Heavy-Tailed Flow Matching via Random Clocks

Zhouhao Yang*
Department of Applied Mathematics and Statistics
Johns Hopkins University
zyang145@jhu.edu

Yezhen Wang*
School of Computing
National University of Singapore
e1154541@u.nus.edu

Kenji Kawaguchi
School of Computing
National University of Singapore
kenji@comp.nus.edu.sg

Vladimir Braverman
Department of Computer Science
Johns Hopkins University
vova@cs.jhu.edu

Haoyang Cao
Department of Applied Mathematics and Statistics
Johns Hopkins University
hycao@jhu.edu

Abstract

Heavy-tailed data arise in many domains where rare events carry disproportionate importance, such as imbalanced image datasets, financial returns, and weather extremes. Standard diffusion and flow-matching models typically begin from Gaussian noise or Gaussian source distributions, which yield tractable training targets but provide a poor inductive match for heavy-tailed data. We propose Heavy-Tailed Flow Matching via Random Clocks (HTFM), a framework that portrays heavy-tailed sources as mixtures of clock-conditioned Gaussian sources. Conditioning on a given clock path, the source distribution and flow are Gaussian; marginalizing over the clock gives a Gaussian scale mixture covering Gaussian, α-stable, and Student-t families. To make the clock-conditioned vector field practical, we encode the path-valued clock using truncated logsignature features, allowing the velocity field to adapt to the realized conditional space with negligible overhead. Empirically, on 2D imbalanced α-stable mixtures, CIFAR10-LT, and HRRR weather fields, HTFM improves mode coverage, sample quality, and tail-statistic recovery over Gaussian flow matching and competitive heavy-tailed baselines, while retaining the low-NFE sampling advantage of flow matching. Moreover, the random-clock formulation further provides a practical tail-control interface: by varying only the clock law or tail parameter, the same architecture can calibrate the "heaviness" of generated tails across different distribution families.

1 Introduction

In many real-world applications, rare or atypical observations carry disproportionately large impact relative to typical ones [9, 45]: hurricanes and heatwaves in operational weather and climate models [46, 13, 42, 16], the tail of asset returns in financial risk modeling [10, 14, 11], and minority classes in long-tailed image datasets [6, 36, 51]. A generative model is useful in such regimes only insofar as it reproduces these rare events with the right frequency and magnitude.

This requirement is difficult for two related reasons. First, extreme observations are rare by construction, so empirical training data contains few samples from the relevant tail regions [16]. Second, many deep generative modeling pipelines begin from simple light-tailed randomness: variational autoencoders and generative adversarial networks commonly use Gaussian or uniform latent

*Equal contribution.

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

priors [25, 15], while state-of-the-art diffusion and flow-matching models typically use Gaussian noising processes or Gaussian source distributions [20, 48, 33]. When the data distribution has substantially heavier tails, this light-tailed source is a poor inductive match: the model must learn tail behavior through the learned transport, which can make rare events difficult to reproduce accurately.

A natural response is to replace the Gaussian source or noise with a heavy-tailed alternative. In diffusion and score-based modeling, Lévy-Itô models [50, 44] use α-stable noise but require nonlocal fractional-score machinery and are tied to a particular stable family. Denoising Lévy probabilistic models (DLPM) [47] adapt the DDPM [20] construction to α-stable noise and use a denoising objective with robust aggregation over heavy-tailed randomness. Student-t models [41] include γ-divergence-based t-EDM with variance preconditioning ($\nu > 2$) and ordinary squared-error noise prediction in t-Flow. Other constructions specialize the source or noise to heavy-tailed families such as Cauchy noise [32] or finite-activity jump diffusions [4]. Related tail limitations have also been studied for normalizing flows [21, 19]. These approaches demonstrate the value of heavy-tailed generative mechanisms, but they are typically built around a specific tail family and the corresponding score, kernel, sampler, or loss aggregation. In practice, the appropriate tail mechanism is rarely known in advance; a method built for one heavy-tailed family may be misspecified for another.

This leaves a gap for flow matching [33, 35, 1]. Flow matching learns an ordinary differential equation (ODE) that transports a source distribution to the data distribution by regressing tractable velocity targets along prescribed probability paths, rather than simulating a long reverse diffusion chain. This often enables high-quality sampling with a substantially lower number of function evaluations (NFE), and rectified-flow and optimal-transport variants further sharpen this sample-efficiency advantage [34, 49, 43, 27]. A heavy-tailed analogue of flow matching should therefore preserve the low-NFE advantages of flow matching while allowing source laws with heavier tails than Gaussians. Existing heavy-tailed diffusion and flow constructions make important progress, but their targets or divergences are usually derived separately for each chosen heavy-tailed family.

In this paper, we propose *Heavy-Tailed Flow Matching via Random Clocks* (HTFM). A random clock is a nondecreasing path-valued latent variable that controls the conditional covariance of the source distribution. The source is Gaussian given the clock; additionally conditioning on a data endpoint yields a Gaussian affine interpolation and its tractable velocity target. After marginalizing over the clock, the source and flow can be heavy-tailed.

The random-clock viewpoint gives a unified construction for several representative tail mechanisms. A deterministic clock recovers vanilla Gaussian flow matching; an $\alpha/2$-stable subordinator clock yields α-stable source marginals; and an inverse-gamma clock recovers Student-t marginals for any $\nu > 0$, including infinite-variance $\nu \leq 2$; finite-risk guarantees additionally require the assumptions in Section 3.1. Across these choices, the endpoint-conditioned target is derived in the same clock-conditioned Gaussian space, avoiding family-specific score formulas, direct use of generally unavailable α-stable densities, or Student-t-specific kernels.

The remaining difficulty is that the pathwise optimal vector field is determined by the realized clock path, an infinite-dimensional object. Passing the full path into the velocity field would amount to an operator-learning problem [28] and would undermine the practical simplicity of flow sampling. We instead characterize the clock path by a finite-dimensional truncated logsignature feature [37, 8]. Signature coordinates approximate continuous functions on compact path sets as the order increases and support kernels, feature maps, and differentiable neural components [31, 26, 22, 23, 39]. Clock conditioning thus adds no velocity-field evaluations.

Empirically, our experiments support three findings about the random clock: (i) heavy-tailed clocks improve over the Gaussian-clock variant across benchmarks; (ii) different clock families behave differently even at matched polynomial tail-decay rates, indicating that the clock family is a meaningful design knob beyond tail heaviness; and (iii) explicit clock conditioning of the velocity field provides gains beyond simply sampling from a heavy-tailed source.

The contributions of our paper can be summarized as follows: 1. **A random-clock framework for heavy-tailed flow matching.** We introduce HTFM, a flow-matching framework in which a path–valued random clock controls the conditional source covariance. Conditioning on the realized clock gives a Gaussian conditional source and a tractable endpoint-conditioned affine target; marginalizing over clocks recovers Gaussian scale-mixture families, including Gaussian, α-stable, and Student-t sources. 2. **Signature-based clock conditioning.** We encode the path-valued clock using truncated

logsignature features, allowing the neural velocity field to adapt to the realized conditional Gaussian space without additional velocity-field evaluations. This provides a finite-dimensional interface for clock-conditioned vector fields. 3. **Empirical results.** Experiments on 2D imbalanced α-stable mixtures, CIFAR10-LT, and HRRR weather fields show that heavy-tailed clock laws improve over the Gaussian-clock variant and that the clock family remains important even at matched polynomial tail-decay rates. On CIFAR10-LT, HTFM compares favorably with existing heavy-tailed diffusion and ODE-based baselines, especially at low NFE; ablations further show that signature conditioning improves over using a heavy-tailed source alone.

2 Preliminaries

2.1 Flow matching

Flow matching learns a time-dependent vector field whose ODE transports a source distribution p_0 to the data distribution p_{data}. Let $X_t \in \mathbb{R}^d$ denote the state of the flow at time $t \in [0,1]$, a stochastic process with $X_0 \sim p_0$ and marginal law $p_t = \mathrm{Law}(X_t)$. Concretely, consider the ODE flow

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = u_t(X_t), \qquad t \in [0,1], \qquad X_0 \sim p_0, \tag{2.1}$$

whose marginal density evolves according to the continuity equation. The goal is to choose u_t so that $p_1 = p_{\mathrm{data}}$. Given the probability flow $(p_t)_{t\in[0,1]}$ and its marginal velocity field u_t, the ideal marginal flow-matching (FM) objective is

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{t\sim\mathrm{Unif}(0,1)}\mathbb{E}_{X_t\sim p_t}\Big[\big\|u_\theta(X_t,t) - u_t(X_t)\big\|_2^2\Big]. \tag{2.2}$$

However, the marginal field u_t is usually intractable. Conditional flow matching (CFM) therefore conditions on a data endpoint $X_1 \sim p_{\mathrm{data}}$ and specifies an endpoint-conditioned path $p_t(\cdot \mid X_1)$ with tractable conditional velocity $v_t(\cdot \mid X_1)$. The corresponding objective is

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{t\sim\mathrm{Unif}(0,1)}\mathbb{E}_{X_1\sim p_{\mathrm{data}}}\,\mathbb{E}_{X_t\sim p_t(\cdot|X_1)}\Big[\big\|u_\theta(X_t,t) - v_t(X_t \mid X_1)\big\|_2^2\Big]. \tag{2.3}$$

Under standard L^2-integrability and differentiability assumptions, the conditional and marginal objectives differ only by a θ-independent constant, so their gradients agree [33, 49]. This L^2-projection principle is the basic mechanism we preserve in our framework proposed later in Section 3.

2.2 Heavy-tailed noises as Gaussian scale mixtures

Gaussian and heavy-tailed laws are primarily distinguished by their tail decay. Gaussian tails are exponentially small, whereas many heavy-tailed laws exhibit only polynomial decay. Such laws are often better suited to modeling rare but non-negligible large deviations.

A convenient latent-Gaussian framework for symmetric heavy-tailed modeling is provided by Gaussian scale mixtures. Let V be a latent scale variable and let $G \sim \mathcal{N}(0, I_d)$ be independent of V. A random vector $X \in \mathbb{R}^d$ is called a Gaussian scale mixture if it admits a representation of the form $X = L(V)G$, where $L(V) \in \mathbb{R}^{d\times d}$ is a measurable random matrix. Equivalently,

$$X \mid V \sim \mathcal{N}\big(0, \Sigma(V)\big), \qquad \Sigma(V) := L(V)L(V)^\top. \tag{2.4}$$

Thus, conditioning on the latent variable V, the law remains Gaussian, whereas marginalizing over specific choices of V can recover a heavy-tailed marginal law. The mechanism is simple: rare large values of the latent scale generate broad conditional Gaussians, and the mixture assigns non-negligible probability to large deviations even though every conditional component is light-tailed. Representative examples include symmetric α-stable laws and Student-t laws; their concrete scale-mixture forms are recalled in Appendix A.3.

This latent-Gaussian viewpoint suggests a natural strategy for flow matching. Conditioning on both the mixing object and a data endpoint gives a Gaussian interpolation and tractable target. The L^2 projection argument additionally requires square-integrable regression residuals under the data, time, and mixing laws (Section 3.1). This conditional-Gaussian perspective is also closely related to [47], which uses a conditioned representation of α-stable noise to build tractable denoising targets for diffusion models.

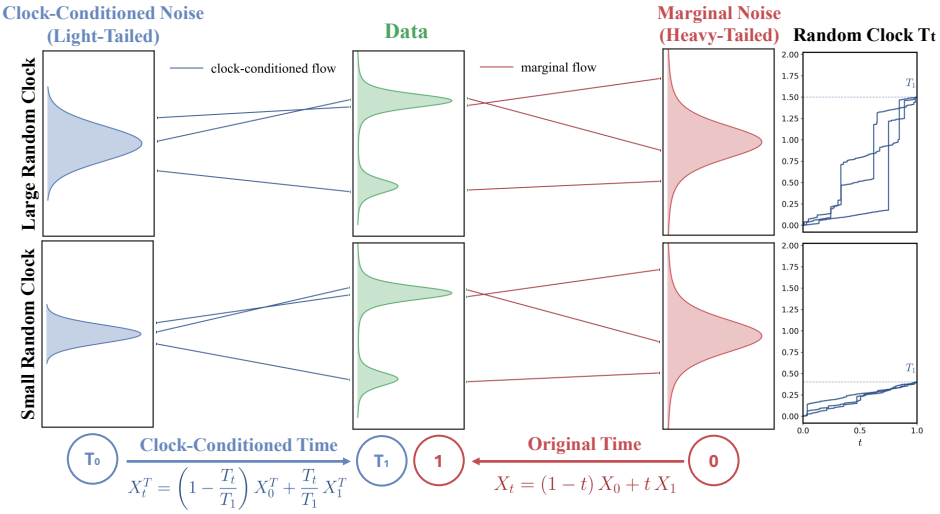

Figure 1: Overview of random-clock conditioning for heavy-tailed flow matching. The right column shows two realized clock paths: a large clock realization (top) and a small clock realization (bottom). For each fixed clock path T, the source distribution is conditionally Gaussian, and the model learns a clock-conditioned flow from this light-tailed conditional source to the data distribution. Larger clock realizations induce broader conditional source distributions, while smaller clock realizations induce more concentrated ones. After marginalizing over clock paths, the mixture of these conditional Gaussian sources yields a heavy-tailed marginal source distribution.

The Gaussian scale-mixture viewpoint is already broad enough to cover many symmetric heavy-tailed laws and will be the focus of the main development in this paper. A broader latent-Gaussian framework, obtained by allowing a nonzero conditional mean in (2.4) and hence passing from Gaussian scale mixtures to Gaussian mean-variance mixtures, can also accommodate more general families of asymmetric heavy-tailed laws; we defer this extension to Appendix A.4.

3 Heavy-Tailed Flow Matching via Random Clocks

This section develops the random-clock framework for heavy-tailed flow matching; a related diffusion-model extension is given in Appendix B.

The latent-Gaussian representation in Section 2.2 suggests replacing the scalar mixing variable V by a path-valued latent object, if we want to work on flow matching. We call this object a **random clock**. Formally, let $\mathbb{D}([0,1];\mathbb{R}_+) := \{\tau : [0,1] \to \mathbb{R}_+ \mid \tau \text{ is càdlàg and nondecreasing}\}$, equipped with its Borel σ-field. A random clock is a random element on a probability space $(\Omega_T, \mathcal{F}_T, \mathbb{P}_T)$, $T : \Omega_T \to \mathbb{D}([0,1];\mathbb{R}_+)$, and we write $T_{[0,1]} = (T_t)_{t\in[0,1]}$ for the full realized clock path. Throughout, $T_0 = 0$ almost surely.

The clock is sampled independently of the base randomness used to construct the source and target variables, such as a standard Gaussian G and $X_1 \sim p_{\text{data}}$. Clock-dependent sources are nevertheless allowed: for example, $X_0 = L(T_{[0,1]})G$ is conditionally Gaussian given the realized clock path but is not independent of the clock after it is formed.

3.1 Clock-conditioned dynamics and objectives

Motivated by Section 2.2, we formulate heavy-tailed flow matching **pathwise**, that is, conditioning on a realized clock path, and reserve the outer averaging over the random clock for the final objective.

Let $T = (T_t)_{t\in[0,1]}$ be a random clock[2], and fix a realization of the full clock path $T_{[0,1]}$. We write X_t^T, p_t^T, and u_t^T for the clock-conditioned process, its law $p_t^T := \text{Law}(X_t^T)$, and its velocity field

[2]When appearing as a superscript, T indicates the clock path.

under this realized clock path. Due to the independence of the data law and the random clock, we write $X_1^T := X_1$ for the terminal data variable, and $p_{\mathrm{data}}(\cdot \mid T_{[0,1]}) = p_{\mathrm{data}}$. We consider the clock-conditioned ODE

$$\frac{\mathrm{d}X_t^T}{\mathrm{d}t} = u_t^T(X_t^T), \qquad t \in [0,1], \qquad X_0^T \sim p_0^T := p_0(\cdot \mid T_{[0,1]}), \tag{3.1}$$

whose conditional density satisfies the continuity equation $\partial_t p_t^T + \nabla \cdot (p_t^T u_t^T) = 0$. Equivalently, one may view $u_t^T(\cdot)$ as the evaluation at $T_{[0,1]}$ of a measurable map $u_t : \mathbb{R}^d \times \mathbb{D}([0,1]; \mathbb{R}_+) \to \mathbb{R}^d$. Our goal is to construct, for each realized $T_{[0,1]}$, a flow from p_0^T at time $t = 0$ to the terminal law p_{data} at time $t = 1$.

The clock sets the conditional source scale and may modulate the interpolation schedule. The source is Gaussian given the clock; the affine interpolation is Gaussian given both clock and endpoint. Its clock-only marginal need not be Gaussian, so projection requires the integrability conditions below. An illustration is provided in Figure 1.

Training objective. For brevity, after fixing $T_{[0,1]}$ we write $\mathbb{E}_{X_t^T}$ for expectation under p_t^T, and $\mathbb{E}_{X_t^T \mid X_1^T}$ for expectation under $\mathrm{Law}(X_t^T \mid X_1^T, T_{[0,1]})$. The pathwise marginal quadratic risk and the corresponding heavy-tailed flow-matching objective are

$$\mathcal{L}_{\mathrm{HTFM}}(\theta) := \mathbb{E}_{t\sim\mathrm{Unif}(0,1)}\mathbb{E}_T \underbrace{\left[\mathbb{E}_{X_t^T}\left[\|u_\theta(X_t^T, t, T_{[0,1]}) - u_t^T(X_t^T)\|_2^2\right]\right]}_{=: R_t^{\mathrm{marg}}(\theta;T)}. \tag{3.2}$$

The risk $R_t^{\mathrm{marg}}(\theta; T)$ is typically intractable because the pathwise marginal velocity u_t^T is unknown. As in [33, 35], we therefore introduce a tractable endpoint-conditioned formulation: for each fixed realized clock path $T_{[0,1]}$, denote a conditional path $p_t^T(\cdot \mid x_1) := \mathrm{Law}(X_t^T \mid X_1^T = x_1)$ and a corresponding pathwise conditional velocity field $v_t^T(\cdot \mid x_1) : \mathbb{R}^d \to \mathbb{R}^d$ such that, for a.e. $t \in [0,1]$,

$$\begin{aligned} p_t^T(x) &= \int_{\mathbb{R}^d} p_t^T(x \mid x_1)\, p_{\mathrm{data}}(\mathrm{d}x_1), \\ p_t^T(x)\, u_t^T(x) &= \int_{\mathbb{R}^d} p_t^T(x \mid x_1)\, v_t^T(x \mid x_1)\, p_{\mathrm{data}}(\mathrm{d}x_1). \end{aligned} \tag{3.3}$$

The corresponding endpoint-conditioned heavy-tailed flow-matching objective writes

$$\mathcal{L}_{\mathrm{CHTFM}}(\theta) := \mathbb{E}_{t\sim\mathrm{Unif}(0,1)}\mathbb{E}_{T,X_1^T} \underbrace{\left[\mathbb{E}_{X_t^T \mid X_1^T}\left[\|u_\theta(X_t^T, t, T_{[0,1]}) - v_t^T(X_t^T \mid X_1^T)\|_2^2\right]\right]}_{=: R_t^{\mathrm{cond}}(\theta;X_1^T,T)}. \tag{3.4}$$

Conditioning on a clock gives a tractable target, but outer-clock integrability requires a separate argument. For the straight-line flow, $X_t^T = tX_1 + (1-t)X_0^T$, the sampled target is exactly $v_t^T(X_t^T \mid X_1) = X_1 - X_0^T$: the apparent terminal $(1-t)^{-1}$ singularity cancels. The usual square-integrable projection argument is stated in Appendix C.3. The following result establishes the outer-clock guarantee without finite endpoint variances.

Theorem 3.1 (Outer-clock integrability for straight-line HTFM)**.** *Let $X_1 \perp T$, and suppose $X_0^T \mid T \sim \mathcal{N}(0, \Sigma(T))$, with a measurable positive-semidefinite covariance matrix $\Sigma(T)$ and $X_0^T \perp X_1 \mid T$. Assume, for some $p \in (1,2)$,*

$$\mathbb{E}\|X_1\|_2^p < \infty, \qquad \mathbb{E}_T\sqrt{\mathrm{tr}\,\Sigma(T)} < \infty.$$

Let $\tau \sim \mathrm{Unif}(0,1)$ be independent of all other randomness, and set $X_\tau^T = \tau X_1 + (1-\tau)X_0^T$, $Y = X_1 - X_0^T$, and $Z = (X_\tau^T, \tau, T)$. Then, with $u^\star(Z) = \mathbb{E}[Y \mid Z]$,

$$C_\star := \mathbb{E}\|Y - u^\star(Z)\|_2^2 < \infty.$$

Viewing the objectives as functionals of a measurable predictor $u(Z)$, $\mathcal{L}_{\mathrm{CHTFM}}(u) < \infty$ if and only if $\mathcal{L}_{\mathrm{HTFM}}(u) < \infty$. On this common finite-risk domain,

$$\mathcal{L}_{\mathrm{CHTFM}}(u) = \mathcal{L}_{\mathrm{HTFM}}(u) + C_\star. \tag{3.5}$$

The proof in Appendix C.6 takes $a(T) = \sqrt{\operatorname{tr}\Sigma(T)}$; conditional Gaussian moments satisfy the required higher-moment bounds. For the stable area clock and inverse-gamma slope clock, the scale condition holds for $\alpha > 1$ and $\nu > 1$, respectively, covering the reported infinite-variance sources.

Corollary 3.2 (Local population-gradient equivalence)**.** *Under Theorem 3.1, let $P_Z := \mathrm{Law}(Z)$. Suppose the model class has a minimizer $\theta^\star$ with finite conditional risk and $B_r(\theta^\star) \subset \Theta$. Assume $\theta \mapsto u_\theta(z)$ is continuously differentiable on this ball for P_Z-almost every z, and*

$$\mathbb{E}_{Z\sim P_Z}\left[\sup_{\vartheta\in B_r(\theta^\star)} \|D_\theta u_\vartheta(Z)\|_{\mathrm{op}}^2\right] < \infty.$$

Both objectives are finite and differentiable throughout the ball, with $\nabla_\theta \mathcal{L}_{\mathrm{CHTFM}} = \nabla_\theta \mathcal{L}_{\mathrm{HTFM}}$.

The full local risk bound and proof are given in Corollary C.8. These assumptions do not guarantee finite risk for arbitrary parameters or convergence from arbitrary initialization.

3.2 Signature feature for path conditioning

The objective (3.4) fixes the inner conditional regression space through its dependence on the path-valued variable $T_{[0,1]}$. Because $T_{[0,1]}$ is infinite-dimensional, it cannot be passed directly to a neural field; we therefore summarize the realized clock path by a finite-dimensional path-signature feature, taking the truncated logsignature $\ell_m(T)$ defined below. Path signatures originate in rough path theory [37, 8] and have been widely used as finite-dimensional encodings of stream-valued data in machine learning, both as differentiable neural-network layers [22, 23] and as kernels and feature maps for sequential data [31, 26]; they are compact and expressive. The full signature separates paths up to tree-like equivalence, and increasing truncation orders yield universal approximation on suitable compact path families (Proposition D.1 in Appendix D.1).

Time-augmented signature. Let $\mathbf{T}_s := (s, T_s) \in \mathbb{R}^2$, $s \in [0,1]$, be the time-augmented clock path. Its truncated signature of order m is the iterated-integral tensor

$$\phi_m(T) := S^{(\le m)}(\mathbf{T}) = \left(1, \left(\int_{0<u_1<\cdots<u_k<1} \mathrm{d}\mathbf{T}_{u_1} \otimes \cdots \otimes \mathrm{d}\mathbf{T}_{u_k}\right)_{k=1}^{m}\right) \in \mathbb{R}^{D_m}. \tag{3.6}$$

Time augmentation captures how clock mass is allocated in physical time. For càdlàg clocks observed on a discretization $0 = t_0 < t_1 < \cdots < t_N = 1$, we linearly interpolate the sampled time-augmented points $\big((t_i, T_{t_i})\big)_{i=0}^N$ and take the truncated signature of the resulting piecewise linear path; this is the operational meaning of $\phi_m(T)$ used throughout the paper. For a more compact representation we additionally use the truncated logsignature $\ell_m(T) := \log_\otimes\big(\phi_m(T)\big)$, computed in the truncated tensor algebra. The dimension reduction is substantial: for $q = 2$ and $m = 3$, the raw signature has 15 coordinates whereas the nonconstant logsignature has only 5. Precise jump conventions, dimension formulas, and the lead–lag and piecewise-constant variants are deferred to Appendix D.

Network conditioning. Let $\psi(T)$ denote the implemented clock feature. For stable clocks, we standardize both channels of $(t_i, T_{t_i})_{i=0}^N$ along time within each path ($N = 200$), then compute its order-m logsignature; $\ell_m(T)$ above instead denotes the raw-path feature. For affine Student-t clocks, $\psi(T) = \log V_\nu$. A two-layer SiLU MLP (hidden/output width 128) embeds $\psi(T)$; the embedding is concatenated with time at every residual block. HTFM-G omits this module (Appendix E.4). Replacing the full clock argument in (3.4) by $\psi(T)$ gives the practical objective

$$\mathcal{L}^{\mathrm{sig}}_{\mathrm{CHTFM}}(\theta) := \mathbb{E}_{t\sim\mathrm{Unif}(0,1)}\mathbb{E}_{T,X_1^T}\left[\mathbb{E}_{X_t^T|X_1^T}\left[\left\|u_\theta(X_t^T, t, \psi(T)) - v_t^T(X_t^T \mid X_1^T)\right\|_2^2\right]\right]. \tag{3.7}$$

We use untruncated minibatch losses, clipping the aggregated gradient norm at 5.0; neither clocks nor per-sample losses are truncated.

Computational overhead. The logsignature is computed once per clock path at cost $\mathcal{O}(Nq^m)$, with $q = 2$, $m = 1$ by default, and $m = 2$ in the order ablation. The feature embedding is reused across solver steps; conditioning introduces no additional velocity-field evaluation.

3.3 Flow design and transport interpretation

We now specify the affine flow family used in this paper. Let $\Sigma : \mathbb{D}([0,1];\mathbb{R}_+) \to \mathbb{S}_+^d$ be a measurable map into the cone of symmetric positive semidefinite $d \times d$ matrices. Suppose that, conditioned on the realized clock path, $X_0^T \sim \mathcal{N}\big(0, \Sigma(T_{[0,1]})\big)$, and that X_0^T and X_1^T are conditionally independent given $T_{[0,1]}$. Let α_t^T and β_t^T be clock-dependent schedules such that, for almost every realized clock path, $t \mapsto \alpha_t^T$ and $t \mapsto \beta_t^T$ are absolutely continuous and satisfy $\alpha_0^T = 0,\ \beta_0^T = 1,\ \alpha_1^T = 1,\ \beta_1^T = 0,\ \beta_t^T > 0$ for $t \in [0,1)$. The clock-conditioned affine path is then

$$X_t^T = \alpha_t^T X_1^T + \beta_t^T X_0^T, \qquad t \in [0,1]. \tag{3.8}$$

Under (3.8), for each $x_1 \in \mathbb{R}^d$ and each realized clock path $T_{[0,1]}$, the conditional law of X_t^T is Gaussian (derivation in Appendix C.2):

$$p_t^T(\cdot \mid x_1) = \mathrm{Law}(X_t^T \mid X_1^T = x_1) = \mathcal{N}\big(\alpha_t^T x_1,\ (\beta_t^T)^2 \Sigma(T_{[0,1]})\big). \tag{3.9}$$

The random clock thus influences the interpolation through the conditional source covariance $\Sigma(T_{[0,1]})$ and through the schedule pair (α_t^T, β_t^T); in both cases the interpolation law is Gaussian conditional on the clock and the endpoint. For a.e. $t \in [0,1)$, differentiating (3.8) at fixed clock realization and substituting $X_0^T = (X_t^T - \alpha_t^T X_1^T)/\beta_t^T$ yields the explicit endpoint-conditioned target

$$v_t^T(x \mid x_1) := \mathbb{E}\left[\frac{\mathrm{d}X_t^T}{\mathrm{d}t} \,\middle|\, X_t^T = x,\ X_1^T = x_1\right] = \frac{\frac{\mathrm{d}\beta_t^T}{\mathrm{d}t}}{\beta_t^T}\, x + \left(\frac{\mathrm{d}\alpha_t^T}{\mathrm{d}t} - \frac{\frac{\mathrm{d}\beta_t^T}{\mathrm{d}t}}{\beta_t^T}\alpha_t^T\right) x_1. \tag{3.10}$$

This is the same algebraic target as in Gaussian flow matching, with the deterministic schedules replaced by their clock-dependent versions. Training samples are generated by drawing T and then X_t^T from (3.8); the model regresses $u_\theta(X_t^T, t, \psi(T))$ toward (3.10) in (3.7). For generation, we sample T, compute $\psi(T)$, draw $X_0^T \sim \mathcal{N}(0, \Sigma(T_{[0,1]}))$, and use a numerical ODE solver on $\mathrm{d}X_t^T = u_{\theta^*}(X_t^T, t, \psi(T))\,\mathrm{d}t$.

We now specialize to the straight-line flow used in our main experiments; a more general signature-linear schedule family is deferred to Appendix D.1.

Straight-line flow. Set $\alpha_t^T = t, \beta_t^T = 1-t$. Then (3.10) reduces to $v_t^T(x \mid x_1) = \frac{x_1 - x}{1-t}$, $t \in [0,1)$, and the endpoint-conditioned target has no explicit clock dependence. The clock still enters through the conditional law (3.9) via the source covariance, and hence through the projection defining the pathwise marginal optimum $u_t^T(x) = \mathbb{E}[v_t^T(X_t^T \mid X_1^T) \mid X_t^T = x]$. A network that does not observe the clock can therefore only learn a clock-averaged field, whereas the clock-conditioned formulation learns the family of optima indexed by the realized clock path. This straight-line schedule cleanly recovers the heavy-tailed marginal source families used in our experiments, and additionally admits a pathwise W_p-geodesic interpretation under a p-optimal endpoint coupling that remains valid in the infinite-second-moment regime $p < 2$. The independent training coupling need not be optimal. See Appendix C.4 for the additional theorem.

Heavy-tailed marginals from clock laws. With the straight-line schedule, choosing the clock law and source covariance recovers the named heavy-tailed marginal families used in our experiments. Throughout this paragraph, X_t denotes the clock-marginal variable obtained by first sampling T and then sampling X_t^T in the corresponding clock-conditioned space. Equivalently, $\mathrm{Law}(X_t \mid X_1 = x_1) = \mathbb{E}_T\left[\mathrm{Law}(X_t^T \mid X_1^T = x_1, T_{[0,1]})\right]$.[3] For the representative examples, we use the area-clock covariance $\Sigma_A(T_{[0,1]}) := 2A(T) I_d$, where $A(T) := \int_0^1 T_s\,\mathrm{d}s$. The following theorem states that the same clock-conditioned Gaussian construction can recover three commonly used source families: Gaussian, α-stable, and Student-t. The proof is deferred to Appendix C.5.

Theorem 3.3 (Recovered clock marginals)**.** *Assume the affine path* (3.8)*, the clock-independent straight-line schedule, and the covariance map $\Sigma_A(T_{[0,1]})$. For notational brevity, write $p_t(\cdot \mid x_1) := \mathrm{Law}(X_t \mid X_1 = x_1)$. For every $t \in [0,1)$ and $x_1 \in \mathbb{R}^d$, the following identities hold:*

[3] When $\mathbb{E}_T$ is applied to laws, it denotes the clock mixture: $(\mathbb{E}_T \mu^T)(B) := \int \mu^T(B)\,\mathbb{P}_T(\mathrm{d}T)$ for every Borel set $B \subseteq \mathbb{R}^d$. This does not require finite moments of the clock-marginal law.

(i) *If $T_t = t$, then $p_t(\cdot \mid x_1) = \mathcal{N}(\alpha_t x_1, \beta_t^2 I_d)$, with sub-Gaussian tail decay.*

(ii) *Let $\alpha \in (0,2)$, $\rho = \alpha/2$, and let T be a ρ-stable subordinator normalized by $\mathbb{E}[e^{-sT_t}] = \exp(-(\rho+1)ts^\rho/2)$. Then $p_t(\cdot \mid x_1)$ is a location-shifted isotropic α-stable law with location $\alpha_t x_1$, scale β_t. In particular, its radial tail satisfies $\mathbb{P}(\|X_t - \alpha_t x_1\|_2 > r \mid X_1 = x_1) = \Theta(r^{-\alpha}),\ r \to \infty$.*

(iii) *Let $\nu > 0$. If $V_\nu \sim \mathrm{InvGamma}(\nu/2, \nu/2)$ and $T_t = tV_\nu$, then $p_t(\cdot \mid x_1)$ is a multivariate Student-t law with location $\alpha_t x_1$, scale $\beta_t^2 I_d$, ν degrees of freedom. In particular, its radial tail satisfies $\mathbb{P}(\|X_t - \alpha_t x_1\|_2 > r \mid X_1 = x_1) = \Theta(r^{-\nu}),\ r \to \infty$.*

Beyond clock-independent straight-line schedules. The schedules α_t^T, β_t^T may themselves depend on the realized clock path through its truncated signature; concrete signature-linear and VE-type parameterizations and a near-linear boundary-corrected variant are collected in Appendices D.1, D.2, and A.5. The endpoint constraints leave the source law at $t = 0$ unchanged. At intermediate times, however, clock-dependent schedules generally yield endpoint-conditioned Gaussian mixtures outside the exact α-stable and Student-t families in Theorem 3.3. A detailed comparison of the random-clock formulation with DLPM, LIM, and Student-t-based heavy-tailed diffusion families, together with a discussion of data-adaptive clock laws, is given in Appendix A.2.

4 Experiments

We evaluate HTFM systematically over different tail distributions with varying tail-heaviness, on a 2D synthetic dataset, an imbalanced image dataset, and a weather forecast benchmark. Appendices E and F provide full experimental settings and additional ablations. [4]

4.1 Experimental Setup

Datasets. We use a nine-component, imbalanced 2D α-stable mixture on a 3×3 grid with data stability index $\alpha_D \in \{1.5, 1.8\}$ [47]; CIFAR10-LT, a 32×32 image dataset with imbalance ratio $\rho = 0.01$ [6, 29]; and hourly HRRR VIL weather fields cropped to 128×128 over the Central US [5, 13, 41].

Metrics. We use precision–recall f_1^{pr} on the 2D mixture [47] and FID against the full CIFAR10-LT training split [18]. For Few classes, *Worst/Macro* coverage is the minimum/mean generated-to-real class-frequency ratio capped at one; *TailMass* is the total generated-to-real Few-class mass ratio (target one). Recall and Kernel Inception Distance (KID) are macro-averaged within Many/Medium/Few (M/Md/F) groups. On HRRR, kurtosis ratio (KR), skewness ratio (SR), and tail KS measure pixel-intensity discrepancies [42, 41]. *Max KS* compares per-field maxima; *Eccentricity KS* compares the eccentricities of the largest 8-connected components exceeding the real 99.9th-percentile intensity. *Event rate* measures the fraction of fields with an exceedance, targeting the real frequency. Definitions and protocols are in Appendices E.3 and F.5.

Methods and replication. We compare HTFM-G, HTFM-α, and HTFM-t from Theorem 3.3. Original benchmark sweeps average three training seeds; five-seed FID results appear in Table 6 (Appendix E.3).

4.2 Source-Tail Choice across Domains

We train separately for each clock law and tail parameter with fixed base backbones and training protocols. Sampling budgets are NFE $= 10$ for the 2D mixture, 100 for CIFAR10-LT, and 50 for HRRR VIL.

Heavy-tailed clocks improve over HTFM-G throughout (Table 1). At $p = 1.7$, HTFM-t performs best on the 2D mixture and HTFM-α on CIFAR10-LT and HRRR: clock-family structure matters alongside the tail exponent.

[4] Code: `https://github.com/BIGKnight/Heavy-tailed-Flow-Matching`.

Table 1: HTFM main results across benchmarks. p is the polynomial tail-decay exponent (p_α for HTFM-α and p_ν for HTFM-t). Bold indicates the best result in each column.

Method		2D synthetic ($f_1^{\mathrm{pr}}\uparrow$)		CIFAR10-LT	HRRR VIL		
Variant	p	α_D=1.5	α_D=1.8	FID ↓	KR ↓	SR ↓	KS ↓
HTFM-G	∞	0.888	0.893	19.29	0.72	0.68	0.42
HTFM-α	1.6	0.943	0.934	16.19	**0.04**	0.11	0.11
	1.7	0.941	0.923	16.10	0.47	**0.06**	0.04
	1.8	0.931	0.921	**15.89**	0.06	0.31	**0.03**
HTFM-t	1.7	**0.962**	**0.959**	16.31	0.56	0.14	0.14
	2.0	0.954	0.950	16.76	0.43	0.52	0.21
	3.0	0.943	0.939	16.14	0.16	0.41	0.16

4.3 Comparison with Baselines

Table 2 compares HTFM with LIM [50, 44], DLPM/DLIM [47], and t-Flow [41] under matched base backbones and training budgets, sweeping stable, Student-t, and Gaussian sources. Table 6 (Appendix E.3) reports five-seed means and standard deviations for a subset of these settings.

Table 2: Main CIFAR10-LT results. We report FID at each NFE while sweeping the polynomial tail-decay parameter p. Source-family columns indicate evaluated configurations.

Method	Evaluated Source Families			NFE	Polynomial Tail Decay Exponent							
	α-stable	Student-t	Gaussian		$p_\alpha=1.6$	$p_\alpha=1.7$	$p_\alpha=1.8$	$p_\alpha=1.9$	$p_\nu=1.7$	$p_\nu=2.0$	$p_\nu=3.0$	$p=\infty$
Baselines								*Heavy Tailed*				*Gaussian*
LIM	✓	✗	✗	20	135.9	130.2	126.8	132.5	–	–	–	–
				100	34.70	34.11	32.40	34.10	–	–	–	–
				500	20.22	19.18	18.23	20.60	–	–	–	–
DLPM	✓	✗	✓	20	29.16	28.55	32.49	30.96	–	–	–	35.49
				100	18.18	18.23	18.29	18.74	–	–	–	22.28
				500	15.26	14.91	15.35	15.43	–	–	–	19.47
DLIM	✓	✗	✓	20	21.30	23.90	22.75	24.50	–	–	–	25.45
				100	19.69	20.38	18.46	19.60	–	–	–	21.63
				500	17.22	17.73	17.90	18.14	–	–	–	19.92
t-Flow	✗	✓	✗	20	–	–	–	–	–	–	22.49	–
				100	–	–	–	–	–	–	19.05	–
				500	–	–	–	–	–	–	17.66	–
Our Framework								*Heavy Tailed*				*Gaussian*
HTFM	✓	✓	✓	20	17.52	17.13	16.91	16.76	19.91	20.48	19.35	22.67
				100	16.19	16.10	15.89	16.21	16.31	16.76	16.14	19.29
				500	15.27	15.04	14.99	15.67	15.97	16.51	15.38	19.04

HTFM covers Gaussian, stable, and Student-t sources with a common objective. At low NFE, it improves substantially over LIM and DLPM; at larger NFE, its FID becomes comparable to DLPM. HTFM achieves lower FID than the ODE-based DLIM and t-Flow samplers across the reported budgets.

4.4 Minority Classes and Spatial Extremes

We complement global sample-quality metrics with minority-class coverage on CIFAR10-LT and event-level statistics on HRRR VIL.

CIFAR10-LT. Table 3 evaluates 50,000 generated images per model at NFE $=100$. HTFM improves minority-class coverage and Few-class Recall over DDPM and DLPM while maintaining comparable class-conditional KID. The gains extend across the Few classes, but their aggregate mass is moderately overestimated: coverage improves without fully resolving class-frequency calibration. Appendix F.5 provides the full class histograms and evaluation protocol.

Table 3: Minority-class evaluation on CIFAR10-LT at NFE $= 100$. Bold marks metric-wise optima, with TailMass ranked by distance to one. Class-conditional metrics are point estimates.

Method	Tail exp.	Worst / Macro ↑	TailMass	Recall M/Md/F ↑	KID M/Md/F ↓
HTFM-α	$p_\alpha = 1.6$	**1.000 / 1.000**	1.204	0.726 / 0.712 / 0.788	**0.012 / 0.035 / 0.050**
HTFM-α	$p_\alpha = 1.7$	**1.000 / 1.000**	1.168	0.727 / 0.715 / 0.787	0.014 / 0.039 / 0.054
HTFM-α	$p_\alpha = 1.8$	0.971 / 0.993	1.256	**0.730** / 0.717 / 0.787	0.013 / 0.036 / 0.053
HTFM-t	$p_\nu = 2.0$	0.975 / 0.994	1.258	0.727 / **0.720** / 0.790	0.013 / 0.037 / 0.054
HTFM-t	$p_\nu = 3.0$	0.948 / 0.987	1.211	0.727 / 0.719 / **0.794**	0.013 / 0.037 / 0.053
DLPM	$p_\alpha = 1.6$	0.651 / 0.868	0.817	0.708 / 0.695 / 0.759	0.015 / 0.038 / 0.052
DLPM	$p_\alpha = 1.7$	0.702 / 0.885	**0.841**	0.696 / 0.712 / 0.772	0.013 / 0.039 / 0.055
DLPM	$p_\alpha = 1.8$	0.642 / 0.840	0.783	0.703 / 0.715 / 0.761	0.013 / 0.041 / 0.058
LIM	$p_\alpha = 1.7$	0.702 / 0.925	1.680	0.713 / 0.679 / 0.762	0.023 / 0.043 / 0.057
DDPM	$p = \infty$	0.549 / 0.721	0.684	0.691 / 0.717 / 0.758	0.013 / 0.041 / 0.062

HRRR VIL. Table 4 evaluates 20,000 generated fields per model at NFE $= 50$. HTFM best matches the marginal intensity statistics and extreme-component eccentricity. Its event frequency is closest to the real reference, despite under-generating exceedances. Peak-intensity recovery is more mixed: HTFM improves on DDPM and DLPM, but LIM better matches the distribution of per-field maxima.

Table 4: HRRR-VIL event-level and spatial evaluation. Bold marks the best generated result per metric; event rate is ranked by distance to the real 20.37%.

Method	Tail exp.	KR ↓	SR ↓	Tail KS ↓	Max KS ↓	Ecc. KS ↓	Event rate
Real test data	–	–	–	–	–	–	20.37%
HTFM-α	$p_\alpha = 1.7$	**0.470**	**0.064**	**0.043**	0.664	**0.073**	**16.92%**
DDPM	$p = \infty$	0.930	0.825	0.277	0.863	0.126	11.49%
LIM	$p_\alpha = 1.7$	0.823	0.991	0.438	**0.584**	0.143	75.50%
DLPM	$p_\alpha = 1.7$	0.674	0.990	0.337	0.845	0.101	14.21%

4.5 Ablation: Source and Clock Conditioning

Table 5 separates source choice from clock input. At $m = 0$, we still sample a clock and then a conditional Gaussian source, but withhold clock features from the network. This is different from LIM, which instead uses a marginal Lévy-score target and a different sampler.

Both the α-stable source and clock conditioning improve FID. The full order sweep (Appendix F.4) favors $m = 1$ over $m = 0$, with no consistent further gain at $m = 2$. The direct scalar comparison in that appendix favors T_1 over $A(T)$ and their pair. These results support low-order conditioning for the tested straight-line flows; gains from higher-order features under other schedules remain to be established.

Table 5: Within-FM FID on CIFAR10-LT (NFE$= 100$; stable $p_\alpha = 1.7$). Here m is the normalized-path logsignature order; $A(T)$ is a direct scalar input.

Source	Clock input	FID ↓
Gaussian	None	19.29
α-stable	$m = 0$	16.65
	$m = 1$	**16.10**
	$A(T)$	17.86
	$m = 2$	16.46

5 Conclusion

We introduced Heavy-Tailed Flow Matching via Random Clocks (HTFM), which represents different families of heavy-tailed sources as mixtures of clock-conditioned Gaussian sources within a single flow-matching objective. Using truncated logsignature features, HTFM conditions the velocity field on the realized clock without additional velocity-field evaluations. Across 2D synthetic data, CIFAR10-LT, and HRRR weather fields, HTFM improves over the Gaussian-clock variant in rare-event fidelity, with competitive low-NFE results against heavy-tailed baselines.

Acknowledgments and Disclosure of Funding

Haoyang Cao is partially supported by the Data Science and AI Institute (DSAI) startup funding at Johns Hopkins University. Kenji Kawaguchi is supported by the Air Force Office of Scientific Research under Award No. FA2386-24-1-4011 and also partially supported by the Singapore Ministry of Education Academic Research Fund Tier 1 (Award No. T1 251RES2509).

The authors declare no competing interests.

References

[1] Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

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

[3] O. Barndorff-Nielsen, J. Kent, and M. Sørensen. Normal variance-mean mixtures and z distributions. *International Statistical Review*, 50(2):145–159, 1982.

[4] Adrian Baule. Generative modelling with jump-diffusions. *arXiv preprint arXiv:2503.06558*, 2025.

[5] Stanley G. Benjamin, Stephen S. Weygandt, John M. Brown, Ming Hu, Curtis R. Alexander, Tatiana G. Smirnova, Joseph B. Olson, Eric P. James, David C. Dowell, Georg A. Grell, Haidao Lin, Steven E. Peckham, Tracy Lorraine Smith, William R. Moninger, Jaymes S. Kenyon, and Geoffrey S. Manikin. A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Monthly Weather Review*, 144(4):1669–1694, 2016.

[6] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, 2019.

[7] Hansheng Chen, Kai Zhang, Hao Tan, Zexiang Xu, Fujun Luan, Leonidas Guibas, Gordon Wetzstein, and Sai Bi. Gaussian mixture flow matching models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 9783–9802. PMLR, 2025.

[8] Ilya Chevyrev and Andrey Kormilitzin. A primer on the signature method in machine learning. In *Signature Methods in Finance: An Introduction with Computational Applications*, pages 3–64. Springer, 2025.

[9] Stuart Coles. *An introduction to statistical modeling of extreme values*. Springer Series in Statistics. Springer, 2001.

[10] Rama Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236, 2001.

[11] Rama Cont, Mihai Cucuringu, Renyuan Xu, and Chao Zhang. Tail-GAN: Learning to simulate tail risk scenarios. *Management Science*, 72(4):2917–2936, 2026.

[12] Christa Cuchiero, Francesca Primavera, and Sara Svaluto-Ferro. Universal approximation theorems for continuous functions of càdlàg paths and Lévy-type signature models. *Finance and Stochastics*, 29(2):289–342, 2025.

[13] David C Dowell, Curtis R Alexander, Eric P James, Stephen S Weygandt, Stanley G Benjamin, Geoffrey S Manikin, Benjamin T Blake, John M Brown, Joseph B Olson, Ming Hu, et al. The High-Resolution Rapid Refresh (HRRR): An hourly updating convection-allowing forecast model. Part I: Motivation and system description. *Weather and Forecasting*, 37(8):1371–1395, 2022.

[14] Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling extremal events: for insurance and finance*, volume 33 of *Stochastic Modelling and Applied Probability*. Springer Science & Business Media, 2013.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

[16] Gaby J Gründemann, Nick van de Giesen, Lukas Brunner, and Ruud van der Ent. Rarest rainfall events will see the greatest relative increase in magnitude under future climate change. *Communications Earth & Environment*, 3(1):235, 2022.

[17] Yunrui Guan, Krishna Balasubramanian, and Shiqian Ma. Mirror flow matching with heavy-tailed priors for generative modeling on convex domains. In *International Conference on Learning Representations*, 2026.

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

[19] Tennessee Hickling and Dennis Prangle. Flexible tails for normalizing flows. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 23155–23178. PMLR, 2025.

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[21] Priyank Jaini, Ivan Kobyzev, Yaoliang Yu, and Marcus Brubaker. Tails of Lipschitz triangular flows. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4673–4681. PMLR, 2020.

[22] Patrick Kidger, Patric Bonnier, Imanol Perez Arribas, Cristopher Salvi, and Terry Lyons. Deep signature transforms. *Advances in Neural Information Processing Systems*, 32, 2019.

[23] Patrick Kidger and Terry J. Lyons. Signatory: differentiable computations of the signature and logsignature transforms, on both CPU and GPU. In *International Conference on Learning Representations*, 2021.

[24] Junwan Kim, Jiho Park, Seonghu Jeon, and Seungryong Kim. Better source, better flow: Learning condition-dependent source distribution for flow matching. *arXiv preprint arXiv:2602.05951*, 2026.

[25] Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

[26] Franz J. Kiraly and Harald Oberhauser. Kernels for sequentially ordered data. *Journal of Machine Learning Research*, 20(31):1–45, 2019.

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

[28] Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

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

[30] Junho Lee, Kwanseok Kim, and Joonseok Lee. Is there a better source distribution than Gaussian? Exploring source distributions for image flow matching. *arXiv preprint arXiv:2512.18184*, 2025.

[31] Daniel Levin, Terry Lyons, and Hao Ni. Learning from the past, predicting the statistics for the future, learning an evolving system. *arXiv preprint arXiv:1309.0260*, 2013.

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

[33] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[34] Qiang Liu. Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*, 2022.

[35] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[36] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2537–2546, 2019.

[37] Terry Lyons. Rough paths, signatures and the modelling of functions on streams. *arXiv preprint arXiv:1405.4537*, 2014.

[38] Frank J. Massey, Jr. The Kolmogorov–Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

[39] James Morrill, Cristopher Salvi, Patrick Kidger, and James Foster. Neural rough differential equations for long time series. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7829–7838. PMLR, 2021.

[40] Fouad Oubari and Mathilde Mougeot. Correcting source mismatch in flow matching with radial-angular transport. *arXiv preprint arXiv:2604.04291*, 2026.

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

[42] Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

[43] Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky T. Q. Chen. Multisample flow matching: Straightening flows with minibatch couplings. *arXiv preprint arXiv:2304.14772*, 2023.

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

[45] Sidney I Resnick. *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer, 2007.

[46] Sonia I Seneviratne, Xuebin Zhang, Muhammad Adnan, Wafae Badi, Claudine Dereczynski, Alejandro Di Luca, Subimal Ghosh, Iskhaq Iskandar, James Kossin, Sophie Lewis, et al. Weather and climate extreme events in a changing climate. In *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 1513–1766. Cambridge University Press, 2021.

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

[48] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[49] 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. *arXiv preprint arXiv:2302.00482*, 2023.

[50] Eun Bi Yoon, Keehun Park, Sungwoong Kim, and Sungbin Lim. Score-based generative models with Lévy processes. *Advances in Neural Information Processing Systems*, 36:40694–40707, 2023.

[51] Tianjiao Zhang, Huangjie Zheng, Jiangchao Yao, Xiangfeng Wang, Mingyuan Zhou, Ya Zhang, and Yanfeng Wang. Long-tailed diffusion models with oriented calibration. In *The Twelfth International Conference on Learning Representations*, 2024.

Appendix Contents

A Additional Discussions . . . 16

A.1 Related work . . . 16

A.2 Detailed comparison with existing heavy-tailed models . . . 17

A.3 Representative Gaussian scale mixtures . . . 18

A.4 Gaussian mean-variance mixtures and asymmetric extensions . . . 18

A.5 Other admissible clock-conditioned paths . . . 19

B All-clock extension to diffusion models . . . 21

C Additional results and proofs of main theorems . . . 24

C.1 A note on conditioning on the clock history . . . 24

C.2 Conditional Gaussian law of the affine path . . . 24

C.3 Detailed pathwise equivalence result and proof . . . 25

C.4 Clock-conditioned straight-line geodesic . . . 25

C.5 Proof of Theorem 3.3 . . . 26

C.6 Outer-clock integrability and local finite-risk gradients . . . 27

D Additional Information on Log-Signature . . . 31

D.1 Signature-linear schedules . . . 33

D.2 VE-type signature-linear schedule . . . 34

E Experimental Setup . . . 35

E.1 Dataset . . . 36

E.2 Baselines . . . 36

E.3 Evaluation . . . 36

E.4 Network Architecture . . . 38

E.5 Training . . . 38

E.6 Sampling . . . 39

E.7 Compute Resources . . . 39

F Ablation Studies . . . 39

F.1 Effect of Flow Type . . . 39

F.2 Effect of Number of Function Evaluations (NFE) . . . 40

F.3 Effect of ODE Solver . . . 40

F.4 Signature order and direct scalar-summary conditioning . . . 41

F.5 Additional rare-event evaluation . . . 41

G Visualizations . . . 43

G.1 2D toy distributions: clock-degree sweep . . . 43

G.2 2D $S\alpha S$ mixture: baseline comparison at NFE=20 . . . 43

G.3 CIFAR10-LT: generated samples . . . 43

A Additional Discussions

A.1 Related work

Heavy-tailed diffusion and flow models. Most diffusion and flow-matching generative models use Gaussian reference distributions or Gaussian perturbation kernels. Several recent works replace this Gaussian mechanism by heavy-tailed alternatives. Lévy score models use α-stable Lévy processes and replace the ordinary score by a fractional score, leading to nonlocal reverse-time dynamics and deterministic fractional probability-flow samplers [50, 44]. DLPM and DLIM instead build a discrete-time α-stable analogue of DDPM/DDIM; their key practical step is a Gaussian scale-mixture augmentation that makes the conditional reverse kernels Gaussian while the marginal noising law remains heavy-tailed [47]. Student-t-based models include family-specific diffusion constructions and t-Flow, which uses ordinary squared noise prediction along a Student-t straight-line interpolation [41]. Cauchy Diffusion studies the Cauchy case in speech synthesis, using Cauchy noise and a ratio-of-Gaussians view to support stochastic and deterministic sampling in a DDPM-style vocoder [32]. Related jump-diffusion generative models consider finite-activity jumps and marginal-matching reverse dynamics [4]. Our random-clock formulation follows the same broad motivation of using heavy-tailed perturbation families, but differs in where the tractability is imposed: the source is Gaussian conditional on the realized clock, and the affine interpolation is Gaussian conditional on both the clock and the data endpoint. This gives an explicit conditional velocity target, while marginalizing the clock produces the heavy-tailed source or endpoint-conditioned path family.

Latent Gaussian mixtures and random time changes. Gaussian scale mixtures provide a classical route from conditionally Gaussian models to heavy-tailed marginals [2]. More general normal variance-mean mixtures can also be interpreted as Brownian motions with drift stopped at independent random times, and include many skewed and heavy-tailed families [3]. This latent-Gaussian perspective is exactly the mechanism exploited by DLPM for α-stable noising [47]. Our construction extends this idea from a scalar mixing variable to a path-valued clock. The clock can determine the conditional source covariance, the interpolation schedule, or both. Thus, the model keeps an explicit affine target derived from the clock-and-endpoint-conditioned Gaussian law, while the outer clock law encodes family choice, tail heaviness, and path-dependent allocation of noise.

Source and path design. Learning a condition-dependent source jointly with the velocity field [24] uses semantic side information to adapt initialization to a task. HTFM instead samples a latent clock independently of the endpoint and prescribes its law to control the source tails; a learned, condition-dependent clock could combine the two approaches. SourceFM [30] studies norm mismatch, directional coverage, and path entanglement in image flow matching, combining norm-aligned training with directionally pruned sampling while retaining Gaussian training support. HTFM instead changes the asymptotic radial tail law and conditions the field on the realized source scale. RAFM [40] matches the source radial law to the data and uses radius-preserving spherical transport. It is appropriate when the target radial distribution can be estimated reliably; HTFM uses a parametric clock law and an explicit affine target from the clock-and-endpoint-conditioned Gaussian law.

Mirror flow matching. Mirror Flow Matching [17] addresses generation on convex domains using a regularized mirror map and a Student-t prior in the dual space. At the source-law level, its unconstrained Student-t flow is closest to a clock-blind HTFM-t model: both use a marginal Student-t source without exposing its inverse-gamma scale to the field. HTFM additionally conditions on that latent scale and includes Gaussian, stable, and Student-t sources within one objective. Conversely, HTFM does not supply the convex-domain feasibility guarantees of mirror geometry.

Gaussian mixture flow matching. GMFlow [7] predicts dynamic Gaussian-mixture parameters for the conditional velocity distribution, trains with a KL-divergence objective, and uses specialized GM-ODE/SDE solvers. HTFM changes a different component: its source is a continuous Gaussian scale mixture, while the network predicts a single velocity and sampling uses an ordinary ODE. A fixed finite Gaussian-mixture source with finite component covariances has Gaussian asymptotic tails, even though it can approximate a broad range of densities on bounded regions. The source-

scale and predictive mixture mechanisms are complementary; a direct GMFlow comparison would also change the network output, objective, and sampler.

Signature and logsignature transforms. Path signatures originate in rough path theory and give an order-sensitive sequence of iterated-integral features for streams [37, 8]. In statistics and machine learning, signature features have been used as universal, truncated representations for stream-valued inputs [31], as kernels for sequential data [26], and as differentiable neural-network components [22, 23]. Logsignature features are especially relevant when one wants a compact nonredundant representation of path increments; for example, neural rough differential equations summarize long driving signals over subintervals by logsignatures [39]. Recent approximation results also justify signature-linear representations of path functionals on suitable compact path sets, including càdlàg settings [12]. Our use of signatures is narrower: the clock path is not the generated sample path, but a latent conditioning environment. We use low-order logsignature features to give the neural velocity field finite-dimensional access to the realized random clock, and, optionally, to parameterize clock-dependent affine schedules.

A.2 Detailed comparison with existing heavy-tailed models

The related-work overview in Appendix A.1 surveys heavy-tailed diffusion and flow constructions. Here we elaborate on the technical contrast between the random-clock formulation and the closest prior families.

DLPM and Lévy-based diffusion. DLPM [47] obtains heavy-tailed diffusion transitions via α-stable innovations and a robust outer power applied to empirical losses. As shown in Theorem 3.3(ii), our stable-subordinator clock with the area-clock covariance reproduces the same α-stable marginal source family. The two constructions differ at two levels. *At the flow-matching target level*, DLPM averages over noise realizations and learns a clock-marginal vector field, whereas conditioning on the realized clock path turns the regression into a clock-indexed family of pathwise optima u_t^T; a network with access to the clock can learn the appropriate optimum for each noise realization, and the random-clock view can therefore be read as a refinement of the DLPM stable-noise construction with the same marginal family but a better-specified flow-matching target. We expect empirical gains when different clock realizations induce meaningfully different optima, and the formulation also avoids direct use of the generally unavailable α-stable density. *At the loss level*, the outer robust power $z \mapsto z^r$ with $0 < r < 1$ used by DLPM destroys the additive projection decomposition in Proposition C.2; in contrast, the random-clock formulation retains the quadratic objective and applies the L^2-projection argument to a square-integrable residual under the finite-risk assumptions in Appendix C.6. It treats any robustification over sampled clocks as an estimator-level stabilization rather than a change of the population target. Compared with Lévy score and Lévy–Itô diffusion models [50, 44], the flow-matching construction further avoids a nonlocal fractional-score target: after conditioning on the clock, the endpoint-conditioned target remains the explicit affine expression (3.10).

Student-t diffusion and t-Flow. Student-t diffusion and t-Flow models replace Gaussian perturbations by Student-t heavy-tailed perturbations [41]. Their variance-normalized t-EDM parameterization requires $\nu > 2$, whereas t-Flow uses ordinary squared noise prediction along a Student-t straight-line interpolation. The Student-t source and interpolation themselves are defined for every $\nu > 0$; existence of the population conditional mean and finiteness of the training risk require separate moment assumptions. HTFM exposes the latent clock or scale to the velocity field and provides one construction for Gaussian, stable, and Student-t sources. In particular, the inverse-gamma clock in Theorem 3.3(iii) is defined for every $\nu > 0$, while the outer-clock finite-risk result in Appendix C.6 applies to Student-t clocks with $\nu > 1$, including the infinite-variance range $1 < \nu \leq 2$, under its stated data-moment assumptions.

Toward data-adaptive clocks. The most general use of random clocks would not prescribe the clock law in advance. Instead, one could estimate a clock distribution from data, or infer a data-dependent posterior over clock paths with an amortized model, and then use the inferred clock as the conditioning variable in (3.7). This would turn the clock into a learned latent description of local tail behavior rather than a hand-chosen family such as a stable subordinator or inverse-

gamma slope. Developing such data-adaptive clock inference requires additional identifiability and estimation assumptions, so we leave it as a future extension.

A.3 Representative Gaussian scale mixtures

We recall two standard Gaussian scale-mixture examples that are only mentioned briefly in Section 2.2. For $\alpha \in (0, 2]$, a centered symmetric isotropic α-stable random vector $S_\alpha \in \mathbb{R}^d$ is characterized by

$$\mathbb{E}\left[e^{i\langle \xi, S_\alpha\rangle}\right] = \exp\bigl(-\|\xi\|_2^\alpha/2\bigr), \qquad \xi \in \mathbb{R}^d. \tag{A.1}$$

The case $\alpha = 2$ reduces to the Gaussian law, whereas for $\alpha < 2$ the law is heavy-tailed, with radial tail probability

$$\mathbb{P}\left(\|S_\alpha\|_2 > r\right) \asymp r^{-\alpha}, \qquad r \to \infty.$$

With this normalization, S_α admits the Gaussian scale-mixture representation

$$S_\alpha \overset{d}{=} \sqrt{V_\alpha}\, G,$$

where $G \sim \mathcal{N}(0, I_d)$ and V_α is a positive $\alpha/2$-stable mixing variable independent of G, with Laplace transform

$$\mathbb{E}[e^{-sV_\alpha}] = \exp\left(-2^{\alpha/2-1} s^{\alpha/2}\right), \qquad s \geq 0.$$

Likewise, a centered d-dimensional Student-t random vector Z_ν with $\nu > 0$ degrees of freedom and identity scale has density

$$f_{\nu,d}(x) = \frac{\Gamma\bigl((\nu+d)/2\bigr)}{\Gamma(\nu/2)\,(\nu\pi)^{d/2}} \left(1 + \frac{\|x\|_2^2}{\nu}\right)^{-(\nu+d)/2}, \qquad x \in \mathbb{R}^d, \tag{A.2}$$

and satisfies

$$\mathbb{P}\left(\|Z_\nu\|_2 > r\right) \asymp r^{-\nu}, \qquad r \to \infty.$$

It can be written as

$$Z_\nu \overset{d}{=} \sqrt{V_\nu}\, G, \qquad V_\nu \sim \mathrm{InvGamma}(\nu/2, \nu/2),$$

with V_ν independent of G. These examples make explicit why conditioning on the latent scale leaves a Gaussian problem, while marginalizing the latent scale produces polynomial tails.

A.4 Gaussian mean-variance mixtures and asymmetric extensions

The main text focuses on centered Gaussian scale mixtures, which already cover a broad class of symmetric heavy-tailed laws. A natural extension is to allow the latent variable to affect both the conditional mean and the conditional covariance, leading to the class of *Gaussian mean-variance mixtures*.

Let V be a latent random variable taking values in a measurable space V, and let $G \sim \mathcal{N}(0, I_d)$ be independent of V. A random vector $X \in \mathbb{R}^d$ is called a Gaussian mean-variance mixture if it admits a representation of the form

$$X = m(V) + L(V)G, \tag{A.3}$$

where $m : \mathsf{V} \to \mathbb{R}^d$ and $L : \mathsf{V} \to \mathbb{R}^{d\times d}$ are measurable. Equivalently,

$$X \mid V \sim \mathcal{N}\bigl(m(V),\, \Sigma(V)\bigr), \qquad \Sigma(V) := L(V)L(V)^\top. \tag{A.4}$$

The heavy-tailed or asymmetric nature of the marginal law is produced by marginalizing over the latent variable V, while conditional on V the distribution remains Gaussian.

The centered Gaussian scale-mixture subclass from the main text is recovered by setting $m(V) \equiv 0$. When $m(V) \neq 0$, the latent variable affects not only the covariance but also the location, and the resulting marginal law may be asymmetric. Thus Gaussian mean-variance mixtures provide a natural route to skewed heavy-tailed extensions.

In the random-clock setting, the corresponding extension would be to replace the centered conditional source law

$$X_0^T \sim \mathcal{N}\bigl(0, \Sigma(T_{[0,1]})\bigr)$$

by the more general form

$$X_0^T \sim \mathcal{N}\big(m(T_{[0,1]}),\, \Sigma(T_{[0,1]})\big), \tag{A.5}$$

where $m : \mathbb{D}([0,1];\mathbb{R}_+) \to \mathbb{R}^d$ is a measurable mean functional. Under the affine interpolation

$$X_t^T = \alpha_t X_1^T + \beta_t X_0^T,$$

we then have

$$X_t^T \mid X_1^T = x_1 \sim \mathcal{N}\big(\alpha_t x_1 + \beta_t m(T_{[0,1]}),\, \beta_t^2 \Sigma(T_{[0,1]})\big). \tag{A.6}$$

Hence the source remains Gaussian conditional on the clock, and the affine interpolation remains Gaussian conditional on both the clock and the data endpoint, preserving the latent-Gaussian construction.

Moreover, the explicit affine target formula is unaffected by this extension. Indeed, for a.e. $t \in [0,1)$, the identity

$$X_0^T = \frac{X_t^T - \alpha_t X_1^T}{\beta_t}$$

still holds pathwise, so substituting into

$$\frac{\mathrm{d}X_t^T}{\mathrm{d}t} = \frac{\mathrm{d}\alpha_t}{\mathrm{d}t} X_1^T + \frac{\mathrm{d}\beta_t}{\mathrm{d}t} X_0^T$$

yields

$$\mathbb{E}\left[\frac{\mathrm{d}X_t^T}{\mathrm{d}t} \,\middle|\, X_t^T = x,\ X_1^T = x_1\right] = \frac{\frac{\mathrm{d}\beta_t}{\mathrm{d}t}}{\beta_t}\, x + \left(\frac{\mathrm{d}\alpha_t}{\mathrm{d}t} - \frac{\frac{\mathrm{d}\beta_t}{\mathrm{d}t}}{\beta_t}\alpha_t\right) x_1. \tag{A.7}$$

Thus, even in the more general mean-variance mixture setting, the affine clock-conditioned target retains the same closed-form expression as in the centered case. What changes is the induced marginal family after integrating out the clock, which may now be skewed as well as heavy-tailed.

This extension suggests that the random-clock framework is not limited to symmetric Gaussian scale mixtures. By allowing a nonzero conditional mean functional $m(T_{[0,1]})$, one can in principle accommodate a broader class of asymmetric heavy-tailed laws while preserving conditional Gaussianity. We do not pursue this direction further in the main text.

A.5 Other admissible clock-conditioned paths

The main text focuses on the straight-line interpolation, which is the most common choice in conditional flow matching and already suffices for the heavy-tailed framework developed here. More generally, however, the same clock-conditioned construction applies to a broader class of affine paths.

Let $\alpha, \beta : [0,1] \to \mathbb{R}$ be absolutely continuous functions such that

$$\alpha_0 = 0, \qquad \beta_0 = 1, \qquad \alpha_1 = 1, \qquad \beta_1 = 0,$$

and assume $\beta_t > 0$ for all $t \in [0,1)$. Define

$$X_t^T = \alpha_t X_1^T + \beta_t X_0^T, \qquad t \in [0,1]. \tag{A.8}$$

As in Section 3.3, if $X_0^T \sim \mathcal{N}(0, \Sigma(T_{[0,1]}))$ and X_0^T and X_1^T are conditionally independent given $T_{[0,1]}$, then

$$p_t^T(\cdot \mid x_1) = \mathrm{Law}(X_t^T \mid X_1^T = x_1) = \mathcal{N}\big(\alpha_t x_1,\, \beta_t^2 \Sigma(T_{[0,1]})\big). \tag{A.9}$$

Moreover, for a.e. $t \in [0,1)$,

$$\frac{\mathrm{d}X_t^T}{\mathrm{d}t} = \frac{\mathrm{d}\alpha_t}{\mathrm{d}t} X_1^T + \frac{\mathrm{d}\beta_t}{\mathrm{d}t} X_0^T, \qquad X_0^T = \frac{X_t^T - \alpha_t X_1^T}{\beta_t},$$

so the corresponding conditional flow-matching target is

$$v_t^T(x \mid x_1) = \frac{\frac{\mathrm{d}\beta_t}{\mathrm{d}t}}{\beta_t}\, x + \left(\frac{\mathrm{d}\alpha_t}{\mathrm{d}t} - \frac{\frac{\mathrm{d}\beta_t}{\mathrm{d}t}}{\beta_t}\alpha_t\right) x_1. \tag{A.10}$$

Thus any admissible pair (α, β) yields an explicit clock-conditioned path/target pair. These pathwise identities do not by themselves guarantee a finite time-integrated quadratic risk; the outer-clock integrability theorem in Appendix C.6 concerns the straight-line schedule. Other schedules require their own time-integrability bounds.

VP-type path. A natural diffusion-inspired choice is the variance-preserving schedule. Let $\bar{\alpha} : [0,1] \to [0,1]$ be absolutely continuous with

$$\bar{\alpha}_0 = 0, \qquad \bar{\alpha}_1 = 1,$$

and assume $\bar{\alpha}_t < 1$ for $t \in [0,1)$. We also require $t \mapsto \sqrt{1-\bar{\alpha}_t^2}$ to be absolutely continuous, as needed for an admissible affine schedule. Define

$$\alpha_t = \bar{\alpha}_t, \qquad \beta_t = \sqrt{1-\bar{\alpha}_t^2}.$$

Then

$$X_t^T = \bar{\alpha}_t X_1^T + \sqrt{1-\bar{\alpha}_t^2}\, X_0^T, \tag{A.11}$$

which is the analogue of a variance-preserving interpolation. In this case, (A.10) becomes

$$v_t^T(x \mid x_1) = -\frac{\bar{\alpha}_t \frac{\mathrm{d}\bar{\alpha}_t}{\mathrm{d}t}}{1-\bar{\alpha}_t^2}\, x + \frac{\frac{\mathrm{d}\bar{\alpha}_t}{\mathrm{d}t}}{1-\bar{\alpha}_t^2}\, x_1. \tag{A.12}$$

VE-type path. A variance-exploding analogue can also be accommodated within the affine framework. Let $\sigma : [0,1] \to [0,1]$ be an absolutely continuous schedule such that

$$\sigma_0 = 1, \qquad \sigma_1 = 0, \qquad \sigma_t > 0 \qquad \text{for } t \in [0,1).$$

Set

$$\alpha_t = 1 - \sigma_t, \qquad \beta_t = \sigma_t.$$

Then the corresponding affine interpolation is

$$X_t^T = (1-\sigma_t)X_1^T + \sigma_t X_0^T, \qquad t \in [0,1]. \tag{A.13}$$

Conditioning on $X_1^T = x_1$ and the realized clock path $T_{[0,1]}$, we have

$$p_t^T(\cdot \mid x_1) = \mathrm{Law}(X_t^T \mid X_1^T = x_1) = \mathcal{N}\big((1-\sigma_t)x_1,\ \sigma_t^2 \Sigma(T_{[0,1]})\big). \tag{A.14}$$

Thus the conditional variance is directly controlled by the scale schedule σ_t^2. Although our interpolation is written in the generative direction from source to data, this corresponds to a variance-exploding schedule when viewed backward from $t = 1$ to $t = 0$.

Differentiating (A.13) gives

$$\frac{\mathrm{d}X_t^T}{\mathrm{d}t} = -\frac{\mathrm{d}\sigma_t}{\mathrm{d}t}X_1^T + \frac{\mathrm{d}\sigma_t}{\mathrm{d}t}X_0^T, \qquad X_0^T = \frac{X_t^T - (1-\sigma_t)X_1^T}{\sigma_t}.$$

Substituting into (A.10), or directly into $\frac{\mathrm{d}X_t^T}{\mathrm{d}t}$, yields the conditional flow-matching target

$$v_t^T(x \mid x_1) = \frac{\frac{\mathrm{d}\sigma_t}{\mathrm{d}t}}{\sigma_t}(x - x_1), \qquad t \in [0,1). \tag{A.15}$$

A simple concrete choice is the polynomial schedule

$$\sigma_t = (1-t)^\gamma, \qquad \gamma > 0,$$

for which

$$X_t^T = \big(1-(1-t)^\gamma\big)X_1^T + (1-t)^\gamma X_0^T,$$

and

$$v_t^T(x \mid x_1) = -\frac{\gamma}{1-t}(x - x_1). \tag{A.16}$$

More generally, any absolutely continuous decreasing schedule σ_t with the above endpoint conditions yields an admissible VE-type path within the same clock-conditioned affine framework.

B All-clock extension to diffusion models

In this appendix we record how the random-clock viewpoint extends to score-based diffusion models. The point of this subsection is slightly different from the flow-matching construction in the main text. Flow matching only needs a clock-conditioned interpolation and an explicit velocity target; it does not require reversing a stochastic process. Diffusion models do require a reverse-time dynamics, and for general càdlàg clocks this dynamics is not, in general, a purely local score SDE. The correct all-clock statement is a reverse-kernel formulation. The familiar local score SDE is recovered on the continuous part of the clock, and in particular for absolutely continuous clocks.

Let W be a standard d-dimensional Brownian motion independent of the clock T. Given drift $f : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ and deterministic diffusion matrix $G : [0,1] \to \mathbb{R}^{d\times d}$, consider the clock-conditioned forward dynamics

$$\mathrm{d}X_t = f_t(X_t)\,\mathrm{d}t + G_t\,\mathrm{d}W_{T_t}, \qquad X_0 \sim p_0(\cdot \mid T_{[0,1]}). \tag{B.1}$$

Assume that X_0 is independent of W conditional on T. For a fixed realized clock path, $M_t := W_{T_t}$ is a square-integrable martingale with predictable quadratic-covariation matrix $\langle M\rangle_t = T_t I_d$. If T is continuous, its quadratic covariation also equals $T_t I_d$; for a jump clock, $[M]_t = T_t^c I_d + \sum_{0<s\le t} \Delta M_s \Delta M_s^\top$, where T^c is the continuous part of T. Each jump is conditionally Gaussian with covariance $\Delta T_t I_d$. Write $a_t := G_t G_t^\top$ and $p_t^T := \mathrm{Law}(X_t \mid T_{[0,1]})$. Whenever p_t^T has a density, we define the clock-conditioned score by

$$s_t^T(x) := \nabla_x \log p_t^T(x). \tag{B.2}$$

As in the main text, the density is conditioned on the full clock path. This is the right object for clock-dependent initial laws such as $X_0 = L(T_{[0,1]})G$; in general it cannot be replaced by conditioning only on $T_{[0,t]}$.

We first state the reverse dynamics in a form that is valid for arbitrary realized càdlàg clock paths. For $0 \le s < t \le 1$, let $P_{s,t}^T(x, \mathrm{d}y)$ denote the conditional forward transition kernel of (B.1) from time s to time t, given the full clock path. Thus $p_s^T P_{s,t}^T = p_t^T$. The reverse transition from t back to s is the Bayes kernel associated with this joint law.

Theorem B.1 (All-clock reverse dynamics in kernel form)**.** *Fix a realized clock path $T \in \mathbb{D}([0,1];\mathbb{R}_+)$. Assume that the forward equation* (B.1) *is well posed and defines a Markov process conditional on T, that regular conditional transition kernels $P_{s,t}^T$ exist, and that $p_s^T P_{s,t}^T = p_t^T$ for $0 \le s < t \le 1$. For $s < t$, define a reverse kernel $R_{t,s}^T(y, \mathrm{d}x)$ by the measure identity*

$$p_t^T(\mathrm{d}y)\, R_{t,s}^T(y, \mathrm{d}x) = p_s^T(\mathrm{d}x)\, P_{s,t}^T(x, \mathrm{d}y) \tag{B.3}$$

on $\mathbb{R}^d \times \mathbb{R}^d$. Equivalently, when transition densities exist,

$$R_{t,s}^T(y, \mathrm{d}x) = \frac{p_s^T(x) p_{s,t}^T(y \mid x)}{p_t^T(y)}\,\mathrm{d}x. \tag{B.4}$$

Let $(Y_r)_{r\in[0,1]}$ be the reverse-time Markov process initialized by $Y_0 \sim p_1^T$ and, for $0 \le r < r' \le 1$, using the transition

$$\mathrm{Law}(Y_{r'} \in \mathrm{d}x \mid Y_r = y, T_{[0,1]}) = R_{1-r,\,1-r'}^T(y, \mathrm{d}x). \tag{B.5}$$

Then the reverse process is conditional marginal-matching:

$$\mathrm{Law}(Y_r \mid T_{[0,1]}) = p_{1-r}^T, \qquad r \in [0,1]. \tag{B.6}$$

Proof. Fix $0 \le s < t \le 1$. Integrating (B.3) over $y \in \mathbb{R}^d$ gives, for every measurable set $A \subseteq \mathbb{R}^d$,

$$\int_{\mathbb{R}^d} R_{t,s}^T(y, A)\, p_t^T(\mathrm{d}y) = \int_A P_{s,t}^T(x, \mathbb{R}^d)\, p_s^T(\mathrm{d}x) = p_s^T(A).$$

Thus a single reverse step transports p_t^T back to p_s^T. Taking $t = 1 - r$ and $s = 1 - r'$, this says that if $Y_r \sim p_{1-r}^T$, then the transition (B.5) gives $Y_{r'} \sim p_{1-r'}^T$. Since $Y_0 \sim p_1^T$, iterating this identity over any finite partition of $[0,1]$ yields $\mathrm{Law}(Y_r \mid T_{[0,1]}) = p_{1-r}^T$ for every $r \in [0,1]$. The Markov kernels are consistent because they are the regular conditional kernels of the same forward Markov law read backward in time, so the corresponding reverse process exists by the usual extension construction. □

At a forward jump time t, the exact-time convention in (B.5) assigns Y_{1-t} the law p_t^T and its right limit the law p_{t-}^T. Thus the time-reversed version is left-continuous at reversed jumps. A càdlàg reverse version instead uses forward left limits, with the corresponding one-sided marginal convention. The reverse jump kernel below maps the post-jump law p_t^T to the pre-jump law p_{t-}^T.

The theorem is deliberately stated without a score. A score is a local object, whereas a jump of the clock produces a finite, nonlocal update of the law. This can be seen explicitly. Let t be a jump time of the realized clock and set $h := \Delta T_t = T_t - T_{t-} > 0$. Since the drift term has no atom in physical time, the forward jump is

$$X_t = X_{t-} + G_t Z_t, \qquad Z_t \sim \mathcal{N}(0, hI_d),$$

so the jump kernel is $Q_t^T(x, \mathrm{d}y) = \mathcal{N}(y; x, ha_t)(\mathrm{d}y)$. Therefore

$$p_t^T(\mathrm{d}y) = \int Q_t^T(x, \mathrm{d}y)\, p_{t-}^T(\mathrm{d}x),$$

and the reverse jump is the Bayes kernel characterized by

$$p_t^T(\mathrm{d}y) R_{t,t-}^T(y, \mathrm{d}x) = p_{t-}^T(\mathrm{d}x) Q_t^T(x, \mathrm{d}y). \tag{B.7}$$

When p_{t-}^T has a density and a_t is nonsingular, this becomes the density formula

$$R_{t,t-}^T(y, \mathrm{d}x) = \frac{p_{t-}^T(x)\varphi_{ha_t}(y-x)}{\int p_{t-}^T(z)\varphi_{ha_t}(y-z)\,\mathrm{d}z}\,\mathrm{d}x, \tag{B.8}$$

where φ_{ha_t} is the Gaussian density with covariance ha_t. Thus reverse jumps are posterior sampling steps. They cannot be represented exactly by the local drift term $a_t s_t^T\,\mathrm{d}T_t$.

We next recover the usual score-based formula from the kernel theorem on the continuous part of the clock. Write

$$T_t^c := T_t - \sum_{0<s\le t} \Delta T_s$$

for the continuous part of the realized clock, and define the reversed continuous clock $\widehat{T}_r^c := T_1^c - T_{1-r}^c$. At reverse times corresponding to jumps of T, the process uses the jump kernel (B.7); between these jumps, or if the clock is continuous, the reverse dynamics have the following local representation.

Corollary B.2 (Local score form on the continuous clock part)**.** *Assume the hypotheses of Theorem B.1. In addition, assume that the continuous part of the forward law admits strictly positive smooth densities, that the usual decay and integrability conditions for integration by parts hold, and that the local reverse equations below are well posed. On the continuous clock part, with $\tau = 1 - r$, the reverse process has the local score representation*

$$\mathrm{d}Y_r = -f_\tau(Y_r)\,\mathrm{d}r + a_\tau s_\tau^T(Y_r)\,\mathrm{d}\widehat{T}_r^c + G_\tau\,\mathrm{d}\bar{W}_{\widehat{T}_r^c}, \tag{B.9}$$

where $\bar{W}$ is a standard Brownian motion. If T is continuous, this local equation alone gives $\mathrm{Law}(Y_r \mid T_{[0,1]}) = p_{1-r}^T$ for all $r \in [0,1]$. If T has jumps, it must be supplemented by the reverse jump kernels (B.7) *at the corresponding reverse jump times.*

For the continuous part, the deterministic probability-flow equation is

$$\mathrm{d}Z_r = -f_\tau(Z_r)\,\mathrm{d}r + \frac{1}{2} a_\tau s_\tau^T(Z_r)\,\mathrm{d}\widehat{T}_r^c. \tag{B.10}$$

When T has jumps, an exact deterministic flow across a jump additionally requires choosing a transport map from p_t^T to p_{t-}^T; this choice is not canonical and is not determined by the score alone.

Proof. We prove the continuous part; the jump updates are already identified by Theorem B.1. Let $\psi \in C_c^\infty(\mathbb{R}^d)$. Along the continuous part of the forward dynamics, the conditional law satisfies the weak equation

$$\mathrm{d}\int \psi(x) p_t^T(x)\,\mathrm{d}x = \int f_t(x)\cdot\nabla\psi(x) p_t^T(x)\,\mathrm{d}x\,\mathrm{d}t + \frac{1}{2}\int \mathrm{Tr}(a_t\nabla^2\psi(x)) p_t^T(x)\,\mathrm{d}x\,\mathrm{d}T_t^c.$$

Set $q_r = p^T_{1-r}$ and $\tau = 1 - r$. Since $\widehat{T}^c_r = T^c_1 - T^c_{1-r}$, this forward weak equation gives

$$\mathrm{d}\int \psi(x) q_r(x)\,\mathrm{d}x = -\int f_\tau(x)\cdot\nabla\psi(x) q_r(x)\,\mathrm{d}x\,\mathrm{d}r - \frac{1}{2}\int \mathrm{Tr}(a_\tau\nabla^2\psi(x)) q_r(x)\,\mathrm{d}x\,\mathrm{d}\widehat{T}^c_r.$$

Now let $\rho_r = \mathrm{Law}(Y_r \mid T_{[0,1]})$ for the local SDE (B.9). Its weak equation is

$$\begin{aligned}\mathrm{d}\int \psi(x)\rho_r(x)\,\mathrm{d}x = &-\int f_\tau(x)\cdot\nabla\psi(x)\rho_r(x)\,\mathrm{d}x\,\mathrm{d}r\\ &+\int a_\tau s^T_\tau(x)\cdot\nabla\psi(x)\rho_r(x)\,\mathrm{d}x\,\mathrm{d}\widehat{T}^c_r\\ &+\frac{1}{2}\int \mathrm{Tr}(a_\tau\nabla^2\psi(x))\rho_r(x)\,\mathrm{d}x\,\mathrm{d}\widehat{T}^c_r.\end{aligned}$$

Substituting $\rho_r = q_r = p^T_\tau$ and using $p^T_\tau s^T_\tau = \nabla p^T_\tau$, integration by parts gives

$$\int a_\tau s^T_\tau\cdot\nabla\psi\, p^T_\tau\,\mathrm{d}x = \int a_\tau\nabla p^T_\tau\cdot\nabla\psi\,\mathrm{d}x = -\int \mathrm{Tr}(a_\tau\nabla^2\psi)p^T_\tau\,\mathrm{d}x.$$

Therefore the two $\mathrm{d}\widehat{T}^c_r$-terms combine to $-\frac{1}{2}\int \mathrm{Tr}(a_\tau\nabla^2\psi)p^T_\tau\,\mathrm{d}x\,\mathrm{d}\widehat{T}^c_r$, which matches the weak equation for q_r. With matching initial law at $r = 0$, uniqueness gives the stated marginal matching.

For the deterministic equation (B.10), the weak equation has the same drift term but only $\frac{1}{2}\int a_\tau s^T_\tau\cdot\nabla\psi\,\zeta_r\,\mathrm{d}x\,\mathrm{d}\widehat{T}^c_r$ in the clock direction. Substituting $\zeta_r = q_r$ and applying the same integration by parts yields exactly the same weak equation for q_r. Hence $\mathrm{Law}(Z_r \mid T_{[0,1]}) = p^T_{1-r}$ on the continuous part. □

If the clock is absolutely continuous, $T_t = \int_0^t \ell_s\,\mathrm{d}s$, then $\mathrm{d}\widehat{T}^c_r = \ell_\tau\,\mathrm{d}r$, and (B.9) reduces to

$$\mathrm{d}Y_r = \big[-f_\tau(Y_r) + \ell_\tau a_\tau s^T_\tau(Y_r)\big]\mathrm{d}r + \sqrt{\ell_\tau}\,G_\tau\,\mathrm{d}\bar{B}_r, \qquad \tau = 1 - r,$$

while the probability flow becomes

$$\mathrm{d}Z_r = \big[-f_\tau(Z_r) + \tfrac{1}{2}\ell_\tau a_\tau s^T_\tau(Z_r)\big]\mathrm{d}r.$$

This is the classical reverse-SDE/probability-flow structure with the diffusion rate modulated by the realized clock rate.

The learning target associated with the local continuous dynamics is the clock-conditioned score s^T_t. For the learning objectives below, we restrict to absolutely continuous clocks $T_t = \int_0^t \ell_s\,\mathrm{d}s$, so the score is needed at Lebesgue-almost-every physical time. A pathwise score-matching objective is

$$\mathcal{L}_{\mathrm{DSM}}(\theta) := \mathbb{E}_{t\sim\mathrm{Unif}(0,1)}\omega_t\,\mathbb{E}_T\left[\mathbb{E}_{X_t|T}\left[\big\|s_\theta(X_t,t,\eta_T) - s^T_t(X_t)\big\|_2^2\right]\right], \tag{B.11}$$

where $\omega_t > 0$ is an optional weighting schedule. The all-clock kernel theorem remains valid beyond this learning setting. For a singular continuous clock, uniform sampling of physical time does not identify the score on the singular part of $\mathrm{d}T^c$; time sampling must instead account for that clock measure. For clocks with jumps, exact reversal additionally requires the nonlocal kernels in (B.7), or an approximation to them.

In practice, the marginal score s^T_t is typically unavailable. One may therefore use a clock-conditioned denoising score-matching objective whenever the forward transition kernel given $(X_0, T_{[0,1]})$ is tractable:

$$\begin{aligned}&\mathcal{L}_{\mathrm{CDSM}}(\theta)\\ :=&\mathbb{E}_{t\sim\mathrm{Unif}(0,1)}\omega_t\,\mathbb{E}_{X_0,T}\left[\mathbb{E}_{X_t|X_0,T}\left[\big\|s_\theta(X_t,t,\eta_T) - \nabla_{x_t}\log p_t(x_t \mid X_0, T_{[0,1]})\big|_{x_t=X_t}\big\|_2^2\right]\right].\end{aligned} \tag{B.12}$$

Under the same square-integrability and projection conditions as standard denoising score matching, the conditional score in (B.12) projects onto the marginal score s^T_t after conditioning on $(X_t, T_{[0,1]})$.

For linear-Gaussian forward schedules, the conditional kernel is Gaussian. If

$$X_t \mid (X_0, T_{[0,1]}) \sim \mathcal{N}\big(m_t(X_0, T_{[0,1]}),\, Q_t(T_{[0,1]})\big),$$

with $Q_t(T_{[0,1]})$ positive definite, then

$$\nabla_{x_t} \log p_t(x_t \mid X_0, T_{[0,1]}) = -\, Q_t(T_{[0,1]})^{-1}\big(x_t - m_t(X_0, T_{[0,1]})\big),$$

and (B.12) becomes

$$\begin{aligned}&\mathcal{L}_{\mathrm{CDSM}}(\theta)\\ :=&\mathbb{E}_{t\sim\mathrm{Unif}(0,1)}\omega_t\, \mathbb{E}_{X_0,T}\left[\mathbb{E}_{X_t|X_0,T}\left[\left\| s_\theta(X_t,t,\eta_T) + Q_t(T_{[0,1]})^{-1}\big(X_t - m_t(X_0,T_{[0,1]})\big)\right\|_2^2\right]\right].\end{aligned} \tag{B.13}$$

C Additional results and proofs of main theorems

C.1 A note on conditioning on the clock history

The main construction conditions on the full realized clock path $T_{[0,1]}$. In general this conditioning cannot be replaced by conditioning only on the past clock history $T_{[0,t]}$, because the clock-conditioned source may itself depend on future clock values such as T_1. The following lemma records the additional non-anticipativity condition under which such a reduction is valid.

Lemma C.1 (Clock-history reduction under non-anticipative initialization)**.** *Consider the forward construction* (B.1)*. Assume that, for every $t \in [0,1]$, the initial variable X_0 is conditionally independent of $\sigma(T_{(t,1]})$ given $\sigma(T_{[0,t]})$, and that the Brownian motion driving the SDE is independent of the clock and of the base randomness used to form X_0. Then, for every $t \in [0,1]$,*

$$\mathrm{Law}(X_t \mid T_{[0,1]}) = \mathrm{Law}(X_t \mid T_{[0,t]}) \qquad \textit{a.s.}$$

Equivalently, whenever both sides admit conditional densities,

$$p_t(\cdot \mid T_{[0,1]}) = p_t(\cdot \mid T_{[0,t]}) \qquad \textit{a.s.}$$

Proof. Fix $t \in [0,1]$ and set

$$\mathcal{G}_t := \sigma(X_0, T_{[0,t]}) \vee \sigma(W_u : 0 \le u \le T_t).$$

By non-anticipativity of the forward dynamics, X_t is measurable with respect to $\mathcal{G}_t$. Since T_t is $\sigma(T_{[0,t]})$-measurable and the Brownian motion W is independent of the entire clock, $\sigma(W_u : 0 \le u \le T_t)$ is conditionally independent of $\sigma(T_{(t,1]})$ given $\sigma(T_{[0,t]})$. By assumption the same is true for X_0. Hence $\mathcal{G}_t$, and therefore X_t, is conditionally independent of $\sigma(T_{(t,1]})$ given $\sigma(T_{[0,t]})$. For any bounded measurable test function φ,

$$\mathbb{E}\left[\varphi(X_t) \,\middle|\, \sigma(T_{[0,1]})\right] = \mathbb{E}\left[\varphi(X_t) \,\middle|\, \sigma(T_{[0,t]})\right] \qquad \text{a.s.},$$

which proves the claim. □

The non-anticipativity assumption in Lemma C.1 is essential. For example, if $X_0 \mid T_{[0,1]} \sim \mathcal{N}(0, T_1 I_d)$, then already at $t = 0$ the conditional law given $T_{[0,1]}$ is $\mathcal{N}(0, T_1 I_d)$, whereas the conditional law given $T_{[0,0]}$ is the corresponding mixture over T_1. These laws are generally different. This is why the main flow-matching construction consistently uses $p_t^T = \mathrm{Law}(X_t^T) = \mathrm{Law}(X_t \mid T_{[0,1]})$ rather than replacing it by a past-clock conditional law.

C.2 Conditional Gaussian law of the affine path

We verify the conditional Gaussian law (3.9) stated in the main text. Recall the assumptions: conditional on the realized clock path $T_{[0,1]}$, $X_0^T \sim \mathcal{N}(0, \Sigma(T_{[0,1]}))$; X_0^T and X_1^T are conditionally independent given $T_{[0,1]}$; and the affine path is $X_t^T = \alpha_t^T X_1^T + \beta_t^T X_0^T$ with deterministic schedules α_t^T, β_t^T given $T_{[0,1]}$.

Proof of (3.9). Fix $T_{[0,1]}$ and $X_1^T = x_1$. Under this conditioning, α_t^T and β_t^T are deterministic, and the conditional independence assumption gives $X_0^T \mid (X_1^T = x_1,\, T_{[0,1]}) \sim \mathcal{N}(0, \Sigma(T_{[0,1]}))$. Hence

$$X_t^T \mid (X_1^T = x_1,\, T_{[0,1]}) = \alpha_t^T x_1 + \beta_t^T\, X_0^T \,\big|\, (X_1^T = x_1,\, T_{[0,1]})$$

is an affine transformation of a Gaussian, with mean $\alpha_t^T x_1$ and covariance $(\beta_t^T)^2\Sigma(T_{[0,1]})$. Therefore

$$X_t^T \,\big|\, (X_1^T = x_1,\, T_{[0,1]}) \sim \mathcal{N}\big(\alpha_t^T x_1,\, (\beta_t^T)^2\Sigma(T_{[0,1]})\big),$$

which is precisely (3.9). □

C.3 Detailed pathwise equivalence result and proof

Proposition C.2 (Detailed pathwise conditional–marginal equivalence)**.** *Assume* (3.3) *and suppose that, for every θ in the parameter class and for a.e. $t \in [0,1]$,*

$$\mathbb{E}\left[\|u_\theta(X_t^T, t, T_{[0,1]})\|_2^2 + \|v_t^T(X_t^T \mid X_1^T)\|_2^2 \,\middle|\, T_{[0,1]}\right] < \infty \qquad \textit{a.s.}$$

Then, for a.e. t, there exists a nonnegative $\sigma(T_{[0,1]})$-measurable random variable $C_t(T)$, independent of θ, such that

$$\mathbb{E}\left[R_t^{\mathrm{cond}}(\theta; X_1^T, T) \,\middle|\, T_{[0,1]}\right] = R_t^{\mathrm{marg}}(\theta; T) + C_t(T).$$

Consequently, whenever the risks and $C_t(T)$ are integrable over (t,T) with $t \sim \mathrm{Unif}(0,1)$,

$$\mathcal{L}_{\mathrm{CHTFM}}(\theta) = \mathcal{L}_{\mathrm{HTFM}}(\theta) + \mathbb{E}_{t\sim\mathrm{Unif}(0,1)}\mathbb{E}_T\left[C_t(T)\right].$$

In particular, if the additive constant is finite, the two population objectives have the same minimizers. If u_θ is differentiable in θ and differentiation may be exchanged with the conditional expectations, the clock expectation, and the uniform-time expectation, then

$$\nabla_\theta \mathcal{L}_{\mathrm{CHTFM}}(\theta) = \nabla_\theta \mathcal{L}_{\mathrm{HTFM}}(\theta).$$

Proof of Proposition C.2. Fix a time t for which the assumptions hold. By (3.3),

$$u_t^T(x) = \mathbb{E}\left[v_t^T(X_t^T \mid X_1^T) \,\middle|\, X_t^T = x\right]$$

for p_t^T-a.e. x, where expectations are taken in the fixed-clock conditional probability space. The same identity is understood in the weak sense when densities are not used. Thus, within the clock-conditioned environment indexed by T, $u_t^T(X_t^T)$ is the L^2-projection of the pathwise conditional velocity $v_t^T(X_t^T \mid X_1^T)$ onto the sigma-field $\sigma(X_t^T)$. Applying the conditional Pythagorean identity gives

$$\begin{aligned}\mathbb{E}\left[R_t^{\mathrm{cond}}(\theta; X_1^T, T) \,\middle|\, T_{[0,1]}\right] &= R_t^{\mathrm{marg}}(\theta; T) + \mathbb{E}\left[\left\|v_t^T(X_t^T \mid X_1^T) - u_t^T(X_t^T)\right\|_2^2 \,\middle|\, T_{[0,1]}\right] \\ &= R_t^{\mathrm{marg}}(\theta; T) + C_t(T),\end{aligned}$$

where

$$C_t(T) := \mathbb{E}\left[\left\|v_t^T(X_t^T \mid X_1^T) - u_t^T(X_t^T)\right\|_2^2 \,\middle|\, T_{[0,1]}\right] \geq 0.$$

If the risks and $C_t(T)$ are integrable over (t,T) with $t \sim \mathrm{Unif}(0,1)$, then Fubini's theorem yields

$$\mathcal{L}_{\mathrm{CHTFM}}(\theta) = \mathcal{L}_{\mathrm{HTFM}}(\theta) + \mathbb{E}_{t\sim\mathrm{Unif}(0,1)}\mathbb{E}_T\left[C_t(T)\right]. \tag{C.1}$$

The last term is independent of θ. Hence, whenever it is finite, the two population objectives have the same minimizers. If u_θ is differentiable in θ and differentiation may be exchanged with the conditional expectations, the clock expectation, and the uniform-time expectation, then $\nabla_\theta \mathcal{L}_{\mathrm{CHTFM}}(\theta) = \nabla_\theta \mathcal{L}_{\mathrm{HTFM}}(\theta)$. □

C.4 Clock-conditioned straight-line geodesic

Theorem C.3 (Clock-conditioned straight-line geodesic)**.** *Fix $p \in [1,\infty)$. For a.e. realized clock path T, assume $p_0(\cdot \mid T_{[0,1]}), p_{\mathrm{data}} \in \mathcal{P}_p(\mathbb{R}^d)$, and let π^T be a p-optimal coupling between them. Then $\mu_t^T := \big((1-t)x_0 + tx_1\big)_{\#}\pi^T$, $t \in [0,1]$, is a constant-speed W_p-geodesic from $p_0(\cdot \mid T_{[0,1]})$ to p_{data}.*

Proof. Fix a realized clock path T for which $p_0(\cdot \mid T_{[0,1]}) \in \mathcal{P}_p(\mathbb{R}^d)$, and abbreviate

$$\mu_0 := p_0(\cdot \mid T_{[0,1]}), \qquad \mu_1 := p_{\mathrm{data}}, \qquad \mu_r := \big((1-r)x_0 + rx_1\big)_{\#}\pi^T, \quad r \in [0,1].$$

Then $\mu_t^T = \mu_t$. By assumption $\mu_0, \mu_1 \in \mathcal{P}_p(\mathbb{R}^d)$, so the p-optimal transport problem is well posed and $W_p(\mu_0, \mu_1) < \infty$.

For any $r, r' \in [0,1]$, the pushforward

$$\big((1-r)x_0 + rx_1,\ (1-r')x_0 + r'x_1\big)_{\#}\pi^T$$

is a coupling of μ_r and $\mu_{r'}$. Hence, using the optimality of π^T at the last step,

$$W_p(\mu_r, \mu_{r'})^p \le \int \|(r-r')(x_1 - x_0)\|^p \, d\pi^T = |r-r'|^p \, W_p(\mu_0, \mu_1)^p. \tag{C.2}$$

Now fix $0 \le r \le r' \le 1$. Applying (C.2) to the pairs $(0, r)$, (r, r'), and $(r', 1)$, we obtain

$$\begin{aligned} &W_p(\mu_0, \mu_r) + W_p(\mu_r, \mu_{r'}) + W_p(\mu_{r'}, \mu_1) \\ &\le r\, W_p(\mu_0, \mu_1) + (r' - r)\, W_p(\mu_0, \mu_1) + (1 - r')\, W_p(\mu_0, \mu_1) \\ &= W_p(\mu_0, \mu_1). \end{aligned}$$

On the other hand, the triangle inequality yields

$$W_p(\mu_0, \mu_1) \le W_p(\mu_0, \mu_r) + W_p(\mu_r, \mu_{r'}) + W_p(\mu_{r'}, \mu_1).$$

Therefore equality holds throughout, and in particular,

$$W_p(\mu_r, \mu_{r'}) = |r - r'|\, W_p(\mu_0, \mu_1) \qquad \text{for all } r, r' \in [0,1].$$

Thus $(\mu_t^T)_{t\in[0,1]}$ is a constant-speed W_p-geodesic between $p_0(\cdot \mid T_{[0,1]})$ and p_{data}. □

Remark C.4. *For each fixed clock path with finite covariance matrix, $p_0(\cdot \mid T_{[0,1]}) = \mathcal{N}(0, \Sigma(T_{[0,1]}))$ belongs to $\mathcal{P}_q(\mathbb{R}^d)$ for every $q \ge 1$. Thus Theorem C.3 is pathwise in T: after conditioning on the clock, the source side does not restrict the choice of p, and the admissible p is governed by the data moment assumption. This does not imply that the clock-marginal source law has all moments; after integrating out T, stable or Student-t clock choices can again produce heavy-tailed sources with limited or infinite moments. The classical W_2-statement is recovered when the data law has a finite second moment and the argument is applied pathwise.*

C.5 Proof of Theorem 3.3

Proof of Theorem 3.3. By (3.9) and the source covariance $\Sigma_A(T_{[0,1]}) := 2A(T)I_d$, where $A(T) := \int_0^1 T_s \, \mathrm{d}s$, for every fixed clock path and endpoint $X_1^T = x_1$,

$$X_t^T \mid X_1^T = x_1 \sim \mathcal{N}\big(\alpha_t x_1,\ 2\beta_t^2 A(T) I_d\big), \qquad A(T) = \int_0^1 T_s \, \mathrm{d}s.$$

Thus the clock-marginal conditional law is a Gaussian scale mixture.

For the deterministic clock $T_t = t$, $A(T) = 1/2$, so the covariance above is $\beta_t^2 I_d$. This proves the Gaussian case.

For the stable case, let $\alpha \in (0,2)$, $\rho = \alpha/2$, and let T be the ρ-stable subordinator in the theorem. With $A(T) = \int_0^1 T_s \mathrm{d}s$, the integration-by-parts identity for Lebesgue–Stieltjes integrals gives

$$A(T) = \int_0^1 (1-s) \, \mathrm{d}T_s.$$

Using the independent increments of the subordinator, for any $\lambda \ge 0$,

$$\mathbb{E}\left[e^{-\lambda A(T)}\right] = \exp\left(-\frac{\rho+1}{2} \int_0^1 (\lambda(1-s))^\rho \, \mathrm{d}s\right) = \exp\left(-\frac{\lambda^\rho}{2}\right). \tag{C.3}$$

Therefore, for any $\xi \in \mathbb{R}^d$,

$$\mathbb{E}\left[e^{i\langle \xi, X_t\rangle} \mid X_1 = x_1\right] = \exp\big(i\alpha_t\langle \xi, x_1\rangle\big)\, \mathbb{E}_T\left[\exp\left(-\beta_t^2 A(T)\|\xi\|_2^2\right)\right]$$
$$= \exp\big(i\alpha_t\langle \xi, x_1\rangle\big)\, \exp\left(-\frac{\beta_t^\alpha \|\xi\|_2^\alpha}{2}\right). \tag{C.4}$$

This is the characteristic function of a location-shifted isotropic α-stable law with location $\alpha_t x_1$ and scale β_t, under the normalization used in the main text. The radial tail asymptotics $\mathbb{P}(\|X_t - \alpha_t x_1\|_2 > r \mid X_1 = x_1) = \Theta(r^{-\alpha})$ as $r \to \infty$ then follow from the standard tail behavior of isotropic α-stable laws [45].

For the Student-t case, $T_t = tV_\nu$ gives $A(T) = V_\nu/2$, and hence $\Sigma_A(T_{[0,1]}) = V_\nu I_d$. Since $V_\nu \sim \mathrm{InvGamma}(\nu/2, \nu/2)$, marginalizing over V_ν yields

$$p_{X_t|X_1=x_1}(x) = \int_0^\infty \frac{1}{(2\pi\beta_t^2 v)^{d/2}} \exp\left(-\frac{\|x - \alpha_t x_1\|_2^2}{2\beta_t^2 v}\right) \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} v^{-\nu/2-1} e^{-\nu/(2v)}\, \mathrm{d}v$$
$$= \frac{\Gamma((\nu+d)/2)}{\Gamma(\nu/2)(\nu\pi)^{d/2}\beta_t^d} \left(1 + \frac{\|x - \alpha_t x_1\|_2^2}{\nu\beta_t^2}\right)^{-(\nu+d)/2}. \tag{C.5}$$

This is the multivariate Student-t density with location $\alpha_t x_1$, scale $\beta_t^2 I_d$, and ν degrees of freedom, whose radial tail satisfies $\mathbb{P}(\|X_t - \alpha_t x_1\|_2 > r \mid X_1 = x_1) = \Theta(r^{-\nu})$ as $r \to \infty$ [45]. □

C.6 Outer-clock integrability and local finite-risk gradients

This subsection gives the outer-clock result and complete proofs for the straight-line flow. It complements the pathwise projection result in Appendix C.3; it does not assert finite population risk for arbitrary network parameters.

C.6.1 Notation and objectives

Let $\mathcal{T} := T_{[0,1]} \sim P_T$ be the realized clock path, $X_1 \sim p_{\mathrm{data}}$, and $X_0^{\mathcal{T}} \sim p_0(\cdot \mid \mathcal{T})$. Let $\tau \sim \mathrm{Unif}(0,1)$ be independent of all other randomness and set

$$X_\tau^{\mathcal{T}} = \tau X_1 + (1-\tau)X_0^{\mathcal{T}}, \qquad Y^{\mathcal{T}} := X_1 - X_0^{\mathcal{T}}. \tag{C.6}$$

At sampled points, $v_\tau^{\mathcal{T}}(X_\tau^{\mathcal{T}} \mid X_1) = Y^{\mathcal{T}}$; hence no residual $(1-\tau)^{-1}$ singularity remains.

For fixed $(t, \mathcal{T})$, write $\mathbb{E}_{X_t^{\mathcal{T}}}$ for expectation under $p_t^{\mathcal{T}} := \mathrm{Law}(X_t^{\mathcal{T}} \mid \mathcal{T})$, and $\mathbb{E}_{X_t^{\mathcal{T}}|X_1}$ for expectation under $\mathrm{Law}(X_t^{\mathcal{T}} \mid X_1, \mathcal{T})$, as in the paper. Define

$$u_t^{\mathcal{T}}(X_t^{\mathcal{T}}) := \mathbb{E}_{X_1|X_t^{\mathcal{T}},\mathcal{T}}\big[v_t^{\mathcal{T}}(X_t^{\mathcal{T}} \mid X_1)\big]. \tag{C.7}$$

Let $Z := (X_\tau^{\mathcal{T}}, \tau, \mathcal{T})$, $P_Z := \mathrm{Law}(Z)$, and $u^\star(Z) := u_\tau^{\mathcal{T}}(X_\tau^{\mathcal{T}})$. For any measurable full-clock predictor u,

$$\mathcal{L}_{\mathrm{CHTFM}}(u) := \mathbb{E}_{\tau\sim\mathrm{Unif}(0,1)}\mathbb{E}_{\mathcal{T},X_1}\left[\mathbb{E}_{X_\tau^{\mathcal{T}}|X_1}\left\|u(Z) - Y^{\mathcal{T}}\right\|^2\right], \tag{C.8}$$

$$\mathcal{L}_{\mathrm{HTFM}}(u) := \mathbb{E}_{\tau\sim\mathrm{Unif}(0,1)}\mathbb{E}_{\mathcal{T}}\left[\mathbb{E}_{X_\tau^{\mathcal{T}}}\left\|u(Z) - u^\star(Z)\right\|^2\right]. \tag{C.9}$$

A feature-conditioned network is included by setting $u_\theta(Z) := u_\theta(X_\tau^{\mathcal{T}}, \tau, \eta(\mathcal{T}))$. For parameterized fields, write $\mathcal{L}_{\mathrm{CHTFM}}(\theta) := \mathcal{L}_{\mathrm{CHTFM}}(u_\theta)$, and similarly for $\mathcal{L}_{\mathrm{HTFM}}$.

Assumption C.5 (Primitive heavy-tail conditions)**.** *Assume:*

(H1) *For some $p \in (1,2)$, $M_p := \mathbb{E}_{X_1}\|X_1\|^p < \infty$.*

(H2) *$X_1 \perp \mathcal{T}$ and $X_0^{\mathcal{T}} \perp X_1 \mid \mathcal{T}$.*

(H3) *There are a clock-measurable $a(\mathcal{T}) \geq 0$, $q > p/(p-1)$, and $C_q < \infty$ such that, P_T-a.s.,*

$$\mathbb{E}_{X_0^{\mathcal{T}}|\mathcal{T}}\left\|X_0^{\mathcal{T}}\right\|^2 \leq a(\mathcal{T})^2, \qquad \mathbb{E}_{X_0^{\mathcal{T}}|\mathcal{T}}\left\|X_0^{\mathcal{T}}\right\|^q \leq C_q a(\mathcal{T})^q, \qquad \mathbb{E}_{\mathcal{T}} a(\mathcal{T}) < \infty. \tag{C.10}$$

Both $\mathbb{E}_{X_1}\|X_1\|^2$ *and* $\mathbb{E}_{\mathcal{T}}\mathbb{E}_{X_0^{\mathcal{T}}|\mathcal{T}}\left\|X_0^{\mathcal{T}}\right\|^2$ *may be infinite.*

Theorem C.6 (Finite outer residual and exact finite-risk domain)**.** *Under Assumption C.5, the clock-averaged projection residual*

$$\begin{aligned} C_\star &:= \mathbb{E}_{\tau\sim\mathrm{Unif}(0,1)}\mathbb{E}_{\mathcal{T}}[C_\tau(\mathcal{T})], \\ C_t(\mathcal{T}) &:= \mathbb{E}_{X_1}\left[\mathbb{E}_{X_t^{\mathcal{T}}|X_1}\left\|Y^{\mathcal{T}} - u_t^{\mathcal{T}}(X_t^{\mathcal{T}})\right\|^2\right]. \end{aligned} \tag{C.11}$$

is finite. More precisely, for a constant $K = K(p,q,C_q) < \infty$*,*

$$C_\star \le K(1+M_p)\mathbb{E}_{\mathcal{T}}[a(\mathcal{T})] < \infty. \tag{C.12}$$

Proposition C.7 (Exact finite-risk domain)**.** *Under Assumption C.5, for every measurable predictor* u*,*

$$\mathcal{L}_{\mathrm{CHTFM}}(u) < \infty \quad\Longleftrightarrow\quad \mathcal{L}_{\mathrm{HTFM}}(u) < \infty, \tag{C.13}$$

and, on this exact finite-risk domain,

$$\mathcal{L}_{\mathrm{CHTFM}}(u) = \mathcal{L}_{\mathrm{HTFM}}(u) + C_\star. \tag{C.14}$$

Equivalently, the finite-risk prediction class is $u^\star + L^2(P_Z;\mathbb{R}^d)$*. Neither* $Y^{\mathcal{T}}$ *nor* $u^\star(Z)$ *must be square-integrable; only their difference is.*

Corollary C.8 (Local finite-risk basin)**.** *Assume Theorem C.6. Let* $\Theta \subset \mathbb{R}^{d_\theta}$*, and suppose:*

(P1) *Some* $\bar\theta \in \Theta$ *satisfies* $\mathcal{L}_{\mathrm{CHTFM}}(\bar\theta) < \infty$*, and the extended-valued objective* $\mathcal{L}_{\mathrm{CHTFM}} : \Theta \to [0,\infty]$ *attains a minimum at* $\theta^\star$*.*

(P2) *For some* $r > 0$*,* $B_r(\theta^\star) \subset \Theta$*,* $\theta \mapsto u_\theta(z)$ *is continuously differentiable on this ball for* P_Z*-a.e.* z*, and*

$$J_r(z) := \sup_{\vartheta\in B_r(\theta^\star)} \|D_\theta u_\vartheta(z)\|_{\mathrm{op}}, \qquad K_r^2 := \mathbb{E}_{Z\sim P_Z}[J_r(Z)^2] < \infty. \tag{C.15}$$

Set $\varepsilon_\star^2 := \mathcal{L}_{\mathrm{HTFM}}(\theta^\star) < \infty$*. Then every* $\theta \in B_r(\theta^\star)$ *satisfies*

$$\sqrt{\mathcal{L}_{\mathrm{HTFM}}(\theta)} \le \varepsilon_\star + K_r\|\theta-\theta^\star\|, \qquad \mathcal{L}_{\mathrm{CHTFM}}(\theta) \le C_\star + \left(\varepsilon_\star + K_r\|\theta-\theta^\star\|\right)^2 < \infty. \tag{C.16}$$

At every interior point of the ball, both gradients are finite and

$$\begin{aligned} \nabla_\theta\mathcal{L}_{\mathrm{CHTFM}}(\theta) &= 2\mathbb{E}_{\tau\sim\mathrm{Unif}(0,1)}\mathbb{E}_{\mathcal{T},X_1}\left[\mathbb{E}_{X_\tau^{\mathcal{T}}|X_1}\left[D_\theta u_\theta(Z)^\top\left(u_\theta(Z) - Y^{\mathcal{T}}\right)\right]\right] \\ &= 2\mathbb{E}_{\tau\sim\mathrm{Unif}(0,1)}\mathbb{E}_{\mathcal{T}}\left[\mathbb{E}_{X_\tau^{\mathcal{T}}}\left[D_\theta u_\theta(Z)^\top\left(u_\theta(Z) - u^\star(Z)\right)\right]\right] = \nabla_\theta\mathcal{L}_{\mathrm{HTFM}}(\theta). \end{aligned} \tag{C.17}$$

Moreover, $\|\nabla_\theta\mathcal{L}_{\mathrm{CHTFM}}(\theta)\| \le 2K_r\sqrt{\mathcal{L}_{\mathrm{CHTFM}}(\theta)}$*. If* $u_{\theta^\star} = u^\star$*, then* $\varepsilon_\star = 0$ *and* $\mathcal{L}_{\mathrm{CHTFM}}(\theta) \le C_\star + K_r^2\|\theta-\theta^\star\|^2$*.*

Remark C.9 (Scope)**.** *Theorem C.6 is distributional and proves the missing outer integrability. Corollary C.8 is parameterization-dependent: proximity to* $\theta^\star$ *is insufficient without the tail-compatible condition* $J_r \in L^2(P_Z)$*. No global finite-loss or convergence claim from arbitrary initialization is made.*

C.6.2 Supporting lemmas

Lemma C.10 (Choice of exponents)**.** *Let* $p \in (1,2)$ *and* $q > p/(p-1)$*. There exists* $\beta \in (2-p,1)$ *such that*

$$\beta > \frac{q(2-p)}{q-p}. \tag{C.18}$$

For $\kappa := \beta/(2-p) > 1$*,*

$$2 - \kappa p \le \beta, \qquad q + \kappa(2-q) \le \beta. \tag{C.19}$$

Proof. The condition on q is equivalent to $q(2-p)/(q-p) < 1$, and this lower bound exceeds $2-p$; hence β exists. Since $\beta = \kappa(2-p)$, the first inequality is equivalent to $\kappa \ge 1$, and the second to $q(2-p) \le \beta(q-p)$. □

Lemma C.11 (Threshold estimator through high-moment noise)**.** *Let $W \sim P_W$ and $N \sim P_N$ be independent, with*

$$\mathbb{E}_W \|W\|^p \le M_p, \qquad \mathbb{E}_N \|N\|^2 \le \lambda^2, \qquad \mathbb{E}_N \|N\|^q \le C_q \lambda^q. \tag{C.20}$$

Choose (β, κ) as in Lemma C.10 and define

$$\Gamma_\lambda(y) := \begin{cases} y, & 0 \le \lambda \le 1, \\ y\,\mathbf{1}_{\{\|y\| > \lambda^\kappa\}}, & \lambda > 1. \end{cases} \tag{C.21}$$

Then, for $C = C(p, q, C_q) < \infty$,

$$\mathbb{E}_{W,N} \|\Gamma_\lambda(W+N) - W\|^2 \le C(1 + M_p)\varphi_\beta(\lambda), \qquad \varphi_\beta(\lambda) := \lambda^2 \mathbf{1}_{\{\lambda \le 1\}} + \lambda^\beta \mathbf{1}_{\{\lambda > 1\}}. \tag{C.22}$$

Proof. For $\lambda \le 1$, the error is N, so the bound follows from $\mathbb{E}_N \|N\|^2 \le \lambda^2$. Let $\lambda > 1$, $R = \lambda^\kappa$, and $A = \{\|W + N\| > R\}$. Then

$$\mathbb{E}_{W,N} \|\Gamma_\lambda(W+N) - W\|^2 = \mathbb{E}_{W,N}[\|N\|^2 \mathbf{1}_A] + \mathbb{E}_{W,N}[\|W\|^2 \mathbf{1}_{A^c}]. \tag{C.23}$$

Since $A \subset \{\|W\| > R/2\} \cup \{\|N\| > R/2\}$, independence, Markov's inequality, and (C.20) give

$$\mathbb{E}_{W,N}[\|N\|^2 \mathbf{1}_A] \le 2^p M_p \lambda^2 R^{-p} + 2^{q-2} C_q \lambda^q R^{2-q}. \tag{C.24}$$

On $\{\|W\| \le 2R\}$, $\|W\|^2 \le (2R)^{2-p} \|W\|^p$. On $A^c \cap \{\|W\| > 2R\}$, $\|N\| > \|W\|/2 > R$, hence

$$\mathbb{E}_{W,N}[\|W\|^2 \mathbf{1}_{A^c}] \le 2^{2-p} M_p R^{2-p} + 4 C_q \lambda^q R^{2-q}. \tag{C.25}$$

Combining the last two displays and using $R = \lambda^\kappa$ with (C.19) yields (C.22). □

Lemma C.12 (Clock-aware finite-risk comparison field)**.** *Under Assumption C.5, there is a measurable field b such that*

$$\mathbb{E}_{\tau \sim \mathrm{Unif}(0,1)} \mathbb{E}_{\mathcal{T}, X_1} \left[\mathbb{E}_{X_\tau^{\mathcal{T}} | X_1} \left\| b(X_\tau^{\mathcal{T}}, \tau, \mathcal{T}) - Y^{\mathcal{T}} \right\|^2 \right] \le K(1 + M_p) \mathbb{E}_{\mathcal{T}} a(\mathcal{T}) < \infty. \tag{C.26}$$

Proof. For fixed $(t, \mathcal{T}) \in (0,1) \times D([0,1]; \mathbb{R}_+)$,

$$\frac{X_t^{\mathcal{T}}}{t} = X_1 + N_{t,\mathcal{T}}, \qquad N_{t,\mathcal{T}} := \frac{1-t}{t} X_0^{\mathcal{T}}, \qquad \lambda(t, \mathcal{T}) := a(\mathcal{T}) \frac{1-t}{t}. \tag{C.27}$$

Conditionally on $\mathcal{T}$, X_1 and $N_{t,\mathcal{T}}$ are independent, and (H3) gives

$$\mathbb{E}_{N_{t,\mathcal{T}} | \mathcal{T}} \|N_{t,\mathcal{T}}\|^2 \le \lambda(t, \mathcal{T})^2, \qquad \mathbb{E}_{N_{t,\mathcal{T}} | \mathcal{T}} \|N_{t,\mathcal{T}}\|^q \le C_q \lambda(t, \mathcal{T})^q. \tag{C.28}$$

Define, for $t \in (0,1)$,

$$b(x, t, \mathcal{T}) := \frac{\Gamma_{\lambda(t,\mathcal{T})}(x/t) - x}{1-t}. \tag{C.29}$$

Because $Y^{\mathcal{T}} = (X_1 - X_t^{\mathcal{T}})/(1-t)$, Lemma C.11, applied conditionally on $\mathcal{T}$, yields

$$\mathbb{E}_{X_1} \left[\mathbb{E}_{X_t^{\mathcal{T}} | X_1} \left\| b(X_t^{\mathcal{T}}, t, \mathcal{T}) - Y^{\mathcal{T}} \right\|^2 \right] \le \frac{C(1 + M_p)}{(1-t)^2} \varphi_\beta \left(a(\mathcal{T}) \frac{1-t}{t} \right). \tag{C.30}$$

For fixed $a > 0$, the change of variables $\lambda = a(1-t)/t$ gives

$$\int_0^1 \frac{\varphi_\beta(a(1-t)/t)}{(1-t)^2}\, dt = a \int_0^\infty \frac{\varphi_\beta(\lambda)}{\lambda^2}\, d\lambda = a \left(1 + \frac{1}{1-\beta} \right). \tag{C.31}$$

The case $a = 0$ is immediate. Taking $\mathbb{E}_{\tau \sim \mathrm{Unif}(0,1)} \mathbb{E}_{\mathcal{T}}$ in (C.30) proves the claim. □

Lemma C.13 (Projection from a finite-risk reference)**.** *Let $(Y, Z) \sim P_{Y,Z}$, assume $\mathbb{E}_Y \|Y\| < \infty$, and suppose a Z-measurable $b(Z)$ satisfies $\mathbb{E}_{Y,Z} \|Y - b(Z)\|^2 < \infty$. Define $u^\star(Z) := \mathbb{E}_{Y|Z}[Y]$. Then*

$$\mathbb{E}_{Y,Z} \|Y - u^\star(Z)\|^2 \le \mathbb{E}_{Y,Z} \|Y - b(Z)\|^2 < \infty. \tag{C.32}$$

Moreover, for every measurable $u(Z)$,

$$\mathbb{E}_{Y,Z} \|u(Z) - Y\|^2 < \infty \quad \Longleftrightarrow \quad \mathbb{E}_Z \|u(Z) - u^\star(Z)\|^2 < \infty, \tag{C.33}$$

and on this domain

$$\mathbb{E}_{Y,Z} \|u(Z) - Y\|^2 = \mathbb{E}_Z \|u(Z) - u^\star(Z)\|^2 + \mathbb{E}_{Y,Z} \|Y - u^\star(Z)\|^2. \tag{C.34}$$

Proof. Let $E_b := Y - b(Z) \in L^2(P_{Y,Z})$. Since b is Z-measurable,

$$Y - u^\star(Z) = E_b - \mathbb{E}_{Y|Z}[E_b], \tag{C.35}$$

so (C.32) follows from the L^2-contraction of conditional expectation. If $u - u^\star \in L^2(P_Z)$, the cross term in

$$u(Z) - Y = (u(Z) - u^\star(Z)) + (u^\star(Z) - Y) \tag{C.36}$$

vanishes because

$$\mathbb{E}_{Y,Z} \langle u(Z) - u^\star(Z), u^\star(Z) - Y \rangle = \mathbb{E}_Z \left\langle u(Z) - u^\star(Z), \mathbb{E}_{Y|Z}[u^\star(Z) - Y] \right\rangle = 0. \tag{C.37}$$

This proves (C.34); the converse in (C.33) follows by subtracting the already square-integrable residual. □

Lemma C.14 (Jacobian envelope implies local $L^2(P_Z)$-regularity)**.** *Under (P2), every $\theta \in B_r(\theta^\star)$ satisfies*

$$\|u_\theta - u_{\theta^\star}\|_{L^2(P_Z)} \le K_r \|\theta - \theta^\star\|. \tag{C.38}$$

Proof. For P_Z-a.e. z, the fundamental theorem of calculus gives

$$u_\theta(z) - u_{\theta^\star}(z) = \int_0^1 D_\theta u_{\theta^\star + s(\theta - \theta^\star)}(z)(\theta - \theta^\star)\, ds. \tag{C.39}$$

Hence $\|u_\theta(z) - u_{\theta^\star}(z)\| \le J_r(z) \|\theta - \theta^\star\|$; square and take $\mathbb{E}_{Z \sim P_Z}$. □

Lemma C.15 (Differentiation and gradient projection)**.** *Suppose (P2) holds and $\mathcal{L}_{\mathrm{CHTFM}}(\theta) < \infty$ throughout $B_r(\theta^\star)$. Then both objectives are differentiable at every interior point, their gradients are the two expressions in* (C.17)*, and they are equal.*

Proof. Fix an interior θ and direction v. For small s,

$$\frac{\|u_{\theta+sv}(Z) - u_\theta(Z)\|}{|s|} \le J_r(Z) \|v\|. \tag{C.40}$$

The conditional-loss difference quotient is dominated by

$$J_r(Z) \|v\| \left(2 \left\|u_\theta(Z) - Y^{\mathcal{T}}\right\| + |s| J_r(Z) \|v\|\right), \tag{C.41}$$

which is integrable under the joint training law by Cauchy–Schwarz. Dominated convergence gives the first gradient formula; the marginal formula is identical. Finally,

$$\begin{aligned} &\mathbb{E}_{\tau \sim \mathrm{Unif}(0,1)} \mathbb{E}_{\mathcal{T}, X_1} \left[\mathbb{E}_{X_\tau^{\mathcal{T}} | X_1} \left[D_\theta u_\theta(Z)^\top \left(Y^{\mathcal{T}} - u^\star(Z)\right)\right]\right] \\ &\quad = \mathbb{E}_{Z \sim P_Z} \left[D_\theta u_\theta(Z)^\top \mathbb{E}_{Y^\tau | Z}\left[Y^{\mathcal{T}} - u^\star(Z)\right]\right] = 0, \end{aligned} \tag{C.42}$$

where the product is integrable because both factors are in L^2. Thus the gradients coincide. □

C.6.3 Proofs of the main results

Proof of Theorem C.6. Since $p > 1$, $\mathbb{E}_{X_1} \|X_1\| < \infty$. Also,

$$\mathbb{E}_{\mathcal{T}} \mathbb{E}_{X_0^{\mathcal{T}} | \mathcal{T}} \left\|X_0^{\mathcal{T}}\right\| \le \mathbb{E}_{\mathcal{T}} \left(\mathbb{E}_{X_0^{\mathcal{T}} | \mathcal{T}} \left\|X_0^{\mathcal{T}}\right\|^2\right)^{1/2} \le \mathbb{E}_{\mathcal{T}} a(\mathcal{T}) < \infty. \tag{C.43}$$

Hence $Y^{\mathcal{T}} \in L^1$. Lemma C.12 constructs a full-clock measurable b with finite joint risk and the bound in (C.26). Lemma C.13, applied under the joint law of $(Y^{\mathcal{T}}, Z)$, then gives (C.12), (C.13), and (C.14). Finally, $\mathcal{L}_{\mathrm{HTFM}}(u) = \|u - u^\star\|^2_{L^2(P_Z)}$, so the domain is $u^\star + L^2(P_Z; \mathbb{R}^d)$. □

Proof of Corollary C.8. Optimality gives $\mathcal{L}_{\mathrm{CHTFM}}(\theta^\star) \le \mathcal{L}_{\mathrm{CHTFM}}(\bar{\theta}) < \infty$, so Theorem C.6 yields $\varepsilon_\star^2 = \mathcal{L}_{\mathrm{HTFM}}(\theta^\star) < \infty$. For $\theta \in B_r(\theta^\star)$, Lemma C.14 and the triangle inequality in $L^2(P_Z)$ give

$$\sqrt{\mathcal{L}_{\mathrm{HTFM}}(\theta)} = \|u_\theta - u^\star\|_{L^2(P_Z)} \le K_r \|\theta - \theta^\star\| + \varepsilon_\star. \tag{C.44}$$

Theorem C.6 then gives the conditional-loss bound in (C.16). Lemma C.15 gives gradient equality and finiteness. The stated gradient-norm bound follows from Cauchy–Schwarz and $\mathbb{E}_{Z \sim P_Z} \|D_\theta u_\theta(Z)\|^2_{\mathrm{op}} \le K_r^2$. □

C.6.4 Specialization to HTFM clocks

Suppose $X_0^{\mathcal{T}} \mid \mathcal{T} \sim \mathcal{N}(0, \Sigma(\mathcal{T}))$. Taking $a(\mathcal{T}) := \sqrt{\operatorname{tr} \Sigma(\mathcal{T})}$ gives

$$\mathbb{E}_{X_0^{\mathcal{T}}|\mathcal{T}} \left\|X_0^{\mathcal{T}}\right\|^2 = a(\mathcal{T})^2, \qquad \mathbb{E}_{X_0^{\mathcal{T}}|\mathcal{T}} \left\|X_0^{\mathcal{T}}\right\|^q \le \mathbb{E}_{G\sim\mathcal{N}(0,I_d)} \|G\|^q \; a(\mathcal{T})^q. \tag{C.45}$$

Thus (H3) reduces to $\mathbb{E}_{\mathcal{T}} \sqrt{\operatorname{tr} \Sigma(\mathcal{T})} < \infty$.

For the stable area clock, $\Sigma(\mathcal{T}) = 2A(T) I_d$, so $a(\mathcal{T}) = \sqrt{2dA(T)}$. Since $A(T)$ is positive stable with index $\alpha/2$,

$$\mathbb{E}_{\mathcal{T}} \sqrt{A(T)} < \infty \quad \Longleftrightarrow \quad \alpha > 1, \tag{C.46}$$

which includes $1 < \alpha < 2$. For the Student-t clock, $T_s = sV_\nu$, $V_\nu \sim \mathrm{InvGamma}(\nu/2, \nu/2)$, and $a(\mathcal{T}) = \sqrt{dV_\nu}$. The inverse-Gamma moment formula gives

$$\mathbb{E}_{V_\nu} \sqrt{V_\nu} < \infty \quad \Longleftrightarrow \quad \nu > 1, \tag{C.47}$$

which includes $1 < \nu \le 2$. If the data tail exponent is $\alpha_D \in (1, 2)$, (H1) holds for any $1 < p < \alpha_D$; hence the theorem also covers the infinite-variance synthetic targets.

Theorem C.6 remains valid when the network uses a feature $\eta(\mathcal{T})$, because such a predictor is still measurable with respect to the full-clock input Z. The local corollary additionally requires the feature-conditioned model class to contain a finite-risk predictor. This is natural when the feature reveals the conditional scale (e.g., $\log V_\nu$); it need not hold for an overly compressed or clock-blind class.

C.6.5 Finite-risk model classes and limitations

Proposition C.16 (A sufficient finite-competitor condition)**.** *Let b be the comparison field from Lemma C.12. If some $\bar\theta \in \Theta$ satisfies*

$$\mathbb{E}_{Z\sim P_Z} \|u_{\bar\theta}(Z) - b(Z)\|^2 < \infty, \tag{C.48}$$

then $\mathcal{L}_{\mathrm{CHTFM}}(\bar\theta) < \infty$.

Proof. Both $u_{\bar\theta}(Z) - b(Z)$ and $b(Z) - Y^{\mathcal{T}}$ are square-integrable under the joint training law, hence so is their sum. □

Remark C.17 (Why parameter proximity alone is insufficient)**.** *Let $X \sim P_X$ satisfy $\mathbb{E}_X|X| < \infty$ and $\mathbb{E}_X X^2 = \infty$, set $Y = X$, and take $u_\theta(X) = \theta X$. Although $\theta^\star = 1$ has zero risk, every $\theta \neq 1$ satisfies*

$$\mathbb{E}_X |u_\theta(X) - Y|^2 = (\theta - 1)^2 \mathbb{E}_X X^2 = \infty. \tag{C.49}$$

Here $D_\theta u_\theta(X) = X \notin L^2(P_X)$, so (P2) fails. The Jacobian envelope is therefore a substantive tail-compatibility condition, not ordinary pointwise smoothness.

The main theorem proves integrability only after averaging over $(\tau, \mathcal{T})$; it does not give a uniform fixed-time bound, and $t = 0$ may remain singular. The condition $p > 1$, together with the source assumptions, ensures the first moments needed for the ordinary conditional mean and enters the time-integrability estimates; existence of a conditional mean itself only requires integrability. The local corollary does not justify global training from arbitrary initialization or prove that an unrestricted neural architecture automatically satisfies (P1)–(P2).

D Additional Information on Log-Signature

This appendix fixes the signature-feature convention used in Section 3.2. Here $\ell_m(T)$ denotes the raw-path logsignature; the implemented feature $\psi(T)$ includes the path standardization described in that section for stable clocks, and equals $\log V_\nu$ for affine Student-t clocks. For completeness we keep the generic notation η_T throughout this appendix and discuss alternative choices (raw signatures, lead–lag transforms, or concatenations with simple summaries). The feature η_T has two roles: it conditions the learned vector field $u_\theta(X_t^T, t, \eta_T)$, and, when a clock-dependent trajectory is used, it may also parameterize the interpolation schedules α_t^T and β_t^T. Throughout this appendix,

the feature is a summary of the full realized clock path $T_{[0,1]}$, sampled before drawing the clock-conditioned source and before solving the generation ODE.

Let $\mathbf{z} : [0,1] \to \mathbb{R}^q$ be a continuous path of bounded variation. Its signature is the tensor series

$$S(\mathbf{z}) = \left(1, \int_{0<u_1<1} \mathrm{d}\mathbf{z}_{u_1}, \int_{0<u_1<u_2<1} \mathrm{d}\mathbf{z}_{u_1} \otimes \mathrm{d}\mathbf{z}_{u_2}, \dots\right),$$

and its truncation to order m is denoted by $S^{(\le m)}(\mathbf{z})$. The truncated logsignature is

$$\mathrm{LogSig}^{(\le m)}(\mathbf{z}) := \log_{\otimes}\big(S^{(\le m)}(\mathbf{z})\big),$$

where $\log_{\otimes}$ is computed in the truncated tensor algebra. Unlike the raw signature, the logsignature lies in the truncated free Lie algebra and removes many algebraic redundancies among tensor coordinates.

In our application, $q = 2$ and the path is the time-augmented clock $\mathbf{T}_s = (s, T_s)$. Since the random clocks considered in the paper may be càdlàg and may have jumps, we do not rely on a canonical signature of the jump path itself. Instead, for a discretization $0 = t_0 < t_1 < \cdots < t_N = 1$, we form the sampled time-augmented path

$$\big(\mathbf{T}_{t_0}, \mathbf{T}_{t_1}, \dots, \mathbf{T}_{t_N}\big) = \big((t_0, T_{t_0}), (t_1, T_{t_1}), \dots, (t_N, T_{t_N})\big)$$

and compute signatures after linearly interpolating these sampled points. With this convention,

$$\phi_m(T) := S^{(\le m)}(\mathbf{T}^{\mathrm{lin}})$$

is the truncated signature feature used in (3.6), where $\mathbf{T}^{\mathrm{lin}}$ denotes the piecewise linear interpolation of the sampled time-augmented clock. A logsignature version of the same feature is

$$\ell_m(T) := \mathrm{LogSig}^{(\le m)}(\mathbf{T}^{\mathrm{lin}}).$$

Thus the model feature η_T may be chosen as $\phi_m(T)$, as $\ell_m(T)$, or as a concatenation of either representation with simple clock summaries such as T_1 or $\int_0^1 T_s \mathrm{d}s$. When invoking the signature-linear approximation result in Proposition D.1, one should take the raw signature coordinates $\phi_m(T)$, because the proposition is stated for linear functionals of signatures. Logsignature coordinates give a compact conditioning representation; in the stable-clock implementation these are computed from the standardized path to obtain $\psi(T)$, rather than using the raw-path $\ell_m(T)$ directly; using logsignature coordinates in the learned schedules should therefore be understood as a practical finite-dimensional parameterization, not as the literal raw-signature universality statement.

Under the piecewise linear convention, a jump in the sampled clock is represented by a steep linear segment between two consecutive observation times. Other conventions, such as inserting explicit vertical jump segments or using a lead–lag transform, would define different finite-dimensional features. They do not change the flow-matching identities in the main text: the source is Gaussian conditional on $T_{[0,1]}$, and the affine interpolation is Gaussian conditional on both $T_{[0,1]}$ and the data endpoint, as in (3.9), and the endpoint-conditioned target remains the affine expression (3.10). The feature η_T only summarizes the realized clock environment for the network and, when used, for the schedule parameterization.

The first logsignature level is the increment of the time-augmented path, so it records $(1, T_1 - T_0)$, which is $(1, T_1)$ under the standing assumption $T_0 = 0$. Higher levels encode how clock mass is allocated over physical time. For example, in two dimensions the second Lie level corresponds to the signed area between the physical-time coordinate and the clock coordinate; for monotone clocks this separates paths with the same terminal value T_1 but different temporal allocations of their increments. This is why low-order signature information can represent path-dependent functionals such as the area-clock quantity $\int_0^1 T_s \mathrm{d}s$ used in Section 3.3.

The dimension reduction from signature to logsignature can be substantial. For a q-dimensional path, the raw truncated signature has $1 + \sum_{k=1}^{m} q^k$ coordinates including the constant coordinate, whereas the nonconstant truncated logsignature has dimension

$$\sum_{k=1}^{m} \ell_q(k), \qquad \ell_q(k) := \frac{1}{k} \sum_{r \mid k} \mu(r) q^{k/r},$$

where μ is the Möbius function. Thus, for the time-augmented clock ($q = 2$) and truncation order $m = 3$, the raw nonconstant signature has dimension $2 + 4 + 8 = 14$, while the logsignature has dimension $2 + 1 + 2 = 5$. This compactness is the main reason that the experiments use truncated logsignature features for η_T when conditioning the neural velocity field.

For clarity, the $m = 3$ calculation is an illustrative dimension count; the main stable-clock experiments use $m = 1$, and the order ablation uses $m \in \{0, 1, 2\}$, with $m = 0$ withholding the clock feature. The degree-wise Lie dimensions at $q = 2$ are $2, 1, 2$: two increments, one signed-area coordinate, and two third-order Lie-bracket coordinates. For the unnormalized piecewise linear path (s, T_s^{lin}), write $A_N(T) := \int_0^1 T_s^{\mathrm{lin}}\,\mathrm{d}s$. Its degree-two coordinate is

$$a_{12}(T) = \tfrac{1}{2}\left(\int_0^1 s\,\mathrm{d}T_s^{\mathrm{lin}} - \int_0^1 T_s^{\mathrm{lin}}\,\mathrm{d}s\right) = \tfrac{1}{2}T_1 - A_N(T).$$

Thus $(T_1, a_{12}(T))$ determines the area of the interpolated path. At a finite grid resolution, this area need not equal $A(T)$ for the original jump clock. For an affine clock $T_s = sV_\nu$, the full path is determined by V_ν, and the implementation directly conditions on $\log V_\nu$.

D.1 Signature-linear schedules

This subsection gives the schedule details that are summarized in Section 3.3. The compactness assumption in Proposition D.1 below is the usual setting for uniform approximation. For heavy-tailed clocks, which need not live on a compact path set globally, the proposition applies on compact truncations or high-probability compact subsets of the discretized clock space. Standardizing the features does not itself make their support compact. The time augmentation is important: it records the order in which clock mass is accumulated, so path functionals such as $\int_0^1 T_s\mathrm{d}s$ can be represented by low-order signature information.

Proposition D.1 (Signature-linear approximation of clock functionals)**.** *Fix the sampling grid $0 = t_0 < \cdots < t_N = 1$ and equip its time-augmented, piecewise linear clock paths, rooted at $(0, 0)$, with the Euclidean topology of the vertex values $(T_{t_0}, \ldots, T_{t_N})$. Let $\mathcal{K}$ be a compact subset on which the full signature separates points. Let $F : [0, 1] \times \mathcal{K} \to \mathbb{R}$ be continuous. Then for every $\varepsilon > 0$, there exist a truncation level m and a continuous coefficient curve $c : [0, 1] \to \mathbb{R}^{D_m}$ such that*

$$\sup_{t\in[0,1]}\sup_{T\in\mathcal{K}} |F(t, T) - \langle c(t), \phi_m(T)\rangle| < \varepsilon. \tag{D.1}$$

In particular, clock-dependent schedule functionals such as $T \mapsto \alpha_t(T)$ and $T \mapsto \beta_t(T)$ can be represented, up to arbitrary uniform accuracy on compact clock families, by linear functionals of a sufficiently high-order truncated signature:

$$\alpha_t(T) \approx \langle a(t), \phi_m(T)\rangle, \qquad \beta_t(T) \approx \langle b(t), \phi_m(T)\rangle. \tag{D.2}$$

Proof. Each signature coordinate is continuous in the vertex values on the fixed grid. Consider finite sums of functions $c_w(t)S_w(T)$, where w indexes a signature coordinate and c_w is continuous on $[0, 1]$. The shuffle identity makes these sums an algebra, and the level-zero coordinate supplies constants and arbitrary continuous functions of t. This algebra separates points of $[0, 1] \times \mathcal{K}$: coefficient functions separate distinct times, and signatures separate distinct clock paths. Stone–Weierstrass therefore gives uniform density in $C([0, 1] \times \mathcal{K})$. Each approximant uses finitely many words; taking m to be their maximum length yields the stated coefficient curve $c(t)$. □

The truncation level may depend on F and ε. The proposition does not claim universality at a fixed level m, nor does uniform approximation alone control time derivatives of the schedules.

The proposition motivates representing clock-dependent interpolation coefficients by signature-linear functions. A direct form is

$$\alpha_t^{T,m} = \langle a(t), \phi_m(T)\rangle, \qquad \beta_t^{T,m} = \langle b(t), \phi_m(T)\rangle, \tag{D.3}$$

where $a(t), b(t) \in \mathbb{R}^{D_m}$ are coefficient curves. If these curves are absolutely continuous in t, then for a fixed realized clock path the physical-time derivatives are explicit:

$$\frac{\mathrm{d}\alpha_t^{T,m}}{\mathrm{d}t} = \left\langle \frac{\mathrm{d}a(t)}{\mathrm{d}t}, \phi_m(T)\right\rangle, \qquad \frac{\mathrm{d}\beta_t^{T,m}}{\mathrm{d}t} = \left\langle \frac{\mathrm{d}b(t)}{\mathrm{d}t}, \phi_m(T)\right\rangle. \tag{D.4}$$

Thus the flow-matching target differentiates the scalar-time coefficient curves, not the clock path itself.

The direct form (D.3) does not automatically enforce the endpoint constraints $\alpha_0^T = 0,\ \beta_0^T = 1,\ \alpha_1^T = 1,\ \beta_1^T = 0,\ \beta_t^T > 0$ for $t \in [0,1)$. A boundary-corrected near-linear schedule is obtained as follows. Let $\lambda : [0,1] \to \mathbb{R}$ be absolutely continuous with $\lambda_0 = 0$, and set

$$A_t(T) := \langle a(t), \eta_T \rangle, \qquad B_t(T) := \langle b(t), \eta_T \rangle. \tag{D.5}$$

Then define

$$\alpha_t^T = t + \lambda_t(1-t)A_t(T), \qquad \beta_t^T = 1 - t + \lambda_t(1-t)B_t(T) = (1-t)\big(1 + \lambda_t B_t(T)\big). \tag{D.6}$$

The endpoint constraints are automatic:

$$\alpha_0^T = 0, \quad \alpha_1^T = 1, \quad \beta_0^T = 1, \quad \beta_1^T = 0. \tag{D.7}$$

The path is admissible whenever

$$1 + \lambda_t B_t(T) > 0, \qquad t \in [0,1). \tag{D.8}$$

Writing

$$\frac{\mathrm{d}A_t(T)}{\mathrm{d}t} = \left\langle \frac{\mathrm{d}a(t)}{\mathrm{d}t}, \eta_T \right\rangle, \qquad \frac{\mathrm{d}B_t(T)}{\mathrm{d}t} = \left\langle \frac{\mathrm{d}b(t)}{\mathrm{d}t}, \eta_T \right\rangle,$$

the required derivatives are

$$\frac{\mathrm{d}\alpha_t^T}{\mathrm{d}t} = 1 + \left(\frac{\mathrm{d}\lambda_t}{\mathrm{d}t}(1-t) - \lambda_t \right) A_t(T) + \lambda_t(1-t)\frac{\mathrm{d}A_t(T)}{\mathrm{d}t}, \tag{D.9}$$

$$\frac{\mathrm{d}\beta_t^T}{\mathrm{d}t} = -\big(1 + \lambda_t B_t(T)\big) + (1-t)\left(\frac{\mathrm{d}\lambda_t}{\mathrm{d}t} B_t(T) + \lambda_t \frac{\mathrm{d}B_t(T)}{\mathrm{d}t} \right), \tag{D.10}$$

$$\frac{\frac{\mathrm{d}\beta_t^T}{\mathrm{d}t}}{\beta_t^T} = -\frac{1}{1-t} + \frac{\frac{\mathrm{d}\lambda_t}{\mathrm{d}t} B_t(T) + \lambda_t \frac{\mathrm{d}B_t(T)}{\mathrm{d}t}}{1 + \lambda_t B_t(T)}. \tag{D.11}$$

At a sampled interpolation point, the apparent division by β_t^T cancels algebraically:

$$v_t^T(X_t^T \mid X_1^T) = \frac{\mathrm{d}\alpha_t^T}{\mathrm{d}t} X_1^T + \frac{\mathrm{d}\beta_t^T}{\mathrm{d}t} X_0^T.$$

This cancellation does not itself imply time-integrability near $t = 1$. Finite quadratic risk for a general schedule requires additional bounds on its derivatives and the endpoint moments. The outer-clock theorem in Appendix C.6 establishes the result for the straight-line schedule; merely excluding the single time $t = 1$ from sampling would not establish it for other schedules.

A symmetric special case uses a single clock-dependent time warp. Let

$$C_t(T) := \langle c(t), \eta_T \rangle, \qquad \gamma_t^T := t + \lambda_t(1-t)C_t(T), \tag{D.12}$$

and set

$$\alpha_t^T = \gamma_t^T, \qquad \beta_t^T = 1 - \gamma_t^T = (1-t)\big(1 - \lambda_t C_t(T)\big). \tag{D.13}$$

This preserves the straight-line form $X_t^T = \gamma_t^T X_1^T + (1-\gamma_t^T)X_0^T$, while allowing the realized clock path to accelerate or decelerate the interpolation. The endpoint-conditioned target simplifies to

$$v_t^T(x \mid x_1) = \frac{\frac{\mathrm{d}\gamma_t^T}{\mathrm{d}t}}{1 - \gamma_t^T}(x_1 - x), \qquad t \in [0,1). \tag{D.14}$$

D.2 VE-type signature-linear schedule

For the clock-aware trajectory ablation, we use a variance-exploding (VE)-type affine path written in the signature-linear form (D.3). This is a deterministic affine flow path inspired by VE noise schedules, not an additional stochastic SDE. Since all clocks in this experiment satisfy $T_1 > 0$, use the fixed-grid interpolant to define $\bar{T}_s := T_s^{\mathrm{lin}}/T_1$ and the normalized time-augmented path $\bar{\mathbf{T}}_s = (s, \bar{T}_s)$. Let $\phi_m(\bar{T})$ denote the raw truncated signature of $\bar{\mathbf{T}}$. For $m \geq 2$, let $e_{\mathtt{ct}}$ and $e_{\mathtt{tc}}$ be

the unit vectors selecting the level-two signature coordinates with words $(\mathtt{c},\mathtt{t})$ and $(\mathtt{t},\mathtt{c})$, where $\mathtt{t}$ denotes physical time and $\mathtt{c}$ denotes the normalized clock coordinate. Then

$$\left\langle e_{\mathtt{ct}}, \phi_m(\bar{T})\right\rangle = \int_{0<u<v<1} \mathrm{d}\bar{T}_u\,\mathrm{d}v = \int_0^1 \bar{T}_u\,\mathrm{d}u, \tag{D.15}$$

$$\left\langle e_{\mathtt{tc}}, \phi_m(\bar{T})\right\rangle = \int_{0<u<v<1} \mathrm{d}u\,\mathrm{d}\bar{T}_v = 1 - \int_0^1 \bar{T}_v\,\mathrm{d}v. \tag{D.16}$$

Thus the signed clock-shape score

$$R_m(T) := \left\langle r_m, \phi_m(\bar{T})\right\rangle, \qquad r_m := e_{\mathtt{ct}} - e_{\mathtt{tc}}, \tag{D.17}$$

satisfies

$$R_m(T) = 2\int_0^1 \bar{T}_s\,\mathrm{d}s - 1. \tag{D.18}$$

It is zero for a linear normalized clock, positive when clock mass accumulates early, negative when it accumulates late, and bounded by $[-1,1]$ for nondecreasing normalized clocks.

We use the normalized linear VE decay $\sigma_{\mathrm{VE}}(t) = 1-t$ and introduce independent signature-linear clock-shape corrections:

$$\alpha_t^{T,m} = t + \lambda_\alpha t(1-t)R_m(T), \qquad \beta_t^{T,m} = (1-t)\big(1+\lambda_\beta t R_m(T)\big). \tag{D.19}$$

Equivalently, with $e_\emptyset$ denoting the constant signature coordinate,

$$a_{\mathrm{VE}}(t) = te_\emptyset + \lambda_\alpha t(1-t)r_m, \tag{D.20}$$

$$b_{\mathrm{VE}}(t) = (1-t)e_\emptyset + \lambda_\beta t(1-t)r_m, \tag{D.21}$$

so that

$$\alpha_t^{T,m} = \left\langle a_{\mathrm{VE}}(t), \phi_m(\bar{T})\right\rangle, \qquad \beta_t^{T,m} = \left\langle b_{\mathrm{VE}}(t), \phi_m(\bar{T})\right\rangle. \tag{D.22}$$

In the experiments we use $\lambda_\alpha = \lambda_\beta = 1/4$. More generally, since $|R_m(T)| \le 1$, any $|\lambda_\beta| < 1$ gives $\beta_t^{T,m} \ge (1-t)(1-|\lambda_\beta|) > 0$ for $t < 1$, so the affine path is admissible.

Writing $R_m = R_m(T)$, the derivatives are

$$\frac{\mathrm{d}\alpha_t^{T,m}}{\mathrm{d}t} = 1 + \lambda_\alpha(1-2t)R_m, \qquad \frac{\mathrm{d}\beta_t^{T,m}}{\mathrm{d}t} = -1 + \lambda_\beta(1-2t)R_m, \tag{D.23}$$

and hence

$$L_t^{T,m} := \frac{\frac{\mathrm{d}\beta_t^{T,m}}{\mathrm{d}t}}{\beta_t^{T,m}} = \frac{-1+\lambda_\beta(1-2t)R_m}{(1-t)\big(1+\lambda_\beta t R_m\big)}. \tag{D.24}$$

Substituting these quantities into (3.10) gives

$$\begin{aligned} v_t^{T,m}(x \mid x_1) &= L_t^{T,m}x + \left(\frac{\mathrm{d}\alpha_t^{T,m}}{\mathrm{d}t} - L_t^{T,m}\alpha_t^{T,m}\right)x_1 \\ &= L_t^{T,m}x + \Big[1 + \lambda_\alpha(1-2t)R_m - L_t^{T,m}\big(t + \lambda_\alpha t(1-t)R_m\big)\Big]\,x_1, \qquad t \in [0,1). \end{aligned} \tag{D.25}$$

When $\lambda_\alpha = \lambda_\beta = 0$, this reduces to the linear VE path $\alpha_t = t$, $\beta_t = 1-t$. The one-warp path is recovered only in the special symmetric case $\lambda_\beta = -\lambda_\alpha$; the ablation above uses the more general affine form $X_t^T = \alpha_t^{T,m}X_1^T + \beta_t^{T,m}X_0^T$.

E Experimental Setup

This section documents the full configuration shared across all reported HTFM and baseline runs. Unless explicitly noted, any HTFM clock variant (Gaussian, α-stable, Student-t) uses the same backbone and optimizer for a given dataset; the only differences are the random-clock law and the optional path-feature conditioning. All experiments are implemented in PyTorch.

E.1 Dataset

We use three benchmarks at increasing scale.

2D imbalanced α-stable mixture. A 9-component isotropic α-stable mixture in $\mathbb{R}^2$. Components sit on a uniformly spaced 3×3 grid with spacing 2.0, each component is a 2-D isotropic α-stable variable with scale 0.1, and the mixture weights are $\{0.18, 0.08, 0.14, 0.06, 0.09, 0.10, 0.16, 0.07, 0.12\}$ (range 0.06–0.18). We sweep the per-component stability index $\alpha_D \in \{1.5, 1.8\}$ and draw 32,000 training samples per setting; an additional 24,000-sample reference set is held out for evaluation.

CIFAR10-LT. The long-tailed CIFAR-10 split [6] of the original CIFAR-10 train set [29]. With imbalance ratio $\rho = 0.01$, head count 5000, and 10 classes, the per-class counts follow a geometric schedule and total $\approx 12{,}406$ training images at 32×32. A deterministic subset is precomputed and reused across all runs to keep the baselines identical at the dataset level. Random horizontal flips are the only augmentation.

HRRR VIL. Vertically Integrated Liquid (VIL) hourly analyses from the High-Resolution Rapid Refresh weather system [5], cropped to a 128×128 window over the Central US following the protocol of [41]. Train/test split mirrors theirs: years $\{2019, 2020\}$ for training ($\approx$ 17,400 valid hourly fields after dropping NaN slots) and year $\{2021\}$ for test ($\approx$ 8,700). Per-channel z-score statistics are computed on the training years and applied to both splits.

E.2 Baselines

We compare HTFM against four heavy-tailed generative baselines, each run under the same backbone and training budget as HTFM (see Sections E.4 and E.5):

- **LIM** [50, 44]: Lévy-Itô stochastic differential equation with α-stable driving noise; trained with a fractional-score regression loss and sampled with the model's α-stable reverse SDE solver at the evaluation budgets reported in each table or figure.
- **DLPM** [47]: discrete-time α-stable analogue of DDPM; uses 4000 training-time reverse steps with the Gaussian scale-mixture augmentation, sampling via its native stochastic predictor.
- **DLIM** [47]: deterministic α-stable sampler sharing the DLPM predictor backbone; we set η=0 for deterministic generation.
- **t-Flow** [41]: heavy-tailed flow with a Student-t source. We evaluate the finite-variance setting $p_\nu = 3$; this experimental choice does not imply that the source distribution or the t-Flow formulation requires $p_\nu > 2$.

E.3 Evaluation

We evaluate the 2D mixture with a precision–recall score, CIFAR10-LT with FID, and HRRR VIL with tail statistics.

2D mixture. We compute the precision–recall f_1^{pr} score following [47]. Let $\mathcal{X}_{\mathrm{ref}} = \{x_i\}_{i=1}^N$ and $\mathcal{X}_{\mathrm{gen}} = \{y_j\}_{j=1}^N$ be the reference and generated sample sets, each containing $N = 24{,}000$ points. For a set $\mathcal{X}$ and a point z, write $r_k(z;\mathcal{X})$ for the distance from z to its k-th nearest neighbor in $\mathcal{X}$, and define the k-NN manifold

$$\mathcal{M}_k(\mathcal{X}) \;=\; \bigcup_{z\in\mathcal{X}} B\big(z,\, r_k(z;\mathcal{X})\big),$$

Each sample thus defines an ℓ_2-ball whose radius is its k-th nearest-neighbor distance. Precision measures the fraction of generated samples within the reference manifold, and recall measures the fraction of reference samples within the generated manifold:

$$\mathrm{P} \;=\; \frac{1}{|\mathcal{X}_{\mathrm{gen}}|}\sum_{y\in\mathcal{X}_{\mathrm{gen}}} \mathbf{1}\big[y \in \mathcal{M}_k(\mathcal{X}_{\mathrm{ref}})\big], \qquad \mathrm{R} \;=\; \frac{1}{|\mathcal{X}_{\mathrm{ref}}|}\sum_{x\in\mathcal{X}_{\mathrm{ref}}} \mathbf{1}\big[x \in \mathcal{M}_k(\mathcal{X}_{\mathrm{gen}})\big],$$

and the score is their harmonic mean

$$f_1^{\mathrm{pr}} \;=\; \frac{2\,\mathrm{P}\,\mathrm{R}}{\mathrm{P}+\mathrm{R}}\,.$$

We use $k = 10$ throughout; higher f_1^{pr} is better.

CIFAR10-LT. We compute Fréchet Inception Distance (FID) [18] using the 2048-dimensional features from the final average-pooling layer of a pre-trained Inception-V3 network. We extract these features for each real and generated 32×32 image. Let $(\mu_{\mathrm{data}}, \Sigma_{\mathrm{data}})$ and $(\mu_{\mathrm{gen}}, \Sigma_{\mathrm{gen}})$ be the empirical mean and covariance of the two feature sets. FID is the squared 2-Wasserstein distance between the fitted Gaussians:

$$\mathrm{FID} \;=\; \|\mu_{\mathrm{data}} - \mu_{\mathrm{gen}}\|_2^2 \;+\; \mathrm{Tr}\big(\Sigma_{\mathrm{data}} + \Sigma_{\mathrm{gen}} - 2(\Sigma_{\mathrm{data}}\Sigma_{\mathrm{gen}})^{1/2}\big). \tag{E.1}$$

The reference set contains all 12,406 images in the long-tail training split. At each listed NFE budget, we generate 12,406 images using the EMA model. Lower FID is better.

Variation across training seeds. Table 6 reports the mean ± standard deviation of CIFAR10-LT FID across five independent training seeds at NFE 20, 100, and 500. The standard deviations summarize variation across training runs. This evaluation covers a subset of the settings in Table 2. The $p_\alpha = 1.9$, HTFM-t $p_\nu = 1.7$, and t-Flow results in the main table are reported as means.

Table 6: CIFAR10-LT FID, mean ± standard deviation over five training seeds. Gaussian DDPM and DDIM correspond to the Gaussian limits of DLPM and DLIM, respectively.

Method (tail setting)	NFE= 20	NFE= 100	NFE= 500
LIM ($p_\alpha = 1.6$)	135.9 ± 2.9	34.70 ± 0.47	20.22 ± 0.25
LIM ($p_\alpha = 1.7$)	130.2 ± 2.4	34.11 ± 0.51	19.18 ± 0.14
LIM ($p_\alpha = 1.8$)	126.8 ± 3.0	32.40 ± 0.66	18.23 ± 0.16
DLPM ($p_\alpha = 1.6$)	29.16 ± 0.37	18.18 ± 0.18	15.26 ± 0.14
DLPM ($p_\alpha = 1.7$)	28.55 ± 0.32	18.23 ± 0.23	14.91 ± 0.13
DLPM ($p_\alpha = 1.8$)	32.49 ± 0.29	18.29 ± 0.20	15.35 ± 0.17
DDPM	35.49 ± 0.48	22.28 ± 0.27	19.47 ± 0.24
DLIM ($p_\alpha = 1.6$)	21.30 ± 0.30	19.69 ± 0.16	17.22 ± 0.16
DLIM ($p_\alpha = 1.7$)	23.90 ± 0.33	20.38 ± 0.11	17.73 ± 0.18
DLIM ($p_\alpha = 1.8$)	22.75 ± 0.41	18.46 ± 0.19	17.90 ± 0.14
DDIM	25.45 ± 0.39	21.63 ± 0.17	19.92 ± 0.18
HTFM-α ($p_\alpha = 1.6$)	17.52 ± 0.34	16.19 ± 0.098	15.27 ± 0.11
HTFM-α ($p_\alpha = 1.7$)	17.13 ± 0.30	16.10 ± 0.13	15.04 ± 0.12
HTFM-α ($p_\alpha = 1.8$)	16.91 ± 0.26	15.89 ± 0.19	14.99 ± 0.19
HTFM-t ($p_\nu = 2.0$)	20.48 ± 0.24	16.76 ± 0.12	16.51 ± 0.097
HTFM-t ($p_\nu = 3.0$)	19.35 ± 0.21	16.14 ± 0.10	15.38 ± 0.13
HTFM-G	22.67 ± 0.36	19.29 ± 0.25	19.04 ± 0.20

HRRR VIL. Following [41], we evaluate the intensity distribution with kurtosis ratio, skewness ratio, and tail KS. We generate 20,000 fields from each model and use approximately 8,760 test fields as the reference. We flatten each set of fields into one-dimensional intensity samples and compute the following statistics. Lower values are better for all three.

Kurtosis Ratio (KR). Let k_{data} and k_{sim} denote the empirical kurtosis, or normalized fourth central moment, of the reference and generated intensities. We measure their relative discrepancy as

$$\mathrm{KR} \;=\; \left|1 - \tfrac{k_{\mathrm{sim}}}{k_{\mathrm{data}}}\right|. \tag{E.2}$$

Skewness Ratio (SR). Skewness is the normalized third central moment and measures distributional asymmetry. With s_{data} and s_{sim} denoting the empirical skewness of the reference and generated intensities, we compute

$$\mathrm{SR} \;=\; \left|1 - \tfrac{s_{\mathrm{sim}}}{s_{\mathrm{data}}}\right|. \tag{E.3}$$

Tail Kolmogorov–Smirnov (KS). The two-sample KS statistic [38] is the maximum absolute difference between two empirical CDFs. We compute it on generated and reference samples above the 99.9-th percentile for the right tail and below the 0.1-th percentile for the left tail. For two-sided distributions, we average the two tail statistics. For VIL, we evaluate the right tail.

E.4 Network Architecture

2D mixture. A small MLP with 4 residual blocks of width 64, SiLU activations, GroupNorm, and a learnable time embedding of size 32. The same backbone is shared across HTFM and the FM baselines.

CIFAR10-LT. The DDPM-style U-Net of [20]: 128 base channels, channel multipliers $(1, 2, 2, 2)$, two residual blocks per scale, 4 attention heads, self-attention at resolution 16, dropout 0. The same base U-Net is used by HTFM, LIM, DLPM, DLIM, and t-Flow. HTFM additionally uses the path-feature conditioning module described below, so the shared backbone does not imply identical total parameter counts.

HRRR VIL. The NCSN++ backbone of [48] on 128×128 single-channel inputs, with base channels 32, channel multipliers $(1, 2, 2, 4, 4)$, four residual blocks per scale, 4 attention heads, self-attention at resolution 16, FIR $(1, 3, 3, 1)$ up/down-sampling, swish nonlinearities, GroupNorm, and BigGAN-style residual blocks. Sigma range $[0.01, 50]$ on 1000 noise scales when used with score-based baselines.

Path-feature conditioning (HTFM-specific). HTFM's velocity network conditions on a clock-derived *path feature* ϕ that summarizes the realized clock trajectory of each sample. The feature source depends on the random clock:

- **Student-t clock** $T_t{=}tV_\nu$, $V_\nu \sim \mathrm{InvGamma}(\nu/2, \nu/2)$. The path is fully determined by the single scalar V_ν, so we take $\phi{=}\log V_\nu \in \mathbb{R}$ ($d_\phi{=}1$); no signature is needed.
- **Stable subordinator clock** (index $\alpha/2$). We discretize the time-augmented path $\{(t_i, T_{t_i})\}_{i=0}^{N}$ on a fixed $N{=}200$ grid, standardize each of the two channels along time, and take its truncated logsignature of order m (default $m{=}1$, i.e. $\phi \in \mathbb{R}^2$ collecting the increments of the two channels). The construction matches Section 3.2.
- **Deterministic clock** $T_t{=}t$ (HTFM-G, $p{=}\infty$). No path feature; the conditioning module is omitted entirely and the network reduces to a standard time-conditioned velocity field.

The raw feature $\phi \in \mathbb{R}^{d_\phi}$ is mapped through a shared two-layer MLP $\phi_{\mathrm{path}} \colon \mathbb{R}^{d_\phi} \to \mathbb{R}^{128}$ with SiLU activations, dropout 0, and an internal hidden width of 128. The 128-dimensional path embedding is injected into the velocity network at *every* residual block by concatenation with the time embedding. In the original block the time embedding $t_{\mathrm{emb}} \in \mathbb{R}^{d_t}$ is mapped to a per-channel scale/shift via a linear $W_t \in \mathbb{R}^{C_{\mathrm{out}}\times d_t}$; with path conditioning this is replaced by a wider linear $W_{t,\phi} \in \mathbb{R}^{C_{\mathrm{out}}\times(d_t+128)}$ acting on the concatenation $[t_{\mathrm{emb}} \,\|\, \phi_{\mathrm{path}}]$. The same concatenation-at-every-block scheme is used identically by the DDPM U-Net (CIFAR10-LT), the NCSN++ U-Net (HRRR VIL), and the residual-MLP backbone (2D mixture). The LIM, DLPM, DLIM, and t-Flow backbones have no analogous trajectory-summary input and receive only the time embedding.

For completeness, our implementation also supports a zero-initialized FiLM modulation interface that produces a per-channel $(\mathrm{scale}, \mathrm{shift})$ from ϕ_{path} and applies it at the middle block. All HTFM rows reported in the paper use only the concatenation interface; FiLM is left disabled.

E.5 Training

All runs use AdamW with default β's. Per-dataset hyperparameters are listed below.

2D mixture. Learning rate 5×10^{-3}, batch size 1024, EMA decay 0.99, total number of epochs 100.

CIFAR10-LT. Learning rate 2×10^{-4} with 500 warmup steps and a step-LR schedule (decay factor 0.99 every 1000 steps); batch size 100; EMA decay 0.9999 used for evaluation; gradient clipping at ℓ_2 norm 5.0. Models are trained for 2000 epochs.

HRRR VIL. Learning rate 10^{-4} with 1000 warmup steps and no LR decay; weight decay 0; batch size 128; EMA decay 0.9999; gradient clipping at 5.0. Models are trained for 1500 epochs.

HTFM-specific knobs. For HTFM the random clock is one of: deterministic $T_t = t$ (Gaussian limit, HTFM-G), an $(\alpha/2)$-stable subordinator with source-tail exponent $p_\alpha \in \{1.6, 1.7, 1.8, 1.9\}$ on CIFAR10-LT and $p_\alpha \in \{1.6, 1.7, 1.8\}$ on HRRR, or a Student-t random-slope clock $T_t = tV_\nu$, $V_\nu \sim \mathrm{InvGamma}(\nu/2, \nu/2)$, with decay degree $p_\nu \in \{1.7, 2.0, 3.0\}$ on CIFAR10-LT and HRRR (extended to $p_\nu = 1.5$ on the 2D mixture to include an infinite-variance source). The training loss is the clock-conditioned L^2 flow-matching objective with velocity prediction.

Objective and gradient stabilization. We use the untruncated minibatch estimator of $\mathcal{L}^{\mathrm{sig}}_{\mathrm{CHTFM}}$, with neither clock truncation nor per-sample loss truncation. Gradient clipping is applied after minibatch aggregation: for the aggregated gradient g, the update uses $g \min\{1, 5/\|g\|_2\}$, with the zero gradient left unchanged. This bounds its norm while preserving its direction. Almost-sure finiteness of each minibatch loss does not by itself establish finite population risk at arbitrary parameters; the separate sufficient conditions are given in Appendix C.6. In exploratory comparisons, Catoni-type aggregation, median-of-means, and direct truncation of the conditional risk underperformed global gradient clipping. Selectively downweighting rare large-scale clock realizations is a possible explanation, rather than a demonstrated causal mechanism.

E.6 Sampling

All HTFM and FM-based variants integrate the learned velocity field with a fixed-step ODE solver; LIM and DLPM use their native stochastic reverse samplers, while DLIM uses its deterministic reverse sampler.

2D mixture. Deterministic Euler ODE solver; $24{,}000$ samples generated per setting and compared to a held-out $24{,}000$-sample reference set.

CIFAR10-LT. HTFM and t-Flow use the deterministic Euler ODE solver. LIM and DLPM use their native stochastic samplers; DLIM uses the deterministic (η=0) variant. Higher-order solvers are reported only in the solver ablation (Appendix F.3). Unless otherwise stated, all FID results use the final EMA checkpoint after the fixed training budget.

HRRR VIL. Heun's second-order ODE solver at 25 reverse steps (NFE $= 50$ per sample), following [41]. Tail statistics are evaluated on $20{,}000$ generated samples.

E.7 Compute Resources

All experiments were conducted on a mix of NVIDIA L40S (48GB) and A100 (80GB) GPUs. CIFAR10-LT runs use up to 8 L40S GPUs in parallel, while HRRR VIL runs use A100 80G GPUs to accommodate the larger 128×128 single-channel inputs and the NCSN++ backbone. The total training budget is fixed by the per-dataset epoch counts reported in Appendix E.5 (100 epochs for the 2D mixture, 2000 epochs for CIFAR10-LT, and 1500 epochs for HRRR VIL). Sampling cost is dominated by the listed NFE budgets and the ODE/SDE solver choices in Appendix E.6.

F Ablation Studies

Unless otherwise stated, ablations are conducted on CIFAR10-LT because it exposes both global generation quality and tail-class coverage while remaining more economical than the HRRR sweep. The same protocol can be repeated on 2D toy distributions for visualization and on HRRR for the final robustness check.

F.1 Effect of Flow Type

We next ablate the interpolation *schedule* itself, holding the random clock and the source covariance fixed. We compare two flow types implemented in our codebase: (i) the standard *straight-line* interpolation $X_t^T = tX_1 + (1-t)X_0^T$ uses a clock-independent interpolation schedule and moves linearly in physical time t; and (ii) a *signature-linear VE* schedule whose interpolation coefficients perturb the straight-line coefficients by endpoint-preserving clock-dependent terms, as specified in (D.19). The parameters $\lambda_\alpha, \lambda_\beta$ control the size of these corrections; the two coefficients need not

Table 7: Flow-type ablation on CIFAR10-LT under HTFM-α at $p_\alpha = 1.7$. Entries are FID at NFE$= 100$ with Euler sampling. Columns give the velocity network's logsignature order, distinct from the order used to define the schedule statistic in (D.19).

Flow type	Level 0	Level 1	Level 2
Straight line	16.65	16.10	16.46
Signature-linear VE	16.43	16.19	16.71

sum to one. Comparing these two schedules under a matched signature feature isolates the value of clock-aware versus clock-blind interpolation. We run this ablation on the HTFM-α family at $p_\alpha = 1.7$ as a representative slice of the heavy-tailed regime.

F.2 Effect of Number of Function Evaluations (NFE)

We next vary the number of function evaluations (NFE) used by the sampler, holding training fixed and changing only the sampling budget. This ablation measures each method's quality–cost tradeoff: Figure 2 plots FID against NFE for HTFM-α, DLPM, and DLIM at three decay degrees $p_\alpha \in \{1.6, 1.7, 1.8\}$.

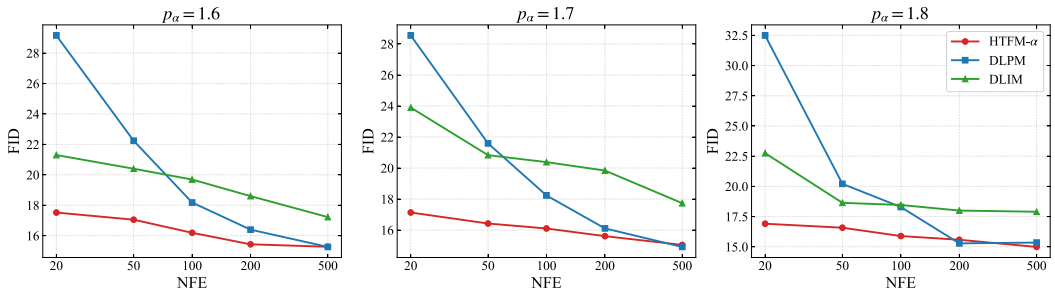

Figure 2: NFE-vs-FID curves on CIFAR10-LT. One panel per decay degree $p_\alpha \in \{1.6, 1.7, 1.8\}$; lines compare HTFM-$\alpha$ (ours) with the DLPM and DLIM baselines. Lower FID is better.

Three trends are visible across all three panels. **(i) HTFM-α dominates at low NFE.** The red curve sits well below both baselines from the smallest sampling budget, and the gap to DLPM is largest precisely where the sampling budget is smallest. **(ii) DLPM closes the gap only at high NFE.** The blue curve falls steeply with more solver steps and crosses or matches HTFM-α near the right edge of the budget range, indicating that the two methods become broadly comparable at the largest evaluated sampling budgets. **(iii) DLIM plateaus.** The green curve flattens early and stays above both other methods for the rest of the range, so its deterministic sampler does not scale further with extra NFE on this benchmark. Together these curves give the same message as the main-text CIFAR10-LT comparison: HTFM-α is broadly comparable to the strongest α-stable diffusion baseline at large evaluated budgets, while outperforming it under tight NFE budgets.

F.3 Effect of ODE Solver

This ablation compares the integration scheme used during sampling, holding the trained model fixed and varying the solver and its configured sampling budget. We compare four fixed-step ODE solvers implemented in our codebase: explicit *Euler* (1st order), *Heun* and *midpoint* (2nd-order single-step Runge–Kutta), and Adams–Bashforth *AB2* (2nd-order multi-step). Euler uses one new velocity evaluation per step, while standard Heun and midpoint updates use two. AB2 reuses the preceding velocity evaluation and requires only one new evaluation per step after startup; its total NFE also depends on the startup procedure. Thus AB2 should not be assigned the same two-evaluations-per-step conversion as Heun and midpoint. Table 8 reports nominal configured budgets, not independently verified equal-NFE budgets. These results therefore do not establish an equal-compute ranking of the solvers. We report HTFM-α at $p_\alpha = 1.7$ as a representative slice.

Table 8: ODE-solver ablation on CIFAR10-LT under HTFM-α at $p_\alpha = 1.7$. Each cell reports FID ($12{,}406$ generated vs $12{,}406$ reference, EMA, lower is better) at the original nominal sampling-budget labels. **Bold** indicates the lowest reported FID in each column.

Solver	Budget 20	Budget 50	Budget 100
Euler (1st order)	20.80	17.92	16.10
Heun (2nd order)	18.55	17.44	15.87
Midpoint (2nd order)	**17.82**	16.51	**15.81**
AB2 (2nd order, multi-step)	18.02	**16.47**	15.82

F.4 Signature order and direct scalar-summary conditioning

Table 9 gives the full signature-order sweep. At $m = 0$, clock features are withheld while the conditional-Gaussian sampling construction is retained. Stable-clock features follow the per-path standardization in Appendix E.4; for affine Student-t clocks, the scalar feature is $\log V_\nu$ and higher orders are omitted.

Table 9: Full signature-order ablation on CIFAR10-LT at NFE$= 100$. Bold indicates the best result in each row.

Variant	Tail exp.	$m = 0$	$m = 1$	$m = 2$
HTFM-α	$p_\alpha = 1.6$	16.78	**16.19**	16.21
	$p_\alpha = 1.7$	16.65	**16.10**	16.46
	$p_\alpha = 1.8$	16.49	**15.89**	15.94
HTFM-t	$p_\nu = 1.7$	18.72	**16.31**	–
	$p_\nu = 2.0$	17.30	**16.76**	–
	$p_\nu = 3.0$	16.58	**16.14**	–

For the stable clock at $p_\alpha = 1.7$, we compare T_1, the area $A(T) = \int_0^1 T_s \, \mathrm{d}s$, and their pair on CIFAR10-LT. For the raw time-augmented piecewise linear path, T_1 is the nonconstant degree-one coordinate, and its degree-two signed-area coordinate together with T_1 determines the interpolated area $A_N(T)$ (Appendix D). These direct summaries are distinct from the logsignature computed after per-path channel standardization. JS is the Jensen–Shannon divergence between the real and generated class histograms. M/Md/F are the frequency groups defined above.

Table 10: Direct scalar-summary ablation at stable exponent $p_\alpha = 1.7$.

Conditioning	FID ↓	JS ↓	KID M/Md/F ↓
T_1	16.10	0.0013	0.014 / 0.039 / 0.054
$A(T)$	17.86	0.0026	0.015 / 0.042 / 0.059
$(A(T), T_1)$	16.46	0.0019	0.014 / 0.042 / 0.057

For this straight-line flow, T_1 performs best among the tested summaries; the additional area information does not improve these metrics. The logsignature supplies a compact, extensible clock representation, rather than a requirement that every flow use high-order features. Whether higher orders benefit more strongly path-dependent schedules remains future work.

F.5 Additional rare-event evaluation

CIFAR10-LT protocol. We generate $50{,}000$ samples per model at NFE$= 100$ and assign class labels with a balanced CIFAR ResNet-18 trained on the full $50{,}000$-image CIFAR-10 training set. The frequency groups are Many (M): airplane, automobile, bird; Medium (Md): cat, deer, dog; and Few (F): frog, horse, ship, truck. Recall and Kernel Inception Distance (KID) are computed per class using 100 real and 100 generated samples over 100 no-replacement repetitions, then macro-averaged within each frequency group. These repetitions are evaluation subsamples, not independent training seeds. This protocol is separate from the $12{,}406$-sample global-FID protocol in Appendix E.3.

Let p_c and $\widehat{p}_c$ be the real and generated class proportions. For the Few-class set $\mathcal{F}$, define

$$r_c := \min\{1, \widehat{p}_c/p_c\}, \qquad \text{Worst} := \min_{c\in\mathcal{F}} r_c, \qquad \text{Macro} := \frac{1}{|\mathcal{F}|}\sum_{c\in\mathcal{F}} r_c,$$

and $\text{TailMass} := \sum_{c\in\mathcal{F}} \widehat{p}_c / \sum_{c\in\mathcal{F}} p_c$. Worst and Macro coverage are better near one; TailMass has target one and also exposes excess minority-class mass.

Table 3 in Section 4.4 reports the results for this protocol.

HTFM improves worst-case minority coverage and Few-class Recall over the reported DDPM and DLPM baselines, while KID remains comparable. It moderately over-generates Few-class mass; therefore coverage alone should not be read as exact class-distribution calibration.

Complete class histograms. Table 11 gives the complete generated-versus-real class proportions. The four Few classes contain 504 of the $12{,}406$ real training images. The reported rounded percentages can sum to slightly more or less than 100. The full histogram distinguishes broad minority coverage from a redistribution concentrated in one class.

Table 11: Class proportions ($\%$) on CIFAR10-LT. Air denotes airplane, Auto denotes automobile, and Few is the sum over frog, horse, ship, and truck. The last row is the real long-tailed training distribution.

Method	Air	Auto	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck	Few
HTFM-α ($p_\alpha = 1.6$)	40.27	21.35	15.15	9.18	5.13	4.03	2.15	1.16	0.88	0.70	4.89
HTFM-α ($p_\alpha = 1.7$)	40.47	21.60	15.16	8.93	5.16	3.93	2.12	1.14	0.77	0.72	4.75
HTFM-α ($p_\alpha = 1.8$)	39.97	21.36	15.22	8.89	5.23	4.23	2.37	1.09	0.91	0.73	5.10
HTFM-t ($p_\nu = 2.0$)	40.69	21.22	15.45	8.81	4.98	3.74	2.52	1.09	0.90	0.60	5.11
HTFM-t ($p_\nu = 3.0$)	41.10	21.18	15.67	8.45	4.94	3.74	2.30	1.06	0.85	0.71	4.92
DLPM ($p_\alpha = 1.6$)	39.46	20.56	16.33	11.22	5.42	3.69	1.22	0.92	0.75	0.44	3.33
DLPM ($p_\alpha = 1.7$)	41.25	18.79	16.35	10.83	5.39	3.98	1.31	0.94	0.73	0.43	3.41
DLPM ($p_\alpha = 1.8$)	40.04	21.02	16.86	10.10	5.11	3.69	1.20	0.93	0.69	0.36	3.18
LIM ($p_\alpha = 1.7$)	36.34	22.45	12.80	12.19	5.38	4.02	3.26	0.79	1.95	0.83	6.83
DDPM	41.61	19.62	17.52	9.69	4.86	3.92	1.03	0.95	0.52	0.29	2.78
Real LT training split	40.30	24.16	14.48	8.68	5.20	3.12	1.87	1.12	0.67	0.40	4.06

HRRR-VIL event protocol. We retain the original HRRR evaluation setting: $20{,}000$ generated fields per model and NFE$= 50$, with the setup in Appendix E. We fix an exceedance threshold q at the 99.9th percentile of all pixels in the real test set and apply the same threshold to every real and generated 128×128 field. We identify 8-connected components in the exceedance mask $\mathbf{1}_{\{X_{ij}>q\}}$. *Max KS* compares the real and generated distributions of per-field maxima. *Eccentricity KS* compares the eccentricities of the largest exceedance component, restricting this metric to fields containing at least one component. If $\lambda_1 \geq \lambda_2 \geq 0$ are the eigenvalues of the component pixel-coordinate covariance and $\lambda_1 > 0$, its eccentricity is $\sqrt{1-\lambda_2/\lambda_1}$: zero is circular and one is elongated. For a zero-variance component ($\lambda_1 = 0$, including a singleton), we set eccentricity to zero. Such fields remain included in the event-rate calculation; only fields without an exceedance component are excluded from eccentricity KS. *Event rate* is the percentage of fields with at least one exceedance component. These metrics assess peak intensity, component shape, and event frequency. KR, SR, and tail KS retain the definitions in Appendix E.3.

Table 4 in Section 4.4 reports the results for this protocol.

HTFM has the smallest KR, SR, tail KS, and eccentricity KS among these methods, and its event rate is closest to the real reference. Its Max KS improves over DDPM and DLPM, while LIM has the smaller Max KS. Thus these results retain a limitation in event-level peak calibration.

G Visualizations

This section collects qualitative visualizations that complement the quantitative results in the main text.

G.1 2D toy distributions: clock-degree sweep

Figure 3 compares HTFM-α samples generated under clocks of different decay degrees (rows: $p = \infty$, $p_\alpha = 1.8$, $p_\alpha = 1.5$) against ground-truth samples (top row) across five 2D toy targets (columns: `checkerboard`, `two_moons`, `swiss_roll`, `rings`, `olympic_rings`). The same backbone, training schedule, and NFE budget are used for every cell; only the random-clock law is swept.

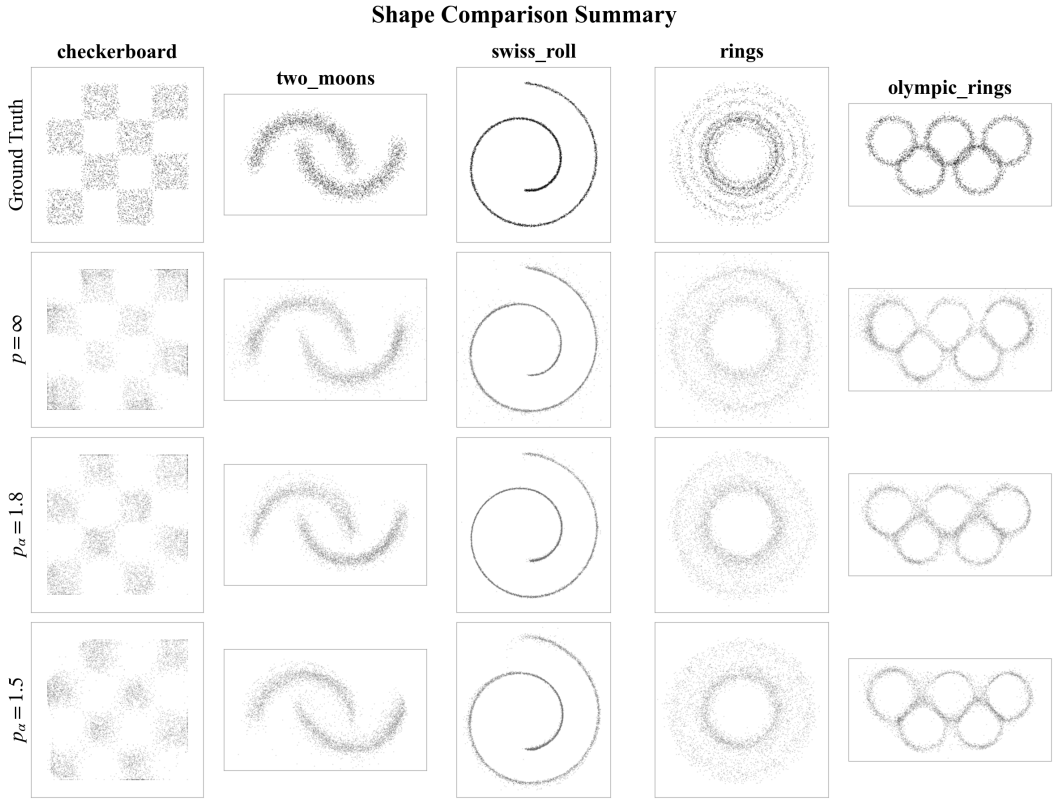

Figure 3: HTFM-α samples on five 2D toy distributions (`checkerboard`, `two_moons`, `swiss_roll`, `rings`, `olympic_rings`) under decay exponents $p \in \{\infty, 1.8, 1.5\}$ (for finite p, the source has stability index $\alpha = p$ and the stable-subordinator clock has index $\alpha/2$; the top row is ground truth). Same backbone, training, and NFE across cells.

G.2 2D $S\alpha S$ mixture: baseline comparison at NFE$=20$

Figure 4 compares HTFM against DLPM and DLIM on the 2D imbalanced α-stable mixture at NFE$=20$, on two stability indices $\alpha_{data} \in \{1.5, 1.8\}$ (rows). Each panel shows either generated samples (blue, the three right columns) or real samples (orange, the leftmost column). HTFM reproduces the imbalanced grid pattern more cleanly.

G.3 CIFAR10-LT: generated samples

Figure 5 shows generated 32×32 samples from HTFM-α on CIFAR10-LT, organised so that each column collects samples from a single CIFAR-10 class.

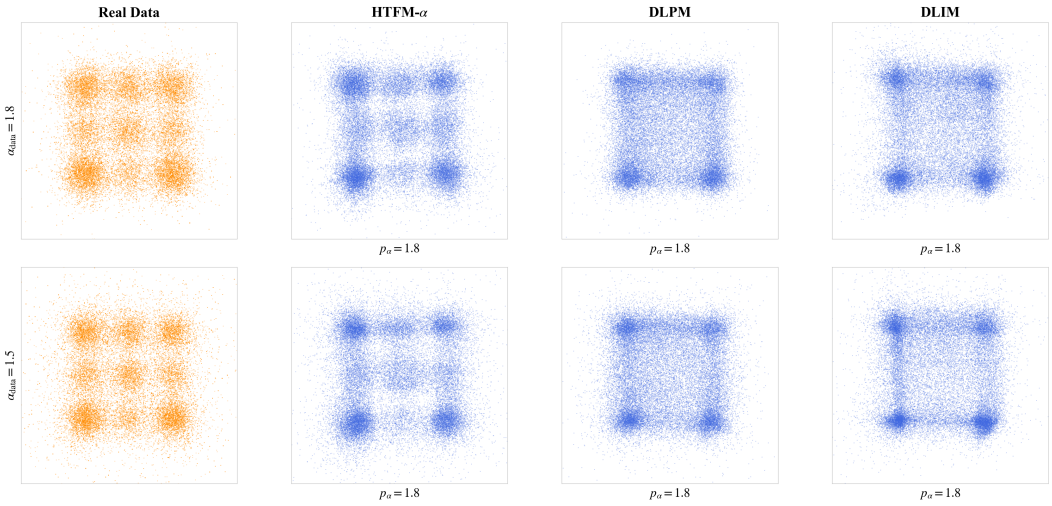

Figure 4: Generated samples on the 2D imbalanced α-stable mixture at NFE=20. Rows: $\alpha_{data} = 1.8$ (top), $\alpha_{data} = 1.5$ (bottom). Columns: real data, HTFM, DLPM, DLIM. The labels below generated panels give the source tail parameter p_α.

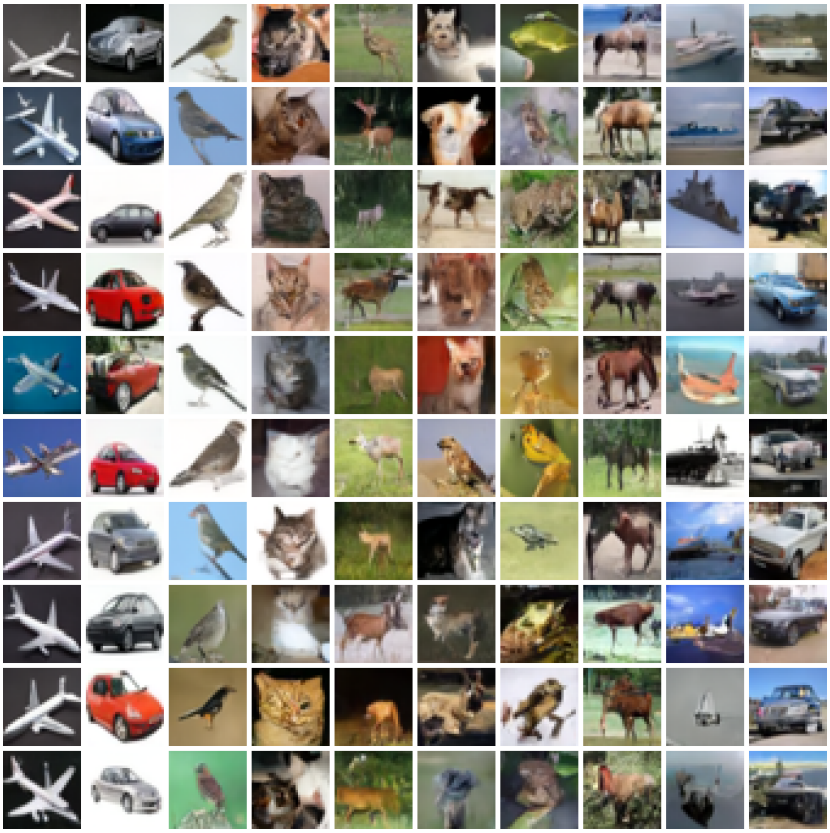

Figure 5: CIFAR10-LT generated samples from HTFM-α. Each column corresponds to one CIFAR-10 class; from left to right: *airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck*. This is also the CIFAR10-LT head-to-tail order, i.e. per-class training-sample counts decrease from the leftmost to the rightmost column. Sample fidelity broadly tracks the per-class training-set size: the head classes *airplane* and *automobile* (leftmost two columns) produce the sharpest, most coherent images, while progressively rarer classes (rightward) exhibit more artefacts and weaker class identity.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Contributions are summarized at the end of Section 1, matching the abstract.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Appendix C.6 states the moment, finite-risk model-class, and Jacobian-envelope assumptions needed for outer-clock integrability and local gradient equality, and does not guarantee finite loss at arbitrary initialization. Appendix D.1 states signature-linear approximation only on compact path sets. Appendix F.5 reports remaining minority-mass and event-maximum calibration errors. Data-adaptive clocks and benefits from higher-order path features remain extensions.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: All theoretical results are numbered, cross-referenced, and proved in full. Theorem 3.1 and Corollary 3.2 are proved through the general outer-clock result and local risk bound in Appendix C.6. The pathwise Proposition C.2 is stated and proved in Appendix C.3; Theorem 3.3 is proved in Appendix C.5. Supporting results—Lemma C.1, Theorem C.3, Theorem B.1, Corollary B.2, and Proposition D.1—are also stated with assumptions and proved.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: Appendix E reports datasets and splits (Appendix E.1), baselines (Appendix E.2), evaluation metrics (Appendix E.3), backbones and the HTFM path-feature module (Appendix E.4), training hyperparameters and clock-law settings (Appendix E.5), and ODE samplers (Appendix E.6). Code and reproduction instructions: `https://github.com/BIGKnight/Heavy-tailed-Flow-Matching`.

Guidelines:

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

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: Training, sampling, and evaluation code and reproduction instructions: `https://github.com/BIGKnight/Heavy-tailed-Flow-Matching`. All datasets are public: CIFAR-10 [29] with the long-tailed split of [6], and the HRRR VIL archive [5, 13]; preprocessing follows Appendix E.1.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Section 4.1 summarizes datasets, tasks, metrics, and methods. Appendix E reports data splits (Appendix E.1), backbones (Appendix E.4), AdamW hyperparameters, learning-rate schedules, EMA, gradient clipping, and epochs (Appendix E.5), and sampler/NFE choices (Appendix E.6).

   Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: Table 6 reports CIFAR10-LT FID as mean $\pm$ standard deviation over five independent training seeds. These standard deviations quantify training-run variability, not standard errors or confidence intervals. The original benchmark sweeps report three-seed means; no missing standard deviation is inferred for settings outside the five-seed table. The 100 repeated class-conditional evaluation subsamples are distinct from independent training runs. Additional training-seed uncertainty is not claimed for the new HRRR spatial metrics.

Guidelines:

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

8. **Experiments compute resources**

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

Answer: [No]

Justification: Appendix E.7 reports GPU types, memory, and the CIFAR10-LT GPU count; Appendix E.5 reports per-dataset epoch budgets. Per-run wall-clock training times and the total compute used for the research project are not reported.

Guidelines:

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

9. **Code of ethics**

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

Answer: [Yes]

Justification: The work is methodological and uses only public datasets (CIFAR-10, HRRR VIL); no human subjects, crowdsourcing, or sensitive personal data are involved.

Guidelines:

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

10. **Broader impacts**

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

Answer: [Yes]

Justification: Improved fidelity for rare events and minority categories may support applications such as weather modelling and imbalanced-image generation; downstream benefits in financial risk or recognition are not evaluated here. Negative impacts are those generic to high-quality generative models (e.g., synthetic-media misuse); HTFM does not specifically lower this barrier beyond standard flow matching.

Guidelines:

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

11. **Safeguards**

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

Answer: [N/A]

Justification: CIFAR10-LT models operate at 32×32 and are not realistic-image generators; HRRR VIL models output single-channel meteorological fields. No scraped or sensitive data is released.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All existing assets are credited: CIFAR-10 [29] with its long-tailed split [6], the HRRR/VIL archive [5, 13], and baselines LIM [50, 44], DLPM/DLIM [47], and t-Flow [41]. CIFAR-10 is publicly distributed for research; HRRR is a public US National Weather Service archive. Baselines are reproduced under the licenses of their public-code repositories.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: HTFM code, dependencies, run configurations, and reproduction instructions: `https://github.com/BIGKnight/Heavy-tailed-Flow-Matching`. No new datasets or pretrained weights are released.

    Guidelines:

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

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

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

Answer: [N/A]

Justification: No crowdsourcing or human-subject research.

Guidelines:

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

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

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

Answer: [N/A]

Justification: No human-subject research; IRB approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: LLMs are not used as a component of the methodology; any usage was limited to copy-editing and formatting.

Guidelines:

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