TRAINING-STATE SHAPING FOR LOW-NFE FLOW MATCHING: GAINS ACROSS SOLVERS AND BETTER COUPLINGS FOR DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Flow matching trains on prescribed interpolation paths, while generation follows trajectories of the learned velocity field. We develop training-state shaping (TSS), a post-training approach that improves few-step sampling and the data efficiency of teacher-generated pairs. In pixel-space models that obtain velocity from clean-image prediction, we reshape training inputs using an exponential-moving-average model while retaining clean targets. We extend TSS to native velocity prediction by pairing each shaped path with its exact tangent. On ImageNet, TSS improves low-NFE generation across JiT model scales and solvers. With JiT-H/16 and DPM++2M, it achieves FIDs of 6.65 and 4.24 at 10 and 20 NFE, respectively. In latent space, deterministic reparameterization of training time explains most of the gain. With fixed pools of 200k pairs per source and matched student recipes, shaped-teacher data lets reflow and flow-map students reach target FID with fewer updates, equivalently fewer sample presentations at fixed batch size. We measure this advantage across multiple target FIDs and two teacher pairs. The original flow-map extensions reach similar one-call EMA quality at 100k updates. A separate pool-prefix study finds better student quality from fewer TSS-generated pairs than from a larger vanilla pool. An endpoint-FID-matched control shows that endpoint FID alone does not rank downstream data value.

1 INTRODUCTION

Flow matching generates images by integrating a learned velocity field (Lipman et al., 2023; Albergo et al., 2025). A network can predict velocity directly or parameterize the field through a clean-data estimate. JiT (Li & He, 2025) uses the latter approach in pixel space: its network predicts a clean image from a noisy state, and the sampler converts this prediction into a velocity. We develop our method primarily in this clean-image, or x-prediction, setting.

With either parameterization, training samples states from prescribed conditional interpolation paths, while generation follows trajectories of the learned velocity field. We investigate these training states as a design choice for few-step generation. Training-state shaping (TSS) modifies the inputs during post-training. For JiT, the clean-target variant TSS-X shifts them toward a detached prediction from an exponential moving average (EMA) of the model weights while retaining the clean-image target x. The sampler and inference cost remain unchanged.

We then extend shaping to native velocity prediction by modifying the supervision together with the input. A velocity recovered from the retained clean-image target need not equal the tangent of the shaped path. Our transport-consistent construction pairs each shaped path with its exact tangent and converts the pair to the model's native prediction head. The full construction includes an EMA-guided normal component. Its EMA-free time-reparameterization special case is TSS-T, where T stands for time. In our latent velocity-prediction experiments, TSS-T explains most of the low-call gain.

We evaluate training-state shaping both as a sampling intervention and as a way to train teachers that supply synthetic data. Three matched JiT training seeds improve low-call generation under Heun and DPM++2M, with small regressions at higher budgets. Downstream, we measure data efficiency in two ways: updates to a reference FID on fixed pools, and student quality when trained on smaller pools. TSS-generated pairs let reflow and flow-map students reach the reference earlier

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.

under matched student recipes. Each consumer initially has two student streams sharing teachers and pools; an additional flow-map comparison uses a second teacher training seed and newly generated pools. A vanilla pool with slightly better endpoint FID still trains worse reflow students, so that scalar alone does not predict the value of these pairs.

The contributions are a distinction between clean-target state re-indexing and exact-path training, matched tests of low-call gains and their components, and a controlled data-efficiency analysis. With 200k pairs per source and the same student recipe, TSS-generated pairs reach target FID with fewer updates and sample presentations. At the endpoint-derived one-call references, interpolated vanilla-to-TSS update ratios are 2.23 to 2.38 for the original flow-map streams and 1.19 to 1.23 for reflow. Separate pool-prefix tests find that 50k TSS-X pairs outperform the 200k vanilla reference in Table 4. The original flow-map one-call EMA endpoints differ by less than 0.4 FID at the measured 100k endpoint. TSS supplies training data to these one-step students.

2 Background and related work

From training paths to sampling trajectories. Conditional flow matching constructs an interpolant $z_t = I_t(x, \epsilon)$ and regresses its velocity $u_t = \partial_t I_t$ (Lipman et al., 2023; Albergo et al., 2025). On the linear path $z_t = (1-t)\epsilon + tx$, this target is $x - \epsilon$. Squared-error regression estimates the marginal field $v^\star(z,t) = \mathbb{E}[u_t \mid z_t = z, t]$, whose direction can change along a trajectory. Straight conditional lines can thus induce curved sampling trajectories even with an exact field. This motivates studying where the field receives supervision. JiT (Li & He, 2025) recovers velocity from a clean-image prediction as $(f_\theta(z,t) - z)/(1-t)$, providing a pixel-space setting in which training states can change while the clean target stays fixed.

Choosing the training geometry. The noise–data coupling determines which endpoints are paired; the interpolant determines which states and velocities lie between them. Minibatch optimal-transport methods change the coupling to straighten flows (Pooladian et al., 2023; Tong et al., 2024), while Metric Flow Matching changes path geometry to account for the data manifold (Kapuśniak et al., 2024). These choices suggest treating the input distribution of a pretrained model as a post-training variable. TSS-X uses a detached EMA prediction to move training states and retains JiT's clean target. For a model trained directly on velocity, changing a path also changes its derivative. AuxPath-FM establishes conditional flow matching for endpoint-vanishing additive paths (Peng & Gao, 2026); within that family, our transport-consistent construction supplies an EMA direction, a trajectory-level trust bound and native-head adapters. This connects the pixel-space intervention to exact-path velocity supervision.

Self-conditioning feeds a prediction through an additional network input (Chen et al., 2023; Stärk et al., 2024). In TSS, the EMA determines the training state while clean-image or exact-tangent supervision defines the target. The sampling network keeps its original input signature.

Changing the numerical sampler. For a fixed field, the solver determines how evaluations advance the sample. DPM-Solver++ improves numerical integration (Lu et al., 2025); VSFM studies inference-time velocity profiles and rescheduling (Bondar, 2026). A gain under one sampler may depend on its integration rule. We therefore evaluate TSS checkpoints under Heun and DPM++2M and use a discretely normalized velocity profile as a frozen-model control. These comparisons separate changes to training states from changes to checkpoint sampling.

From a sampling field to student training data. An alternative to integrating local velocities is to learn a map over a finite time interval. MeanFlow and improved MeanFlow learn interval-average fields (Geng et al., 2025; 2026); pixel MeanFlow expresses the field through a generalized denoised-image prediction and adds perceptual supervision (Lu et al., 2026a). Teacher-generated data provides another route to such maps. Rectified Flow trains on noise–endpoint couplings produced by a learned flow, and Re-MeanFlow studies rectified couplings and self-distillation (Liu et al., 2023; Zhang et al., 2025). CMT also uses intermediate rollout states: finite differences along frozen-teacher trajectories initialize later flow-map training (Hu et al., 2026). These methods make the teacher's outputs part of the student's training distribution. We ask whether improving a teacher at low NFE produces more useful synthetic pairs, measured by student updates and pool size under matched recipes (Section 5.4). Appendix A discusses consistency, self-conditioning and representation objectives in more detail.

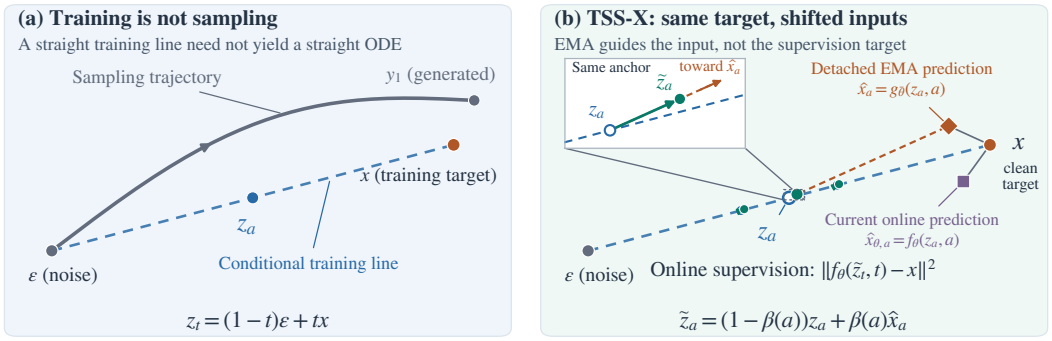

Figure 1: Training states and model predictions. Left: conditional training lines differ from sampling trajectories. Right: online and EMA predictions query the same z_a; gray segments measure their residuals to the clean target x. The box and inset show the same anchor; the orange ray passes through $\tilde{z}_a$ toward the EMA estimate $\hat{x}_a$. Green points use $\beta_{\max} = 0.05$; clean x remains the target. Schematic prediction locations illustrate the construction; relative accuracy requires measurement.

3 Training-State Shaping

3.1 Clean-Target State Re-Indexing

Even an exact marginal flow can follow curved trajectories despite straight conditional training lines (Figure 1). TSS-X shapes the training inputs while sampling follows the learned field. At anchor $a \in (0, 1)$, let $\hat{x}_a = g_{\bar{\theta}}(z_a, a, c)$ be a detached EMA prediction with class condition c. TSS-X trains on

$$\tilde{z}_t = z_t + \beta(t)(\hat{x}_a - z_t), \qquad \mathcal{L}_{\text{TSS-X}} = \mathbb{E}\big[w(t)\|f_\theta(\tilde{z}_t, t, c) - x\|_2^2\big]. \tag{1}$$

We use $\beta(t) = \beta_{\max} \exp(-(t-a)^2/(2\sigma_\beta^2))$, with $a = 0.5$, $\beta_{\max} = 0.05$ and $\sigma_\beta = 0.08$. The continuation's absolute-epoch annealing schedule has already saturated, so shaping is at full strength from its first update. This state re-indexing construction estimates the conditional clean-image mean $\mathbb{E}[x \mid \tilde{z}_t, t, c]$; exact-path velocity training instead requires the shaped path's tangent.

Decomposing the direction. Let $D_t = x - z_t$ and $e_a = \hat{x}_a - x$. Project the error onto the base tangent u_a:

$$P_a = u_a \frac{\langle e_a, u_a \rangle_M}{\langle u_a, u_a \rangle_M}, \qquad N_a = e_a - P_a, \qquad \hat{x}_a - z_t = D_t + P_a + N_a. \tag{2}$$

Here M is a positive diagonal metric; experiments use $M = I$. D, P and N are algebraic components of the shaping direction. On a linear path, the D-only state $z_t + \beta(t)(x - z_t)$ equals $z_{t+\beta(t)(1-t)}$, a time relabeling of the same conditional line. This clean-target control is distinct from TSS-T below, which also changes the tangent target.

4 Transport-Consistent Shaping and Its Controls

4.1 A Shaped Path with an Exact Target

For any differentiable conditional path, choose a detached direction d_a and scale s_a that are constant along each sampled trajectory. With a compact, endpoint-safe C^1 profile $b(t)$, define

$$z_t^\star = z_t + b(t) s_a d_a, \qquad u_t^\star = u_t + b'(t) s_a d_a. \tag{3}$$

Our linear-path instance uses $d_a = \lambda_D D_a + \lambda_N N_a$, where $D_a = x - z_a$. An x-prediction head supplies $\hat{x}_a$ directly; a native-velocity head gives $\hat{x}_a = z_a + (1-a)\hat{u}_a$ before the error projection (Appendix Eq. (25)). The full construction sets $\lambda_D = \lambda_N = 1$. For $q(v) = 3v^2 - 2v^3$, the profile is

$$b(t) = \begin{cases} \eta q((t - a + h)/h), & a - h \le t \le a, \\ \eta q((a + h - t)/h), & a < t \le a + h, \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

with $a = 0.5$, $h = 0.25$ and $\eta = 0.05$, ramped over 50 updates in the latent setting. Both b and b' vanish at the support boundary. The trust scale is

$$s_a = \min\left(1, \frac{\rho \ \mathrm{RMS}(u_a)}{\max\{\|b'\|_\infty \mathrm{RMS}(d_a), \varepsilon\}}\right), \qquad \rho = 0.25, \quad \varepsilon = 10^{-12}. \tag{5}$$

We set $s_a = 1$ if $\|b'\|_\infty = 0$. RMS is computed over a sample's components. Because d_a and s_a do not vary with training time, $u_t^\star = dz_t^\star/dt$ and $\sup_t \mathrm{RMS}(u_t^\star - u_t) \le \rho \mathrm{RMS}(u_a)$. Both state and tangent return to the base path where $b = b' = 0$. Under the regularity assumptions in Appendix B, the conditional mean of $u_t^\star$ defines a valid marginal flow with the original endpoint marginals.

Native prediction heads. For a linear schedule $z_t = \alpha(t)x + \sigma(t)\epsilon$, let $\Delta = \alpha\sigma' - \sigma\alpha' \neq 0$. The shaped state and velocity determine effective endpoints and a native target:

$$\begin{bmatrix} x^\star \\ \epsilon^\star \end{bmatrix} = \begin{bmatrix} \alpha & \sigma \\ \alpha' & \sigma' \end{bmatrix}^{-1} \begin{bmatrix} z_t^\star \\ u_t^\star \end{bmatrix}, \qquad y_t^\star = c_x(t)x^\star + c_\epsilon(t)\epsilon^\star. \tag{6}$$

The transport-consistent regression is $\mathcal{L}_{\mathrm{TC}} = \mathbb{E}[w_y(t)\|f_\theta(z_t^\star, t, c) - y_t^\star\|_2^2]$. The EMA anchor is converted by inverting $\left[\begin{smallmatrix} \alpha(a) & \sigma(a) \\ c_x(a) & c_\epsilon(a) \end{smallmatrix}\right]$ on $(z_a, \widehat{y}_a)$; this system must also be nonsingular. These adapters cover clean-image, noise, physical-velocity and diffusion-v heads in pixel or Euclidean latent coordinates. Loss weighting and empirical performance remain specific to each parameterization. Appendix A gives the expanded conversions and component controls; Appendix H gives algorithms.

Proposition 1 (Exact-path regression)**.** *If the trajectory-fixed direction and scale satisfy the conditions above, then $u_t^\star = \partial_t z_t^\star$. If $u_t^\star$ is integrable and differentiation under expectation is valid for smooth compactly supported test functions, $v^\star(z,t) = \mathbb{E}[u_t^\star \mid z_t^\star = z]$ satisfies the continuity equation for the marginal law of $z_t^\star$. The endpoint marginals agree with the base path.*

The proof differentiates the path and its expectation against a test function (Appendix B). This is a population statement; finite-capacity regression and numerical sampling remain empirical questions.

4.2 TSS-T: DETERMINISTIC TIME REPARAMETERIZATION

Setting $\lambda_D = 1$ and $\lambda_N = 0$ gives $d_a = D_a = (1-a)u$ on the linear path and a deterministic scale $s_D = \min\{1, \rho/(\|b'\|_\infty(1-a))\}$. Then

$$\phi(t) = t + (1-a)s_D b(t), \qquad z_t^\star = z_{\phi(t)}, \qquad u_t^\star = \phi'(t)u. \tag{7}$$

This time reparameterization preserves both endpoints, with $|\phi' - 1| \le \rho < 1$. For the tested full-strength profile, $(1-a)\|b'\|_\infty = 0.15 < \rho$, so $s_D = 1$ and $\max_t |\phi'(t) - 1| = 0.15$. TSS-T trains on the modified state and tangent while conditioning on t, using no EMA direction. Comparing it with the full construction tests the normal component under the same trust rule: N_a can increase $\mathrm{RMS}(d_a)$ and reduce the full construction's s_a below one, while $s_D = 1$. This differs from changing state and target one at a time.

The shaped path also gives an exact interval-average target:

$$\overline{u}_{s,t}^\star = \overline{u}_{s,t} + \frac{b(t) - b(s)}{t - s} s_a d_a. \tag{8}$$

The correction is bounded by $\rho \mathrm{RMS}(u_a)$ in RMS and is zero over the full interval $[0,1]$. This identity defines an interface to flow-map targets. Its effect on learned trajectories and one-step generation is an empirical question.

4.3 AN INFERENCE-ONLY COMPANION

The D-only tangent factor $\phi'(\tau) = 1 + (1-a)s_D b'(\tau)$ from Eq. (7) also defines a frozen-field control. At solver nodes τ_i with quadrature weights ω_i, set

$$Z_K = \frac{\sum_i \omega_i \phi'(\tau_i)}{\sum_i \omega_i}, \qquad m_i = \phi'(\tau_i)/Z_K. \tag{9}$$

Table 1: Methods and intervention points. The training variants retain local-field ODE sampling; S_{norm} changes inference only. Targets are expressed in the native prediction head. The final two rows describe related flow-map approaches; their combination with TSS remains untested.

Method / role	Training state	Supervision / use
Standard flow matching (FM)	Base interpolant	Base tangent
Clean-target shaping (TSS-X)	Shift toward detached EMA prediction	Retain clean-image target x
Full transport-consistent TSS	Along-path and EMA-guided normal displacement	Exact shaped-path tangent
Time reparameterization (TSS-T)	Retimed $z_{\phi(t)}$; no EMA	Rescaled tangent $\phi'(t)u$
Tangent-only control	Unchanged base state	Shaped-path tangent
Normalized velocity profile (S_{norm})	No training	Rescale frozen field; no extra calls
Related flow-map approaches		
MeanFlow / pixel MeanFlow (pMF) (Geng et al., 2025; Lu et al., 2026a)	Base interpolant	Average-velocity identity and method-specific losses; direct interval sampling
Teacher-trajectory mid-training (CMT-MF) (Hu et al., 2026)	Frozen-teacher rollout states	Within-trajectory finite differences; initialize flow-map training

S_{norm}multiplies the final physical velocity after classifier-free guidance (CFG) by m_i at the sampler's current state and time, with no explicit state/time remapping or extra network calls. Its discrete multiplier mass is one, preserving the total drift of a constant field. TSS-T instead changes the training states and tangent targets; neither method replaces the inference-time query coordinate by $\phi(t)$. The profile is frozen across pipelines; only Z_K changes with the grid. With one interval, $m_0 = 1$, so this normalization gives no direct one-step modification.

Table 1 summarizes what each construction changes. TSS-X, TSS-T and the full transport-consistent construction keep a local-field sampling interface; MeanFlow and pixel MeanFlow instead learn interval-conditioned fields. Teacher-trajectory mid-training supplies supervision for a flow-map student and is a separate intervention.

5 Experiments

5.1 Protocols and call accounting

We use class-conditional ImageNet and Fréchet inception distance (FID) (Deng et al., 2009; Heusel et al., 2017). Unless noted otherwise, each cell contains 50,000 generated images. JiT-B/16 continuations share an epoch-600 source and run to epoch 700 with global batch 1,024, AdamW and matched training seeds $0, 1, 2$. The base learning rate is 5×10^{-5}. Evaluation uses EMA, CFG 2.5 and guidance interval $[0.1, 1]$. We report the number of function evaluations (NFE), counting each guided velocity-field evaluation once with its conditional and unconditional branches together. The two branches cost twice this many per-image network evaluations. DPM++2M uses one NFE per interval. We evaluate the common 10/20/50-NFE grid on Blackwell; the main latent table focuses on 10 NFE. Heun comparisons and their call accounting remain in Appendix A at the original measured budgets.

The latent study continues a local LightningDiT-B/2 800k state, including Adam and EMA, for 20k updates on the full latent cache with batch 1,024. Three matched seeds compare vanilla, TSS-T and

Table 2: JiT EMA FID50K. Δ = TSS-X minus vanilla. Heun 19/49/99 NFE corresponds to 10/25/50 intervals. Checkpoint and evaluation details appear in Appendices A.4.3 and C.

	DPM++2M				Heun			
Model	NFE	Vanilla	TSS-X	Δ	NFE	Vanilla	TSS-X	Δ
B/16	10	9.6193	8.2636	-1.3557	19	8.2543	7.1112	-1.1431
B/16	20	7.0641	5.9952	-1.0688	49	4.03	4.07	$+0.04$
B/16	50	3.7107	3.8176	$+0.1070$	99	3.6851	3.7163	$+0.0312$
L/16	10	7.9974	6.7743	-1.2231	19	6.6247	5.4201	-1.2046
L/16	20	5.7918	4.7803	-1.0115	49	3.02	2.70	-0.32
L/16	50	2.7905	2.6460	-0.1444	99	2.4767	2.2450	-0.2317
H/16	10	7.6652	6.6524	-1.0128	19	6.3556	5.5335	-0.8221
H/16	20	4.9841	4.2417	-0.7424	49	2.1606	2.1562	-0.0044
H/16	50	1.9881	2.0466	$+0.0584$	99	1.9405	1.8988	-0.0417

full transport-consistent TSS under the original velocity MSE-plus-cosine loss. We use this local configuration for within-platform comparisons. Paired Student-t intervals use three seeds, conditional on shared source, data and evaluation noise, without multiplicity correction. Full protocols and earlier 625-update experiments are in Appendix A.

5.2 Low-call JiT gains across solvers and scales

TSS-X improves DPM++2M in all three B/16 pairs at 10 and 20 NFE (Table 2). Mean FID falls from 9.6193 to 8.2636 at 10 NFE and from 7.0641 to 5.9952 at 20 NFE, with paired differences of -1.3557 and -1.0688. The L/16 and H/16 pairs also improve at both budgets, extending the low-call result across model scales.

For L/16, the paired differences are -1.2231 and -1.0115 at 10 and 20 NFE; for H/16, they are -1.0128 and -0.7424. At 50 NFE, the differences are -0.1444 for L/16, $+0.1070$ for B/16 and $+0.0584$ for H/16. The strongest gains therefore occur at low NFE. Heun comparisons support the same low-call pattern; higher-resolution results and direction-component controls are reported in Appendix A. Those controls favor retaining the full TSS-X direction for JiT.

Inference additivity. TSS-X composes with the inference-only companion S_{norm} under Heun. The frozen S_{norm} profile also improves LightningDiT and RAE within their solvers. Appendix A reports the additivity measurements at their original sampling budgets, alongside solver-versus-companion comparisons, RAE recall and sampling conventions.

5.3 Latent low-call gains at 10 NFE

TSS-T improves generation at 10 NFE in the latent velocity model without an EMA teacher (Table 3). Three matched seeds improve Euler FID from 14.6458 to 13.9658 and DPM++2M from 10.2868 to 9.6469; full transport-consistent TSS gives 13.6358 and 9.6142. Full further improves Euler in all three pairs: full minus TSS-T is -0.3300 with paired 95% interval $[-0.4936, -0.1663]$. The DPM++2M increment is -0.0327 with interval $[-0.2880, +0.2226]$. TSS-T captures most of the low-call gain, with a further EMA-guided benefit at Euler NFE10. The benefit depends on solver and budget: Appendix A.4.4 reports the full sweep, including higher-budget regressions, and all paired readings. Appendix E reports the earlier 625-update factorial.

5.4 Data efficiency of TSS-generated pairs

Matched epoch-700 JiT teachers, vanilla and TSS-X, each generate 200k class-balanced float32 noise–endpoint pairs using CFG 2.5 and 10 DPM++2M field calls. We denote these pools by Vanilla$_{10}$ and TSS-X$_{10}$; subscripts count teacher field calls. Vanilla$_{14}$ is the endpoint-FID-matched control. Prior indices and pool sizes match; endpoints differ. We use a reflow student and a flow-map student trained with the MeanFlow identity via a Jacobian-vector product (JVP), using the boundary-velocity

Table 3: LightningDiT-B/2 EMA FID50K at 10 NFE (11 grid points), averaged over three matched training seeds after 20k full-data continuation updates. Frozen CFG20; common Blackwell evaluation. TSS-T is EMA-free time reparameterization. The full 10/20/50-NFE sweep appears in Table 10.

Solver	NFE	Vanilla	TSS-T	Full transport-consistent TSS
Euler	10	14.6458	13.9658	13.6358
DPM++2M	10	10.2868	9.6469	9.6142

form of the pMF objective without perceptual or auxiliary-velocity losses (Geng et al., 2025; Lu et al., 2026a). Each consumer uses matched initialization, batch size (256) and training recipe. Two streams per consumer share the original teachers and pools; a third flow-map stream uses a second teacher pair. Reflow and the second-teacher-pair runs end at 20k updates; the two original flow-map runs continue to 100k.

Students inherit teacher guidance (CFG 2.5) through the pair endpoints and sample without further guidance. Labels are never dropped, so the paired objective trains no unconditional student field. MeanFlow and pMF likewise incorporate guidance during training to avoid an additional CFG combination at sampling (Geng et al., 2025; Lu et al., 2026a). Each 20k-update run processes 5.12M sample presentations (about four ImageNet-equivalent epochs) and approximates a 10-NFE teacher map; the original Vanilla$_{10}$/TSS-X$_{10}$ pools have endpoint FIDs of 9.4/8.2. We use one-call FID to compare pair quality at a fixed student budget (Appendix A.4.9); Section 5.5 reports post-training costs and sampling quality.

The JiT-B/16 architecture allows both arms to share a vanilla-teacher initialization. The released pMF network has a different architecture and provides PyTorch inference code with training in JAX (Lu et al., 2026b). We evaluate its mature checkpoint separately below (Appendix I.1).

Fewer updates to target quality. TSS pairs let both consumers reach the reference EMA FID with fewer interpolated updates (Figure 2). We estimate the first downward crossing by linear interpolation between measured checkpoints; the logarithmic axis affects only the display. At fixed batch size, fewer updates mean fewer sample presentations, including repeated accesses to the same pairs. Distinct-pair requirements are tested separately below. Using each vanilla run's final FID plus 0.5 as the target, the original flow-map streams give interpolated Vanilla/TSS update ratios of 2.38 and 2.23. The first measured checkpoints to reach these targets are 20k for the TSS-X pool and 40k for the vanilla pool. The reflow ratios are 1.23 and 1.19, although both sources first reach their targets at the measured 20k checkpoint. Figure 2 labels the mean ratios, $2.3\times$ and $1.2\times$. Appendix A.4.9 reports single-reference and multi-target tables; crossings for targets already reached by 5k are left-censored.

A second teacher pair. With the seed-0 teacher pair, the 20k one-call EMA FID difference is -2.6120 (TSS minus vanilla). At this pair's own 20k reference, interpolated update ratios are 1.90 for one call and 2.07 for two calls; both sources first reach the one-call target at the measured 20k checkpoint. This confirms the direction of the result with a second teacher pair at the measured budget; the 100k extension covers the original pair. Students initialize from their respective pair's vanilla teacher. Prior indices and evaluation noise remain shared across pairs; Appendix A.4.10 details these dependencies and execution differences.

Fewer pairs. With 20k student updates, the 50k-pair TSS-X prefix gives lower one- and two-call FID than the 200k vanilla reference in Table 4. The 100k TSS-X prefix comes within 0.2 FID of the full TSS-X pool. Thus a quarter-sized TSS-X pool gives better measured student quality

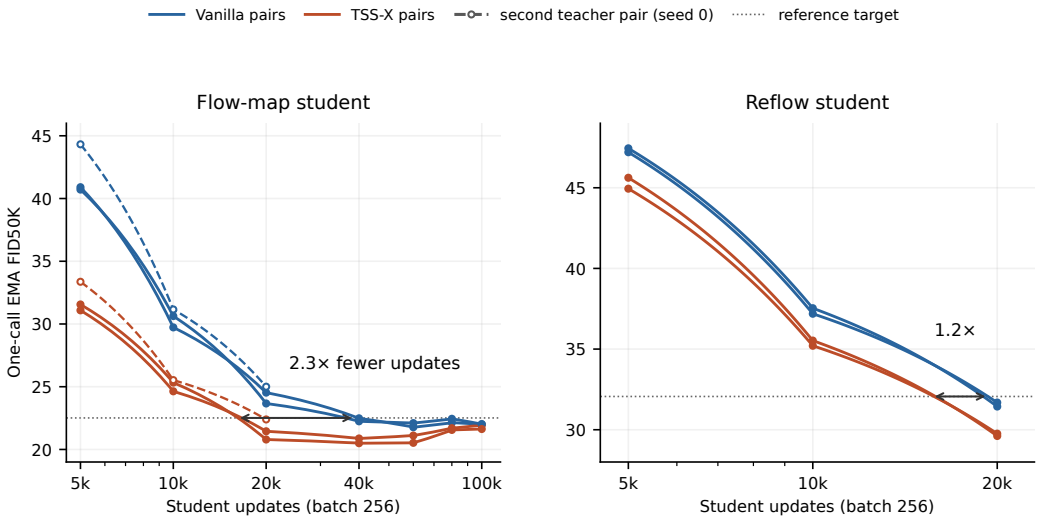

Figure 2: One-call EMA FID versus student updates (log scale). Solid lines show two student streams sharing the original teacher pair; dashed lines with open markers use a second, independently trained teacher pair (seed 0). Markers denote measured checkpoints. For the two solid streams, the dotted line averages the references (each vanilla final FID plus 0.5), arrows join the mean interpolated first-crossing updates, and labels give the mean Vanilla/TSS update ratio. Appendix A.4.9 includes per-stream curves, multi-target tables, two-call flow-map sampling and eight-interval reflow sampling (15 forwards).

than the full vanilla pool under this recipe. The 200k reference runs use an earlier training stream (Appendix A.4.11).

Endpoint quality and coupling geometry. TSS-X pools have slightly higher kNN precision and recall than vanilla pools for both teacher pairs. The 14-call vanilla control pool has slightly better endpoint FID and recall than the TSS-X pool yet trains worse reflow students in both streams. TSS-X's downstream advantage therefore extends beyond the ranking given by endpoint FID. In both teacher pairs, TSS-X endpoints have about 2% lower transport cost $\mathbb{E}\|x_1 - x_0\|^2$ from their generating noise and about 50% higher cosine alignment between noise and endpoints, at equal endpoint energy (Appendix A.4.11). These correlations motivate the straighter-coupling hypothesis for future testing.

Endpoints. On fresh 640k pools with cosine decay to zero, TSS-X is better in all eight endpoint comparisons at 60k updates (one-call EMA 20.44 vs. 21.09; Online 19.70 vs. 20.69; Table 4), with Online updates-to-target ratios near 1.5 to 2. On the reused 200k pools, one-call EMA FID differs by less than 0.4 at 100k updates, after about 128 accesses per pair. For those fixed pools, the measured advantage is earlier attainment along the learning curve (Appendix A.4.9).

5.5 Cost

Direct flow-map post-training. Guided flow-map post-training of the same JiT-B/16 network (arm F) reaches EMA FID 38.26/14.39/8.67/6.70 at 1/2/5/10 NFE after a frozen 43k-update continuation with a timing-calibrated compute proxy. Arm F starts from epoch-600 EMA weights and fresh AdamW; the Gate2B TSS-X reference in Table 5 resumes Online weights and optimizer state. The different loading rules make this a shared-source comparison. Appendix A.5 gives the JiT-specific recipe, all milestones and the contextual NFE comparison.

Table 4: Fewer pairs and a larger pool with learning-rate decay. Same flow-map student recipe and common vanilla initialization; pools are index prefixes of the seed-1 200k pools (50k, 100k) or fresh 640k pools from the same teachers. FID50K at 20k updates (constant LR) or 60k updates (cosine to zero, EMA 0.9999). Sampling without further guidance (teacher CFG 2.5 in the pairs); calls are actual network forwards.

Source	Pairs	Updates	EMA 1	EMA 2	Online 1	Online 2
Vanilla	50k	20k, const.	27.55	20.97	26.61	20.77
TSS-X	50k	20k, const.	23.23	18.47	23.26	18.83
Vanilla	100k	20k, const.	25.48	20.07	24.60	19.63
TSS-X	100k	20k, const.	21.64	17.58	21.98	17.89
Vanilla	200k	20k, const.	24.53	19.63	24.20	19.74
TSS-X	200k	20k, const.	21.45	17.53	21.84	17.90
Vanilla	640k	60k, cosine	21.09	18.19	20.69	18.13
TSS-X	640k	60k, cosine	20.44	17.72	19.70	17.16

TSS-X adds one detached EMA forward per update and leaves inference work unchanged. A forward/backward proxy of $F/2F$ gives a 33.33% per-update increment; matched JiT-B/16 measurements give 4.30–15.50% wall-clock overhead (10.08% mean) and about 1.1% sampled peak-memory growth. The proxy counts operations, while wall time reflects their execution. Downstream costs include teacher post-training, endpoint generation, student training and evaluation. Reducing rollout calls from 19 to 10 saves 47.4% within the generation stage; each cached pool incurs this cost once. Appendix G reports hardware-specific allocations.

6 Discussion and limitations

Teacher sample FID and student learning curves answer different practical questions. The former measures generation at a specified sampling budget; the latter measures how quickly a fixed student recipe learns from a teacher's pairs. Our endpoint-FID-matched control ranks differently under these evaluations. This motivates assessing a teacher both as a generator and as a source of training couplings.

JiT gains hold at low call budgets under two solver families. TSS-T accounts for most latent low-call gains, with an EMA-guided increment at Euler NFE10; transfer to other prediction heads remains to be tested. The heuristic anchor $a = 0.5$ remains a tunable choice; the single-run anchor sweep in Appendix C leaves sensitivity unresolved. Our analysis concerns conditional-path identities. Learned-field accuracy and student-mapping complexity require separate measurements.

Data-efficiency evidence covers two JiT teacher pairs and two pair-trained consumers initialized from a vanilla teacher. Consumer dependence appears in two further recipes: direct one-step fine-tuning removes the reflow advantage and degrades multi-call sampling; adapting a mature pMF checkpoint to fixed pairs degrades both arms, with the TSS-X pool worse (Appendices F and I.1). Our evidence concerns low-NFE sampling and pair quality under matched student budgets. Lower-cost generation poses synthetic-media misuse risks.

7 Conclusion

Training-state shaping improves low-call JiT generation across two solvers; TSS-T explains most latent gains. On fixed 200k-pair pools, matched reflow and flow-map students reach target FID with fewer updates using TSS pairs. A 50k-pair TSS-X prefix outperforms the 200k vanilla reference in Table 4. One-call EMA endpoints are similar at 100k on reused 200k pools, whereas fresh 640k pools with learning-rate decay retain TSS-X gains in all eight endpoint comparisons. The results identify teacher-side training as a way to improve the data efficiency of specific student recipes, measured through sample presentations and pool size.

AI USE STATEMENT

Generative AI assistants were used to help refine hypotheses and experimental designs, implement methods and evaluation code, troubleshoot and manage experiments, interpret results, search and summarize related work, and draft, translate, edit and format the manuscript and figures. The reported numerical results come from model runs and saved evaluation artifacts, not from language-model-generated measurements. Checks of AI-assisted implementations included unit tests, gradient and checkpoint-continuation checks, and comparisons with recorded evaluation outputs. These checks do not replace the authors' responsibility for the final claims, citations, code and disclosures.

REPRODUCIBILITY STATEMENT

The path definitions, target conversions and assumptions are given in Sections 3 and 4, with proofs and adapter details in Appendix B. Implementation settings and provenance are documented in Appendices H and J; the downstream protocols, budget continuations and negative results are detailed in Appendices F, F.1 and I.1. Hardware-specific cost measurements are reported in Appendix G. The experimental descriptions distinguish shared teachers and data pools from independent training streams and retain the reported Online and EMA comparisons.

REFERENCES

Michael S. Albergo, Mark Goldstein, Nicholas M. Boffi, Rajesh Ranganath, and Eric Vanden-Eijnden. Stochastic interpolants with data-dependent couplings. In *International Conference on Machine Learning*, 2024.

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Nicholas M. Boffi, Michael S. Albergo, and Eric Vanden-Eijnden. How to build a consistency model: Learning flow maps via self-distillation. In *Advances in Neural Information Processing Systems*, 2025.

Vitalii Bondar. Velocity scheduled flow matching. *arXiv preprint arXiv:2607.11442*, 2026.

Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. In *International Conference on Learning Representations*, 2023.

Lucas Degeorge, Paul Couairon, Arijit Ghosh, Alexei A. Efros, David Picard, and Vicky Kalogeiton. Balancing frequencies and pixels in flow matching. *arXiv preprint arXiv:2609.02748*, 2026.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *International Conference on Learning Representations*, 2025.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *Advances in Neural Information Processing Systems*, 2025.

Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J. Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30467–30476, 2026.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

Zheyuan Hu, Chieh-Hsin Lai, Yuki Mitsufuji, and Stefano Ermon. CMT: Mid-training for efficient learning of consistency, mean flow, and flow-map models. In *International Conference on Learning Representations*, 2026.

Kacper Kapuśniak, Peter Potaptchik, Teodora Reu, Leo Zhang, Alexander Tong, Michael Bronstein, Avishek Joey Bose, and Francesco Di Giovanni. Metric flow matching for smooth interpolations on the data manifold. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

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.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 22 (4):730–751, 2025. doi: 10.1007/s11633-025-1562-4. arXiv:2211.01095.

Yiyang Lu, Susie Lu, Qiao Sun, Hanhong Zhao, Zhicheng Jiang, Xianbang Wang, Tianhong Li, Zhengyang Geng, and Kaiming He. One-step latent-free image generation with pixel mean flows. *arXiv preprint arXiv:2601.22158*, 2026a.

Yiyang Lu et al. Pixel Mean Flows: Official PyTorch inference implementation. `https://github.com/Lyy-iiis/pMF/tree/990e81a84249dbd68a128accef67eb95621d10b1`, 2026b. Torch branch, commit 990e81a; accessed September 18, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Xin Peng and Ang Gao. Flow matching with arbitrary auxiliary paths. *arXiv preprint arXiv:2605.06364*, 2026.

Yansong Peng, Kai Zhu, Yu Liu, Pingyu Wu, Hebei Li, Xiaoyan Sun, and Feng Wu. FACM: Flow-anchored consistency models. *arXiv preprint arXiv:2507.03738*, 2025.

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 *International Conference on Machine Learning*, 2023.

Amirmojtaba Sabour, Sanja Fidler, and Karsten Kreis. Align your flow: Scaling continuous-time flow map distillation. In *Advances in Neural Information Processing Systems*, 2025.

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

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, 2023.

Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Harmonic self-conditioned flow matching for joint multi-ligand docking and binding site design. In *International Conference on Machine Learning*, 2024.

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.

Qi Yan, Brian Zhang, Yutong Zhang, Daniel Yang, Joshua White, Di Chen, Jiachao Liu, Langechuan Liu, Binnan Zhuang, Shaoshuai Shi, et al. TrajFlow: Multi-modal motion prediction via flow matching. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3923–3928, 2025. doi: 10.1109/IROS60139.2025.11247503.

Jiawei Yang, Zhengyang Geng, Xuan Ju, Yonglong Tian, and Yue Wang. Representation Fréchet loss for visual generation. *arXiv preprint arXiv:2604.28190*, 2026.

Ling Yang, Zixiang Zhang, Zhilong Zhang, Xingchao Liu, Minkai Xu, Wentao Zhang, Chenlin Meng, Stefano Ermon, and Bin Cui. Consistency flow matching: Defining straight flows with velocity consistency. *arXiv preprint arXiv:2407.02398*, 2024.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6613–6623, 2024.

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.

Xinxi Zhang, Shiwei Tan, Quang Nguyen, Quan Dao, Ligong Han, Xiaoxiao He, Tunyu Zhang, Chengzhi Mao, Dimitris Metaxas, and Vladimir Pavlovic. Overcoming the curvature bottleneck in MeanFlow. *arXiv preprint arXiv:2511.23342*, 2025. Version 3, March 2026.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations*, 2026.

Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In *International Conference on Machine Learning*, 2024.

A Detailed methods and experimental comparisons

This appendix gives the full target conversions, component controls and evaluation protocols underlying the main-text summaries. Each evaluation stream is reported separately, with paired comparisons computed within that stream.

A.1 Background and related work

Flow matching and conditional paths. For endpoints x_0 and x_1, a conditional interpolant is $z_t = I_t(x_0, x_1)$ with physical velocity $u_t = \partial_t I_t(x_0, x_1)$. Flow matching regresses a model toward this conditional velocity or an equivalent native target (Lipman et al., 2023; Albergo et al., 2025). JiT uses the linear noise-to-data path

$$z_t = (1-t)\epsilon + tx, \qquad u_t = x - \epsilon, \tag{10}$$

and exposes a clean-x head (Li & He, 2025). Its velocity is recovered as $u_\theta(z,t) = (f_\theta(z,t) - z)/(1-t)$. The architecture follows the ViT and DiT lineage (Dosovitskiy et al., 2021; Peebles & Xie, 2023) and is related to diffusion and score models (Ho et al., 2020; Song et al., 2021; Karras et al., 2022).

Changing paths, couplings, and losses. Minibatch optimal-transport flow matching changes the noise-to-data coupling (Pooladian et al., 2023; Tong et al., 2024). Metric Flow Matching and data-dependent stochastic interpolants modify the conditional path or its geometry (Kapuśniak et al., 2024; Albergo et al., 2024). FACM adds a derivative-based consistency term (Peng et al., 2025). These interventions differ from TSS-X, which keeps clean-target regression but changes its input distribution, and from transport-consistent TSS, which changes a state and its exact tangent together.

AuxPath-FM establishes conditional flow matching for endpoint-vanishing additive auxiliary paths (Peng & Gao, 2026). Within this family, transport-consistent TSS introduces an endpoint-conditioned, detached EMA direction, a per-trajectory trust bound and native-head adapters, evaluated through post-training. VSFM studies unit-mass scalar progress profiles and inference-time rescheduling of pretrained local flows (Bondar, 2026). Our inference component retains the original query times and multiplies the final learned field using a TSS-derived, discretely normalized solver profile.

Consistency, conditioning, and representation objectives. Consistency Models, Consistency Flow Matching, and Align Your Flow learn consistency or teacher-derived targets (Song et al., 2023; Yang et al., 2024; Sabour et al., 2025). Self-conditioning and related flow methods feed predictions back through an added conditioning route (Chen et al., 2023; Stärk et al., 2024; Yan et al., 2025). REPA aligns internal features to a frozen encoder (Yu et al., 2025), while FD-loss uses distribution-level representation statistics (Yang et al., 2026). In TSS-X and full transport-consistent TSS, the EMA supplies the detached anchor direction used to shape training states. The regression targets follow the clean-image or exact-path construction, and the inference-time network signature stays unchanged.

Concurrent frequency-balanced flow matching schedules Fourier-domain supervision toward the pixel-domain velocity loss (Degeorge et al., 2026). It changes the regression metric and training curriculum rather than the sampled state, tangent, or inference field. We borrow radial-spectrum measurements as diagnostics only and do not infer that spectral changes cause our FID changes.

Flow maps and few-step generation. MeanFlow and improved MeanFlow replace local-velocity prediction with an interval-average field (Geng et al., 2025; 2026). Pixel MeanFlow parameterizes this field through a generalized denoised-image output and adds pixel-space perceptual supervision in its reported system (Lu et al., 2026a). CMT-MF regresses fixed finite differences from actual frozen-teacher trajectories to initialize later MeanFlow training (Hu et al., 2026). Re-MeanFlow studies rectified couplings and MeanFlow self-distillation (Zhang et al., 2025). Our downstream study builds on synthetic-pair training to measure the incremental value of a TSS-trained teacher under matched rollout and student protocols. For a direct objective such as MeanFlow, transport-consistent TSS can replace the conditional state and instantaneous target inside the diagonal regression. A teacher-trajectory objective requires actual rollouts from a trained TSS teacher or an explicitly shaped solver. The downstream effect of either interface requires measurement. Shortcut models add step-size

conditioning (Frans et al., 2025); Rectified Flow, progressive distillation, DMD, and SiD use reflow, distillation, or adversarial training (Liu et al., 2023; Salimans & Ho, 2022; Yin et al., 2024; Zhou et al., 2024). Flow-map self-distillation learns a two-time map (Boffi et al., 2025). Equal-compute comparisons with these full training pipelines remain open.

A.2 Training-state shaping

A.2.1 TSS-X: clean-target state re-indexing

Let $a \in (0, 1)$ be a fixed anchor and let $g_{\bar{\theta}}$ be the detached EMA model. In JiT, the anchor prediction is

$$\widehat{x}_a = g_{\bar{\theta}}(z_a, a, c), \qquad z_a = (1-a)\epsilon + ax. \tag{11}$$

For a nonnegative localized profile $\beta(t)$, TSS-X changes the state to

$$\boxed{\widetilde{z}_t = z_t + \beta(t)(\widehat{x}_a - z_t)} \tag{12}$$

and retains JiT's clean-target regression:

$$\mathcal{L}_{\text{TSS-X}} = \mathbb{E}\Big[w(t)\,\|f_\theta(\widetilde{z}_t, t, c) - x\|_2^2\Big]. \tag{13}$$

The main JiT experiments use $\beta(t) = \beta_{\max}\exp(-(t-a)^2/(2\sigma_\beta^2))$, $a = 0.5$, $\beta_{\max} = 0.05$, and $\sigma_\beta = 0.08$. The continuation uses `correction_warmup` $= 0$ and records `beta_anneal_epochs` $= 30$, but it resumes at absolute epoch 601, after this absolute-epoch schedule has saturated. TSS-X therefore operates at full strength from the first continuation update.

Equation (13) is a coherent regression objective. At the population optimum,

$$f^\star(\widetilde{z}_t, t, c) = \mathbb{E}[x \mid \widetilde{z}_t, t, c]. \tag{14}$$

This construction estimates the conditional clean-image mean under the re-indexed states. Transport-consistent path training instead requires the derivative of the path traced by $\widetilde{z}_t$.

A.2.2 D/P/N decomposition

The submitted warp direction admits a useful decomposition. Define the clean-path shift $D_t = x - z_t$ and the anchor endpoint error $e_a = \widehat{x}_a - x$. For a positive diagonal metric M, split the error relative to the base-path tangent u_a:

$$P_a = u_a \frac{\langle e_a, u_a\rangle_M}{\langle u_a, u_a\rangle_M}, \qquad N_a = e_a - P_a. \tag{15}$$

The reported LightningDiT experiments use the Euclidean choice $M = I$; the diagonal form records the more general decomposition rather than an additional tuned component. Then

$$\widehat{x}_a - z_t = D_t + P_a + N_a. \tag{16}$$

These labels describe algebraic components of the shaping direction. On the linear path, the D-only state

$$z_t + \beta(t)(x - z_t) = z_{t+\beta(t)(1-t)}$$

is an exact time relabeling along the same conditional line. The mechanism experiments in Section 5.2 test which combinations survive post-training rather than inferring a mechanism from direction magnitude alone.

A.3 Transport-consistent TSS

A.3.1 General construction and the tested anchor direction

Let $z_t = I_t(\xi)$ be any differentiable conditional path, with trajectory variables ξ and tangent $u_t = \partial_t I_t(\xi)$. Transport-consistent TSS requires a detached direction $d_a = d(\xi, I_a, u_a, g_{\bar{\theta}})$ and scale s_a that are fixed with respect to the sampled training time. The construction below then applies to the path regardless of how d_a is obtained. Our experiments use a linear schedule and the following EMA-derived D+N instance.

For the linear path and a physical-velocity teacher $\widehat{u}_a$, the anchor prediction gives the effective endpoint estimate and error

$$\widehat{x}_a = z_a + (1-a)\widehat{u}_a, \qquad e_a = \widehat{x}_a - x = (1-a)(\widehat{u}_a - u_a). \tag{17}$$

After the metric projection in Eq. (15), we use the trajectory-fixed direction

$$d_a = \lambda_D D_a + \lambda_N N_a, \qquad D_a = x - z_a. \tag{18}$$

Full transport-consistent TSS sets $\lambda_D = \lambda_N = 1$; the time-only variant below sets $\lambda_N = 0$. D+N has a native-velocity construction. On JiT, the three-seed test in Table 5 favors retaining the original TSS-X direction.

Let $q(v) = 3v^2 - 2v^3$, $l = a - h$, and $r = a + h$, with the support inside $[0, 1]$. We use the compact C^1 profile

$$b(t) = \begin{cases} \eta q((t-l)/(a-l)), & l \le t \le a, \\ \eta q((r-t)/(r-a)), & a < t \le r, \\ 0, & \text{otherwise.} \end{cases} \tag{19}$$

Both b and b' vanish at the support boundary. The tested LightningDiT setting uses $a = 0.5$, $h = 0.25$, and $\eta = 0.05$, ramped over the first 50 continuation updates.

Transport-consistent TSS uses the per-trajectory trust scale

$$s_a = \min\left(1, \frac{\rho \ \mathrm{RMS}(u_a)}{\max\{\|b'\|_\infty \ \mathrm{RMS}(d_a), \varepsilon\}}\right). \tag{20}$$

Here $\mathrm{RMS}(q) = \sqrt{\|q\|_2^2/m}$ for an m-component sample and $\varepsilon = 10^{-12}$ matches the numerical implementation; if $\|b'\|_\infty = 0$, we set $s_a = 1$. We use $\rho = 0.25$. The shaped state and its training velocity are

$$\boxed{z_t^\star = z_t + b(t) s_a d_a, \qquad u_t^\star = u_t + b'(t) s_a d_a.} \tag{21}$$

The direction and trust scale depend on the sampled trajectory and fixed anchor, but not on the sampled training time.

A.3.2 Schedule and prediction-head adapters

Consider a differentiable linear schedule

$$z_t = \alpha(t)x + \sigma(t)\epsilon, \qquad u_t = \alpha'(t)x + \sigma'(t)\epsilon. \tag{22}$$

If $\Delta(t) = \alpha(t)\sigma'(t) - \sigma(t)\alpha'(t) \neq 0$, the shaped state-velocity pair uniquely determines effective endpoints:

$$\begin{bmatrix} x^\star \\ \epsilon^\star \end{bmatrix} = \begin{bmatrix} \alpha & \sigma \\ \alpha' & \sigma' \end{bmatrix}^{-1} \begin{bmatrix} z_t^\star \\ u_t^\star \end{bmatrix}. \tag{23}$$

A native linear head $y_t = c_x(t)x + c_\epsilon(t)\epsilon$, including clean-$x$, noise, physical velocity, and diffusion-v, receives

$$y_t^\star = c_x(t)x^\star + c_\epsilon(t)\epsilon^\star. \tag{24}$$

The EMA anchor uses the same algebra in reverse. Given its native prediction $\widehat{y}_a$, define

$$\begin{bmatrix} \widehat{x}_a \\ \widehat{\epsilon}_a \end{bmatrix} = \begin{bmatrix} \alpha(a) & \sigma(a) \\ c_x(a) & c_\epsilon(a) \end{bmatrix}^{-1} \begin{bmatrix} z_a \\ \widehat{y}_a \end{bmatrix}, \qquad e_a = \widehat{x}_a - x, \tag{25}$$

whenever $\alpha(a)c_\epsilon(a) - \sigma(a)c_x(a) \neq 0$. Using this e_a in Eq. (15) constructs the D+N direction without an x-prediction output. For the tested linear schedule and a physical-velocity head, Eq. (25) reduces to $\widehat{x}_a = z_a + (1-a)\widehat{u}_a$. The transport-consistent TSS objective is

$$\mathcal{L}_{\mathrm{TC}} = \mathbb{E}\left[w_y(t) \, \|f_\theta(z_t^\star, t, c) - y_t^\star\|_2^2\right]. \tag{26}$$

The represented $(z_t^\star, u_t^\star)$ is independent of the exposed head. Native MSE values need not agree because $w_y(t)$ can differ across parameterizations. All operations above take place in the model's state coordinates. Thus x may be a pixel tensor or a Euclidean latent; transport-consistent TSS uses neither a pixel-space loss nor a decoder-specific operation.

A.3.3 TSS-T: ANALYTIC TIME REPARAMETERIZATION

On the linear path, $D_a = x - z_a = (1-a)u$. With $\lambda_D = 1$ and $\lambda_N = 0$, the trust factor is deterministic for a given profile and the shaped state becomes

$$\phi(t) = t + (1-a)s_D b(t), \qquad z_t^\star = z_{\phi(t)}, \qquad u_t^\star = \phi'(t)u, \tag{27}$$

where $s_D = \min\{1, \rho/(\|b'\|_\infty(1-a))\}$. Endpoint safety preserves $\phi(0) = 0$ and $\phi(1) = 1$, and the trust bound gives $|\phi' - 1| \le \rho < 1$ in our setting. TSS-T trains on the reparameterized state and tangent while conditioning the model on the training time t. It requires training and no EMA direction. It is distinct from the inference-only multiplier S_{norm}, which leaves the frozen model's query times unchanged. The full construction versus TSS-T tests the inclusion of the normal component under the same trust rule; it is not the same contrast as state-only versus tangent-only in the short-budget factorial.

A.3.4 FACTORIAL COMPONENT CONTROLS

To separate the state and tangent interventions, define binary state and target switches $S, T \in \{0, 1\}$:

$$z_t^{(S)} = z_t + S\, b(t)s_a d_a, \qquad u_t^{(T)} = u_t + T\, b'(t)s_a d_a. \tag{28}$$

Vanilla uses $(S, T) = (0, 0)$, state-only uses $(1, 0)$, tangent correction-only uses $(0, 1)$, and full transport-consistent TSS uses $(1, 1)$. Only the full arm among these non-vanilla controls supplies the exact tangent of its input path. The correction-only arm trains on a modified native target at the base state, isolating the target intervention in the latent comparison.

A.3.5 OPERATIONAL PROPERTIES

Proposition 2 (Exact transport consistency)**.** *If d_a and s_a are fixed along a sampled trajectory, then $u_t^\star = dz_t^\star/dt$. Where $b = b' = 0$, both state and velocity equal the base path.*

Corollary 1 (Valid marginal flow)**.** *Let $p_t^\star$ be the marginal law of $z_t^\star$. Assume $u_t^\star$ is integrable and differentiation under the expectation is valid for smooth compactly supported test functions. Then*

$$v^\star(z, t) = \mathbb{E}[u_t^\star \mid z_t^\star = z]$$

satisfies the continuity equation for $p_t^\star$. Because $b(0) = b(1) = 0$, the shaped and base paths have the same endpoint marginals.

Proposition 3 (Parameterization covariance)**.** *When the schedule and anchor systems above are nonsingular, they map one shaped state-velocity pair and one anchor prediction to unique native targets and an endpoint estimate. Optimization and empirical performance also depend on the loss and model parameterization.*

Proposition 4 (Bounded correction)**.** *The trust scale in Eq.* (20) *gives*

$$\sup_t \mathrm{RMS}(u_t^\star - u_t) \le \rho\ \mathrm{RMS}(u_a).$$

Proposition 5 (Interval-map identity)**.** *For $s < t$, the exact average velocity satisfies*

$$\overline{u}_{s,t}^\star = \overline{u}_{s,t} + \frac{b(t) - b(s)}{t - s} s_a d_a. \tag{29}$$

Moreover,

$$\mathrm{RMS}(\overline{u}_{s,t}^\star - \overline{u}_{s,t}) \le \|b'\|_\infty s_a\, \mathrm{RMS}(d_a) \le \rho\ \mathrm{RMS}(u_a).$$

If $b(0) = b(1) = 0$, the analytic conditional path's exact full-interval displacement and average-velocity target are unchanged.

The last identity defines interval-map targets for transport-consistent TSS. Intermediate finite differences change while the exact $[0, 1]$ displacement stays intact. The identity concerns the analytic conditional training path; learned-ODE length, curvature, solver error and downstream student quality require empirical evaluation.

A.3.6 GRID-NORMALIZED INFERENCE-FIELD SHAPING

For the linear noise-to-data path, the D-only direction satisfies $D_a = x - z_a = (1-a)u$. In this specialization, Eq. (20) reduces to $s_D = \min\{1, \rho/(\|b'\|_\infty(1-a))\}$. Its shaped-path tangent therefore contains the scalar factor

$$w(\tau) = 1 + (1-a)s_D\, b'(\tau), \tag{30}$$

Here $w(\tau) = \phi'(\tau)$ in Eq. (7), and τ denotes physical data time. We use this factor to define an inference-only companion for a frozen learned field. Let τ_i be its solver query nodes and let ω_i be the exact quadrature weight of each field call. We set

$$Z_K = \frac{\sum_i \omega_i w(\tau_i)}{\sum_i \omega_i}, \qquad m_i = \frac{w(\tau_i)}{Z_K}, \tag{31}$$

and multiply the final physical velocity by m_i before the solver update. For Euler, $\omega_i = \tau_{i+1} - \tau_i$; for JiT's mixed solver, the weights are those of its nine Heun intervals and final Euler interval. We call this discretely mass-normalized profile S_{norm}. Equation (31) makes the weighted multiplier mass exactly one, preserving the discrete endpoint of a constant field.

The multiplier is applied after conversion from the native prediction head to physical velocity and after CFG has formed the final field. It changes no state or model query time and adds no model evaluation. For a model using noise time q, such as the RAE pipeline below, we evaluate the same frozen profile at $\tau = 1 - q$. All experiments use $a = 0.5$, half-width 0.25, peak 0.05, and $\rho = 0.25$; only Z_K is recomputed from the declared solver grid. The model is still queried at τ_i; a full time reparameterization would also query the field at the mapped time. These operations differ for a general non-autonomous field. With a single quadrature interval, normalization gives $m_0 = 1$, leaving the one-step update unchanged.

Table 1 in the main text summarizes these intervention points and distinguishes the training constructions from the inference-only control and related flow-map approaches.

A.4 EXPERIMENTS

A.4.1 PROTOCOLS AND SOLVER ACCOUNTING

The JiT experiments use class-conditional ImageNet (Deng et al., 2009) and report FID (Heusel et al., 2017). The primary Gate-2B comparison continues one JiT-B/16 epoch-600 source to epoch 700 for 100 epochs with global batch size 1,024, base learning rate 5×10^{-5}, and matched training seeds $0, 1, 2$. Each final cell contains 50,000 generated images at generation seed 0, CFG 2.5, guidance interval $[0.1, 1]$, and EMA weights. Training uses AdamW (Loshchilov & Hutter, 2019); sampling uses classifier-free guidance (Ho & Salimans, 2022).

For JiT Heun, S solver steps comprise $S-1$ Heun intervals and one final Euler interval: $2S-1$ CFG-combined field evaluations and $2(2S-1)$ raw conditional/unconditional forwards. Heun-10 and Heun-50 therefore use 38 and 198 raw forwards. The DPM++2M comparisons (Lu et al., 2025) use $L = 9, 19, 49$ field evaluations, or 18/38/98 raw forwards. We label each table explicitly; identical nominal step counts across solvers are not assumed to mean equal compute.

LightningDiT's upstream argument specifies grid points: a K-point Euler grid has $K-1$ field evaluations. Our full-budget study continues one local LightningDiT-B/2 800k state, including optimizer and EMA, for 20k updates on the full 1,281,168-row latent cache with global batch 1,024. Three matched seeds compare vanilla, TSS-T, and full transport-consistent TSS under the original velocity MSE-plus-cosine loss. All nine training runs use two A100 GPUs each; all 60 evaluations use a common H100 environment, including six source-checkpoint references. Euler and DPM++2M grids 10/25/50 use 9/24/49 field evaluations, or 18/48/98 raw forwards with the frozen CFG20, first-three-channel convention. This is an internally controlled platform, not reproduction of an official B/2 headline. Each cell is FID50K. The earlier 625-update, batch16 factorial on 10,000 latents remains a separate component diagnostic. Paired Student-t intervals use three training seeds and two degrees of freedom; they are conditional on the shared source, data and evaluation noise, without multiplicity correction.

Table 5: Audited Gate-2B result on ImageNet 256×256. Values are mean $\pm$ sample SD across three matched training seeds. Each cell uses 50,000 images, generation seed 0, CFG 2.5, and EMA weights. Paired differences are method minus vanilla with 95% Student-t intervals in brackets. H10 and H50 denote Heun solver steps.

Budget	Arm	FID	Paired Δ [95% CI]
H10	Vanilla	8.2543 ± 0.0271	reference
	TSS-X (D+P+N)	$\mathbf{7.1112 \pm 0.1301}$	-1.1431 ± 0.1560 $[-1.5306, -0.7556]$
	D+N	7.2824 ± 0.1127	-0.9719 ± 0.1398 $[-1.3192, -0.6246]$
H50	Vanilla	$\mathbf{3.6851 \pm 0.0261}$	reference
	TSS-X (D+P+N)	3.7163 ± 0.0807	$+0.0312 \pm 0.0561$ $[-0.1082, +0.1705]$
	D+N	3.8556 ± 0.0374	$+0.1704 \pm 0.0559$ $[+0.0317, +0.3092]$

Table 6: JiT comparison at Heun NFE19/99 and DPM++2M NFE9/19/49. B/16 rows are three-seed means; L/H rows are single pairs. Each guided field evaluation counts as one NFE (two CFG branches). Δ = TSS-X minus vanilla. †: L/16 vanilla epoch700 versus TSS-X epoch699; these are not the later matched-epoch700 DPM++2M results.

Model	Solver	NFE	Vanilla	TSS-X	Δ
B/16	Heun	19	8.2543	7.1112	-1.1431
B/16	Heun	99	3.6851	3.7163	$+0.0312$
B/16	DPM++2M	9	10.2044	8.9102	-1.2942
B/16	DPM++2M	19	7.2249	6.1026	-1.1223
B/16	DPM++2M	49	3.7107	3.8152	$+0.1045$
L/16†	Heun	19	6.6247	5.4201	-1.2046
L/16†	Heun	99	2.4767	2.2450	-0.2317
H/16	Heun	19	6.3556	5.5335	-0.8221
H/16	Heun	99	1.9405	1.8988	-0.0417

A.4.2 AUDITED JIT-B/16 EFFICACY

Table 5 is the primary JiT-B/16 result. TSS-X improves Heun-10 in every matched training seed. Its mean paired change is -1.1431 FID, and the 95% interval excludes zero. At Heun-50, the mean paired change is $+0.0312$, a 0.84% arm-mean degradation, and the interval crosses zero. The result supports a strong low-step effect with a small high-step trade-off.

D+N also improves Heun-10, but its mean Heun-50 degradation is 4.63% and fails the predeclared replacement rule. TSS-X beats D+N at Heun-10 in all three seeds, has a 2.35% arm-mean advantage there, and has better mean Heun-50 FID. We therefore retain TSS-X and treat D+N as a mechanism comparator. This is also why the D+N anchor in transport-consistent TSS is a separate transport design choice rather than a retroactive redefinition of the JiT method.

The earlier single-pair scale study gives Heun-10 FID changes $8.42 \to 7.24$, $6.62 \to 5.42$, and $6.36 \to 5.53$ on JiT-B/16, L/16, and H/16. At Heun-5, the relative reductions are 12.6%, 16.0%, and 12.6%. Higher-step behavior varies by scale. The full solver-step table and the historical L/16 one-epoch endpoint offset are in Appendix C. A matched JiT-L/32 512×512 pair improves Heun-10 FID from 6.5844 to 5.6025 using 50,000 images, but it contains one training pair and one solver setting.

Table 7: DPM++2M on three matched JiT-B/16 training seeds. Means of FID 50K, EMA1, CFG 2.5; raw forwards include both CFG branches. All three seedwise differences have the displayed sign at each budget.

Raw forwards	Vanilla	TSS-X	Paired Δ
18	10.2044	8.9102	-1.2942
38	7.2249	6.1026	-1.1223
98	3.7107	3.8152	$+0.1045$

A.4.3 Solver robustness and inference additivity

The DPM++2M sweep evaluates the same three Gate-2B training-seed checkpoint pairs at three fixed call budgets (Table 7). All pairs improve at 18 and 38 raw forwards; all regress slightly at 98. The low-call benefit thus extends to a second solver family. For comparison, Heun-50 uses 198 forwards. This evaluation has its own common prior stream, so paired differences are computed within each evaluation table.

A separate Heun-10 four-cell comparison tests additivity over three training seeds. Mean FIDs are 8.1890 for vanilla, 7.6264 for vanilla plus S_{norm}, 7.1789 for TSS-X, and 6.7039 for TSS-X plus S_{norm}. All four cells use 38 raw forwards; the combined reduction is 18.14%. The incremental S_{norm}changes on TSS-X are $-0.4668, -0.6395, -0.3187$ across seeds, supporting additivity under this protocol. Causal independence would require a separate test. These generated samples follow the paired A100 evaluation protocol rather than the DPM++ stream; seed 0's original per-rank audit payload was not retained after scoring failures, although the complete images and fixed generation configuration were preserved for scoring recovery.

The common-stream seed 1 solver comparison gives vanilla/TSS-X FIDs of 8.3047/7.2664 with Heun, 7.6625/6.6443 with Heun plus S_{norm}, and 7.0375/6.0605 with DPM++2M, all at 38 forwards. Thus DPM++2M is better than Heun plus $S_{\mathrm{norm}}$for both checkpoints, while TSS-X retains a within-solver improvement.

A.4.4 Exact 10/20/50-NFE evaluation

Tables 2 and 10 evaluate frozen checkpoints at exactly 10, 20 and 50 guided field calls on the same RTX PRO 6000 Blackwell hardware. Main-text Table 3 shows the latent 10-NFE subset. Each cell contains 50,000 completed Inception feature vectors. JiT uses EMA1, batch64, per-image prior base43,000,000 and class ID modulo1000; CFG is 2.5 for B/L and 2.2 for H, with interval $[0.1, 1]$. LightningDiT uses EMA, batch32, prior base91,000,000, the frozen CFG20 first-three-channel convention, shift0.3 and its original VA-VAE pixel conversion. Each field call includes both CFG branches. The reference statistics and Inception implementation are unchanged.

JiT-B/16 uses three matched training-seed pairs at epoch700. The L/16 pair also ends at epoch700: its TSS-X model resumes 1,251 updates from the saved epoch699 state, as in the larger-teacher experiment. Both H/16 checkpoints store epoch699. The latent comparison reuses all nine 820k checkpoints from the 20k-update continuation. Tables 8 and 9 retain every paired reading. Heun NFE19/99 and the additional DPM++2M and latent grids are reported separately in Tables 6 and 12; their FIDs are not relabeled to the new grid or averaged across evaluation protocols.

The budget response differs by solver: both TSS-T and full transport-consistent TSS increase FID at 20 NFE; at 50 NFE they improve DPM++2M and increase Euler FID. Table 10 retains the complete mean results alongside the individual-seed readings.

At Euler NFE10, full minus TSS-T is $-0.2895, -0.2944, -0.4060$ across the three seeds, with mean -0.3300 and paired 95% interval $[-0.4936, -0.1663]$. DPM++2M NFE10 instead gives -0.0327 $[-0.2880, +0.2226]$. These new-grid readings show an EMA-guided increment at one operating point; the NFE9 comparison in Table 12 has a different result. The intervals describe three paired seeds and are not corrected for the multiple solver/budget comparisons.

Table 8: All 30 DPM++2M JiT FID50K evaluations underlying Table 2. Δ = TSS-X minus vanilla. The L/16 TSS-X checkpoint completes 1,251 continuation updates from epoch699 to match vanilla at epoch700; H checkpoints both store epoch699. Each row uses matched generation indices and evaluation settings.

Model	Training seed	NFE	Vanilla	TSS-X	Δ
B/16	0	10	9.5972	8.3887	-1.2085
B/16	1	10	9.5181	8.2033	-1.3148
B/16	2	10	9.7427	8.1989	-1.5438
B/16	0	20	7.1995	6.0809	-1.1186
B/16	1	20	6.9768	6.0140	-0.9629
B/16	2	20	7.0158	5.8909	-1.1250
B/16	0	50	3.7298	3.8053	$+0.0754$
B/16	1	50	3.6855	3.7653	$+0.0798$
B/16	2	50	3.7167	3.8824	$+0.1657$
L/16	single pair	10	7.9974	6.7743	-1.2231
L/16	single pair	20	5.7918	4.7803	-1.0115
L/16	single pair	50	2.7905	2.6460	-0.1444
H/16	single pair	10	7.6652	6.6524	-1.0128
H/16	single pair	20	4.9841	4.2417	-0.7424
H/16	single pair	50	1.9881	2.0466	$+0.0584$

Table 9: All 54 latent FID50K evaluations underlying Table 10, including the 10-NFE subset in Table 3. These are new evaluations of the same nine frozen 820k checkpoints, with the same per-ID noise and class labels across all cells. The training recipe and budgets are unchanged.

Training seed	Solver	NFE	Vanilla	TSS-T	Full
20260921	Euler	10	14.5089	13.8372	13.5477
20260922	Euler	10	14.7046	14.1312	13.8368
20260923	Euler	10	14.7239	13.9289	13.5229
20260921	Euler	20	7.1247	7.4137	7.2474
20260922	Euler	20	7.2088	7.4030	7.5278
20260923	Euler	20	7.1909	7.4030	7.2895
20260921	Euler	50	5.8911	6.0215	5.9760
20260922	Euler	50	5.9464	6.1355	6.4584
20260923	Euler	50	5.9351	6.0664	6.1818
20260921	DPM++2M	10	10.2216	9.5583	9.4223
20260922	DPM++2M	10	10.2618	9.7326	9.8021
20260923	DPM++2M	10	10.3771	9.6498	9.6182
20260921	DPM++2M	20	5.8671	6.2166	6.1452
20260922	DPM++2M	20	5.9362	6.3410	6.5129
20260923	DPM++2M	20	5.9346	6.3294	6.2627
20260921	DPM++2M	50	6.6600	6.5733	6.2997
20260922	DPM++2M	50	6.7064	6.5736	6.9792
20260923	DPM++2M	50	6.7008	6.5317	6.6089

A.4.5 What the JiT warp contains

The single-seed factorial screen in Table 11 uses a frozen epoch-625 endpoint and 10,000 images per cell. D, P, N, and P+N fail the predeclared gate. D+P and D+N pass only when a residual is combined with the clean-path shift. A norm-matched Gaussian perturbation does not reproduce the Heun-10 gain. The screen shows a D-plus-residual pattern at this checkpoint; it does not estimate a replicated factorial interaction or isolate a geometric cause.

A separate seed-0, 50,000-image evaluation gives H10/H50 FID $8.5258/3.7530$ for vanilla, $7.4124/3.6392$ for TSS-X, and $7.4199/3.6809$ for D+N. D+N nearly matches TSS-X in this one

Table 10: LightningDiT-B/2 EMA FID50K at NFE 10/20/50 (11/21/51 grid points), with the full-data 20k-update checkpoints from three matched seeds. Frozen CFG20; common Blackwell evaluation. NFE follows Table 2. TSS-T applies EMA-free training-time reparameterization.

Solver	NFE	Vanilla	TSS-T	Full transport-consistent TSS
Euler	10	14.6458	13.9658	13.6358
Euler	20	7.1748	7.4065	7.3549
Euler	50	5.9242	6.0745	6.2054
DPM++2M	10	10.2868	9.6469	9.6142
DPM++2M	20	5.9126	6.2957	6.3069
DPM++2M	50	6.6891	6.5595	6.6293

Table 11: JiT-B/16 D/P/N screen at one training seed. Every cell uses 10,000 images, CFG 2.5, generation seed 0, and the same Heun protocol. The gate requires at least 5% relative H10 improvement and at most 3% H50 degradation against vanilla.

Warp factors	H10 FID	H50 FID	Gate
None (vanilla)	10.9201	6.3353	control
D	10.5267	6.6972	fail
P	11.0941	6.3560	fail
N	10.6680	6.3718	fail
D+P	10.2626	6.4269	pass
D+N	9.8837	6.2332	pass
P+N	10.5737	6.3423	fail
D+P+N (TSS-X)	**9.8697**	**6.1774**	pass
Gaussian, norm matched	10.8062	6.1874	fail

seed, but the replicated result in Table 5 reverses the replacement decision. The evidence supports keeping P in TSS-X under the JiT protocol.

A.4.6 Latent transfer: budget and time-only controls

Table 12 reports the full-data 20k-update experiment. Both TSS-T and full transport-consistent TSS improve low-call FID relative to matched vanilla. For Euler grid 10, full minus vanilla is -1.6527 with paired 95% interval $[-2.3843, -0.9210]$, while time-only minus vanilla is -1.6164 $[-1.9917, -1.2412]$. The additional full-minus-time-only change is only -0.0363 $[-1.1330, +1.0605]$ and has mixed signs. At the same call budget with DPM++2M, full minus time-only is $+0.3039$ $[-0.5444, +1.1522]$, positive in all three pairs. Neither comparison establishes an incremental full-method benefit.

At higher call budgets, full minus vanilla is positive in all three pairs for Euler grids 25/50 and DPM++2M grid 25; DPM++2M grid 50 has mixed signs and a mean difference of $+0.0169$. The predeclared high-call investment tolerance was not met. TSS-T accounts for the low-call gain on this grid; the exact-NFE10 comparison in Appendix A.4.4 additionally finds an EMA-guided increment under Euler. The earlier negative 20k-update endpoint-consistent TSS-X transfer is reported separately because it uses a different recipe.

The earlier endpoint-consistent TSS-X pilot used the target $(x_1 - \widetilde{z}_t)/(1-t)$ in LightningDiT's noise-to-data time convention, with 20k updates and global batch 1,024. Euler grid-10/50 FID changes were approximately $+1.70/+0.73$. This tests the target-level conversion under a native velocity head, MSE-plus-cosine metric, denominator floor and warp profile that jointly differ from the JiT recipe. The source of the negative transfer remains unresolved.

The earlier short-budget factorial in Table 13 answers a different component question. State-only is worse than vanilla in all three seeds, while tangent-only and full transport-consistent TSS improve mean FID by 0.3578 and 0.2320. Full minus tangent-only is $+0.1258$ with interval $[+0.0499, +0.2017]$, and the factorial interaction is -0.0642 $[-0.1347, +0.0062]$. Under that

Table 12: LightningDiT-B/2 continuation at NFE9/24/49: three matched training seeds, 20k updates, batch 1,024, EMA FID 50K, frozen CFG 20. TSS-T denotes the training transform in Eq. (27). Grid sizes 10/25/50 give exactly 9/24/49 NFE; each NFE uses two CFG branches. Table 10 separately reports NFE10/20/50 on Blackwell. Δ = full transport-consistent TSS minus vanilla.

Solver	NFE	Vanilla	TSS-T	Full transport-consistent TSS	Δ
Euler	9	18.3826	16.7662	16.7299	-1.6527
Euler	24	6.9218	7.1761	7.2240	$+0.3022$
Euler	49	5.8932	6.1313	6.1992	$+0.3060$
DPM++2M	9	12.4940	11.2987	11.6026	-0.8914
DPM++2M	24	6.0566	6.4174	6.4721	$+0.4154$
DPM++2M	49	6.5987	6.6217	6.6155	$+0.0169$

Table 13: Earlier 625-update, batch16 all-A100 component factorial on LightningDiT-B/2, distinct from the 20k-update experiment. A 10-point Euler grid uses 9 CFG-combined evaluations. Each cell uses 50,000 images, CFG20, and EMA weights. Differences are paired over three training seeds; brackets give 95% Student-t intervals.

Arm (S, T)	Mean FID	Paired Δ FID	95% interval
Vanilla $(0, 0)$	18.8567	reference	–
State-only $(1, 0)$	19.0467	$+0.1901$	$[+0.1491, +0.2310]$
Tangent-only $(0, 1)$	$\mathbf{18.4989}$	$\mathbf{-0.3578}$	$[-0.4865, -0.2291]$
Full transport-consistent TSS $(1, 1)$	18.6247	-0.2320	$[-0.4110, -0.0529]$

specific short continuation, the target intervention is more useful than the state-only intervention; this does not establish a universal mechanism or substitute for the time-only/full comparison at larger budget. Full transport-consistent TSS remains the only non-vanilla arm of this factorial that supplies an exact shaped-path tangent.

A secondary sweep under the earlier short-budget protocol gives mean paired changes -0.0718, -0.1076, and -0.0881 on 5/20/50-point Euler grids, with only 10,000 images per cell. It must not be pooled with the new FID50K full-budget curve or used to remove its higher-call regressions.

A.4.7 Zero-training transfer of the inference companion

Table 14 tests one frozen S_{norm}profile without retraining any evaluated checkpoint. Each row is a within-checkpoint paired comparison with matched sampling inputs and the same base solver. The JiT row applies the multiplier after clean-image prediction has been converted to physical velocity. The two latent rows apply it directly to a native-velocity field. Thus the table crosses both state representation and exposed prediction head, while isolating only the inference component.

The LightningDiT comparison uses our locally trained seed-0 EMA checkpoint, CFG 20, and the same VSFM-inspired composite clock in both arms. This measures the incremental effect within our local LightningDiT configuration. Its IS changes from 169.9760 ± 3.5903 to 179.6897 ± 3.5943. On three continuations sharing the same upstream initializer, independently generated 10,000-image comparisons give incremental FID gains of 1.2049, 1.2090, and 1.1780. These are descriptive checks across continuations of the shared pretrained model.

The RAE row uses the released DiT^{DH}-XL checkpoint (Zheng et al., 2026), CFG 1, FP32, equal labels, and the official shifted grid ending at noise time $q = 0.001$. We transfer the profile through $\tau = 1 - q$ and perform no tuning. IS improves from 226.0448 to 228.6043, sFID from 8.8614 to 8.5773, and precision from 0.76234 to 0.76424; recall decreases from 0.6219 to 0.6163. The paired RAE result therefore shows a quality–coverage trade-off.

Table 14: Paired 50,000-image transfer of S_{norm} on ImageNet 256×256. "Calls" counts final field evaluations; the JiT calls are CFG-combined and each invokes conditional and unconditional branches. No row adds a model evaluation. Lower FID is better.

Backbone	Calls	Base FID	$+S_{\mathrm{norm}}$	Improvement
JiT-B/16	19	8.2373	7.6251	0.6121 (7.43%)
Pixel / clean-x; mixed Heun, 10 intervals				
LightningDiT-B/2	9	11.9067	10.7208	1.1859 (9.96%)
Latent / velocity; composite clock, 10 points				
RAE DiT^{DH}-XL	9	3.2887	3.1335	0.1552 (4.72%)
Latent / velocity; official Euler, 10 points				

Table 15: Reflow consumer, final EMA FID 50K after 20k updates. Rollout budgets count teacher field evaluations; the second column counts student calls. $\mathrm{Vanilla}_{14}$ is the endpoint-FID-matched control selected without student feedback. $\Delta = \mathrm{TSS\text{-}X}_{10} - \mathrm{Vanilla}_{10}$. Each student stream is paired across all columns.

Stream	Student calls	$\mathrm{Vanilla}_{10}$	$\mathrm{Vanilla}_{14}$	$\mathrm{TSS\text{-}X}_{10}$	$\mathrm{Vanilla}_{19}$	Δ
1	1	31.6854	32.7024	29.7561	32.3649	-1.9293
2	1	31.4385	32.1380	29.6124	31.8548	-1.8261
1	15	14.6269	14.7629	13.5930	13.2008	-1.0339
2	15	14.6910	14.7446	13.5491	13.2715	-1.1419

A.4.8 Synthetic-data transfer and consumer dependence

We generate 200k class-balanced, raw floating-point noise–endpoint pairs per teacher from matched Gate-2B seed 1 epoch 700 EMA1 checkpoints. $\mathrm{Vanilla}_{10}$/$\mathrm{TSS\text{-}X}_{10}$ use 10 DPM++2M field evaluations each; $\mathrm{Vanilla}_{19}$ uses 19, with teacher CFG 2.5. Reflow students share a vanilla initialization and train for 20k updates, batch 256, learning rate 5×10^{-5} and EMA 0.999. Students sample without further guidance (implementation CFG 1; teacher CFG 2.5 is already in the pairs). One-call and eight-interval Heun sampling use 1 and 15 raw forwards. Two matched student training streams vary training randomness while reusing the teachers, endpoint pools and evaluation noise (Table 15).

$\mathrm{TSS\text{-}X}_{10}$ improves both sampling settings relative to $\mathrm{Vanilla}_{10}$ in both streams, including Online weights and the 5k/10k milestones. Mean final EMA one-call FID changes from 31.5620 to 29.6843, and 15-call FID from 14.6589 to 13.5710. Against $\mathrm{Vanilla}_{19}$, $\mathrm{TSS\text{-}X}_{10}$ has better one-call FID but worse 15-call FID: EMA differences are $+0.3922/+0.2775$, and Online differences are $+0.5397/+0.6983$. Checkpoints and weight choices provide correlated readings within each stream.

Endpoint FID is insufficient to rank downstream value. We preselected a vanilla rollout budget from $L \in \{13, 14\}$ using the first 50k training endpoints, requiring FID within 0.3 of $\mathrm{TSS\text{-}X}_{10}$ and choosing the closest candidate. L14 gives 8.1094 versus $\mathrm{TSS\text{-}X}_{10}$'s 8.2179 and is selected without student-result feedback. Completing $\mathrm{Vanilla}_{14}$ in both existing student streams gives $\mathrm{TSS\text{-}X}_{10} - \mathrm{Vanilla}_{14}$ final EMA differences of $-2.9463/-2.5256$ at one call and $-1.1699/-1.1955$ at 15 calls. Corresponding Online differences are $-2.6156/-2.1133$ and $-0.8312/-0.7151$. All 24 correlated milestone/weight/budget comparisons favor $\mathrm{TSS\text{-}X}_{10}$. The two streams share teachers, pools and evaluation noise. The scalar FID match leaves distributional differences and the coupling mechanism unresolved. On a fixed 1,000-pair training subset, $\mathrm{TSS\text{-}X}_{10}$'s one-call endpoint MSE is slightly higher despite its lower FID: regression fit and distribution quality rank the sources differently.

Direct one-step fine-tuning does not retain the advantage. Starting from the second-stream reflow EMA checkpoints, both arms receive the same fixed 10k-update one-step endpoint-regression recipe: 5k raw-RGB MSE updates followed by 5k LPIPS updates against their original JiT teacher pools. The final one-call EMA FIDs are 15.2721/15.3754 for $\mathrm{Vanilla}_{10}$/$\mathrm{TSS\text{-}X}_{10}$, changing the paired difference from -1.8261 to $+0.1033$. The difference was already small at 2.5k during MSE training. Both arms substantially reduce their one-call FID and end at similar quality. Three-forward Heun

FID degrades from 17.7575/16.2613 to 22.5178/33.6833. The TSS advantage thus depends on the consumer stage, and its disappearance begins before the switch to LPIPS.

A second consumer: a flow-map student trained via JVP. Separately, we train JiT-B/16 students on the original Vanilla$_{10}$/TSS-X$_{10}$ endpoint pools using a pMF-identity objective, a common vanilla initialization, and 20k updates per arm in each of two matched student training streams. We preserve each raw floating-point endpoint and its original generating noise, with augmentation disabled. Both sources contain teacher CFG 2.5; the students sample without further guidance (implementation CFG 1). The JiT-specific objective is defined in Section 5.4. Final one-call EMA FID improves from 24.5294 to 21.4469 in the first stream and from 23.6579 to 20.7923 in the second ($-3.0825/-2.8656$); two-call EMA differences are $-2.0938/-1.9770$. Final Online one-call differences are $-2.3650/-1.7146$, and two-call differences are $-1.8407/-0.9511$. All 24 correlated milestone/weight/call comparisons favor TSS-X$_{10}$. Both one-call EMA gaps shrink during training. The streams share teachers, pools, full initialization and evaluation noise, changing only matched training randomness. They confirm the data-source benefit for this JVP consumer at the tested budget. Appendix F.1 retains both full trajectories and the common zero-update baseline.

The JVP benefit narrows with extended training. We continue all four 20k states to 100k with their Online weights, EMA, Adam and random streams intact, retaining the pools, objective, batch size and constant learning rate. All extension evaluations use a common H100 environment, with 16 repeated parent-20k cells to separate environment changes from training. At 100k, one-call EMA FID is $22.0088/21.9359$ for Vanilla$_{10}$/TSS-X$_{10}$ in stream 1 and $22.0153/21.6280$ in stream 2: differences of $-0.0729/-0.3873$, both smaller than the predeclared 0.5 descriptive magnitude. Final two-call EMA differences reverse to $+0.5604/+0.1510$; one-call Online signs disagree. Relative to their common-environment 20k references, vanilla improves by $2.5216/1.6501$, while TSS-X$_{10}$ deteriorates by $0.4870/0.8351$. Both changes contribute to the narrowing gap. Appendix F.1 retains all trajectories, costs and environment references; Appendix A.4.9 analyzes the updates needed to reach target FID.

Fixed-pair adaptation degrades the mature official network. Using the full released pMF-B/16 generation network and common mature weights gives a one-call baseline FID of 3.1543. A separate matched 2,000-update fixed-pair adaptation with an external boundary-u JVP objective degrades terminal EMA FID to 22.7820/24.7594 for Vanilla$_{10}$/TSS-X$_{10}$; TSS-X$_{10}$ is worse by 1.9774, and terminal Online results also favor Vanilla$_{10}$. This adaptation uses a different objective and initialization from both official full training and the JiT-student experiment. Its cause of failure remains unresolved. Appendix I.1 reports every checkpoint and both weight choices.

A.4.9 Data-efficiency definitions, curves and source readings

Scope and reference. This analysis measures updates to target FID on fixed 200k-pair pools, one per data source. The sources share pool size and noise IDs; each teacher supplies its own endpoints. Within each consumer and student training stream, Vanilla$_{10}$ and TSS-X$_{10}$ use the same initialization, optimizer, batch size, update budget, sampling protocol and evaluation noise. Both use batch 256 and learning rate 5×10^{-5}; reflow uses EMA 0.999, and the flow-map consumer uses its frozen JVP recipe. We analyze each consumer under its own fixed JiT training protocol. The students measure the downstream value of TSS-generated data; their objective and architecture are defined in Section 5.4.

For consumer c, stream s and student forward budget n, write $F_V(u)$ and $F_M(u)$ for the measured EMA FID50K at update u. For the single-reference table, define

$$q = F_V(U_c) + 0.5, \qquad U_c = \begin{cases} 20{,}000, & \text{reflow or the seed-0 teacher pair,} \\ 100{,}000, & \text{the original flow-map pair.} \end{cases} \tag{32}$$

The first observed attainment is $u_a^{\mathrm{obs}} = \min\{u \in \mathcal{G}_c : F_a(u) \le q\}$ for arm a. For the first adjacent pair u_i, u_{i+1} with $F_a(u_i) > q \ge F_a(u_{i+1})$, the interpolated crossing is

$$\widehat{u}_a = u_i + \frac{F_a(u_i) - q}{F_a(u_i) - F_a(u_{i+1})}(u_{i+1} - u_i), \qquad R = \widehat{u}_V / \widehat{u}_M. \tag{33}$$

The reflow and seed-0 grids are $\{5, 10, 20\}$k; the original flow-map grid is $\{5, 10, 20, 40, 60, 80, 100\}$k. Interpolated crossings are bracketed by measured checkpoints

and estimate attainment between evaluations. The same ratio applies to pair-access counts because batch size is fixed. Distinct-pair counts, wall time and total pipeline cost require separate accounting.

Multiple targets and the 5k boundary. For an arbitrary target q, use the same first-crossing rule rather than requiring q to equal a final FID plus 0.5. If the first measured FID is already at or below q, the crossing is left-censored: we know only $u_a(q) \leq 5{,}000$. We report this bound because the analysis begins at the first measured checkpoint. If only TSS is censored, the displayed ratio is a lower bound $R(q) \geq \widehat{u}_V(q)/5{,}000$. An unattained target has no interpolated update or finite ratio.

The secondary target-FID curves in Figure 5 restrict each panel to targets bracketed for both sources: from the smaller of their first measured FIDs down to the larger of their best measured FIDs. The 5k boundary is marked separately. We interpolate linearly in raw updates and FID, then plot updates on a log axis; interpolating log updates would produce different values. With nonmonotone training curves, we take the first downward crossing of the unsmoothed curve. The update ratio need not increase monotonically as the target FID decreases. Tables 18 and 19 give the explicit target grid, including left-censored readings omitted from the bracketed curves. These targets are descriptive selections from the completed runs.

Figure 2 instead plots the one-call learning curves, retaining each stream separately through its final checkpoint. The dotted reference and arrow endpoints average the two original streams' reference FIDs and first-crossing updates, respectively. The labels average their individual update ratios: 2.3081 for flow-map and 1.2111 for reflow, rounded to 2.3 and 1.2. The seed-0 teacher pair is shown separately and does not enter those averages. Lines interpolate in raw updates and FID; markers identify measured checkpoints.

Interpretation limits. The interpolated ratios describe completed curves retrospectively. The prior budget extension specified a 0.5 exploratory endpoint-reference reading; we reuse that margin for descriptive interpolation across consumers. Statistical significance and equivalence were not tested in this analysis. First attainment can be followed by regression. Vanilla's final FID can exceed its best FID: its stream-1 one-call FID is 21.7697 at 60k versus 22.0088 at 100k. The TSS-X_{10} one-call curves worsen after 40k. Table 17 shows sensitivity to using the earlier 20k reference instead.

At 100k, each pair receives 128 accesses on average under sampling with replacement. The cause of the later regressions remains unresolved. Teacher post-training, rollout generation, student training and evaluation remain separate costs. The two original streams share teachers, pools, initial weights and evaluation noise. The third flow-map stream has another teacher pair and new endpoints, with the dependencies described in Appendix A.4.10; it has no 100k extension.

Environment and call accounting. Reflow curves use the original matched Blackwell environment. For flow-map students, 5k/10k use the parent H100 NVL evaluations; 20k uses the repeated H100 bridge evaluation, and later cells use that same extension environment. The maximum shift across the 16 repeated parent-20k cells is 0.00751, measuring the environment change under the fixed evaluation protocol. The original NVL 20k results remain in Table 34. The source reader historically labels reflow's eight Heun intervals as `calls=8`; all tables and plots in this analysis instead report the actual 15 student forwards. Flow-map students sample without further guidance (implementation CFG 1; teacher CFG 2.5 is embedded in the pairs), so one-/two-call sampling uses one/two network forwards.

Source and reproducibility. The research reader `data_efficiency_20260924.py` directly reads each run's `FID.json`. Its output is preserved as `data/data_efficiency_20260924.json`. The renderer `scripts/build_data_efficiency.py` builds single-reference tables and full training curves; `scripts/build_target_fid.py` produces the two-panel learning curves, secondary target-FID curves and multi-target tables from the same 14-row snapshot, including the seed-0 rows. The first renderer can refresh the portable snapshot via `-snapshot`; the second reads that snapshot directly. For duplicated 20k flow-map cells, the extension's H100 bridge values take precedence. Historical `mesh` keys and pool identifiers remain provenance identifiers for TSS-X. Tables 20 and 21 list all EMA cells used here; the subsequent consumer tables and Appendix F retain Online results, other controls, full budget curves and unfavorable results.

Table 16: Data efficiency on fixed 200k pools, with batch 256. Each Vanilla/TSS entry gives the first measured / linearly interpolated updates (thousands) to reach q, the matched vanilla final EMA FID plus 0.5. Ratio is Vanilla/TSS interpolated updates. GPU time is accounted for separately. Final budgets are 20k for reflow and 100k for flow-map students. Forwards are student network calls without further guidance (teacher CFG 2.5 in the pairs); 15 corresponds to eight Heun intervals.

Consumer	Stream	Forwards	q	Vanilla	TSS	Ratio
Flow-map	1	1	22.51	40 / 39.7	20 / 16.7	2.38
Flow-map	1	2	18.84	40 / 38.4	10 / 9.8	3.91
Flow-map	2	1	22.52	40 / 36.2	20 / 16.2	2.23
Flow-map	2	2	18.92	40 / 31.8	20 / 10.8	2.94
Reflow	1	1	32.19	20 / 19.1	20 / 15.5	1.23
Reflow	1	15	15.13	20 / 13.7	10 / 7.6	1.80
Reflow	2	1	31.94	20 / 19.2	20 / 16.1	1.19
Reflow	2	15	15.19	20 / 11.9	10 / 7.1	1.69

Table 17: Reference-budget sensitivity: use vanilla's 20k endpoint plus 0.5 instead of its 100k endpoint for the flow-map consumer. Entries and ratios use the same definitions as Table 16; crossings use only the 5k/10k/20k milestones. The seed0 rows use the second teacher pair, which has no 100k continuation. This is a descriptive sensitivity check.

Consumer	Stream	Forwards	q	Vanilla	TSS	Ratio
Flow-map	1	1	25.03	20 / 19.0	10 / 9.7	1.96
Flow-map	1	2	20.13	20 / 17.3	10 / 8.1	2.14
Flow-map	2	1	24.17	20 / 19.3	20 / 12.6	1.53
Flow-map	2	2	19.83	20 / 18.2	10 / 9.0	2.03
Flow-map	seed0	1	25.50	20 / 19.2	20 / 10.1	1.90
Flow-map	seed0	2	20.26	20 / 17.8	10 / 8.6	2.07

A.4.10 A SECOND TEACHER TRAINING SEED

We regenerate both 200k pools from the Gate-2B seed-0 vanilla/TSS-X teachers, both at epoch 700, with the same pinned DPM++2M ten-field-call rollout and teacher CFG 2.5. Their endpoint FID50K is 9.5292/8.3587. Both students start from identical seed-0 vanilla EMA1 weights and identically initialized interval parameters, with fresh AdamW and EMA. Each receives 20k updates under the same fixed-pair JVP recipe, batch 256 and microbatch 16. Training seed and new-parameter seed are 20260929; sample-ID/time bases are 101000000/102000000. The pair-generation prior base 32000000 and evaluation base 84000000 are shared with the original teacher pair. The pools share endpoint noise across teacher pairs, while the common student initialization changes with the vanilla teacher. The comparison therefore changes both the teachers and student starting weights.

Both new arms use four H100 ranks and the same software stack (PyTorch 2.7.1+cu128), rather than the original single-H100-NVL execution. The paired comparisons are hardware-matched within this replication; their absolute FIDs are reported separately by environment. All 26 student evaluations (24 milestones and two shared zero-update cells) and both pool evaluations contain 50,000 actual Inception features. Full checkpoints preserve 186 Adam states, four-rank RNG and continuous effective updates 1–20,000. Both arms have zero nonfinite/rejected updates and zero spikes in 19,900 eligible updates; these single-stream telemetry readings describe successful completion.

The final EMA one-call difference of -2.6120 passes the frozen -0.5 descriptive selection rule. All twelve paired student differences are negative, but the final Online two-call difference is only -0.2687. One-call EMA gaps narrow from -10.9554 to -5.6298 to -2.6120. The additional pair supports the data-efficiency result at a second teacher seed; updates beyond 20k were not measured. The reference in Table 23 uses this pair's 20k endpoint; the original pair has a separate 100k reference. At 10k, TSS-X_{10}'s one-call FID 25.5279 is slightly above 25.5034, so its first observed crossing is 20k.

Table 18: Updates to multiple target EMA FIDs (thousands), one student forward. Crossings use linear interpolation in raw updates and FID. ≤ 5 means attained by the first measured checkpoint; its exact crossing is unknown, so we report a lower bound ($\geq$) on the ratio. Targets were selected retrospectively to describe the measured curves.

Consumer / stream	Target FID	Vanilla	TSS	V/TSS
Flow-map: stream 1	35	7.64	≤ 5.00	≥ 1.53
	30	9.88	5.84	1.69
	27.5	14.29	7.78	1.84
	25	19.10	9.72	1.97
	24	25.18	11.99	2.10
	23	34.94	15.13	2.31
	22.5	39.83	16.70	2.38
Flow-map: stream 2	35	7.83	≤ 5.00	≥ 1.57
	30	10.91	6.25	1.75
	27.5	14.50	8.26	1.76
	25	18.08	10.73	1.69
	24	19.52	12.93	1.51
	23	29.36	15.13	1.94
	22.5	36.39	16.24	2.24
Flow-map: seed0 pair	35	8.54	≤ 5.00	≥ 1.71
	30	11.88	7.15	1.66
	27.5	15.94	8.74	1.82
	26	18.38	9.70	1.90
	25.5	19.19	10.09	1.90
Reflow: stream 1	42	7.60	6.51	1.17
	38	9.60	8.57	1.12
	35	13.98	10.38	1.35
	33	17.61	14.05	1.25
	32.2	19.07	15.52	1.23
Reflow: stream 2	42	7.75	6.79	1.14
	38	9.76	8.77	1.11
	35	14.16	10.89	1.30
	33	17.44	14.27	1.22
	32	19.08	15.96	1.20

The two jobs allocate 36.5826 H100 GPU-hours, including rollout generation, student training, evaluation, startup and IO; logged training accounts for 32.1180 of those hours. No GPU recovery was needed. Historical teacher post-training remains a separate reused cost in a full-pipeline comparison. The reviewed original FID readings are preserved in `data/teacher_seed0_20260924.json`; `scripts/build_teacher_replication.py` generates the following tables.

A.4.11 Fewer pairs, coupling geometry, and a decayed endpoint on larger pools

Protocol. All runs reuse the flow-map/JVP student recipe of Appendix A.4.9 with the common vanilla seed-1 epoch-700 EMA initialization, batch 256, microbatch 16 and four H100 ranks on one node. Pool-prefix students train for 20k updates at constant learning rate 5×10^{-5} on the first 50k or 100k pairs of the seed-1 200k pools (prior indices unchanged, float32 endpoints), so the 50k and 100k pools are accessed about 100 and 50 times on average. The decayed-endpoint students train for 60k updates with cosine decay to zero after a 200-update warmup and EMA 0.9999 on fresh pools of 640k pairs per teacher (prior seed base 33000000, DPM++2M with 10 field calls at CFG 2.5, endpoints stored in float16), which are accessed about 24 times. Milestones are 5k/10k/20k and 10k–60k respectively; the 60k endpoint adds 5- and 10-forward flow-map sampling. Zero-update cells equal those of the original flow-map stream because the initialization is identical. The historical 200k rows are the original stream's 20k cells and were trained with a different random stream and

Table 19: Updates to multiple target EMA FIDs (thousands), two student forwards. Crossings use linear interpolation in raw updates and FID. ≤ 5 means attained by the first measured checkpoint; its exact crossing is unknown, so we report a lower bound ($\geq$) on the ratio. Targets were selected retrospectively to describe the measured curves.

Consumer / stream	Target FID	Vanilla	TSS	V/TSS
Flow-map: stream 1	24	7.60	≤ 5.00	≥ 1.52
	22	9.53	5.67	1.68
	20	18.04	8.28	2.18
	19.5	23.04	8.94	2.58
	19	34.64	9.59	3.61
	18.8	39.28	9.85	3.99
Flow-map: stream 2	24	8.08	≤ 5.00	≥ 1.62
	22	10.52	6.13	1.72
	20	17.62	8.77	2.01
	19.5	19.40	9.43	2.06
	19	29.54	10.38	2.85

Table 20: Reflow complete EMA FID50K cells used in the data-efficiency analysis. $\Delta = \text{TSS-X}_{10} - \text{Vanilla}_{10}$. Forwards are student network calls.

Stream	Forwards	Updates	Vanilla_{10}	TSS-X_{10}	Δ
1	1	5k	47.2022	44.9434	-2.2589
1	1	10k	37.1903	35.2062	-1.9841
1	1	20k	31.6854	29.7561	-1.9293
1	15	5k	16.7966	15.8089	-0.9877
1	15	10k	15.4170	14.4887	-0.9283
1	15	20k	14.6269	13.5930	-1.0339
2	1	5k	47.4442	45.6205	-1.8237
2	1	10k	37.5341	35.5263	-2.0078
2	1	20k	31.4385	29.6124	-1.8261
2	15	5k	16.6121	15.8067	-0.8054
2	15	10k	15.3099	14.3100	-0.9999
2	15	20k	14.6910	13.5491	-1.1419

software stack, so the pool-size comparison across 200k is descriptive. The reviewed cells, training records and pool files are listed in the pool-efficiency review file; no FID was recomputed for this appendix.

Precision, recall and transport cost. kNN precision and recall use $k = 3$ on Inception-2048 features of the first 50k endpoints against 50k real ImageNet training images chosen at evenly spaced indices of the class-sorted manifest with the same center crop. The original definition uses VGG16 features, so we interpret these values within our feature protocol. The transport cost is $\mathbb{E}\|x_1 - x_0\|^2$ per dimension between an endpoint and the noise that generated it, and $\cos(x_1, x_0)$ the mean cosine over the same 8,192 pairs; both are computed from the stored endpoints and regenerated priors without any network call. Lower transport cost and higher alignment are the quantities minibatch-OT couplings target. Here they provide correlational diagnostics for the coupling hypothesis.

Readings. Table 24 shows that TSS-X prefixes of 50k pairs give lower FID than the full 200k vanilla pool at every milestone and both call counts, and that the 100k TSS-X prefix is within 0.2 FID of the full TSS-X pool. Table 25 shows that on the fresh 640k pools with learning-rate decay the TSS-X source is better in all eight endpoint cells (EMA one-call -0.65, Online one-call -0.98), whereas on the reused 200k pools the 100k gaps had closed (Appendix A.4.9); the two settings differ in pool size, precision, learning-rate schedule, EMA decay and budget, so the 640k run is a

Table 21: Flow-map complete EMA FID50K cells used in the data-efficiency analysis. $\Delta = \text{TSS-X}_{10} - \text{Vanilla}_{10}$. Forwards are student network calls.

Stream	Forwards	Updates	Vanilla_{10}	TSS-X_{10}	Δ
1	1	5k	40.9031	31.0885	-9.8146
1	1	10k	29.7279	24.6342	-5.0937
1	1	20k	24.5305	21.4489	-3.0816
1	1	40k	22.4821	20.8754	-1.6067
1	1	60k	21.7697	21.1086	-0.6612
1	1	80k	22.1402	21.6960	-0.4442
1	1	100k	22.0088	21.9359	-0.0729
1	2	5k	26.7008	22.5108	-4.1900
1	2	10k	21.5147	18.6882	-2.8264
1	2	20k	19.6308	17.5326	-2.0982
1	2	40k	18.7691	17.8739	-0.8952
1	2	60k	18.3043	18.1395	-0.1648
1	2	80k	18.3842	18.7221	$+0.3379$
1	2	100k	18.3395	18.9000	$+0.5604$
2	1	5k	40.7126	31.5556	-9.1571
2	1	10k	30.6317	25.3294	-5.3023
2	1	20k	23.6654	20.7929	-2.8725
2	1	40k	22.2435	20.4969	-1.7465
2	1	60k	22.1004	20.5278	-1.5726
2	1	80k	22.4446	21.5569	-0.8877
2	1	100k	22.0153	21.6280	-0.3873
2	2	5k	26.9817	22.8535	-4.1281
2	2	10k	22.1474	19.0646	-3.0828
2	2	20k	19.3305	17.3492	-1.9813
2	2	40k	18.6376	17.6845	-0.9531
2	2	60k	18.4154	17.6329	-0.7826
2	2	80k	18.5590	18.5308	-0.0282
2	2	100k	18.4220	18.5731	$+0.1510$

Table 22: Second teacher training seed (seed 0), all matched student FID50K readings. $\Delta = \text{TSS-X}_{10} - \text{Vanilla}_{10}$. Two additional shared zero-update evaluations give one/two-call FID 299.6919/157.0682; the two initial models and their EMA weights are identical.

Updates	Weights	Calls	Vanilla_{10}	TSS-X_{10}	Δ
5k	Online	1	36.6234	29.3322	-7.2911
5k	Online	2	24.6735	21.4497	-3.2238
5k	Ema	1	44.3198	33.3644	-10.9554
5k	Ema	2	28.0254	23.3886	-4.6368
10k	Online	1	31.2684	25.5358	-5.7326
10k	Online	2	22.3692	19.4472	-2.9220
10k	Ema	1	31.1577	25.5279	-5.6298
10k	Ema	2	22.0685	19.0711	-2.9974
20k	Online	1	23.7495	22.6131	-1.1364
20k	Online	2	18.9695	18.7008	-0.2687
20k	Ema	1	25.0034	22.3914	-2.6120
20k	Ema	2	19.7573	18.1181	-1.6392

new comparison rather than a pool-only ablation. Table 26 shows that TSS-X pools have slightly higher precision and recall in both teacher pairs and in the 640k pools, that the Vanilla_{14} control has lower endpoint FID and higher recall than TSS-X_{10} while training worse students, and that TSS-X

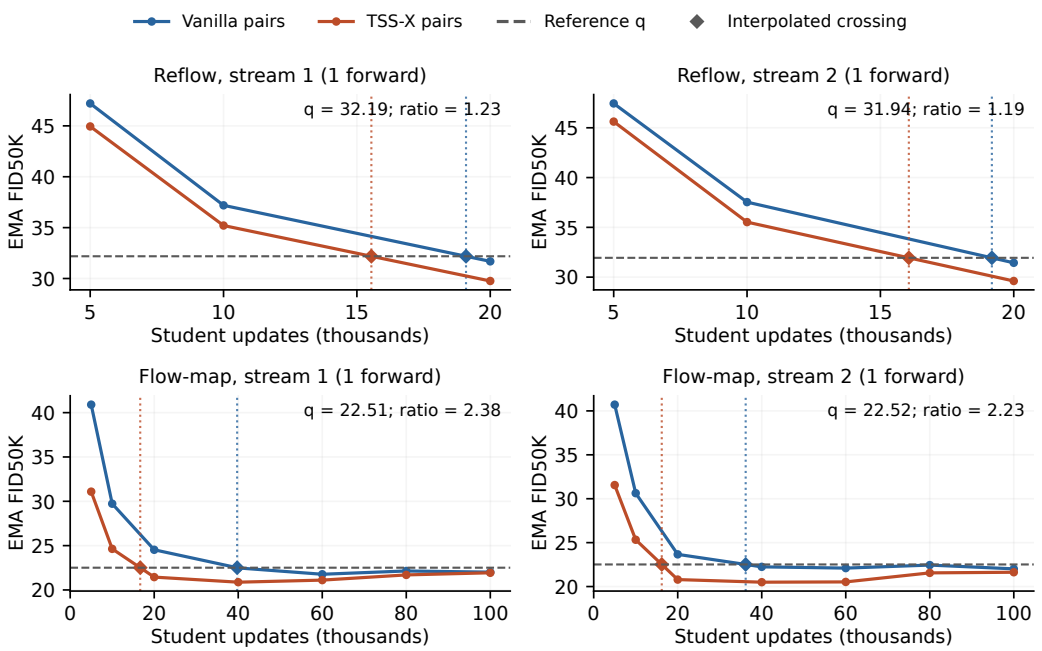

Figure 3: Full one-call EMA FID trajectories for the original two streams of each consumer. Circles are measured checkpoints; dashed horizontal lines are the endpoint-derived references in Table 16. Diamonds and vertical dotted lines mark estimated first crossings between evaluations. The full 100k trajectories show the narrowing gaps and TSS students' deterioration after 40k, with a separate panel for each training stream.

Table 23: Second teacher pair: descriptive updates to its own vanilla 20k EMA FID plus 0.5. Entries are first observed / interpolated updates (thousands), with the same definition as Table 16. This pair has no 100k continuation.

Calls	Reference q	Vanilla	TSS	Interpolated ratio
1	25.5034	20 / 19.19	20 / 10.08	1.90
2	20.2573	20 / 17.84	10 / 8.63	2.07

pools have about 2% lower transport cost and about 50% higher noise–endpoint alignment at equal endpoint energy, in both teacher pairs. Updates-to-target on the 640k runs follow the definitions of Appendix A.4.9: with the vanilla 60k EMA plus 0.5 as reference, the interpolated one-call ratio is 1.21 for EMA weights and 1.99 for Online weights, and 1.34/2.76 at two calls; the EMA ratio is compressed by the 0.9999 decay.

A.4.12 Training cost

TSS-X adds one detached EMA forward per training step and no inference work. If a forward costs F and the online backward costs approximately $2F$, this is a 33.33% analytical increment over a $3F$ training step. Measured JiT-B/16 wall-clock overhead is 4.30–15.50% across three matched pairs (10.08% mean), with about 1.1% sampled peak-memory growth; Appendix G gives each run. S_{norm}adds no training or network evaluation, only a field multiplication per call and one grid normalization. NFE stays unchanged, while these arithmetic operations still contribute to runtime.

For downstream use, teacher post-training, endpoint generation, student training, and evaluation are separate costs. The original three-arm reflow campaign used 7.65/9.19/8.95 Blackwell GPU-hours for Vanilla_{10}/TSS-X_{10}/Vanilla_{19} endpoint generation, student training and evaluation, with different I/O conditions. The matched historical teacher continuations used 238.19/263.09 H100 GPU-hours for vanilla/TSS-X, an increment of 24.91 H100 hours before downstream use. We report H100 and

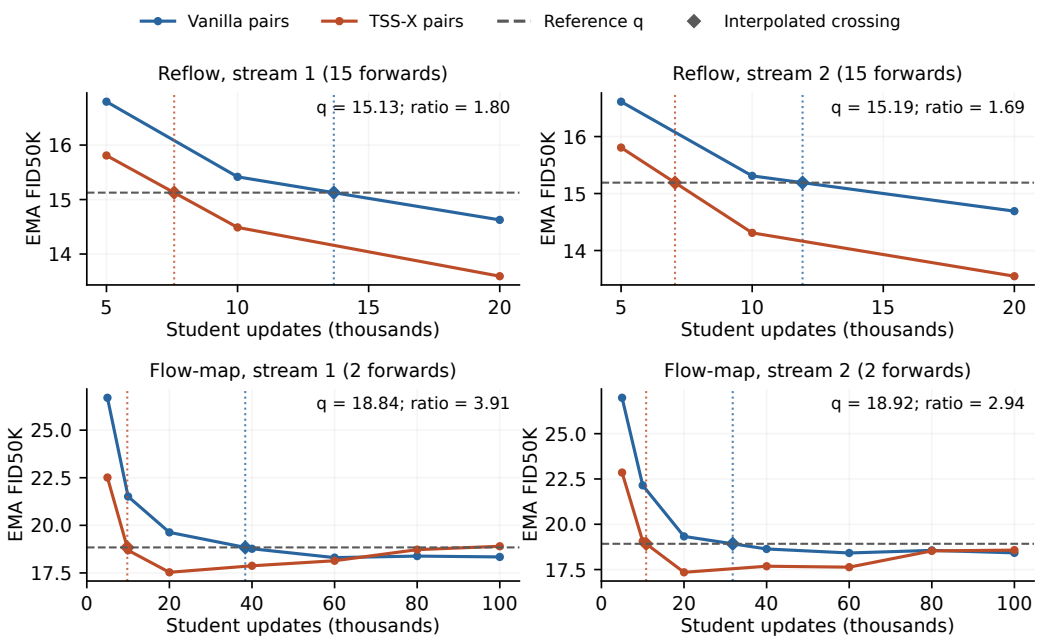

Figure 4: Secondary sampling budgets for the data-efficiency analysis: reflow at eight Heun intervals (15 forwards) and two-call flow-map sampling. Markers and references follow Figure 3. Earlier reference attainment coexists with later regressions: both final two-call flow-map comparisons favor vanilla. Crossings are estimated between measured checkpoints.

Blackwell allocations separately because hardware and I/O conditions affect throughput. Reducing teacher field calls from 19 to 10 saves 47.4% within the rollout stage; generation cost is incurred once per cached pool and amortized over its uses.

A.5 Guided flow-map post-training: a budget-proxy diagnostic

This experiment trains the JiT-B/16 network as a guided flow-map arm (F) on real ImageNet and compares its endpoint with the historical epoch-700 TSS-X and vanilla models. It is separate from the synthetic-pair data-efficiency experiments. Arm F adds a zero-initialized interval embedding and uses the pMF identity with a detached JVP, adaptive weighting and logit-normal time pairs. The instantaneous target is $\omega v_c + (1-\omega)v_u$, where v_c is the conditional data target and v_u is the detached unconditional boundary prediction. We fix $\omega = 2.5$ for non-dropped labels at noise-to-data times $t \leq 0.9$, and sample without a further CFG combination (implementation CFG 1); guidance is embedded in the training target. This JiT-specific recipe omits the official pMF perceptual loss and guidance embedding.

Initial-state audit. The runs use different states from a shared epoch-600 checkpoint file. Arm F loads its EMA1 weights, starts fresh AdamW and initializes its own EMA. Historical Gate2B TSS-X resumes the checkpoint's distinct Online weights, AdamW and EMA states. The flow-map run also uses a 1000-update warmup; Gate2B uses none. These differences in initial weights and optimization state make the historical comparison descriptive. We retain the frozen starting weights and budget.

Budget and evaluation. On the same eight-H100 allocation and global batch 1024, a short timing probe measured 0.51844 seconds per vanilla update and 2.01993 per guided flow-map update, giving $\kappa = 3.89617$. Counting vanilla and TSS-X updates as $3F$ and $4F$, respectively, the frozen rule allocated

$$U = \mathrm{round}_{1000}\left(100\frac{1{,}281{,}167}{1024}\frac{4}{3\kappa}\right) = 43{,}000$$

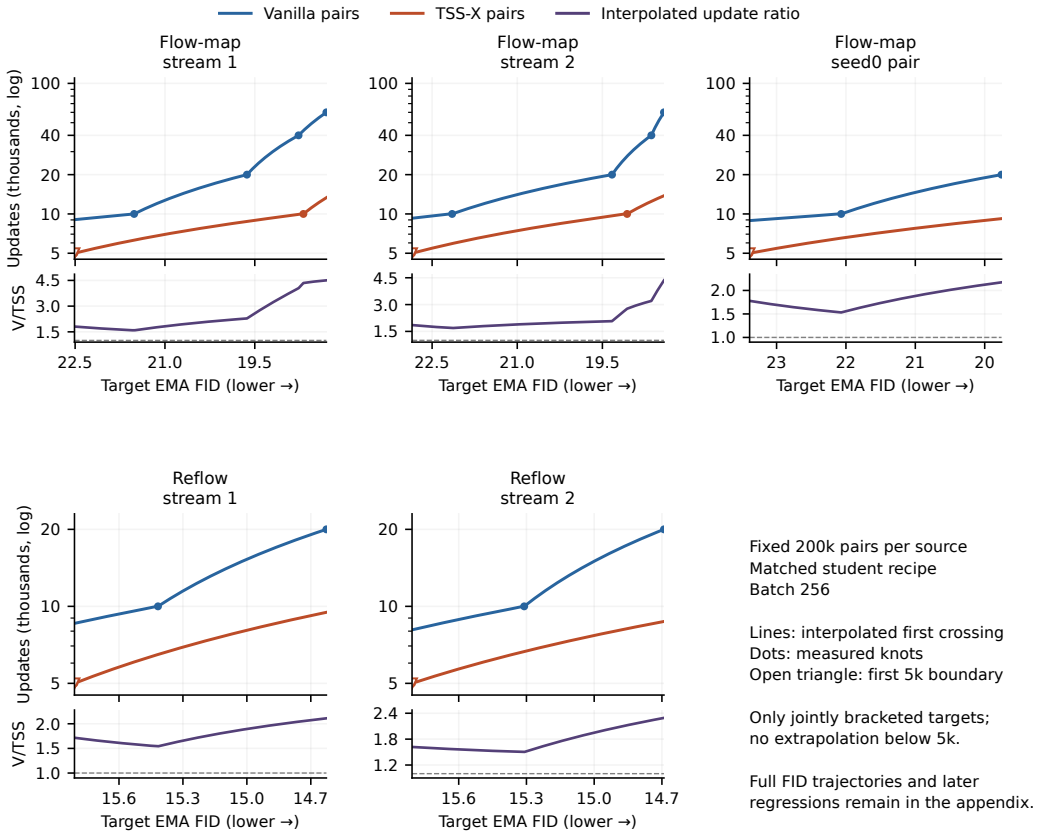

Figure 5: Target-FID first-attainment curves at the secondary sampling budgets: two student forwards for flow-map and eight Heun intervals (15 forwards) for reflow. Unlike Figure 2's learning curves, the horizontal axis is target FID (decreasing to the right), the vertical axis is first-attainment updates (log scale), and lower panels show Vanilla/TSS ratios. Crossings use the same raw-update interpolation; open triangles mark the 5k left-censoring boundary. The additional seed-0 pair ends at 20k; the original pair's later two-call regressions are preserved in Figure 4.

updates to the flow-map run, about 34.37 ImageNet passes. The rounded proxy budget is 1.0043 times the TSS-X reference budget. This forward-equivalent proxy uses the measured timing ratio on that hardware; actual GPU allocations are reported separately below. Learning rate is 2×10^{-4} after warmup, AdamW betas are $(0.9, 0.95)$, weight decay is zero, microbatch is 16 and EMA decay is 0.9999. The run completes all 43k effective updates across two allocations, resuming at update 40,001 with full optimizer and per-rank RNG states.

All 30 evaluations contain 50k samples and use the JiT reference statistics with inception-v3-compat. The historical R1 references use H100 NVL and generation seed base 43,000,000; this run uses H100 and base 84,000,000, giving different evaluation noise. One flow-map NFE costs one backbone evaluation, whereas one CFG-combined JiT NFE costs two branch evaluations. Inference latency also depends on this branch count and the hardware.

Arm F reaches EMA FID 38.2606 at one NFE and 6.6960 at ten NFE. For context, the released pMF model gives one-NFE FIDs of 3.6067 and 3.1543 under our original-setting and prespecified condition-diagnostic evaluations, using its own training recipe and budget. At five backbone evaluations, arm F reaches 8.6660, while historical TSS-X DPM++2M reaches 8.7210 at nine CFG-combined NFE (18 branch evaluations). At 19 NFE, TSS-X reaches 6.0605, at greater inference work than the ten-NFE flow map. Figure 6 shows these operating points under their respective protocols.

Training uses 194.6424 H100 GPU-hours; total allocation, including both probes, evaluation and I/O, is 201.6005 H100 GPU-hours. The failed first probe is included. Historical Gate2B cost records

Table 24: Pool-prefix students: all milestone FID50K readings. Δ = TSS-X − Vanilla at equal pool size.

Pairs	Updates	Weights	Calls	Vanilla	TSS-X	Δ
50k	5k	Online	1	37.1907	29.4648	−7.7259
50k	5k	Online	2	24.6555	21.8510	−2.8045
50k	5k	Ema	1	41.6489	31.4060	−10.2430
50k	5k	Ema	2	26.8191	22.3016	−4.5174
50k	10k	Online	1	29.8379	25.1815	−4.6564
50k	10k	Online	2	21.5194	18.6579	−2.8615
50k	10k	Ema	1	31.8490	25.1986	−6.6504
50k	10k	Ema	2	22.2047	18.7122	−3.4925
50k	20k	Online	1	26.6112	23.2593	−3.3518
50k	20k	Online	2	20.7728	18.8325	−1.9403
50k	20k	Ema	1	27.5537	23.2301	−4.3237
50k	20k	Ema	2	20.9711	18.4674	−2.5037
100k	5k	Online	1	36.3074	28.9810	−7.3264
100k	5k	Online	2	24.3004	22.1025	−2.1979
100k	5k	Ema	1	41.2385	31.5615	−9.6770
100k	5k	Ema	2	26.6718	22.6021	−4.0698
100k	10k	Online	1	29.7903	23.7308	−6.0594
100k	10k	Online	2	21.8198	18.2423	−3.5775
100k	10k	Ema	1	31.0796	24.7625	−6.3171
100k	10k	Ema	2	22.0837	18.6356	−3.4481
100k	20k	Online	1	24.5979	21.9754	−2.6224
100k	20k	Online	2	19.6318	17.8924	−1.7394
100k	20k	Ema	1	25.4769	21.6444	−3.8325
100k	20k	Ema	2	20.0661	17.5800	−2.4860

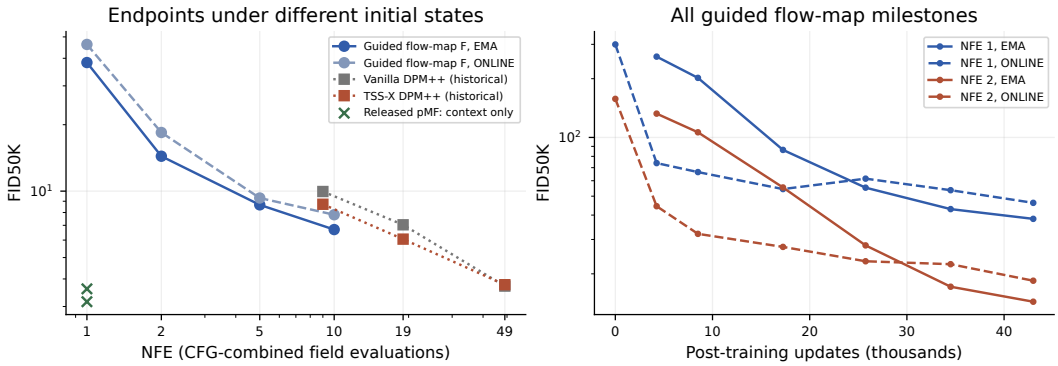

Figure 6: Left: endpoint quality versus actual NFE under the different initialization, training and evaluation conditions described above. Lines connect measured settings. Right: all one/two-NFE flow-map milestones, including unfavorable EMA and non-monotone Online readings.

report 263.0933/238.1867 H100 GPU-hours for TSS-X/vanilla at a different 64-GPU topology; we retain these hardware-specific totals separately from the present eight-GPU budget. No nonfinite or rejected updates occurred, and the recorded spike count was 0/42,900 under the preceding-100-update median rule. Stability comparison would require matched additional runs. Arm F uses real ImageNet data throughout; the synthetic-pair comparisons are reported in Appendix A.4.9.

Table 25: Fresh 640k-pair pools, 60k updates with cosine decay to zero and EMA 0.9999: all milestone FID50K readings and the endpoint at 1/2/5/10 forwards. Δ = TSS-X $-$ Vanilla.

Updates	Weights	Calls	Vanilla	TSS-X	Δ
10k	Online	1	28.5844	24.5003	-4.0841
10k	Online	2	21.4676	18.7078	-2.7598
10k	Ema	1	151.7314	138.0256	-13.7059
10k	Ema	2	81.3709	72.2060	-9.1648
20k	Online	1	23.6527	21.1903	-2.4624
20k	Online	2	19.5529	17.6430	-1.9099
20k	Ema	1	67.8674	59.2847	-8.5828
20k	Ema	2	35.2875	30.5301	-4.7575
30k	Online	1	21.7991	20.7091	-1.0900
30k	Online	2	18.5899	17.7053	-0.8846
30k	Ema	1	34.4844	30.0271	-4.4574
30k	Ema	2	23.2314	20.3349	-2.8964
40k	Online	1	21.1856	20.1153	-1.0703
40k	Online	2	18.3744	17.4472	-0.9272
40k	Ema	1	25.0633	22.5282	-2.5351
40k	Ema	2	19.7512	17.9892	-1.7620
50k	Online	1	20.5497	19.7509	-0.7987
50k	Online	2	18.0275	17.1643	-0.8632
50k	Ema	1	22.1512	20.9171	-1.2341
50k	Ema	2	18.6406	17.7936	-0.8470
60k	Online	1	20.6887	19.7038	-0.9849
60k	Online	2	18.1343	17.1559	-0.9784
60k	Online	5	16.3161	15.2048	-1.1113
60k	Online	10	15.1465	14.0715	-1.0750
60k	Ema	1	21.0873	20.4418	-0.6455
60k	Ema	2	18.1900	17.7241	-0.4659
60k	Ema	5	16.1127	15.6102	-0.5025
60k	Ema	10	14.8386	14.4838	-0.3548

Table 26: Pool quality versus coupling geometry. FID50K on the first 50k endpoints; kNN precision/recall ($k = 3$) on Inception-2048 features of those endpoints against 50k evenly spaced real ImageNet training images; transport cost $\mathbb{E}\|x_1 - x_0\|^2$ per dimension and $\cos(x_1, x_0)$ over the first 8,192 pairs with their generating noise. Empty cells were not computed.

Pool	FID	Precision	Recall	Transport cost	$\cos(x_1, x_0)$
seed-1 Vanilla$_{10}$ (200k)	9.42	0.666	0.376	1.0706	0.064
seed-1 TSS-X$_{10}$ (200k)	8.22	0.681	0.399	1.0495	0.095
seed-1 Vanilla$_{14}$ control (200k)	8.11	0.667	0.421	1.0804	0.075
seed-0 Vanilla$_{10}$ (200k)	9.53	0.670	0.370	1.0768	0.061
seed-0 TSS-X$_{10}$ (200k)	8.36	0.677	0.391	1.0574	0.090
seed-1 Vanilla (640k, fp16)	9.49	0.670	0.376	–	–
seed-1 TSS-X (640k, fp16)	8.22	0.683	0.400	–	–

B Proofs and adapter details

Proof of exact transport consistency. For a sampled trajectory, d_a and s_a are constants with respect to t. Differentiating Eq. (21) gives

$$\frac{d}{dt}\left(z_t + b(t)s_a d_a\right) = u_t + b'(t)s_a d_a = u_t^\star.$$

The compact smoothstep has zero value and derivative at the two support boundaries, so state and tangent agree with the base path outside the support.

Table 27: Guided flow-map arm F at the frozen 43k endpoint. FID50K for both Online and EMA weights.

NFE	Online	EMA
1	46.2190	38.2606
2	18.4407	14.3940
5	9.3078	8.6660
10	7.8166	6.6960

Table 28: Complete one/two-NFE trajectory. The zero-update row is shared by identical EMA and Online weights.

Updates	Online, 1	Online, 2	EMA, 1	EMA, 2
0	299.6396	157.6277	–	–
4250	73.8805	44.4264	259.4692	132.4375
8500	66.4500	32.0581	201.9612	106.2621
17250	54.3765	27.4808	86.3112	55.4159
25750	61.4854	23.1983	55.2827	27.9771
34500	53.6602	22.4077	42.9295	17.1772
43000	46.2190	18.4407	38.2606	14.3940

Proof of marginal-flow validity. For any smooth compactly supported test function φ,

$$\frac{d}{dt}\mathbb{E}[\varphi(z_t^\star)] = \mathbb{E}[\nabla\varphi(z_t^\star)^\top u_t^\star] = \mathbb{E}[\nabla\varphi(z_t^\star)^\top v^\star(z_t^\star, t)].$$

Integration by parts gives the continuity equation in weak form. Endpoint safety gives $z_0^\star = z_0$ and $z_1^\star = z_1$ trajectory by trajectory, so the endpoint marginals also agree.

Proof of parameterization covariance. The determinant of the matrix in Eq. (23) is $\Delta(t)$. If it is nonzero, the inverse exists and gives

$$x^\star = \frac{\sigma' z^\star - \sigma u^\star}{\Delta}, \qquad \epsilon^\star = \frac{-\alpha' z^\star + \alpha u^\star}{\Delta}.$$

Substitution into Eq. (24) gives a unique target for any specified linear native head. At the anchor, the matrix formed by the schedule row (α, σ) and head row (c_x, c_ϵ) is invertible under the second determinant condition, so the teacher output and anchor state recover one effective endpoint pair.

Proof of the correction bound. From Eq. (21),

$$\mathrm{RMS}(u_t^\star - u_t) = |b'(t)| s_a \,\mathrm{RMS}(d_a) \le \|b'\|_\infty s_a \,\mathrm{RMS}(d_a).$$

Let $C = \|b'\|_\infty \,\mathrm{RMS}(d_a)$ and $R = \rho\,\mathrm{RMS}(u_a)$. For $\|b'\|_\infty > 0$, Eq. (20) gives $s_a = \min(1, R/\max\{C, \varepsilon\})$, so $s_a C \le R$ whether the minimum selects 1 or the ratio. If $\|b'\|_\infty = 0$, the correction is zero. This proves the bound for every t, including $d_a = 0$.

Proof of the interval identity. For $s < t$,

$$\frac{z_t^\star - z_s^\star}{t-s} = \frac{z_t - z_s}{t-s} + \frac{b(t) - b(s)}{t-s} s_a d_a.$$

This is also the integral average of $u^\star$ by exact transport consistency. Setting $s = 0$, $t = 1$, and $b(0) = b(1) = 0$ leaves the analytic conditional path's full-interval finite-difference target unchanged. Learned marginal ODEs and CMT teacher rollouts are separate objects. With this conditional-path target fixed, any one-step benefit after transport-consistent TSS pre-training must enter through the learned parameters or representations.

The mean-value theorem also gives

$$\frac{|b(t) - b(s)|}{t-s} s_a \,\mathrm{RMS}(d_a) \le \|b'\|_\infty s_a \,\mathrm{RMS}(d_a) \le \rho\,\mathrm{RMS}(u_a),$$

which proves the interval-correction bound.

C ADDITIONAL JIT RESULTS

C.1 EARLIER SCALE CONTINUATIONS

Table 29: Earlier scale continuations on ImageNet 256×256. All entries are FID from 50,000 images. B/16 and L/16 use CFG 2.5; H/16 uses CFG 2.2. Columns are Heun solver steps. There is one training pair per scale, separate from the three-seed B/16 result in Table 5. The L/16 pair uses vanilla epoch700 and TSS-X epoch699, detailed below.

Scale	Arm	H5	H10	H25	H50
B/16	Vanilla	11.51	8.42	**4.03**	**3.74**
	TSS-X	**10.06**	**7.24**	4.07	3.77
L/16	Vanilla	10.18	6.62	3.02	2.48
	TSS-X	**8.55**	**5.42**	**2.70**	**2.25**
H/16	Vanilla	10.40	6.36	2.1606	1.94
	TSS-X	**9.09**	**5.53**	**2.1562**	**1.90**

The scale study shows H5/H10 gains across model sizes within the JiT pixel-space family. At H25/H50, B/16 is slightly worse, L/16 improves, and H/16 is close to or better than vanilla.

Historical checkpoint endpoints. The L/16 evaluation logs identify vanilla's internal epoch700 and TSS-X's internal epoch699 checkpoint. These Heun FIDs therefore compare different training endpoints. The later TSS-X L/16 checkpoint continued by 1,251 updates to epoch700 was used for the larger-teacher downstream study and the DPM++2M evaluation in Table 2, but was not the source of this Heun table. Both H/16 checkpoints have internal epoch699 despite their epoch700 filenames; their endpoints agree.

C.2 ANCHOR, BASELINE, AND RESOLUTION CONTROLS

At the base learning rate specified in the JiT repository, a CFG sweep over $\{2.0, 2.5, 2.75, 3.0, 3.5, 4.0\}$ gives TSS-X H10 FIDs $10.28, 7.24, 6.44, 6.19, 6.14, 6.40$ and vanilla FIDs $10.94, 8.42, 7.80, 7.43, 7.31, 7.33$. TSS-X is better at every shared CFG. A separate vanilla learning-rate and CFG grid obtains its best observed value, 6.78, at half the base learning rate and CFG 3.5. TSS-X obtains 6.14 at the base learning rate specified in the JiT repository and the same CFG. These are the best observed values in the stated grids.

The anchor sweep uses one continuation per added anchor. At H10, $a = 0.3, 0.5, 0.7$ gives FID $10.96, 7.24, 6.68$. The $a = 0.7$ run also gives 9.79 at H5 and 3.54 at H50. These runs show sensitivity among the three tested values; later anchors remain untested.

The matched JiT-L/32 512×512 control continues both arms from epoch 600 to 700 and evaluates 50,000 images with generation seed 0, CFG 2.5, and H10. Vanilla obtains 6.5844 FID and TSS-X obtains 5.6025, a reduction of 0.9819 or 14.9%. This result has one training pair and one solver setting. The vanilla reference is our matched continuation evaluated at the same sampling budget.

C.3 SEED-0 50,000-IMAGE MECHANISM EVALUATION

In this seed, D+N is within 0.0075 FID of TSS-X at H10. The independent three-seed Gate-2B result is the authority for method selection and retains TSS-X.

C.4 EARLIER TARGET-CORRECTION DIAGNOSTIC

Before the D/P/N analysis, we tested an SI-style derivative correction for the Gaussian TSS-X path. On JiT-B/16, the target-only variant is worse than vanilla at every tested solver step, while state plus correction gives less H5/H10 benefit than TSS-X. On JiT-L/16, TSS-X is better than state plus correction at H5, H10, H25, and H50. These ablations support retaining the original TSS-X target on JiT.

Table 30: Gate-2A mechanism evaluation at one training seed. All entries use 50,000 images, generation seed 0, CFG 2.5, and the shared epoch-600 source and continuation budget. Columns are Heun solver steps.

Arm	Factors	H10 FID	H50 FID
Vanilla	000	8.5258	3.7530
TSS-X	D+P+N	**7.4124**	**3.6392**
D+N	101	7.4199	3.6809
D+P	110	7.7923	3.8972

Table 31: Earlier JiT ingredient ablation. Every cell uses 50,000 images and Heun sampling. B/16 and L/16 use CFG 2.5. H5/H10/H25/H50 are solver steps.

Scale	Arm	H5	H10	H25	H50
B/16	Vanilla	11.51	8.42	4.03	3.74
	TSS-X	**10.06**	**7.24**	4.07	3.77
	State + correction	10.87	7.57	**3.84**	**3.69**
	Correction only	12.48	9.03	4.31	4.26
L/16	Vanilla	10.18	6.62	3.02	2.48
	TSS-X	**8.55**	**5.42**	**2.70**	**2.25**
	State + correction	8.68	5.83	2.91	2.45
	Correction only	9.58	6.39	3.06	2.70

D ADDITIONAL MECHANISM DIAGNOSTICS

A frozen-checkpoint diagnostic decomposes the full warp direction $\widehat{x}_a - z_t$ at the profile peak. At $a = 0.5$, 94.44% of its measured energy is parallel to the clean path and 5.56% is normal. At $a = 0.7$, the split is 91.71% and 8.29%. These sample estimates use 32 ImageNet examples at one JiT-B/16 checkpoint.

We also retain a local Heun truncation diagnostic as correlational evidence. For a base-path state at time t, it compares one coarse step of size $h = 0.1$ with ten fine substeps:

$$\mathrm{err}(t) = \mathbb{E}\left[\|H_{\mathrm{coarse}}(z_t) - H_{\mathrm{fine}}(z_t)\|_2/\sqrt{D}\right].$$

The difference between TSS-X and vanilla is concentrated toward the data end, but this diagnostic does not fully predict the FID ordering of the target-corrected ablation. It is therefore not used as a causal explanation.

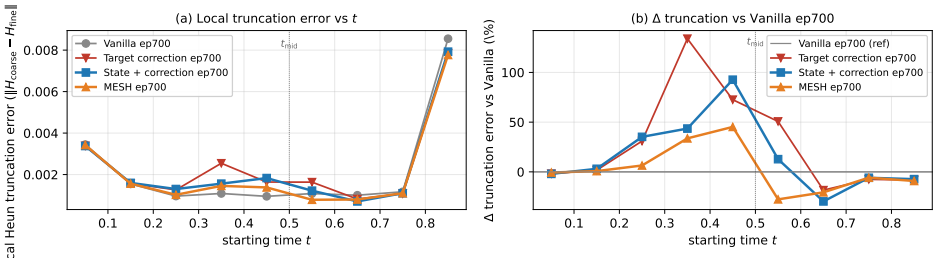

Figure 7: Local Heun truncation diagnostic on one JiT-B/16 checkpoint set. The states are sampled from the training interpolant rather than generated trajectories. The plot is correlational and does not establish trajectory straightness or explain every FID difference.

The teacher-free D shift also demonstrates checkpoint sensitivity. At epoch 610 it would pass the screen with H10/H50 FID $9.8810/6.0444$, but at the predeclared epoch-625 endpoint it degrades to $10.5267/6.6972$. The corresponding vanilla endpoint is $10.9201/6.3353$. Selecting the intermediate checkpoint after inspection would change the mechanism conclusion.

E LATENT COMPONENT-FACTORIAL DETAILS

Table 32 gives the individual cells behind the earlier all-A100 625-update factorial. The intervention switches follow Eq. (28). Every cell uses the same 625-update continuation and 50,000-image, 10-point Euler evaluation, separate from the full-budget time-only/full study.

Table 32: Per-seed LightningDiT-B/2 component factorial. Lower FID is better.

Training seed	Vanilla	State-only	Tangent-only	Full transport-consistent TSS
20260825	18.9057	19.1126	**18.4941**	18.6174
20260826	18.9342	19.1234	**18.6259**	18.7835
20260827	18.7301	18.9041	**18.3766**	18.4732

Across seeds, full minus tangent-only is $+0.1258$ with 95% Student-t interval $[+0.0499, +0.2017]$. The 2×2 interaction, $(\text{full} - \text{state-only}) - (\text{tangent-only} - \text{vanilla})$, is -0.0642 $[-0.1347, +0.0062]$. Its interval includes zero, leaving the interaction unresolved.

F ADDITIONAL DOWNSTREAM BOUNDARIES

L teacher to the same B student. We repeat the fixed 200k-pair, 20k-update reflow design using L/16 teachers and a common B/16 student initialization. The teacher endpoints are matched at epoch 700; TSS-X L/16 is obtained by continuing its full epoch 699 state for exactly 1,251 updates with its original optimizer, EMA and shaping recipe. With EMA students, vanilla/TSS endpoint sources yield one-call FID 54.0564/53.1044 and 15-call FID 19.5735/17.8685. Online values are 51.0282/53.0953 and 20.0232/18.0260. Thus multistep gains hold for both weight choices, while one-call signs disagree. This is one teacher pair and one student stream. Because this stage uses an H100 environment rather than the B-teacher Blackwell environment, its worse absolute student FID cannot be attributed to teacher scale alone.

Initialization-only test in the JiT flow-map network. A separate pixel flow-map experiment initializes the same pMF-identity/JVP target from vanilla or TSS-X Gate-2B epoch 700 EMA1 weights. Added interval parameters share their initialization; both arms start fresh AdamW and their own EMA. Full ImageNet, CFG 1, batch 256, microbatch 16 and 50k updates are fixed. The first 1,250 updates run on Blackwell, followed by matched full-state four-H100 continuations; all 37 final evaluation cells use one common H100 environment.

Table 33: Initialization-only test: 50k-update FID 50K in the JiT pixel flow-map network. One/two calls apply the map directly; Heun with one/two intervals would use one/three calls.

Weights	Calls	Vanilla init.	TSS-X init.	Δ
EMA	1	82.0159	80.3757	-1.6403
Online	1	95.6743	78.5397	-17.1347
EMA	2	51.7024	51.1915	-0.5109
Online	2	57.7356	51.8033	-5.9323

The primary one-call EMA difference is a small positive signal at limited absolute quality. EMA differences at 5k/12.5k/25k/50k are $-4.9162, +1.0867, +0.9683, -1.6403$, changing sign during training. TSS at 42.5k reaches 81.4028 compared with vanilla's 82.0159 at 50k, but vanilla already reaches 81.7974 at 25k; earlier vanilla attainment defeats the proposed 15% update-saving interpretation. The large final Online gap accompanies vanilla's Online increase from 80.2832 at 25k to 95.6743 at 50k.

Both runs complete 50k effective updates with zero nonfinite/rejected updates. Gradient spikes, defined as exceeding ten times the preceding 100-valid-update median, are 3/49,900 in each arm; gradient 99th percentiles are 1.5657/1.5825. The spike counts match in this single-stream comparison.

Table 34: Fixed-pair JVP consumer, two matched student training streams sharing teachers, pools, full initialization and evaluation noise. FID50K at the frozen checkpoints; one and two calls are direct map compositions. Common zero-update Online/EMA weights are identical and give FID 301.5403/158.8908 at one/two calls in both streams.

Stream	Updates	Weights	Vanilla$_{10}$ 1 call	TSS-X$_{10}$ 1 call	Vanilla$_{10}$ 2 calls	TSS-X$_{10}$ 2 calls
1	5k	Online	36.6945	27.8461	24.6054	20.7449
1	5k	EMA	40.9031	31.0885	26.7008	22.5108
1	10k	Online	28.1692	25.1220	20.4954	19.0098
1	10k	EMA	29.7279	24.6342	21.5147	18.6882
1	20k	Online	24.2019	21.8369	19.7373	17.8966
1	20k	EMA	24.5294	21.4469	19.6286	17.5348
2	5k	Online	38.6871	29.2664	26.1277	21.1691
2	5k	EMA	40.7126	31.5556	26.9817	22.8535
2	10k	Online	30.2076	26.3059	22.1996	19.3417
2	10k	EMA	30.6317	25.3294	22.1474	19.0646
2	20k	Online	23.0113	21.2966	18.9773	18.0262
2	20k	EMA	23.6579	20.7923	19.3251	17.3481

The stage consumed 2.42 Blackwell and 92.8794 H100 allocated GPU-hours, including the migration probe and evaluation. The hardware migration changed throughput in both arms. All unfavorable checkpoints and both weight choices are retained.

F.1 Fixed synthetic pairs in the internal JVP consumer

Both students start from the same Gate-2B seed-1 epoch-700 vanilla EMA1 weights, including identical new interval parameters. AdamW is fresh and EMA starts from each identical online state. We retain the host's data-at-zero convention, JiT time query $1 - t$, $t_\epsilon = 0.05$, boundary-velocity JVP tangent, detached correction and math-SDPA backend. The two 200k pools retain their original per-ID Gaussian noise, raw unclipped RGB endpoints and class labels, without augmentation. Both have teacher CFG 2.5; evaluation CFG is 1. No TSS student initialization, midtraining, teacher-field supervision or guidance curriculum is added.

Training uses batch 256, microbatch 16, learning rate 5×10^{-5} with 200-update warmup, AdamW betas $(0.9, 0.95)$, zero weight decay, EMA 0.999 and exactly 20k updates. The time sampler has logit-normal mean -0.4, standard deviation 1 and diagonal proportion 0.5; adaptive weighting uses exponent 1 and offset 0.01 on per-example squared-error sums. Both arms share the sample-ID and time streams. Training and all evaluations use single H100 NVL GPUs in the same software environment. The equal-update comparison is not an equal-total-pipeline-compute experiment.

The predeclared primary candidate rule is a final one-call EMA difference at most -0.5; both -3.0825 and -2.8656 pass this descriptive selection rule. Mean final EMA FID is $24.0936/21.1196$ at one call and $19.4768/17.4415$ at two calls for Vanilla$_{10}$/TSS-X$_{10}$. Relative to the identical zero-update baseline, first-stream final one-call EMA changes are $-277.0109/-280.0934$, and two-call changes are $-139.2622/-141.3559$; the second stream gives $-277.8824/-280.7480$ and $-139.5657/-141.5427$. Absolute quality remains well below that of the released one-step models. Training seeds are 20260925/20260926, sample-ID bases 85000000/87000000 and time bases 86000000/88000000; the new-parameter seed remains 20260925 and evaluation base 84000000. The streams share teachers, pools and full initial weights, including the interval parameters. The main downstream comparison also reports the negative result for direct one-step endpoint fine-tuning.

All four runs complete 20k effective updates without nonfinite or rejected updates. Each has zero spikes in 19,900 eligible updates under the same preceding-100 median rule. First-stream gradient 99th percentiles are 22.5894/19.3554 and JVP-RMS 99th percentiles 1.9699/1.9573; second-stream values are 22.3515/19.4464 and 1.9354/1.9247. This telemetry does not establish improved training stability. Final-100-update raw MSE is 0.0010841/0.0013353 in the first stream and

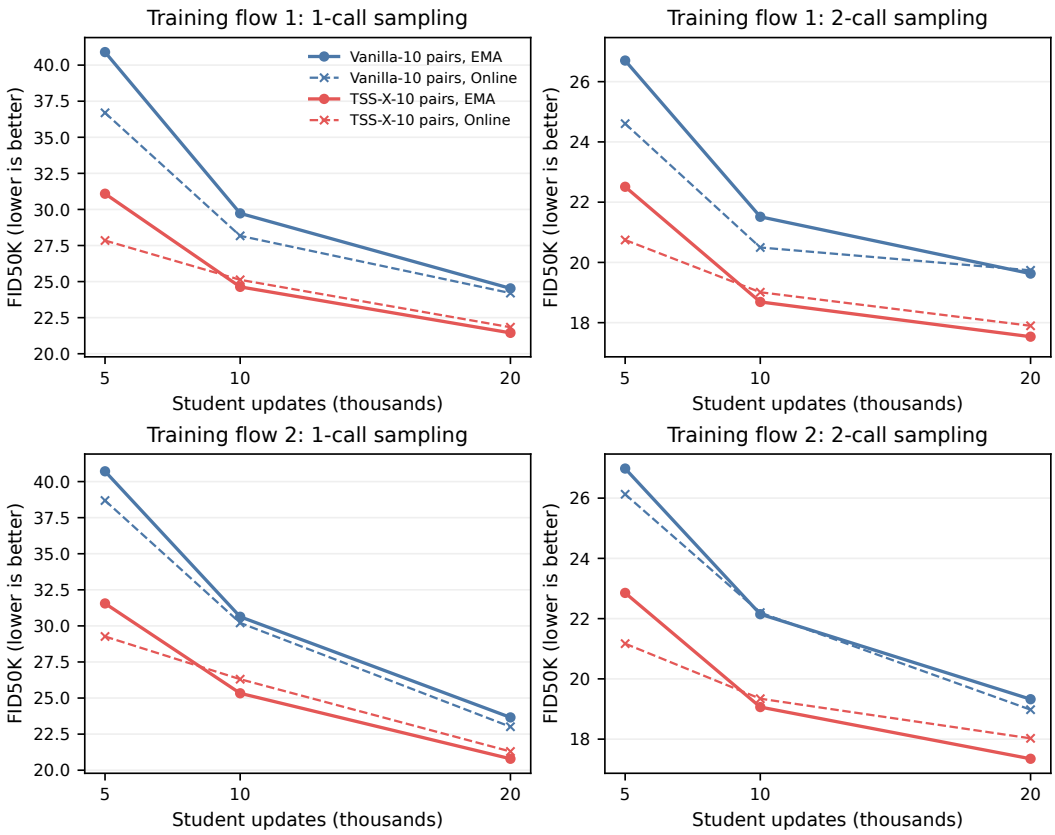

Figure 8: All measured 5k–20k JVP-consumer quality points in both student training streams. TSS-X_{10} remains better, but the one-call EMA gaps narrow. Lines connect observations within each stream; both streams share teachers and pools. The common zero-update baseline is reported in Table 34.

0.0012090/0.0014880 in the second, higher for TSS-X_{10} despite its better FID; this random-time host loss is not a mapping-complexity measure.

First-stream allocation totals are 16.7522/16.3461 H100 NVL GPU-hours for the formal arms plus 0.2433 for isolated smoke, or 33.3417 overall. Logged training including loader waits totals 27.2577 hours within that allocation. The second stream uses 16.1906/16.2747 allocated hours, or 32.4653 total, with 26.1037 logged training hours as a subset and no additional smoke or failed attempts. Across the two streams, 65.8069 allocated GPU-hours include evaluation, checkpoint IO and startup. Historical teacher and pool generation costs require separate accounting when these assets are reused. The one-call EMA gaps shrink from $9.8146/9.1571$ at 5k to $3.0825/2.8656$ at 20k. Appendix A.4.9 analyzes first attainment in student updates; GPU time and total pipeline cost are separate measures.

Predeclared budget-extension protocol. To test whether the 20k advantage persists at a larger student budget, we continue all four runs to 100k updates under a predeclared protocol. Each arm resumes its own 20k Online, EMA, Adam and random-number states. Updates 20,001–100,000 continue the original sample-ID and time streams, constant post-warmup learning rate, batch 256, microbatch 16, objective and EMA, preserving the completed warmup and optimizer state. The raw 200k endpoint pools and their original noise coupling remain unchanged. At 100k, each stored pair receives an expected 128 accesses under sampling with replacement.

All four continuations use four H100 GPUs per arm in a common software environment, after a single-versus-four-GPU gradient check. Because the parent runs used H100 NVL GPUs and a different PyTorch version, the 16 parent-20k evaluation cells are repeated in the new environment. The extension curves use these common-environment references; Table 34 retains the original

Table 35: Budget extension: final 100k FID50K in the common H100 environment. $\Delta = \text{TSS-X}_{10} - \text{Vanilla}_{10}$. The original 20k experiment is retained separately in Table 34. Both one-call EMA differences fall below the predeclared 0.5 descriptive magnitude.

Stream	Weights	Calls	Vanilla_{10}	TSS-X_{10}	Δ
1	EMA	1	22.0088	21.9359	−0.0729
1	Online	1	21.8900	22.3308	+0.4408
1	EMA	2	18.3395	18.9000	+0.5604
1	Online	2	18.3632	19.4614	+1.0982
2	EMA	1	22.0153	21.6280	−0.3873
2	Online	1	22.4391	22.0716	−0.3675
2	EMA	2	18.4220	18.5731	+0.1510
2	Online	2	18.6176	18.9722	+0.3546

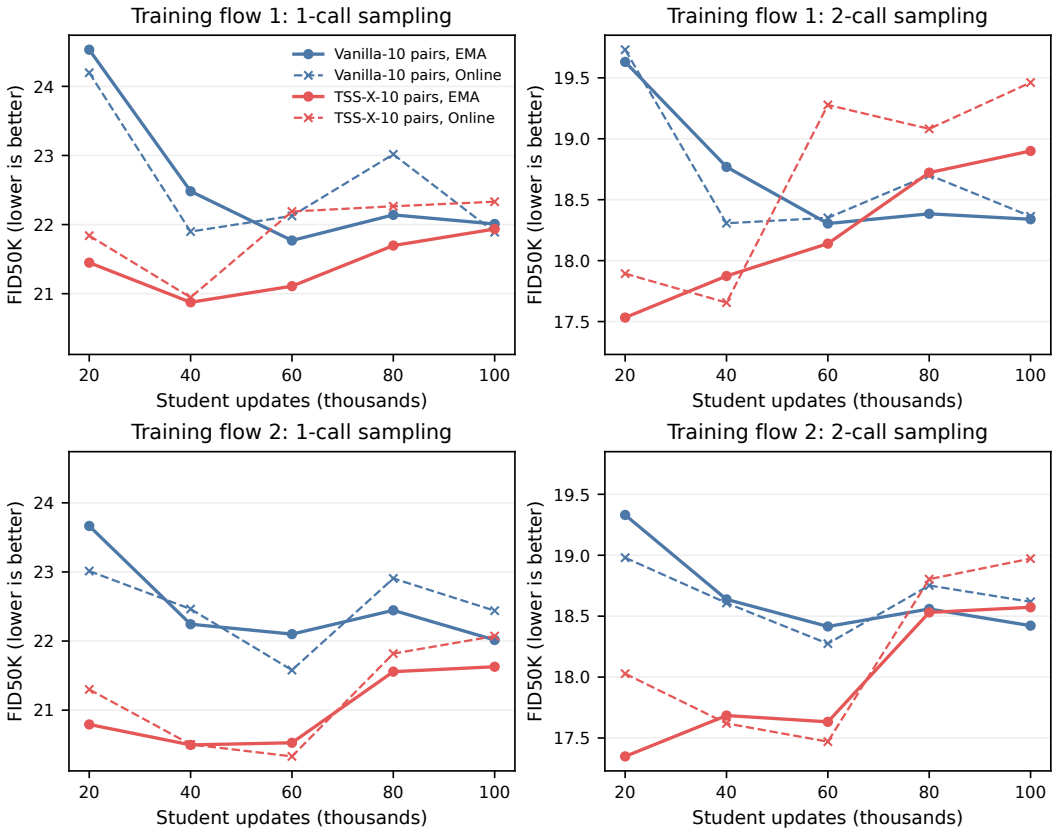

Figure 9: All 80 FID50K cells of the budget extension: 16 parent-20k environment references and 64 new cells at 40k/60k/80k/100k. Both weights and call budgets are shown in each stream. Lines connect measurements within each run. The shared teachers, pools and evaluation noise are unchanged.

measurements. New FID50K evaluations cover 40k, 60k, 80k and 100k, both Online and EMA weights, and one and two map calls, giving 64 new cells plus the 16 environment references.

The primary comparison is the 100k one-call EMA difference in each stream separately, using -0.5 as a descriptive comparison threshold. The 60k/80k differences and each arm's absolute trajectory are reported alongside this endpoint. Appendix A.4.9 defines first attainment and discusses the reference choice and shared data sources.

The maximum absolute FID shift across the 16 parent-environment references is 0.00751. In the common H100 environment, one-call EMA differences at 20k/40k/60k/80k/100k are $-3.0816/-1.6067/-0.6612/-0.4442/-0.0729$ in stream 1 and $-2.8725/-1.7465/-1.5726/-0.8877/-0.3873$ in stream 2. Both TSS-X_{10} students obtain their best measured one-call EMA FID at 40k ($20.8754/20.4969$), then worsen by $1.0605/1.1310$ by 100k. Vanilla improves overall from 20k, but its later curve is not monotone either. At 100k, both two-call weight choices favor vanilla in both streams. These observations limit budget extrapolation of the 20k benefit without identifying a cause, establishing convergence, or proving equal performance.

The predeclared exploratory reference is each stream's vanilla 100k one-call EMA FID with tolerance 0.5. On the observed 20k/40k/60k/80k/100k grid, TSS-X_{10} meets this reference from 20k onward and vanilla from 40k onward in both streams. The 50% difference between first measured qualifying checkpoints describes student update counts. Interpretation depends on the endpoint-derived reference, which exceeds vanilla's best measured quality, TSS-X_{10}'s own deterioration, the unobserved crossings between checkpoints and the separately incurred teacher/pool costs.

Each continuation completes exactly 80k new effective updates with all 186 Adam states at the corresponding checkpoint step. Learning rate remains 5×10^{-5} throughout, with no nonfinite or rejected updates and zero gradient spikes in 80k eligible updates per arm; the preceding-gradient window is restored from the parent state. These telemetry measures are identical across arms. New training time totals 243.2219 H100 GPU-hours, including loader waits; total allocation is 253.2691 H100 GPU-hours including evaluation, IO and both probes. The first probe passed gradient agreement but failed the cold-cache throughput criterion; one bounded cache-warm retest passed. Both costs are retained, with no formal training recovery. The parent phase's 65.8069 H100 NVL allocated GPU-hours and historical teacher/pool costs are recorded separately by hardware and stage. This extends the fixed-data-source comparison within the JiT flow-map protocol.

G Training-cost details

Table 36: Measured JiT-B/16 continuation cost over three matched training seeds. Wall time is seconds. Peak memory is sampled per GPU. The final column is the within-seed TSS-X overhead.

Seed	Vanilla time	TSS-X time	Peak MiB: V / TSS-X	Overhead
0	17,064	19,709	7,487 / 7,575	15.50%
1	13,398	14,799	7,487 / 7,563	10.46%
2	15,964	16,650	7,487 / 7,563	4.30%
Mean paired ratio				10.08%

The vanilla runs total 825.35 GPU-hours and the TSS-X runs total 909.48 GPU-hours in this protocol. Different utilization and runtime conditions produce visible variation across pairs, so the measured 10.08% mean should not be treated as a hardware-independent constant. The analytical 33.33% estimate and measured wall time answer different questions.

H Algorithms

TSS-X for a clean-x head.

```
Input: online model f, EMA model g, endpoints (eps, x), label c
       anchor a, beta_max, sigma_beta

sample t
z_t = (1-t) * eps + t * x
z_a = (1-a) * eps + a * x
with no_grad:
    x_hat = g(z_a, a, c)
beta = beta_max * exp(-(t-a)^2 / (2*sigma_beta^2))
z_mesh = z_t + beta * (x_hat - z_t)
```

```
loss = w(t) * ||f(z_mesh, t, c) - x||^2
```

Transport-consistent TSS with a native linear head.

```
Input: online model f, EMA model g, trajectory endpoints, label c
       anchor a, compact profile b, correction ratio rho

construct anchor state z_a and base tangent u_a
with no_grad:
    y_hat_a = g(z_a, a, c)
recover endpoint estimate from (z_a, y_hat_a)
split endpoint error into metric-parallel P_a and normal N_a
d_a = lambda_D * (x - z_a) + lambda_N * N_a
s_a = min(1, rho*RMS(u_a) / max(max|b'|*RMS(d_a), eps))

sample t independently of trajectory-fixed d_a and s_a
z_star = z_t + b(t)  * s_a * d_a
u_star = u_t + b'(t) * s_a * d_a
convert (z_star, u_star) to native target y_star
loss = w_y(t) * ||f(z_star, t, c) - y_star||^2
```

For TSS-T, omit the EMA query and error projection, set $d_a = D_a$ and use the deterministic scale s_D in Eq. (27). State and target still change together.

S_{norm} for a frozen local-flow sampler.

```
Input: frozen model f, solver field-call times tau[i],
       quadrature weights omega[i], fixed profile b, a, rho

s_D = min(1, rho / (max|b'| * (1-a)))
raw[i] = 1 + (1-a) * s_D * b'(tau[i])
Z = sum_i omega[i] * raw[i] / sum_i omega[i]
m[i] = raw[i] / Z

for each solver field call i:
    v = physical_velocity(f(state, tau[i]))
    v = apply_CFG_if_used(v)
    solver_update_with(m[i] * v)
```

Repeated Heun calls retain their own quadrature weights. For a one-interval Euler solver, $Z = w(\tau_0)$ and $m_0 = 1$.

I Official-checkpoint calibration and fixed-pair adaptation

We separately test the released pMF-B/16 Torch inference checkpoint (Lu et al., 2026a;b) without TSS or further training. Both new cells use the official model, sampler and scorer, 50,000 balanced-class images, seed 42, eight H100 ranks across two nodes, batch 64 per rank, FP32 with TF32, and JiT's ImageNet-256 reference statistics. We record Torch 2.7.1+cu128, torchvision 0.22.1+cu128, NumPy 1.22.3 and torch-fidelity 0.4.0b0; the author's complete software stack is not specified. Only the two sampling conditions below were prespecified; no guidance search was performed.

The original-setting cell remains outside the frozen 3.15 ± 0.10 calibration range. The separate condition diagnostic brings both FID and IS near the reported values in our Torch environment by changing guidance strength and interval together. Their individual effects and the author's historical settings remain unresolved. Original-setting FID changes by only -0.0003 from the earlier A100 run. This calibrates released-checkpoint sampling; the JiT students retain their own architecture, target, guidance and training budget. The two-cell allocation cost was 1.4387 H100 GPU-hours, including loading, scoring and idle ranks; no evaluation retry was required.

Table 37: Calibration of the unmodified official pMF checkpoint. Both original-setting cells use $\omega = 7.5$, interval $[0.1, 0.8]$. The diagnostic jointly changes these to 8.5 and $[0.1, 0.7]$ under the same Torch weights and new execution environment.

Setting	FID50K	IS
Historical: 4 A100 ranks, batch32	3.6070	282.6003
Original: 8 H100 ranks, batch64	3.6067	282.6108
Prespecified condition diagnostic	3.1543	252.9287
Torch README (reported)	3.15	252.9

Table 38: Released pMF-B/16 network adapted with the boundary-u objective on fixed pairs. All rows are FID50K from one matched training stream. The common unmodified one-call baseline is 3.1543. $\Delta = \text{TSS-X}_{10} - \text{Vanilla}_{10}$; positive differences favor vanilla data. A zero-update two-call reference was not evaluated.

Updates	Weights	Calls	Vanilla_{10}	TSS-X_{10}	Δ
250	Online	1	38.7065	38.9780	$+0.2715$
250	EMA	1	3.4177	3.3815	-0.0362
1,000	Online	1	40.4878	46.8905	$+6.4026$
1,000	EMA	1	10.7802	10.5000	-0.2802
2,000	Online	1	32.2100	40.5678	$+8.3578$
2,000	EMA	1	22.7820	24.7594	$+1.9774$
2,000	Online	2	35.3651	43.6446	$+8.2795$
2,000	EMA	2	28.4624	30.1875	$+1.7252$

I.1 Fixed-pair adaptation of the released pMF generation network

We test the complete released pMF-B/16 generation network with its mature weights. Both arms start from the same checkpoint and use the original Vanilla_{10}/TSS-X_{10} 200k floating-point pairs and their generating noise. Student inputs follow the base interpolant. Sampling uses the official Torch network with fixed $\omega = 8.5$ and interval $[0.1, 0.7]$. Zero-update one-call FID50K is 3.154285, agreeing with the prespecified condition diagnostic above. The original-setting O0 calibration failure remains recorded separately.

We adapt this network with the external boundary-u JVP objective used for the JiT consumer: with noise at $t = 1$ and data at $t = 0$, the tangent is stopped $u_\theta(z_t, t, t)$, and the correction differentiates t with r fixed. The released inference weights omit the auxiliary v branch. Our adaptation uses fresh AdamW and omits the official auxiliary and perceptual losses. It retains pMF's fixed guidance tokens, whose semantics differ from the teacher's CFG 2.5.

Both arms use 2,000 updates, batch 256, microbatch 8, learning rate 10^{-6} with 100-update warmup, AdamW betas $(0.9, 0.95)$, zero weight decay and EMA 0.999. The time sampler, adaptive weighting and denominator floor match the internal boundary-u consumer. Training is BF16 with FP32 master weights; evaluation is FP32 with TF32, official per-image seed 42 noise, balanced labels and the same JiT reference statistics and Inception implementation. Training randomness is matched across arms, with sample/time bases 91,000,000/92,000,000. A separate engineering test checks official sampling parity, the actual-network JVP against finite differences and all trainable gradients before formal training. All 17 evaluation cells retain 50,000 feature vectors.

The frozen terminal one-call EMA contrast reverses the JiT-consumer result: TSS-X_{10} is worse by 1.9774, and both arms worsen relative to the mature initial model by $19.6277/21.6051$. All measured adapted one-call values are worse than the common baseline. The small earlier EMA advantages reverse at the terminal comparison; Online results already deteriorate at 250 updates. Both runs complete 2,000 effective updates with finite logged losses and gradients despite the degradation in sample quality. Coupling, objective, optimization, precision and conditioning are possible causes that remain unresolved. This recipe gives a negative result for mature-model adaptation alongside the positive JiT-consumer results.

Total allocation is 6.6950 H100 NVL GPU-hours, including a 20-second failed finite-difference test, its isolated recovery, the shared baseline, both formal arms and all evaluation/IO. Logged training is a 5.4516-hour subset. No training or evaluation retry, additional teacher rollout, guidance search or budget extension follows this result.

J EXPERIMENTAL PROVENANCE AND OPEN EVALUATION SCOPE

The primary JiT-B/16 aggregate is reconstructed from nine completed 64-H100 training runs and 18 raw 50,000-image evaluation cells. The cells were generated in nine 16-H100 evaluation jobs. An independent remote audit and a local recomputation both record status `ok`. The DPM++2M sweep reuses three matched training pairs with a separate common evaluation stream. The full-budget LightningDiT study contains nine 20k-update continuations on two A100 GPUs each and 60 FID50K cells evaluated on H100, including six source-checkpoint references. Its allocations total 21.7378 A100 and 7.2289 H100 GPU-hours, including an initial failed short test. These hardware-specific totals are not interchangeable.

The earlier 625-update LightningDiT factorial contains 12 continuations and 12 matched 50,000-image 10-point Euler cells (9 field evaluations); every arm in every seed was trained and evaluated on NVIDIA A100 80GB GPUs. Its secondary 5-, 20-, and 50-point grid rows use six 10,000-image cells per grid and are interpreted only through within-seed paired contrasts. They are not pooled with the full-budget experiment.

The original cross-pipeline S_{norm}evidence consists of three exact paired 50,000-image comparisons; the later JiT additivity and solver comparisons are described separately in Section A.4.3. Within each row of Table 14, the checkpoint, initial noise, class labels, seed, solver grid, precision, CFG setting, and evaluator are fixed; only the declared field multiplier changes. The JiT experiment uses 19 CFG-combined field calls per image, while the two 10-point Euler experiments use nine calls. Each arm was generated and scored separately rather than reusing features or samples across arms.

Three further evaluations would extend the present evidence:

- Full transport-consistent TSS under the locked JiT-B/16 continuation protocol, to compare it with TSS-X as a common default.
- A replicated tangent-only result beyond LightningDiT-B/2 or its native-velocity head, to test the empirical reach of the adapter equations.
- Replicated TSS benefits in an official average-velocity pipeline or a different student architecture, or benefits from S_{norm}-shaped teacher trajectories. Current JVP evidence covers two student streams for the original JiT teacher pair and one for a second teacher training seed (Appendix A.4.10), with shared prior indices and evaluation noise. Flow-map performance under additional architectures or trajectory interfaces remains an empirical question.