POST-TRAIN THE MAP YOU SOLVE: DEPLOYMENT-ALIGNED SPECIALIZATION FOR FEW-STEP FLOW SAMPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

Flow matching trains a continuous velocity field, but few-step sampling deploys a solver- and budget-specific finite map. We post-train that exact map by differentiating a terminal distribution objective through the deployed solver. A shared field–clock repair is followed by tiny solver-specific residual clocks, activated only when held-out target alignment predicts a gain. At 12 field evaluations on a public conditional-flow receiver trained with C^2OT (conditional optimal transport), shared repair lowers mean CleanFID10k from 46.824 to 7.704; a 120-parameter residual package reaches 7.334 and is favorable in 10/10 retrainings for each active solver. On the official Scalable Interpolant Transformer (SiT) checkpoint, a frozen finite-stage operator improves Heun and fourth-order Runge–Kutta in every paired 50,000-sample ImageNet Fréchet Inception Distance (FID) evaluation and transfers without target-budget fitting to 8 and 24 field evaluations. B-series analysis and measured error orders trace these gains to tableau-dependent finite-step error. The alignment rule also predicts held-out directions on a second architecture and non-image receivers, while an alternative objective and independent feature representations corroborate the effect. The method keeps one field across solvers and adds no field evaluations; held-out evidence determines which residual clocks are stored.

1 INTRODUCTION

Flow Matching (FM) regresses a velocity field along a probability path (Lipman et al., 2022; Tong et al., 2024; Liu et al., 2023; Pooladian et al., 2023). A deployed sampler exposes a different object: the finite numerical map defined by the field, solver tableau, evaluation nodes, and number of function evaluations (NFE). At low NFE, optimizing the continuous model alone leaves this deployment map misaligned with the terminal distribution of interest.

Learned paths and schedules improve integrability (Albergo et al., 2023; Min et al., 2026; Sabour et al., 2024), while inference-time methods redistribute evaluations (Karras et al., 2022; Lu et al., 2022; Zhao et al., 2023; Yuan et al., 2026; Lee et al., 2025). Finite-horizon models instead change the prediction target (Salimans & Ho, 2022; Song et al., 2023; Zhao et al., 2024; Lee et al., 2026). These approaches do not isolate a central deployment question: can one post-train the exact solver–budget map, then specialize it without duplicating the field?

Our post-training procedure has two stages. Stage I repairs a restricted late-field subset and a monotone clock through the exact deployed solver. Stage II freezes the field and allocates a small residual clock to each solver. Held-out alignment selects the shared or residual clock. The procedure requires no dense endpoint teacher or per-sample controller and adds no field evaluations at deployment (Figure 1).

Our contributions are:

- We optimize the exact finite-step deployment map with chunked replay, producing one shared repair and 120 active residual parameters without extra field evaluations.
- Solver-tableau analysis and matched challenges establish solver dependence. A conditional optimal transport (C^2OT) receiver (Cheng & Schwing, 2025) provides a two-stage ladder;

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.

an ImageNet transformer confirms temporal adaptation and cross-budget transfer. Frozen alignment predicts prospective directions; alternative objectives, independent features, and bounded second-order Runge–Kutta transfer test robustness and scope.

2 Related work and scope

Inference-time and lightweight solver accelerators reshape trajectories or solver updates (Shaul et al., 2024; Tan et al., 2026; Jin et al., 2026; Kong et al., 2026); finite-horizon and delay-based models change the prediction or update map (Salimans & Ho, 2022; Song et al., 2023; Zhao et al., 2026; Lee et al., 2026; Kim et al., 2026). A-FloPS reparameterizes diffusion trajectories; ERK-Guid uses embedded Runge–Kutta (RK) error; Decoupled MeanFlow learns a flow map; and SpanLift adds an endpoint-teacher-matched spatial residual. Our method post-trains the exact field–clock–solver program, retains one field across solvers, and uses held-out evidence to choose residual clocks. Differentiable ordinary differential equation (ODE) training provides rollout gradients (Chen et al., 2018; Zhuang et al., 2020). Diffusion Transformers (DiT) and Scalable Interpolant Transformers (SiT) receivers and paired Inception/DINOv2 diagnostics test architecture and metric transfer (Peebles & Xie, 2023; Ma et al., 2024; Parmar et al., 2022; Stein et al., 2023).

3 Method

At fixed NFE, the optimization target is the coupled field–clock–solver program that is actually executed. Our method post-trains this exact deployment map in two stages. Stage I updates a restricted field subset and a monotone clock through the exact rollout; Stage II freezes the field and learns small solver-specific residual clocks. Held-out alignment selects the shared or residual clock, so deployment retains one field and incurs no extra NFE (Figure 1).

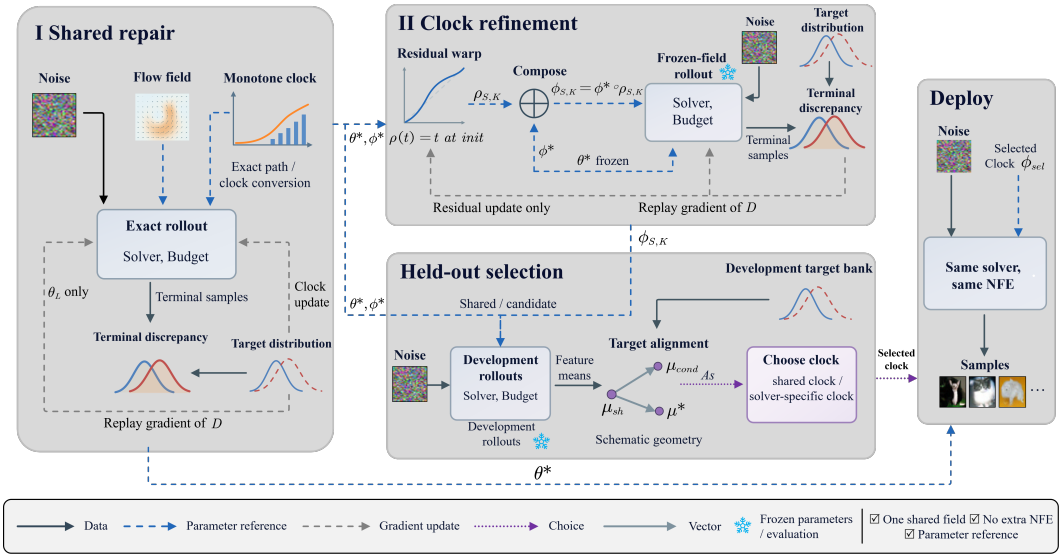

Figure 1: **Deployment-aligned post-training.** Stage I repairs one shared field–clock artifact by differentiating terminal discrepancy through the exact deployed rollout. Stage II freezes the field and learns small residual clock warps. A held-out alignment statistic selects the shared or residual clock for deployment with the same solver and NFE.

3.1 Exact path conversion

Let noise and data be x_0 and x_1, and define

$$x_t = \alpha_\phi(t)x_1 + \sigma_\phi(t)x_0, \qquad s = \alpha + \sigma > 0, \qquad \tau = \alpha/s. \tag{1}$$

Then $y = x_t/s = \tau x_1 + (1-\tau)x_0$ lies on the canonical linear interpolation under the TorchCFM convention $u_\theta(\tau, y, c) \approx x_1 - x_0$. Direct differentiation in forward time gives

$$v_{\theta,\phi}(t, x, c) = \frac{\dot{s}}{s}x + \underbrace{s\dot{\tau}}_{(\sigma\dot{\alpha}-\alpha\dot{\sigma})/s} u_\theta\Big(\tau, \frac{x}{s}, c\Big). \tag{2}$$

The sign follows the stated forward-velocity convention; reversing the target to $u = x_0 - x_1$ reverses it. For $(\alpha, \sigma) = (t, 1-t)$, the converter reduces exactly to $v = u_\theta$.

For a time-only clock, 16 logits ϕ give positive interval masses $p = \mathrm{softmax}(\phi)$; their cumulative sums are the Bernstein controls $0, c_1, \ldots, c_{15}, 1$ of a monotone, fixed-endpoint map $\tau_\phi : [0,1] \to [0,1]$. Thus $\tau_\phi(0) = 0$ and $\tau_\phi(1) = 1$. We reserve ϕ for the logit vector and write τ_ϕ for its induced map. This time-only parameterization is the smallest stable component in matched decompositions; higher-capacity scale and path families are controls.

3.2 The discrete deployment objective

For explicit solver S and exact evaluation budget B, let $F^{S,B}_{\theta,\phi}$ denote the terminal map produced by the actual deployed discrete program. We optimize

$$J_{S,B}(\theta, \phi) = \mathcal{D}\Big((F^{S,B}_{\theta,\phi})_{\#}p_0, p_1\Big) + \lambda_{\mathrm{FM}} L_{\text{path-FM}} + \lambda_A L_{\text{linear-FM}} + \lambda_I \|\phi\|_2^2. \tag{3}$$

$\mathcal{D}$ is the full-covariance Fréchet discrepancy after standardizing 2048-D CleanFID `legacy_pytorch` pool-3 features with a real-bank mean and standard deviation, then projecting them onto fixed Gaussian-QR orthonormal columns (seed 11991). Headline C^{2}OT uses 32 columns, and TorchCFM robustness uses 64. We set $(\lambda_{\mathrm{FM}}, \lambda_A, \lambda_I) = (1, 0.1, 10^{-3})$. The path-FM term regresses $v_{\theta,\phi}$ to $\partial_t x_t$ on the learned path, while linear-FM regresses u_θ to $x_1 - x_0$ on the frozen linear coordinate; both use independently drawn (x_0, x_1, c) pairs and no endpoint teacher. A frozen objective audit replaces $\mathcal{D}$ by sliced-W2 along the same orthonormal axes, without result-dependent scaling. Image claims are confirmed by paired Fréchet Inception Distance (FID) and independent DINOv2 metrics on fresh streams; cross-domain tests freeze θ and use normalized endpoint mean squared error (MSE).

Exact NFE counts field calls, so legal budgets are divisible by two, three, and four for Heun, RK3, and classical RK4. Paired comparisons hold the solver, NFE, inputs, preprocessing, and evaluator fixed.

3.3 Shared adaptation and selective solver refinement

Let $\mathcal{S} = \{\text{Heun}, \text{RK3}, \text{RK4}\}$. Stage I learns one shared artifact by alternating solvers and their compatible training budgets:

$$(\theta^\star, \phi^\star) = \arg\min_{\theta_L, \phi} \; \mathbb{E}_{S\sim\mathcal{S},\, B\sim\mathcal{B}_S} J_{S,B}(\theta, \phi). \tag{4}$$

Here, θ_L contains the time embedding, output blocks 10–11, and output head; the remaining field is frozen.

Stage II freezes the complete field, initializes $\phi_S \leftarrow \phi^\star$ for each solver, and solves

$$\phi_S^\star = \arg\min_{\phi_S} \mathbb{E}_{B\sim\mathcal{B}_S} J_{S,B}(\theta^\star, \phi_S). \tag{5}$$

The fixed-capacity study uses 16 refinement updates per solver and three direct 16-logit clocks. Capacity $K = 0$ retains $\phi^\star$; for $K > 0$, the induced map is composed with a monotone Bernstein warp, $\tau_{\phi_{S,K}}(t) = \tau_{\phi^\star}(\rho_{S,K}(t))$. Zero residual logits make $\rho_{S,K}(t) = t$ exactly, so every capacity starts from the same discrete map. A frozen development rule selects one action per solver before independent evaluation. The C^{2}OT package uses $(0, 32, 8)$ and 120 active residual parameters, with three logits per residual order. Neither construction changes NFE nor evaluates a controller.

Finite-stage temporal operator. For the frozen SiT control at NFE $= 12$, eight logits are centered and exponentiated, then linearly resampled to the seven unique Heun/RK4 stage locations. The resulting positive density values r_j are normalized by the original trapezoid (Heun) or Simpson (RK4)

weights, so that $\sum_j w_j r_j = 1$. Their cumulative quadrature mass gives the query times τ_j; the field evaluations are scaled as $k_j = r_j v_\theta(\tau_j, x_j, c)$. The tableau and state update remain unchanged. Thus the operator preserves unit dose on a constant field and the exact number of field calls; it does not learn a new tableau.

Chunked replay evaluates identical discrete rollouts and stores their terminal cotangents. Parameters are updated only after all chunks. The supplement gives pseudocode, equality checks, and hashes.

3.4 Why specialize, and when to deploy a residual clock

For a frozen field, write the solver terminal map as $z_{\text{end}} = \Psi_S(\tau_\phi; z_0, \theta)$. Under the local quadratic-model assumption $H_S \succ 0$, approximate each solver objective around τ_{ϕ_0} by

$$Q_S(\delta) = q_S + g_S^\top \delta + \tfrac{1}{2}\delta^\top H_S \delta, \qquad H_S \succ 0. \tag{6}$$

The local error depends explicitly on the tableau. For $f_\phi(t, x) = \gamma(t) v(\tau_\phi(t), x)$ and $\gamma = \dot{\tau}_\phi$, its B-series differentials mix $\ddot{\tau}_\phi$, $\dddot{\tau}_\phi$, v_τ, and $D_x v\, v$. Their coefficients depend on the RK tableau. A symbolic witness recovers distinct leading polynomials at h^3, h^4, and h^5 for Heun, RK3, and RK4. For $v(\tau, x) = (1 + 0.7\tau)x$ and the local witness $\psi(t) = t + 0.2t^2 + 0.1t^3$, measured one-step slopes are 3.005, 3.995, and 5.003. Normalized coefficients match the analytic values within 0.16%. The polynomial is an analytic witness, not a deployed clock; the supplement gives the expansion and numeric trace. The same recursion makes g_S and H_S depend on tableau and NFE. Let $\delta_S^* = -H_S^{-1} g_S$ and $\delta_{\text{sh}}^* = \arg\min_\delta \sum_S w_S Q_S(\delta)$, and define $R_{\text{spec}} := \sum_S w_S [Q_S(\delta_{\text{sh}}^*) - Q_S(\delta_S^*)]$. Completing the square gives

$$R_{\text{spec}} = \tfrac{1}{2} \sum_S w_S \|\delta_{\text{sh}}^* - \delta_S^*\|_{H_S}^2 \geq 0, \tag{7}$$

where $\|x\|_{H_S}^2 = x^\top H_S x$. For $w_S > 0$, equality holds exactly when all solvers share a local optimum. Equation 7 permits local conflict but says nothing about held-out quality.

Held-out deployment uses a separate development-stream diagnostic. For image features, with shared and conditioned generated means μ_{sh} and μ_{cond} and target mean μ_{tgt}, define

$$A_S = \cos(\mu_{\text{cond}} - \mu_{\text{sh}},\, \mu_{\text{tgt}} - \mu_{\text{sh}}). \tag{8}$$

Zero-norm displacements and cases with $|A_S| < 0.05$ fall back to the shared action; otherwise, the frozen sign of A_S predicts $\Delta_S = \text{FID}_{\text{sh}} - \text{FID}_{\text{cond}}$. The action analogue uses flattened endpoints and MSE. All predictions and hashes precede disjoint confirmation. The statistic selects but never tunes a residual clock; its extractor is frozen.

3.5 Bounded tableau-coordinate transfer within RK2

An explicit two-stage, second-order RK method is indexed by its internal stage $c_2 \in [1/2, 1]$. Midpoint and Heun provide endpoint artifacts. Let $\omega = (\theta_L, \phi)$ collect their stored late-field tensors and clock logits. This is a separate auxiliary zero-fit transfer audit, rather than the one-shared-field residual-clock deployment. It targets Ralston ($c_2 = 2/3$) and loads

$$\omega_B = (1 - \beta_B)\omega_{\text{midpoint}} + \beta_B \omega_{\text{Heun}}, \qquad \beta_B = \tfrac{1}{3}\,\text{clip}\left(\frac{B - 12}{12}, 0, 1\right). \tag{9}$$

This frozen rule uses no target-receiver loss, samples, or gradients; at NFE $= 8$ and NFE $= 12$ it is exactly the midpoint artifact. A distinct fixed- NFE $= 24$ coordinate probe uses $\beta_{c_2} = 2(c_2 - 1/2)$ to interpolate the same tensors for $c_2 \in \{0.625, 0.70\}$. The probes provide positive checks at two unseen coordinates within the pre-specified region $c_2 \leq 0.75$.

4 Experimental protocol

Receivers. Image experiments use CIFAR-10 (Krizhevsky, 2009) flows: public TorchCFM receivers, stochastic-bridge conditional flow matching (SB-CFM) and variance-preserving conditional flow matching (VP-CFM) pretraining variants, an audited DiT exponential-moving-average (EMA) with FP32 master weights, and the official SiT-XL/2 ImageNet-256 checkpoint (Peebles & Xie,

2023; Ma et al., 2024; Yu et al., 2025). C^{2}OT is the separate public, independently implemented 89.1M-parameter class-conditional CIFAR-10 FM U-Net; C^{2}OT and DiT test post-training stability. Three-seed frozen-field receivers cover Electricity forecasting (Alexandrov et al., 2020; Zhou et al., 2021), projected-Qwen WikiText-2 continuation (Merity et al., 2017; Yang et al., 2025), Spiking Heidelberg Digits (SHD) event continuation (Cramer et al., 2022), and Lorenz-96 forecasting (Lorenz, 1995). Four offline-action receivers independently test alignment for endpoint prediction.

Metrics and comparisons. Development uses fixed projected-Fréchet features. Confirmatory image tests use paired legacy-TensorFlow Fréchet Inception Distance (FID) (Heusel et al., 2017; Parmar et al., 2022): CleanFID10k denotes 10,000 generated images, while ADM-FID50k denotes 50,000 images evaluated with the ImageNet Inception pool-3 protocol used for OpenAI's ablated diffusion model (ADM). Unprefixed FID10k/FID50k denotes the same sample-count convention when the surrounding experiment fixes the evaluator. DINOv2 Fréchet distance (FD) is distinct from Inception FID. The audit adds Kernel Inception Distance (KID), precision, and recall (Bińkowski et al., 2018; Kynkäänniemi et al., 2019; Oquab et al., 2024). The C^{2}OT ladder and cross-solver challenges use 10,000 samples; sequential specialization and SiT FID confirmations use 50,000 unless stated otherwise. A SiT stream varies only generation noise and labels for one frozen checkpoint and operator; five SiT post-training seeds vary the post-training run, not SiT pretraining. C^{2}OT retrainings reuse one pretrained checkpoint and measure adaptation robustness. Rules and metrics are frozen before confirmation. All displayed differences and intervals use stored unrounded values; the score means are rounded separately. Complete budgets, seeds, hashes, and matrices are in the supplement.

Optimization details. The C^{2}OT shared stage updates 3.477M late-field parameters and a 16-logit clock for 192 alternating-solver steps; residual refinement uses 48 steps (16 per solver). Both use AdamW, field/clock learning rates $10^{-5}/10^{-3}$, weight decay 10^{-4}, gradient clip 5, 128 calibration samples, replay batches of 16, and FM batches of 128. The fixed real fit/evaluation banks contain 1,024/2,048 samples. Receiver-specific deviations are listed in the supplement.

5 Results

5.1 Shared repair accounts for most of the C^{2}OT gain

A same-checkpoint C^{2}OT ladder measures the gap between the untouched receiver, Stage I, and Stage II (Table 1). At exact NFE $= 12$, shared field–clock repair lowers mean CleanFID10k from 46.824 to 7.704. Freezing the field and applying the selected solver-specific actions reaches 7.334; Heun keeps the shared clock while RK3 and RK4 receive residual refinement. The ladder attributes most of the improvement to shared repair and the smaller final correction to solver-specific clocks.

A prospectively frozen three-seed audit replaces projected Fréchet training with sliced-W2. It improves common CleanFID10k by 0.823/0.468/0.142 at NFE $= 8, 12, 24$, with all NFE $= 8$ and NFE $= 12$ artifacts improving. Independent DINOv2 evaluation agrees: FD improves by 8.929 and 5.483 at NFE $= 8$ and NFE $= 12$, and KID improves in every seed. At NFE $= 12$, the paired 95% interval is $[0.262, 0.674]$, while DINOv2 recall rises by 0.0082. Removing the path-FM anchor loses this cross-metric consistency. Per-seed values and secondary metrics appear in the supplement.

Temporal parameterization is isolated on the frozen official SiT-XL/2 checkpoint. Each learned arm has eight scalars, 192 updates, balanced Heun/RK4 exposure, and paired NFE $= 12$ streams. Across five fresh adaptations, continuous time beats the grid by 0.115 mean FID in 4/5 pooled contrasts. The subsequent comparison changes only the placement of the same eight scalars. In the original FID50k isolation, continuous/grid/uniform/flexible-path obtain 8.1652/8.1988/8.2331/8.2756, respectively.

Table 1: Same-checkpoint C^{2}OT ladder at NFE $= 12$. Stage-I and Stage-II reductions are paired within solver. Mean FID is the arithmetic mean of the three solver scores; reductions use the stored unrounded scores. Lower CleanFID10k is better.

Solver	Untouched	Shared	Selected	Stage-I reduction	Stage-II reduction
Heun	62.842	7.421	7.421	88.19%	0.00%
RK3	35.585	8.536	7.565	76.01%	11.37%
RK4	42.045	7.156	7.016	82.98%	1.97%
Mean	46.824	7.704	7.334	83.54%	4.81%

Table 2: Three-stream SiT finite-stage confirmation at NFE $= 12$. Entries are means across $n = 3$ paired 50k-sample generation streams from one frozen SiT checkpoint and operator; brackets are two-sided 95% t intervals for control minus finite-stage. Lower ADM-FID50k is better.

Solver	Grid	Continuous	Finite-stage	Grid − finite	Continuous − finite
Heun	9.2210	9.1564	**8.5980**	0.6230 [0.5574,0.6886]	0.5583 [0.4851,0.6315]
RK4	7.1933	7.1832	**7.0450**	0.1483 [0.1346,0.1621]	0.1382 [0.1300,0.1465]

The frozen finite-stage operator improves both temporal controls for Heun and RK4 in every paired ADM-FID50k stream (Table 2). Against the strongest control, mean gains are 0.5583 FID for Heun and 0.1382 for RK4; both paired intervals exclude zero.

Five independent post-training seeds support the same action. Finite-stage risk is lower than the grid in every Heun and RK4 unit, and lower than continuous time in 4/5 and 5/5 units, respectively. The four mean risk reductions span 0.0255–0.0671, and every paired interval is positive.

5.2 Finite-stage budget transfer and shared-clock sampler composition

We freeze the selected eight-logit operator and resample its positive density at each solver's evaluation nodes, without fitting the target budget. It has lower FID than all temporal controls for both solvers at NFE $= 8$ and NFE $= 24$. Against the strongest control, Heun/RK4 gains are 0.3707/0.0312 at NFE $= 8$ and 0.0165/0.0053 at NFE $= 24$. The smaller NFE $= 24$ margin matches the expected contraction of finite-step error; the complete budget matrix is in the supplement.

The shared clock also composes with Stabilized Taylor Orthogonal Runge–Kutta (STORK) on the official SiT checkpoint. At NFE $= 12$, it improves all three ADM-FID50k streams by 0.0307–0.0393 FID, with a mean reduction of 0.0344 (supplement). The correction preserves STORK's training-free solver design and leaves the field unchanged.

5.3 Post-training is tied to the deployed solver

Figure 2a tests the strongest off-diagonal artifact for each NFE $= 12$ solver. Every VP-CFM and corrected-DiT cell is favorable in all replications; C^2OT confirms all solvers, including 2/3 favorable replications for its saturated RK4 row. Relative reductions span 29.8–66.5% for VP-CFM, 19.9–75.1% for DiT, and 3.6–70.4% for C^2OT. Formal order does not identify the solver-specific deployment optimum. The gaps must be measured for each solver.

At NFE $= 24$, VP-CFM retains RK3/RK4 matching and corrected DiT retains Heun/RK3; the other cells approach saturation. We refine small residual clocks from one shared artifact instead of copying the field three times.

5.4 Tiny residual clocks specialize one shared artifact

Each sequential unit starts from a new shared C^2OT artifact, clones its clock, and freezes every field weight. At NFE $= 12$, residual clocks improve all nine paired FID50k cells (Figure 2b), reducing pooled FID by 5.586%, from 5.4797 to 5.1736. At NFE $= 24$, the reduction contracts to 2.513% with 8/9 wins. Heun and RK3 remain positive in every unit, while the 16-logit RK4 residual is nearly saturated.

Refinement uses 25% of Stage I optimizer steps and model-sample evaluations. Field hashes remain unchanged, and pooled inference overhead is 0.00342%.

Three independently trained SB-CFM receivers replicate the sequence, lowering pooled NFE $= 12$ FID50k from 9.6244 to 9.1037 (5.410%, 8/9). After equalizing the final 48 updates, RK3 still improves by 4.455% in all receivers. The selected policy retains the shared action for Heun and RK4.

An equal-start capacity sweep selects $(0, 32, 8)$, preserving the shared Heun clock and activating RK3/RK4 (Figure 3a). Across the active RK3/RK4 comparison, selected residuals reduce pooled CleanFID10k by 7.025%. RK3 improves from 8.3770 to 7.4176, and RK4 from 7.3095 to 7.1669, on independent confirmation streams. Each active residual is favorable in 10/10 retrainings. A separate SB-CFM receiver yields a different response. The two capacity curves differ by both solver and receiver. Complete values are in the supplement.

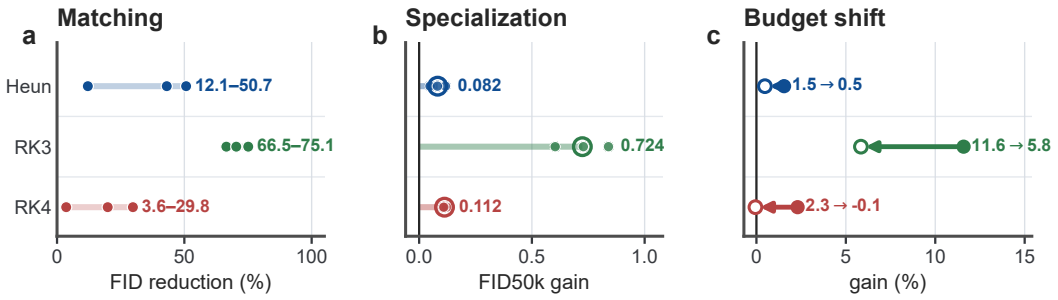

Figure 2: **Few-step efficacy is solver-specific.** Every horizontal row is one solver. (a) Three aligned dots summarize the VP-CFM, DiT, and C^{2}OT cells; the pale segment spans their range. (b) Small dots are paired C^{2}OT units and outlined circles are means. (c) Filled-to-open arrows show contraction from NFE $= 12$ to NFE $= 24$.

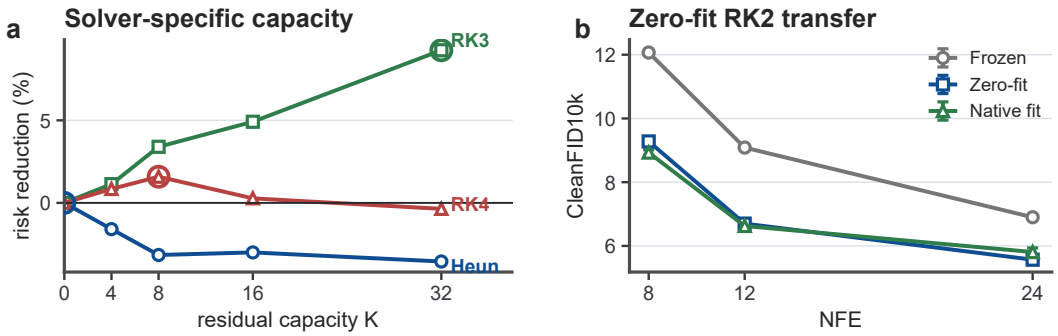

Figure 3: **Residual capacity is solver–receiver specific, while tableau transfer is bounded.** (a) C^{2}OT solvers select different residual capacities; circles mark the frozen actions. (b) Zero-fit Ralston interpolation approaches the native fit; error bars are standard errors across five artifact seeds.

5.5 Selection for unseen solvers and bounded second-order transfer

The SiT isolation withholds RK3 from every learned temporal object. On fresh FID10k streams, the grid improves over uniform time by 0.407%, while the continuous clock remains at the uniform operating point. A single default temporal family is unsuitable for an unseen solver.

The solver–temporal interaction nevertheless permits bounded transfer within the second-order Runge–Kutta family. For Ralston's $c_2 = 2/3$, the budget rule in Eq. 9 improves FID10k over the frozen receiver by 2.789/2.386/1.332 at NFE $= 8, 12, 24$, with 5/5 wins at each budget. It retains 88.9%/97.0%/121.9% of native Ralston gains; the NFE $= 24$ ratio exceeds 100% because zero-fit narrowly outperforms the native fit on that stream. Because $\beta_B = 0$ at NFE $= 8, 12$, these two cells test the midpoint artifact only; the midpoint comparison is made at NFE $= 24$. Separately, the fixed-NFE $= 24$ coordinate probe at $c_2 = 0.625$ and 0.70 beats both frozen and midpoint transfer in every pair without target-receiver fitting. Zero-fit operation holds at the tested coordinates within $c_2 \leq 0.75$.

The finite-stage SiT audit independently favors midpoint risk, 0.6144, over uniform, 0.6268, and continuous time, 0.6226. The two audits support bounded, solver-checked transfer rather than unrestricted tableau interchangeability.

5.6 The effect follows finite-step error and saturates

Against an NFE $= 128$ Heun reference from the same artifact, the NFE $= 8$ risk change splits into $+0.115$ high-NFE transport and -16.614 finite-step excess; using stored unrounded values, these terms explain the -16.500 total. NFE $= 12, 24$ behave similarly. Across all 15 artifact–budget cells, adjoint-weighted Richardson defects improve and raw root-mean-square (RMS) error falls by 46.9%/39.0%/49.7%, locating the gain in finite-step excess.

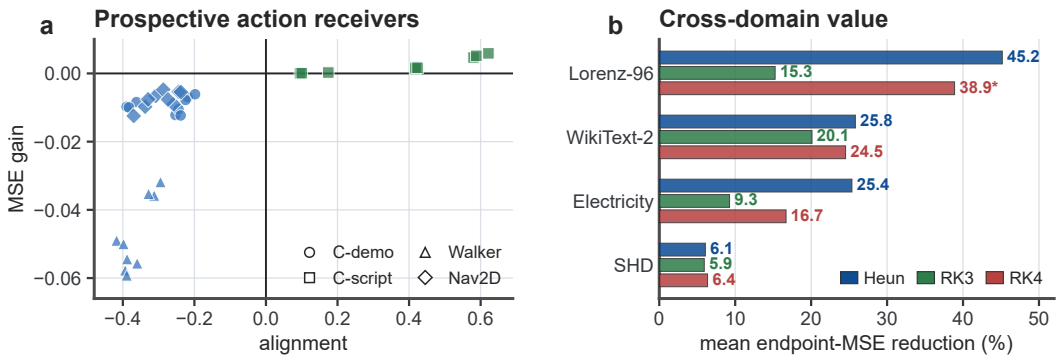

Figure 4: **Frozen alignment selects useful actions across receivers.** The preregistered rule predicts all 36 unseen action cells (a); green and blue denote residual and shared selections. Mean endpoint-MSE reductions across four non-image domains and three solvers remain positive (b); the asterisk marks the single cell selecting a conditioned residual clock.

With the field frozen, pooled clock gain falls from 5.586% at NFE $= 12$ to 2.513% at NFE $= 24$ and 0.250% at NFE $= 36$. Such decay is expected for a finite-step correction, not for a budget-independent distribution shift.

5.7 Target alignment predicts useful specialization

At a fixed SB-CFM field, clock, noise, and target, clock-gradient cosines differ across solvers. Residual displacements follow solver-specific descent in all 9/9 cells; across 18 retrospective SB-CFM/C^{2}OT cells, Eq. 8 correlates with FID gain (Pearson 0.787, Spearman 0.851), motivating prospective selection.

The frozen rule predicts every FID and mean/covariance sign on disjoint 10,000-sample C^{2}OT and DiT streams: C^{2}OT recovers the informative $(+, +, -)$ pattern, and DiT replicates all nine directions. All 30 stream signs repeat from 1,024 to 2,048 development samples, and thresholds from 0 to 0.1 preserve all six decisions. DINOv2 and Contrastive Language–Image Pretraining (CLIP) agree in 10/12 cross-feature checks, including 6/6 for DiT. The retained 64-dimensional projection wins every NFE $= 12$ seed pair against 16 and 32 dimensions; exact 5^5 stream bootstraps exclude zero in every case (Figure 5).

Across four conditional-action receivers, alignment predicts all 36 confirmation-MSE directions: conditioning wins nine Cheetah-scripted cells and the shared clock the other 27 (Figure 4a).

5.8 Frozen-field clock gains extend beyond image generation

With every field frozen, selected clocks reduce normalized endpoint MSE by 5.93–45.17% across all 12 domain–solver means and 36 seed cells (Figure 4b); the shared clock is selected in 11/12. At eight-fold NFE, changes shrink to -0.022%–2.31%, reproducing image-side saturation with task-native losses. MSE is averaged over endpoint dimensions and held-out examples after training-partition affine standardization, so reductions are compared within domains. WikiText-2 also improves three secondary metrics for every solver, showing the effect without image features or trainable field capacity.

5.9 Architecture, quality, and efficiency controls

TorchCFM controls identify time reparameterization as the stable compact path component. A full-field update uses $25.2\times$ as many trainable parameters, whereas the restricted late-field–clock configuration retains the gain.

Corrected DiT provides a selection-free NFE $= 8$ audit. Across five matched 10,000-sample streams, field–clock adaptation reaches CleanFID10k 39.218, versus 129.558 for the frozen EMA and 66.746 for field-only adaptation. Resetting the clock raises the joint artifact to 70.750; the full model then improves FID by 31.531 [30.407,32.655] with 5/5 signs and lowers CLIP-FID from 21.698 to 16.479.

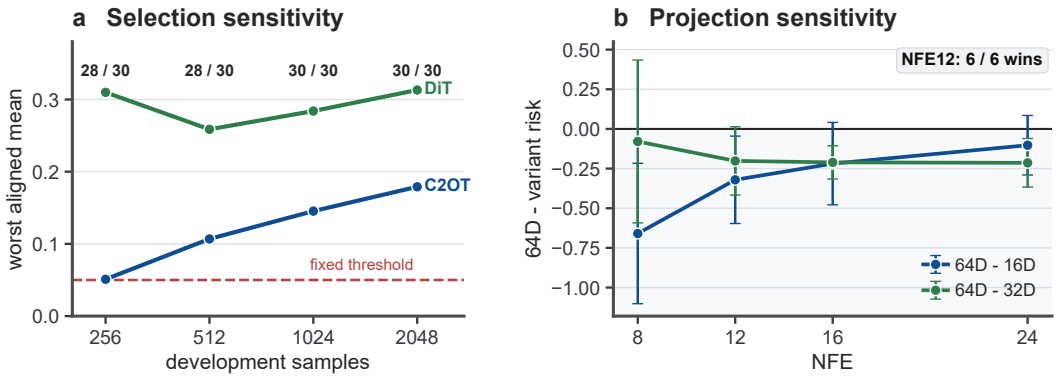

Figure 5: **Deployment decisions are stable to development size and projection resolution.** (a) Each curve is the smallest confirmation-aligned five-stream mean across solvers for one receiver; labels report correct, non-indeterminate directions across all 30 stream cells, and the dashed line is the frozen threshold. (b) Default-64 minus lower-dimensional risk under one frozen 64-dimensional witness; negative values favor the default and bars are 95% t intervals across three adaptation seeds.

The same-checkpoint BézierFlow comparison separates quality from adaptation cost (Min et al., 2026). Its endpoint-teacher portfolio obtains 6.186/5.082/3.964 FID50k at NFE $= 8, 12, 24$, versus 20.562/9.432/5.487 for our teacher-free arm. Our shared artifact needs 253.4 s and 196,608 backbone-sample evaluations, versus 1,465.5 s and 1,940,800 for BézierFlow.

6 Discussion

Restricted field updates repair the shared artifact; a clock then redistributes finite-step error for a particular tableau and budget. B-series terms explain solver-dependent descent directions, while held-out alignment determines whether the displacement is worth deploying. The alternative objective, independent features, frozen-field non-image tests, and the NFE $= 12$–36 decay locate the effect in finite-step error rather than extra field capacity or one image metric. Together, the results separate two roles: shared repair restores a common low-NFE operating point, whereas residual clocks absorb the remaining tableau-dependent error. This division allows one velocity field to support several deployed solvers without treating their finite maps as interchangeable.

The scope is fixed, few-step deployment. Ralston transfer is verified at NFE $= 8, 12, 24$; two additional unseen RK2 coordinates are tested at NFE $= 24$ within $c_2 \leq 0.75$. The receivers span C^{2}OT, SB-CFM, DiT, SiT, four non-image domains, and STORK.

For deployment, adapt one shared artifact at the target low-NFE regime and freeze its field. Estimate alignment for each intended solver and capacity, then store a residual clock only when the diagnostic predicts a gain. Otherwise, use the shared clock unchanged. Interpolate unseen RK2 tableaus only inside the tested interval. At higher NFE, use the shared action because the finite-step excess has contracted.

7 Conclusion

Few-step flow generation is governed by the finite map that is actually solved. Deployment-aligned post-training repairs that map once, then assigns tiny solver-specific clocks when frozen alignment predicts a gain. At NFE $= 12$, the shared C^{2}OT repair lowers mean CleanFID10k by 83.54%. Within the active RK3/RK4 residual-clock scope, the 120-parameter package adds a 7.025% reduction; the 4.81% all-solver mean includes Heun's shared-clock action. On SiT, the finite-stage operator improves both solvers in every three-stream comparison. Cross-budget transfer and independent receiver audits are consistent within the tested low-NFE regime. The sampler adds no field evaluations or model copies and can fall back exactly to the shared clock. Solver-specific temporal calibration therefore becomes a lightweight deployment choice around one reusable velocity field.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript language polishing and code development. The authors reviewed all AI-assisted text and tested and verified all AI-assisted code. The authors take full responsibility for the submitted work.

ETHICS STATEMENT

This work evaluates generative and predictive models using existing benchmark datasets, pretrained checkpoints, and task-level experimental records. It did not recruit human participants or collect personally identifying information. Reported results are aggregate model evaluations. The method may reduce sampling cost but does not remove the misuse risks associated with generative models; deployment remains subject to the access, licensing, privacy, and safety conditions of the underlying data and receivers.

REPRODUCIBILITY STATEMENT

The main paper specifies the exact deployment objective, solver and NFE conventions, parameterization, selection rule, evaluation protocols, and central comparisons. The supplementary material records the replay procedure, hyperparameters, parameter subsets, data preprocessing, feature extractors, sample counts, seeds, paired-stream construction, per-artifact values, statistical intervals, checkpoint and artifact hashes, complete budget and capacity matrices, and representative launch and summarization commands. Together these materials identify the settings needed to reconstruct every headline comparison and distinguish confirmatory evaluations from supporting audits.

REFERENCES

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Alexander Alexandrov, Konstantinos Benidis, Michael Bohlke-Schneider, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, Danielle C. Maddix, Syama Rangapuram, David Salinas, Jasper Schulz, Lorenzo Stella, Ali Caner Türkmen, and Yuyang Wang. GluonTS: Probabilistic and neural time series modeling in python. *Journal of Machine Learning Research*, 21(116):1–6, 2020.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, 2018.

Ho Kei Cheng and Alexander Schwing. The curse of conditions: Analyzing and improving optimal transport for conditional flow-based generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15875–15884, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Cheng_The_Curse_of_Conditions_Analyzing_and_Improving_Optimal_Transport_for_ICCV_2025_paper.html.

Benjamin Cramer, Yannik Stradmann, Johannes Schemmel, and Friedemann Zenke. The heidelberg spiking data sets for the systematic evaluation of spiking neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7):2744–2757, 2022. doi: 10.1109/TNNLS.2020.3044364.

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

Cheng Jin, Zhenyu Xiao, and Yuantao Gu. A-FloPS: Accelerating diffusion models via adaptive flow path sampler. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 22390–22398, 2026. doi: 10.1609/aaai.v40i27.39397.

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

Sihyeon Kim, Seunghun Lee, Vikas Singh, and Hyunwoo J. Kim. Learning to solve generative ODEs beyond the linear span. *arXiv preprint arXiv:2606.08672*, 2026.

Inho Kong, Sojin Lee, Youngjoon Hong, and Hyunwoo J. Kim. Error as signal: Stiffness-aware diffusion sampling via embedded Runge–Kutta guidance. In *International Conference on Learning Representations*, 2026.

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

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, 2019.

Dohoon Lee, Jaehyun Park, Hyunwoo J. Kim, and Kyogu Lee. Multidimensional adaptive coefficient for inference trajectory optimization in flow and diffusion. In *International Conference on Machine Learning*, 2025.

Kyungmin Lee, Sihyun Yu, and Jinwoo Shin. Decoupled MeanFlow: Turning flow models into flow maps for accelerated sampling. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/699124f548e4cb86674b20eae0fd8619-Abstract-Conference.html.

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

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Edward N. Lorenz. Predictability: A problem partly solved. In *Seminar on Predictability*, Reading, UK, 1995. European Centre for Medium-Range Weather Forecasts.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems*, volume 35, pp. 5775–5787, 2022.

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

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Yunhong Min, Juil Koo, Seungwoo Yoo, and Minhyuk Sung. BézierFlow: Learning Bézier stochastic interpolant schedulers for few-step generation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0bd32794b26cfc99214b89313764da8e-Abstract-Conference.html.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11410–11420, 2022.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, 2023.

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, 2023.

Amirmojtaba Sabour, Sanja Fidler, and Karsten Kreis. Align your steps: Optimizing sampling schedules in diffusion models. *arXiv preprint arXiv:2404.14507*, 2024.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

Neta Shaul, Juan Perez, Ricky T. Q. Chen, Ali Thabet, Albert Pumarola, and Yaron Lipman. Bespoke solvers for generative flow models. In *International Conference on Learning Representations*, 2024.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252, 2023.

George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0165. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/0bc795afae289ed465a65a3b4b1f4eb7-Abstract.html.

Zheng Tan, Weizhen Wang, Andrea L. Bertozzi, and Ernest K. Ryu. STORK: Faster diffusion and flow matching sampling by resolving both stiffness and structure-dependence. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/7a641b8ec86162fc875fb9f6456a542f-Abstract-Conference.html.

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.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d9e42b4d7163931f3689d6d6fbaa11d0-Abstract-Conference.html.

Liangyu Yuan, Ruoyu Wang, Tong Zhao, Dingwen Fu, Mingkun Lei, Beier Zhu, and Chi Zhang. Few-step diffusion sampling through instance-aware discretizations. *arXiv preprint arXiv:2603.17671*, 2026.

Bolin Zhao, Xiaoyu Zhang, Yuting Dong, Xin Lu, Wei Lin, and Qunxi Zhu. Delay flow matching. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/5808ba2d4643885430fc8d0f69fd6d73-Abstract-Conference.html.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. UniPC: A unified predictor-corrector framework for fast sampling of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Wenliang Zhao, Minglei Shi, Xumin Yu, Jie Zhou, and Jiwen Lu. FlowTurbo: Towards real-time flow-based image generation with velocity refiner. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0135. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/07c256a163a7559186ec1c71e95b9ec9-Abstract-Conference.html.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021.

Juntang Zhuang, Nicha Dvornek, Xiaoxiao Li, Sekhar Tatikonda, Xenophon Papademetris, and James Duncan. Adaptive checkpoint adjoint method for gradient estimation in neural ODE. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11639–11649, 2020. URL https://proceedings.mlr.press/v119/zhuang20a.html.

SUPPLEMENTARY MATERIAL: POST-TRAIN THE MAP YOU SOLVE: DEPLOYMENT-ALIGNED SPECIALIZATION FOR FEW-STEP FLOW SAMPLING

Anonymous authors
Paper under double-blind review

This supplement provides the complete evidence behind deployment-aligned post-training: exact definitions, seeds, per-artifact values, costs, diagnostic controls, and reproduction commands. Reported values come from promoted JSON summaries whenever available, preserving a direct path from each claim to its machine-readable artifact.

Evidence organization. The main paper follows one claim chain supported by seven evidence roles: exact deployment-map optimization, cross-model solver matching, compact clock refinement, same-checkpoint SiT finite-stage confirmation, target alignment, bounded RK2 transfer, and frozen-field cross-domain validation. The remaining evidence includes complete path-order and factorial cells, bank/projection/anchor and layer sensitivity, universal-artifact budgets, gradient and Jacobian diagnostics, joint Euler/Heun training, ImageNet-256 and conditional stress tests, full-field capacity, secondary precision/recall metrics, cross-domain secondary metrics, cost controls, and reproduction commands. These analyses test the alternatives most relevant to the central mechanism without expanding the headline claim.

1 RELATION TO DIFFERENTIABLE SAMPLER FINE-TUNING

Backpropagating through a generative sampling chain is established by DRaFT (Clark et al., 2024), and memory/accuracy tradeoffs for differentiable ODE solvers are established by Neural ODE and checkpoint-adjoint work (Chen et al., 2018; Zhuang et al., 2020). We build on these tools to optimize a minibatch distribution discrepancy under a fixed solver and exact NFE. The procedure co-adapts a compact clock and restricted field subset, adds analytic FM anchors, and uses exact replay for bounded-memory gradients. This restricted formulation separates shared deployment-map repair from solver-specific specialization.

2 PATH CONVERTER, ORIENTATION, AND FUNCTIONAL DIMENSION

For $x = \alpha x_1 + \sigma x_0$, $s = \alpha + \sigma$, and $\tau = \alpha/s$, $y = x/s = \tau x_1 + (1-\tau)x_0$. Under the TorchCFM forward convention $u_\theta(\tau, y) \simeq x_1 - x_0$,

$$\dot{x} = \frac{\dot{s}}{s} x + s\dot{\tau}\, u_\theta(\tau, x/s), \tag{1}$$

$$s\dot{\tau} = \frac{\sigma\dot{\alpha} - \alpha\dot{\sigma}}{s}. \tag{2}$$

The opposite sign corresponds to $u = x_0 - x_1$ and is not used. Identity-path tests recover $v = u_\theta$ to floating-point precision.

An order-K monotone Bernstein path uses two K-logit softmax-increment vectors. Adding a constant to every logit within either vector changes no increment, so the full path has $2K$ nominal parameters but $2K-2$ functional degrees of freedom. Time-only and learned-grid each have K nominal and $K-1$ functional degrees. Thus cubic full/time have $6/4$ and $3/2$ nominal/functional dimensions; order-16 full/time have $32/30$ and $16/15$.

3 OBJECTIVE AND EXACT REPLAY

For exact-NFE terminal map $F_{\theta,\phi}^{S,B}$,

$$J_{S,B} = \mathcal{D}((F_{\theta,\phi}^{S,B})_{\#}p_0, p_1) + L_{\mathrm{path-FM}} + 0.1L_{\mathrm{linear-FM}} + 10^{-3}\|\phi\|_2^2.$$

$\mathcal{D}$ is a full-covariance projected Fréchet distance. CIFAR runs extract 2048-D CleanFID `legacy_pytorch` pool-3 features, use the 1,024-image real fit bank to fix coordinate-wise mean and standard deviation, and multiply standardized features by orthonormal columns obtained from a reduced QR decomposition of a $2048 \times d$ standard-Gaussian matrix with seed 11991. The C2OT headline uses $d = 32$; TorchCFM robustness uses $d = 64$. The real evaluation bank contains 2,048 disjoint images. A no-gradient full calibration rollout first produces endpoint cotangents $q_i = \partial\mathcal{D}/\partial\hat{x}_{1,i}$. Identical noises are then replayed in chunks, differentiating $\sum_i q_i^\top F_{\theta,\phi}^{S,B}(x_{0,i})$. Since the parameters and discrete map are unchanged between cotangent construction and replay, the summed gradient is the same first-order gradient as retaining the entire graph. Replay batch size is not a statistical hyperparameter.

The headline C2OT shared stage uses AdamW for 192 updates, field/clock learning rates $10^{-5}/10^{-3}$, weight decay 10^{-4}, global gradient clip 5, 128 calibration noises, replay batch 16, and FM batch 128. It updates the time embedding, output blocks 10–11, and output head (3,476,547 field parameters). Stage II freezes every field tensor and runs 48 alternating updates, 16 per solver, with otherwise identical optimization settings.

Exact replay training loop. Each update follows this fixed sequence:

1. select the preregistered NFE budget and draw the full calibration noise batch;
2. run the exact deployed solver without gradients and construct the terminal-distribution endpoint cotangents;
3. replay the identical noises in chunks through the unchanged solver and accumulate each cotangent–endpoint inner-product gradient;
4. accumulate path-FM and linear-field anchor gradients on their independent real/noise batch;
5. clip once, update the permitted path and late-field parameters once, and leave every other parameter frozen;
6. evaluate only on disjoint noises; promotion to image metrics follows the frozen proxy and high-NFE safety gates.

No endpoint, chunk, budget, or arm is selected from a partial image result.

4 PRIMARY FIVE-ARTIFACT EVIDENCE

The statistical unit is the independently post-trained artifact, not an image or a KID subset. Artifact seeds are 20267511–20267515. Each FID evaluation uses generation seed artifact+100, reset for every paired arm and budget. All arms use the same official original-FM checkpoint, 64 updates, exact-Heun NFE8/16 during training, and untouched NFE8/12/24 FID10k evaluation. The full family is displayed to avoid selective reporting. Because $n = 5$, Student-t intervals are accompanied by raw differences and sign counts; they should not be interpreted as high-powered normal-theory inference. No multiplicity adjustment is used, and only the preregistered time/full versus trained-linear contrasts are confirmatory.

Table 1: All primary original-FM FID10k artifact values. Lower is better.

NFE	Artifact seed	Frozen time	Frozen full	Trained linear	Trained time	Trained full	Trained grid
8	20267511	86.715	83.167	28.808	22.359	22.830	28.255
8	20267512	87.040	83.361	28.851	23.033	22.952	27.245
8	20267513	83.518	81.134	27.612	24.116	23.516	30.170
8	20267514	87.954	84.298	27.572	23.771	24.825	29.523
8	20267515	83.168	82.936	28.845	22.926	22.098	27.645
12	20267511	61.389	60.290	12.813	11.222	11.203	12.505

Table 2: Primary paired gains, defined as trained-linear FID minus arm FID. Student-t intervals use the five artifact differences; wins are exact signs.

NFE	Arm	Mean gain	95% CI	Wins
8	Frozen time	-57.342	[-60.217, -54.466]	0/5
8	Frozen full	-54.642	[-56.163, -53.121]	0/5
8	Trained time	5.096	[3.423, 6.770]	5/5
8	Trained full	5.093	[3.066, 7.121]	5/5
8	Trained grid	-0.230	[-2.587, 2.127]	3/5
12	Frozen time	-47.444	[-49.077, -45.811]	0/5
12	Frozen full	-47.390	[-48.450, -46.330]	0/5
12	Trained time	1.407	[0.960, 1.854]	5/5
12	Trained full	1.360	[0.858, 1.862]	5/5
12	Trained grid	0.095	[-0.484, 0.675]	3/5
24	Frozen time	-22.938	[-23.626, -22.250]	0/5
24	Frozen full	-23.744	[-24.305, -23.182]	0/5
24	Trained time	0.712	[0.186, 1.238]	5/5
24	Trained full	0.533	[0.215, 0.850]	5/5
24	Trained grid	0.346	[-0.012, 0.704]	5/5

NFE	Artifact seed	Frozen time	Frozen full	Trained linear	Trained time	Trained full	Trained grid
12	20267512	61.080	60.923	13.305	11.525	11.750	12.813
12	20267513	59.262	59.217	12.839	12.016	11.821	13.440
12	20267514	61.618	61.545	12.915	11.538	12.070	13.074
12	20267515	58.695	59.802	12.954	11.491	11.183	12.516
24	20267511	31.483	31.602	8.191	7.432	7.442	7.695
24	20267512	31.903	32.081	8.302	7.745	7.979	8.174
24	20267513	30.630	32.307	8.238	7.946	7.834	8.096
24	20267514	31.913	32.036	8.855	7.446	7.990	8.069
24	20267515	30.389	32.319	8.041	7.498	7.718	7.861

Table 4: Checkpoint-by-order paired gains (linear minus learned-path FID).

Checkpoint	NFE/path	Mean gain	95% CI	Wins
fm	8/cubic	4.989	[3.141, 6.838]	5/5
fm	8/order 16	9.078	[7.558, 10.598]	5/5
fm	12/cubic	1.376	[0.808, 1.943]	5/5
fm	12/order 16	3.215	[2.691, 3.740]	5/5
fm	24/cubic	0.477	[0.389, 0.564]	5/5
fm	24/order 16	1.015	[0.829, 1.200]	5/5
otcfm	8/cubic	4.129	[3.560, 4.697]	5/5
otcfm	8/order 16	7.355	[6.538, 8.172]	5/5
otcfm	12/cubic	1.099	[0.803, 1.395]	5/5
otcfm	12/order 16	2.984	[2.775, 3.193]	5/5
otcfm	24/cubic	0.564	[0.307, 0.821]	5/5
otcfm	24/order 16	1.293	[0.858, 1.728]	5/5

Table 3: All checkpoint-by-order FID10k artifact values.

Checkpoint	NFE	Artifact seed	Linear	Cubic	Order 16
fm	8	20267711	28.748	22.514	18.395
fm	8	20267712	27.941	22.763	19.531
fm	8	20267713	28.675	23.090	19.153
fm	8	20267714	27.029	24.614	19.736
fm	8	20267715	29.662	24.125	19.849
fm	12	20267711	13.094	11.424	9.442
fm	12	20267712	12.787	11.488	9.745
fm	12	20267713	13.214	11.480	9.813
fm	12	20267714	12.766	12.153	10.199
fm	12	20267715	13.215	11.655	9.801
fm	24	20267711	8.141	7.681	7.275
fm	24	20267712	8.042	7.644	6.822
fm	24	20267713	8.010	7.422	7.135
fm	24	20267714	8.163	7.673	7.137
fm	24	20267715	7.930	7.482	6.843
otcfm	8	20267711	18.882	14.727	12.232
otcfm	8	20267712	19.269	14.963	11.189
otcfm	8	20267713	19.461	16.126	12.475
otcfm	8	20267714	19.769	15.376	12.746
otcfm	8	20267715	19.301	14.846	11.266
otcfm	12	20267711	10.852	9.630	8.119
otcfm	12	20267712	10.915	9.729	7.742
otcfm	12	20267713	10.947	10.197	7.993
otcfm	12	20267714	10.961	9.601	7.864
otcfm	12	20267715	10.708	9.729	7.745
otcfm	24	20267711	8.112	7.415	6.841
otcfm	24	20267712	8.008	7.652	7.060
otcfm	24	20267713	8.204	7.360	6.499
otcfm	24	20267714	8.159	7.626	6.571
otcfm	24	20267715	7.933	7.544	6.979

Table 5: Every absolute projected terminal-risk cell in the factorial. R_{PFO} uses path, field, objective bit order. The $O = 0$ and $O = 1$ frozen-linear cells coincide by construction.

NFE	Seed	Objective	R_{000}	R_{100}	R_{010}	R_{110}
8	20266401	fm	29.179	38.438	31.214	37.850
8	20266401	rollout	29.179	17.920	6.632	5.709
12	20266401	fm	16.933	20.538	18.236	19.972
12	20266401	rollout	16.933	13.612	3.952	3.593
16	20266401	fm	12.702	14.004	13.971	13.580
16	20266401	rollout	12.702	11.354	3.485	3.325
24	20266401	fm	8.284	8.278	8.746	7.840
24	20266401	rollout	8.284	8.484	3.304	3.130
8	20266402	fm	29.534	38.000	30.197	37.635
8	20266402	rollout	29.534	17.542	6.482	5.553
12	20266402	fm	17.251	20.856	17.586	21.043
12	20266402	rollout	17.251	14.228	4.000	3.651
16	20266402	fm	12.845	13.864	12.800	13.589
16	20266402	rollout	12.845	11.961	3.411	3.175
24	20266402	fm	8.346	8.427	8.434	8.095
24	20266402	rollout	8.346	8.315	3.361	3.136
8	20266403	fm	29.150	38.395	29.560	37.538
8	20266403	rollout	29.150	16.872	6.267	5.087
12	20266403	fm	16.844	20.265	17.970	19.851
12	20266403	rollout	16.844	13.469	4.190	3.809
16	20266403	fm	13.106	14.170	14.053	13.914
16	20266403	rollout	13.106	11.937	3.655	3.364
24	20266403	fm	8.357	8.376	8.819	8.068
24	20266403	rollout	8.357	8.607	3.474	3.344

5 CHECKPOINT AND PATH-ORDER TRANSFER

The checkpoint-by-order cross independently trains linear, cubic, and order-16 arms with five artifact seeds on each public TorchCFM checkpoint. Means are paired FID10k values; every cubic and order-16 gain over linear has a 95% artifact-level interval above zero. The cross removes the checkpoint/order confound but does not establish generic curvature: the matched decomposition below shows that the stable order-16 increment is time reparameterization, while full-minus-time and learned-grid intervals include zero.

Table 6: Complete checkpoint-by-order mean comparison. Lower is better.

Checkpoint	Path	NFE8	NFE12	NFE24
original-FM	linear	28.411	13.015	8.057
original-FM	cubic	23.421	11.640	7.581
original-FM	order 16	**19.333**	**9.800**	**7.043**
OT-CFM	linear	19.336	10.877	8.083
OT-CFM	cubic	15.208	9.777	7.519
OT-CFM	order 16	**11.982**	**7.893**	**6.790**

6 MATCHED VSFM REPAIR AND ALL ARTIFACT VALUES

The first matched-VSFM execution correctly shared solver, terminal rollout, late-field subset, updates, and calibration, but a canonical cost audit found 212,992 model sample-evaluations for learned field–clock adaptation and 204,800 for the fixed arms. The difference was one 8,192-sample linear-FM anchor that the trainer applied only when a path object existed. That complete metric run is retained with `complete=false` and cannot support the matched claim.

Before producing a replacement result, we froze seeds 20269011–20269015 and forced two independent FM batches in every arm: active-path FM plus the 0.1-weight linear-FM anchor. All repaired artifacts have 1,416,323 trainable field parameters, 64 updates, exact-Heun NFE8/16 training, and exactly 212,992 model sample-evaluations. Schedule-node hashes are identical across seeds within each arm and field hashes are recorded separately.

Learned-time improvement over VSFM is 11.895 [11.123,12.667] at NFE8 and 2.386 [2.240,2.531] at NFE12, both 5/5. At NFE24 it is -0.948 [-1.216,-0.680], 0/5. With five nonzero pairs, the two-sided

Table 7: Every repaired matched-schedule FID10k value. Each NFE block is uniform/VSFM/learned-time with the identically adapted field.

	NFE8			NFE12			NFE24		
Seed	U	V	L	U	V	L	U	V	L
20269011	29.615	35.976	24.251	13.639	14.206	11.994	8.470	6.264	7.538
20269012	27.585	35.305	24.193	12.763	14.296	11.923	8.241	6.971	7.651
20269013	27.903	34.229	21.397	13.025	13.710	11.299	8.551	6.717	7.682
20269014	28.346	34.422	22.407	13.004	13.998	11.457	8.025	6.530	7.389
20269015	28.668	35.819	24.028	13.257	14.536	12.143	8.133	6.574	7.535
Mean	28.423	35.150	**23.255**	13.138	14.149	**11.763**	8.284	**6.611**	7.559

exact sign value is 0.0625; intervals are reported as effect-size summaries rather than a claim of high-powered asymptotics.

Table 8: Independent DINOv2 metrics for repaired VSFM/learned-time artifacts. Each entry is VSFM/learned clock. Lower FD and KID are better; higher precision and recall are better.

	NFE8				NFE12			
Seed	FD	KID	Prec.	Recall	FD	KID	Prec.	Recall
20269011	681.194/583.526	.9740/.8618	.6143/.6582	.3195/.2925	473.841/450.808	.6079/.6010	.5711/.6210	.4256/.3775
20269012	677.959/581.338	.9657/.8612	.6162/.6519	.3202/.2876	465.291/459.224	.5975/.6160	.5625/.6280	.4243/.3785
20269013	681.807/583.260	.9842/.8678	.6077/.6653	.3506/.3032	477.386/461.010	.6219/.6180	.5712/.6326	.4436/.3805
20269014	675.490/586.594	.9788/.8726	.6157/.6668	.3416/.2918	473.588/459.647	.6100/.6228	.5704/.6273	.4431/.3692
20269015	677.213/585.578	.9695/.8719	.6087/.6516	.3238/.2998	475.867/459.143	.6162/.6183	.5676/.6272	.4338/.3888

At NFE8, the learned clock improves FD by 94.674, KID by 0.1074, and precision by 0.0462, all 5/5, while recall changes by -0.0362, 0/5. At NFE12, FD and precision improve by 15.228 and 0.0587, both 5/5; KID is unresolved and recall changes by -0.0552, 0/5. This is a distribution quality–coverage trade-off, not representation-independent dominance.

7 Order-16 component decomposition

The checkpoint-by-order cross originally compared linear, cubic full path, and order-16 full path. It established a capacity effect but did not localize that effect. The reviewer-requested matched decomposition was frozen on 2026-08-18 before new results: on original-FM and seeds 20267711–20267715, compare linear, order-16 time-only, full path, and learned grid under identical training and FID10k streams. Nominal/functional path dimensions are $0/0$, $16/15$, $32/30$, and $16/15$. Full execution details are in `docs/order16_path_decomposition_protocol.md`.

Table 9: All order-16 component-decomposition FID10k values. Linear and full rows are the already frozen checkpoint-cross artifacts; time and grid are the reviewer-requested matched additions.

NFE	Seed	Linear	Time16	Full16	Grid16
8	20267711	28.748	19.401	18.395	26.693
8	20267712	27.941	19.558	19.531	28.376
8	20267713	28.675	19.270	19.153	28.479
8	20267714	27.029	21.191	19.736	28.515
8	20267715	29.662	20.802	19.849	28.736
12	20267711	13.094	9.515	9.442	12.736
12	20267712	12.787	9.682	9.745	13.099
12	20267713	13.214	9.632	9.813	13.076
12	20267714	12.766	10.259	10.199	13.311
12	20267715	13.215	9.927	9.801	12.909
24	20267711	8.141	7.134	7.275	8.273
24	20267712	8.042	6.996	6.822	8.069

NFE	Seed	Linear	Time16	Full16	Grid16
24	20267713	8.010	7.201	7.135	7.712
24	20267714	8.163	7.037	7.137	8.116
24	20267715	7.930	6.924	6.843	7.765

Time-only beats linear in 5/5 artifacts with paired gains 8.366 [6.538,10.195], 3.212 [2.662,3.763], and 0.999 [0.854,1.144]. Full-path incremental gains over time-only are 0.712 [-0.053,1.477], 0.003 [-0.151,0.157], and 0.016 [-0.148,0.180]. Learned-grid gains over linear are 0.251 [-1.415,1.918], -0.011 [-0.506,0.484], and 0.070 [-0.138,0.278]. Order-16 capacity improves time reparameterization; independent scale/path geometry and learned nodes remain unresolved.

8 METRIC-ALIGNMENT AUDIT

The training loss uses frozen Inception features, so CleanFID improvements can reflect representation alignment. The preregistered independent audit instead uses ImageNet VGG16 conv5 global-average-pooled 512-D features, standardized only with all 10,000 CIFAR-10 test images. This test split is absent from model training, calibration, proxy selection, and headline FID statistics.

Table 10: Five-artifact independent VGG audit; 10,000 paired samples per cell. Intervals are over independently post-trained artifacts. Positive improvement favors full path–field adaptation.

NFE	Metric	Linear mean	Full mean	Paired improvement [95% CI]
8	KID	0.30456	0.23984	0.06472 [0.03195,0.09749]
8	Precision	0.84952	0.85280	0.00328 [-0.00675,0.01331]
8	Recall	0.11646	0.14890	0.03244 [0.02919,0.03569]
12	KID	0.10095	0.08927	0.01168 [0.00008,0.02327]
12	Precision	0.78828	0.79324	0.00496 [-0.00218,0.01210]
12	Recall	0.29066	0.29894	0.00828 [0.00152,0.01504]

KID and recall improve in 5/5 artifacts at both budgets. Precision improves in 4/5, with intervals including zero. Within each artifact, KID uses 100 deterministic subsets of 1,000 only to estimate the KID expectation; those subsets are not treated as training replicates. The fixed grid below predates this extension and shows one prespecified artifact/noise set.

At NFE8, FD, KID, and recall improve in 5/5 artifacts; the precision interval includes zero. At NFE12 all four intervals include zero. Together with VGG, this argues against gross low-NFE Inception metric gaming but does not establish representation-independent dominance (Oquab et al., 2024; Stein et al., 2023).

8.1 REPRESENTATION AUDIT FOR THE DEPLOYMENT STATISTIC

The deployment statistic A_S in the main paper is a direction in feature space, so the extractor is part of its definition. We first repeated the frozen Inception diagnostic with five new generation streams and nested development sizes. At 1,024 and 2,048 samples, all 30 receiver–solver–stream signs match the prior disjoint 10k confirmation. At 256/512 samples, only C2OT–RK4 is weak: its mean remains negative, but two streams are indeterminate at each size.

We then reused the same 2,048 generated samples and replaced Inception by raw DINOv2-L/14 CLS and CLIP ViT-B/32 embeddings. Their reference directions use streaming means over all 50,000 CIFAR-10 training images. Table 12 reports the five-stream means, 95% Student-t intervals, and correct non-indeterminate signs relative to the frozen 10k confirmation. The joint gate requires at least 4/5 correct signs and the expected mean sign in all 12 receiver–solver–feature cells.

The registered abstention threshold is also stable. Sweeping $|A_S| \in \{0, .025, .05, .1\}$ preserves every one of the six deployment decisions with no abstentions. Exact enumeration of all $5^5 = 3125$ stream bootstrap means gives the following sign-separated intervals.

The cross-feature audit agrees in 10/12 cells, including 6/6 for the independent DiT receiver. C2OT–Heun/CLIP and C2OT–RK4/DINOv2 are extractor-specific reversals; across the three feature families,

Table 11: Five-artifact DINOv2 audit. Improvement is linear minus trained full for FD/KID and trained full minus linear for precision/recall.

NFE	Metric	Linear	Trained full	Improvement [95% CI]
8	FD	608.784	581.778	27.005 [17.966,36.045]
8	KID	0.91729	0.86310	0.05419 [0.03462,0.07377]
8	Precision	0.67290	0.66632	-0.00658 [-0.01360,0.00044]
8	Recall	0.27108	0.28728	0.01620 [0.01390,0.01850]
12	FD	459.258	456.204	3.054 [-0.482,6.590]
12	KID	0.61384	0.61256	0.00129 [-0.01033,0.01291]
12	Precision	0.62652	0.62868	0.00216 [-0.00116,0.00548]
12	Recall	0.37928	0.37968	0.00040 [-0.00748,0.00828]

Table 12: Stochastic and representation audit of target alignment at NFE12. Each entry is mean [95% CI]; correct signs/5. The Inception column uses the 2,048-sample audit. DINOv2 and CLIP use the identical five streams.

Receiver	Solver	Inception pool-3	DINOv2-L/14	CLIP ViT-B/32
C2OT	Heun	$+.304$ [.256,.352]; 5/5	$+.322$ [.253,.390]; 5/5	$-.626$ [-.640,-.612]; 0/5
C2OT	RK3	$+.884$ [.872,.897]; 5/5	$+.340$ [.325,.355]; 5/5	$+.887$ [.882,.891]; 5/5
C2OT	RK4	$-.179$ [-.233,-.125]; 5/5	$+.265$ [.201,.329]; 0/5	$-.611$ [-.621,-.601]; 5/5
DiT	Heun	$+.313$ [.265,.361]; 5/5	$+.083$ [.036,.130]; 4/5	$+.714$ [.699,.729]; 5/5
DiT	RK3	$+.834$ [.827,.840]; 5/5	$+.113$ [.100,.126]; 5/5	$+.877$ [.875,.879]; 5/5
DiT	RK4	$+.626$ [.614,.639]; 5/5	$+.166$ [.150,.182]; 5/5	$+.876$ [.874,.877]; 5/5

the majority sign matches the confirmation for all six receiver–solver pairs. We therefore define alignment in the representation frozen before development evaluation, which is the operational setting tested by the deployment rule.

9 Objective, feature, and late-field sensitivity

9.1 Alternative terminal loss and FM anchor

A prospectively frozen audit starts from the same original-FM checkpoint and the paired default seeds 20268611–20268613. The `sliced_w2` arm keeps all regularizers but replaces projected Fréchet training by the sum of one-dimensional W_2^2 distances along 64 frozen orthonormal axes. The `fm0` arm removes the path-FM term while retaining terminal rollout, the independent linear-FM anchor, and identity regularization. Both use the same time-only path, late-field subset, 64 updates, Heun NFE8/16 training, and paired evaluation streams as the default. Positive entries below favor the variant over the default.

The three sliced-W2 NFE12 improvements are 0.3824, 0.5479, and 0.4732. Independent DINOv2 metrics reproduce the direction (Table 15). The no-path-FM arm lacks the same cross-seed and cross-metric agreement, showing that the FM term stabilizes terminal optimization rather than serving as a feature-specific score correction.

All six artifacts, six CleanFID cells, and twelve DINO cells pass the frozen completeness checks. The canonical summary SHA-256 is recorded beside `results/remote/paper_a_terminal_objective_ablation_20260919/summary.json`; the protocol SHA-256 is `0b64b7d2c9defcd9a2315b244e5f4e410226141adb354c91d9361d5ff1336a6f`.

9.2 One-factor sensitivity and field placement

This audit was frozen before seeds 20268611–20268613 were inspected. Every comparable arm uses the same time-only path, exact Heun NFE8/16 training, 64 updates, and fixed 64-dimensional evaluation projection. The default absolute projected risks at NFE8/12/16/24 are 5.6658/3.6948/3.3767/3.1724. Table values are variant minus default, so negative is favorable.

Table 13: Exact five-stream bootstrap audit for the frozen Inception-space alignment statistic. All percentile intervals exclude zero.

Receiver	Solver	Mean alignment	95% interval
C2OT	Heun	$+0.30407$	$[+0.27468, +0.33305]$
C2OT	RK3	$+0.88427$	$[+0.87618, +0.89237]$
C2OT	RK4	-0.17911	$[-0.20988, -0.14222]$
Corrected DiT	Heun	$+0.31309$	$[+0.28259, +0.34223]$
Corrected DiT	RK3	$+0.83350$	$[+0.82925, +0.83729]$
Corrected DiT	RK4	$+0.62649$	$[+0.61785, +0.63363]$

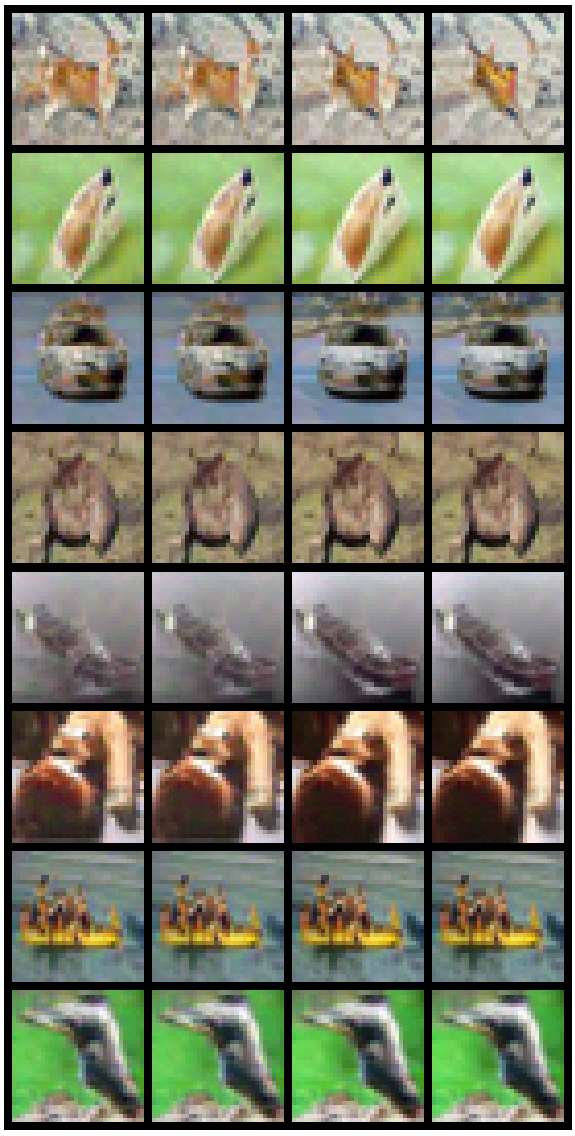

Figure 1: Fixed, non-cherry-picked noises. Columns are linear NFE8, joint NFE8, linear NFE12, and joint NFE12.

The default subset is strictly better at NFE8 than all three smaller nested subsets; adding the time embedding beyond the last-two-block subset is useful. Full-field adaptation is favorable in 3/3 NFE8 streams and uses 25.2 times as many trainable parameters, but its three-seed interval narrowly includes zero and no higher-budget contrast resolves. Mean default/full training times are 414.7/410.7

Table 14: Paired CleanFID10k improvements from the frozen terminal-objective audit. Parentheses report favorable artifacts out of three.

Arm	NFE8	NFE12	NFE24
No path-FM	$0.037\ [-3.615, 3.689]$ (1)	$0.074\ [-0.875, 1.023]$ (1)	$0.131\ [-0.292, 0.554]$ (2)
Sliced-W2	$0.823\ [-0.742, 2.389]$ (3)	$\mathbf{0.468}\ [\mathbf{0.262}, \mathbf{0.674}]$ (3)	$0.142\ [-0.302, 0.586]$ (2)

Table 15: DINOv2 improvement of sliced-W2 training over the paired default. Intervals are artifact-level 95% Student-t intervals; parentheses give favorable signs out of three.

Metric	NFE8	NFE12
FD ↓	$8.929\ [3.385, 14.474]$ (3)	$5.483\ [-0.220, 11.187]$ (3)
KID ↓	$.02306\ [.02022, .02591]$ (3)	$.01324\ [-.00080, .02729]$ (3)
Precision ↑	$-.00103\ [-.01146, .00939]$ (1)	$-.00263\ [-.01201, .00675]$ (1)
Recall ↑	$.00857\ [.00569, .01145]$ (3)	$.00817\ [.00033, .01600]$ (3)

seconds and peak allocated memory is 13.34/13.94 GiB, so the observed advantage of the default is parameter count, not wall clock.

Fixed layer-position controls. The plan fixed three additional positions before their outputs and prohibited using them to replace the default. Table values are default risk minus control risk, so positive favors the control. The substantially larger early/middle controls lose every cell, supporting a late-location effect rather than a parameter-count explanation. Removing the head at nearly matched capacity is suggestively favorable at NFE8, neutral at NFE12, and slightly unfavorable at NFE24. Thus the late region is supported, but the output head is not isolated as universally necessary.

FID-free, exact-capacity distributed placement. As a subsequent capacity-matched refinement, we rank field tensors by median per-parameter update energy computed from three existing full-field development artifacts. This ranking reads neither FID nor any confirmation artifact. The top 97 tensors contain exactly 1,416,323 trainable parameters, matching the default late subset parameter for parameter. Table 18 reports five new paired training artifacts; positive values are default FID minus distributed-subset FID.

The two low-budget intervals exclude zero and all paired signs agree. At NFE24 the mean remains favorable but unresolved, so the refinement is a low-budget placement result rather than a new universal field subset. The canonical summary SHA256 is `b0e646d556ed14b8bdc1b5d9678c2ed8688b55416f2cd66fca15e9ac599878bd`.

Common-metric robustness. The 4,096-real-bank and altered-projection proxy scores were not comparable because they changed the whitening reference or metric subspace. We trained the missing same-seed linear reference and evaluated every frozen arm with common FID10k and DINOv2 streams.

Common-witness projection audit. We also re-evaluated the 16-, 32-, and 64-dimensional training variants under one frozen 64-dimensional witness. This comparison uses identical generation banks, target moments, whitening, and projection columns. Table 19 reports default-64 minus lower-dimensional risk, so negative values favor the default configuration. At the primary NFE12 budget, default-64 wins all six paired seed comparisons. This common-witness result supports the retained 64-dimensional configuration without comparing incompatible native objectives.

Table 20 reports linear minus variant FID for the broader one-factor audit, so positive values favor field–clock adaptation.

All 72 FID pairs favor the variant (8 arms $\times$ 3 budgets $\times$ 3 artifacts). Table 21 gives DINO improvements versus the same linear artifact as mean(signs out of three); lower-is-better metrics use linear minus variant and higher-is-better metrics use variant minus linear.

Every frozen construction preserves the primary NFE8 FID/FD/KID/recall direction, while precision falls for all eight. NFE12 DINO is projection-seed dependent. Full field is a stronger distribution-

Table 16: Three-artifact one-factor and nested-subset audit on the common proxy. The bracket is the paired 95% interval at NFE8.

Arm	Field params	ΔR_8 [95% CI]	ΔR_{12}	ΔR_{16}	ΔR_{24}
FM weight 0.5	1.416M	-0.171 [-0.429,0.086]	-0.040	+0.007	+0.055
FM weight 2	1.416M	-0.013 [-0.319,0.293]	-0.013	-0.002	+0.027
Anchor weight 0	1.416M	-0.045 [-0.208,0.118]	-0.035	-0.013	+0.023
Anchor weight 0.3	1.416M	-0.170 [-0.659,0.318]	-0.049	-0.020	+0.007
Identity weight 0	1.416M	+0.154 [0.068,0.240]	+0.041	+0.011	+0.063
Identity weight 0.01	1.416M	-0.155 [-0.250,-0.059]	-0.040	-0.017	+0.036
Head only	0.0037M	+7.147 [6.314,7.980]	+5.491	+4.083	+2.120
Last block + head	0.546M	+1.306 [0.875,1.737]	+0.646	+0.343	+0.060
Last two blocks + head	1.088M	+0.496 [0.098,0.895]	+0.225	+0.079	-0.008
Default + time embedding	1.416M	0	0	0	0
Full field	35.746M	-1.016 [-2.037,0.004]	-0.273	-0.151	+0.040

Table 17: Capacity/position controls in projected terminal risk. Intervals are paired over the same three seeds.

Control	Field params	Δ_8	Δ_{12}	Δ_{24}
Early blocks 0–1 + head	4.401M	-0.932 [-1.270,-0.594]	-0.760 [-0.848,-0.672]	-0.552 [-0.629,-0.474]
Middle blocks 5–6 + head	5.255M	-0.413 [-0.786,-0.040]	-0.453 [-0.609,-0.296]	-0.390 [-0.544,-0.235]
Late blocks 10–11, no head	1.413M	+0.193 [-0.188,0.575]	+0.032 [-0.075,0.140]	-0.036 [-0.072,-0.001]

quality point but uses 25.2 times the parameters and retains a precision trade-off. The defensible claim is low-budget feature-construction robustness, not representation-independent dominance.

10 Finite-step objective versus segmented straightness

We compare the terminal rollout objective with an equal-capacity segmented flow-map straightness control on three fresh CIFAR-100 streams. Both arms use the same six-parameter monotone cubic path, 1.416M-parameter late-field subset, 64 updates, data, anchors, and evaluation. The control also follows the image-level stabilization rules reported by Learning Straight Flows: 16 FM-only warmup updates, linear annealing of the straightness weight, field learning rate 10^{-4}, and path learning rate 5×10^{-6}.

The learned straightness path is neutral relative to evaluating its own updated field on the linear path (6.6568 versus 6.6596 at NFE12 and 2.5962 versus 2.5909 at NFE24). Thus the result supports objective alignment in this lightweight post-training regime, not superiority of our spline geometry. It is not an official reproduction: the original interpolant, optimal-field estimator, training scale, and architecture differ. We retain it as a projected-feature mechanism ablation and do not promote the collapsed control to FID.

11 Factorial definitions and absolute values

Let $R(P, F, O; B)$ be projected terminal Fréchet risk, where P, F, O encode spline path, adapted field, and rollout objective. Relative path gain is $100[R(0, F, O) - R(1, F, O)]/R(0, F, O)$. At fixed objective,

$$I_O = [R(0, 1, O) - R(1, 1, O)] - [R(0, 0, O) - R(1, 0, O)], \qquad T = I_1 - I_0.$$

Positive I_O means field adaptation makes path learning more valuable in absolute risk units; positive T means rollout strengthens that complementarity beyond FM. Table 5 contains every absolute seed-level cell. These are mechanism estimates from three streams, not FID or a promotion gate. At NFE8/12/16, T is negative in all seeds, so the result is low-budget substitution rather than universal synergy. At NFE24, field adaptation enables a useful path under both objectives, while T remains negative.

Table 18: Exact-capacity delta-energy placement on five fresh artifacts. The bracketed summary is the paired 95% Student-t interval.

Seed	NFE8		NFE12		NFE24	
	Default	Distributed	Default	Distributed	Default	Distributed
20272511	23.114	21.763	11.682	11.214	7.883	7.781
20272512	23.244	20.947	11.489	10.772	7.449	7.472
20272513	22.572	21.130	11.317	10.738	7.606	7.482
20272514	22.656	21.353	10.888	10.735	7.467	7.474
20272515	22.674	21.679	11.320	10.931	7.910	7.834
Mean	22.852	21.374	11.339	10.878	7.663	7.608
Gain	1.478 [0.872, 2.084] (5/5)		.461 [.198, .724] (5/5)		.054 [−.027, .136] (3/5)	

Table 19: Projection-dimension sensitivity under one frozen 64-dimensional witness. Parentheses report default wins among three adaptation seeds.

NFE	Default64 − proj16 [95% CI]	Default64 − proj32 [95% CI]
8	−.6593 [−1.1018, −.2168] (3/3)	−.0791 [−.5922, .4341] (2/3)
12	−.3207 [−.5962, −.0452] (3/3)	−.2017 [−.4169, .0136] (3/3)
16	−.2187 [−.4784, .0410] (3/3)	−.2107 [−.3157, −.1057] (3/3)
24	−.1026 [−.2902, .0850] (3/3)	−.2136 [−.3666, −.0607] (3/3)

12 Universal artifact: exact meaning of the gap

The 128-update universal artifact alternates NFE8/16 and is compared with two 64-update specialists, matching total update count. Define

$$G_B = 100\frac{R_{\mathrm{univ}}(B) - R_{\mathrm{ref}}(B)}{R_{\mathrm{ref}}(B)},$$

where R is projected terminal risk. At NFE8/16 the reference is the matching specialist; elsewhere it is the lower-risk specialist selected by the earlier frozen 512-sample evaluation. Hence "oracle envelope" refers to this fixed surrogate reference, not CleanFID. Gaps at NFE6/8/10/12/16/24 are $-6.63, -7.85, -9.93, -11.44, -7.17, +0.90$ percent, with paired wins $3/3, 3/3, 3/3, 3/3, 3/3, 2/3$. A fresh 64-update universal control has mean gap $+8.21$ percent, so the result requires matched total portfolio compute.

For the read-only decomposition,

$$R_{\mathrm{univ}}(B) - R_{\mathrm{ref}}(B) = [R_{\mathrm{univ}}(128) - R_{\mathrm{ref}}(128)] + [E_{\mathrm{univ}}(B) - E_{\mathrm{ref}}(B)],$$

where $E(B) = R(B) - R(128)$. The universal artifact wins all 18 audited seed-by-budget cells. Combined gradient cosines for NFE8–12, 8–16, and 12–16 are 0.226 ± 0.257, -0.083 ± 0.158, and 0.014 ± 0.479; path-only 12–16 is negative in 3/3 seeds. The shared model succeeds despite this local gradient conflict.

13 Conditional fidelity and higher-resolution stress test

The C2OT audit uses 10,000 balanced requested labels and a real-only CIFAR-10 ResNet-18 with 93.47% test accuracy. Generated images never train the classifier. Agreement/cFID for linear versus joint are 92.02/0.642 versus 92.55/0.555 at NFE8, 93.87/0.291 versus 93.99/0.292 at NFE12, and 94.60/0.201 versus 94.51/0.192 at NFE24. Thus the overall FID gain is not explained by conditional collapse.

At NFE8 the mean gain is 0.0344 with interval [0.0103,0.0586], but its relative magnitude is only 0.093%. Artifact and generation stream are paired one-to-one rather than crossed, fitting uses the released ADM statistics, and no independent feature space was evaluated. The correct conclusion is a tiny target-aware stress-test effect, not resolved modern-architecture transfer.

Table 20: P1 common FID10k paired improvements with 95% intervals over three artifacts. Every raw pair is retained in the canonical summary.

Arm	NFE8	NFE12	NFE24
Default	5.075 [3.007,7.142]	1.405 [0.999,1.810]	.673 [.505,.841]
Path identity 0	4.423 [3.419,5.427]	1.155 [.662,1.649]	.577 [.163,.990]
Path identity .01	5.886 [3.175,8.597]	1.625 [1.143,2.106]	.545 [-.140,1.231]
Bank 4096	6.628 [5.005,8.251]	1.780 [1.028,2.532]	.503 [-.143,1.149]
Projection 256	6.308 [2.973,9.643]	1.685 [.863,2.507]	.802 [.364,1.239]
Projection seed 11992	6.020 [4.592,7.448]	1.601 [1.255,1.946]	.630 [-.579,1.838]
Projection seed 11993	5.595 [4.780,6.409]	1.316 [-.192,2.823]	.612 [-.242,1.465]
Full field	11.779 [9.129,14.428]	3.881 [2.127,5.634]	.820 [-.674,2.313]

Table 21: P1 common DINOv2 improvements versus paired trained linear.

	NFE8				NFE12			
Arm	FD	KID	Prec.	Recall	FD	KID	Prec.	Recall
Default	25.245(3)	.0519(3)	-.0149(0)	.0153(3)	2.311(2)	.0007(2)	.0005(2)	-.0017(1)
Path identity 0	22.047(3)	.0465(3)	-.0141(0)	.0178(3)	1.349(2)	.0006(2)	.0011(2)	-.0009(1)
Path identity .01	27.470(3)	.0536(3)	-.0140(0)	.0194(3)	1.331(1)	-.0036(1)	.0040(2)	-.0024(1)
Bank 4096	31.681(3)	.0620(3)	-.0128(0)	.0201(3)	1.527(2)	-.0035(1)	.0026(2)	.0000(2)
Projection 256	33.703(3)	.0636(3)	-.0206(0)	.0212(3)	1.518(2)	-.0041(1)	-.0016(0)	-.0015(1)
Projection seed 11992	26.069(3)	.0463(3)	-.0171(0)	.0147(3)	-1.119(1)	-.0106(1)	.0031(3)	-.0023(1)
Projection seed 11993	20.169(3)	.0323(3)	-.0151(0)	.0118(3)	-7.240(0)	-.0221(0)	.0016(2)	-.0038(1)
Full field	89.022(3)	.1807(3)	-.0445(0)	.0258(3)	39.539(3)	.0705(3)	-.0048(0)	-.0024(1)

14 Adjoint-weighted propagated-defect audit

After freezing the main artifacts, we evaluate five fresh 32-noise analysis streams at Heun NFE8/12/24. For every interval, the local estimate is the one-step/two-half-step difference divided by three. We propagate the terminal projected-Fréchet cotangent backward through the exact coarse solver and report both the raw absolute contraction and $|\lambda^\top d|/(\|\lambda\|\sqrt{D})$, which removes terminal-gradient scale. No artifact is retrained or selected.

Field–clock adaptation improves over frozen in all 15 seed–budget cells. Unit-adjoint paired improvement intervals are [0.000366,0.000682] at NFE8 and [0.000144,0.000418] at NFE12; NFE24 improves 32.3%. Raw absolute adjoint contractions fall from 0.5692/0.3701/0.1567 to 0.2726/0.2067/0.0777. Thus the result is not solely smaller terminal-gradient scale. Disabling the clock on the same joint field worsens the normalized NFE8 statistic in 4/5 streams, but the interval includes zero and NFE12/24 are unresolved. This bounds the mechanistic claim: joint training lowers terminal-direction-relevant local error, while the clock's large image-level effect also changes distributional coverage in ways not captured by this first-order scalar.

15 Corrected official-DiT confirmation

The initially used upstream endpoint was invalidated by a BF16 EMA-accumulator and time-response audit. We resumed the official 65.65M-parameter model to one million updates and used the corrected EMA. Its 50k scores are 4.524 under the upstream script's default `clip_vit_b_32` metric and 18.521 under Inception CleanFID. Five development seeds (20268101–20268105) were used only to diagnose the frozen gate. Five disjoint holdout seeds (20268201–20268205), paired with generation seeds 20269201–20269205, then repeated the fixed three-arm protocol and alone authorized image metrics.

Against field-only, field–clock adaptation improves Inception FID by 27.528 [26.002,29.054] and CLIP-FID by 3.619 [2.253,4.985], with 5/5 wins for both. It improves DINO recall by 0.0170 [0.0075,0.0265], but loses 23.26 DINO-FD, 0.0794 KID, and 0.0261 precision, each consistently across the five streams. Against frozen, however, its DINO-FD and KID improve by 12.85% and

Table 22: Projected terminal risk for the literature-stabilized, equal-capacity straightness control and finite-step terminal training. Lower is better.

NFE	Straightness	Finite-step	Relative gain	Paired wins
8	12.7979	**2.3569**	81.58%	3/3
12	6.6568	**1.6123**	75.78%	3/3
24	2.5962	**1.2878**	50.40%	3/3

Table 23: All target-aware SelfFlow ImageNet-256 FID10k pairs.

NFE	Artifact	Generation	Uniform	Learned	Gain
8	20267811	20267911	36.9621	36.9544	0.0078
8	20267812	20267912	36.8759	36.8516	0.0243
8	20267813	20267913	37.2533	37.2123	0.0410
8	20267814	20267914	37.2404	37.2008	0.0395
8	20267815	20267915	36.4549	36.3953	0.0596
12	20267811	20267911	21.7258	21.6641	0.0617
12	20267812	20267912	21.5956	21.5785	0.0171
12	20267813	20267913	21.5491	21.4962	0.0528
12	20267814	20267914	21.8656	21.8931	-0.0276
12	20267815	20267915	20.9433	21.0299	-0.0866

23.01%, and recall rises by 0.1212. These independent metrics define a coverage– precision boundary rather than invalidating the FID confirmation.

The minimal clock-off control reuses the same saved NFE8 noises and changes no parameter. It raises Inception-FID from 39.218 to 70.750, a paired degradation of 31.531 [30.407,32.655] with 5/5 signs. We include this single control only to show that the path parameters remain functionally active; the full cross-pair and runtime-control analysis is reserved for a separate study and is not part of this paper's claim set.

15.1 Budget generalization of the DiT intervention

Without retraining, five disjoint generation streams at each of Heun NFE12 and NFE24 repeat the independently trained field-only and deployed field–clock arms on all six metrics. Means are:

Field–clock adaptation outperforms field-only on both FID families, DINOv2-FD/KID, and recall in all five streams at both budgets; precision is lower in all ten pairs. This endpoint comparison establishes cross-budget architecture transfer without making a separate claim about the internal allocation between clock and field.

16 Additional preregistered FID50k and CIFAR-100 transfer

Before the unified operating-point matrix was inspected, six image folders were frozen for 50,000-sample confirmation. On original-FM, trained linear versus field–clock gives 26.811/20.782 at NFE8 (22.49% improvement) and 11.624/10.244 at NFE12 (11.87%). On the independent C2OT implementation, the corresponding NFE12 values are 8.982/8.150 (9.26%). These single-artifact confirmations are valid but do not estimate adaptation variance; the repeated NFE8 study below supplies that missing statistical unit for the headline original-FM contrast.

The separate CIFAR-100 transfer uses three independently adapted artifacts. It passes a frozen high-accuracy backbone gate (CleanFID10k 16.377 at Heun NFE100), then gives:

The CIFAR-100 result is positive at low NFE and has an explicit NFE24 crossover. It complements rather than replaces the independent C2OT model-family result in the main paper.

17 Artifact-level repeated FID50k

The original-FM NFE8 trained-linear versus field–clock contrast is repeated over five independently post-trained artifacts. Each artifact is evaluated with the same two fresh generation seeds, 20268801

Table 24: Five-artifact propagated-defect means. F/L/TF are frozen, field-only linear, and field–clock. Lower is better.

	Unit-adjoint weighted sum			Raw Richardson RMS sum		
NFE	F	L	TF	F	L	TF
8	.001975	.001507	**.001451**	.03573	.01964	**.01896**
12	.001216	.000992	**.000935**	.01924	.01177	**.01174**
24	.000530	**.000356**	.000359	.00861	**.00424**	.00433

Table 25: Per-holdout NFE8 image metrics. Lower is better except precision and recall. F/Fld/TF denote frozen, field-only, and field–clock.

Seed	Inception F/Fld/TF	CLIP F/Fld/TF	DINO-FD F/Fld/TF	Recall F/Fld/TF
20268201	128.54/67.15/38.64	32.07/20.23/16.87	979.82/842.28/858.58	.148/.263/.286
20268202	130.01/65.61/39.98	32.33/19.82/16.14	986.45/824.84/855.33	.145/.258/.273
20268203	129.36/65.49/38.50	32.24/20.46/15.79	981.08/829.75/847.63	.153/.248/.274
20268204	129.90/67.70/39.68	32.11/18.96/17.05	988.64/835.55/870.71	.163/.255/.263
20268205	129.98/67.78/39.29	32.10/21.02/16.55	989.11/843.72/860.18	.138/.245/.258

and 20268802; each cell contains 50,000 samples and uses Heun plus streaming float64 first and second Inception moments. Generation streams are repeats within an artifact, not additional independent training units.

After averaging generation streams inside each artifact, the paired improvement is 5.020 [3.402,6.640], with 5/5 signs; individual improvements range from 3.548 to 6.170. The mean absolute change caused by switching the generation seed is only 0.069 for trained linear and 0.079 for field–clock. This resolves adaptation randomness for the NFE8 contrast, but does not turn the remaining single-artifact FID50k arms or budgets into replicated comparisons.

18 Euler/Heun joint training

A frozen robustness experiment compares five Euler-native artifacts, five Heun-native artifacts, and five joint artifacts that alternate solver every update while cycling NFE8/16 independently. Every artifact uses 64 updates, the same time-only path and late-field subset, and exactly 212,992 model sample-evaluations. Lower projected terminal risk is better.

Joint training improves over frozen in all 30 solver–seed–budget cells. It is nevertheless classified gray by the preregistered gate: Euler NFE8/12 gain retention is below 80%, and Euler NFE24 degradation is 2.76%, slightly above the 2% safety threshold. The result is a usable positive dual-solver Pareto, not one artifact matching both native optima; no solver-weight search follows.

19 Unified baseline and cost ledger

The single-artifact FID50k matrix is included to compare absolute operating points, not to estimate training randomness. At NFE8/12/24: frozen uniform is 122.284/72.402/31.603; SharpEuler signal plus Heun is 82.232/37.920/9.662; VSFM plus Heun is 77.190/23.131/4.642; trained linear is 26.786/11.018/6.267; trained time/field is 20.562/9.432/5.487; and BézierFlow is 6.186/5.082/3.964. BézierFlow is the clear absolute-quality reference. Our shared artifact defines the teacher-free, low-adaptation-cost operating point quantified below.

The time/field arm uses 212,992 model sample-evaluations: 98,304 no-gradient terminal calls, 98,304 replay-gradient calls, 8,192 path-FM calls, and 8,192 linear-anchor calls. It also performs 16,384 feature forwards and 8,192 feature backwards. Five runs average 392.5 seconds and 13.34 GiB peak allocated memory. The trained-linear arm uses 204,800 model evaluations, 321.1 seconds, and 13.08 GiB. Full path and learned grid average 410.8 and 337.9 seconds. These counts are sample-forwards, not FLOPs.

Table 26: Corrected-DiT endpoint image-budget generalization. Each cell is field-only/field–clock.

Metric	Direction	NFE12	NFE24
Inception-FID	lower	54.79/35.25	42.35/33.02
CLIP-FID	lower	18.15/16.24	15.51/13.88
DINOv2-FD	lower	747.86/724.38	730.88/663.14
DINOv2-KID	lower	1.211/1.157	1.153/1.022
Precision	higher	.739/.717	.710/.690
Recall	higher	.289/.326	.352/.379

Table 27: CIFAR-100 paired FID10k. Lower is better.

NFE	Trained linear	Field–clock	Relative gain	Wins
8	42.244	**37.612**	10.96%	3/3
12	27.432	**26.349**	3.95%	3/3
24	**19.872**	19.928	-0.28%	1/3

On one RTX 6000D, batch 256, three warmups, and ten synchronized repeats, the same adapted field with a linear path takes 1.394/2.090/4.186 ms per image at NFE8/12/24; spline rollout takes 1.422/2.136/4.272 ms. The corresponding path overhead is 2.00/2.02/2.03%, while peak allocated memory remains approximately 1559 MiB for both. Under the identical timing protocol, BézierFlow takes 2.094/3.140/6.282 ms per image and approximately 2010 MiB, making our shared artifact about 1.47 times faster online at each budget. These timings quantify a cost–quality Pareto only and do not override BézierFlow's stronger FID.

20 Reproduction commands and artifact map

All experiments use the provided environment through its `python` executable; no additional environment is required. The project root is denoted by `<project-root>`. Representative commands are:

```
bash scripts/launch_reviewer_artifact_training.sh launch-main
bash scripts/queue_reviewer_artifact_fid.sh launch
bash scripts/launch_reviewer_artifact_training.sh \
  launch-cross-throttled
bash scripts/queue_reviewer_cross_fid.sh launch
bash scripts/launch_order16_decomposition.sh launch
bash scripts/queue_order16_decomposition_fid.sh launch
bash scripts/launch_corrected_dit_transfer_fid.sh launch
bash scripts/launch_corrected_dit_transfer_dinov2.sh launch
bash scripts/launch_paper_a_sensitivity.sh launch
PAPER_A_VSFM_ROOT=results/paper_a_gap/matched_vsfm_exact \
PAPER_A_VSFM_SEED_START=20269011 \
PAPER_A_VSFM_FORCE_LINEAR_ANCHOR=1 \
  bash scripts/launch_paper_a_matched_vsfm.sh launch
bash scripts/launch_paper_a_sensitivity_linear_reference.sh launch
bash scripts/launch_paper_a_repeated_fid50k.sh launch
bash scripts/launch_paper_a_multi_solver_joint.sh launch
bash scripts/launch_paper_a_c2ot_residual_clock_capacity.sh launch
bash scripts/launch_paper_a_c2ot_solverwise_capacity_\
confirmation.sh launch
bash scripts/launch_paper_a_c2ot_solverwise_capacity_\
fid10k.sh launch
bash scripts/launch_paper_a_c2ot_solverwise_capacity_\
seed_robustness.sh launch
bash scripts/launch_paper_a_c2ot_posttraining_motivation.sh launch
bash scripts/launch_paper_a_terminal_objective_ablation.sh launch
```

Table 28: Every repeated FID50k cell. Each entry is trained-linear/field–clock.

Artifact seed	Generation 20268801	Generation 20268802	Artifact-mean gain
20267511	26.8157/20.6552	26.9067/20.7270	6.170
20267512	26.7505/20.8823	26.8093/20.9724	5.853
20267513	25.6051/22.0466	25.6723/22.1353	3.548
20267514	25.0356/21.3732	25.0840/21.4453	3.651
20267515	26.9658/21.0881	27.0453/21.1603	5.881
Mean	26.269/21.249		5.020

Table 29: Aggregate multi-solver result. Gain retention is the joint gain over frozen divided by the native gain; the last column is joint degradation from the native artifact.

Solver	NFE	Frozen	Native	Joint	Retention	Joint/native
Euler	8	5.4175	3.9519	4.8981	35.5%	+23.96%
Euler	12	4.3962	3.2925	3.6208	69.9%	+10.01%
Euler	24	3.5529	2.9463	3.0245	93.1%	+2.76%
Heun	8	28.5397	5.6511	8.5333	87.4%	+51.71%
Heun	12	16.6345	3.7929	4.8108	92.1%	+26.99%
Heun	24	8.3811	3.3268	3.2477	101.5%	-2.34%

```
python scripts/summarize_paper_a_terminal_objective_ablation.py
python scripts/summarize_paper_a_sensitivity_common_metrics.py
python scripts/summarize_paper_a_repeated_fid50k.py
python scripts/summarize_paper_a_multi_solver_joint.py
python scripts/audit_reviewer_gap_closure.py
python scripts/audit_paper_a_final_closeout.py
```

Long jobs are launched through tmux. Canonical aggregate files are `results/remote/main_artifact_statistics.json`, `order_cross_statistics.json`, `results/remote/literature_straightness_control_summary.json`, `results/remote/matched_latency.json`, `results/reviewer_gap/multi_artifact_dinov2/summary.json`, `results/paper_a_gap/sensitivity/summary.json`, `results/paper_a_gap/matched_vsfm/canonical_summary.json` (retained with `complete=false` after the exact-compute audit), `results/paper_a_gap/matched_vsfm_exact/canonical_summary.json`, `results/paper_a_gap/sensitivity/common_metric_summary.json`, `results/paper_a_gap/layer_position/canonical_summary.json`, `results/paper_a_gap/repeated_fid50k/canonical_summary.json`, `results/paper_a_gap/multi_solver/canonical_summary.json`, `results/paper_a_gap/final_closeout_audit.json`, `results/paper_a/c2ot_residual_clock_capacity_20260911/summary.json`, `results/paper_a/c2ot_solverwise_capacity_confirmation_20260911/summary.json`, `results/paper_a/c2ot_solverwise_capacity_fid10k_20260911/summary.json`, `results/paper_a/c2ot_solverwise_capacity_seed_robustness_20260911/summary.json`, `results/paper_a/c2ot_posttraining_motivation_20260914/summary.json`, `results/remote/paper_a_terminal_objective_ablation_20260919/summary.json`, `reviewer_gap/c2ot/conditional_fidelity.json`, and `reviewer_gap/self_flow/fid10k_summary.json`. The legacy audit checks the reviewer-gap conditions; the final closeout audit also verifies P0–P3 completion/gates, seed overlap, artifact hashes, citations, and manuscript evidence markers.

21 High-order solver-matching extension

After the original Euler/Heun audit, we froze explicit Heun, RK3, and classical RK4 implementations under exact field-call accounting. Matrix columns denote the post-training solver, and rows denote the deployment solver. We evaluate the diagonal property before CleanFID. Full protocols and source hashes are in `docs/paper_a_high_order_solver_matching_protocol_20260905`.

Table 30: All Euler evaluations. Each NFE block is frozen/native/joint.

Seed	NFE8			NFE12			NFE24		
20268911	5.6496	4.2406	5.3005	4.5848	3.3824	3.7604	3.5295	2.8932	3.1183
20268912	5.3741	3.9299	5.0590	4.4881	3.2540	3.5905	3.7072	2.7959	3.0159
20268913	5.4242	3.7649	4.8286	4.3138	3.2621	3.7566	3.6094	2.9991	3.0674
20268914	5.8096	4.2788	5.0821	4.5880	3.4488	3.6056	3.5810	3.0545	3.0637
20268915	4.8300	3.5454	4.2203	4.0066	3.1154	3.3910	3.3375	2.9888	2.8570

Table 31: All Heun evaluations. Each NFE block is frozen/native/joint.

Seed	NFE8			NFE12			NFE24		
20268911	28.4211	6.2673	8.9429	16.2238	3.9462	5.0652	8.1708	3.4154	3.4000
20268912	28.8020	5.2343	8.2959	16.7167	3.5418	4.5082	8.6299	3.4346	3.2129
20268913	28.3300	5.7912	8.6226	16.7782	3.9639	5.0064	8.7202	3.4951	3.4309
20268914	28.6763	5.9411	8.4879	17.0969	4.0262	4.8726	8.2961	3.1095	3.1187
20268915	28.4693	5.0216	8.3170	16.3569	3.4864	4.6013	8.0883	3.1793	3.0760

`md`, `docs/paper_a_dit_solver_matching_protocol_20260906.md`, and `docs/paper_a_c2ot_high_order_solver_matrix_protocol_20260905.md`.

Table 32: Matched-versus-strongest-off-diagonal CleanFID10k at exact NFE12.

Family	deployment	matched	challenger	wins
VP-CFM	Heun	12.9441	22.7570	2/2
	RK3	33.8311	100.9206	2/2
	RK4	10.4056	14.8274	2/2
DiT	Heun	26.4844	53.6956	3/3
	RK3	14.8219	59.4546	3/3
	RK4	15.3664	19.1872	3/3
C2OT	Heun	10.4758	11.9217	3/3
	RK3	7.2830	24.5893	3/3
	RK4	9.9599	10.3294	2/3

The C2OT proxy matrix at NFE12 is

	Heun-trained	RK3-trained	RK4-trained
Heun deploy	**1.0293**	2.8724	1.1244
RK3 deploy	1.8596	**0.8616**	2.0142
RK4 deploy	1.0268	1.9921	**0.9905**

and every diagonal wins all 3/3 adaptation seeds. At NFE24, C2OT RK3 remains diagonal in 3/3 units, Heun prefers the RK4-trained artifact, and RK4 is tied with RK3-trained to within 0.000394 mean proxy risk. VP-CFM similarly retains RK3/RK4 matching while Heun saturates. The corrected DiT NFE24 means for matched versus the frozen challenger are 8.4312/10.4647 (Heun), 5.7260/8.4778 (RK3), and 6.5325/6.5747 (RK4), with 3/3, 3/3, and 2/3 wins.

22 Sequential refinement from a shared field–clock base

Each of three new units trains a shared order-16 clock and 3,476,547 late-field parameters for 192 alternating-solver updates. The field is then frozen and hashed. The shared clock is copied exactly into three residual clocks, and only their 48 logits are refined for 48 updates (16 per solver). Training budgets are Heun 8/16, RK3 9/18, and RK4 8/16. Unit seeds are 20281951–20281953 for the base, 20282951–20282953 for refinement, and 20283951–20283953 for the held-out proxy test. All field hashes remain unchanged.

At NFE12 the solver means are 5.3086/5.2265, 6.2473/5.5236, and 4.8832/4.7708 for Heun, RK3, and RK4. Pooled FID50k falls from 5.4797 to 5.1736 (5.586%, 9/9). At NFE24, the corresponding

Table 33: Every sequential C2OT CleanFID50k pair. Each cell is shared base/refined clock under identical noise, labels, field, solver, and exact NFE.

NFE, unit	Heun	RK3	RK4
12, 1	5.2744/5.2261	6.3989/5.7956	5.2589/5.1370
12, 2	5.4940/5.3773	6.3433/5.6149	4.7833/4.6773
12, 3	5.1573/5.0762	5.9998/5.1602	4.6075/4.4980
24, 1	4.3654/4.3605	6.1452/5.8318	4.8464/4.8687
24, 2	3.9014/3.8748	5.4757/5.1513	3.9506/3.9361
24, 3	3.5588/3.5341	5.2229/4.8763	3.6564/3.6555

means are 3.9419/3.9232, 5.6146/5.2865, and 4.1511/4.1534. Pooled FID falls from 4.5692 to 4.4544 (2.513%, 8/9). Both complete matrices pass their frozen quality and artifact-integrity gates.

22.1 Cost and runtime ledger

Table 34: Sequential post-training cost.

Quantity	Shared stage	Refinement	Refinement/base
Optimizer steps	192	48	25.0%
Rollout sample-evaluations	614,400	153,600	25.0%
Trainable field parameters	3,476,547	0	0%
Stored clock scalars	16	48	+32
Mean measured time	1,933.6 s	457.3 s	23.65%

Alternating CUDA-synchronized timing uses the same field, noise, labels, batch 512, NFE12, three warmup pairs, and 20 repetitions in alternating order. Mean unit-median wall overhead is 0.00037% for Heun, 0.01992% for RK3, and -0.01003% for RK4; the pooled value is 0.00342%. The range over individual unit–solver cells is -0.0233% to $+0.0512\%$. These differences establish runtime parity rather than a speedup.

22.2 Corrected high-budget protocol

The initial NFE32 request is incompatible with the three-stage RK3 solver. The implementation rejected all three RK3 cells before sampling. Six partial Heun/RK4 cells are excluded, and the corrected common budget is NFE36, since 36 is divisible by 2, 3, and 4. The fresh NFE36 FID10k means are 6.0197/6.0106 for Heun (0.151%, 2/3), 6.7427/6.7044 for RK3 (0.569%, 3/3), and 6.2197/6.2197 for RK4 (0.0006%, 2/3). Pooled reduction is 0.250% with 7/9 signs, below the frozen 0.5% FID50k-promotion threshold. No dose is retuned and no NFE36 FID50k is run.

23 Budget-conditioned transfer within explicit RK2

The load-time coordinate rule is

$$\beta(c_2, B) = 2(c_2 - 1/2)\,\mathrm{clip}((B-12)/12, 0, 1).$$

It interpolates midpoint and Heun artifacts and uses no target-receiver objective. For Ralston ($c_2 = 2/3$), the paired FID10k results are:

Table 35: Frozen Ralston receiver, zero-fit candidate, and native Ralston.

NFE	Frozen	Candidate	Native	Gain	retention
8	12.0680	9.2789	8.9291	2.7892 (5/5)	88.9%
12	9.0864	6.7005	6.6266	2.3860 (5/5)	97.0%
24	6.9003	5.5688	5.8084	1.3315 (5/5)	121.9%

Two previously unseen coordinates confirm the safe interval at NFE24: $c_2 = 0.625$ gives frozen/midpoint/coordinate FID 7.1356/5.7928/5.5146, and $c_2 = 0.70$ gives 6.6686/5.9969/5.6243.

Coordinate transfer beats both controls in 5/5 pairs for each receiver. The full $[0.5, 1]$ claim is rejected because the high-stage end of the midpoint–Heun interpolation edge is not covered. At $c_2 = 0.875$, a three-anchor local mesh reduces the excess over frozen by 0.0807/0.0581 at NFE24/32, but remains 0.0333/0.0477 above frozen. Residual correction still leaves a fresh $c_2 = 0.85$, NFE24 cell 0.0147 above frozen. Raw sample-space and frozen Inception-feature Richardson discrepancies were also tested as label-free load-time selectors. Neither passed its preregistered development gate: at NFE24 both select $\beta = 0$, whereas the best candidate on the frozen grid is $\beta = 0.25$. We therefore retain $c_2 \leq 0.75$ as the confirmed whitelist and make no label-free safety claim outside it. The development summaries are `results/remote/paper_a_20260905/vp_rk2_defect_selector/development_summary.json` and `results/remote/paper_a_20260905/vp_rk2_feature_defect_selector/development_summary.json`.

24 Frozen-field receivers beyond images

24.1 Matched clock protocol

We test a common conditional linear OT-CFM receiver on real time series, natural-text latents, spoken-event streams, and chaotic dynamics. The conditional field is a width-256 multilayer perceptron with five deterministic time features. Each of three receiver seeds uses 5,000 base updates with batch 1,024. After base training, we freeze every field parameter and train two equal-capacity clock families for 400 updates with batch 512. The shared clock has 48 parameters. The solver-conditioned family has three 16-parameter residual clocks. Both cycle through Heun NFE8/16, RK3 NFE9/18, and RK4 NFE8/16. The clock loss is normalized endpoint MSE plus 10^{-3} clock regularization.

For each solver, development selects the learned clock with lower mean endpoint MSE. The arm, predicted direction, dataset hash, split hashes, and seeds are written to a content-hashed lock before confirmation. Confirmation compares the promoted arm with the identity clock under paired source noise and exact field- call accounting. Table 36 gives every low-budget mean. All 12 domain–solver cells and all 36 seed–solver pairs improve.

Table 36: Low-budget frozen-field receiver results. Values are mean normalized endpoint MSE over three receiver seeds. S and C denote the shared and solver-conditioned clocks.

Receiver	Solver/NFE	Clock	Identity	Selected	Reduction
Electricity	Heun/8	S	1.060820	0.791689	25.370%
	RK3/9	S	0.729409	0.661883	9.258%
	RK4/8	S	0.917158	0.764016	16.698%
WikiText-2/Qwen	Heun/8	S	2.505180	1.858252	25.824%
	RK3/9	S	2.498770	1.996172	20.114%
	RK4/8	S	2.497162	1.884440	24.537%
Lorenz-96 H20	Heun/8	S	0.020826	0.011418	45.173%
	RK3/9	S	0.009095	0.007707	15.258%
	RK4/8	C	0.011579	0.007076	38.889%
SHD event continuation	Heun/8	S	1.237843	1.162686	6.072%
	RK3/9	S	1.256236	1.181787	5.926%
	RK4/8	S	1.254831	1.174998	6.362%

The shared clock is selected in 11/12 cells. A conditioned residual clock is selected only for Lorenz-96 with RK4. This breadth primarily supports frozen-field time reparameterization; solver-specific residual clocks provide an optional residual correction rather than a universal advantage.

24.2 Frozen cross-budget evaluation

Each low-budget selection is reevaluated at two, four, and eight times its original NFE without refitting. Table 37 reports relative endpoint-MSE reductions over the identity clock. Gains decay across every receiver–solver series. At the largest budget, 11/12 means remain nonnegative; Lorenz-96 RK3 is effectively tied at -0.022%.

24.3 Representations, secondary metrics, and audit status

Electricity uses the official GluonTS `electricity_nips` files. Each example maps 24 context values to 24 forecast values. Context and target are divided by the larger of the context mean absolute

Table 37: Relative endpoint-MSE reduction for frozen clocks across NFE. The columns multiply the solver-compatible base budgets NFE8/9/8.

Receiver	Solver	1×	2×	4×	8×
Electricity	Heun	25.370	15.340	6.680	2.307
	RK3	9.258	3.350	0.530	0.036
	RK4	16.698	7.440	2.860	0.748
WikiText-2/Qwen	Heun	25.824	17.740	7.400	2.124
	RK3	20.114	7.790	1.139	0.034
	RK4	24.537	14.340	3.310	0.270
Lorenz-96 H20	Heun	45.173	15.520	4.505	1.183
	RK3	15.258	1.412	0.028	−0.022
	RK4	38.889	6.336	0.297	0.013
SHD	Heun	6.072	4.710	2.230	0.719
	RK3	5.926	3.120	0.960	0.116
	RK4	6.362	3.780	1.100	0.152

value and 1.0; log scale is appended to the condition. The train/calibration/development/confirmation counts are 40,000/6,000/6,000/2,590. The confirmation examples are the final forecast windows from the official test entries.

WikiText-2 maps 16 prefix tokens to four continuation tokens. Frozen Qwen3-0.6B token embeddings are projected to 16 dimensions by a Gaussian matrix with seed 21001071 and divided by the square root of the hidden dimension. The split counts are 20,400/3,600/3,000/3,000, with official validation used for development and official test used for confirmation. At low NFE, endpoint cosine improves for every solver. Gaussian Fréchet distance falls by 35.7/49.7/44.6%, and sliced-W2 falls by 86.0/89.7/89.0% for Heun/RK3/RK4. This receiver measures continuous token-embedding continuation, not perplexity or discrete text generation.

Lorenz-96 uses dimension 10, forcing 8, RK4 simulation step 0.01, a 500-step burn-in, four conditioned states, and a 20-step horizon. The 64/16/16/16 train/calibration/development/confirmation trajectories have disjoint initial conditions. SHD bins the first and second halves of each 1.4-second event stream into 16×16 log-count targets and conditions on the first half, digit, and speaker. Its split counts are 5,701/1,224/1,231/2,264.

The initial SHD preprocessing divided 19 target bins with zero training variance by 10^{-6} and produced invalid MSE near 2.6×10^{7}. The repaired receiver drops 2 unsupported observation bins and 19 unsupported target bins using only training statistics, with support threshold 10^{-4}. Because the official test split had already been exposed by the invalid run, the repaired SHD result is supportive implementation evidence rather than a prospective blind confirmation. Its Gaussian Fréchet and sliced-W2 means improve for all three solvers, while endpoint cosine decreases slightly. Electricity and Lorenz-96 have mixed secondary marginal metrics, so their claim is restricted to the primary forecasting MSE.

25 Independent SB-CFM replication

25.1 Receivers and direct post-training effect

Three SB-CFM ($\sigma = 0.5$) backbones were independently pretrained for 400k updates with seeds 20291204–20291206. Their untouched NFE100 CleanFID50k values are 11.3585, 9.8094, and 10.0683 (mean 10.4121). No SB-specific clock order, dose, loss weight, or iteration count was selected after these receivers were opened.

The first experiment applies the same exact-deployment-map post-training used in the main study. Table 38 reports fresh FID50k streams. The NFE8 and NFE12 gates pass. NFE24 FID10k reverses, retaining the finite-budget boundary.

25.2 Sequential clock refinement and equal-compute isolation

The two-stage experiment uses 192 alternating Stage-I field–clock updates, then freezes the field and performs 48 clock-only updates, 16 per solver. Pooled NFE12 FID50k falls from 9.6244 to 9.1037

Table 38: Independent SB-CFM post-training. Positive reduction is better than the untouched receiver; parentheses are positive receiver counts.

NFE	Frozen	Field only	Reduction	Field–clock	Reduction
8	10.0334	9.5709	4.61% (2/3)	9.6954	3.37% (2/3)
12	7.8825	7.2844	7.59% (3/3)	7.3007	7.38% (3/3)

(5.410%, 8/9). The sole negative cell is Heun on seed 20291206, a 0.007 FID regression; every RK3/RK4 cell improves.

Because this comparison gives only the conditioned arm the final 48 updates, a separate control reloads Stage I and gives the shared clock the same 48 updates, with identical field bytes, batches, noises, solver exposures, and optimizer budget. Table 39 reports two independently seeded distributional readouts. The original projected-feature promotion gate failed and was not relaxed; the displayed CleanFID streams were authorized by a separately frozen held-out protocol after the independent DINOv2 readout.

Table 39: SB-CFM equal-compute frozen-field comparison at NFE12. Positive CleanFID gain is continued shared minus solver-conditioned.

Solver	Shared FID50k	Conditioned FID50k	Gain	Wins
Heun	7.6664	7.6775	-0.0111 (-0.145%)	1/3
RK3	8.8081	8.4157	$+0.3924$ ($+4.455\%$)	3/3
RK4	10.8820	11.0429	-0.1609 (-1.479%)	1/3
Pooled	9.1188	9.0454	$+0.0735$ ($+0.805\%$)	5/9

On 10,000-sample DINOv2 features, conditioning improves FD in 8/9 cells and KID in 9/9; receiver-averaged FD/KID gains are positive for all three backbones. Precision is essentially tied and recall falls by 0.00277. Thus the equal- compute evidence isolates a robust RK3 benefit and a small pooled distributional effect, not unconditional Pareto dominance for all solvers.

26 Specialization regret and solver-clock geometry

Let $z_K = \Psi_S(\phi; z_0, \theta)$ be the terminal map of explicit Runge–Kutta solver $S = (A, b, c)$ under a monotone clock ϕ. Differentiating the stage recursion gives a clock sensitivity $J_S = \partial\Psi_S/\partial\phi$ that depends on the tableau: clock changes alter step lengths and internal stage times, and solver-specific stage Jacobians propagate both changes. For terminal loss ℓ,

$$g_S = J_S^\top \nabla_z \ell(z_K) + g_{\mathrm{anchor},S}, \qquad H_S \simeq J_S^\top H_z J_S + H_{\mathrm{anchor},S} + \lambda R.$$

Both the local gradient and curvature depend on solver and budget.

Around a shared clock, let $Q_S(\delta) = q_S + g_S^\top \delta + \frac{1}{2}\delta^\top H_S \delta$ with $H_S \succ 0$, individual optimum $\delta_S^* = -H_S^{-1} g_S$, and weighted shared optimum

$$\delta_{\mathrm{sh}}^* = -\left(\sum_S w_S H_S\right)^{-1} \sum_S w_S g_S.$$

Completing the square proves

$$\sum_S w_S\{Q_S(\delta_{\mathrm{sh}}^*) - Q_S(\delta_S^*)\} = \frac{1}{2}\sum_S w_S(\delta_{\mathrm{sh}}^* - \delta_S^*)^\top H_S(\delta_{\mathrm{sh}}^* - \delta_S^*) \geq 0.$$

Equality holds exactly when all supported solvers share the same local optimum. This is a local training-objective identity. It does not imply that a separate residual clock necessarily improves held-out FID.

The read-only SB-CFM diagnostic verifies that the geometry is not degenerate. At identical Stage-I field, clock, noise, and target, the mean gradient cosines are -0.357 (Heun/RK3), 0.189 (Heun/RK4),

and 0.003 (RK3/RK4). Learned residual-clock displacement aligns with the negative specialization residual in all 9/9 receiver–solver cells. Raw gradient magnitude nevertheless ranks Heun first, whereas only RK3 has a stable equal-compute FID50k gain. Gradient magnitude is not a reliable deployment predictor.

Against a common high-NFE endpoint, the NFE12 relative RMS becomes 1.53% worse for Heun but 2.92% and 1.62% better for RK3 and RK4. RK4 provides a counterexample: numerical closeness improves while FID worsens. A BF16 and FP32 reference ladder also fails the preregistered asymptotic-contraction gates, so the paper makes no continuous-limit recovery claim. Twelve analytic-order, exact-NFE, and interface tests pass. The observed behavior belongs to the discrete map and is not explained by an RK implementation error.

27 Distributional localization and prospective target alignment

27.1 Matched retrospective decompositions

For generated Inception mean μ_{sh}, conditioned mean μ_{cond}, and real mean μ_*, the diagnostic is

$$A_S = \cos(\mu_{\mathrm{cond}} - \mu_{\mathrm{sh}}, \mu_* - \mu_{\mathrm{sh}}).$$

Positive gains in Table 40 mean that the solver-conditioned clock improves the equal-compute/equal-capacity shared control. Every source FID is exactly recovered by its float64 decomposition.

Table 40: Retrospective FID decomposition. Mean and covariance columns sum to the total gain. DiT values are numerically stable supportive evidence; see text.

Family	Solver	Total	Mean	Covariance	Alignment	Wins
SB-CFM	Heun	-0.0111	$+0.0163$	-0.0274	$+0.078$	1/3
	RK3	$+0.3924$	$+0.1442$	$+0.2481$	$+0.421$	3/3
	RK4	-0.1609	-0.0824	-0.0785	-0.279	1/3
C2OT	Heun	-0.5309	-0.1956	-0.3353	-0.127	0/3
	RK3	$+5.0745$	$+2.7647$	$+2.3098$	$+0.866$	3/3
	RK4	-0.6741	-0.2744	-0.3997	-0.289	0/3
DiT	Heun	$+0.3475$	$+0.1185$	$+0.2290$	$+0.168$	2/3
	RK3	$+12.1669$	$+6.6081$	$+5.5589$	$+0.834$	3/3
	RK4	$+3.2185$	$+1.0217$	$+2.1967$	$+0.607$	3/3

For SB-CFM, RK3 improves both components in 3/3 receivers; RK4 worsens both pooled components. Equal-capacity C2OT gives complete solver-wise separation: RK3 improves both components in 3/3, while Heun/RK4 worsen both in 3/3. In corrected DiT, conditioning improves total FID in 8/9 cells and both component signs match total FID in 9/9. The sole alignment mismatch has cosine 0.0033 and is direction-neutral. Across the 18 matched SB-CFM/C2OT cells, alignment and total-FID gain have descriptive Pearson/Spearman correlations 0.787/0.851. Because these FIDs were already exposed, this association is explanatory only.

The DiT historical-reproduction gate was fixed at 10^{-8} and failed: the largest arm offset was 6.68×10^{-5}. A separately frozen replay is bitwise stable in the current evaluator, and the largest treatment-gain drift from the historical files is 1.153×10^{-4}, only 0.0441% of the smallest effect. No sign is at risk, but the original integrity failure remains recorded; hence this retrospective DiT table is supportive rather than an exact historical-reproduction pass.

27.2 Fresh prospective image receivers

The alignment rule and threshold $|A_S| \geq 0.05$ were frozen before training fresh C2OT seed 20990954 and corrected-DiT seed 20990955. Both comparisons use 48 clock parameters per arm, matched late-field capacity, matched solver/NFE exposure, and disjoint 2,048-sample development and 10,000-sample confirmation streams. Predictions and development hashes were written before confirmation.

C2OT passes its primary and strong gates: 3/3 total signs, 6/6 component signs, and all testable alignment signs replicate, including the negative RK4 result. DiT also gives 3/3 total and 6/6 component sign replication. Its preregistered primary gate also required at least one positive and one negative prediction. The observed $(+,+,+)$ pattern makes the primary flag false even though every directional prediction is correct. We use DiT only as supportive cross-architecture replication.

Table 41: Prospective NFE12 confirmation. Gain is shared minus conditioned CleanFID10k. "Components" counts development mean/covariance signs retained on confirmation.

Family	Solver	Dev. gain	Confirm. gain	Dev. alignment	Components
C2OT	Heun	$+0.2322$	$+0.1480$	$+0.294$	2/2
	RK3	$+4.7259$	$+5.0727$	$+0.889$	2/2
	RK4	-0.5657	-0.5724	-0.169	2/2
DiT	Heun	positive	$+1.0630$	positive, testable	2/2
	RK3	positive	$+11.8943$	positive, testable	2/2
	RK4	positive	$+3.2983$	positive, testable	2/2

27.3 Offline conditional-action receivers

The supplied archive `<dataset-root>/FlowMatching_datasets_2026-09-09.zip` has SHA-256 `0b1129ecef946c5855dfcd0b6048c7bd30494cbde5e07339a1995bb4c2b4b29c`. For each of its four assets, three fresh width-256 conditional-action flows are trained for 8,000 updates. The field is then frozen. One order-48 shared clock and three order-16 residual clocks receive 600 matched updates and are tested with Heun, RK3, and RK4 at NFE12. The action-space alignment uses all held-out sample/action coordinates and predicts the sign of normalized-action MSE.

Table 42: Offline-action confirmation. Positive C–S gain means conditioning beats the shared clock; C–I is conditioned post-training versus identity.

Asset	Mean C–S gain	C wins	Mean C–I gain	Direction matches
Cheetah demos	-0.009325	0/9	$+0.003701$	9/9
Cheetah scripted	$+0.002309$	9/9	$+0.003385$	9/9
Walker scripted	-0.047655	0/9	$+0.028325$	9/9
Nav2D	-0.007627	0/9	$+0.003112$	9/9

All 36 development alignments are testable and all 36 predict the disjoint confirmation direction. The preregistered mechanism-supporting gate nevertheless fails because conditioning improves pooled MSE on only one of four assets. This is precisely the deployment distinction. Clock post-training is favorable over identity on average in every asset, but solver-specific residual clocks should replace the equal-capacity shared clock only when alignment predicts a benefit. Because the archive lacks episode IDs, rewards, next states, and environment versions, the experiment makes no online-control, return, success-rate, or trajectory-held-out claim. The two Cheetah assets also have partly incomplete provenance.

28 Latest artifact map

The additional canonical sources are:

- `results/remote/paper_a_20260905/solver_matching_fid/canonical_summary.json`;
- `results/remote/paper_a_20260906/dit_solver_matching/fid_summary.json`;
- `results/remote/paper_a_20260906/c2ot_solver_matching_fid/canonical_summary.json`;
- `results/remote/paper_a_c2ot_sequential_solver_clock_fid50k_20260907/summary.json`;
- `results/remote/paper_a_c2ot_sequential_solver_clock_nfe24_fid50k_20260907/summary.json`;
- `results/remote/paper_a_c2ot_sequential_solver_clock_runtime_20260907/summary.json`;

- results/remote/paper_a_c2ot_sequential_solver_clock_high_nfe_fid10k_corrected_20260907/summary.json;
- results/remote/paper_a_20260905/rk2_budget_conditioned_fid10k_summary.json;
- results/remote/paper_a_20260905/safe_rk2_interval_fid10k_summary.json;
- results/remote/paper_a/sb_cfm_posttraining_20260909/summary.json;
- results/remote/paper_a_sb_cfm_sequential_solver_clock_fid50k_20260909/summary.json;
- results/remote/paper_a_sb_cfm_equal_compute_clock_control_heldout_fid50k_20260909/summary.json;
- results/remote/paper_a_sb_cfm_solver_clock_gradient_geometry_20260909/summary.json;
- results/remote/paper_a_sb_cfm_fid_component_mechanism_20260909/summary.json;
- results/remote/paper_a_c2ot_equal_capacity_fid_components_20260909/summary.json;
- results/remote/paper_a_dit_equal_capacity_fid_components_20260909/summary.json;
- results/remote/paper_a/c2ot_prospective_alignment_20260909/summary.json;
- <project-root>/results/paper_a_dit_prospective_alignment_20260909/summary.json;
- results/remote/paper_a/offline_action_receivers_20260909/summary.json; and
- results/remote/paper_a/multidomain_latent_receivers_20260910/consolidated_summary.json.

Protocol documents beside these files record frozen seeds, gates, and launch commands. The cross-domain protocol and audit report are docs/paper_a_multidomain_protocol_20260910.md and docs/paper_a_multidomain_experiment_report_20260910.md. The consolidated solver-clock mechanism audit is docs/paper_a_solver_clock_mechanism_diagnostics_evidence_20260909.md; the frozen prospective protocols are docs/paper_a_c2ot_prospective_alignment_protocol_20260909.md, docs/paper_a_dit_prospective_alignment_protocol_20260909.md, and docs/paper_a_offline_action_receiver_protocol_20260909.md.

29 Full base-effect component isolation table

The main text now summarizes this result alongside the claim-driven figures. For completeness, Table 43 retains the full original component-isolation comparison and paired intervals.

Table 43: Five-artifact TorchCFM component isolation. Values are mean FID10k; Δ is the paired improvement over terminal-trained linear, with a 95% artifact-level interval.

Artifact	NFE8	NFE12	NFE24	Paired improvement $\Delta_{8/12/24}$
Trained linear	28.338	12.965	8.325	–
Trained field–clock	**23.241**	**11.558**	**7.613**	5.096 / 1.407 / 0.712
Trained full-path–field	23.244	11.605	7.793	5.093 / 1.360 / 0.533
Trained grid–field	28.568	12.870	7.979	−0.230 / 0.095 / 0.347

Field–clock intervals are [3.423,6.770], [0.960,1.854], and [0.186,1.238]. Full-path intervals are reported in the preceding component ablations.

30 SAME-CHECKPOINT MOTIVATION AND SOLVER-SPECIFIC RESIDUAL CAPACITY

Stage I of the C2OT experiment adapts one shared monotone order-16 clock and a late field subset for exact NFE12 deployment. Stage II freezes that receiver and composes the shared clock with a monotone Bernstein residual, $\tau_{S,K}(t) = \tau_{\mathrm{shared}}(\rho_{S,K}(t))$. The residual has $3K$ trainable scalars and is initialized as the exact identity, so $K = 0$ is the shared-clock action rather than a separately fitted control. Every uniform-K development arm uses 48 updates, exactly 16 per solver, and the same paired 2,048-sample stream.

Table 44: Equal-start residual-clock capacity sweep at exact NFE12. Positive gain means lower projected Fréchet risk than the shared clock.

K	Scalars	Heun	RK3	RK4	Pooled rel.	Sign
4	12	-0.007407	+0.005519	+0.003895	+0.142%	$(-,+,+)$
8	24	-0.014710	+0.016540	+0.007324	+0.647%	$(-,+,+)$
16	48	-0.013986	+0.023931	+0.001269	+0.792%	$(-,+,+)$
32	96	-0.016543	+0.045017	-0.001674	+1.894%	$(-,+,-)$

No single positive capacity is safe for all solvers: Heun degrades at every K, RK3 improves monotonically through $K = 32$, and RK4 peaks at $K = 8$ before reversing at $K = 32$. The action vector $(K_{\mathrm{Heun}}, K_{\mathrm{RK3}}, K_{\mathrm{RK4}}) = (0, 32, 8)$ was frozen before opening a fresh confirmation stream. Table 45 reports both the new proxy stream and the separately seeded CleanFID10k evaluation.

Table 45: Frozen solver-specific action: independent NFE12 proxy and CleanFID10k. Heun needs no generated comparison because its selected action is exactly the shared clock.

Solver	Action	Proxy rel.	Shared FID	Selected FID	FID rel.
Heun	$K = 0$	0.000%	–	–	–
RK3	$K = 32$	+5.534%	8.3770	7.4176	+11.453%
RK4	$K = 8$	+1.162%	7.3095	7.1669	+1.951%
Pooled/mean	$(0, 32, 8)$	+2.266%	7.8433	7.2922	+7.025%

The frozen proxy checks all pass: on seed 20993951, RK3 risk changes from 0.545918 to 0.515709, RK4 from 0.542570 to 0.536268, Heun is identical at 0.522695 by construction, and pooled risk changes from 0.537061 to 0.524891. The CleanFID cells also pass the field-tensor, field-hash, initialization, source-artifact, parameter-count, update-ledger, and paired-generation checks.

Table 46: Optimization-seed robustness of the two active residual clocks. Each clock is retrained from the same Stage-I receiver and evaluated on the fixed independent 2,048-sample NFE12 stream.

Solver	K	Positive	Mean gain	Relative	SD
RK3	32	10/10	0.03964	7.656%	0.00252
RK4	8	10/10	0.00642	1.317%	0.00256

The two-sided Wilson 95% interval for each 10/10 sign rate is $[0.7225, 1.0000]$; gain ranges are 0.03628–0.04413 for RK3 and 0.00326–0.01185 for RK4. This excludes a single favorable refinement initialization as the explanation, while the independent CleanFID test remains the distributional confirmation.

Table 47: Same-checkpoint NFE12 CleanFID10k ladder. Comparisons are paired within each solver; solver means are descriptive because generation seeds differ across solvers.

Solver	Untouched	Shared	Selected	Stage-I rel.	Stage-II rel.
Heun	62.8422	7.4206	7.4206	88.19%	0.00%
RK3	35.5854	8.5360	7.5653	76.01%	11.37%
RK4	42.0449	7.1564	7.0156	82.98%	1.97%
Mean	46.8242	7.7043	7.3339	83.54%	4.81%

Across the full ladder, mean CleanFID falls by 39.4903, or 84.34%. Stage I measures the continuous-training versus fixed-discretization gap on one released receiver; Stage II shows that the remaining gap has solver-dependent effective dimension. Stage I changes 3,476,547 late-field parameters (about 3.9% of the 89.1M receiver) plus a 16-scalar clock over 192 updates. The deployed Stage II action adds 120 active residual parameters: 96 for RK3 ($K = 32$) and 24 for RK4 ($K = 8$). Residual refinements start bitwise equal to the shared clock, preserve the Stage-I field exactly, and use disjoint development, confirmation, image, and robustness streams. The scope remains limited to one frozen C2OT receiver, one CleanFID10k stream per active solver, and no selected-action FID50k or second-backbone confirmation.

31 Same-checkpoint SiT temporal isolation and complete capacity

31.1 Matched temporal families on an official SiT receiver

We isolate the temporal parameterization on the official SiT-XL/2 ImageNet-256 checkpoint (SHA256 `e645b5cd82d1d645cfd9b7c96041b56b2a0aa61faed90c0869e79be02ab55aca`). The entire field is frozen. The learned-grid, flexible-path, and continuous families each expose eight trainable scalars, run for 192 updates, and receive 96 Heun and 96 RK4 updates. All FID cells use exact NFE12, CFG 1.5, batch 128, the same role-reset random stream, and the ADM TensorFlow Inception pool-3 evaluator with the official ImageNet-256 reference. Development freezes one grid and one continuous action per solver before either FID stream is opened: Heun selects grid-conditioned and continuous-shared; RK4 selects grid-shared and continuous-conditioned. The full matrix below is retained so exploratory arm winners cannot replace those actions post hoc.

Table 48: Complete same-checkpoint SiT temporal-parameterization matrix at exact NFE12. FID10k averages seeds 21108001–21108002; FID50k uses seed 21109001. Lower is better.

Samples	Solver	Uniform	Grid-sh.	Grid-cond.	Path	Cont.-sh.	Cont.-cond.
10k	Heun	11.9452	11.9495	11.9099	11.9250	11.8622	11.8101
10k	RK4	10.0430	10.0094	10.0234	10.1521	10.0208	10.0027
50k	Heun	9.2562	9.2588	9.2184	9.2444	9.1611	9.1216
50k	RK4	7.2101	7.1792	7.1856	7.3068	7.1949	7.1693

Using only the frozen primary actions, pooled FID50k is 8.2331 for uniform time, 8.1988 for the learned grid, 8.2756 for the flexible path, and 8.1652 for continuous reparameterization. The continuous clock improves over the selected grid for both solvers: 9.1611 versus 9.2184 for Heun and 7.1693 versus 7.1792 for RK4. The exploratory conditioned Heun clock reaches 9.1216 but is not substituted for the preregistered continuous-shared result.

Five-artifact FID50k replication. We next repeated only the frozen continuous–grid contrast with independent adaptation seeds 21112001–21112005 and paired generation seeds 21113001–21113005. All 20 solver–family cells passed the artifact and metric integrity audit. Table 49 reports continuous minus grid, so negative values favor continuous.

Table 49: Independent SiT ADM FID50k replication at exact NFE12. The pooled difference averages Heun and RK4 within each adaptation seed before inference.

Adaptation seed	Heun grid/cont.	RK4 grid/cont.	Pooled Δ
21112001	9.1529/8.8302	7.2096/7.1757	-0.1784
21112002	9.2137/9.2861	7.2270/7.2310	$+0.0383$
21112003	9.1910/8.7076	7.1539/7.2833	-0.1770
21112004	9.1698/9.1310	7.2831/7.2272	-0.0473
21112005	9.2118/8.8755	7.2666/7.1775	-0.2127
Mean	9.1878/8.9661	7.2280/7.2189	-0.1154
95% CI for Δ	$[-0.5075, 0.0640]$	$[-0.1141, 0.0959]$	$[-0.2478, 0.0170]$
Directions	4/5	3/5	4/5

The mean direction agrees with the original isolation, but neither the pooled nor solver-specific interval excludes zero. Consequently the replication supports a favorable trend and rejects a claim of seed-stable strict continuous-clock dominance over learned grids.

31.2 Finite-stage robustness, budget transfer, and STORK stacking

The finite-stage operator was tested separately from the continuous–grid contrast. Table 50 reports three paired ADM-FID50k streams at exact NFE12. Every stream favors the finite-stage operator for both solvers and both controls. Against grid, mean reductions are 0.6230 FID for Heun and 0.1483 for RK4. Against continuous, the reductions are 0.5583 and 0.1382. All four paired 95% intervals exclude zero. The machine-readable summary is `results/paper_a_sit_stage_aware_fid_replication_20260920/summary.json`.

Table 50: Three-stream SiT finite-stage confirmation at exact NFE12. Each row uses paired noise and labels; lower ADM-FID50k is better.

Stream	Solver	Grid	Continuous	Stage
21120001	Heun	9.2231	9.1556	**8.5700**
21122001	Heun	9.2704	9.2206	**8.6583**
21122002	Heun	9.1696	9.0929	**8.5658**
21120001	RK4	7.1752	7.1662	**7.0273**
21122001	RK4	7.2301	7.2172	**7.0760**
21122002	RK4	7.1748	7.1663	**7.0317**

Five independently post-trained artifacts provide a separate seed-stability audit on projected risk. The finite-stage operator beats grid in all five Heun and all five RK4 cells. It beats continuous in 4/5 Heun and 5/5 RK4 cells. Mean reductions are 0.0671 [0.0225,0.1117] and 0.0255 [0.0014,0.0497] against grid. They are 0.0541 [0.0095,0.0987] and 0.0269 [0.0028,0.0511] against continuous. The machine-readable summary is `results/paper_a_sit_stage_seed_replication_20260920/summary.json`.

Independent scale-only fit. The complete operator learns both query times and positive quadrature density. We therefore fitted the density alone with the same eight logits, 192 updates, and five preregistered seeds. Table 51 compares each scale-only artifact with its same-seed complete operator on one fresh paired stream. The complete operator has no detected advantage: the mean scale-only-minus-complete differences are -0.0010 for Heun and -0.0090 for RK4, and both intervals include zero. The density-only fit therefore occupies the same operating range under this proxy design, supporting it as a parsimonious component for future deployment studies. The machine-readable summary is `results/paper_a_sit_independent_components_20260921/scale_only_all_summary.json`.

Table 51: Independently fitted scale-only density versus the matched complete finite-stage operator. Negative differences favor scale-only; lower projected risk is better.

Adaptation seed	Heun difference	RK4 difference
21119001	$+0.0129$	-0.0033
21119002	-0.0920	-0.0435
21119003	$+0.0707$	$+0.0397$
21119004	$+0.0019$	-0.0112
21119005	$+0.0014$	-0.0265
Mean [95% CI]	-0.0010 [$-.0735, .0715$]	-0.0090 [$-.0478, .0299$]
Complete wins	4/5	1/5

The selected seed-21119001 stage density was next frozen and resampled without target-budget fitting. Table 52 reports the full four-cell matrix on fresh stream 21126001. All four target cells favor the resampled finite-stage operator, so the preregistered proxy gate authorizes a future FID promotion without changing the operator or the target stream.

Table 52: Zero-fit finite-stage transfer across budgets. Values are projected Fréchet risks on 4,096 paired samples; lower is better.

Budget	Solver	Uniform	Grid	Continuous	Stage
NFE8	Heun	2.4723	2.4359	2.3812	**2.0105**
NFE8	RK4	1.3939	1.3769	1.3570	**1.3258**
NFE24	Heun	0.5823	0.5836	0.5808	**0.5643**
NFE24	RK4	0.5258	0.5220	0.5217	**0.5164**

For a direct external positioning check, the official STORK implementation was run unchanged at exact NFE12 with and without the frozen shared clock. The two new streams and the original stream all favor stacking (Table 53); the mean paired change is -0.0344 FID. The complete metadata, proxy audit, and FID cells are retained under `results/paper_a_sit_stork_replication_20260921/`.

Table 53: Three-stream STORK stacking audit at exact NFE12. Each pair uses the same noise and labels; lower ADM-FID50k is better.

Stream	STORK	STORK+clock	Δ
21120001	6.7231	**6.6900**	-0.0331
21125001	6.7434	**6.7041**	-0.0393
21125002	6.6773	**6.6466**	-0.0307

The unseen-tableau stage audit uses a separate frozen stream (21127001) and does not alter the earlier RK2 coordinate-interpolation claim. At exact NFE12, the midpoint stage transfer gives 0.6144 versus 0.6268 for uniform and 0.6226 for the shared continuous clock. Ralston is retained as an unpromoted control in the complete result file; the main claim therefore remains limited to solver-checked transfer rather than a universal tableau guarantee.

Unseen RK3. All temporal objects above were trained with Heun/RK4 exposure only. A fresh two-seed RK3 FID10k holdout gives 10.2121 for uniform time, 10.1705 for the shared grid (+0.407%), 10.4280 for the flexible path (-2.115%), and 10.2133 for the shared continuous clock (-0.011%). The proxy ranks the continuous clock first, but FID retains only the grid gain. This is a direct boundary on zero-fit transfer and motivates action selection for an unseen solver.

31.3 Defect redistribution rather than scalar defect minimization

From each deployed coarse state, we compare one full step with two half steps and propagate both alternatives to the endpoint. Table 54 reports signed reductions relative to uniform time; negative values denote an increase. No one scalar diagnostic improves in every beneficial FID cell.

Table 54: SiT defect-profile changes relative to uniform time. Terminal impact is the finite endpoint-impact surrogate used in the frozen diagnostic.

Solver	Clock	Local defect	Endpoint RMS	Terminal impact
Heun	shared	+0.86%	−0.09%	+1.86%
Heun	conditioned	−0.61%	−0.54%	+16.75%
RK4	shared	+3.33%	+0.42%	−29.54%
RK4	conditioned	+2.64%	+1.12%	−79.61%

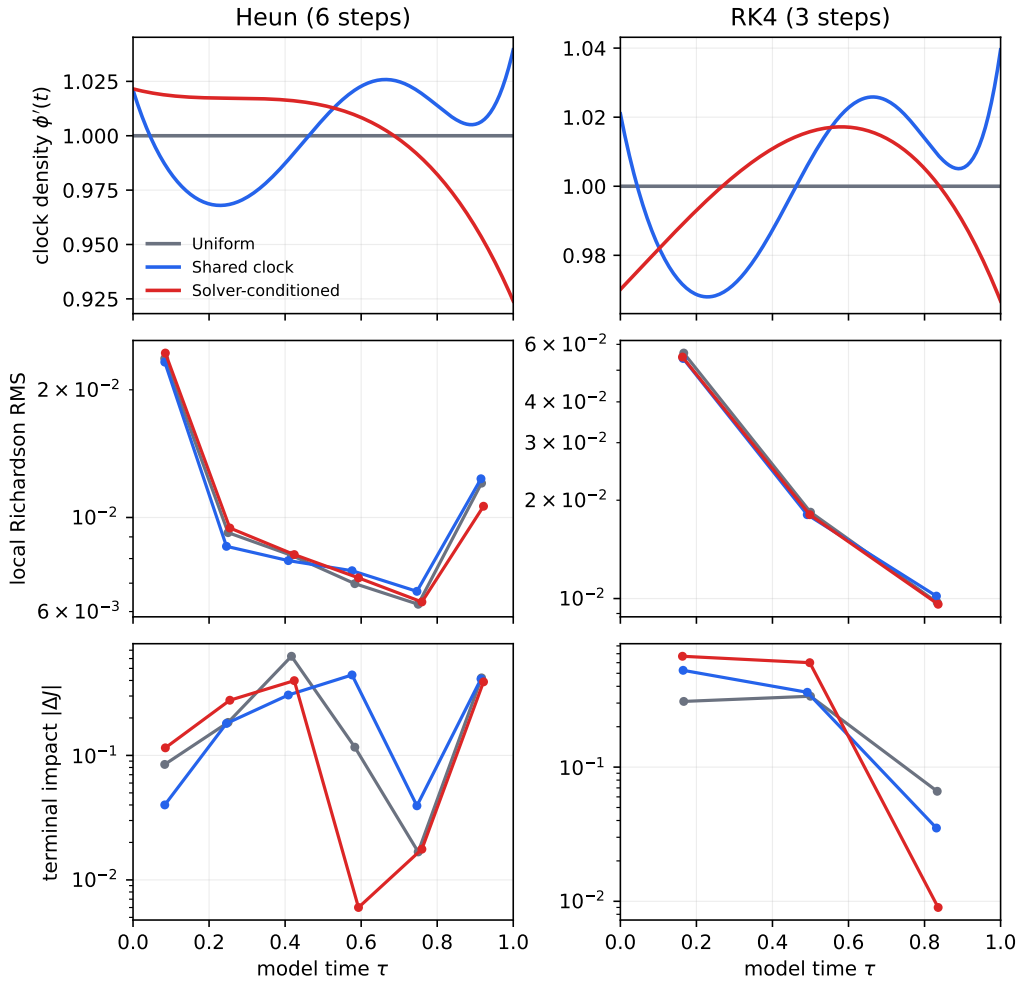

Figure 2: Clock density, local Richardson defect, propagated endpoint defect, and terminal-impact profiles for the same-checkpoint SiT isolation. The learned clock changes where error enters the terminal map; it does not uniformly reduce every diagnostic.

31.4 RK-TABLEAU INTERACTION WITH CONTINUOUS REPARAMETERIZATION

Let $f_\phi(t,x) = q(t)v(\phi(t),x)$ and $q = \phi'$. Augmenting the state with $t' = 1$ gives the standard non-autonomous B-series expansion

$$\Psi_{S,h} - \Phi_h = h^{p+1} \sum_{|\mathcal{T}|=p+1} \big(a_S(\mathcal{T}) - a_\star(\mathcal{T})\big) F_\phi(\mathcal{T}) + O(h^{p+2}).$$

Here $f_t = q'v + q^2 v_\tau$, $f_x f = q^2 D_x v\, v$, and $f_{tt} = q''v + 3qq'v_\tau + q^3 v_{\tau\tau}$. The elementary differentials thus mix ϕ'', ϕ''', v_τ, and $D_x v\, v$, while their coefficients are polynomials in the RK tableau. For an order-two method, defining $m_2 = b^T c^{\odot 2}$ and $m_A = b^T Ac$, the method-minus-exact order-h^3 term is

$$h^3\{(\tfrac{1}{2}m_2 - \tfrac{1}{6})f_{tt} + (m_2 - \tfrac{1}{3})f_{tx}f + (\tfrac{1}{2}m_2 - \tfrac{1}{6})f_{xx}(f,f) + (m_A - \tfrac{1}{6})f_x f_t + (m_A - \tfrac{1}{6})f_x^2 f\} + O(h^4).$$

Heun, explicit midpoint, and Ralston RK2 have different m_2 despite identical formal order. A symbolic scalar witness likewise recovers distinct leading polynomials at h^3, h^4, and h^5 for Heun, RK3, and RK4. This does not prove a globally optimal clock; it identifies the tableau-dependent terms that make a universal specialized clock non-generic.

For a direct numerical check, set $v(\tau,x) = (1 + 0.7\tau)x$, $\phi(t) = t + 0.2t^2 + 0.1t^3$, and $x(0) = 1$. The exact solution is $x(h) = \exp(\phi(h) + 0.35\phi(h)^2)$. High-precision one-step errors over $h = 2^{-3}, \ldots, 2^{-12}$ recover both the expected powers and analytic leading coefficients (Table 55).

Table 55: Measured companion to the symbolic B-series witness. The observed coefficient is $(x_{\mathrm{RK}} - x_{\mathrm{exact}})/h^{p+1}$ at $h = 2^{-12}$.

Solver	Expected power	Measured slope	Analytic coeff.	Observed coeff.
Heun	3	3.0046	-0.0466667	-0.0467311
RK3	4	3.9950	$+0.0087500$	$+0.0087369$
RK4	5	5.0034	-0.0077500	-0.0077580

All coefficient errors are below 0.16%. The complete precision-100 trace and implementation hash are stored in `results/paper_a_bseries_numeric_witness_20260918/summary.json`.

31.5 SIX-POINT RESIDUAL-CAPACITY CURVE

The earlier C2OT sweep selected $(0, 32, 8)$ on that receiver. A separate SB-CFM receiver supplies the complete $K \in \{0, 4, 8, 16, 32, 64\}$ FID10k curve at NFE12 (Table 56). It reveals a different but compatible hierarchy: Heun is nearly saturated, RK3 peaks at intermediate capacity, and RK4 continues improving through $K = 64$.

Table 56: Complete SB-CFM residual-capacity CleanFID10k curve. $K = 0$ is the shared-clock action. Lower is better.

K	Heun	RK3	RK4
0	10.1637	11.0559	14.1587
4	10.1547	10.8701	13.2349
8	10.1572	10.9820	12.7558
16	10.1526	10.8361	13.3326
32	10.1278	**10.6167**	12.0206
64	**10.1265**	10.6343	**11.1802**

The best relative gains over $K = 0$ are 0.366%, 3.973%, and 21.036% for Heun, RK3, and RK4. Proxy and FID select the same capacity for RK3/RK4 but not for Heun, whose absolute difference between $K = 32$ and $K = 64$ is only 0.00135 FID. The result supports solver-dependent capacity while also showing that its numerical value is receiver dependent.

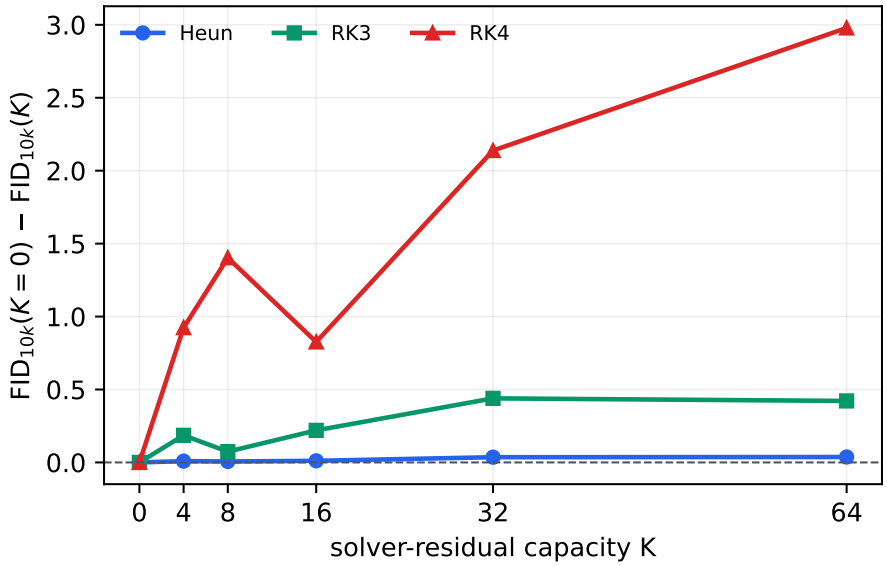

Figure 3: Complete residual-capacity curve for the independently trained SB-CFM receiver. Curves report FID10k change from the shared $K = 0$ action.

32 Scope controls and selection rationale

The complete control matrix explains the final method design. Time-only and full cubic paths occupy the same operating point, making the smaller time-only controller the preferred shared parameterization. Learned grids remain a matched temporal baseline, while the SiT isolation places the continuous clock at the best mean FID50k point. Cross-solver transfer and joint training show why the deployed map is conditioned on the intended tableau rather than represented by one universal artifact. On C2OT, the shared Heun clock is already saturated across the tested residual orders, so the selected action is exactly $K = 0$; RK3 and RK4 receive the capacities supported by their held-out alignment. At high NFE the residual action similarly returns to zero capacity as finite-step excess contracts. These controls motivate the paper's compact policy: one shared field, a zero-cost fallback, and solver-specific capacity only where the frozen deployment evidence supports it.

References

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, 2018.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=1vmSEVL19f.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0165. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/0bc795afae289ed465a65a3b4b1f4eb7-Abstract.html.

Juntang Zhuang, Nicha Dvornek, Xiaoxiao Li, Sekhar Tatikonda, Xenophon Papademetris, and James Duncan. Adaptive checkpoint adjoint method for gradient estimation in neural ODE. In *Proceedings*

of the 37th International Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, pp. 11639–11649, 2020. URL https://proceedings.mlr.press/v119/zhuang20a.html.