MODEL-AWARE SCHEDULES IMPROVE GENERATION VIA FIBERWISE OPTIMAL TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

In generative modeling with diffusion and flow matching, schedules control the signal and noise coefficients that mix data and noise along affine probability paths. Minimizing a kinetic action defined on coefficient paths, motivated by optimal transport, helps explain strong baselines but remains model-agnostic and ignores prediction error. Here we introduce a model-aware schedule construction that uses fiberwise optimal transport to improve generation quality. At a fixed time and state on the probability path, compatible signal/noise decompositions form an affine fiber. We define a fiberwise prediction risk by averaging optimal-transport costs between the true and predictor-induced decompositions within these fibers. On a fixed coefficient curve, combining this risk with coefficient-path kinetic action yields a closed-form optimal time allocation. This construction extends to general linear prediction targets, and the risk profile can be estimated from an early baseline checkpoint. We evaluate denoising diffusion probabilistic models (DDPMs) and flow matching across prediction targets, training configurations, risk-estimation checkpoints, datasets, and architectures. Our model-aware schedules consistently outperform strong baselines, reducing flow-matching FID by 38.6% on CIFAR-10 at 16 function evaluations. Each model-agnostic kinetic baseline determines its own kinetic reference coordinate. In these coordinates, fiberwise-risk profiles from independently trained models in different settings align closely after normalization to unit area. The resulting schedule deformations used in training also align, suggesting empirical universality across the evaluated models and settings. A frozen analytic allocation template retains most of the model-aware gain without further risk estimation or model-specific fitting.

1 INTRODUCTION

Diffusion and flow-matching models learn to generate samples from noise (Ho et al., 2020; Lipman et al., 2023). During training, a data sample x_0 and reference noise ϵ are commonly combined along an affine probability path with state $x_t = m_t x_0 + s_t \epsilon$ at time t. A schedule specifies how the signal and noise coefficients m_t, s_t evolve. A kinetic action defined on the path of coefficients, motivated by dynamical optimal transport (Benamou & Brenier, 2000), integrates the squared speed of these coefficients over time. On the respective fixed coefficient curves, its minimization yields cosine-type diffusion schedules and the standard conditional optimal-transport (Cond-OT) parameterization for flow matching (Shaul et al., 2023; Ikeda et al., 2025). However, the coefficient-path kinetic action is model-agnostic and ignores prediction error. Its connection to optimal transport motivates a model-dependent correction based on transport between signal/noise decompositions.

Here, we propose a model-aware schedule construction that improves generation quality. At fixed time t, different signal/noise decompositions can lead to the same given state $x_t = x$. All such decompositions form an affine fiber over x. We define a fiberwise prediction risk by averaging the optimal-transport costs between the true and predictor-induced decompositions within these fibers. This transport is base-preserving: it keeps x fixed and preserves the distribution p_t of x_t. We first train a model under a baseline schedule. From a fixed baseline checkpoint, we estimate the risk profile, i.e., the risk as a function of position along the baseline coefficient curve. On a fixed coefficient curve, combining this risk with coefficient-path kinetic action yields a closed-form optimal time allocation. This allocation defines our model-aware schedule, which is held fixed throughout training. The risk term favors less schedule time in higher-risk regions, while the kinetic term pe-

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

nalizes excessively rapid traversal (Figure 1). To compare generation quality with the baseline, we train a new model from scratch under our model-aware schedule, with all other settings matched.

We also compare risk profiles across models and settings in the kinetic reference coordinates determined independently by the model-agnostic kinetic baselines. After unit-area normalization, fiberwise-risk profiles from independently trained models align closely. The resulting allocation deformations used in training also align. This agreement is not assumed by the construction. We use the shared risk shape and the closed-form allocation rule to construct a frozen analytic allocation template (Figure 2; Section 6.4). These findings suggest that our model-aware schedule design can provide a reusable allocation structure in both diffusion and flow matching.

Our contributions are as follows:

- We formulate model-aware scheduling as minimum coefficient-path kinetic action under a budget on schedule-time-integrated fiberwise prediction risk. The fixed-curve specialization yields a closed-form optimal time allocation for one-shot schedule construction, using a risk profile that can be estimated from an early baseline checkpoint.
- We define the fiberwise prediction risk used in our model-aware schedule construction through base-preserving optimal transport on signal/noise decomposition fibers. Under the normalized Euclidean product metric, the risk reduces to the s_t^2-weighted noise-prediction mean squared error (MSE) and extends to general linear prediction targets, enabling direct risk estimation from the corresponding prediction errors.
- Across prediction targets, training configurations, datasets, and architectures, our schedules consistently improve strong DDPM and flow-matching baselines, including a 38.6% relative reduction in Fréchet Inception Distance (FID) (Heusel et al., 2017) at 16 function evaluations. Improvements also persist across multiple numerical solvers.
- We introduce kinetic reference coordinates defined by model-agnostic kinetic baselines. In these coordinates, shared normalized risk shapes and allocation deformations suggest empirical universality across the evaluated models and settings. Diagnostics on pretrained diffusion transformer (DiT) and 2-Rectified Flow (2-RF) models extend the normalized-risk agreement to larger conditional latent models with distinct architectures and generative constructions. The frozen analytic allocation template retains most of the model-aware improvement.

2 Related Work

Adjacent design choices at different levels. Diffusion and flow-matching models involve distinct design choices: the inference-time solver grid; the training-time sampling distribution and loss weighting; and the probability path and its traversal. Inference-time methods select grids using discretization-error criteria, trajectory regularity, or conditional-entropy change (Sabour et al., 2024; Chen et al., 2024; Stancevic et al., 2025). BOSS uses dynamic programming for grid selection and also fine-tunes the velocity field to its selected grid (Nguyen et al., 2024). Xu et al. (2026) design training-time sampling distributions and loss weights using estimated optimal loss values. InfoNoise adapts the training-time sampling distribution using an online estimate of the conditional-entropy-rate profile (Raya et al., 2026). We change only the traversal of a fixed coefficient curve and use the resulting schedule during retraining and sampling; the other choices remain complementary.

Schedule and probability-path design. Hang et al. (2025) design importance sampling in log signal-to-noise ratio (SNR) for continuous diffusion training rather than the discrete DDPM setting evaluated here. Constant Rate Scheduling reparameterizes variance-preserving (VP) diffusion schedules to equalize a chosen rate of distributional change. Its model-dependent procedure updates the schedule online, with training experiments under ϵ-prediction (Okada et al., 2026). Our construction instead estimates fiberwise prediction risk once from a baseline checkpoint and holds the schedule fixed during retraining. On prescribed paths, LayoutFlow tests a sine traversal for layout generation (Guerreiro et al., 2024), while Tsimpos et al. (2025) optimize traversal for a uniform-in-time spatial Lipschitz bound. Concurrent work uses polynomial velocity profiles motivated by Euler's local truncation error (Bondar, 2026). Another concurrent approach uses loss-quantile traversal that slows at high conditional-flow-matching loss (Tania & Khan, 2026). Chen et al. (2025) instead optimize interpolation coefficients for averaged squared drift Lipschitzness. Most of these

methods address either diffusion or flow matching. Our formulation applies across general linear prediction targets in both DDPM and flow matching. It balances coefficient-path kinetics against schedule-time-integrated fiberwise prediction risk. Improvements also persist across multiple numerical solvers. Empirical universality across the evaluated models and settings supports a frozen analytic allocation template. It can be used directly as a schedule without further risk estimation or model-specific fitting.

Kinetic and optimal-transport perspectives. Motivated by dynamical optimal transport (Benamou & Brenier, 2000), Shaul et al. (2023) derive expressions for conditional and marginal kinetic energy of affine probability paths. With independent endpoints and matched second moments, the normalized conditional energy is coefficient-path kinetic action, minimized by Cond-OT. They optimize the marginal kinetic energy over both the coefficient curve and its traversal. The resulting paths are evaluated in flow-matching experiments using independent endpoints rather than minibatch OT coupling. On ImageNet-64, the estimated optimum nearly coincides with the standard Cond-OT coefficient path. We instead minimize coefficient-path kinetic action over traversals of a fixed coefficient curve under a budget on schedule-time-integrated fiberwise prediction risk. Ikeda et al. (2025) also recover the Cond-OT and exact cosine parameterizations by minimizing coefficient-path kinetic action, with the latter subject to the VP constraint $m_t^2 + s_t^2 = 1$. Under their stated assumptions, they derive a speed-accuracy relation between marginal Wasserstein action and a specific Wasserstein generation-error sensitivity. Base-preserving Wasserstein formulations restrict transport to corresponding fibers over a shared base marginal (Peszek & Poyato, 2023; Chemseddine et al., 2025). We apply this structure to signal/noise decomposition fibers. The resulting risk measures prediction discrepancy at a fixed state and supplies the model-dependent term in our schedule-design objective.

3 Background

3.1 Affine Probability Paths and Schedule Coordinates

To describe diffusion and flow-matching schedules in a common framework, we work with affine probability paths

$$x_t = m_t x_0 + s_t \epsilon, \qquad x_0 \sim p_{\text{data}}, \quad \epsilon \sim \mathcal{N}(0, I), \qquad t \in [0, 1], \tag{1}$$

with marginals $p_t = \text{Law}(x_t)$. A schedule specifies a parameterized coefficient path $t \mapsto (m_t, s_t)$. The underlying coefficient curve is the geometric curve traced in the (m, s)-plane, independent of traversal speed. For VP/DDPM, write the squared signal coefficient as $\bar{\alpha}_t$. Then $m_t = \sqrt{\bar{\alpha}_t}$ and $s_t = \sqrt{1 - \bar{\alpha}_t}$, so $m_t^2 + s_t^2 = 1$. Standard VP/DDPM uses independent x_0 and ϵ. Under our data-to-noise convention, the standard Cond-OT parameterization is $m_t = 1 - t$ and $s_t = t$. With the endpoint marginals fixed, different couplings of data and noise samples (x_0, ϵ) can change their joint law and hence the prediction problem, while preserving the affine coefficient representation above (Ho et al., 2020; Lipman et al., 2023; Pooladian et al., 2023; Tong et al., 2024).

3.2 Coefficient-Path Kinetics and Standard Baselines

Let $\gamma(t) = (m_t, s_t)$ be a given coefficient path. For a fixed endpoint pair (x_0, ϵ), differentiating $x_t = m_t x_0 + s_t \epsilon$ with respect to t gives the path velocity $\dot{x}_t = \dot{m}_t x_0 + \dot{s}_t \epsilon$. Under standard regularity conditions, the expected pathwise kinetic action $\int_0^1 \mathbb{E}\|\dot{x}_t\|^2 \, \mathrm{d}t$ is an upper bound on the marginal Wasserstein action of p_t. This marginal action integrates the squared 2-Wasserstein speed of p_t over time and upper-bounds the endpoint quadratic transport cost. In the settings covered by speed-accuracy relations, it also upper-bounds a Wasserstein generation-error sensitivity (Benamou & Brenier, 2000; Shaul et al., 2023; Ikeda et al., 2025). Prior kinetic analyses use the Euclidean action of the coefficient path to explain strong baseline schedules (Shaul et al., 2023; Ikeda et al., 2025). We adopt this action and refer to it as *coefficient-path kinetics*:

$$\mathcal{J}_{\text{kin}}(\gamma) = \int_0^1 \left(\dot{m}_t^2 + \dot{s}_t^2\right) \mathrm{d}t. \tag{2}$$

For independent data/noise endpoints with matched second moments, $\mathcal{J}_{\text{kin}}$ equals the normalized expected pathwise kinetic action. With unequal second moments or coupled endpoints, the pathwise action is upper-bounded by a fixed multiple of $\mathcal{J}_{\text{kin}}$ (Appendix A.1).

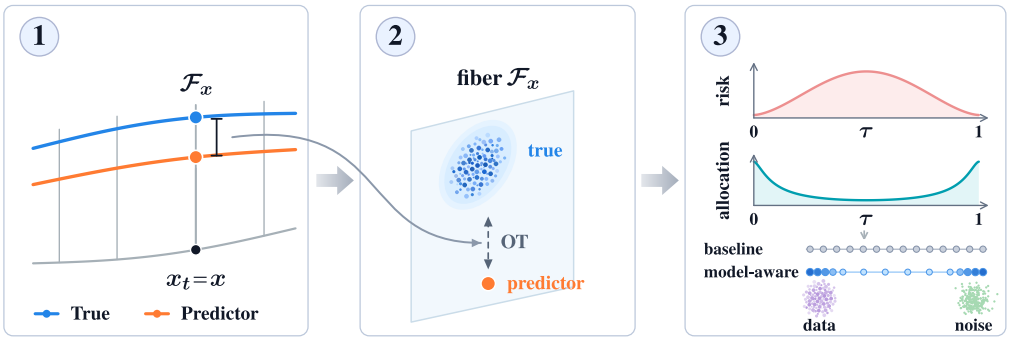


Figure 1: Fiberwise prediction risk and model-aware allocation. **1:** At $x_t = x$, the true and predictor-induced decompositions differ vertically within $\mathcal{F}_x$. **2:** The true conditional law and predictor-induced point mass define a base-preserving fiberwise transport cost, averaged over $x \sim p_t$. **3:** The resulting risk and coefficient-path kinetics determine schedule-time allocation on the fixed curve, with profile shapes based on the empirical findings; marker spacing is schematic.

For fixed coefficient endpoints $(1, 0)$ and $(0, 1)$, minimizing $\mathcal{J}_{\text{kin}}$ gives a model-agnostic kinetic baseline with constant-speed traversal of the shortest admissible coefficient curve. We call the normalized time of this baseline the *kinetic reference coordinate* τ. The resulting baselines are the standard Cond-OT parameterization and, under the VP constraint, the exact cosine parameterization (Shaul et al., 2023; Ikeda et al., 2025). For VP, the exact cosine parameterization is $m_\tau = \cos(\pi\tau/2)$ and $s_\tau = \sin(\pi\tau/2)$. This parameterization has appeared in prior diffusion work (Salimans & Ho, 2022). Our DDPM baselines use the standard offset-normalized cosine schedule introduced by Nichol & Dhariwal (2021), which closely approximates this kinetic baseline (Appendix A.2).

4 Fiberwise Prediction Risk via Optimal Transport

Section 3.2 defines the kinetic reference coordinate along each baseline coefficient curve. We now define the fiberwise prediction risk through base-preserving optimal transport between the true and predictor-induced decompositions over each fixed base state (Figure 1, panels 1 and 2).

4.1 Decomposition Fibers and Fiberwise Optimal Transport

To compare decompositions at a fixed state, we first describe their geometry. Write $x_t = u_t + n_t$, with $u_t = m_t x_0$ and $n_t = s_t \epsilon$. We define the affine fibers $\mathcal{F}_x$ over states x, the decomposition bundle $\mathcal{B}$ collecting these fibers, and the projection pr onto the base state:

$$\mathcal{F}_x = \{(u, n) : u + n = x\}, \qquad \mathcal{B} = \{(x, u, n) : u + n = x\}, \qquad \text{pr}(x, u, n) = x. \tag{3}$$

We equip the decomposition bundle with the normalized Euclidean product metric

$$g_{\mathcal{B}} := \frac{1}{2}\big(\|\mathrm{d}u\|^2 + \|\mathrm{d}n\|^2\big). \tag{4}$$

With the fiber coordinate $z = (u - n)/2$, this becomes

$$g_{\mathcal{B}} = \frac{1}{4}\|\mathrm{d}x\|^2 + \|\mathrm{d}z\|^2. \tag{5}$$

This gives an orthogonal base-fiber splitting. For a fixed endpoint pair, the true decomposition (u_t, n_t) evolves with time. The horizontal component of its velocity, orthogonal to the fiber, has squared norm $\|\dot{x}_t\|^2/4$. In contrast, true and predictor-induced decompositions at the same (t, x) differ only in the vertical direction (Appendix B.1).

For noise prediction, the true and predictor-induced decompositions are

$$(u_t, n_t) = (m_t x_0, s_t \epsilon), \qquad (u_t^\theta, n_t^\theta) = \big(x_t - s_t \epsilon_\theta(x_t, t), s_t \epsilon_\theta(x_t, t)\big). \tag{6}$$

At fixed t, the observed state $x_t = x$ may be compatible with different true signal/noise decompositions. Let μ_t^x denote their conditional law given $x_t = x$. In contrast, a deterministic predictor returns a single decomposition for the same (x, t), inducing the point mass

$$\nu_t^{\theta,x} = \delta_{\zeta_\theta(x,t)}, \qquad \zeta_\theta(x,t) = \big(x - s_t\epsilon_\theta(x,t), s_t\epsilon_\theta(x,t)\big) \in \mathcal{F}_x. \tag{7}$$

With $W_{2,\mathcal{B}}$ denoting the 2-Wasserstein distance on $\mathcal{F}_x$ induced by $g_{\mathcal{B}}$, we define the *fiberwise prediction risk* by the base-preserving transport cost

$$\mathcal{W}_{\mathrm{fib}}^2(t) := \int W_{2,\mathcal{B}}^2\big(\mu_t^x, \nu_t^{\theta,x}\big)\, p_t(\mathrm{d}x). \tag{8}$$

Intuitively, $\mathcal{W}_{\mathrm{fib}}^2(t)$ measures the average squared distance between true and predicted signal/noise decompositions that sum to the same observed state. Within each fiber, every true decomposition must be transported to the single predicted decomposition. The optimal-transport cost is therefore the conditional mean squared vertical discrepancy. Averaging over $x \sim p_t$ gives the corresponding expectation under the joint law of (x_0, ϵ). With $z_t = (u_t - n_t)/2$ and $z_t^\theta = (u_t^\theta - n_t^\theta)/2$, we write this risk as

$$\begin{aligned} R(t) &:= \mathcal{W}_{\mathrm{fib}}^2(t) = D_{\mathrm{fib}}^2(t) := \mathbb{E}\|z_t - z_t^\theta\|^2 = s_t^2 e_t, \\ e_t &:= \mathbb{E}\|\epsilon_\theta(x_t,t) - \epsilon\|^2. \end{aligned} \tag{9}$$

This identity makes the risk directly estimable from prediction errors. The transport defining $R(t)$ is restricted to each fixed-(t, x) fiber, leaving the base marginal p_t unchanged (Appendix B.2).

4.2 General Linear Prediction Targets and Target-Coordinate Invariance

To express the same risk for other prediction targets, consider a general linear prediction target:

$$y_t = a_t x_0 + b_t \epsilon, \qquad \Delta_t = m_t b_t - s_t a_t \neq 0. \tag{10}$$

A prediction $y_\theta(x_t, t)$ induces the noise coordinate

$$\epsilon_\theta^{\mathrm{ind}}(x_t,t) = \frac{-a_t x_t + m_t y_\theta(x_t,t)}{\Delta_t}. \tag{11}$$

Since $\epsilon_\theta^{\mathrm{ind}} - \epsilon = \frac{m_t}{\Delta_t}(y_\theta - y_t)$, the fiberwise risk can be written as

$$R(t) = s_t^2 \mathbb{E}\|\epsilon_\theta^{\mathrm{ind}}(x_t,t) - \epsilon\|^2 = \frac{m_t^2 s_t^2}{\Delta_t^2}\mathbb{E}\|y_\theta(x_t,t) - y_t\|^2. \tag{12}$$

Thus changing the prediction target changes only the coordinate expression of the same fiberwise risk. This expression applies to the DDPM and flow-matching targets considered here. This invariance does not imply matching risk profiles across separately trained models, different settings, or coefficient curves. The conditional and latent-state extension used for the pretrained-checkpoint diagnostics is given in Appendix D.3. For comparisons across models and settings, we express the fiberwise risk as $R(\tau)$ in the corresponding kinetic reference coordinate. Section 5 shows how this profile enters schedule-time allocation.

5 Model-Aware Schedule Design

5.1 Risk-Constrained Kinetic Formulation

To incorporate prediction risk into schedule design, we minimize coefficient-path kinetic action subject to a budget B_R on schedule-time-integrated fiberwise prediction risk:

$$\min_\gamma \ \mathcal{J}_{\mathrm{kin}}(\gamma) \qquad \text{subject to} \qquad \int_0^1 R(t)\,\mathrm{d}t \le B_R. \tag{13}$$

Under the conditions stated in Section 3.2 and Appendix A.1, this can also be viewed as minimizing a tractable upper bound on the corresponding Wasserstein generation-error sensitivity under a model-dependent risk budget. Introducing a Lagrange multiplier $\lambda \ge 0$ for the risk constraint gives, up to an additive constant independent of the schedule, the penalized objective

$$\mathcal{J}_\lambda(\gamma) = \mathcal{J}_{\mathrm{kin}}(\gamma) + \lambda \int_0^1 R(t)\,\mathrm{d}t. \tag{14}$$

The orthogonal base-fiber splitting in Eq. (5) separates changes in the state from changes in its decomposition at a fixed state. Under this splitting, a fixed multiple of $\mathcal{J}_{\mathrm{kin}}$ upper-bounds expected horizontal kinetic action (Appendix B.1), while R is a model-dependent vertical squared-distance potential. These terms do not form a single dynamical optimal-transport action. Rather than specifying B_R, we parameterize the tradeoff by λ; the attained risk integral gives the corresponding budget. We solve a one-shot fixed-curve problem with a fixed risk profile estimated from a baseline checkpoint, optimizing only traversal. Empirically, re-estimating risk after model-aware training changes the time allocations only marginally, supporting this one-shot construction (Appendix Table 15).

5.2 Fixed-Curve Model-Aware Allocation

To solve Eq. (13) explicitly on a fixed coefficient curve, let $\gamma_0(\tau)$ denote the baseline coefficient path in its kinetic reference coordinate. Under $t = \Phi(\tau)$, the *schedule-time allocation density* is

$$w(\tau) = \Phi'(\tau) > 0, \qquad \int_0^1 w(\tau)\,\mathrm{d}\tau = 1. \tag{15}$$

Thus $w(\tau)\,\mathrm{d}\tau$ is the schedule time assigned to a reference interval of width $\mathrm{d}\tau$. Let $L'(\tau) = \|\gamma_0'(\tau)\|$ denote the coefficient-space speed of γ_0, and write the previously estimated fiberwise-risk profile as $R(\tau)$. We optimize w while holding this profile fixed. The constrained problem becomes

$$\begin{aligned} \min_{w>0} \quad & \int_0^1 \frac{L'(\tau)^2}{w(\tau)}\,\mathrm{d}\tau \\ \text{subject to} \quad & \int_0^1 R(\tau)w(\tau)\,\mathrm{d}\tau \le B_R, \qquad \int_0^1 w(\tau)\,\mathrm{d}\tau = 1. \end{aligned} \tag{16}$$

Introducing the corresponding multiplier $\lambda \ge 0$ gives the penalized form

$$\mathcal{J}_\lambda[w] = \int_0^1 \left[\frac{L'(\tau)^2}{w(\tau)} + \lambda R(\tau)w(\tau)\right]\mathrm{d}\tau. \tag{17}$$

The Karush–Kuhn–Tucker (KKT) stationarity condition gives the closed-form solution:

$$w^\star(\tau) = \frac{L'(\tau)}{\sqrt{\eta + \lambda R(\tau)}}, \tag{18}$$

where η enforces normalization. This rule converts an estimated risk profile into a time allocation. Appendix C gives the derivation, discrete allocation rule, and related invariance properties.

With $L'(\tau)$ constant in the kinetic reference coordinate, higher $R(\tau)$ means less schedule time when $\lambda > 0$. The kinetic term penalizes arbitrarily rapid traversal. Hence R and λ jointly determine the allocation deformation: the change in schedule-time allocation relative to the baseline traversal (Figure 1, panel 3). Matching normalized risk-profile shapes need not yield matching allocation deformations: Eq. (18) also depends on absolute risk scale and λ relative to the kinetic term.

5.3 Instantiation for Diffusion and Flow Matching

For DDPM, we retain the standard offset-normalized cosine schedule as a practical approximation to the VP kinetic baseline. We deform only its traversal, using baseline time as the practical VP reference coordinate. Appendix A.2 gives the coordinate audit.

For flow matching, the standard Cond-OT time parameter is already the exact reference coordinate along $\gamma_0(\tau) = (1-\tau, \tau)$. Under $t = \Phi(\tau)$,

$$x_t = (1-\tau)x_0 + \tau\epsilon, \qquad \frac{\mathrm{d}x_t}{\mathrm{d}t} = \frac{\epsilon - x_0}{w(\tau)}, \qquad \tau = \Phi^{-1}(t). \tag{19}$$

Thus reparameterization changes both the state-time map and the linear velocity prediction target, not merely the training-time sampling distribution. Appendix D.1 details both constructions.

Table 1: CIFAR-10 DDPM FID (epoch 400; three paired seeds, mean ± standard deviation).

Sampler	NFE	**Baseline** FID ↓	**Model-aware (Ours)** FID ↓	ΔFID ↑	%Impr. ↑
DPM++3M	16	9.64 ± 0.26	8.06 ± 0.16	1.58 ± 0.11	$\mathbf{16.4 \pm 0.7}\%$
DPM++3M	32	7.69 ± 0.26	6.79 ± 0.27	0.91 ± 0.12	$11.8 \pm 1.5\%$
DPM++3M	64	6.83 ± 0.24	6.19 ± 0.27	0.64 ± 0.07	$9.3 \pm 1.2\%$
DDIM	16	11.27 ± 0.19	9.98 ± 0.23	1.29 ± 0.20	$11.4 \pm 1.7\%$
DPM++2M	16	9.81 ± 0.26	8.29 ± 0.14	1.51 ± 0.14	$15.4 \pm 1.1\%$

Table 2: ImageNet-64 DDPM FID at the 800k checkpoint.

Sampler	NFE	**Baseline** FID ↓	**Model-aware (Ours)** FID ↓	ΔFID ↑	%Impr. ↑
DPM++3M	16	25.01	23.26	1.75	$\mathbf{7.0}\%$
DPM++3M	32	20.59	20.13	0.46	2.3%
DPM++3M	64	20.05	19.50	0.55	2.7%
DDIM	16	27.46	26.55	0.91	3.3%
DPM++2M	16	24.90	23.39	1.51	6.1%

6 Experiments

6.1 Experimental Setup

We evaluate unconditional generation on CIFAR-10 and ImageNet-64 (Krizhevsky, 2009; van den Oord et al., 2016) using U-Net backbones and include the U-ViT-S/2 vision-transformer backbone (Bao et al., 2023) as an architecture-family control on CIFAR-10. Primary DDPM experiments use ϵ-prediction. Our primary flow-matching setting is conditional flow matching with minibatch OT endpoint coupling (OT-CFM), using a uniform training-time sampling distribution and velocity prediction. Baseline and model-aware models are trained from scratch under their respective schedules (Appendices D.1 and D.2). Section 6.3 reports robustness and ablation studies. Separately, we perform risk-profile and allocation diagnostics on public DiT-XL/2 and pre-distillation InstaFlow 2-RF checkpoints, without retraining or evaluating a modified generative schedule (Appendix D.3).

Following established evaluation practice (Nichol & Dhariwal, 2021; Karras et al., 2022), we compare generation quality using 50,000-sample FID at matched numbers of function evaluations (NFE). In each FID comparison, all other training and sampling settings are the same, and results are reported at matched training stages. We define $\Delta\mathrm{FID} = \mathrm{FID}_{\mathrm{baseline}} - \mathrm{FID}_{\mathrm{model\text{-}aware}}$, with positive values indicating improvement. %Impr. denotes relative FID reduction. Main CIFAR-10 results give mean and standard deviation over three pre-specified paired seeds; ImageNet-64 uses one paired seed. All other single-seed CIFAR-10 studies use the same preselected seed from this set. Coarse FID sweeps in the two primary CIFAR-10 settings set $\lambda = 220$ for DDPM and $\lambda = 450$ for flow matching (see Appendix D.1 for the numerical convention). These sweeps precede the cross-system risk-allocation analysis (Appendix E.6). We reuse the corresponding values without retuning across prediction targets, endpoint couplings, training-time sampling distributions, risk-estimation checkpoints, datasets, and architectures.

6.2 Main Results

DDPM. Our model-aware schedule reduces FID in all five reported sampling configurations on both CIFAR-10 and ImageNet-64 (Tables 1 and 2). On CIFAR-10, the FID reduction reaches 16.4% with DPM++3M at 16 NFE. FID improves across the full 3×3 sampler-NFE grid for all three paired seeds (Appendix Table 5). Cosine also outperforms the linear-β schedule throughout the grid (Appendix Table 6).

Flow matching. All four CIFAR-10 sampling configurations improve for each of the three paired seeds, with mean relative FID reductions of 6.9%–38.6%. All seven ImageNet-64 configurations improve; midpoint and Heun3 improve by 24.7%–29.8% (Tables 3 and 4).

Table 3: CIFAR-10 flow-matching FID (epoch 900; three paired seeds, mean ± standard deviation).

Integrator	NFE	**Baseline** FID ↓	**Model-aware (Ours)** FID ↓	ΔFID ↑	%Impr. ↑
Midpoint	16	7.62 ± 0.20	4.68 ± 0.13	2.95 ± 0.09	$\mathbf{38.6 \pm 0.6}\%$
Midpoint	32	5.52 ± 0.10	4.72 ± 0.10	0.79 ± 0.09	$14.4 \pm 1.6\%$
Midpoint	64	4.45 ± 0.07	4.15 ± 0.09	0.31 ± 0.11	$6.9 \pm 2.3\%$
Heun3	15	7.19 ± 0.17	4.88 ± 0.27	2.32 ± 0.12	$32.2 \pm 2.3\%$

Table 4: ImageNet-64 flow-matching FID at the 600k checkpoint.

Integrator	NFE	**Baseline** FID ↓	**Model-aware (Ours)** FID ↓	ΔFID ↑	%Impr. ↑
Midpoint	16	41.52	30.11	11.40	27.5%
Midpoint	32	38.75	29.08	9.67	25.0%
Midpoint	64	36.99	27.87	9.12	24.7%
Heun3	15	41.38	29.04	12.34	**29.8**%
Euler	16	40.29	36.50	3.79	9.4%
Euler	32	38.47	32.26	6.21	16.1%
Euler	64	37.33	30.34	6.99	18.7%

6.3 Robustness and Ablations

Model and training settings. On CIFAR-10, our model-aware schedules improve FID across all tested sampling configurations for both DDPM v-prediction and the U-ViT-S/2 architecture control. They also improve FID under changes to flow-matching endpoint coupling and the training-time sampling distribution (Appendices E.3, E.4, and E.5).

Training-stage, tradeoff-weight, and risk-estimation robustness. Gains persist at the tested training stages and for the tested nonzero λ values in both model families (Appendix E.6). Using epoch-100 rather than late-stage risk estimates changes final FID by only 0.02 for DDPM and 0.06 for flow matching (Appendix Tables 10 and 11).

Finite-step integration. For flow matching, our model-aware schedules improve FID across all tested higher-order integrators and budgets on CIFAR-10, but worsen FID with Euler. On ImageNet-64, they improve FID with Euler at all three budgets, indicating setting-specific degradation rather than systematic incompatibility with first-order integration (Table 4; Appendix E.8).

6.4 Shared Risk-Profile Shapes and Allocation Deformations

In the kinetic reference coordinates determined independently by the model-agnostic baselines, the CIFAR-10 DDPM ϵ-prediction, DDPM v-prediction, and OT-CFM risk profiles align closely after unit-area normalization. Their allocation deformations also align closely, with more schedule time near the endpoints than in the middle (Figure 2; Appendix Table 13). The DDPM and flow-matching λ values were selected separately based on FID, not to align these deformations.

To characterize this agreement and examine its scope, we fit analytic forms to the two primary CIFAR-10 baseline risk profiles and their corresponding model-aware allocation densities (DDPM ϵ-prediction and OT-CFM). The resulting *frozen risk-shape reference* is the unit-area normalization of $\sin^{5/4}(\pi\tau)$. The *frozen analytic allocation template* uses the reciprocal-square-root form in Eq. (18): $w_{\mathrm{ana}}(\tau) \propto \left(1 + 4\sin^{5/4}(\pi\tau)\right)^{-1/2}$. We compare unit-area risk profiles and allocation densities in the remaining settings with the respective frozen reference profiles without refitting, using Pearson correlation, relative L^2 error, and total variation (TV) (Appendix E.10).

Appendix E.10 extends both findings across targets, checkpoints, architectures, datasets, endpoint couplings, and flow-matching training-time sampling distributions. Using pretrained checkpoints without retraining, we also observe the shared normalized risk shape in larger conditional latent DiT-XL/2 and InstaFlow 2-RF models (Appendices D.3 and E.10).

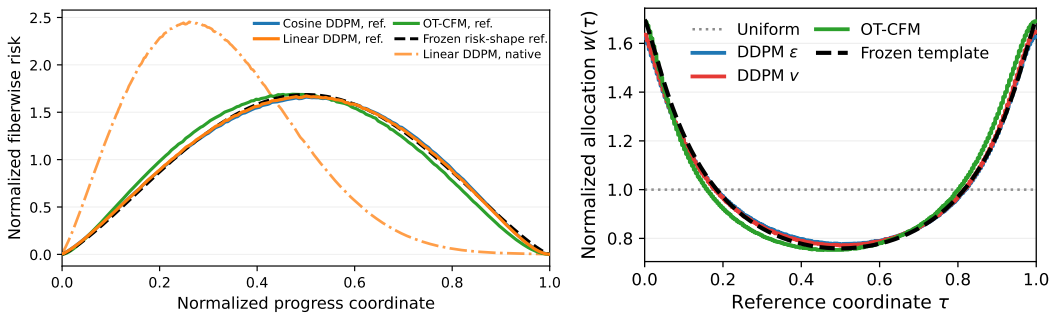


Figure 2: Unit-area fiberwise-risk profiles (**left**) and model-aware allocation densities (**right**). All use kinetic reference coordinates except the dash-dotted linear-DDPM profile in native model time. Dashed black curves denote the frozen reference profiles.

The linear-β DDPM control recovers this shared risk shape in the VP kinetic reference coordinate, but not in native model time (Figure 2, left). Linear and cosine DDPM follow the same VP coefficient curve but assign different native model times to the same coefficient pairs. At each native model time t, the linear-DDPM risk $R(t)$ is evaluated at the corresponding coefficient pair (m_t, s_t). We express the profile in the practical VP reference coordinate by matching each risk value to the cosine-baseline time with the same coefficient pair. The linear-DDPM profile agrees weakly with the risk-shape reference in native model time (Pearson 0.3616, TV 38.00%), but closely matches it in the practical VP reference coordinate (Pearson 0.9994, TV 0.84%; Appendix Table 14). In the corresponding kinetic reference coordinates, risk profiles re-estimated from the primary CIFAR-10 DDPM and OT-CFM model-aware checkpoints retain the shared normalized risk shape. The recomputed allocations also retain the shared deformation (Appendix Table 15).

An empirical Bayes decomposition indicates that the shared normalized risk shape is carried predominantly by predictor-dependent excess risk (Appendix E.12). Overall, these results suggest empirical universality across the evaluated models and settings. The reference coordinate, normalized risk shape, and allocation deformation connect the empirical comparison to our formulation. Coefficient-path kinetics determines the reference coordinate, and fiberwise optimal transport defines the model-dependent risk. Their fixed-curve tradeoff determines the allocation deformation.

On CIFAR-10, the analytic template recovers 98.1%, 98.0%, and 75.8% of the model-aware FID gains for DDPM ϵ-prediction, DDPM v-prediction, and OT-CFM, respectively. On ImageNet-64, the corresponding recovery fractions are 109.7% for DDPM and 76.2% for OT-CFM. Each recovery fraction compares gains over the same baseline, at the same training stage and with all other training and sampling settings matched. The analytic template thus distills rather than replaces the model-aware construction: the latter identifies the shared normalized risk shape, maps it to an allocation, and retains substantial additional flow-matching gains (Appendix E.11).

7 Discussion and Limitations

Our experiments show that combining coefficient-path kinetics with fiberwise prediction risk to optimize traversal can improve generation quality across models, training settings, and sampling settings. The shared risk shapes and allocation deformations suggest empirical universality and motivate the analytic template, which provides a directly usable schedule without further risk estimation or model-specific fitting. The additional flow-matching gains from the model-aware schedules highlight the value of setting-specific risk information.

The shared profile shape and quantitative cross-system agreement remain theoretically unexplained. A unified path-fiber optimal-transport formulation and an analysis of how the coefficient curve and training dynamics jointly shape fiberwise risk and allocation deformations may offer an explanation. Extending the one-shot construction beyond fixed coefficient curves remains future work. Our training and FID experiments use unconditional CIFAR-10 and ImageNet-64 for controlled comparisons. Pretrained-checkpoint diagnostics extend to large conditional latent DiT and 2-RF models, whose retraining and end-to-end evaluation under the resulting schedules remain future work.

AI USE STATEMENT

ChatGPT was used to assist with text editing, coding, the generation and refinement of Figure 1, and literature research. ChatGPT was not involved in research ideation, the development of theoretical results, research methodology, experimental design, or selection and verification of citations. All AI-assisted outputs were reviewed, revised where necessary, and verified by the authors.

REPRODUCIBILITY STATEMENT

The assumptions and derivations underlying our formulation are provided in Appendices A, B, and C. Appendix D.1 and Algorithm 1 detail risk estimation, schedule construction, and the training pipeline. Appendix D.2 specifies the architectures, optimization, and evaluation settings, while Appendix D.3 describes the pretrained-checkpoint diagnostic protocols. Additional experimental results and diagnostics are reported in Appendix E. Anonymized source code is provided as part of the supplementary material.

REFERENCES

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A ViT backbone for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22669–22679, 2023. doi: 10.1109/CVPR52729.2023.02171. URL https://openaccess.thecvf.com/content/CVPR2023/html/Bao_All_Are_Worth_Words_A_ViT_Backbone_for_Diffusion_Models_CVPR_2023_paper.html.

Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000. doi: 10.1007/s002110050002.

Vitalii Bondar. Velocity scheduled flow matching. *arXiv preprint arXiv:2607.11442*, 2026. URL https://arxiv.org/abs/2607.11442.

Jannis Chemseddine, Paul Hagemann, Gabriele Steidl, and Christian Wald. Conditional Wasserstein distances with applications in Bayesian OT flow matching. *Journal of Machine Learning Research*, 26(141):1–47, 2025. URL https://www.jmlr.org/papers/v26/24-0586.html.

Defang Chen, Zhenyu Zhou, Can Wang, Chunhua Shen, and Siwei Lyu. On the trajectory regularity of ODE-based diffusion sampling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7905–7934. PMLR, 2024. URL https://proceedings.mlr.press/v235/chen24bm.html.

Yifan Chen, Eric Vanden-Eijnden, and Jiawei Xu. Lipschitz-guided design of interpolation schedules in generative models. *arXiv preprint arXiv:2509.01629*, 2025. URL https://arxiv.org/abs/2509.01629.

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

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 2024. URL https://proceedings.mlr.press/v235/esser24a.html.

Julian Jorge Andrade Guerreiro, Naoto Inoue, Kento Masui, Mayu Otani, and Hideki Nakayama. LayoutFlow: Flow matching for layout generation. In *Computer Vision – ECCV 2024*, volume 15094 of *Lecture Notes in Computer Science*, pp. 56–72. Springer, 2024. doi: 10.1007/978-3-031-72764-1_4. URL https://doi.org/10.1007/978-3-031-72764-1_4.

Tiankai Hang, Shuyang Gu, Jianmin Bao, Fangyun Wei, Dong Chen, Xin Geng, and Baining Guo. Improved noise schedule for diffusion training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4796–4806, 2025. doi: 10.1109/ICCV51701.2025.00456. URL https://openaccess.thecvf.com/content/ICCV2025/html/Hang_Improved_Noise_Schedule_for_Diffusion_Training_ICCV_2025_paper.html.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Kotaro Ikeda, Tomoya Uda, Daisuke Okanohara, and Sosuke Ito. Speed-accuracy relations for diffusion models: Wisdom from nonequilibrium thermodynamics and optimal transport. *Physical Review X*, 15(3):031031, July 2025. doi: 10.1103/x5vj-8jq9. URL https://doi.org/10.1103/x5vj-8jq9.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/a98846e9d9cc01cfb87eb694d946ce6b-Abstract-Conference.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Sangyun Lee, Zinan Lin, and Giulia Fanti. Improving the training of rectified flows. In *Advances in Neural Information Processing Systems*, volume 37, pp. 63082–63109. Curran Associates, Inc., 2024. doi: 10.52202/079017-2014. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7343a5c976f8399880b695267f1f9e9f-Abstract-Conference.html.

Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5404–5411. IEEE, 2024. doi: 10.1109/WACV57701.2024.00532. URL https://openaccess.thecvf.com/content/WACV2024/html/Lin_Common_Diffusion_Noise_Schedules_and_Sample_Steps_Are_Flawed_WACV_2024_paper.html.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

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

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. InstaFlow: One step is enough for high-quality diffusion-based text-to-image generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=1k4yZbbDqX.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 22 (4):730–751, 2025. doi: 10.1007/s11633-025-1562-4. URL https://doi.org/10.1007/s11633-025-1562-4.

Bao Nguyen, Binh Nguyen, and Viet Anh Nguyen. Bellman optimal stepsize straightening of flow-matching models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Iyve2ycvGZ.

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

Shuntaro Okada, Kenji Doi, Ryota Yoshihashi, Hirokatsu Kataoka, and Tomohiro Tanaka. Constant rate scheduling: A general framework for optimizing diffusion noise schedule via distributional change. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=Pjq6kdvMBj.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023. doi: 10.1109/ICCV51070.2023.00387. URL https://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html.

Jan Peszek and David Poyato. Heterogeneous gradient flows in the topology of fibered optimal transport. *Calculus of Variations and Partial Differential Equations*, 62(9):258, 2023. doi: 10.1007/s00526-023-02601-8.

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. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28100–28127. PMLR, 2023. URL https://proceedings.mlr.press/v202/pooladian23a.html.

Gabriel Raya, Bac Nguyen, Georgios Batzolis, Yuhta Takida, Dejan Stancevic, Naoki Murata, Chieh-Hsin Lai, Yuki Mitsufuji, and Luca Ambrogioni. Noise scheduling as information-guided allocation in diffusion training. *arXiv preprint arXiv:2602.18647*, 2026. URL https://arxiv.org/abs/2602.18647.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. doi: 10.1109/CVPR52688.2022.01042. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28. URL https://doi.org/10.1007/978-3-319-24574-4_28.

Amirmojtaba Sabour, Sanja Fidler, and Karsten Kreis. Align your steps: Optimizing sampling schedules in diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42947–42975. PMLR, 2024. URL https://proceedings.mlr.press/v235/sabour24a.html.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *The Tenth International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TIdIXIpzhoI.

Neta Shaul, Ricky T. Q. Chen, Maximilian Nickel, Matthew Le, and Yaron Lipman. On kinetic optimal probability paths for generative models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 30883–30907. PMLR, 2023. URL https://proceedings.mlr.press/v202/shaul23a.html.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *The Ninth International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Dejan Stancevic, Florian Handke, and Luca Ambrogioni. Entropic time schedulers for generative diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1474. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3ed10616ecdd776be283c0a45cf9332d-Abstract-Conference.html.

Airin Akter Tania and Md Raihan Khan. Difficulty-calibrated interpolation paths for conditional flow matching. *arXiv preprint arXiv:2608.21286*, 2026. URL https://arxiv.org/abs/2608.21286.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=CD9Snc73AW.

Panos Tsimpos, Ren Zhi, Jakob Zech, and Youssef Marzouk. Optimal scheduling of dynamic transport. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 5441–5505. PMLR, 2025. URL https://proceedings.mlr.press/v291/tsimpos25a.html.

Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1747–1756. PMLR, 2016. URL https://proceedings.mlr.press/v48/oord16.html.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

Yixian Xu, Shengjie Luo, Liwei Wang, Di He, and Chang Liu. Diagnosing and improving diffusion models by estimating the optimal loss value. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=X7JfjLKKLQ.

A Coefficient-Path Kinetics and Reference-Coordinate Details

A.1 Relating Pathwise Action to Coefficient-Path Kinetics under General Endpoint Joint Laws

This appendix relates expected pathwise kinetic action to coefficient-path kinetics under general endpoint joint laws. We first show that the pathwise action is upper-bounded by a fixed multiple of coefficient-path kinetics, then identify cases of exact proportionality.

General endpoint joint laws. For a fixed joint law of (x_0, ϵ) with finite second moments,

$$\dot{x}_t = \dot{m}_t x_0 + \dot{s}_t \epsilon. \tag{A.1}$$

Let

$$A = \mathbb{E}\|x_0\|^2, \qquad B = \mathbb{E}\|\epsilon\|^2, \qquad C = \mathbb{E}\langle x_0, \epsilon\rangle. \tag{A.2}$$

Then

$$\mathbb{E}\|\dot{x}_t\|^2 = A\dot{m}_t^2 + 2C\dot{m}_t\dot{s}_t + B\dot{s}_t^2 = \dot{\gamma}(t)^\top G\dot{\gamma}(t), \tag{A.3}$$

where

$$G = \begin{pmatrix} A & C \\ C & B \end{pmatrix}. \tag{A.4}$$

The matrix G is positive semidefinite because its quadratic form is an expected squared norm. Let $c_G > 0$ denote its largest eigenvalue. Then $\dot{\gamma}(t)^\top G\dot{\gamma}(t) \le c_G(\dot{m}_t^2 + \dot{s}_t^2)$. Integrating this inequality over $t \in [0, 1]$ gives

$$\int_0^1 \mathbb{E}\|\dot{x}_t\|^2 \,\mathrm{d}t = \int_0^1 \dot{\gamma}(t)^\top G\dot{\gamma}(t) \,\mathrm{d}t \le c_G \mathcal{J}_{\mathrm{kin}}(\gamma). \tag{A.5}$$

For a fixed endpoint joint law, c_G is independent of traversal. Minimizing $\mathcal{J}_{\mathrm{kin}}$ therefore minimizes this upper bound, but not necessarily the pathwise action itself.

Independent endpoints. For independent endpoints, the centered noise gives $C = 0$, so $c_G = \max(A, B)$. Standard VP/DDPM is one example of this setting. If the endpoint second moments also match ($A = B$), then

$$\int_0^1 \mathbb{E}\|\dot{x}_t\|^2 \,\mathrm{d}t = A\mathcal{J}_{\mathrm{kin}}(\gamma). \tag{A.6}$$

This recovers the normalized relation used in the main text. Under the same conditions, $\mathbb{E}\|x_t\|^2 = A(m_t^2 + s_t^2)$, so the VP constraint preserves the common second-moment scale.

Fixed Cond-OT coefficient curve. On the fixed Cond-OT coefficient curve, write

$$m_t = 1 - r(t), \qquad s_t = r(t), \qquad r(0) = 0, \quad r(1) = 1. \tag{A.7}$$

Then

$$\mathbb{E}\|\dot{x}_t\|^2 = (A + B - 2C)\dot{r}(t)^2, \tag{A.8}$$

where

$$A + B - 2C = \mathbb{E}\|x_0 - \epsilon\|^2 \ge 0. \tag{A.9}$$

Since $\mathcal{J}_{\mathrm{kin}}(\gamma) = 2\int_0^1 \dot{r}(t)^2 \,\mathrm{d}t$, the pathwise action satisfies

$$\int_0^1 \mathbb{E}\|\dot{x}_t\|^2 \,\mathrm{d}t = (A + B - 2C)\int_0^1 \dot{r}(t)^2 \,\mathrm{d}t = \frac{A + B - 2C}{2}\mathcal{J}_{\mathrm{kin}}(\gamma). \tag{A.10}$$

This identity holds for any fixed endpoint joint law. Along this curve, endpoint moments and coupling change only the proportionality factor relating pathwise action to coefficient-path kinetics. For a fixed nondegenerate endpoint joint law with $A + B - 2C > 0$, replacing coefficient-path kinetics by pathwise action therefore only rescales the relative weighting of action and risk.

Cond-OT optimality without fixing the coefficient curve. For any absolutely continuous coefficient path with $\gamma(0) = (1, 0)$ and $\gamma(1) = (0, 1)$, the Cauchy–Schwarz inequality gives

$$\mathcal{J}_{\mathrm{kin}}(\gamma) \ge \left\| \int_0^1 \dot{\gamma}(t) \,\mathrm{d}t \right\|^2 = 2. \tag{A.11}$$

Applying the same inequality to each sample path and averaging gives

$$\int_0^1 \mathbb{E}\|\dot{x}_t\|^2 \,\mathrm{d}t \ge \mathbb{E}\left\| \int_0^1 \dot{x}_t \,\mathrm{d}t \right\|^2 = \mathbb{E}\|\epsilon - x_0\|^2 = A + B - 2C. \tag{A.12}$$

The standard Cond-OT coefficient path $\gamma(t) = (1 - t, t)$ attains both lower bounds. Thus, for each fixed endpoint joint law, it minimizes both coefficient-path kinetics and pathwise action when only the coefficient endpoints are prescribed.

A.2 Practical Approximation to the VP Kinetic Reference Coordinate

For the VP quarter-circle coefficient curve, the exact VP kinetic reference coordinate is

$$\tau_{\mathrm{VP}}(t) = \frac{2}{\pi} \arccos \sqrt{\bar{\alpha}(t)}, \tag{A.13}$$

with $\tau_{\mathrm{VP}} = 0$ at the data endpoint and $\tau_{\mathrm{VP}} = 1$ at the noise endpoint.

The standard offset-normalized cosine schedule used for our DDPM baselines parameterizes the same VP coefficient curve through

$$\varphi(t) = \frac{\pi}{2} \frac{t + s_{\mathrm{off}}}{1 + s_{\mathrm{off}}}, \qquad \bar{\alpha}(t) = \frac{\cos^2 \varphi(t)}{\cos^2 \varphi(0)}, \tag{A.14}$$

where s_{off} is the cosine offset. The cosine phase $\varphi(t)$ is affine in baseline time t. However, normalizing $\cos^2 \varphi(t)$ by $\cos^2 \varphi(0)$ makes $\tau_{\mathrm{VP}}(t)$ differ from t. We retain t as the practical VP reference coordinate. For $s_{\mathrm{off}} = 0.008$,

$$\|t - \tau_{\mathrm{VP}}(t)\|_\infty = 6.996 \times 10^{-3}. \tag{A.15}$$

Thus the practical and exact VP reference coordinates differ by less than 0.007 at every time $t \in [0, 1]$. At the same time t, the standard offset-normalized cosine schedule corresponds to exact VP reference coordinate $\tau_{\mathrm{VP}}(t)$, whereas the exact cosine parameterization corresponds to reference coordinate t. Since the exact cosine parameterization has constant coefficient-space speed $\pi/2$, the Euclidean distance between the two schedules' coefficient pairs at that time is at most $(\pi/2)|\tau_{\mathrm{VP}}(t) - t|$. The small coordinate discrepancy therefore implies uniformly close coefficient values for the two schedules. Appendix E.10 reports the corresponding DDPM coordinate controls. Using the exact rather than practical VP reference coordinate changes the DDPM risk-shape comparisons only marginally.

B Decomposition-Bundle Geometry and Fiberwise Optimal-Transport Details

B.1 Orthogonal Base-Fiber Splitting

We give the tangent-space calculation behind Section 4.1. Let d denote the state dimension. For a tangent vector with components $(\delta u, \delta n)$, the normalized Euclidean product metric $g_{\mathcal{B}}$ introduced in the main text induces the squared norm

$$\|(\delta u, \delta n)\|_{\mathcal{B}}^2 := \frac{1}{2} \left(\|\delta u\|^2 + \|\delta n\|^2 \right). \tag{B.1}$$

A point of $\mathcal{B}$ is a triple (x, u, n) satisfying $x = u + n$, so the pair (u, n) determines the full point. Equivalently, using the fiber coordinate $z = (u - n)/2$, the pair (x, z) determines u and n through

$$u = \frac{x}{2} + z, \qquad n = \frac{x}{2} - z. \tag{B.2}$$

The corresponding tangent-vector components satisfy

$$\delta u = \frac{1}{2} \delta x + \delta z, \qquad \delta n = \frac{1}{2} \delta x - \delta z, \tag{B.3}$$

and hence

$$\frac{1}{2} \left(\|\delta u\|^2 + \|\delta n\|^2 \right) = \frac{1}{4} \|\delta x\|^2 + \|\delta z\|^2. \tag{B.4}$$

The differential $\mathrm{d}\,\mathrm{pr}$ maps a tangent direction to the corresponding change in the base state. In the (u, n) coordinates on $\mathcal{B}$, $\mathrm{d}\,\mathrm{pr}(\delta u, \delta n) = \delta u + \delta n = \delta x$. The vertical tangent space V consists of directions that leave x fixed. Its orthogonal complement with respect to $g_{\mathcal{B}}$ is the horizontal tangent space H. In the (u, n) coordinates, these spaces are

$$V = \ker \mathrm{d}\,\mathrm{pr} = \{(a, -a) : a \in \mathbb{R}^d\}, \qquad H = V^\perp = \{(b, b) : b \in \mathbb{R}^d\}. \tag{B.5}$$

In the (x, z) coordinates, the same spaces are

$$V = \{(\delta x, \delta z) : \delta x = 0\}, \qquad H = \{(\delta x, \delta z) : \delta z = 0\}. \tag{B.6}$$

Thus vertical directions vary z at fixed x, whereas horizontal directions vary x at fixed z, giving the orthogonal base-fiber splitting.

A base displacement δx has horizontal lift $(\delta x, 0)$ in the (x, z) coordinates. Equivalently, its components in the (u, n) coordinates are

$$\left(\frac{\delta x}{2}, \frac{\delta x}{2}\right), \tag{B.7}$$

with squared norm $\frac{1}{4}\|\delta x\|^2$. For the true lifted path $\ell_t = (x_t, u_t, n_t)$, let $\dot{\ell}_t^H$ denote the horizontal component of its velocity $\dot{\ell}_t$. In the (x, z) coordinates, this component is $(\dot{x}_t, 0)$, so

$$\|\dot{\ell}_t^H\|_{\mathcal{B}}^2 = \frac{1}{4}\|\dot{x}_t\|^2. \tag{B.8}$$

Expected horizontal kinetic energy is therefore one quarter of expected pathwise kinetic energy. This does not imply that the full lifted tangent $\dot{\ell}_t$ is horizontal: the true decomposition may also move in the fiber direction as t varies. Rather, the splitting isolates the base-motion component from same-base decomposition discrepancies, which are purely vertical. For the true and predictor-induced lifts, which share the same base state,

$$\mathbb{E}\|z_t - z_t^\theta\|^2 = D_{\mathrm{fib}}^2(t), \tag{B.9}$$

so the fiberwise prediction discrepancy is purely vertical. Appendix A.1 upper-bounds the pathwise action by a fixed multiple of $\mathcal{J}_{\mathrm{kin}}$. Expected horizontal kinetic action is one quarter of the pathwise action and therefore satisfies the corresponding upper bound. Our objective combines $\mathcal{J}_{\mathrm{kin}}$ with a vertical squared-distance potential; it is not the full kinetic action of the lifted path.

B.2 Full Fiberwise Optimal-Transport Derivation

This subsection gives the full disintegration argument behind the base-preserving fiberwise optimal-transport construction in Section 4.1. We present the derivation for noise prediction. For a general linear prediction target with $\Delta_t \neq 0$, the same argument applies after replacing ϵ_θ by the induced noise prediction $\epsilon_\theta^{\mathrm{ind}}$ defined in Eq. (11).

Recall the decomposition bundle

$$\mathcal{B} = \{(x, u, n) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d : u + n = x\}, \qquad \mathrm{pr}(x, u, n) = x, \tag{B.10}$$

with fiber

$$\mathcal{F}_x = \mathrm{pr}^{-1}(x) = \{(u, n) : u + n = x\}. \tag{B.11}$$

Fix an arbitrary time t for the derivation below. The true and predictor-induced lifts

$$\ell_t = (x_t, u_t, n_t), \qquad \ell_t^\theta = (x_t, u_t^\theta, n_t^\theta) \tag{B.12}$$

are random points in $\mathcal{B}$. The predictor-induced lift is a deterministic function of the random state x_t. Both lifts satisfy

$$\mathrm{pr}(\ell_t) = \mathrm{pr}(\ell_t^\theta) = x_t. \tag{B.13}$$

Their laws are

$$\mu_t = \mathrm{Law}(x_t, u_t, n_t), \qquad \nu_t^\theta = \mathrm{Law}(x_t, u_t^\theta, n_t^\theta), \tag{B.14}$$

where

$$u_t = m_t x_0, \qquad n_t = s_t \epsilon, \tag{B.15}$$

and

$$n_t^\theta = s_t \epsilon_\theta(x_t, t), \qquad u_t^\theta = x_t - n_t^\theta. \tag{B.16}$$

Both measures project to the same base marginal $p_t = \mathrm{Law}(x_t)$. Disintegrating over x expresses each measure in terms of this shared marginal and its own conditional decomposition laws on the fibers $\mathcal{F}_x$:

$$\mu_t(\mathrm{d}x, \mathrm{d}u, \mathrm{d}n) = p_t(\mathrm{d}x)\, \mu_t^x(\mathrm{d}u, \mathrm{d}n), \tag{B.17}$$

$$\nu_t^\theta(\mathrm{d}x, \mathrm{d}u, \mathrm{d}n) = p_t(\mathrm{d}x)\, \nu_t^{\theta,x}(\mathrm{d}u, \mathrm{d}n). \tag{B.18}$$

For each fixed base point x, μ_t^x is the conditional decomposition law supported on $\mathcal{F}_x$:

$$\mu_t^x = \mathrm{Law}\big((m_t x_0, s_t\epsilon) \mid x_t = x\big). \tag{B.19}$$

For a deterministic predictor,

$$\nu_t^{\theta,x} = \delta_{\zeta_\theta(x,t)}, \tag{B.20}$$

where

$$\zeta_\theta(x,t) = \big(x - s_t\epsilon_\theta(x,t), s_t\epsilon_\theta(x,t)\big) \in \mathcal{F}_x. \tag{B.21}$$

Let

$$d_{\mathcal{B}}^2\big((u,n),(u',n')\big) = \frac{1}{2}\big(\|u-u'\|^2 + \|n-n'\|^2\big) \tag{B.22}$$

denote the squared distance on each fiber induced by the normalized Euclidean product metric $g_{\mathcal{B}}$. We define the fiberwise prediction risk by

$$\mathcal{W}_{\mathrm{fib}}^2(t) := \int W_{2,\mathcal{B}}^2\big(\mu_t^x, \nu_t^{\theta,x}\big)\, p_t(\mathrm{d}x). \tag{B.23}$$

Because $\nu_t^{\theta,x}$ is a Dirac mass, its coupling with μ_t^x is unique and therefore optimal. Hence

$$W_{2,\mathcal{B}}^2\big(\mu_t^x, \nu_t^{\theta,x}\big) = \int_{\mathcal{F}_x} d_{\mathcal{B}}^2\big((u,n), \zeta_\theta(x,t)\big)\, \mu_t^x(\mathrm{d}u, \mathrm{d}n). \tag{B.24}$$

Equivalently,

$$W_{2,\mathcal{B}}^2\big(\mu_t^x, \nu_t^{\theta,x}\big) = \mathbb{E}\left[d_{\mathcal{B}}^2\big((u_t,n_t),(u_t^\theta,n_t^\theta)\big) \,\middle|\, x_t = x\right]. \tag{B.25}$$

Averaging over $x \sim p_t$ gives

$$\mathcal{W}_{\mathrm{fib}}^2(t) = \frac{1}{2}\mathbb{E}\left[\|u_t - u_t^\theta\|^2 + \|n_t - n_t^\theta\|^2\right]. \tag{B.26}$$

Since the true and predictor-induced decompositions lie in the same fiber,

$$u_t + n_t = u_t^\theta + n_t^\theta = x_t, \tag{B.27}$$

and therefore

$$u_t - u_t^\theta = -(n_t - n_t^\theta). \tag{B.28}$$

It follows that

$$\begin{aligned} R(t) := \mathcal{W}_{\mathrm{fib}}^2(t) = D_{\mathrm{fib}}^2(t) &= \mathbb{E}\|n_t - n_t^\theta\|^2 \\ &= s_t^2\mathbb{E}\|\epsilon - \epsilon_\theta(x_t,t)\|^2 = s_t^2 e_t. \end{aligned} \tag{B.29}$$

The fiberwise risk $R(t)$ can equivalently be formulated as an optimal-transport problem on the decomposition bundle $\mathcal{B}$. Write

$$\omega = (x,u,n), \qquad \omega' = (x',u',n'). \tag{B.30}$$

Define the extended-valued base-preserving cost

$$c(\omega,\omega') = \begin{cases} d_{\mathcal{B}}^2\big((u,n),(u',n')\big), & x = x', \\ +\infty, & x \neq x'. \end{cases} \tag{B.31}$$

Let $\Pi(\mu_t, \nu_t^\theta)$ denote the set of couplings of μ_t and ν_t^θ. Then

$$\mathcal{W}_{\mathrm{fib}}^2(t) = \inf_{\Gamma \in \Pi(\mu_t, \nu_t^\theta)} \int c(\omega,\omega')\, \Gamma(\mathrm{d}\omega, \mathrm{d}\omega'). \tag{B.32}$$

Because $c(\omega,\omega') = +\infty$ whenever $x \neq x'$, any finite-cost coupling must pair points with the same base state. Conditioning on this shared state x gives a coupling between μ_t^x and $\nu_t^{\theta,x}$ within $\mathcal{F}_x$. Thus the formulation on $\mathcal{B}$ is exactly equivalent to the base-preserving fiberwise transport definition of $R(t)$.

C Fixed-Curve Optimization Details

C.1 Constrained Formulation and KKT Solution

Let

$$\mathcal{K}[w] := \int_0^1 \frac{L'(\tau)^2}{w(\tau)}\,\mathrm{d}\tau, \qquad \mathcal{E}_R[w] := \int_0^1 R(\tau)w(\tau)\,\mathrm{d}\tau. \tag{C.1}$$

The fixed-curve formulation of Section 5.2 is

$$\min_{w>0} \mathcal{K}[w] \qquad \text{subject to} \qquad \mathcal{E}_R[w] \le B_R, \qquad \int_0^1 w(\tau)\,\mathrm{d}\tau = 1. \tag{C.2}$$

The kinetic term is convex in w for $w > 0$, and both constraints are linear. Under the usual regularity and strict-feasibility conditions, the KKT conditions characterize the optimum. With multipliers $\lambda \ge 0$ and η, the Lagrangian is

$$\mathcal{L}[w, \lambda, \eta] = \int_0^1 \left[\frac{L'(\tau)^2}{w(\tau)} + \lambda R(\tau)w(\tau) + \eta w(\tau)\right]\mathrm{d}\tau - \lambda B_R - \eta. \tag{C.3}$$

Pointwise stationarity gives

$$-\frac{L'(\tau)^2}{w(\tau)^2} + \lambda R(\tau) + \eta = 0, \tag{C.4}$$

and hence

$$w^\star(\tau) = \frac{L'(\tau)}{\sqrt{\eta + \lambda R(\tau)}}. \tag{C.5}$$

The normalization condition determines η, and complementary slackness gives

$$\lambda\big(\mathcal{E}_R[w^\star] - B_R\big) = 0. \tag{C.6}$$

Thus an inactive risk constraint has $\lambda = 0$ and recovers the kinetic baseline, whereas $\lambda > 0$ implies an active risk constraint and selects a point on the kinetic-risk tradeoff. For normalized w, the η-dependent terms cancel. The Lagrangian then equals $\mathcal{J}_\lambda[w] - \lambda B_R$, where $\mathcal{J}_\lambda$ is the penalized objective used in the main text.

C.2 Discrete Interval Allocation

For a partition $0 = \tau_0 < \cdots < \tau_K = 1$, define the coefficient-space arc length L_k and the schedule time Δt_k allocated to each interval by

$$L_k := \int_{\tau_k}^{\tau_{k+1}} L'(u)\,\mathrm{d}u, \qquad \Delta t_k := \Phi(\tau_{k+1}) - \Phi(\tau_k), \tag{C.7}$$

and let r_k denote the interval-average fiberwise risk. The discrete fixed-curve kinetic and schedule-time-integrated risk terms are

$$\mathcal{K}_{\mathrm{disc}} = \sum_k \frac{L_k^2}{\Delta t_k}, \qquad \mathcal{E}_{R,\mathrm{disc}} = \sum_k r_k \Delta t_k. \tag{C.8}$$

The constrained problem therefore becomes

$$\begin{aligned} \min_{\Delta t_k > 0} \quad & \sum_k \frac{L_k^2}{\Delta t_k} \\ \text{subject to} \quad & \sum_k r_k \Delta t_k \le B_R, \qquad \sum_k \Delta t_k = 1. \end{aligned} \tag{C.9}$$

For fixed $\lambda \ge 0$, the corresponding penalized objective is

$$\mathcal{J}_\lambda^{\mathrm{disc}} = \sum_k \frac{L_k^2}{\Delta t_k} + \lambda \sum_k r_k \Delta t_k. \tag{C.10}$$

The KKT stationarity condition gives the allocation rule, with η determined by normalization:

$$\Delta t_k^\star = \frac{L_k}{\sqrt{\eta + \lambda r_k}}, \qquad \sum_k \Delta t_k^\star = 1, \tag{C.11}$$

which is the coefficient-space allocation rule underlying the one-shot construction in Appendix D.1.

C.3 Effective Geometry and Risk-Recalibration Properties

The kinetic baseline has constant coefficient-space speed. To give the model-aware optimum $w^\star$ a similar interpretation, let $\mathrm{d}L = L'(\tau)\,\mathrm{d}\tau$ denote the coefficient-space arc-length element. We define the risk-modified line element $\mathrm{d}\widetilde{L}$ by

$$\mathrm{d}\widetilde{L} := \frac{\mathrm{d}L}{\sqrt{\eta + \lambda R(\tau)}}. \tag{C.12}$$

The allocation formula in Eq. (C.5) then gives $\mathrm{d}\widetilde{L}/\mathrm{d}\tau = w^\star(\tau)$. Under this allocation, schedule time satisfies $\mathrm{d}t = w^\star(\tau)\,\mathrm{d}\tau$, so

$$\frac{\mathrm{d}\widetilde{L}}{\mathrm{d}t} = 1. \tag{C.13}$$

Thus the optimal allocation gives constant-speed traversal with respect to $\mathrm{d}\widetilde{L}$. The cumulative risk-modified length from the starting point gives the schedule time t at each reference location τ. This time assignment acts as an effective model-dependent clock along the fixed coefficient curve.

The allocation is also invariant to positive affine recalibration of the risk profile. If

$$\widetilde{R}(\tau) = aR(\tau) + b, \qquad a > 0, \tag{C.14}$$

then choosing

$$\widetilde{\lambda} = \frac{\lambda}{a}, \qquad \widetilde{\eta} = \eta - \frac{\lambda b}{a} \tag{C.15}$$

gives

$$\widetilde{\eta} + \widetilde{\lambda}\widetilde{R}(\tau) = \eta + \lambda R(\tau). \tag{C.16}$$

Thus $w^\star$ is unchanged and remains normalized. Consequently, $\widetilde{\eta}$ is the normalization multiplier for the recalibrated risk $\widetilde{R}$ and tradeoff weight $\widetilde{\lambda}$. An additive risk offset b is absorbed by the normalization multiplier. A positive risk rescaling by a can be compensated by replacing λ with λ/a.

D Model-Aware Construction and Experimental Details

D.1 Model-Aware Schedule Construction and Training Pipeline

All model-aware experiments use the same one-shot construction pipeline. We first train a model under the standard baseline schedule and use a fixed baseline checkpoint to estimate the fiberwise-risk profile along the baseline coefficient path. From the estimated risk profile, we construct the model-aware reparameterization using the closed-form allocation rule and then train a new model from scratch under the resulting fixed schedule. The baseline checkpoint is used only for schedule construction: neither the estimated risk profile nor the schedule is updated during model-aware training.

Throughout the numerical implementation and experiments, risk estimates use per-coordinate averaging over the d scalar components of the image or latent state. The numerical estimate $\widehat{D}^2_{\mathrm{fib}}$ therefore estimates $D^2_{\mathrm{fib}}/d = R/d$. We compute this estimate using the corresponding specialization of Eq. (12), with $\widehat{\mathbb{E}}[\|\cdot\|^2]$ denoting the average squared residual over both sampled endpoint pairs and the d scalar state components. In this subsection, r_k denotes the corresponding interval risk estimate, and λ denotes its tradeoff weight. All reported λ values follow this convention. With the kinetic term unchanged, a numerical weight λ corresponds to a weight λ/d for the theoretical risk R.

For model-aware schedule construction, we use the reference partition defined in Appendix C.2, with $K = 199$ for DDPM and $K = 200$ for flow matching. In both cases, the intervals have equal width in τ. Let $\tau_k^\circ = (\tau_k + \tau_{k+1})/2$ be the midpoint of interval k. We approximate the per-coordinate interval-average fiberwise risk by the midpoint Monte Carlo estimate

$$r_k \approx \widehat{D}^2_{\mathrm{fib}}(\tau_k^\circ). \tag{D.1}$$

Algorithm 1 One-shot model-aware schedule construction

Require: Baseline coefficient path γ_0, baseline checkpoint with parameters θ_{ref}, tradeoff λ, partition $0 = \tau_0 < \cdots < \tau_K = 1$

1: **for** $k = 0, \ldots, K-1$ **do**
2: Set representative location $\tau_k^\circ = (\tau_k + \tau_{k+1})/2$
3: Estimate the midpoint risk $r_k \leftarrow \widehat{D}^2_{\mathrm{fib}}(\tau_k^\circ; \theta_{\mathrm{ref}})$
4: Compute the interval coefficient-space arc length L_k
5: **end for**
6: Solve for η such that $\Delta t_k^\star = L_k/\sqrt{\eta + \lambda r_k}$ and $\sum_k \Delta t_k^\star = 1$
7: Construct $t = \Phi(\tau)$ from the cumulative $\{\Delta t_k^\star\}_{k=0}^{K-1}$
8: Realize Φ in the model-family-specific training schedule
9: Train a new model from scratch with Φ fixed

Applying Eq. (C.11) to these risk estimates gives the interval allocations. Their cumulative sums then define the monotone piecewise-linear reparameterization $t = \Phi(\tau)$.

Algorithm 1 summarizes the coefficient-space construction using L_k. The practical DDPM implementation replaces L_k with cosine-phase increments, as described below. For the straight Cond-OT curve, L_k is available in closed form.

DDPM. For DDPM, we retain the standard offset-normalized cosine schedule of Appendix A.2. Its baseline time τ is the practical VP reference coordinate. With $\varphi(\tau)$ denoting the cosine phase,

$$\bar{\alpha}(\tau) = \frac{\cos^2 \varphi(\tau)}{\cos^2 \varphi(0)}, \qquad \gamma_0(\tau) = \left(\sqrt{\bar{\alpha}(\tau)}, \sqrt{1 - \bar{\alpha}(\tau)}\right). \tag{D.2}$$

On this unit-radius VP coefficient curve, the interval coefficient-space arc length is

$$L_k = \arccos \sqrt{\bar{\alpha}(\tau_{k+1})} - \arccos \sqrt{\bar{\alpha}(\tau_k)}. \tag{D.3}$$

In the practical DDPM implementation, we replace L_k with the cosine-phase increment

$$\Delta\varphi_k := \varphi(\tau_{k+1}) - \varphi(\tau_k) \tag{D.4}$$

and use

$$\Delta t_k^\star = \frac{\Delta\varphi_k}{\sqrt{\eta + \lambda r_k}}, \qquad \sum_k \Delta t_k^\star = 1. \tag{D.5}$$

Because $\varphi(\tau)$ is affine in τ, setting $\lambda = 0$ gives $\Delta t_k^\star = \tau_{k+1} - \tau_k$ and thus recovers the standard cosine baseline timing. Nonconstant fiberwise risk deforms the allocation over the practical VP reference coordinate.

For the VP coefficient curve, this replacement has a small total approximation error in coefficient-space arc length. To quantify this error, note that on this unit-radius curve, $\arccos(\cos\varphi/\cos\varphi(0))$ gives the coefficient-space arc length from the data endpoint to phase φ. For $\varphi(0) < \varphi < \pi/2$, its derivative satisfies

$$\frac{\mathrm{d}}{\mathrm{d}\varphi} \arccos \frac{\cos\varphi}{\cos\varphi(0)} = \frac{\sin\varphi}{\sqrt{\sin^2\varphi - \sin^2\varphi(0)}} \geq 1. \tag{D.6}$$

Hence $L_k \geq \Delta\varphi_k$ on every interval. Since the curve has total arc length $\pi/2$, the total absolute approximation error relative to this length is

$$\frac{\sum_k |L_k - \Delta\varphi_k|}{\sum_k L_k} = 1 - \frac{\varphi(1) - \varphi(0)}{\pi/2} = \frac{s_{\mathrm{off}}}{1 + s_{\mathrm{off}}} \approx 0.79\% \tag{D.7}$$

for $s_{\mathrm{off}} = 0.008$.

For the primary ϵ-prediction setting, the midpoint risk is

$$r_k = \left(1 - \bar{\alpha}_{\tau_k^\circ}\right) \widehat{\mathbb{E}}\left[\left\|\epsilon_\theta(x_{\tau_k^\circ}, \tau_k^\circ) - \epsilon\right\|^2\right]. \tag{D.8}$$

For the v-prediction robustness setting (Salimans & Ho, 2022), let $t_{\mathrm{zsnr}}(\tau)$ denote the interpolated model-time coordinate under the zero-terminal-SNR-rescaled scheduler, defined by

$$\bar{\alpha}^{\mathrm{zsnr}}\big(t_{\mathrm{zsnr}}(\tau)\big) = \bar{\alpha}(\tau). \tag{D.9}$$

With

$$\begin{aligned} x_\tau &= \sqrt{\bar{\alpha}(\tau)}\, x_0 + \sqrt{1-\bar{\alpha}(\tau)}\, \epsilon, \\ v_\tau &= \sqrt{\bar{\alpha}(\tau)}\, \epsilon - \sqrt{1-\bar{\alpha}(\tau)}\, x_0, \end{aligned} \tag{D.10}$$

the midpoint risk is

$$r_k = \bar{\alpha}(\tau_k^\circ)\big(1-\bar{\alpha}(\tau_k^\circ)\big)\widehat{\mathbb{E}}\left[\left\| v_\theta\big(x_{\tau_k^\circ}, t_{\mathrm{zsnr}}(\tau_k^\circ)\big) - v_{\tau_k^\circ}\right\|^2\right]. \tag{D.11}$$

After constructing the model-aware traversal, we realize it on the native $T = 1000$ diffusion grid. For model-time boundaries $t_j = j/T$, let

$$\tau_j = \Phi^{-1}(t_j), \qquad \varphi_j = \varphi(\tau_j). \tag{D.12}$$

The discrete diffusion coefficients are obtained directly from consecutive phase values:

$$\alpha_j = \frac{\cos^2 \varphi_j}{\cos^2 \varphi_{j-1}}, \qquad \beta_j = 1-\alpha_j, \qquad j = 1, \dots, T, \tag{D.13}$$

with the standard 0.999 upper cap on β_j.

Both the baseline and model-aware v-prediction runs additionally apply the same zero-terminal-SNR rescaling before the final β_j cap. For a discrete diffusion schedule, let $q_j = \sqrt{\bar{\alpha}_j}$. We rescale

$$q_j^{\mathrm{zsnr}} = q_{\mathrm{first}} \frac{q_j - q_{\mathrm{last}}}{q_{\mathrm{first}} - q_{\mathrm{last}}}, \qquad \bar{\alpha}_j^{\mathrm{zsnr}} = \big(q_j^{\mathrm{zsnr}}\big)^2. \tag{D.14}$$

The per-step coefficients are reconstructed from the rescaled cumulative coefficients, with the resulting β_j values capped at 0.999 for numerical stability. Because this cap is applied after rescaling, the realized terminal SNR is small but nonzero: 2.278×10^{-9} for the baseline and 8.512×10^{-10} for the model-aware schedule with $\lambda = 220$. For v-prediction evaluation, both baseline and model-aware runs use trailing timestep spacing (Lin et al., 2024).

In both prediction settings, training uses the resulting fixed diffusion coefficients, with discrete model-time indices sampled uniformly as in the corresponding baseline.

Flow matching. For flow matching, we keep the Cond-OT coefficient curve fixed and use its standard parameterization as the baseline:

$$\gamma_0(\tau) = (1-\tau, \tau), \qquad x_\tau = (1-\tau)x_0 + \tau\epsilon. \tag{D.15}$$

For velocity prediction $v_\tau = \epsilon - x_0$, the midpoint risk is

$$r_k = {\tau_k^\circ}^2 (1-\tau_k^\circ)^2 \widehat{\mathbb{E}}\left[\left\| v_\theta(x_{\tau_k^\circ}, \tau_k^\circ) - (\epsilon - x_0)\right\|^2\right]. \tag{D.16}$$

For the straight Cond-OT coefficient curve, the interval coefficient-space arc length is

$$L_k = \sqrt{2}\,(\tau_{k+1} - \tau_k). \tag{D.17}$$

We use the normalized Euclidean coefficient-space convention of the main text. On the fixed Cond-OT curve, the pathwise action is proportional to $\mathcal{J}_{\mathrm{kin}}$; this constant factor is absorbed into the kinetic-risk tradeoff convention (Appendix A.1).

The resulting interval allocations define a piecewise-linear reparameterization $t = \Phi(\tau)$. During model-aware training, model time t is drawn from the same training-time sampling distribution as in the corresponding baseline. To construct the state and velocity target, we map t to

$$\tau = \Phi^{-1}(t). \tag{D.18}$$

Although the training-time sampling distribution in model time t is unchanged, the reparameterization generally changes the induced distribution over the reference coordinate τ. The state and velocity target are then

$$x_t = (1-\tau)x_0 + \tau\epsilon, \qquad \frac{\mathrm{d}x_t}{\mathrm{d}t} = \frac{\epsilon - x_0}{w(\tau)}, \qquad w(\tau) = \Phi'(\tau). \tag{D.19}$$

Hence reparameterization changes both the state-time map and the corresponding velocity target. Within each comparison between the baseline and model-aware schedules, the endpoint coupling and training-time sampling distribution are held fixed; only the traversal of the prescribed coefficient curve is changed.

The Monte Carlo budgets and computational cost of the one-time risk-profile estimation are reported in Appendix E.7.

D.2 Architectures, Training, and Evaluation Details

Model architectures. The primary experiments use U-Net backbones (Ronneberger et al., 2015). For CIFAR-10 DDPM, we use a Diffusers-style unconditional U-Net (von Platen et al., 2022) with four resolution levels, channel widths $(128, 128, 256, 256)$, and two layers per block. The architecture-family control replaces this backbone with the CIFAR-10 U-ViT-S/2 architecture (Bao et al., 2023): 32×32 inputs, patch size 2, embedding dimension 512, 13 Transformer blocks (`depth=12` in the implementation), and 8 attention heads. It is unconditional and uses the same ϵ-prediction target. For CIFAR-10 flow matching, we follow the TorchCFM CIFAR-10 image-generation configuration, with 128 base channels, two residual blocks per resolution, channel multipliers $(1, 2, 2, 2)$, and four-head attention at 16×16 resolution (Tong et al., 2024).

For ImageNet-64, DDPM and flow matching use the same unconditional U-Net architecture at 64×64 resolution, with channel widths $(192, 384, 384, 768)$ and two layers per block. Attention is used at 32×32, 16×16, and 8×8 resolutions, with attention-head dimension 8; we use 32 GroupNorm groups and no dropout.

Optimization. All DDPM training runs use $T = 1000$ diffusion steps. The CIFAR-10 DDPM U-Net and U-ViT runs use the same optimization recipe: AdamW with batch size 128, learning rate 2×10^{-4}, weight decay 10^{-4}, and an exponential moving average (EMA) with decay 0.9999. The CIFAR-10 flow-matching models use Adam with batch size 128, learning rate 2×10^{-4}, zero weight decay, 5000 warmup steps, gradient clipping at 1.0, and EMA decay 0.9999. For CIFAR-10, one epoch corresponds to 390 optimizer updates.

On ImageNet-64, both model families use AdamW with learning rate 10^{-4}, zero weight decay, EMA decay 0.9999, effective batch size 128, and mixed-precision training. The DDPM runs use two GPUs with per-GPU batch size 32 and two gradient-accumulation steps. The flow-matching runs instead use two GPUs with per-GPU batch size 64 and no gradient accumulation, so each minibatch-OT coupling is formed within a 64-sample per-GPU batch.

Evaluation. All reported FIDs use 50,000 generated samples. FID differences, relative reductions, and gain-recovery percentages are computed from unrounded FIDs. Within each comparison between the baseline and model-aware schedules, we match the evaluation protocol, generated-sample seed, sampler or integrator, and NFE budget. Unless otherwise specified, single-seed FID evaluations use DPM++3M for DDPM and midpoint integration for flow matching, both at 16 NFE. For CIFAR-10, FID is computed against the full training split. For ImageNet-64, we use the training split of the Downsampled ImageNet 64×64 dataset distributed through Academic Torrents and compute FID against its full validation split. Exact epochs or update counts are stated with the corresponding results.

D.3 Pretrained Conditional Latent-Model Diagnostics

Conditional latent-state extension. In this subsection, x denotes the checkpoint-native latent state rather than a pixel-space observation, and c denotes a class label or text prompt. We allow an arbitrary joint endpoint law

$$(x_0, \epsilon, c) \sim \rho, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{D.20}$$

and average the fiberwise risk over this joint law. For a conditional predictor, we regard (x, c) as the fixed base variable, replace the conditional fiber laws μ_t^x and $\nu_t^{\theta,x}$ by $\mu_t^{x,c}$ and $\nu_t^{\theta,x,c}$, and otherwise retain the same decomposition fiber $\mathcal{F}_x$. The identities of Section 4 therefore remain unchanged. Each checkpoint is evaluated in its own native latent representation; no coordinate-wise alignment between the DiT and 2-RF latent spaces is assumed.

DiT-XL/2. We use the public `facebook/DiT-XL-2-256` checkpoint, a 675M class-conditional latent diffusion transformer on ImageNet-256 (Peebles & Xie, 2023). We select 64 examples from a fixed seed-42 shuffle of an ImageNet-1K validation subset and apply the center-cropping procedure used in DiT at 256×256 (Deng et al., 2009; Dhariwal & Nichol, 2021). We then draw one posterior sample per example using the checkpoint's variational autoencoder (VAE) and apply its native scaling factor. Each latent is paired with independent Gaussian noise and its class label. For risk estimation, we query the checkpoint at 100 distinct diffusion timesteps $t(\tau)$. These are selected so that their reference coordinates are nearest to uniformly spaced midpoints in the same practical VP reference coordinate used in the other diffusion comparisons. The checkpoint is evaluated without classifier-free guidance; from its eight output channels we retain only the four ϵ-prediction channels and exclude the learned-variance channels. The corresponding fiberwise risk is

$$R_{\mathrm{DiT}}(\tau) = \big(1 - \bar{\alpha}_{t(\tau)}\big)\mathbb{E}\,\|\epsilon_\theta(x_\tau, t(\tau), c) - \epsilon\|^2 . \tag{D.21}$$

InstaFlow 2-RF. We use the public pre-distillation checkpoint `XCLiu/2_rectified_flow_from_sd_1_5`, a 0.9B text-conditional latent 2-Rectified Flow model (Liu et al., 2023; 2024). To construct a public model-induced coupling, we select 64 nonempty prompts from a fixed seed-42 shuffle of the COCO 2017 training-caption stream and sample one Gaussian noise latent per prompt. Using the same prompt and noise, the Stable Diffusion 1.5 teacher produces the paired endpoint with a second-order DPM-Solver++ sampler at 25 steps and guidance scale 5.0. Both endpoints remain in the native Stable Diffusion latent scale, without VAE decoding or re-encoding. This defines a reconstructed reflow coupling induced by the teacher under a public COCO prompt distribution. It differs from the original InstaFlow training coupling and the minibatch OT endpoint coupling used in our OT-CFM experiments (Lin et al., 2014; Rombach et al., 2022; Lu et al., 2025; Ho & Salimans, 2022).

We evaluate the released 2-RF predictor with its own text encoder and the raw text condition, without classifier-free guidance. In the common data-to-noise coordinate,

$$x_\tau = (1-\tau)x_0 + \tau\epsilon, \qquad \tau = 0 \text{ at the teacher endpoint}, \quad \tau = 1 \text{ at noise}. \tag{D.22}$$

The official model timestep is 1000τ: it equals 1000 at the noise endpoint and decreases toward 0 along the native noise-to-data sampling direction. The released predictor uses the corresponding native velocity orientation $x_0 - \epsilon$, so the data-to-noise fiberwise risk can be evaluated equivalently as

$$R_{\mathrm{2RF}}(\tau) = \tau^2(1-\tau)^2\mathbb{E}\,\|v_\theta(x_\tau, 1000\tau, c) - (x_0 - \epsilon)\|^2 . \tag{D.23}$$

Flipping both the prediction and target gives the identical squared residual in the data-to-noise velocity orientation.

Estimation and allocation comparison. Each profile uses the same 64 fixed endpoint pairs at all 100 reference locations. Nested subsample comparisons at $N \in \{16, 32, 64\}$ show that the reported shape metrics have stabilized by $N = 64$. Per-sample squared residuals are averaged over latent coordinates. We compare normalized risk profiles and allocation densities using the metrics in Appendix E.10. Allocation densities are constructed from the raw risks, using the previously fixed family values $\lambda = 220$ for DiT and $\lambda = 450$ for 2-RF. For DiT, we additionally apply the affine risk-recalibration property of Appendix C.3 to diagnose the observed absolute-scale shift.

E Additional Experimental Results and Diagnostics

E.1 Complete CIFAR-10 DDPM Sampler-NFE Grid

Table 5 reports the complete sampler-NFE grid underlying the main CIFAR-10 DDPM ϵ-prediction results. The main text uses DPM++3M as the primary sampler at 16, 32, and 64 NFE and reports DDIM and DPM++2M at 16 NFE as cross-sampler checks. Here we include the remaining DDIM and DPM++2M budgets for completeness (Song et al., 2021; Lu et al., 2025). The model-aware schedule improves for all three paired seeds in every one of the nine configurations, with mean relative FID reductions ranging from 9.2% to 16.4%.

Table 5: CIFAR-10 DDPM FID on the complete sampler-NFE grid (epoch 400; three paired seeds, mean $\pm$ standard deviation).

Sampler	NFE	**Baseline** FID ↓	**Model-aware (Ours)** FID ↓	ΔFID ↑	%Impr. ↑
DDIM	16	11.27 ± 0.19	9.98 ± 0.23	1.29 ± 0.20	$11.4 \pm 1.7\%$
DDIM	32	8.14 ± 0.23	7.39 ± 0.23	0.75 ± 0.14	$9.2 \pm 1.7\%$
DDIM	64	7.27 ± 0.23	6.44 ± 0.21	0.83 ± 0.07	$11.4 \pm 0.8\%$
DPM++2M	16	9.81 ± 0.26	8.29 ± 0.14	1.51 ± 0.14	$15.4 \pm 1.1\%$
DPM++2M	32	7.90 ± 0.27	7.00 ± 0.25	0.90 ± 0.12	$11.4 \pm 1.5\%$
DPM++2M	64	6.97 ± 0.24	6.30 ± 0.27	0.67 ± 0.07	$9.7 \pm 1.1\%$
DPM++3M	16	9.64 ± 0.26	8.06 ± 0.16	1.58 ± 0.11	$\mathbf{16.4 \pm 0.7}\%$
DPM++3M	32	7.69 ± 0.26	6.79 ± 0.27	0.91 ± 0.12	$11.8 \pm 1.5\%$
DPM++3M	64	6.83 ± 0.24	6.19 ± 0.27	0.64 ± 0.07	$9.3 \pm 1.2\%$

Table 6: CIFAR-10 DDPM FID for the cosine and linear schedules at epoch 400. Single-seed results.

Sampler	NFE	FID (Cosine) ↓	FID (Linear) ↓
DDIM	16	11.08	13.77
DDIM	32	7.96	10.30
DDIM	64	7.05	8.66
DPM++2M	16	9.60	13.93
DPM++2M	32	7.73	11.34
DPM++2M	64	6.78	9.59
DPM++3M	16	9.43	13.54
DPM++3M	32	7.52	11.10
DPM++3M	64	6.64	9.36

E.2 Linear-β DDPM Control

Table 6 reports results for the linear-β DDPM control used in Section 6.2. The β_t values are linearly spaced from 10^{-4} to 2×10^{-2}. Architecture, optimizer, batch size, training horizon, data pipeline, and all other training settings are identical to the cosine-DDPM baseline; only the diffusion schedule is changed. All results use epoch-400 EMA weights.

The linear schedule is consistently weaker than cosine across all nine sampler-NFE configurations, supporting our use of cosine as the primary DDPM baseline in the main text.

E.3 DDPM v-Prediction Results

We additionally evaluate the model-aware schedule under v-prediction to test prediction-target robustness within DDPM and provide an additional instance of the general linear-target formulation. Both baseline and model-aware runs use matched zero-terminal-SNR rescaling, and evaluation uses trailing timestep spacing. We retain the same $\lambda = 220$ used for ϵ-prediction. We report DPM-Solver++ multistep results. In the Hugging Face Diffusers `DDIMScheduler` implementation used in our experiments (von Platen et al., 2022), DDIM updates with trailing timestep spacing do not always end at the next grid timestep at these NFE budgets. We therefore omit DDIM from this robustness table.

Table 7 shows that the model-aware schedule improves all six reported DPM-Solver++ configurations across both second- and third-order multistep solvers, with relative FID reductions from 7.4% to 14.7%. The gains are therefore preserved under a different prediction target.

E.4 CIFAR-10 U-ViT Architecture-Family Transfer

We evaluate architecture-family transfer by replacing the CIFAR-10 DDPM U-Net with U-ViT-S/2. We construct the model-aware schedule from the U-ViT baseline risk profile with the same $\lambda = 220$,

Table 7: CIFAR-10 DDPM v-prediction FID at epoch 400 under matched zero-terminal-SNR rescaling and trailing timestep spacing. Single-seed results.

Sampler	NFE	**Baseline** FID ↓	**Model-aware (Ours)** FID ↓	ΔFID ↑	%Impr. ↑
DPM++2M	16	11.60	9.94	1.66	14.3%
DPM++2M	32	9.24	8.27	0.97	10.5%
DPM++2M	64	8.03	7.40	0.63	7.9%
DPM++3M	16	11.31	9.65	1.66	**14.7%**
DPM++3M	32	8.98	8.02	0.96	10.7%
DPM++3M	64	7.85	7.27	0.58	7.4%

Table 8: CIFAR-10 U-ViT-S/2 architecture-transfer FID at epoch 600. Single-seed results.

Sampler	NFE	**Baseline** FID ↓	**Model-aware (Ours)** FID ↓	ΔFID ↑	%Impr. ↑
DDIM	16	15.38	14.17	1.21	7.8%
DPM++2M	16	13.50	12.05	1.46	10.8%
DPM++3M	16	13.18	11.57	1.61	**12.2%**
DPM++3M	32	10.52	9.70	0.83	7.9%
DPM++3M	64	9.17	8.69	0.48	5.2%

without architecture-specific retuning, and then hold it fixed. Both baseline and model-aware FIDs are evaluated at 600 training epochs using the same sampler-NFE configurations as the main CIFAR-10 DDPM results.

Table 8 shows improvements in all five reported sampler-NFE configurations, including a 12.2% relative FID reduction with DPM++3M at 16 NFE. Thus the model-aware improvement persists after changing the architecture family without architecture-specific retuning.

E.5 Flow-Matching Coupling and Training-Time Sampling Robustness

We test whether the model-aware improvement persists under changes to endpoint coupling and the training-time sampling distribution. Within each matched 400-epoch setting, the baseline and model-aware runs use the same coupling and training-time sampling distribution. We retain $\lambda = 450$ throughout without retuning as a robustness test.

The logit-normal control uses $t = \mathrm{sigmoid}(\xi) = 1/(1 + e^{-\xi})$ with $\xi \sim \mathcal{N}(0, 1)$, whereas the RF++ U-shaped control uses the training-time sampling density $p_{\mathrm{time}}(t) \propto e^{4t} + e^{4(1-t)}$ on $t \in [0, 1]$. No additional clipping is applied to the sampled model times (Esser et al., 2024; Lee et al., 2024).

Table 9 shows improvements in all four settings, with relative FID reductions from 31.3% to 39.4%. The gains persist under independent endpoint coupling and alternative training-time sampling distributions, while preserving the baseline ordering across the tested settings.

E.6 Training-Stage and Tradeoff-Weight Sensitivity

We selected the operating values using preliminary single-seed, 10,000-sample FID sweeps in the two primary CIFAR-10 settings, before the cross-system risk-allocation analysis. For DDPM, the sweep used 16-step DDIM at epoch 320 and considered $\lambda \in \{155, 195, 210, 220, 230, 250\}$. For flow matching, the sweep used 16-NFE midpoint integration at epoch 400 and considered $\lambda \in \{100, 420, 430, 450, 460, 470, 480, 490, 500\}$. Among the tested values, $\lambda = 220$ and $\lambda = 450$ gave the lowest FID for DDPM and flow matching, respectively.

We further examine how generation quality varies with training stage and tradeoff weight. For the tradeoff-weight comparisons, we use representative nonzero λ values from the preliminary sweeps. Figure 3 summarizes the corresponding CIFAR-10 results for both DDPM and flow matching.

For DDPM ϵ-prediction, the model-aware schedule consistently improves over the cosine baseline from epochs 280 through 400. At epoch 320, all tested nonzero values $\lambda \in \{155, 220, 250\}$ also improve over cosine, although the magnitude of the gain varies across the tested weights.

Table 9: CIFAR-10 flow-matching FID under changes in endpoint coupling and the training-time sampling distribution, using matched 400-epoch settings. Single-seed results.

Variation	Setting	**Baseline** FID ↓	**Model-aware (Ours)** FID ↓	%Impr. ↑
Reference	OT-CFM	8.60	5.57	35.2%
Coupling	Independent-CFM	8.82	6.06	31.3%
Time sampling	Logit-normal	8.27	5.01	**39.4%**
Time sampling	U-shaped (RF++)	8.97	6.12	31.8%

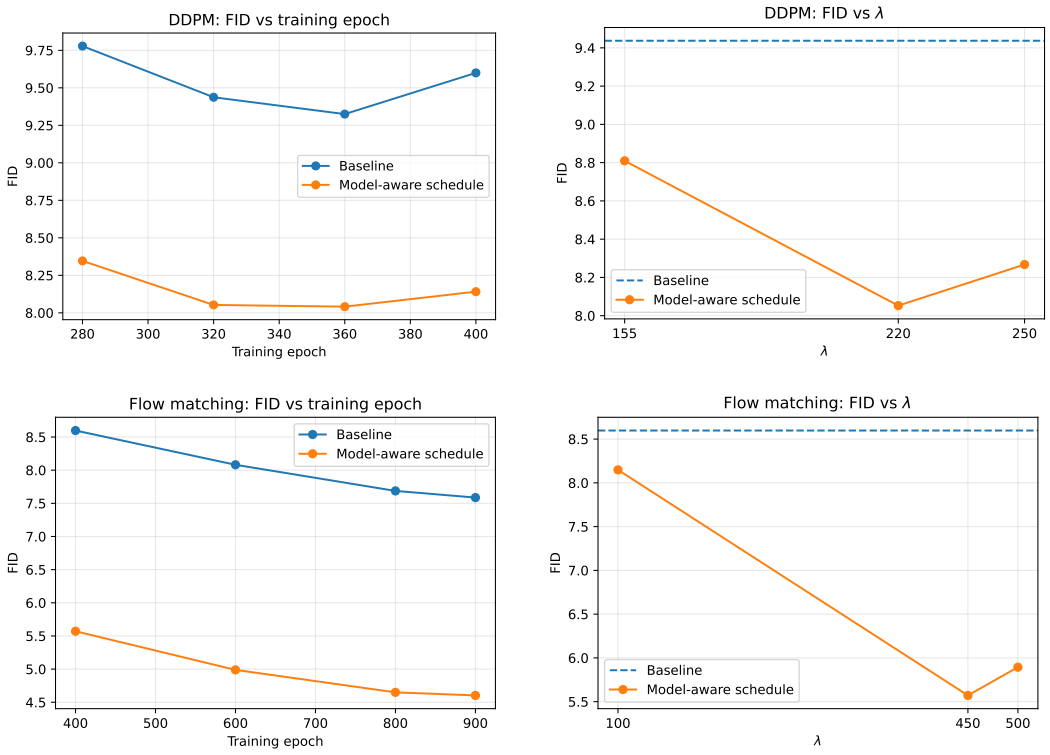


Figure 3: FID ↓ sensitivity to training stage and λ on CIFAR-10. From top left to bottom right: DDPM training stage, DDPM λ, flow-matching training stage, and flow-matching λ. DDPM uses ϵ-prediction with DPM++3M at 16 NFE; flow matching uses midpoint at 16 NFE. The λ panels show only tested nonzero values; dashed lines denote the corresponding baselines, and connecting lines are visual guides.

For flow matching, the model-aware schedule yields an approximately 3-FID improvement throughout the tested range from epochs 400 to 900. At epoch 400, all tested nonzero values $\lambda \in \{100, 450, 500\}$ improve over the standard Cond-OT parameterization. Together, these results show that the observed gains are not tied to a single training stage or an isolated nonzero choice of λ.

E.7 Risk-Profile Checkpoint Sensitivity

Tables 10 and 11 give the complete results for the early-checkpoint studies summarized in the main text. In both model families, using the risk profile estimated at epoch 100 retains essentially the full improvement obtained with the later estimate, changing the final FID by only 0.02 for DDPM and 0.06 for flow matching.

The slightly lower DDPM FID obtained with the epoch-100 estimate is not interpreted as evidence that earlier checkpoints are systematically preferable; rather, these results indicate that the estimated risk profile is sufficiently stable across substantially different training stages for the resulting

Table 10: DDPM FID sensitivity to the checkpoint used to estimate the risk profile. FIDs are evaluated at epoch 400. Single-seed results.

Schedule	FID ↓	ΔFID ↑
Cosine baseline	9.43	–
Model-aware, epoch 320 estimate	7.90	1.52
Model-aware, epoch 100 estimate	7.88	1.55

Table 11: Flow-matching FID sensitivity to the checkpoint used to estimate the risk profile in the 400-epoch reference OT-CFM setting. Single-seed results.

Schedule	FID ↓	ΔFID ↑
Baseline	8.60	–
Model-aware, epoch 400 estimate	5.57	3.03
Model-aware, epoch 100 estimate	5.63	2.96

schedule to remain effective. Table 13 further supports the stability of normalized risk shapes and allocation deformations across the tested risk-estimation checkpoints.

Risk-estimation cost. We use $12{,}800$ Monte Carlo samples per reference interval: 199 intervals for DDPM (2.55M sample-level forward evaluations) and 200 for flow matching (2.56M). The same respective budgets are used on CIFAR-10 and ImageNet-64. On ImageNet-64, a full DDPM run at 800k updates with effective batch size 128 processes 102.4M training sample instances, while a 600k-update flow-matching run processes 76.8M. Thus risk estimation amounts to only about 2.5% and 3.3% of these sample counts, respectively, and consists only of forward evaluations. This estimation cost is incurred only once.

E.8 Flow-Matching Finite-Step Integration Diagnostic

Table 12 provides the complete finite-step integration diagnostic underlying the discussion in Section 6.3. We compare several integrators over increasing NFE budgets to separate schedule effects from numerical behavior specific to coarse integration grids.

The unusually poor low-NFE behavior of standard Heun2 is closely tied to its terminal-stage evaluation. Replacing only the final Heun2 step with a midpoint step reduces the baseline FID from 52.97 to 7.00 at 16 NFE and from 20.14 to 5.20 at 32 NFE. The discrepancy also shrinks substantially under grid refinement: the baseline gap between standard and last-midpoint Heun2 decreases from 45.97 FID at 16 NFE to 1.77 FID at 64 NFE.

RK4 exhibits a related dependence on grid resolution, with its baseline FID improving from 30.95 at 16 NFE to 3.73 at 64 NFE. The latter is lower than the corresponding midpoint baseline of 4.45, and the model-aware schedule further reduces it to 3.52. Euler instead yields higher FID under the model-aware schedule at all three CIFAR-10 budgets. For comparison, midpoint and Heun2 (last-mid.) yield lower FID than Euler at all three budgets under both the baseline and model-aware schedules. On ImageNet-64, however, the model-aware schedule improves Euler FID at all three tested budgets (Table 4). Thus the degradation observed on CIFAR-10 is setting-specific rather than a systematic incompatibility with first-order integration.

E.9 ImageNet-64 Flow-Matching Qualitative Samples

Figure 4 complements the 16-NFE midpoint result in Table 4. With the same 64 fixed sampling seeds, the model-aware grid exhibits more coherent object and scene structure, consistent with the FID reduction from 41.52 to 30.11.

Table 12: Finite-step integration diagnostic for CIFAR-10 flow matching at epoch 900. Each entry reports Baseline → Model-aware FID ↓ (relative improvement ↑). Single-seed results. Heun2 (last-mid.) replaces the final Heun2 step with a midpoint step to diagnose low-NFE terminal-stage sensitivity.

Integrator	16 NFE	32 NFE	64 NFE
Midpoint	$7.59 \to 4.60$ (39.3%)	$5.52 \to 4.63$ (16.1%)	$4.45 \to 4.04$ (9.2%)
Heun2	$52.97 \to 33.66$ (36.5%)	$20.14 \to 9.95$ (50.6%)	$6.04 \to 3.76$ (37.8%)
Heun2 (last-mid.)	$7.00 \to 4.44$ (36.5%)	$5.20 \to 4.07$ (21.8%)	$4.27 \to 3.93$ (8.0%)
RK4	$30.95 \to 12.84$ (58.5%)	$9.55 \to 4.43$ (53.6%)	$\mathbf{3.73} \to \mathbf{3.52}$ (5.5%)
Euler	$9.53 \to 9.98$ (-4.7%)	$6.98 \to 7.90$ (-13.2%)	$5.47 \to 5.80$ (-6.0%)

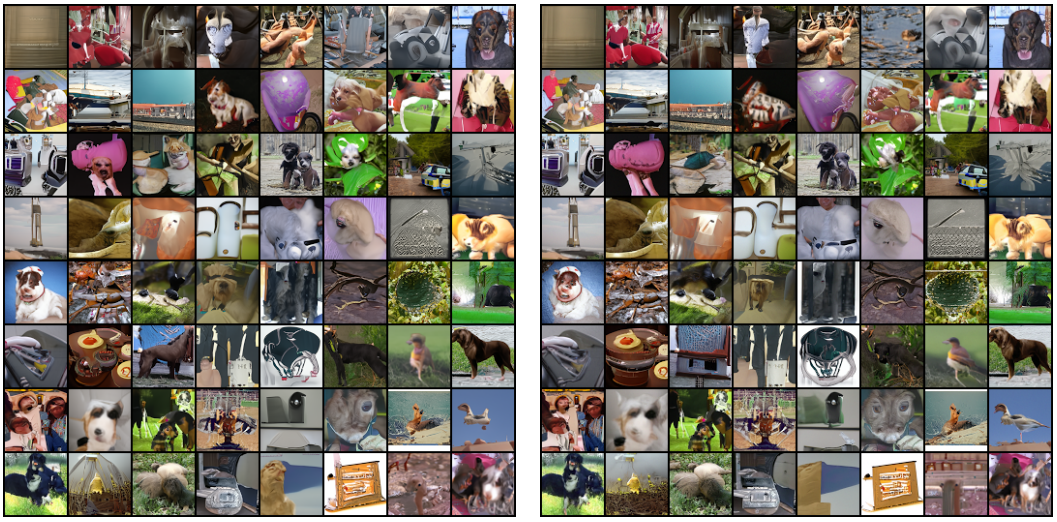


Figure 4: Uncurated samples from ImageNet-64 OT-CFM models at the 600k checkpoint and 16 NFE using midpoint integration. **Left:** standard Cond-OT schedule. **Right:** our model-aware schedule with $\lambda = 450$. Both grids use the same 64 fixed sampling seeds.

E.10 Risk-Profile and Allocation-Density Agreement

This appendix quantifies and extends the risk-profile and allocation-density agreement reported in Section 6.4.

Comparison protocol and metrics. Figure 5 shows representative CIFAR-10 fiberwise-risk profiles in their respective kinetic reference coordinates. Their absolute magnitudes differ substantially across systems, motivating a shape comparison after normalization.

For cross-system shape comparison, we normalize each risk profile to unit area,

$$\bar{R}(\tau) = \frac{R(\tau)}{\int_0^1 R(u)\,\mathrm{d}u}. \tag{E.1}$$

Model-aware allocation densities are constructed from the raw risk profiles with the corresponding λ values, not from $\bar{R}$.

To characterize the agreement reported in Section 6.4, we fit the reference profiles using the epoch-320 DDPM ϵ-prediction and epoch-400 OT-CFM baseline risk profiles on CIFAR-10, together with their corresponding model-aware allocation densities. The exponent $5/4$ and coefficient 4 are fitted over simple candidate values. Neither parameter is theoretically distinguished; the fitted profiles are empirical approximations rather than uniquely identified parametric laws. The risk-shape reference and analytic template remain fixed in all subsequent comparisons.

For both normalized-risk and allocation-density comparisons, let f denote the empirical profile normalized to unit area and g the corresponding frozen reference profile. We report Pearson correlation

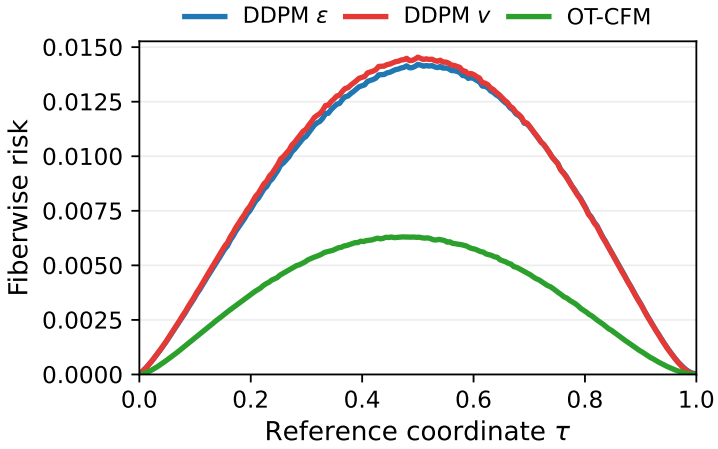


Figure 5: Raw fiberwise-risk profiles for representative CIFAR-10 settings in their respective reference coordinates. Absolute risk magnitudes differ substantially across systems.

together with

$$E_{L^2}^{\mathrm{rel}}(f,g)=\frac{\|f-g\|_{L^2([0,1])}}{\|f\|_{L^2([0,1])}},\qquad \mathrm{TV}(f,g)=\frac{1}{2}\|f-g\|_{L^1([0,1])}. \tag{E.2}$$

For risk profiles, $f=\bar{R}$ and g is the normalized $\sin^{5/4}(\pi\tau)$ reference; for allocation densities, $f=w$ and $g=w_{\mathrm{ana}}$.

To quantify mirror asymmetry independently of either frozen reference profile, we additionally report

$$A_{\mathrm{TV}}(f)=\frac{1}{2}\int_0^1 |f(\tau)-f(1-\tau)|\,\mathrm{d}\tau. \tag{E.3}$$

This quantity vanishes for a profile symmetric about $\tau=1/2$.

Agreement across models and settings. Table 13 compares the normalized risk profiles and allocation densities with the frozen reference profiles. In our CIFAR-10 DDPM experiments, both profiles retain close agreement with their respective references across prediction targets, risk-estimation checkpoints, and the U-Net and U-ViT architectures. For flow matching, both findings persist across endpoint couplings, training-time sampling distributions, and risk-estimation checkpoints. The ImageNet-64 experiments extend this agreement to a second dataset for both model families.

Pretrained-checkpoint diagnostics. The DiT-XL/2 and InstaFlow 2-RF checkpoints evaluated in Appendix D.3 also retain close normalized-risk agreement with the risk-shape reference. These are diagnostics without retraining or FID evaluation under a modified schedule. Under the raw-risk allocation construction, the previously fixed flow-matching value $\lambda=450$ also yields a 2-RF allocation with high correlation to the frozen analytic template, whereas the fixed diffusion value $\lambda=220$ does not transfer directly to DiT (Table 13).

Fitting the DiT raw risk by an affine rescaling of the $\sin^{5/4}(\pi\tau)$ shape and applying the risk-recalibration property of Appendix C.3 predicts $\lambda=34.7$. At this value, the DiT allocation attains Pearson 0.9861, relative L^2 error 4.40%, TV 1.44%, and $A_{\mathrm{TV}}=2.04\%$. The predicted value is close to the one-dimensional TV- and relative-L^2-optimal values, 35.3 and 37.6, respectively. Thus the DiT discrepancy is primarily an absolute risk-scale shift rather than a failure of the normalized shape.

DDPM coordinate controls. The linear-β DDPM control compares the same pointwise risk values in native model time and in the practical VP reference coordinate. We match each risk value to the cosine-baseline time with the same coefficient pair. Unit-area normalization is performed separately in the two coordinates. A separate check examines whether the practical approximation to

Table 13: Agreement with the frozen risk-shape reference and frozen analytic allocation template, together with mirror asymmetry A_{TV}. CIFAR-10 DDPM uses the U-Net backbone unless U-ViT-S/2 is specified; unqualified OT-CFM uses uniform training-time sampling. Allocation densities use $\lambda = 220$ for diffusion and $\lambda = 450$ for flow matching unless otherwise specified. The DiT $\lambda = 34.7$ row uses the same risk profile as the preceding row. $E_{L^2}^{\mathrm{rel}}$, TV, and A_{TV} are reported in percent.

	Normalized risk				Allocation density			
Setting	Pearson ↑	$E_{L^2}^{\mathrm{rel}}$ ↓	TV ↓	A_{TV}	Pearson ↑	$E_{L^2}^{\mathrm{rel}}$ ↓	TV ↓	A_{TV}
CIFAR-10 DDPM ϵ, ep. 320	0.9991	2.20	1.12	1.43	0.9988	1.98	0.74	0.72
CIFAR-10 DDPM ϵ, ep. 100	0.9993	3.18	1.51	1.40	0.9969	2.16	0.66	0.99
CIFAR-10 DDPM v	0.9994	1.72	0.85	0.96	0.9985	1.67	0.62	0.72
CIFAR-10 DDPM ϵ, U-ViT-S/2	0.9995	1.61	0.82	1.22	0.9988	1.61	0.58	0.69
CIFAR-10 OT-CFM, ep. 400	0.9917	6.39	3.22	6.41	0.9909	3.78	1.50	2.91
CIFAR-10 OT-CFM, ep. 100	0.9905	6.98	3.54	7.01	0.9897	5.48	2.24	3.37
CIFAR-10 OT-CFM, logit-normal	0.9906	6.69	3.40	6.70	0.9901	3.73	1.51	3.00
CIFAR-10 OT-CFM, U-shaped (RF++)	0.9921	6.34	3.19	6.33	0.9912	3.90	1.55	2.90
CIFAR-10 Independent-CFM	0.9989	2.25	1.08	1.42	0.9953	7.12	2.67	0.82
ImageNet-64 DDPM ϵ	0.9948	4.92	2.54	5.02	0.9944	2.74	1.19	2.17
ImageNet-64 OT-CFM	0.9878	7.78	3.96	7.86	0.9888	6.94	2.84	3.70
Pretrained conditional latent-model diagnostics								
DiT-XL/2, ImageNet-256, $\lambda = 220$	0.9930	8.70	4.37	5.19	0.5223	88.00	36.18	26.43
DiT-XL/2, $\lambda = 34.7$	–	–	–	–	0.9861	4.40	1.44	2.04
InstaFlow 2-RF, COCO 2017, $\lambda = 450$	0.9835	9.92	4.52	6.34	0.9812	21.35	8.68	0.58

Table 14: CIFAR-10 DDPM risk-shape coordinate controls. "Practical VP reference" and "exact VP reference" denote, respectively, the practical approximation to and exact form of the VP kinetic reference coordinate. For the direct linear-cosine comparison, $E_{L^2}^{\mathrm{rel}}$ uses the L^2 norm of the cosine-DDPM profile as its denominator. $E_{L^2}^{\mathrm{rel}}$ and TV are reported in percent.

Comparison	Coordinate	Pearson ↑	$E_{L^2}^{\mathrm{rel}}$ ↓	TV ↓
Cosine DDPM vs. risk-shape reference	Practical VP reference	0.9991	2.20	1.12
Cosine DDPM vs. risk-shape reference	Exact VP reference	0.9988	2.37	1.10
Linear DDPM vs. risk-shape reference	Native model time	0.3616	64.38	38.00
Linear DDPM vs. risk-shape reference	Practical VP reference	0.9994	1.67	0.84
Linear DDPM vs. risk-shape reference	Exact VP reference	0.9995	1.64	0.75
Linear vs. cosine DDPM	Practical VP reference	0.9998	0.93	0.45

the VP kinetic reference coordinate affects the shape comparisons. We therefore also recompute the cosine- and linear-DDPM profiles in the exact VP reference coordinate defined in Appendix A.2. Table 14 reports both coordinate controls.

The direct linear-cosine comparison is tight, confirming that the two normalized risk profiles closely agree with each other rather than only agreeing separately with the risk-shape reference. Using the exact rather than the practical VP reference coordinate changes the cosine- and linear-DDPM comparisons with the risk-shape reference only marginally. Together with the coordinate-level audit in Appendix A.2, this result supports using the practical VP reference as a close approximation to the exact VP kinetic reference coordinate in the experiments.

Post-training allocation stability. We also test whether the shared normalized risk shape persists after model-aware training and whether re-estimating risk changes the one-shot allocation. Let R_0 denote the baseline-checkpoint risk used to construct the one-shot allocation w_1. After training with w_1, we estimate R_1 from the resulting model-aware checkpoint and compute the diagnostic allocation w_2 from the same closed-form rule with the original λ. We compute R_1 using the general-target expression in Section 4.2. The result is expressed in the same kinetic reference coordinate as R_0. Let $\Phi_i(\tau) = \int_0^\tau w_i(u)\,\mathrm{d}u$, $i \in \{1, 2\}$. For OT-CFM, let $v_{\theta,1}(x, t)$ denote the predictor

trained with the schedule defined by w_1. At $t = \Phi_1(\tau)$, the input state is $x_\tau = (1-\tau)x_0 + \tau\epsilon$, and the training target is $y_t = (\epsilon - x_0)/w_1(\tau)$. The corresponding general-target coefficients are $a_t = -1/w_1(\tau)$ and $b_t = 1/w_1(\tau)$, giving $\Delta_t = 1/w_1(\tau)$. Substituting these coefficients into Eq. (12) yields

$$R_1(\tau) = \tau^2(1-\tau)^2\mathbb{E}\left\|w_1(\tau)v_{\theta,1}(x_\tau, \Phi_1(\tau)) - (\epsilon - x_0)\right\|^2. \tag{E.4}$$

For DDPM, R_0 and R_1 are estimated from the epoch-320 baseline and epoch-400 model-aware checkpoints, respectively; for OT-CFM, both checkpoints are at epoch 400. Table 15 reports the absolute risk-scale ratio $(\int_0^1 R_1(\tau)\,\mathrm{d}\tau)/(\int_0^1 R_0(\tau)\,\mathrm{d}\tau)$, together with $\mathrm{TV}(\bar{R}_1, \bar{R}_0)$, $\mathrm{TV}(w_2, w_1)$, and $\|\Phi_2 - \Phi_1\|_\infty$. These metrics separate changes in risk scale from changes in normalized risk shape, and quantify the resulting changes in allocation density and the cumulative map.

Table 15: Post-training allocation stability on the two primary CIFAR-10 settings. TV values are reported in percent.

	Risk		Allocation	
Setting	Scale ratio	TV ↓	TV ↓	$\|\Phi_2 - \Phi_1\|_\infty$ ↓
DDPM ϵ-pred.	0.9988	0.161	0.051	4.57×10^{-4}
OT-CFM	1.0192	0.532	0.284	1.45×10^{-3}

The model-aware checkpoints retain normalized risk profiles close to those of the baseline checkpoints, extending the shared-shape evidence beyond baseline models. Recomputing the allocations from these post-training risks changes them only marginally, preserving the shared allocation deformation. This diagnostic supports the one-shot construction, which does not require risk re-estimation after model-aware training.

Interpreting the normalized risk shape. Across models and settings, the shared normalized risk shape reflects the combination of s_τ^2 and the corresponding noise-prediction MSE, rather than either alone; for velocity prediction, this MSE uses the induced noise prediction. In the exact VP reference coordinate, $s_\tau^2 = \sin^2(\pi\tau/2)$, whereas $s_\tau^2 = \tau^2$ for the standard Cond-OT parameterization. Both factors are monotone and largest at the noise endpoint, unlike the risk-shape reference. For $0 < \tau < 1$, dividing that reference by the respective factors yields different noise-prediction MSE shapes.

Mirror asymmetry. In our CIFAR-10 and ImageNet-64 experiments, the tested OT-CFM settings exhibit systematically stronger mirror asymmetry than DDPM and Independent-CFM: A_{TV} ranges from 6.33% to 7.86% for normalized risk and from 2.90% to 3.70% for allocation density (Table 13). For any symmetric reference g,

$$\mathrm{TV}(f, g) \geq \frac{1}{2}A_{\mathrm{TV}}(f), \tag{E.5}$$

so profile asymmetry places a direct lower bound on its discrepancy from either symmetric frozen reference profile. For normalized risk, the observed OT-CFM TV values lie close to the lower bound $A_{\mathrm{TV}}/2$; for allocation density, the same lower bound remains substantial but leaves a larger residual in some settings.

The skew also has a consistent direction across the tested OT-CFM settings. Their normalized-risk centers of mass,

$$\mathrm{COM}(f) = \int_0^1 \tau f(\tau)\,\mathrm{d}\tau, \tag{E.6}$$

lie on the data side (0.480–0.483), whereas their allocation centers of mass lie on the noise side (0.510–0.512). This reversal is consistent with the reciprocal-square-root risk-to-allocation relation, which reduces allocation where the estimated risk is larger. The model-aware allocations thus retain setting-specific asymmetry absent from the symmetric analytic template. The CIFAR-10 symmetrization control suggests that retaining this asymmetry can improve FID (Appendix E.11).

Table 16: FID and model-aware gain recovery of the frozen analytic allocation template on CIFAR-10 and ImageNet-64. All CIFAR-10 rows use epoch-400 checkpoints; the ImageNet-64 DDPM and OT-CFM rows use the 800k and 600k checkpoints, respectively. Single-seed results. Gain recovery is measured relative to the improvement from the baseline to the model-aware schedule and computed from unrounded FIDs.

Setting	**Baseline** FID ↓	**Model-aware (Ours)** FID ↓	Analytic FID ↓	G_{rec} ↑
CIFAR-10 DDPM ϵ-pred.	9.43	7.90	7.93	98.1%
CIFAR-10 DDPM v-pred.	11.31	9.65	9.68	98.0%
CIFAR-10 OT-CFM	8.60	5.57	6.30	75.8%
ImageNet-64 DDPM ϵ-pred.	25.01	23.26	23.09	109.7%
ImageNet-64 OT-CFM	41.52	30.11	32.83	76.2%

E.11 Frozen Analytic Allocation Template: Functional Validation

We next evaluate whether the frozen analytic allocation template defined in Section 6.4 is functionally useful when used directly as a schedule. Its normalized density is

$$w_{\mathrm{ana}}(\tau) \propto \left(1 + 4\sin^{5/4}(\pi\tau)\right)^{-1/2}, \qquad \int_0^1 w_{\mathrm{ana}}(\tau)\,\mathrm{d}\tau = 1. \tag{E.7}$$

Its cumulative map defines a fixed monotone reparameterization:

$$\Phi_{\mathrm{ana}}(\tau) = \int_0^\tau w_{\mathrm{ana}}(u)\,\mathrm{d}u. \tag{E.8}$$

We realize the same analytic template in the native numerical representation of each model family. Across both CIFAR-10 and ImageNet-64, the primary DDPM realization integrates w_{ana} directly at the final 1000-step diffusion resolution and applies the baseline-preserving cosine-phase construction of Appendix D.1. For both datasets, the primary flow-matching realization uses the same 200-bin piecewise-linear reparameterization as the corresponding model-aware schedule. These primary realizations are used for all functional-validation results reported in Table 16. On CIFAR-10, we additionally evaluate a 199-bin DDPM realization matched to the estimated model-aware schedule and a smooth flow-matching realization obtained from the continuous cumulative map.

To quantify how much of the model-aware improvement is retained, we define the recovered-gain fraction as

$$G_{\mathrm{rec}} = \frac{\mathrm{FID}_{\mathrm{baseline}} - \mathrm{FID}_{\mathrm{analytic}}}{\mathrm{FID}_{\mathrm{baseline}} - \mathrm{FID}_{\mathrm{model\text{-}aware}}}. \tag{E.9}$$

We leave this ratio unclipped, so $G_{\mathrm{rec}} > 1$ indicates that the analytic schedule attains a lower FID than the corresponding model-aware schedule.

Table 16 reports the CIFAR-10 and ImageNet-64 results. These results clarify that the analytic template does not supplant the model-aware formulation. For DDPM, its near-complete gain recovery shows that the shared allocation deformation accounts for most of the model-aware FID improvement. However, the analytic template is not an independently motivated heuristic schedule. The cross-system risk-allocation analysis identifies the shared normalized risk shape, while the fixed-curve formulation in Section 5.2 yields the analytic template's reciprocal-square-root dependence on risk. For flow matching, the 75.8% and 76.2% recovery fractions show that the full model-aware construction retains substantial additional gains from setting-specific risk structure.

For CIFAR-10 flow matching, we further examine whether the remaining gap to the model-aware schedule is explained primarily by the setting-specific asymmetry visible in both panels of Figure 2: in the normalized-risk profile on the left and the induced allocation density on the right. Let Δt_k denote the 200 interval allocations of the $\lambda = 450$ model-aware schedule. We symmetrize these allocations by pairwise averaging: $\Delta t_k^{\mathrm{sym}} = (\Delta t_k + \Delta t_{199-k})/2$, $k = 0, \ldots, 199$. The resulting OT-CFM schedule attains an FID of 5.84, compared with 5.57 for the original model-aware schedule and 6.30 for the analytic template. Relative to the symmetrized model-aware schedule, the analytic template recovers 83.1% of the FID improvement over the baseline. Symmetrizing the model-aware allocation therefore narrows the gap, but does not account for it completely.

Finally, using the CIFAR-10 settings, we test whether the observed benefit is sensitive to the numerical realization of the same analytic template. For DDPM ϵ-prediction, the direct 1000-step realization gives an FID of 7.93, compared with 8.05 for the 199-bin realization. For OT-CFM, the 200-bin piecewise-linear realization gives 6.30, compared with 6.47 for the smooth realization. All four remain substantially better than their respective baselines. We view these differences as numerical realization effects, rather than evidence for a generally preferred discretization.

E.12 Empirical Bayes Decomposition of Fiberwise Risk

Independent-CFM is included in the coupling-robustness experiments of Section 6.3, and its normalized risk and allocation shapes are included in the unified comparisons of Appendix E.10. Here we use the independent endpoint structure of DDPM and Independent-CFM to examine which part of the fiberwise-risk profile carries the shared broad shape. To do so, we apply an empirical Bayes decomposition along the corresponding reference coordinate τ. Let

$$p_{\mathrm{emp}}(x_0) = \frac{1}{50{,}000} \sum_{j=1}^{50{,}000} \delta(x_0 - x_j) \tag{E.10}$$

be the finite empirical CIFAR-10 training distribution. In this subsection, outer expectations are over $x_0 \sim p_{\mathrm{emp}}$ and independent $\epsilon \sim \mathcal{N}(0, I)$, whereas $\mathbb{E}_{\mathrm{emp}}[\cdot \mid x_\tau]$ denotes the corresponding empirical-posterior conditional expectation. For the affine observation $x_\tau = m_\tau x_0 + s_\tau \epsilon$, the posterior mean of x_0 is

$$\mathbb{E}_{\mathrm{emp}}[x_0 \mid x_\tau = x] = \frac{\sum_j x_j \exp\!\big(-\|x - m_\tau x_j\|^2/(2s_\tau^2)\big)}{\sum_j \exp(-\|x - m_\tau x_j\|^2/(2s_\tau^2))}. \tag{E.11}$$

For the fiberwise risk R, the standard squared-error Bayes decomposition gives

$$R(\tau) = P(\tau) + M(\tau), \tag{E.12}$$

where P is the Bayes-risk component under the empirical data distribution p_{emp} and M is the Bayes-excess component. For a fixed empirical data distribution and probability path, P is independent of the trained predictor, whereas M measures the predictor-dependent excess risk above the empirical Bayes predictor.

For DDPM ϵ-prediction,

$$\begin{aligned} R_{\mathrm{DDPM}}(\tau) &= (1 - \bar{\alpha}_\tau)\mathbb{E}\left\|\epsilon - \epsilon_\theta(x_\tau, \tau)\right\|^2, \\ P_{\mathrm{DDPM}}(\tau) &= \bar{\alpha}_\tau \mathbb{E}\left\|x_0 - \mathbb{E}_{\mathrm{emp}}[x_0 \mid x_\tau]\right\|^2. \end{aligned} \tag{E.13}$$

For Independent-CFM,

$$x_\tau = (1 - \tau)x_0 + \tau\epsilon, \tag{E.14}$$

and

$$\begin{aligned} R_{\mathrm{ICFM}}(\tau) &= \tau^2(1 - \tau)^2 \mathbb{E}\left\|v_\theta(x_\tau, \tau) - (\epsilon - x_0)\right\|^2, \\ P_{\mathrm{ICFM}}(\tau) &= (1 - \tau)^2 \mathbb{E}\left\|x_0 - \mathbb{E}_{\mathrm{emp}}[x_0 \mid x_\tau]\right\|^2. \end{aligned} \tag{E.15}$$

In both cases,

$$M(\tau) = R(\tau) - P(\tau). \tag{E.16}$$

The posterior means are evaluated against the same finite empirical training distribution, and the decomposition uses the same reference-bin risk estimates used for schedule construction.

Table 17 summarizes the integrated component shares and the agreement of their normalized shapes with the total fiberwise-risk profile.

The Bayes-excess component accounts for 96.32% of the integrated DDPM risk and 93.74% of the Independent-CFM risk. Its normalized shape also closely follows the total risk in both settings, with Pearson correlations of 0.9923 and 0.9809 and TV discrepancies of 3.25% and 5.57%, respectively. By contrast, the Bayes-risk component has weakly negative shape correlation with the total risk. Thus the broad risk-profile shape shared across these two independently coupled systems is carried predominantly by the Bayes-excess component—the part that depends on the learned predictor—rather than by the Bayes-risk component alone.

Table 17: Bayes-risk/Bayes-excess decomposition of the CIFAR-10 fiberwise-risk profiles. Integrated shares are computed under the finite empirical training distribution. Shape metrics compare unit-area profiles.

Setting	$\frac{\int_0^1 P}{\int_0^1 R}$ (%)	$\frac{\int_0^1 M}{\int_0^1 R}$ (%)	Pearson(R, P)	Pearson(R, M)	TV(R, M) (%)
DDPM ϵ-pred.	3.68	96.32	-0.4333	0.9923	3.25
Independent-CFM	6.26	93.74	-0.3514	0.9809	5.57

This decomposition characterizes the empirical fiberwise risk used for schedule construction under the finite CIFAR-10 training distribution. The reported component shares and shape agreement do not by themselves establish the same findings under the underlying data distribution.