathbb{E}[e_i^v \mid X_t = x] = m_i - v_{\theta,i}, \qquad \mathbb{E}[(e_i^v)^2 \mid X_t = x] = {\sigma_i}^2 + (m_i - v_{\theta,i})^2.$$

Thus squared residuals measure conditional variance at an accurate mean and also include squared mean bias otherwise.

A practical check groups reference states at matched times by their inputs or predicted variances. If the predicted moments equal their conditional targets, averaging the preceding identities within a group gives zero mean residual and agreement between mean predicted variance and mean squared residual. These comparisons assess the fitted moments without requiring repeated observations at an identical continuous state. Equation 52 gives the corresponding endpoint comparison by multiplying both variance and squared residual by $(1-t)^2$. This is the basis of the variance–residual plots in Section 5.2.

A complementary check compares Gaussian NLLs at a shared predicted mean. Evaluating covariance models on the same reference states keeps the prediction residuals fixed, so the covariance-weighted residual term and log determinant in Eq. 40 jointly assess how each covariance fits those residuals. Appendix H.4 applies this comparison within the centered coordinate subspace, using the learned molecular covariance and a covariance that depends only on time and is fitted on validation data.

D.3 Readouts along generated trajectories

At a visited state $Y_t = y$, the model approximates the reference endpoint posterior $p(X_1 \mid X_t = y)$. Its mean $\mu_\theta(y,t) = y + (1-t)v_\theta(y,t)$ predicts the endpoint, and its covariance $\Sigma_\theta{}^x(y,t) = (1-t)^2\Sigma_\theta(y,t)$ describes the dispersion of compatible data endpoints. Both readouts are available under ODE and SDE sampling. Variance maps and standard-deviation contours around the predicted means display these readouts throughout generation, as in Section 5.2.

The same velocity covariance also determines the SDE diffusion tensor $a_\theta = \lambda(t)\Sigma_\theta$. Thus the endpoint factor $(1-t)^2$ converts velocity dispersion into endpoint dispersion, while $\lambda(t)$ scales the stochastic perturbations used for sampling. Appendix C.1 develops this noise construction. The moment and likelihood checks above assess the posterior readouts on reference states with known endpoints.

E Conditional Generation

This appendix extends the posterior learning, sampling, and diagnostics of Sections 4.1–4.3 to an external generation input C. After defining the conditional reference distribution, we develop posterior learning and endpoint readouts, construct ODE and SDE sampling, and assess the readouts through posterior diagnostics. The condition can specify a class, observation, or reference structure. It is held fixed within each generation trajectory, while the sampled endpoint pair $Z = (X_0, X_1)$ specifies the reference interpolation.

E.1 Conditional reference construction

Let $C \sim p_C$ and $X_1 \mid C = c \sim p_{\text{data}}(\cdot \mid c)$. Draw an independent source $X_0 \sim p_0 = \mathcal{N}(0, I)$, so the joint law is

$$p_C(dc)\, p_0(dx_0)\, p_{\text{data}}(dx_1 \mid c). \tag{53}$$

For each pair, retain the reference interpolation and velocity target (Lipman et al., 2023; 2024)

$$X_t = (1-t)X_0 + tX_1, \qquad V = X_1 - X_0, \qquad p_t(\cdot \mid c) = \text{Law}(X_t \mid C = c). \tag{54}$$

At fixed c, the marginal velocity is

$$v(x,t,c) = \mathbb{E}[V \mid X_t = x, C = c]. \tag{55}$$

Applying the conditional-expectation calculation in Appendix A.2 within each c gives $\partial_t p_t(\cdot \mid c) + \nabla_x \cdot (p_t(\cdot \mid c) v(\cdot, t, c)) = 0$. The source density at $t = 0$ is p_0 for every condition. The affine identity $X_1 = X_t + (1-t)V$ holds for each reference pair and provides the posterior correspondence used next.

E.2 Learning conditional velocity posteriors

At fixed $t < 1$, the information supplied by the reference state after observing C is

$$I(X_t; X_1 \mid C) = H(X_1 \mid C) - H(X_1 \mid X_t, C). \tag{56}$$

Model the remaining endpoint posterior $p(x_1 \mid x, t, c)$ by $q_\theta^x(x_1 \mid x, t, c)$. Under finite conditional entropies and cross-entropies and a model density positive on the posterior support, the decomposition in Appendix B.1 applies conditionally and gives

$$\begin{aligned} \mathbb{E}[-\log q_\theta^x(X_1 \mid X_t, t, C)] &= H(X_1 \mid X_t, C) \\ &\quad + \mathbb{E}_{X_t, C}\, \mathrm{KL}(p(\cdot \mid X_t, C) \,\|\, q_\theta^x(\cdot \mid X_t, t, C)). \end{aligned} \tag{57}$$

Subtracting this identity from $H(X_1 \mid C)$ gives the conditional counterpart of the variational lower bound in Eq. 9 (Barber & Agakov, 2003; Poole et al., 2019; Mohamed & Rezende, 2015):

$$I(X_t; X_1 \mid C) \geq H(X_1 \mid C) + \mathbb{E} \log q_\theta^x(X_1 \mid X_t, t, C). \tag{58}$$

The reference construction fixes the CMI. Minimizing the conditional posterior NLL reduces the expected posterior KL and tightens this bound, as in Appendix B.1.

Holding (x, t, c) fixed, the affine change of variables has the same Jacobian as in Appendix B.2:

$$q_\theta^x(x_1 \mid x, t, c) = \frac{q_\theta((x_1 - x)/(1-t) \mid x, t, c)}{(1-t)^d}. \tag{59}$$

Consequently the endpoint and velocity NLLs differ by $d \log(1-t)$, and their true conditional entropies obey $H(X_1 \mid X_t, C) = H(V \mid X_t, C) + d \log(1-t)$. The conditional velocity model is

$$\begin{aligned} (v_\theta, \ell_\theta) &= F_\theta(x, t, c), \qquad \gamma_\theta = \exp \ell_\theta, \qquad \Sigma_\theta = \operatorname{diag} \gamma_\theta, \\ q_\theta(v \mid x, t, c) &= \mathcal{N}(v; v_\theta(x, t, c), \Sigma_\theta(x, t, c)). \end{aligned} \tag{60}$$

With $t \sim r$ independent of (C, X_0, X_1), the training objective becomes

$$\begin{aligned} \mathcal{L}_{\mathrm{PFM}}^C &= \frac{1}{2}\mathbb{E}\left[(V - v_\theta)^\top {\Sigma_\theta}^{-1} (V - v_\theta) + \log\det \Sigma_\theta\right] \\ &= \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{d} \left(\frac{(V_i - v_{\theta,i})^2}{\gamma_{\theta,i}} + \log \gamma_{\theta,i} \right)\right]. \end{aligned} \tag{61}$$

where all outputs are evaluated at (X_t, t, C). Under the finite-expectation conditions of Appendix B.3, the complete endpoint NLL is

$$\mathbb{E}[-\log q_\theta^x(X_1 \mid X_t, t, C)] = \mathcal{L}_{\mathrm{PFM}}^C + d\, \mathbb{E}_t \log(1-t) + \frac{d}{2}\log(2\pi). \tag{62}$$

The conditional risk in Appendix B.4 applies at each (x, t, c). With freely varying means and positive diagonal variances and finite, strictly positive true coordinate variances, its optimum satisfies

$$v_\theta(x, t, c) = v(x, t, c), \qquad \gamma_{\theta,i}(x, t, c) = \mathrm{Var}(V_i \mid X_t = x, C = c). \tag{63}$$

Fixing $\Sigma_\theta = I$ reduces Eq. 61 to $\frac{1}{2}\mathbb{E}\|V - v_\theta(X_t, t, C)\|^2$, the FM objective for conditional generation. Algorithm 1 implements this formulation by sampling (C, X_1) jointly, drawing independent X_0, t, and passing C to both posterior heads in the same NLL update.

The affine map in Eq. 59 converts these velocity moments into the conditional endpoint-posterior readouts corresponding to Eq. 15:

$$\begin{aligned} \mu_\theta(x, t, c) &= x + (1-t) v_\theta(x, t, c), \\ {\Sigma_\theta}^x(x, t, c) &= (1-t)^2 \Sigma_\theta(x, t, c). \end{aligned} \tag{64}$$

These readouts require no additional head and are available under both samplers below.

E.3 Posterior-driven ODE and SDE sampling

Fix $C = c$ throughout a generation trajectory. For the independent Gaussian source, Appendix A.4 applied to $p_t(\cdot \mid c)$ gives

$$s(x,t,c) = \nabla_x \log p_t(x \mid c) = \frac{tv(x,t,c) - x}{1-t}, \qquad \widehat{s}_\theta(x,t,c) = \frac{tv_\theta(x,t,c) - x}{1-t}. \tag{65}$$

The mean defines the ODE, initialized from $Y_0 \sim \mathcal{N}(0, I)$:

$$dY_t = v_\theta(Y_t, t, c)\, dt. \tag{66}$$

To scale stochastic perturbations by learned uncertainty, apply the posterior normalization principle of Eq. 17. At fixed c, requiring isotropic noise after the same normalization gives the diffusion tensor and its Brownian coefficient,

$$a_\theta(x,t,c) = \lambda(t)\Sigma_\theta(x,t,c), \qquad B_\theta(x,t,c) = \sqrt{\lambda(t)}\,\mathrm{diag}\,\sqrt{\gamma_\theta(x,t,c)}. \tag{67}$$

Appendix C.1 derives this construction. Adding diffusion requires the corresponding drift correction, so the shared ODE–SDE correspondence determines the full drift and sampler:

$$\begin{aligned} b_\theta(x,t,c) &= v_\theta(x,t,c) + \tfrac{1}{2}\nabla_x \cdot a_\theta(x,t,c) + \tfrac{1}{2}a_\theta(x,t,c)\widehat{s}_\theta(x,t,c), \\ dY_t &= b_\theta(Y_t,t,c)\, dt + B_\theta(Y_t,t,c)\, dW_t. \end{aligned} \tag{68}$$

All spatial derivatives act on x, with c held fixed. At $\lambda \equiv 0$ the SDE reduces to Eq. 66. Algorithms 2 and 3 evaluate every field and readout at the same input c throughout sampling.

With the exact conditional velocity and score, the compatibility proof in Appendix A.5 applies to $p_t(\cdot \mid c)$ under its stated regularity assumptions. The resulting ODE and SDE have these conditional marginals. Likewise, the proof of Proposition 4.2 in Appendix C.3 applies at each fixed c: for two densities evolving under the same exact conditional coefficients, the relative-entropy dissipation is weighted by $a_\theta(\cdot, t, c)$. This is the conditional counterpart of the dynamics developed in Appendix C.

E.4 Posterior diagnostics

At a generated state $Y_t = y$, the endpoint readouts in Eq. 64 approximate the moments of $p(X_1 \mid X_t = y, C = c)$. They describe the data endpoints compatible with that reference state and generation condition, following the interpretation in Section 4.3.

For a conditional reference pair, construct X_t from independent noise and a data endpoint with condition C, as in Appendix E.1. Define $e_i^v = V_i - v_{\theta,i}(X_t, t, C)$ and $e_i^x = X_{1,i} - \mu_{\theta,i}(X_t, t, C)$. The identity $e_i^x = (1-t)e_i^v$ holds pointwise. Conditioning the residual identities of Appendix D.2 on both X_t and C gives the same mean-residual and variance–residual checks at matched times. Evaluation uses data endpoints not seen during training, as in Appendix D.1. To assess covariance fit, compare conditional Gaussian NLLs on the same reference inputs (X_t, t, C) with a shared predicted mean, following Appendix D.2. Appendix H.4 applies this comparison to molecular conformations. This completes the conditional formulation of posterior learning, generation, and diagnostics in Sections 4.1–4.3. Appendix H specifies how each task supplies its generation condition.

F Endpoint-Mean Parameterization

This appendix develops direct endpoint-mean prediction as an alternative parameterization of Section 4.1. On the same reference path, $X_1 = X_t + (1-t)V$ makes endpoint and velocity means interchangeable for $t < 1$ (Lipman et al., 2024). The variance head continues to predict velocity log variance. We define this model and its likelihood, identify its conditional moment targets, and construct the fields used by the samplers and diagnostics in Sections 4.2–4.3.

F.1 Direct endpoint-mean model

We develop the alternative to direct velocity prediction introduced in Section 4.1. An endpoint network F_θ^x outputs

$$(\mu_\theta, \ell_\theta) = F_\theta^x(x,t), \qquad \gamma_\theta = \exp \ell_\theta, \qquad q_\theta^x = \mathcal{N}(\mu_\theta, (1-t)^2\,\mathrm{diag}\,\gamma_\theta). \tag{69}$$

Here ℓ_θ is the velocity log variance. Inverting Eq. 10 induces

$$v_\theta(x,t) = \frac{\mu_\theta(x,t) - x}{1-t}, \qquad V - v_\theta(X_t,t) = \frac{X_1 - \mu_\theta(X_t,t)}{1-t}. \tag{70}$$

F.2 Likelihood and conditional moments

Substituting the residual in Eq. 70 into Eq. 13 gives

$$\mathcal{L}^x_{\mathrm{PFM}} = \frac{1}{2}\mathbb{E}\sum_i \left[\frac{(X_{1,i} - \mu_{\theta,i})^2}{(1-t)^2\gamma_{\theta,i}} + \log\gamma_{\theta,i}\right]. \tag{71}$$

All outputs are evaluated at (X_t, t). Algorithm 1 applies with the endpoint network and this integrand. Under the finite-expectation conditions in Appendix B.3, the complete objective is

$$\mathbb{E}[-\log q^x_\theta(X_1 \mid X_t, t)] = \mathcal{L}^x_{\mathrm{PFM}} + d\,\mathbb{E}_t \log(1-t) + \frac{d}{2}\log(2\pi). \tag{72}$$

Thus Eqs. 41 and 72 express the same change of variables for the two mean parameterizations. Setting $\gamma_\theta = \mathbf{1}$ in Eq. 71 yields endpoint-parameterized FM with squared-error weight $(1-t)^{-2}$, corresponding to Eq. 45 in velocity coordinates.

We connect the endpoint network to Proposition 4.1. At fixed (x,t) with $t < 1$, the transformation between candidate endpoint and velocity means is bijective. Under the moment conditions of Proposition 4.1, the conditional population optimum of Eq. 71 has

$$\mu_\theta = \mathbb{E}[X_1 \mid X_t = x], \qquad \gamma_{\theta,i} = \mathrm{Var}(V_i \mid X_t = x).$$

Equation 38 maps this velocity variance to the diagonal endpoint covariance.

F.3 Generation dynamics

Substitution of Eq. 70 into Eq. 16 gives

$$\widehat{s}_\theta = \frac{t(\mu_\theta - x)/(1-t) - x}{1-t} = \frac{t\mu_\theta - x}{(1-t)^2}, \qquad a_\theta = \lambda\,\mathrm{diag}\,\gamma_\theta = \frac{\lambda(t)}{(1-t)^2}{\Sigma_\theta}^x. \tag{73}$$

With these fields and $B_\theta = \sqrt{\lambda}\,\mathrm{diag}\,\sqrt{\gamma_\theta}$, the samplers are

$$dY_t = \frac{\mu_\theta(Y_t,t) - Y_t}{1-t}\,dt,$$
$$dY_t = \left[\frac{\mu_\theta - x}{1-t} + \tfrac{1}{2}\nabla_x\cdot a_\theta + \tfrac{1}{2}a_\theta\widehat{s}_\theta\right](Y_t,t)dt + B_\theta(Y_t,t)dW_t.$$

Both initialize from $\mathcal{N}(0,I)$. Their discrete versions use Algorithms 2 and 3 after constructing the fields in Eq. 73.

F.4 Readouts and parameterization correspondence

The endpoint mean μ_θ is predicted directly, while $v_\theta = (\mu_\theta - x)/(1-t)$ is derived. The variance outputs represent $\Sigma_\theta = \mathrm{diag}\exp\ell_\theta$ in velocity coordinates and ${\Sigma_\theta}^x = (1-t)^2\Sigma_\theta$ in endpoint coordinates. Thus both parameterizations supply the same types of posterior readouts in Section 4.3.

When the two mean parameterizations satisfy $\mu_\theta(x,t) = x + (1-t)v_\theta(x,t)$ and use the same velocity covariance field and schedule $\lambda(t)$, they yield the same velocity, score, diffusion tensor, and drift. The likelihood relation follows from Appendix B.3, and the residual relation follows from Appendix D.1. These field and likelihood correspondences explain the two mean parameterizations: the main text develops direct velocity prediction, while the ImageNet implementation in Appendix H.2 uses direct endpoint-mean prediction.

For conditional generation, evaluate $F^x_\theta(x,t,c)$ under the reference construction in Appendix E.1. Holding c fixed leaves the affine transformation unchanged. The likelihood averages over (C, X_1, X_0, t), and the fields above enter the conditional samplers and diagnostics of Appendices E.3–E.4.

G Training and Numerical Sampling

This appendix turns the objective and continuous dynamics of Sections 4.1–4.2 into training and sampling algorithms. The network $F_\theta(x,t)$ outputs a velocity mean v_θ and log variance ℓ_θ, with $\gamma_\theta = \exp \ell_\theta$ and $\Sigma_\theta = \operatorname{diag} \gamma_\theta$. We specify the joint training update, compute the variance derivatives required by the SDE drift, and then discretize both samplers. The final two subsections derive the effect of randomized derivatives on a sampling step and explain the terminal time and endpoint output. Multiplication $\odot$, exponentials, and square roots of vectors are elementwise.

Figure 8 expands Figure 1 with the application readouts, connecting the algorithms below to the task implementations in Appendix H.

G.1 Training update

We expand the single-loss training procedure in Section 4.1. Since Σ_θ is diagonal, its inverse has entries $\exp(-\ell_{\theta,i})$ and its log determinant is $\sum_i \ell_{\theta,i}$. Equation 13 is therefore

$$\mathcal{L}_{\mathrm{PFM}} = \frac{1}{2}\mathbb{E} \sum_i \left[\exp(-\ell_{\theta,i})(V_i - v_{\theta,i})^2 + \ell_{\theta,i}\right]. \tag{74}$$

The variance head produces an unconstrained value $\rho_{\theta,i}$, which is mapped to the log variance used in this loss by

$$\ell_{\theta,i} = \ell_- + (\ell_+ - \ell_-)\,\mathrm{sigmoid}(\rho_{\theta,i}), \qquad \gamma_{\theta,i} = \exp \ell_{\theta,i}. \tag{75}$$

For finite $\rho_{\theta,i}$, this gives an open variance interval. Interior optima in Eq. 44 are attained, while boundary optima are approached as $\rho_{\theta,i} \to \pm\infty$. The same bounded variance supplies the likelihood, diffusion tensor, and endpoint readouts. Table 7 lists the task-specific bounds. Algorithm 1 averages the loss over a minibatch and differentiates through this mapping and the complete shared model.

Algorithm 1 (Joint posterior training)**.**
Input: network F_θ, data law, parameter-independent time law r, and an optimizer.
Repeat until the specified training budget is reached:

1. Draw a minibatch of endpoints $X_1 \sim p_{\mathrm{data}}$. Independently draw $X_0 \sim \mathcal{N}(0, I)$ and $t \sim r$ for each item.
2. Form $X_t = (1-t)X_0 + tX_1$ and $V = X_1 - X_0$.
3. Evaluate $(v_\theta, \ell_\theta) = F_\theta(X_t, t)$ once and average the integrand of Eq. 74 over the minibatch.
4. Backpropagate this loss through both heads and the shared backbone (Rumelhart et al., 1986), and update θ with the optimizer.

Output: trained parameters θ.

The resulting model supplies the mean, variance, and derived score used by the samplers below. Appendix E.2 gives its conditional-generation extension.

G.2 Variance derivatives for the SDE drift

For a trained model, the ODE uses v_θ directly. The SDE also uses $\widehat{s}_\theta = (tv_\theta - x)/(1-t)$ and $a_\theta = \lambda(t) \operatorname{diag} \gamma_\theta$ for a chosen nonnegative schedule λ. The compatible drift derived in Appendix C.2 is

$$b_\theta = v_\theta + \frac{\lambda(t)}{2}\left(D + \gamma_\theta \odot \widehat{s}_\theta\right), \qquad D_i = \partial_{x_i}\gamma_{\theta,i}.$$

To compute the required derivatives at fixed time, define

$$(J_\gamma)_{ij} = \partial_{x_j}\gamma_{\theta,i}, \qquad D = \operatorname{diag} J_\gamma, \qquad D_i = \gamma_{\theta,i}\partial_{x_i}\ell_{\theta,i}. \tag{76}$$

Rows index variance outputs and columns index input coordinates. Here diag extracts a matrix diagonal. Applied to a vector, as in Eq. 11, it instead constructs a diagonal matrix. Appendix B.3 explains why J_γ can be dense.

Exact derivatives. A Jacobian–vector product (JVP) computes $J_\gamma \zeta$ for a direction ζ. A vector–Jacobian product (VJP), written as a column vector, computes $J_\gamma^\top \zeta$. For the standard basis vector

e_i,

$$D_i = (J_\gamma e_i)_i = (J_\gamma^\top e_i)_i. \tag{77}$$

Using the d basis directions gives the exact diagonal without storing the full Jacobian.

Randomized derivatives. Let $\zeta^{(1)}, \dots, \zeta^{(m)}$ be independent Rademacher vectors, whose coordinates independently take $+1$ and -1 with equal probability. The JVP estimator (Han et al., 2026) is

$$\widehat{D} = \frac{1}{m} \sum_{k=1}^{m} \zeta^{(k)} \odot J_\gamma \zeta^{(k)}. \tag{78}$$

The estimator is unbiased, with coordinate variance

$$\mathbb{E}[\widehat{D} \mid x, t] = D, \qquad \mathrm{Var}(\widehat{D}_i \mid x, t) = \frac{1}{m} \sum_{j \neq i} (J_\gamma)_{ij}^2. \tag{79}$$

Replacing $J_\gamma \zeta$ in Eq. 78 by $J_\gamma^\top \zeta$ gives the VJP estimator. It remains unbiased, with coordinate variance $m^{-1} \sum_{j \neq i} (J_\gamma)_{ji}^2$. Appendix G.4 derives these moments and their effect on the sampling update.

Model parameters remain fixed during sampling, while differentiation with respect to the current state retains the full variance-field dependence. The SDE therefore requires these derivative evaluations in addition to network forwards.

G.3 Numerical sampling

We discretize the samplers in Eqs. 16 and 19. Let $0 = t_0 < \cdots < t_K = T$ and $h_k = t_{k+1} - t_k$. Initialize $y_0 \sim \mathcal{N}(0, I)$. Euler ODE sampling uses (Lipman et al., 2024)

$$y_{k+1} = y_k + h_k v_\theta(y_k, t_k). \tag{80}$$

This update evaluates the mean at the beginning of each interval. Appendix G.5 explains the terminal-time convention.

Algorithm 2 (PFM ODE sampling)**.**
Input: trained F_θ, a grid ending at $T \leq 1$.
Draw $y_0 \sim \mathcal{N}(0, I)$.
For $k = 0, \dots, K-1$:

1. Set $h_k = t_{k+1} - t_k$ and evaluate $v_\theta(y_k, t_k)$.
2. If posterior diagnostics are requested, also record Σ_θ, μ_θ, and $\Sigma_\theta{}^x$ using Eq. 15 at this state.
3. Apply Eq. 80.

Output: y_K, with any recorded posterior readouts.

For SDE sampling, let $\xi_k \sim \mathcal{N}(0, I)$ be independent across steps and independent of initialization. The Brownian increment is represented by $\sqrt{h_k}\xi_k$. Euler–Maruyama gives (Song et al., 2021; Albergo et al., 2025)

$$y_{k+1} = y_k + h_k b_\theta(y_k, t_k) + \sqrt{h_k \lambda(t_k) \gamma_\theta(y_k, t_k)} \odot \xi_k. \tag{81}$$

Its Brownian step covariance is $h_k a_\theta(y_k, t_k)$. Algorithm 3 computes the drift from the same posterior outputs, using exact or randomized derivatives as detailed in Appendix G.2.

Algorithm 3 (PFM SDE sampling)**.**
Input: trained F_θ, schedule λ, derivative procedure, and a grid ending at $T < 1$.
Draw $y_0 \sim \mathcal{N}(0, I)$.
For $k = 0, \dots, K-1$:

1. Set $h_k = t_{k+1} - t_k$ and evaluate $(v_\theta, \ell_\theta) = F_\theta(y_k, t_k)$.
2. Compute $\gamma_\theta = \exp \ell_\theta$ and $\widehat{s}_\theta = (t_k v_\theta - y_k)/(1 - t_k)$.
3. Compute $D_i = \partial_{y_i} \gamma_{\theta,i}$, or its estimate $\widehat{D}_i$ from fresh probes. Form $b_\theta = v_\theta + \lambda(t_k)(D + \gamma_\theta \odot \widehat{s}_\theta)/2$, substituting $\widehat{D}$ when used.
4. Draw an independent $\xi_k \sim \mathcal{N}(0, I)$ and apply Eq. 81. Posterior readouts at the evaluated state follow from Eq. 15.

Output: y_K, with any recorded posterior readouts. At steps with $\lambda(t_k) = 0$, use Eq. 80 directly.

Setting $\lambda \equiv 0$ removes the Brownian term and both drift corrections, so Algorithm 3 reduces to Algorithm 2 without variance derivatives. Both implement the same trained model from Section 4.1.

FM sampling. For the FM rows of Table 1, apply Euler–Maruyama to Eq. 32 on the same grid:

$$\begin{aligned} y_{k+1} = y_k + h_k \left[v_\eta(y_k, t_k) + \tfrac{1}{2}\kappa(t_k)\widehat{s}_\eta(y_k, t_k) \right] \\ + \sqrt{h_k \kappa(t_k)}\xi_k. \end{aligned} \tag{82}$$

Each step uses the velocity and its derived score from Appendix A.4, with independent $\xi_k \sim \mathcal{N}(0, I)$. The prescribed diffusion is isotropic and has no spatial derivative term. Choosing $\kappa = 0$ gives Euler sampling of the FM ODE, completing the discrete baseline construction from Appendix A.6.

G.4 Effect of derivative estimation

We first derive the estimator moments in Eq. 79, then use them to quantify the additional randomness in Algorithm 3. For one probe at fixed (x, t), the error in coordinate i expands as

$$\zeta_i (J_\gamma \zeta)_i - D_i = \sum_{j \neq i} (J_\gamma)_{ij} \zeta_i \zeta_j.$$

Each summand has zero mean. Since $\zeta_i^2 = 1$ and $\mathbb{E}[\zeta_j \zeta_l] = 0$ for distinct j, l, the second moment is

$$\mathbb{E}\left[\left(\sum_{j \neq i} (J_\gamma)_{ij} \zeta_i \zeta_j \right)^2 \right] = \sum_{j \neq i} (J_\gamma)_{ij}^2.$$

Averaging independent probes divides this variance by m, giving Eq. 79. For VJP probes, the same calculation exchanges rows and columns.

Let $\mathcal{F}_k$ contain initialization and all randomness preceding step k, so y_k is measurable with respect to $\mathcal{F}_k$. Draw the current probes independently of this history and the current Brownian vector ξ_k. Replacing D by $\widehat{D}$ in Eq. 81 adds

$$\varepsilon_k = \frac{h_k \lambda(t_k)}{2} \left[\widehat{D}(y_k, t_k) - D(y_k, t_k) \right]. \tag{83}$$

By Eq. 79, the JVP version satisfies

$$\mathbb{E}[\varepsilon_k \mid \mathcal{F}_k] = 0, \qquad \mathrm{Var}(\varepsilon_{k,i} \mid \mathcal{F}_k) = \frac{h_k^2 \lambda(t_k)^2}{4m} \sum_{j \neq i} (J_\gamma(y_k, t_k))_{ij}^2. \tag{84}$$

For VJP probes the sum instead uses $(J_\gamma)_{ji}^2$. This identifies the additional step randomness introduced by the derivative procedure in Algorithm 3, separately from its Brownian increment.

G.5 Terminal time and endpoint output

This subsection specifies the terminal convention for the samplers in Section 4.2 and Algorithms 2–3. Euler ODE sampling may end at $T \leq 1$, while all coefficient evaluations occur at $t_k < 1$. The

SDE construction evaluates the score containing $(1-t)^{-1}$ and uses $T < 1$. Both algorithms return y_K at their chosen stopping time T. At $T < 1$, Eq. 15 also supplies the endpoint-mean readout $\mu_\theta(y_K, T)$.

For the continuous-time endpoint statement, suppose a single compatible process is defined on $[0, 1)$ and satisfies Appendix A.5 on every $[0, T]$ with $T < 1$. Its law is then p_t for every $t < 1$. The reference interpolation satisfies

$$X_t - X_1 = (1-t)(X_0 - X_1) \longrightarrow 0 \quad \text{almost surely as } t \uparrow 1.$$

For every bounded continuous test function ψ, dominated convergence gives

$$\int \psi \, dp_t = \mathbb{E}[\psi(X_t)] \longrightarrow \mathbb{E}[\psi(X_1)] = \int \psi \, dp_{\text{data}}.$$

This proves weak convergence of probability measures,

$$\text{Law}(Y_t) = p_t \Rightarrow p_{\text{data}} \quad \text{as } t \uparrow 1. \tag{85}$$

H SUPPLEMENTARY EXPERIMENTS

This appendix specifies how the formulation in Sections 4.1–4.3 is instantiated for the four tasks in Section 5. The task sections explain the state and conditioning choices, extend the image and action examples and point-cloud moment checks, and detail the molecular covariance comparison. Appendix H.6 connects one-dimensional sampling trajectories to the evolution of density perturbations. Appendix H.7 then examines the covariance choice, stochasticity scale, and sampling cost introduced in Section 5.3. Diagnostic intervals describe held-out-data bootstrap resampling, and displayed trajectory times t are rounded to one decimal.

H.1 IMPLEMENTATION DETAILS

Table 7 summarizes the model and training settings. Its log-variance bounds specify (ℓ_-, ℓ_+) in Eq. 75, whose effect on diagonal Gaussian moment recovery is derived in Appendix B.4. All four models jointly learn mean and variance through the PFM NLL, with the task-specific parameterizations described below; the ShapeNet latent model also uses a KL regularizer. Conditions follow Appendix E and remain implicit in the readout notation.

Table 7: Core implementation settings for the four generation tasks in Section 5. Parameter counts include both PFM output heads and trainable conditioning networks. Batch sizes are effective global batches. Maximum learning rates are reached after warmup, where used. The bounds apply to velocity log-variance outputs, including the pre-projection atomwise scales for molecules.

Setting	ImageNet-256	ShapeNet	GEOM-QM9	PushT
Generated state	256×256 RGB image	2,048 3D points	Centered atom coordinates	16 two-dimensional actions
Condition	Class label	Shape latent	Molecular graph and atom/bond features	Two image/agent-position observations
Backbone	JiT-B/16 (Li & He, 2026)	Pointwise network with PointNet encoder and latent prior (Luo & Hu, 2021)	SO(3)-equivariant TorchMD Transformer (Thölke & De Fabritiis, 2022; Hassan et al., 2024)	Image encoder and conditional 1D U-Net (Chi et al., 2025)
PFM parameters (M)	132.90	3.88	9.09	262.71
Model dimensions	12 blocks; width 768; 12 heads; patch 16	Hidden widths 128, 256, 512, 256, 128	20 layers; width 160; 8 heads	U-Net channels 512, 1,024, 2,048
Optimizer	AdamW (Loshchilov & Hutter, 2019)	Adam (Kingma & Ba, 2015)	AdamW	AdamW
Maximum learning rate	2×10^{-4}	2×10^{-3}	7×10^{-4}	10^{-4}
Warmup	5 epochs	None	2 epochs per 200-epoch stage	500 updates
Learning-rate decay	Constant after warmup	Linear to 10^{-4} over updates 200k–400k, then constant	Cosine to 10^{-8} in each 200-epoch stage	Cosine to zero
Effective batch	1,024	128	512	64
Training length	600 epochs	800k updates	400 epochs	3,050 epochs
Log-variance bounds	$[-1, 1]$	$[-3, 3]$	$[-4, 4]$	$[-3, 3]$

PFM SDE sampling uses the linear schedule $\lambda(t) = \lambda_0(1-t)$. The isotropic SDE ablation in Appendix H.7.1 uses the same temporal form, $\kappa(t) = \kappa_0(1-t)$. For each covariance-ablation comparison, we set $\kappa_0 = \lambda_0$, so the two SDEs share the same scalar noise schedule.

H.2 IMAGE GENERATION

For ImageNet (Deng et al., 2009; Russakovsky et al., 2015), the class label supplies C, and the JiT-B/16 model (Li & He, 2026; Dosovitskiy et al., 2021) generates images in pixel space. Its mean head predicts the endpoint directly using the parameterization in Appendix F. Its second head predicts velocity log-variance, so $v_\theta = (\mu_\theta - x)/(1-t)$ and ${\Sigma_\theta}^x = (1-t)^2\Sigma_\theta$ give the same generation fields and diagnostics as the main formulation.

CIFAR-10 (Krizhevsky, 2009) extends the image-generation comparison in Section 5.1, with a lower FID for the evaluated SDE (Table 8). Figure 9 shows generated images together with spatial variance

maps, following the main ImageNet presentation. Image metrics use 50,000 class-balanced samples per setting, with 250 integration steps for ImageNet and 50 for CIFAR-10. The CIFAR-10 SDE uses $\lambda_0 = 0.5$. Images show endpoint-mean readouts near the end of sampling; their variance maps come from an earlier state on the same trajectory and share a contrast transform.

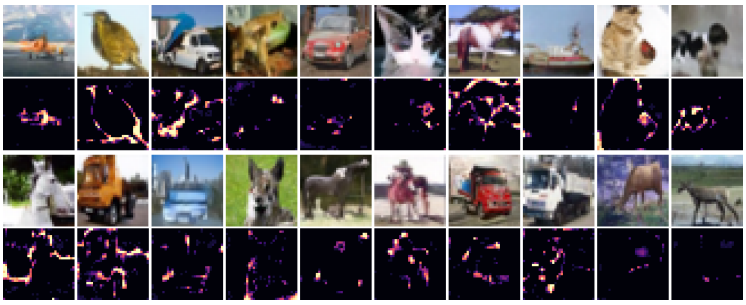


Figure 9: CIFAR-10 generation with posterior readouts. Twenty endpoint means are arranged in two rows of ten, each paired with its variance map directly below, as in Figure 2.

Table 8: CIFAR-10 FID.

Method	FID ↓
FM	3.75
Diffusion	4.12
PFM ODE	3.49
PFM SDE	3.03

H.3 Point-cloud generation

Each category uses a pointwise mean/variance network conditioned on a shape latent C (Luo & Hu, 2021). The PointNet encoder (Qi et al., 2017), latent prior, and PFM network are jointly trained with the PFM NLL and latent KL regularization of weight 10^{-3}. The KL term aligns the encoder distribution with the prior used to supply generation conditions (Kingma & Welling, 2014). The NLL is averaged over shapes, points, and coordinates, whereas the encoder-to-prior KL uses the full latent vector and is averaged over shapes. For a given encoder, the joint distribution of shape latents and data defines the conditional reference law in Appendix E.1. Generation samples C from the learned latent prior, shared by the ODE and SDE samplers of that category. Reference diagnostics condition on encodings of held-out shapes.

The generation visualizations in Figure 3 track the same points over time with a fixed camera and scale. The variance–residual correspondence persists across Chair and Airplane at six reference times, with more visible departures in low-variance and sparsely populated bins (Figure 10). This extends the selected-time comparisons in Section 5.2.

H.4 Molecular conformation generation

The molecular graph and atom/bond features supply C; generation predicts atom coordinates with their centroid fixed at zero. For N atoms, let $\mathcal{H}$ be this $3(N-1)$-dimensional subspace and P its orthogonal projector. The equivariant network (Thölke & De Fabritiis, 2022; Hassan et al., 2024) predicts a centered velocity and one positive, rotation-invariant variance scale $\gamma_{\theta,i}^{\mathrm{atom}}$ per atom. Writing the pre-projection covariance as G_θ gives

$$P = \left(I_N - \frac{\mathbf{1}\mathbf{1}^\top}{N}\right) \otimes I_3, \qquad G_\theta = \mathrm{diag}_{i=1}^{N}\left(\gamma_{\theta,i}^{\mathrm{atom}} I_3\right), \\ \Sigma_\theta = P G_\theta P, \qquad \Sigma_\theta{}^{x} = (1-t)^2 \Sigma_\theta. \tag{86}$$

This projected covariance replaces the ambient diagonal covariance in Eq. 12. Because centering couples atom coordinates, the projection introduces correlations between atoms. The reference noise is a standard Gaussian projected by P, and the NLL uses the covariance inverse and determinant restricted to $\mathcal{H}$, normalized by $3(N-1)$. The same posterior-driven construction gives

$$a_\theta = \lambda \Sigma_\theta, \qquad B_\theta = \sqrt{\lambda}\, P G_\theta{}^{1/2}, \\ b_\theta = v_\theta + \tfrac{1}{2}\left(\nabla_{\mathcal{H}} \cdot a_\theta + a_\theta \widehat{s}_\theta\right), \qquad \widehat{s}_\theta = \frac{t\, v_\theta - x}{1-t} \in \mathcal{H}. \tag{87}$$

Here $\nabla_{\mathcal{H}}$ denotes differentiation within the centered subspace. Both drift and noise stay in $\mathcal{H}$; numerical updates are recentered. Equations 86–87 apply the posterior readouts and compatible drift of Sections 4.2–4.3 to the molecular geometry.

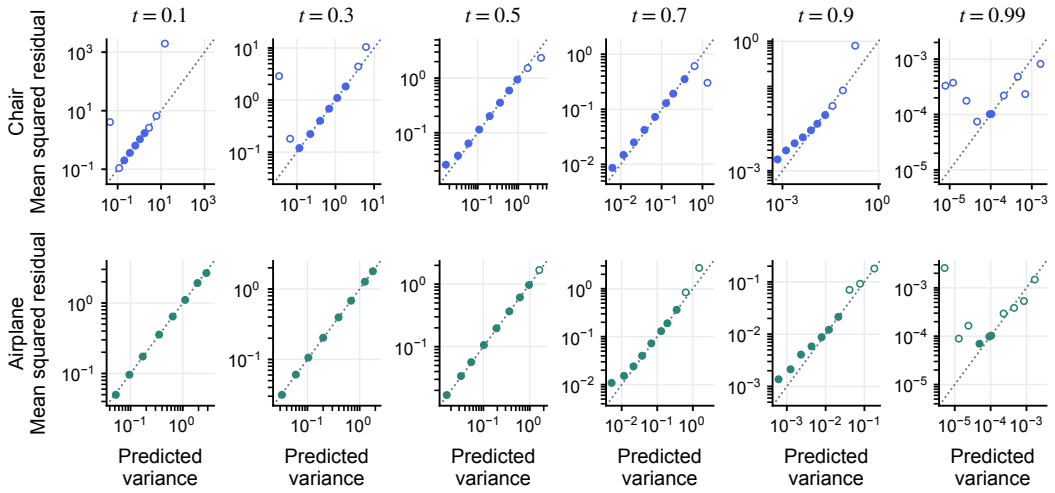


Figure 10: Point-cloud posterior diagnostics over six reference times: Chair (top) and Airplane (bottom). Axes compare bin-averaged endpoint coordinate variance and mean squared residual. Dotted lines indicate equality. Hollow markers indicate sparse bins.

To evaluate the divergence in Eq. 87, let $\gamma_\theta^{\mathrm{pre}} = \operatorname{diag} G_\theta \in \mathbb{R}^{3N}$ collect the pre-projection scales, repeating each atom's scale over its three coordinates. Let J_{pre} be its input Jacobian at fixed time and molecular condition. Since P is constant,

$$\nabla_{\mathcal{H}} \cdot a_\theta = \lambda(t) P \operatorname{diag}(J_{\mathrm{pre}} P).$$

The diagonal-extraction procedures in Appendix G.2 apply to $J_{\mathrm{pre}}P$, whose JVP and VJP are $J_{\mathrm{pre}}(P\zeta)$ and $PJ_{\mathrm{pre}}^\top\zeta$, respectively. Multiplying the exact diagonal or its randomized estimate by $\lambda(t)P$ gives the divergence used in the drift.

GEOM-QM9 coverage uses a 0.5 Å matching threshold. Performance metrics use chirality correction without force-field relaxation (Hassan et al., 2024). The visualizations and posterior checks use model readouts before this correction. For molecular sampling, $\lambda_0 = 0.3$ is selected by precision matching distance in the evaluated sweep.

To assess the learned covariance at a shared mean, Figure 6(d) compares it with a time-only baseline that is isotropic on $\mathcal{H}$. At each reference time, the baseline's scalar velocity variance is fitted by averaging the squared velocity residual norm divided by $3(N-1)$ over validation molecules, with equal weight per molecule. Both models use PFM's predicted mean. Gaussian NLL is evaluated in velocity coordinates, omits common constants, and is normalized by the subspace dimension $3(N-1)$. The learned covariance gives lower NLL at every evaluated time, supporting a better conditional Gaussian fit than a single variance scale per time. This comparison complements the variance–residual checks in Section 5.2 by evaluating covariance allocation while holding the mean fixed.

H.5 Action generation on PushT

The policy conditions on two observations, each containing a 96×96 image and the two-dimensional agent position; the image encoder uses 84×84 crops. Within each planning call, this condition C is fixed while the conditional 1D U-Net (Chi et al., 2025) generates a 16-step sequence of two-dimensional action targets. The controller executes eight steps and replans from new observations (Chi et al., 2025). This instantiates Appendix E within the control loop shown in Section 5.1. Control scores average the episode maximum normalized overlap over 50 test initial states; the table reports ODE and SDE scores at 20 and 50 integration steps. Appendix H.7.2 describes the noise-scale comparison at these two budgets. Figures 11 and 12 complement the two main examples with four additional episodes. Ellipses show endpoint action standard deviations in workspace coordinates, using the same scale as the actions.

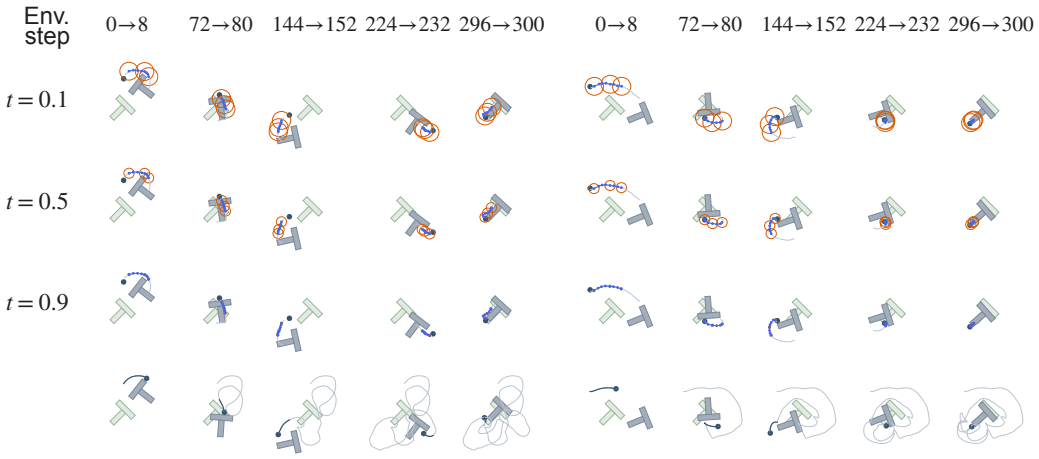


Figure 11: PushT posterior readouts for two additional episodes, following Figure 5. Columns indicate the environment steps from planning to execution, and t denotes generation time. Upper rows show endpoint action means μ_θ (blue) and standard-deviation ellipses from ${\Sigma_\theta}^x$ (orange). The bottom row shows execution.

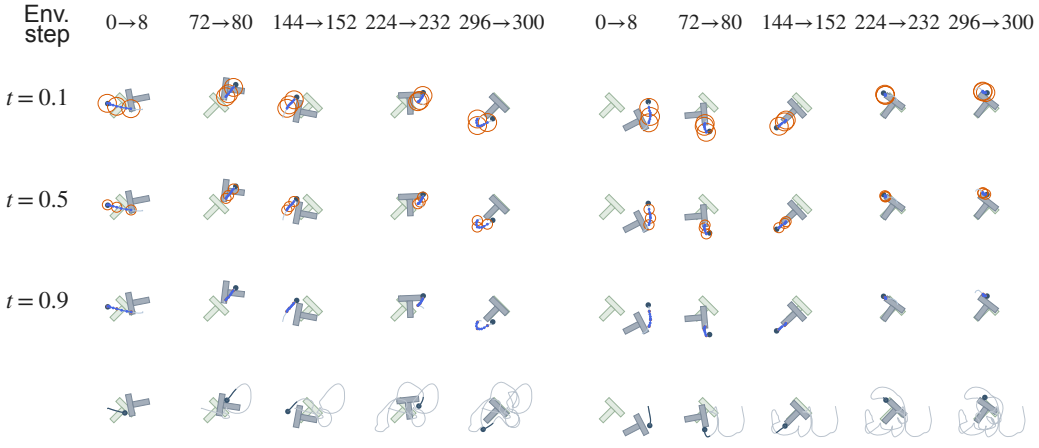


Figure 12: PushT posterior readouts for two additional episodes, following Figure 5. Columns indicate the environment steps from planning to execution, and t denotes generation time. Upper rows show endpoint action means μ_θ (blue) and standard-deviation ellipses from ${\Sigma_\theta}^x$ (orange). The bottom row shows execution.

H.6 One-dimensional generation dynamics

This section develops three views of the sampling construction in Section 4.2: transport along individual trajectories, adjustment of stochasticity within one posterior model, and the evolution of density perturbations. A one-dimensional Gaussian mixture provides the common target, with component weights $(0.25, 0.45, 0.30)$, means $(-2.5, -0.4, 2.0)$, and standard deviations $(0.22, 0.40, 0.28)$. Sampling starts from $\mathcal{N}(0, 1)$. PFM uses the linear reference path of Eq. 1, while diffusion uses a variance-preserving noising process with $f(x, \tau) = -6x$ and $g = \sqrt{12}$. Both networks have three hidden layers of width 128 with SiLU activations and are trained for 60,000 updates. PFM uses Eq. 13 with velocity log-variance bounded in $[-6, 4]$, and diffusion uses denoising score matching as in Appendix A.3.

Figure 13 compares diffusion and PFM under ODE and SDE sampling. Backgrounds show the respective analytic reference densities, while trajectories follow the learned fields. For PFM, the posterior mean supplies deterministic transport. The SDE pairs posterior-scaled noise with the drift correction in Eq. 19, so both terms change as stochasticity is introduced. The right panels show all four samplers resolving the three target modes. Sampling uses Euler or Euler–Maruyama with 2,000 uniform steps to $T = 0.999$, including the variance derivative in the full PFM drift. Generated densities use 20,000 samples.

To separate the learned noise allocation from its overall strength, Figure 14 holds the PFM checkpoint fixed and varies λ_0 in $\lambda(t) = \lambda_0(1 - t)$. In Eq. 18, $\Sigma_\theta(x, t)$ supplies the local covariance and $\lambda(t)$ scales it. All settings share initial particles and paired Brownian increments on an 8,000-step grid. The six scales connect deterministic transport at $\lambda_0 = 0$ to increasingly stochastic motion under the same learned covariance field, evaluated along each sampler's trajectory. This gives a trajectory view of the adjustment studied in Section 5.3.

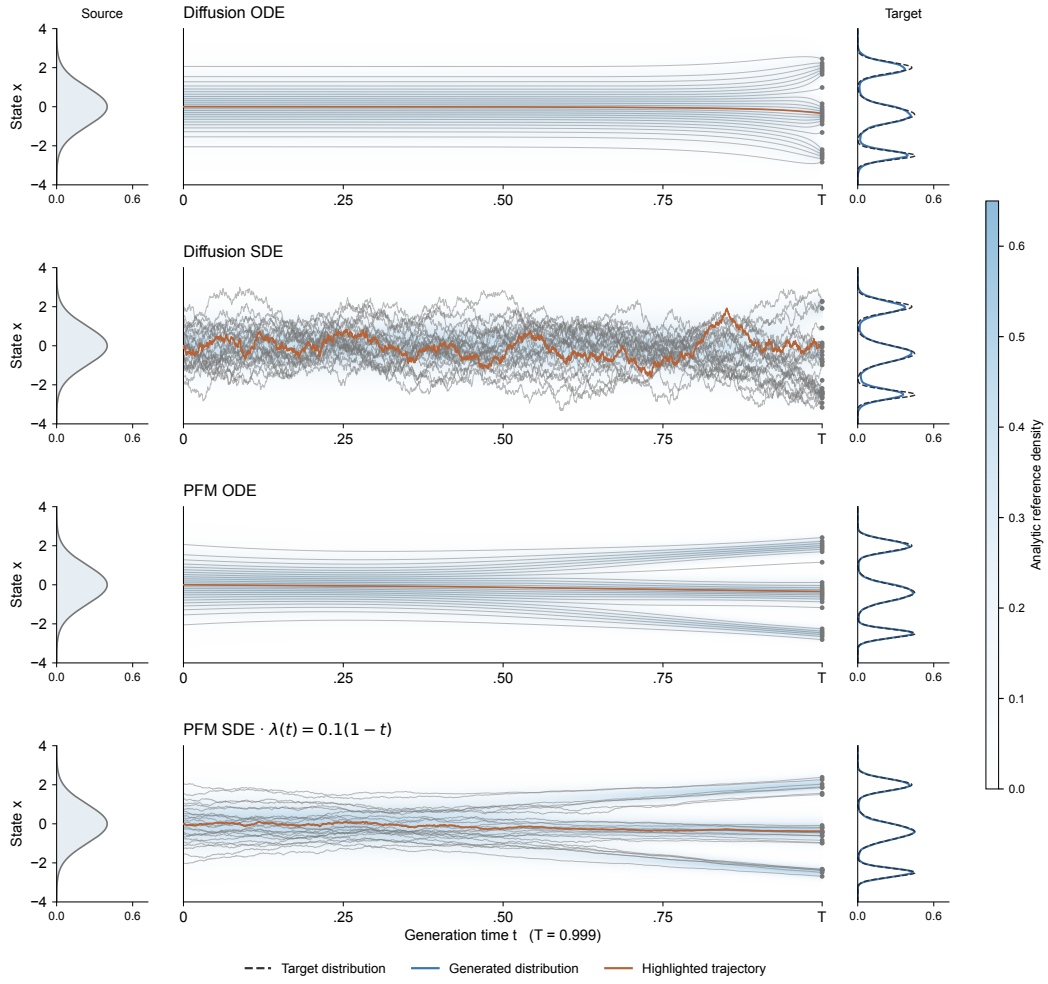


Figure 13: Learned diffusion and PFM dynamics on a one-dimensional Gaussian mixture. Gray curves are sampled trajectories, with one highlighted in orange. Backgrounds show each model family's analytic reference density. Right panels compare generated densities (blue) with the target (dashed). The PFM SDE uses $\lambda(t) = 0.1(1 - t)$. All rows use 2,000 steps to $T = 0.999$.

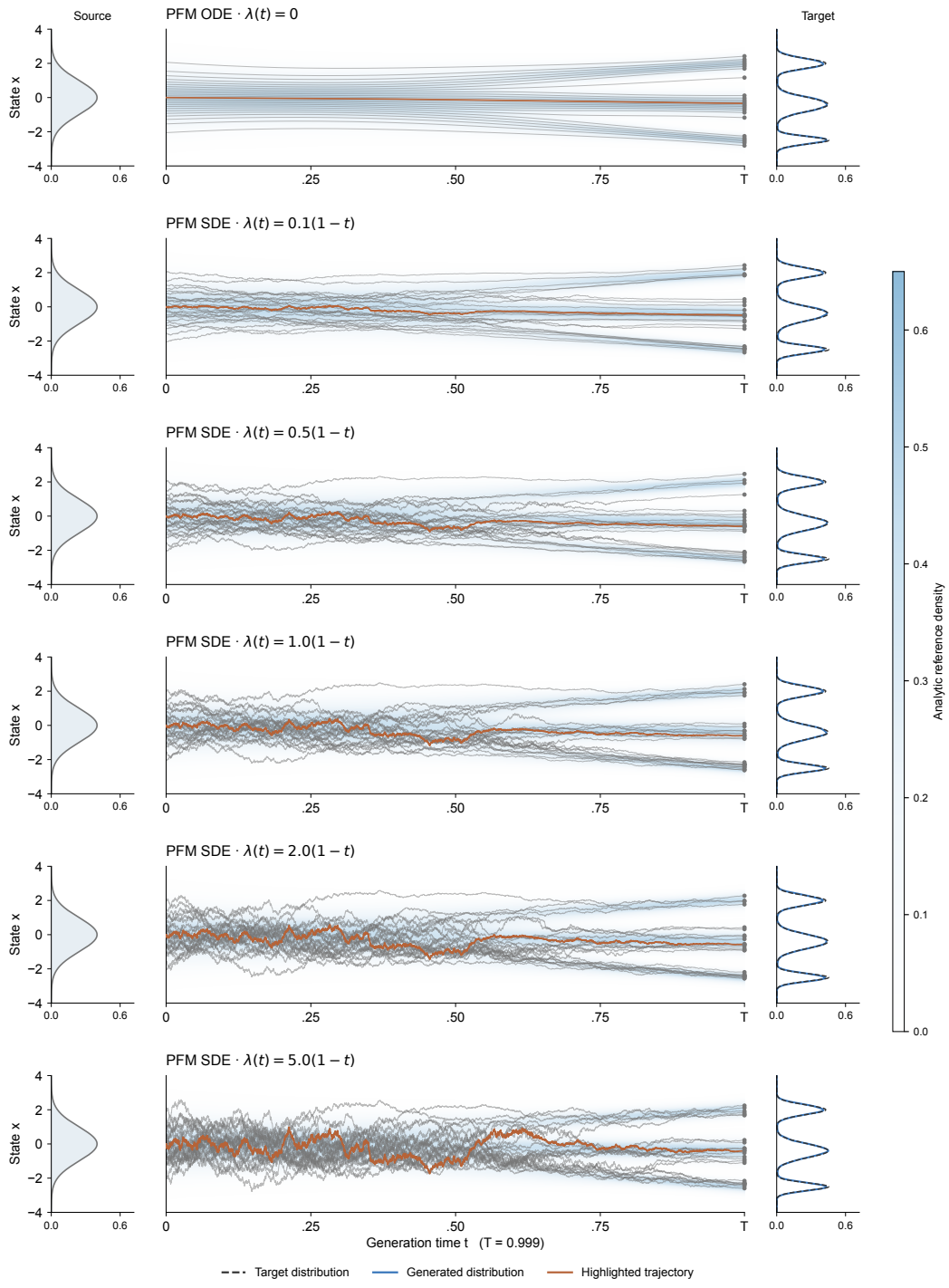


Figure 14: PFM trajectories for $\lambda(t) = \lambda_0(1-t)$, using one checkpoint, shared initial particles, and paired Brownian increments. Rows use $\lambda_0 \in \{0, 0.1, 0.5, 1, 2, 5\}$ with 8,000 steps to $T = 0.999$. The zero-noise row is the PFM ODE. Colors and density curves follow Figure 13.

The trajectory views lead to a distributional question: how does an initial density perturbation evolve under the compatible dynamics? To connect this question to Proposition 4.2, we now use the mixture's analytic conditional velocity moments and score on the same PFM reference path, with $\lambda(t) = 1 - t$. For each sampler, p_t and $\widetilde{p}_t$ share these coefficients and start from $p_0 = \mathcal{N}(0, 1)$ and $\widetilde{p}_0 = \mathcal{N}(0.5, 1)$, giving initial KL 0.125. The shifted initialization makes the density discrepancy visible while keeping the dynamics fixed. We evaluate the SDE density evolution through Eq. 49 up to $T = 0.999$ and use KL conservation for the ODE.

Figure 15(a,b) visualizes these dynamics from the same initial samples drawn from $\widetilde{p}_0$, against the reference density p_t. The ODE transports each sample deterministically, while the PFM SDE adds fluctuations scaled by the posterior variance. Panel (c) quantifies the resulting distributional difference. The ODE preserves KL at 0.125 even as the distributions change shape. Under the PFM SDE, KL decreases to 0.02128, showing contraction of the initial density perturbation. Panel (d) connects this decrease to the integrated posterior-weighted dissipation in Eq. 51: both reach approximately 0.10373 at T. Thus the trajectory and density views link stochastic motion along individual paths to the relative-entropy dissipation established in Proposition 4.2.

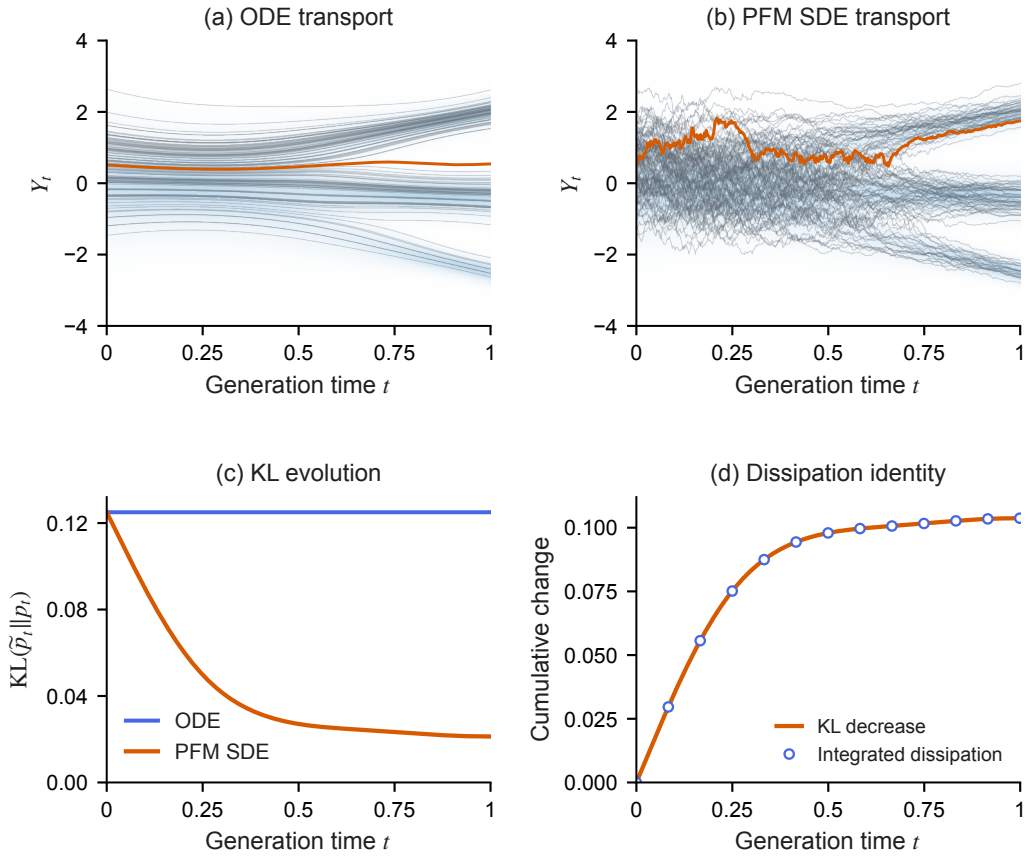


Figure 15: Density perturbation under analytic PFM dynamics on the Gaussian-mixture reference path of Figures 13–14, with $p_0 = \mathcal{N}(0, 1)$ and $\widetilde{p}_0 = \mathcal{N}(0.5, 1)$. (a,b) ODE and PFM SDE trajectories from shared samples of $\widetilde{p}_0$, with $\lambda(t) = 1 - t$ for the SDE. Backgrounds show p_t, and orange highlights one trajectory. (c) Density-based KL evolution. (d) SDE KL decrease and integrated dissipation from Eq. 51.

H.7 Ablation Studies and Sampling Cost

This section expands Section 5.3 around the two factors in $a_\theta = \lambda(t)\Sigma_\theta$. The covariance field determines local noise allocation, and the schedule controls its overall strength. We first compare covariance choices at a shared mean model, then vary stochasticity within the posterior-driven sampler, and finally compare sampling times.

H.7.1 Role of posterior covariance

PFM ODE and PFM SDE use the same learned velocity v_θ, with the SDE additionally using Σ_θ to allocate noise. An isotropic alternative allows us to examine this allocation while keeping the mean model fixed. We replace the posterior-driven diffusion tensor by $a_{\mathrm{iso}}(t) = \kappa(t)I$ with the shared scalar schedule $\kappa(t) = \lambda(t)$ specified in Appendix H.1. Using the same derived score $\widehat{s}_\theta$ from Eq. 16, the correspondence in Eq. 7 gives

$$dY_t = \left[v_\theta(Y_t, t) + \frac{\kappa(t)}{2} \widehat{s}_\theta(Y_t, t) \right] dt + \sqrt{\kappa(t)}\, dW_t. \tag{88}$$

The spatial divergence of this time-only tensor is zero. For molecular conformations, isotropy is defined within the centered subspace $\mathcal{H}$ of Appendix H.4: the ambient tensor is $\kappa(t)P$ and the noise is $\sqrt{\kappa(t)}P\, dW_t$. All three samplers use the same trained PFM model, with conditions fixed as in Appendix E. The isotropic variant thus changes how the PFM model is sampled, rather than using the separately trained diffusion model.

Table 9 organizes this comparison using the task metrics from Section 5.1. PFM ODE provides the deterministic reference, while the two SDEs compare prescribed isotropic noise with posterior-dependent noise and their corresponding drift corrections. The PFM ODE and PFM SDE entries reproduce the main generation results. Relative to isotropic sampling, PFM SDE reduces ImageNet FID from 24.31 to 22.28 and raises the PushT score from 0.882 to 0.914 at 50 steps. The isotropic alternative gives smaller gains on ImageNet, Chair, and PushT, but lowers molecular coverage and worsens Airplane MMD and 1-NNA relative to PFM ODE. With the trained PFM model fixed, posterior-driven SDE sampling improves generation quality over isotropic SDE sampling across all four tasks.

Table 9: Covariance choice for sampling a shared PFM model. PFM ODE and PFM SDE entries reproduce Tables 4, 5, 3, and 2. PFM + isotropic SDE uses Eq. 88. Metrics follow the corresponding main tables. Best values are bold.

ImageNet-256

Sampler	FID ↓	IS ↑
PFM ODE	24.45	65.44
PFM + isotropic SDE	24.31	65.53
PFM SDE	**22.28**	**68.06**

PushT

Sampler	20 steps ↑	50 steps ↑
PFM ODE	0.827	0.878
PFM + isotropic SDE	0.833	0.882
PFM SDE	**0.848**	**0.914**

ShapeNet (CD). Ideal 1-NNA: 50%.

Sampler	Chair			Airplane		
	MMD×10^3 ↓	COV (%) ↑	1-NNA (%)	MMD×10^3 ↓	COV (%) ↑	1-NNA (%)
PFM ODE	11.57	50.98	59.95	3.23	50.13	62.70
PFM + isotropic SDE	11.46	51.34	59.68	3.32	50.32	62.78
PFM SDE	**10.50**	**51.58**	**58.23**	**3.06**	**51.80**	**60.23**

GEOM-QM9

Sampler	COV-R (%) ↑	COV-P (%) ↑	MAT-R (Å) ↓	MAT-P (Å) ↓
PFM ODE	96.44	93.99	0.075	0.108
PFM + isotropic SDE	96.14	93.58	0.078	0.123
PFM SDE	**98.37**	**96.42**	**0.063**	**0.081**

H.7.2 INFERENCE-TIME STOCHASTICITY

The stochasticity adjustment in Section 5.3 varies λ_0 in the schedule of Appendix H.1 while keeping the learned posterior covariance field fixed. Since λ depends only on time, the drift correction in Eq. 19 scales linearly with λ_0, while the Brownian coefficient scales with $\sqrt{\lambda_0}$. The adjustment therefore changes the full compatible dynamics, with covariance retaining its role in local noise allocation.

The quality trends in Figure 7 distinguish this adjustable strength from the learned covariance: increasing the scale does not give a uniform improvement, and the two PushT integration budgets favor different ranges. Thus the scale provides a sampling choice within one trained model whose effect depends on the task and integration budget. Appendix H.6 gives the corresponding trajectory view: its six settings show how changing this scale alters individual paths under the same learned covariance field.

H.7.3 SAMPLING TIME

Table 10 expands the cost comparison in Section 5.3 across the four generation tasks. FM ODE and Diffusion SDE pair each reference model with its named sampler, while PFM ODE and PFM SDE use the same trained posterior model. Relative to PFM ODE, PFM SDE takes $1.16\times$ as long on ImageNet-256, $1.31\times$ on ShapeNet, $1.33\times$ on GEOM-QM9, and $1.17\times$ on PushT. PFM ODE and FM ODE times differ by at most 0.05 ms per sample. Together with the generation results in Section 5.1, these measurements describe the quality and time tradeoff between the two PFM samplers.

A PFM forward evaluation jointly returns the mean and variance. PFM ODE uses the mean for transport, while the full PFM SDE also evaluates the variance-field derivatives required by its drift correction. Appendix G.2 describes their exact and randomized computation.

Table 10: Sampling time in milliseconds per generated sample on NVIDIA B200. Columns distinguish each model and its sampler. PushT timings use 50 integration steps.

Task	FM ODE	Diffusion SDE	PFM ODE	PFM SDE
ImageNet-256	3.43	3.78	3.42	3.98
ShapeNet	1.87	2.11	1.88	2.47
GEOM-QM9	2.18	2.42	2.17	2.89
PushT	4.25	4.49	4.30	5.04