MOMENT Z-FLOW: LEARNING PATH MOMENTS FOR ONE-STEP GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

One-step generation requires predicting finite-time transport in a single network evaluation. We introduce Moment Z-Flow, which jointly learns normalized temporal moments of a velocity path using local differential supervision and cocycle consistency across intervals. A normalized generating-function readout combines these moments into a sampling update. Its complex parameter controls temporal amplitude and phase, while the deployed computation remains real-linear and preserves constant velocity. Our analysis relates local prediction error, interval consistency, and readout bias. At 100k updates on CIFAR-10, Moment Z-Flow achieves one-step FID-50K 3.54, compared with 3.84 for the MeanFlow reference. Across two training seeds, cocycle supervision reduces mean FID from 4.04 to 3.61. On ImageNet-256, the model achieves one-step FID-50K 2.44 after 400k updates. These results support joint path-moment prediction and interval-consistent training for fast generative transport.

1 INTRODUCTION

Generative transport connects noise to data through a continuous evolution, but a sampler can access that evolution only through a finite number of network evaluations. Diffusion and flow models usually learn local dynamics and recover a sample through repeated numerical updates (Ho et al., 2020; Song et al., 2021; Lipman et al., 2023). When the inference budget is reduced to a single evaluation, the network must instead predict a quantity that summarizes a finite interval. This makes the choice of *what to predict* central to one-step generation.

Existing approaches offer several answers. Consistency models learn endpoint mappings, shortcuts learn interval-conditioned displacements, and flow map models learn transport between two times (Song et al., 2023; Frans et al., 2025; Boffi et al., 2025a). MeanFlow expresses this transport through an interval-average velocity (Geng et al., 2025a); distillation provides another route to learning a fast sampler from an existing model (Salimans & Ho, 2022). We study a complementary representation question: *can jointly predicting several temporal summaries of a path improve the sample obtained from one network evaluation?*

An exact interval average already determines the endpoint, so additional moments are not needed to reconstruct the exact displacement. Their potential value arises in learning. A finite network need not approximate different path summaries with the same errors. Predicting them jointly provides auxiliary supervision and a family of outputs that can be combined at inference. The central hypothesis is that this joint representation can support a more effective readout than its average-velocity component alone.

We introduce *Moment Z-Flow*, which predicts normalized temporal moments of velocity along a transport characteristic. Higher moments emphasize different portions of the same interval. Although their definitions involve path integrals, every moment satisfies a local differential identity involving a first material derivative. We use this hierarchy to construct a simulation-free training surrogate with conditional data–noise pairs and Jacobian–vector products. A single real-valued backbone predicts the entire moment vector.

A normalized generating-function readout converts this vector into a sampling update. The zeroth-moment readout recovers the interval-average prediction; a nonzero transform parameter activates the higher moments without another network evaluation. The complex parameter packages the

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.

moments into an analytic transform with a weighted cocycle under interval composition, while specifying real coefficients for sampling. Our analysis shows that the readout preserves constant velocity and separates its error into estimation error and a temporal-reweighting bias. This distinction explains why the readout can help a learned model without improving an already exact interval average.

At 100k updates on CIFAR-10, Moment Z-Flow achieves one-step FID 3.54 versus 3.84 for MeanFlow, and the sCM extension reaches 3.58. Across two seeds, cocycle supervision reduces mean FID from 4.04 to 3.61. On ImageNet-256, the model achieves one-step FID 2.44 at 400k updates.

The contributions are threefold: a path-moment representation connecting local differential supervision to cocycle consistency across intervals; a normalized one-call readout with explicit preservation, approximation, and stability properties; and evaluations of the combined method, including paired readout comparisons across training runs.

2 Related work

Diffusion, flows, and finite-time maps. Diffusion models and score-based SDEs generate samples through iterative denoising or probability-flow ODEs (Ho et al., 2020; Song et al., 2021); EDM separates training, preconditioning, and sampling choices (Karras et al., 2022). Flow Matching, Rectified Flow, and stochastic interpolants learn transport from tractable conditional paths (Lipman et al., 2023; Liu et al., 2023; Albergo et al., 2025). Flow map matching directly models finite-time transport (Boffi et al., 2025a). Its self-distillation extension organizes objectives into Eulerian, Lagrangian, and progressive families (Boffi et al., 2025b), while Align Your Flow scales continuous-time flow-map distillation (Sabour et al., 2025). These works already establish two-time parameterizations and local map identities; our focus is the additional representation supplied by weighted path-velocity integrals.

Consistency and shortcut learning. Consistency models map points along a trajectory to a common endpoint (Song et al., 2023). Improved consistency training and continuous-time sCM refine the objective, parameterization, and stabilization of this construction (Song & Dhariwal, 2024; Lu & Song, 2025). Consistency Trajectory Models learn maps between arbitrary trajectory times (Kim et al., 2024); Shortcut Models condition on step size and combine local supervision with self-consistency (Frans et al., 2025). Together, these methods motivate endpoint and interval-conditioned representations. Their training costs depend on initialization as well as the consistency objective; published sCT experiments use pretrained diffusion initialization.

Distribution matching versus path moments. Generative Moment Matching Networks use maximum mean discrepancy to match generated and data distributions (Li et al., 2015). Inductive Moment Matching bootstraps distribution matching across interpolant times for few-step generation (Zhou et al., 2025). Their distributional moments differ from our temporal moments $m_k = \int_0^1 \tau^k w(\tau)\, d\tau$ along a characteristic. Progressive distillation compresses a pretrained sampler (Salimans & Ho, 2022); DMD and DMD2 instead optimize distribution matching with teacher-derived score information (Yin et al., 2024a;b). Recent Mean Flow Distillation uses averaged velocities for distilling a pretrained flow model (Zhao et al., 2026). These teacher-based settings differ from our from-scratch CIFAR-10 moment training.

MeanFlow objectives and training dynamics. MeanFlow learns interval-average velocity through a local differential identity (Geng et al., 2025a). iMF revisits the regression objective and guidance conditioning (Geng et al., 2025b); Modular MeanFlow explores gradient modulation and warmup (You et al., 2025); Kim et al. (2026) study interactions between instantaneous and average-velocity learning and propose a progressive training strategy. These alternatives make training protocol an essential control for assessing additional moment heads. The local moment component below retains the original MeanFlow-style conditional-velocity JVP direction.

Higher-order information and readout. HOMO studies high-order supervision for shortcut diffusion (Chen et al., 2025), and high-order MeanFlow introduces average-acceleration fields (Cao et al., 2025). Our moment index denotes powers of normalized time weighting velocity, rather than

successive state derivatives; each local moment identity uses a first material derivative. The distinction we investigate is the use of jointly predicted temporal integrals and their fixed readout. For fixed complex parameters, our readout is real-linear; our ablations examine readout selection and cocycle supervision separately.

3 MOMENT Z-FLOW

Moment Z-Flow combines path-moment prediction, local and cocycle supervision, and a normalized sampling readout. Figure 1 summarizes these components.

3.1 REPRESENTING FINITE-TIME TRANSPORT

Let $x \sim p_{\text{data}}$ and $\epsilon \sim \mathcal{N}(0, I)$ be independent. For $0 \le r \le t \le 1$, we use the noise-increasing convention

$$z_t = (1-t)x + t\epsilon, \qquad v_c = \epsilon - x, \qquad v_\star(z,t) = \mathbb{E}[v_c \mid z_t = z]. \tag{1}$$

Data lie at $t = 0$, noise at $t = 1$, and sampling proceeds toward smaller times. Assume that $v_\star$ is continuous in time, locally Lipschitz in space, and sufficiently differentiable for the material derivatives below to exist. For a characteristic X_s satisfying $\dot{X}_s = v_\star(X_s, s)$ and an interval $h = t - r > 0$, define

$$m_k(X_t, r, t) = \frac{1}{h^{k+1}} \int_r^t (s-r)^k v_\star(X_s, s)\, ds = \int_0^1 \tau^k w(\tau)\, d\tau, \tag{2}$$

where $w(\tau) = v_\star(X_{r+h\tau}, r+h\tau)$. These are temporal moments along a marginal characteristic, distinct both from constant-velocity conditional interpolants and from distributional moment matching (Zhou et al., 2025).

Proposition 1 (Local moment hierarchy)**.** *With r fixed and $\mathcal{D}_t = \partial_t + v_\star \cdot \nabla_z$,*

$$(k+1)m_k + h\mathcal{D}_t m_k = v_\star(X_t, t), \qquad k \ge 0. \tag{3}$$

At the diagonal, $m_k(X_t, r, t) \to v_\star(X_r, r)/(k+1)$ as $t \downarrow r$. In particular, $X_r = X_t - h m_0(X_t, r, t)$.

The result follows by differentiating $h^{k+1} m_k$ along the characteristic. Its zeroth-order case is the MeanFlow identity. Appendix A gives the derivation and a boundedness condition that excludes a singular homogeneous solution.

3.2 LEARNING MOMENTS THROUGH LOCAL IDENTITIES

In the MeanFlow-based CIFAR-10 implementation, a single SongUNet backbone, using the diffusion-network implementation of Karras et al. (2022), outputs $m_{0,\theta}$ and higher residual channels $\rho_{k,\theta}$. This implementation uses orders $k = 0, \dots, K$ with $K = 3$ and

$$m_{k,\theta} = \frac{m_{0,\theta}}{k+1} + h\rho_{k,\theta}, \quad 1 \le k \le K. \tag{4}$$

Provided the residual outputs remain bounded, this parameterization enforces the required ratios between heads as $h \to 0$. Learning the boundary velocity itself remains the task of $m_{0,\theta}$. The backbone is conditioned on (t, h).

At a conditional interpolant z_t, we replace the unknown marginal velocity in the derivative direction by v_c. All time derivatives below hold r fixed:

$$\begin{aligned} \widehat{\mathcal{D}}_t m_{k,\theta} &= \partial_t m_{k,\theta}|_r + (\nabla_z m_{k,\theta}) v_c, \\ y_k &= \mathrm{sg}\left[\frac{v_c - h\widehat{\mathcal{D}}_t m_{k,\theta}}{k+1}\right], \qquad e_k = \|m_{k,\theta} - y_k\|^2_{\text{mean}}. \end{aligned} \tag{5}$$

Here $\|\cdot\|^2_{\text{mean}}$ averages squared residuals over image coordinates for each example, and sg denotes stop-gradient. The expectation below is over data, noise, and sampled intervals. At update j, the local moment loss is

$$\mathcal{L}_{\text{moment},j} = \mathbb{E}\left[\frac{e_0}{(\mathrm{sg}(e_0) + \delta)^p} + \frac{\lambda_j}{K}\sum_{k=1}^{K} \frac{e_k}{(\mathrm{sg}(e_k) + \delta)^p}\right], \quad p = 0.75, \quad \delta = 10^{-3}, \tag{6}$$

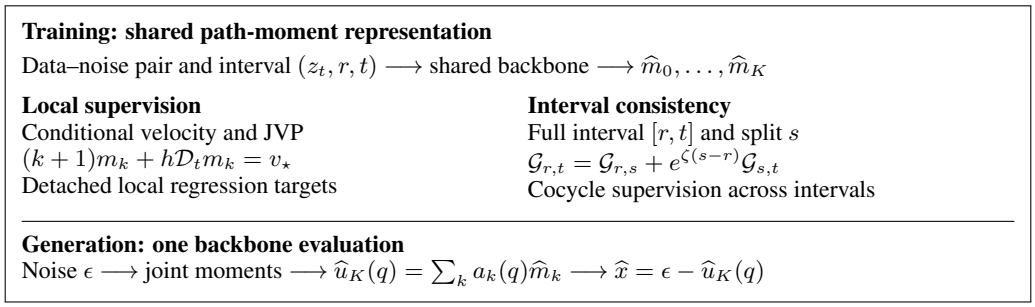


Figure 1: Moment Z-Flow combines local moment supervision and cocycle consistency during training. The cocycle shown is the exact structural identity; training uses the detached EMA construction in Sec. 3.4. Generation uses one joint prediction and a normalized real-linear readout.

where λ_j increases to 1 over the first 5,000 updates. For $K = 0$, the higher-order sum is omitted. Differentiation holds r fixed: the JVP direction is $(v_c, 0, 1)$ for an interface (z, r, t), or $(v_c, 1, 1)$ for (z, t, h). Treating h as fixed would implement a different derivative.

Equation (3) characterizes exact moments; Eqs. (5)–(6) specify the local moment component of training. In unweighted regression, conditional averaging recovers the marginal JVP direction at a fixed network. Appendix D derives how adaptive weighting and target detachment affect the conditional regression update.

3.3 From joint predictions to a one-step sample

Let $\widehat{m}_0, \ldots, \widehat{m}_K$ denote the network's moment predictions. We combine them using a truncated exponential generating function. Its parameter $q \in \mathbb{C}$ defines a smooth family of coefficients anchored at the zeroth-moment prediction when $q = 0$. To preserve constant velocity, we normalize the transform:

$$d_K(q) = \sum_{k=0}^{K} \frac{q^k}{k!(k+1)}, \qquad \widehat{u}_K(q) = \mathrm{Re}\left[\frac{\sum_{k=0}^{K} q^k \widehat{m}_k / k!}{d_K(q)}\right], \quad d_K(q) \neq 0. \tag{7}$$

The one-step sample is $\widehat{x} = \epsilon - \widehat{u}_K(q)$, evaluated at $(r, t) = (0, 1)$. For a shorter interval the displacement is $h\widehat{u}_K(q)$; the multi-step diagnostics state their interval-specific readout conventions separately. For fixed q, the coefficients can be computed once; sampling then requires only a weighted sum of the joint outputs. We refer to $q = 0$ as the center readout and to the chosen nonzero q as the active readout.

3.4 Complex generating functions and the cocycle structure

The complex coordinate gives the moment representation a compositional interpretation. Along one characteristic, define the unnormalized exponential transform

$$\mathcal{G}_{r,t}(\zeta) = \int_r^t e^{\zeta(u-r)} v_\star(X_u, u)\, du = \sum_{k=0}^{\infty} \frac{\zeta^k h^{k+1}}{k!} m_k(X_t, r, t). \tag{8}$$

For $r < s < t$, splitting the integral yields the weighted cocycle

$$\mathcal{G}_{r,t}(\zeta) = \mathcal{G}_{r,s}(\zeta) + e^{\zeta(s-r)} \mathcal{G}_{s,t}(\zeta). \tag{9}$$

The subintervals follow the same characteristic, with intermediate state X_s. At $\zeta = 0$, this is displacement additivity. For $\zeta = \alpha + i\omega$, the exponential separates into an amplitude weight $e^{\alpha(u-r)}$ and a phase factor $e^{i\omega(u-r)}$. The imaginary coordinate therefore supplies oscillatory temporal weighting within the same representation; the cocycle itself also holds for real ζ.

With $q = \zeta h$, the numerator of Eq. (7) is the degree-K approximation to $\mathcal{G}_{r,t}(q/h)/h$. This connects local moment supervision, interval composition, and the sampling readout. Z-Flow uses cocycle consistency during training alongside the local moment component in Eq. (6). The identity above describes the exact path transform and provides the mathematical basis for interval consistency. Appendix D.7 proves associativity, coefficient recovery, and finite-order error bounds.

Coefficient-level training objective. Cocycle supervision uses the coefficient residual $C_{k,\theta}$ in Appendix D.8, with an EMA teacher on the composed branch. In training, $X_t = z_t$ is the conditional interpolant from Eq. (1); no exact characteristic is simulated. Sample $s = r + \xi(t-r)$ with $\xi \sim \mathcal{U}[0.1, 0.9]$ and construct $\widetilde{X}_s = \mathrm{sg}[X_t - (t-s)m_{0,\bar{\theta}}(X_t, s, t)]$. The student predicts the full-interval moments; both subinterval predictions use the detached EMA teacher. Let $T_{k,\bar{\theta}}$ be the composed right-hand side of Eq. (44), divided by h^{k+1} and evaluated at this intermediate state. The corresponding objective is

$$\mathcal{L}_{\mathrm{coc}} = \frac{1}{K+1}\sum_{k=0}^{K} \mathbb{E}\big\|(k+1)\big[m_{k,\theta}(X_t, r, t) - \mathrm{sg}(T_{k,\bar{\theta}})\big]\big\|^2_{\mathrm{mean}}, \tag{10}$$

$$\mathcal{L}_{\mathrm{total},j} = \mathcal{L}_{\mathrm{moment},j} + \gamma_j \mathcal{L}_{\mathrm{coc}}, \qquad \gamma_j = 0.05\min\{j/5000, 1\}. \tag{11}$$

The factor $k+1$ compensates for the diagonal scale of each moment. We use teacher decay 0.9999 and one split per example, skipping diagonal intervals for the cocycle term. All student heads receive gradients through the full-interval branch; teacher and intermediate-state branches are detached. Equation (11) gives the MeanFlow-based objective. The sCM extension retains its native consistency objective and adds higher-moment and cocycle supervision; initialization and evaluation settings are given in Appendix B.

4 Readout Properties

We analyze exact preservation, prediction error, and the effect of truncation order.

4.1 Real equivalence and constant-velocity preservation

Proposition 2 (Real readout)**.** *For real-valued moment tensors and fixed q with $d_K(q) \neq 0$, define*

$$a_k(q) = \mathrm{Re}\left[\frac{q^k}{k! d_K(q)}\right]. \tag{12}$$

Then $\widehat{u}_K(q) = \sum_{k=0}^{K} a_k(q)\widehat{m}_k$ and $\sum_{k=0}^{K} a_k(q)/(k+1) = 1$. Thus the readout returns v whenever all $\widehat{m}_k = v/(k+1)$.

Linearity of the real part proves the result. For $K = 3$ and $q = -1+0.5i$, the coefficients are approximately $(1.546168, -1.369276, 0.402921, 0.016652)$. Complex notation parameterizes a restricted family of real coefficient vectors; it does not enlarge the fixed sampler's arithmetic expressivity. A comparison of real q and complex q consequently tests different real weights, not whether complex arithmetic is necessary.

4.2 Estimation error and temporal bias

Let $\widehat{m}_k = m_k + \eta_k$. At a fixed interval, the error relative to the true average velocity decomposes exactly as

$$\widehat{u}_K(q) - m_0 = \underbrace{\sum_{k=0}^{K} a_k(q)\eta_k}_{\text{moment-estimation error}} + \underbrace{\int_0^1 \left(\sum_{k=0}^{K} a_k(q)\tau^k - 1\right) w(\tau)\, d\tau}_{\text{temporal-reweighting bias}}. \tag{13}$$

The reweighting kernel integrates to zero after subtracting 1, so the second term vanishes for constant velocity. It need not vanish otherwise, even for perfect moments. This explains why a nonzero-q readout is a calibration rule rather than a guaranteed improvement to exact integration. Finite-model errors could interact favorably with this bias, but our FID measurements do not identify that mechanism.

For any constant vector $\bar{v}$, the triangle inequality gives

$$\|\widehat{u}_K - m_0\| \leq \sum_{k=0}^{K} |a_k|\,\|\eta_k\| + \int_0^1 \left|\sum_{k=0}^{K} a_k\tau^k - 1\right| \|w(\tau) - \bar{v}\|\, d\tau. \tag{14}$$

Multiplying by h bounds endpoint error from the same X_t; one-step generation uses $h = 1$. Small $|d_K(q)|$ amplifies estimation error through large coefficients. Appendix D analyzes denominator stability.

4.3 The role of truncation order

For bounded velocity $\|w(\tau)\| \le V$, the exact exponential transform is

$$Z(q) = \int_0^1 e^{q\tau} w(\tau)\, d\tau = \sum_{k=0}^{\infty} \frac{q^k m_k}{k!}, \qquad \left\| Z(q) - \sum_{k=0}^{K} \frac{q^k m_k}{k!} \right\| \le \frac{V e^{|q|} |q|^{K+1}}{(K+1)!}. \tag{15}$$

Increasing K controls approximation of $Z(q)$, while moment-estimation error and nonzero-q reweighting bias remain distinct. Appendix D gives normalized remainder bounds.

5 Experiments

Our evaluation separates the effect of joint training from the effect of the readout. Base-method pairs compare training recipes; within-checkpoint ablations isolate the output combination; cost and structural controls qualify the interpretation.

5.1 Experimental setup

We study unconditional CIFAR-10 (Krizhevsky, 2009) at 32×32 resolution. The main moment model starts from random initialization and trains for $100{,}000$ updates with global batch $1{,}024$ on eight H100 GPUs. It has $55{,}953{,}164$ parameters; the native MeanFlow reference has $55{,}932{,}419$. Both follow the native MeanFlow recipe, with the moment model adding moment outputs, local moment supervision, and cocycle consistency. All Z-Flow variants, including MeanFlow + Z-Flow, sCM + Z-Flow, and the ImageNet model, use the same cocycle settings. We abbreviate Moment Z-Flow as Z-Flow in tables and figure labels. Comparisons match training updates rather than GPU-hours.

We report FID with $50{,}000$ generated images (FID-50K) (Heusel et al., 2017), KID (Bińkowski et al., 2018), and precision/recall (Kynkäänniemi et al., 2019). Evaluation uses Inception-2048 features through TorchMetrics, with the CIFAR-10 training images as the real reference; precision/recall use at most $10{,}000$ features. FID-5K is used for readout screening and is not compared numerically to FID-50K. Appendix B details optimization, evaluation, and random seeds.

Our CIFAR-10 runs use the primary EMA at their fixed 100k-update endpoint, not a selected checkpoint from a longer run. Seed 0 is the discovery run: FID-5K selects the readout, and the subsequent FID-50K stream is not independent of screening. Seed 1 includes a distinct confirmation noise set. We do not pool those evaluations into a training-population confidence interval.

Table 2 compares MeanFlow with its Z-Flow extension at matched updates and NFE, but unequal training compute. For the MeanFlow pair, "+ Z-Flow" denotes from-scratch joint $K = 3$ moment training with cocycle consistency and readout $q = -1 + 0.5i$.

5.2 Consistency-model comparison at 100k

Table 1 includes the 100k consistency-training (sCT) result from the sCM framework, Appendix E.3, Table 3 of Lu & Song (2025). It uses DDPM++ without AdaGN, dropout 0.20, and pretrained diffusion initialization. Its 100k count covers consistency training, not pretraining.

Both sCM conditions use 100k consistency-stage updates at batch 128, excluding diffusion pretraining, whereas each MeanFlow-based run processes 102.4 million examples. The sCM extension obtains FID 3.58; cross-family update counts therefore should not be interpreted as equal training cost.

Table 1: CIFAR-10 FID at 100k updates. The MeanFlow rows form a matched-update, one-step comparison. The lower block gives additional consistency-model results: published sCT from Lu & Song (2025), Table 3, and sCM + Z-Flow. Both sCM conditions use pretrained diffusion initialization and batch 128; the published baseline is not a new reproduction.

Method	Initialization	Batch	FID ↓
MeanFlow	Random	1024	3.84
MeanFlow + Z-Flow	Random	1024	3.54
sCM: sCT (published)	Pretrained diffusion	128	3.97
sCM + Z-Flow	Pretrained diffusion	128	3.58

Table 2: Seed 0 at 100k updates and NFE$=1$. Z-Flow rows share a checkpoint and evaluation protocol, differing in readout. KID is unscaled; dashes denote unavailable metrics. Training compute is not matched to MeanFlow.

Model / readout	FID-50K ↓	KID ↓	Precision ↑	Recall ↑
Native MeanFlow	3.8422	0.001497	0.6918	0.5808
Z-Flow center	4.3304	0.002160	0.6936	0.5648
Z-Flow, $q = -1$	3.7986	0.001371	0.7100	0.5625
Z-Flow, $q = -1 + 0.5i$	3.54	—	—	—

5.3 Isolating the contribution of the readout

The active readout achieves FID 3.54, compared with 4.3304 for the same checkpoint's center and 3.8422 for native MeanFlow: reductions of 18.3% and 7.9%, respectively. All FID evaluations use the same evaluator, 50k sample count, reference statistics, and preprocessing. A numerical equivalence check also supports the real-linear implementation in Eq. (12) (Appendix B).

Figure 2 illustrates the three main readouts; distribution-level comparisons use Table 2.

5.4 Readout comparison across training runs

The active readout $q = -1 + 0.5i$ achieves FID 3.54 and 3.68 for seeds 0 and 1, compared with center FIDs 4.3304 and 4.4281. The respective reductions are 18.3% and 16.9% (Table 6).

5.5 Contribution of cocycle supervision

Table 3 compares training with and without cocycle supervision, retaining the moment representation and active readout. Cocycle supervision reduces mean FID from 4.04 to 3.61, a 10.6% reduction, with improvements in both seeds. Appendix E.5 relates local prediction error, interval inconsistency, and readout bias at fixed q.

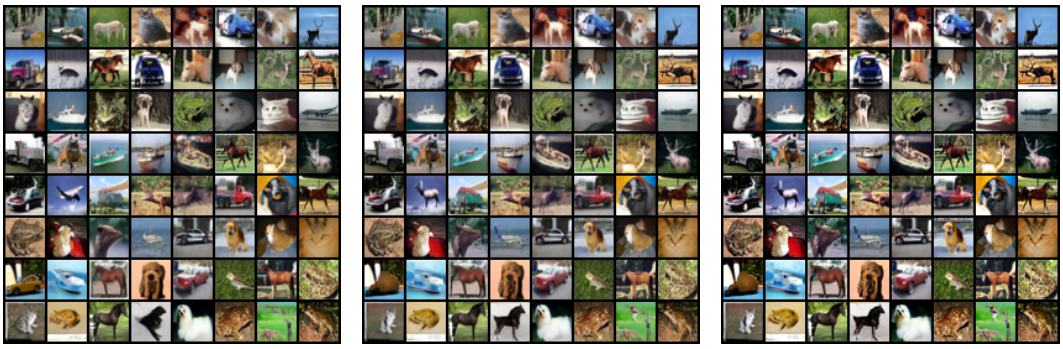
5.6 Training and inference cost

Joint moment prediction adds $20{,}745$ parameters (0.037%), with training cost $1.17\times$ that of MeanFlow (17.3% higher). Sampling uses one backbone evaluation; Table 4 summarizes the parameter counts, training cost, and measured inference latency.

5.7 ImageNet results

On ImageNet-256 (Deng et al., 2009), Z-Flow achieves a **one-step FID-50K of 2.44 after 400k training steps**. Evaluation uses 50k class-balanced samples with CFG scale 1.0 and one generation-network evaluation per image. Figure 3 shows samples from the 400k-step model. The ImageNet run uses cocycle supervision with the same cocycle settings as the CIFAR-10 Z-Flow variants. We report this result separately from the controlled CIFAR-10 readout comparisons.

Appendix B gives the ImageNet training and evaluation protocols; we do not claim a matched-protocol ranking against published ImageNet methods.



MeanFlow
FID-50K: 3.8422

Z-Flow center
FID-50K: 4.3304

Z-Flow active
FID-50K: 3.54

Figure 2: CIFAR-10 samples at NFE$=1$, with 64 images per panel and a shared visualization seed. Center and active use the same seed-0 Z-Flow checkpoint at 100k updates. The active readout uses $q = -1 + 0.5i$. FID values refer to the separate 50,000-sample evaluations in Table 2.

Table 3: Cocycle ablation on CIFAR-10 at 100k updates with one-step active readout $q = -1+0.5i$. Both conditions use the same training budget and evaluation protocol.

Training objective	Seed 0	Seed 1	Mean FID ↓
Without cocycle	4.01	4.07	4.04
With cocycle	3.54	3.68	3.61

6 Discussion and Limitations

The readout comparisons and cocycle ablation support complementary roles for interval-consistent training and output combination. They do not identify whether gains arise primarily from smaller head errors or more favorable error correlations. The center readout remains weaker than the native MeanFlow reference, consistent with the training interactions studied by Kim et al. (2026).

Our bounds concern represented fields and readouts, not convergence of the stochastic optimizer or FID. Fixed complex parameters define real coefficient vectors; their utility does not require complex arithmetic. The two-seed evaluation does not establish training-population uncertainty, and seed 0 shares evaluation noise with readout screening. Equal-GPU-hour controls, order ablations, and paired ImageNet readout comparisons would further isolate the sources and scope of improvement.

7 Conclusion

Moment Z-Flow represents finite-time transport through jointly predicted path moments. Local supervision anchors the transport, cocycle consistency relates adjacent intervals, and a normalized readout produces a one-step sample. The model achieves CIFAR-10 FID 3.54 at 100k updates and ImageNet-256 FID 2.44 at 400k updates. The two-seed cocycle ablation reduces mean CIFAR-10 FID from 4.04 to 3.61, supporting interval-consistent moment learning as a practical approach to one-step generation.

Table 4: Training and inference cost at 100k updates. GPU-hours measure training. Batch-128 latency measurements are descriptive and were not collected in an isolated hardware benchmark.

Quantity	Native MeanFlow	Z-Flow
Trainable parameters	55.932M	55.953M
Training updates	100,000	100,000
Training GPU-hours	48.15	56.48
Backbone calls per sample	1	1
Batch-128 latency	62.2 ms	62.5 ms



Figure 3: Sixteen ImageNet-256 samples from **Z-Flow trained for 400k steps**. The model achieves **one-step FID-50K 2.44**. The grid illustrates generated samples; FID is computed separately.

AI USE STATEMENT

Generative AI tools assisted with research-code design, literature search, experimental interpretation, mathematical exposition, manuscript drafting and editing, and preparation of LaTeX tables and the method schematic. Numerical results were obtained through model training and evaluation.

REPRODUCIBILITY STATEMENT

Section 3 specifies the moment parameterization, local regression targets, cocycle objective, and sampling readout. Appendix B gives training configurations and evaluation settings for CIFAR-10 and ImageNet-256. Appendices A, D, and E provide derivations and state the assumptions underlying the error and convergence bounds. The experiments distinguish fixed-checkpoint readout comparisons, cocycle ablations, and comparisons with published baselines.

REFERENCES

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. https://arxiv.org/abs/2303.08797.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. https://arxiv.org/abs/1801.01401.

Nicholas M. Boffi, Michael S. Albergo, and Eric Vanden-Eijnden. Flow map matching with stochastic interpolants: A mathematical framework for consistency models. *Transactions on Machine Learning Research*, 2025a. https://arxiv.org/abs/2406.07507.

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*, 2025b. https://arxiv.org/abs/2505.18825.

Yang Cao, Yubin Chen, Zhao Song, and Jiahao Zhang. Towards high-order mean flow generative models: Feasibility, expressivity, and provably efficient criteria. *arXiv preprint arXiv:2508.07102*, 2025. https://arxiv.org/abs/2508.07102.

Bo Chen, Chengyue Gong, Xiaoyu Li, Yingyu Liang, Zhizhou Sha, Zhenmei Shi, Zhao Song, and Mingda Wan. High-order matching for one-step shortcut diffusion models. *arXiv preprint arXiv:2502.00688*, 2025. https://arxiv.org/abs/2502.00688.

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*, 2009. https://www.image-net.org/static_files/papers/imagenet_cvpr09.pdf.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *International Conference on Learning Representations*, 2025. https://arxiv.org/abs/2410.12557.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025a. https://arxiv.org/abs/2505.13447.

Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J. Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. *arXiv preprint arXiv:2512.02012*, 2025b. Version 2, 2026. https://arxiv.org/abs/2512.02012.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017. https://arxiv.org/abs/1706.08500.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, 2020. https://arxiv.org/abs/2006.11239.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the Design Space of Diffusion-Based Generative Models. In *Advances in Neural Information Processing Systems*, 2022. https://arxiv.org/abs/2206.00364.

Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency Trajectory Models: Learning Probability Flow ODE Trajectory of Diffusion. In *International Conference on Learning Representations*, 2024. https://arxiv.org/abs/2310.02279.

Jin-Young Kim, Hyojun Go, Lea Bogensperger, Julius Erbach, Nikolai Kalischek, Federico Tombari, Konrad Schindler, and Dominik Narnhofer. Understanding, Accelerating, and Improving MeanFlow Training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. https://arxiv.org/abs/2511.19065.

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

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, 2019. https://arxiv.org/abs/1904.06991.

Yujia Li, Kevin Swersky, and Rich Zemel. Generative Moment Matching Networks. In *International Conference on Machine Learning*, 2015. https://arxiv.org/abs/1502.02761.

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. https://arxiv.org/abs/2210.02747.

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. https://arxiv.org/abs/2209.03003.

Cheng Lu and Yang Song. Simplifying, Stabilizing and Scaling Continuous-Time Consistency Models. In *International Conference on Learning Representations*, 2025. https://arxiv.org/abs/2410.11081.

William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. *arXiv preprint arXiv:2212.09748*, 2023. https://arxiv.org/abs/2212.09748.

Amirmojtaba Sabour, Sanja Fidler, and Karsten Kreis. Align Your Flow: Scaling Continuous-Time Flow Map Distillation. In *Advances in Neural Information Processing Systems*, 2025. https://arxiv.org/abs/2506.14603.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. https://arxiv.org/abs/2202.00512.

Yang Song and Prafulla Dhariwal. Improved Techniques for Training Consistency Models. In *International Conference on Learning Representations*, 2024. https://arxiv.org/abs/2310.14189.

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. https://arxiv.org/abs/2011.13456.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, 2023. https://proceedings.mlr.press/v202/song23a.html.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T. Freeman, and Taesung Park. One-step Diffusion with Distribution Matching Distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a. https://arxiv.org/abs/2311.18828.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T. Freeman. Improved Distribution Matching Distillation for Fast Image Synthesis. In *Advances in Neural Information Processing Systems*, 2024b. https://arxiv.org/abs/2405.14867.

Haochen You, Baojing Liu, and Hongyang He. Modular MeanFlow: Towards Stable and Scalable One-Step Generative Modeling. *arXiv preprint arXiv:2508.17426*, 2025. https://arxiv.org/abs/2508.17426.

An Zhao, Shengyuan Zhang, Zhongjian Sun, Yixiang Zhou, Zejian Li, Ling Yang, Tianrun Chen, and Lingyun Sun. Mean Flow Distillation: Robust and Stable Distillation for Flow Matching Models. *arXiv preprint arXiv:2606.11155*, 2026. https://arxiv.org/abs/2606.11155.

Linqi Zhou, Stefano Ermon, and Jiaming Song. Inductive Moment Matching. In *International Conference on Machine Learning*, 2025. https://arxiv.org/abs/2503.07565.

A Derivations and scope of the identities

A.1 Moment hierarchy and the diagonal boundary

Fix r and a sufficiently regular characteristic X_s. Let $I_k(t) = \int_r^t (s-r)^k v_\star(X_s, s)\, ds$. The fundamental theorem of calculus yields $I_k'(t) = h^k v_\star(X_t, t)$. Since $I_k = h^{k+1} m_k$, the product rule gives

$$(k+1)h^k m_k + h^{k+1}\mathcal{D}_t m_k = h^k v_\star(X_t, t). \tag{16}$$

Dividing by h^k proves Eq. (3). Changing variables to $s = r + h\tau$ gives Eq. (2); continuity and local boundedness permit the limit

$$\lim_{h\to 0} m_k = \int_0^1 \tau^k v_\star(X_r, r)\, d\tau = \frac{v_\star(X_r, r)}{k+1}. \tag{17}$$

At $k = 0$, integration of the characteristic gives $X_t - X_r = h m_0$.

Conversely, along a fixed characteristic a solution of the local hierarchy satisfies $\frac{d}{dt}(h^{k+1} m_k) = h^k v_\star(X_t, t)$. If m_k remains bounded as $h \to 0$, the integration constant is zero and the integral definition follows. Without such a boundary condition, a term $Ch^{-(k+1)}$ solves the homogeneous equation and need not represent a path moment. The learned parameterization imposes the ratios at the diagonal when its residual channels remain bounded. It does not impose the exact characteristic differential equation away from the diagonal.

A.2 Finite generating-polynomial identity

Define $F_K(q) = \sum_{k=0}^K m_k q^k$. Summing the hierarchy yields

$$h\mathcal{D}_t F_K(q) + \partial_q\big(qF_K(q)\big) = v_\star(X_t, t) \sum_{k=0}^K q^k. \tag{18}$$

Holomorphy in q is automatic because F_K is a polynomial. It does not impose an additional Cauchy–Riemann training constraint on the image backbone. For example, on any auxiliary circle of radius $R > 0$,

$$m_0 = F_K(0) = \frac{1}{2\pi}\int_0^{2\pi} F_K(Re^{i\phi})\, d\phi. \tag{19}$$

The contour average extracts the zeroth coefficient. This ordinary generating polynomial differs from the factorial-weighted exponential transform used by the sampler.

A.3 Readout equivalence and error decomposition

Because each $\widehat{m}_k$ is real, taking the real part of Eq. (7) commutes with multiplication by $\widehat{m}_k$ and finite summation. This proves Eq. (12). Moreover,

$$\sum_{k=0}^K \frac{a_k(q)}{k+1} = \mathrm{Re}\left[\frac{1}{d_K(q)} \sum_{k=0}^K \frac{q^k}{k!(k+1)}\right] = 1. \tag{20}$$

Writing $A_q(\tau) = \sum_{k=0}^K a_k(q)\tau^k$ gives $\int_0^1 A_q(\tau)\, d\tau = 1$. Substitution of $\widehat{m}_k = m_k + \eta_k$ and the moment integrals proves Eq. (13). Subtracting any constant $\bar{v}$ inside the zero-integral bias kernel and applying the triangle inequality gives Eq. (14). A simpler coefficient bound is

$$|a_k(q)| \le \frac{|q|^k}{k!|d_K(q)|}, \qquad \left\|\sum_k a_k \eta_k\right\| \le \frac{\sum_k |q|^k \,\|\eta_k\| / k!}{|d_K(q)|}. \tag{21}$$

These deterministic statements require no distributional assumptions on η_k. Comparing the active and center errors additionally requires their magnitudes and correlations.

A.4 A NONCONSTANT-VELOCITY EXAMPLE

For $w(\tau) = b + c\tau$ with constant vectors b, c, exact moments are $m_k = b/(k+1) + c/(k+2)$. The exact average is $m_0 = b + c/2$, whereas the active readout is

$$u_K(q) = b + c \sum_{k=0}^{K} \frac{a_k(q)}{k+2}. \tag{22}$$

Its bias is $c(\sum_k a_k/(k+2) - 1/2)$, which need not vanish. For $K = 3$, $q = -1+0.5i$, the coefficient of c is approximately 0.42072, rather than 0.5. Thus even exact higher moments do not make this nonzero-q rule an exact unweighted interval integral. This example exhibits the temporal bias of a nonzero-q readout on a linear velocity path.

A.5 TRANSFORM REMAINDER

The scalar Taylor remainder obeys

$$\left| e^{q\tau} - \sum_{k=0}^{K} \frac{(q\tau)^k}{k!} \right| \le \frac{e^{|q|\tau}(|q|\tau)^{K+1}}{(K+1)!} \le \frac{e^{|q|}|q|^{K+1}}{(K+1)!}. \tag{23}$$

Multiplying by $\|w(\tau)\| \le V$, integrating, and using uniform convergence on $\tau \in [0, 1]$ proves Eq. (15). The deployed readout also divides by $d_K(q)$. Controlling its error requires accounting for this denominator and the temporal bias, in addition to the unnormalized remainder above.

B TRAINING AND EVALUATION DETAILS

Table 5: Training configuration for the MeanFlow-based CIFAR-10 experiments.

Item	Setting
Dataset / resolution	CIFAR-10, 32×32, unconditional
Initialization	Random; no pretrained MeanFlow weights
Architecture	SongUNet with joint moment residual outputs
Moment orders	$k = 0, 1, 2, 3$
Training budget	100,000 updates; global batch 1,024
Devices / microbatch	Eight H100 GPUs; 128 examples per GPU
Optimizer	Adam, learning rate 6×10^{-4}; betas $(0.9, 0.999)$
Learning-rate warmup	9,600 updates
Dropout / augmentation	0.2 / EDM augmentation enabled
Primary EMA	Decay 0.9999; final checkpoint
Additional EMA decays	0.99995, 0.9996; primary evaluation uses 0.9999
Higher-order loss	Weight 1 after 5,000-update warmup
Adaptive normalization	$p = 0.75$, $\delta = 0.001$
Cocycle consistency	Weight 0.05; 5k linear warmup; shared across Z-Flow variants
Time sampler	Native v1; $(\mu_t, \sigma_t) = (-0.6, 1.6)$; $(\mu_r, \sigma_r) = (-4, 1.6)$; ratio 0.75
Precision / execution	FP32, no AMP; uncompiled
External pretrained teacher / simulated cache	None

The time sampler follows the endpoint and diagonal-sampling implementation of https://github.com/Gsunshine/py-meanflow, pinned to commit `1f6d72d94247c8fdeb89489acca1a8007a6baf6c`. Executed batch and precision settings are given above. Table 4 gives total training cost for each model.

CIFAR-10 evaluation uses TorchMetrics Inception-2048, with TorchMetrics 1.7.4 and torch-fidelity 0.4.0 in the standard evaluation environment. Images are mapped from $[-1, 1]$ to $[0, 1]$, clamped, and quantized by flooring to 8-bit levels for the main comparison. FID uses 50,000 generated images and the CIFAR-10 training reference. KID uses 100 subsets of 1,000 features per evaluation, and values in Table 2 are not multiplied by 1,000. Precision/recall use $k = 3$ and at most 10,000 features. Some later confirmation evaluations use CPU feature caches; comparisons use values produced under the same within-comparison protocol.

Cocycle ablation. Z-Flow uses the EMA-target cocycle loss in Sec. 3.4. The no-cocycle ablation sets its coefficient to zero while retaining moment supervision and the active readout.

sCM comparison. The sCM + Z-Flow experiment uses the same pretrained diffusion initialization, batch size 128, and one-step sampling setting as the sCT reference in Table 3 of Lu & Song (2025). The native consistency objective is retained alongside higher-moment and cocycle supervision. The 100,000 updates count the consistency-training stage and exclude diffusion pretraining. The baseline FID is taken from Lu & Song (2025).

ImageNet training and sampling. The ImageNet-256 model uses DiT-XL/2 (Peebles and Xie, 2023) with an $8\times$ VAE spatial downsampling factor. The global batch size is 256. Training uses AdamW with learning rate 10^{-4}, betas $(0.9, 0.95)$, and zero weight decay, with a 10,000-update learning-rate warmup and EMA decay 0.9999. Training runs for 400,000 updates. The cocycle loss and its schedule follow Sec. 3.4. One-step sampling achieves FID 2.44. These training settings are distinct from the CIFAR-10 SongUNet configuration above.

ImageNet evaluation protocol. We evaluate the EMA checkpoint at 400k updates using 50,000 generated images, with 50 samples for each of the 1,000 ImageNet classes and random seed 42. Sampling uses one full-interval update with $q = -1 + 0.5i$ and classifier-free guidance scale 1.0, so no unconditional prediction is mixed into the class-conditional output. Each image requires one generation-network evaluation; VAE decoding is excluded from NFE.

Training and decoding use the same `stabilityai/sd-vae-ft-mse` checkpoint and latent scaling convention. The checkpoint is available at https://huggingface.co/stabilityai/sd-vae-ft-mse. FID is computed with the ADM TensorFlow evaluation suite against the statistics distributed in `VIRTUAL_imagenet256_labeled.npz`, using the evaluator and reference batch from https://github.com/openai/guided-diffusion/tree/main/evaluations. This protocol gives the ImageNet-256 FID-50K of 2.44 reported in Sec. 5.7.

Randomness and uncertainty. Training seeds vary initialization and training randomness; evaluation seeds vary the generated noise for a fixed checkpoint. Seeds 0 and 1 provide independent training runs. Periodic and final evaluations can use different noise streams, as distinguished in Table 6.

Readout implementation. For a fixed q, compute the coefficients in Eq. (12) once and apply them to the real moment tensors. At NFE$= 1$, the interval is $(r, t) = (0, 1)$. For the same trained checkpoint, changing the readout does not retrain the backbone. The numerical equivalence check gives a maximum sample-space floating-point difference of about 1.43×10^{-6} over 4,096 samples, with 106 out of 12,582,912 quantized channel values differing because of rounding. These are numerical implementation effects, not distinct mathematical samplers.

Training procedure. For each batch, sample data, independent noise, and native interval times. Construct z_t and v_c; evaluate the joint moments and their JVP with the lower endpoint held fixed; form detached local targets; train with the local moment component in Eq. (6) and cocycle consistency; and update the primary EMA. For the local moment component, JVP outputs are detached in the target, while the prediction branch supplies the parameter gradient. Computing all heads in a single JVP operation does not remove the extra training work associated with their outputs.

C Supplementary readout results

C.1 Multi-step sampling diagnostics

With uniform grids and the interval readout $q_h = h(-1 + 0.5i)$, active Z-Flow obtains FID 3.67892 at two evaluations and 5.98954 at four, versus MeanFlow's 3.27492 and 4.50703. After limited grid/readout screening, the selected Z-Flow results are 3.17357 and 3.25698, still above MeanFlow's 3.11566 and 3.06454. The search budgets differ and the confirmation stream is not independent of screening; we report these as diagnostics, not a new controlled benchmark.

Table 6: Z-Flow FID at 100k updates and one backbone evaluation. Final evaluations share the evaluator, sample count, reference statistics, and preprocessing. Periodic measurements use separate evaluation-noise streams. Dashes denote unavailable measurements.

Training seed	Evaluation	Center	$q = -1 + 0.5i$	$q = -1$
0	Final evaluation	4.3304	3.54	3.7986
1	Periodic	4.4144	—	—
2	Periodic	4.3007	—	—
1	Final evaluation	4.4281	3.68	3.9102

Table 7: CIFAR-10 FID-50K at one backbone evaluation, using seed-0 checkpoints trained for 100k updates. Z-Flow center and active share the same checkpoint; the active readout uses $q = -1+0.5i$. Training compute is not matched to MeanFlow.

Model / readout	1 NFE FID-50K ↓
MeanFlow	3.8422
Z-Flow center	4.3304
Z-Flow active	3.54

Table 7 summarizes the one-step results. The multi-step diagnostics above indicate that one-step calibration does not automatically transfer to shorter intervals; exact moment composition does not guarantee improved composition of learned readouts.

D Additional mathematical derivations

This section develops consequences of the representation used in Section 3. All identities concern marginal characteristics unless explicitly stated otherwise. They do not assume that the trained network satisfies the identities exactly. Vector norms below are Euclidean, with their induced matrix norms; finite second moments are assumed wherever expectations of squared errors appear.

D.1 Coordinate changes and conditional regression

Let $f_\theta(z, r, t) = g_\theta(z, t, h)$ with $h = t - r$. The chain rule at fixed r gives

$$\partial_t f_\theta|_r = \partial_t g_\theta + \partial_h g_\theta, \qquad \partial_r f_\theta|_t = -\partial_h g_\theta, \tag{24}$$

$$\mathcal{D}_t f_\theta = \nabla_z g_\theta \, v_\star + \partial_t g_\theta + \partial_h g_\theta. \tag{25}$$

This proves the equivalence of JVP directions $(v_c, 0, 1)$ in (z, r, t) coordinates and $(v_c, 1, 1)$ in (z, t, h) coordinates. A JVP with direction $(v_c, 1, 0)$ in the latter coordinates omits $\partial_h g_\theta$.

Fix network parameters and condition on (z_t, r, t). Assume interval sampling does not reveal additional information about (x, ϵ) beyond (z_t, t), so that $\mathbb{E}[v_c \mid z_t, r, t] = v_\star(z_t, t)$. For a single head, write $J_k = \nabla_z m_{k,\theta}$ and $b_k = \partial_t m_{k,\theta}|_r$. The numerical value of its detached target is

$$Y_k = \frac{(I - hJ_k)v_c - hb_k}{k+1}, \qquad \bar{Y}_k = \mathbb{E}[Y_k \mid z_t, r, t] = \frac{v_\star - h\mathcal{D}_t m_{k,\theta}}{k+1}. \tag{26}$$

Thus the conditional mean of the unweighted training residual is exactly the hierarchy residual divided by $k + 1$:

$$m_{k,\theta} - \bar{Y}_k = \frac{(k+1)m_{k,\theta} + h\mathcal{D}_t m_{k,\theta} - v_\star}{k+1}. \tag{27}$$

This statement is about averaging at fixed parameters, not convergence of the learning algorithm.

The conditional squared loss contains an additional variance term. With $\Sigma_c = \mathrm{Cov}(v_c \mid z_t, r, t)$,

$$\mathbb{E}[\|m_{k,\theta} - Y_k\|^2 \mid z_t, r, t] = \|m_{k,\theta} - \bar{Y}_k\|^2 + \mathrm{tr}\,\mathrm{Cov}(Y_k \mid z_t, r, t), \tag{28}$$

$$\mathrm{Cov}(Y_k \mid z_t, r, t) = \frac{(I - hJ_k)\Sigma_c(I - hJ_k)^\top}{(k+1)^2}. \tag{29}$$

Both the conditional mean target and its variance depend on the current network. Detaching the target removes their derivative contributions from the implemented update. For unweighted regression, that update averages to $2(\partial_\theta m_{k,\theta})^\top (m_{k,\theta} - \bar{Y}_k)$, conditional on the inputs. It is therefore not the full gradient of the squared hierarchy residual.

Adaptive weighting further changes this relation. If $W = (e_k + \delta)^{-p}$ is numerically evaluated but detached, the averaged residual entering the update is

$$\mathbb{E}[W(m_{k,\theta} - Y_k) \mid z_t, r, t] = \mathbb{E}[W \mid z_t, r, t] m_{k,\theta} - \mathbb{E}[WY_k \mid z_t, r, t]. \tag{30}$$

In general $\mathbb{E}[WY_k] \neq \mathbb{E}[W]\bar{Y}_k$, because W depends on the sampled residual. Consequently, the conditional mean identity does not by itself establish consistency of the adaptively weighted optimizer.

D.2 Hierarchy residuals control moment errors along a path

Fix a characteristic and let $\widetilde{m}_k$ be a differentiable approximation with residual

$$R_k(s) = (k+1)\widetilde{m}_k(X_s, r, s) + (s-r)\mathcal{D}_t\widetilde{m}_k(X_s, r, s) - v_\star(X_s, s). \tag{31}$$

Subtracting the exact hierarchy and defining $e_k = \widetilde{m}_k - m_k$ yields

$$\frac{d}{ds}\big[(s-r)^{k+1} e_k(X_s, r, s)\big] = (s-r)^k R_k(s). \tag{32}$$

If e_k is bounded at the diagonal, integration gives the exact residual representation

$$e_k(X_t, r, t) = \frac{1}{h^{k+1}} \int_r^t (s-r)^k R_k(s)\, ds, \qquad \|e_k\| \leq \frac{\sup_{s\in[r,t]} \|R_k(s)\|}{k+1}. \tag{33}$$

Setting $R_k = 0$ proves uniqueness among bounded solutions. If only a boundary at $r + \varepsilon$ is available, an additional term $(\varepsilon/h)^{k+1} e_k(X_{r+\varepsilon}, r, r+\varepsilon)$ remains, and the integral starts at $r + \varepsilon$. The bound requires residual control along the marginal path; a small minibatch training loss does not automatically supply such uniform control.

D.3 Local expansion and the residual-head parameterization

Suppose $a(s) = v_\star(X_s, s)$ is twice continuously differentiable on the interval. Taylor expansion about r gives

$$m_k = \frac{a(r)}{k+1} + \frac{h\dot{a}(r)}{k+2} + O(h^2), \qquad \dot{a}(r) = \mathcal{D}_t v_\star(X_r, r). \tag{34}$$

Subtracting $m_0/(k+1)$ and dividing by h shows that the exact residual corresponding to Eq. (4) satisfies

$$\rho_k^\star = \frac{m_k - m_0/(k+1)}{h} = \frac{k}{2(k+1)(k+2)} \mathcal{D}_t v_\star(X_r, r) + O(h). \tag{35}$$

Thus bounded residual heads are compatible with smooth exact moments, rather than forcing higher moments to be redundant away from the diagonal.

For $A_q(\tau) = \sum_k a_k(q)\tau^k$, define $c_1(q) = \int_0^1 \tau A_q(\tau)\, d\tau = \sum_k a_k(q)/(k+2)$. Using $\int A_q = 1$, the exact-moment readout obeys

$$u_K(q) - m_0 = h\big(c_1(q) - \tfrac{1}{2}\big)\mathcal{D}_t v_\star(X_r, r) + O(h^2). \tag{36}$$

Its endpoint bias is therefore $O(h^2)$ for fixed coefficients on smooth short intervals. This is a local exact-moment statement, not a high-order accuracy guarantee for the learned one-step model at $h = 1$.

D.4 Normalized transform truncation and denominator stability

Let $P_K(z) = \sum_{k=0}^K z^k/k!$, $Z_K(q) = \int_0^1 P_K(q\tau) w(\tau)\, d\tau$, and

$$d_\infty(q) = \int_0^1 e^{q\tau}\, d\tau = \begin{cases} (e^q - 1)/q, & q \neq 0, \\ 1, & q = 0. \end{cases} \tag{37}$$

Integrating the pointwise exponential remainder, and using $\|w\| \le V$, gives the sharper bounds

$$\varepsilon_K(q) = \frac{e^{|q|}|q|^{K+1}}{(K+1)!(K+2)}, \qquad |d_\infty - d_K| \le \varepsilon_K, \quad \|Z - Z_K\| \le V\varepsilon_K. \tag{38}$$

If $|d_\infty| > \varepsilon_K$, then $|d_K| \ge |d_\infty| - \varepsilon_K > 0$. The quotient identity

$$\frac{Z_K}{d_K} - \frac{Z}{d_\infty} = \frac{Z_K - Z}{d_K} + \frac{Z(d_\infty - d_K)}{d_K d_\infty} \tag{39}$$

therefore implies

$$\left\| \frac{Z_K}{d_K} - \frac{Z}{d_\infty} \right\| \le \frac{V\varepsilon_K}{|d_K|} \left(1 + \frac{e^{|q|}}{|d_\infty|} \right). \tag{40}$$

Taking the real part cannot increase this norm. This controls truncation of a normalized exponential readout; its limit can still differ from the unweighted moment m_0.

A simple sufficient zero-free region follows from

$$|d_K(q) - 1| \le \sum_{k=1}^{K} \frac{|q|^k}{k!(k+1)} \le e^{|q|} - 1. \tag{41}$$

Hence $|q| < \log 2$ guarantees $d_K(q) \neq 0$ for every K. At the experimental setting $K = 3$, $q = -1 + 0.5i$, direct evaluation gives $d_3(q) = 0.614583 + 0.140625i$ and $|d_3(q)| \approx 0.630467$. The sufficient region therefore excludes some valid parameters. In contrast, d_∞ vanishes at $q = 2\pi i n$ for nonzero integers n.

D.5 An exact criterion for mean-squared readout improvement

Fix a path, let $B_q = \sum_k a_k m_k - m_0$, and suppose the estimation errors η_k are random. Define $\mu_k = \mathbb{E}\eta_k$ and $C_{k\ell} = \mathbb{E}\langle \eta_k - \mu_k, \eta_\ell - \mu_\ell \rangle$. Expansion of the square gives

$$\mathbb{E}\|\widehat{u}_K - m_0\|^2 = \left\| B_q + \sum_k a_k \mu_k \right\|^2 + \sum_{k,\ell} a_k a_\ell C_{k\ell}, \tag{42}$$

$$\mathbb{E}\|\widehat{m}_0 - m_0\|^2 = \|\mu_0\|^2 + C_{00}. \tag{43}$$

The active readout improves this conditional mean-squared error exactly when the first expression is smaller than the second. Correlated errors can make a weighted combination useful, but the reweighting bias and nonzero error means can reverse the gain. This criterion explains which quantities would need measurement to substantiate an error-cancellation mechanism. It does not convert the observed FID gains into a proof of that mechanism.

D.6 Exact composition of moments across adjacent intervals

For $r < s < t$, write $a = s - r$, $b = t - s$, $h = a + b$, and use the same characteristic for all intervals. Splitting the integral at s and expanding $(\tau - r)^k = ((\tau - s) + a)^k$ on $[s, t]$ gives

$$h^{k+1} m_k(X_t, r, t) = a^{k+1} m_k(X_s, r, s) + \sum_{j=0}^{k} \binom{k}{j} a^{k-j} b^{j+1} m_j(X_t, s, t). \tag{44}$$

For $k = 0$, this reduces to additivity of displacement: $h m_0(X_t, r, t) = a m_0(X_s, r, s) + b m_0(X_t, s, t)$. The required intermediate state is $X_s = X_t - b m_0(X_t, s, t)$, not an independently drawn conditional interpolant. Equation (44) gives the exact moment-level composition identity underlying interval consistency. Replacing X_s by a learned update introduces state error. Section 3.4 specifies the EMA intermediate state and detached teacher targets used in training.

D.7 Complex cocycles, phase weighting, and finite-order composition

Fix a characteristic with integrable velocity on a finite interval and suppress the dependence on its terminal state. The transform $\mathcal{G}_{r,t}(\zeta)$ in Eq. (8) is entire: the exponential series converges uniformly on each compact subset of the complex plane, and the finite interval provides an integrable dominating bound. Its derivatives recover every path moment,

$$\partial_\zeta^k \mathcal{G}_{r,t}(\zeta)\big|_{\zeta=0} = \int_r^t (u-r)^k v_\star(X_u, u)\, du = h^{k+1} m_k(X_t, r, t). \tag{45}$$

Thus the complex transform packages the moment hierarchy into one analytic object; analyticity does not require a separate Cauchy–Riemann loss.

Weighted composition and associativity. Split the integral at s. On the second subinterval, $e^{\zeta(u-r)} = e^{\zeta(s-r)} e^{\zeta(u-s)}$, which proves Eq. (9). If $r < s < t < u$, either grouping of adjacent intervals gives

$$\mathcal{G}_{r,u} = \mathcal{G}_{r,s} + e^{\zeta(s-r)} \mathcal{G}_{s,t} + e^{\zeta(t-r)} \mathcal{G}_{t,u}. \tag{46}$$

Equivalently, the pairs (ℓ, G) compose as $(a, A) \star (b, B) = (a + b, A + e^{\zeta a} B)$, with identity $(0, 0)$. The relation $e^{\zeta(a+b)} = e^{\zeta a} e^{\zeta b}$ proves associativity. This weighted additive cocycle describes the unnormalized path transform; the normalized sampling map is analyzed separately.

Rescaling the normalized coordinate. Write $Z_{r,t}(q) = \mathcal{G}_{r,t}(q/h)/h$, $a = s - r$, and $b = t - s$. Keeping the physical transform parameter $\zeta = q/h$ fixed gives

$$Z_{r,t}(q) = \frac{a}{h} Z_{r,s}\left(\frac{a}{h} q\right) + \frac{b}{h} e^{qa/h} Z_{s,t}\left(\frac{b}{h} q\right). \tag{47}$$

The subinterval parameters must therefore scale with their lengths; using the same dimensionless q on all intervals describes a different transform. The same formula holds for $d_\infty(q) = \int_0^1 e^{q\tau}\, d\tau$ in place of Z, because it is the constant-velocity transform. Dividing each interval by its own denominator consequently does not preserve the unnormalized additive law. The real sampling readout must be analyzed with its normalization and its state dependence intact.

What the imaginary coordinate contributes. For real velocity and $\zeta = \alpha + i\omega$,

$$\operatorname{Re} \mathcal{G}_{r,t}(\alpha + i\omega) = \int_r^t e^{\alpha(u-r)} \cos(\omega(u-r)) v_\star(X_u, u)\, du, \tag{48}$$

$$\operatorname{Im} \mathcal{G}_{r,t}(\alpha + i\omega) = \int_r^t e^{\alpha(u-r)} \sin(\omega(u-r)) v_\star(X_u, u)\, du. \tag{49}$$

Real and imaginary coordinates encode amplitude-weighted cosine and sine summaries. Translation of the interval origin multiplies their joint complex value by $e^{\zeta a}$, combining amplitude scaling with phase rotation. At $\alpha = 0$, this multiplier has modulus one, so translation alone does not amplify an error in the subinterval transform. This does not control the sampling denominator, which can have zeros. For the finite readout, varying the imaginary part changes the real coefficient family $a_k(q)$; the network can still predict entirely real tensors and use real arithmetic at deployment. The empirical effect of interval consistency is evaluated in Table 3.

Exact finite-order coefficient identity. Define

$$\mathcal{G}_{r,t}^{[K]}(\zeta) = \sum_{k=0}^{K} \frac{\zeta^k h^{k+1}}{k!} m_k(X_t, r, t), \tag{50}$$

and let $\Pi_{\le K}$ retain Taylor coefficients of degrees at most K at zero. Differentiating the exact cocycle, or matching its Taylor coefficients, yields

$$\mathcal{G}_{r,t}^{[K]} = \Pi_{\le K}\left[\mathcal{G}_{r,s}^{[K]} + e^{\zeta a} \mathcal{G}_{s,t}^{[K]}\right]. \tag{51}$$

The coefficient of ζ^k is precisely Eq. (44). This identifies moment composition as the finite-order coefficient form of the exponential cocycle. Without the projection, the exponential factor creates higher-degree terms, so the unprojected equality is generally false for a finite polynomial.

Truncation and prediction error. If $\|v_\star(X_u, u)\| \le V$, define

$$E_K(\ell, \zeta) = \frac{V\ell e^{|\zeta|\ell}(|\zeta|\ell)^{K+1}}{(K+1)!(K+2)}. \tag{52}$$

Integrating the exponential remainder gives $\|\mathcal{G}_{r,t} - \mathcal{G}^{[K]}_{r,t}\| \le E_K(h, \zeta)$. The unprojected composition defect therefore obeys

$$\left\|\mathcal{G}^{[K]}_{r,t} - \mathcal{G}^{[K]}_{r,s} - e^{\zeta a}\mathcal{G}^{[K]}_{s,t}\right\| \le E_K(h, \zeta) + E_K(a, \zeta) + e^{a\,\mathrm{Re}\,\zeta}E_K(b, \zeta). \tag{53}$$

For learned transforms evaluated at the same exact states, add their prediction-error bounds $\epsilon_{r,t} + \epsilon_{r,s} + e^{a\,\mathrm{Re}\,\zeta}\epsilon_{s,t}$ to the right-hand side. Replacing the intermediate state by a learned sample requires an additional state-error bound. These identities provide a structural interpretation of cocycle supervision and quantify finite-order composition error. The bounds concern represented transforms and hold independently of the optimizer and loss weighting.

D.8 Coefficient-level cocycle residuals

The exact composition law supplies an explicit residual without treating the finite polynomial as an exact exponential transform. Set $a = s - r$, $b = t - s$, and $h = a + b$. For predicted moments and a chosen intermediate state $\widetilde{X}_s$, define

$$\begin{aligned} C_{k,\theta} = \widehat{m}_k(X_t, r, t) &- \left(\frac{a}{h}\right)^{k+1} \widehat{m}_k(\widetilde{X}_s, r, s) \\ &- \sum_{j=0}^{k} \binom{k}{j} \left(\frac{a}{h}\right)^{k-j} \left(\frac{b}{h}\right)^{j+1} \widehat{m}_j(X_t, s, t), \qquad 0 \le k \le K. \end{aligned} \tag{54}$$

With exact moments and $\widetilde{X}_s = X_s$, every residual vanishes by Eq. (44). For any nonnegative weights w_k, the quantity $\mathcal{R}_{\mathrm{coc}} = \mathbb{E}\sum_{k=0}^{K} w_k \|C_{k,\theta}\|^2$ measures coefficient-level inconsistency. Section 3.4 uses these composition coefficients with detached EMA targets to define the training loss.

State-error sensitivity. Let the left-interval moment predictor be spatially Lipschitz with constant L_k. Holding the other evaluations fixed gives

$$\|C_{k,\theta}(\widetilde{X}_s) - C_{k,\theta}(X_s)\| \le (a/h)^{k+1} L_k \|\widetilde{X}_s - X_s\|. \tag{55}$$

If the intermediate state is constructed as $\widetilde{X}_s = X_t - b\widehat{m}_0(X_t, s, t)$, its error is exactly $b\|\widehat{m}_0(X_t, s, t) - m_0(X_t, s, t)\|$. Thus interval consistency couples moment accuracy to state accuracy. Vanishing cocycle residual alone is insufficient to recover the data transport: identically zero moment predictions also satisfy the composition equations. Local velocity supervision anchors the transport, whereas the cocycle constrains compatibility across intervals. The two terms therefore constrain target accuracy and interval compatibility, respectively.

E Conditional convergence and the benefit of joint readouts

The following results identify three potential advantages of the construction: stable recovery of path moments from small hierarchy residuals, convergent transport under controlled approximation error, and reduced estimation variance through joint prediction. They concern the represented fields and readouts. They establish conditional consistency and stability of the representation, rather than convergence of the combined stochastic training procedure. In particular, the vanishing-q result describes a consistent limit of the readout family, while the experiments use a fixed nonzero q.

E.1 Mean-square consistency of moment recovery

Proposition 3 (Residual-to-moment consistency)**.** *Let $\widetilde{m}^{(n)}_k$ be a sequence of differentiable moment approximations, bounded at the diagonal along almost every characteristic. Let $R^{(n)}_k$ be their hierarchy residuals as in Eq. (33). For a distribution of paths and fixed $r < t$, suppose*

$$\sup_{s\in[r,t]} \mathbb{E}\|R^{(n)}_k(s)\|^2 \le \varepsilon^2_{k,n}. \tag{56}$$

Then

$$\mathbb{E}\|\widetilde{m}_k^{(n)} - m_k\|^2 \le \frac{\varepsilon_{k,n}^2}{(k+1)^2}. \tag{57}$$

In particular, vanishing hierarchy residuals in this sense imply mean-square convergence of each fixed-order moment.

Proof. The residual representation can be written as $\widetilde{m}_k^{(n)} - m_k = \int_0^1 \tau^k R_k^{(n)}(r + h\tau)\, d\tau$. The density $(k+1)\tau^k$ integrates to one. Jensen's inequality and Tonelli's theorem give

$$\mathbb{E}\|\widetilde{m}_k^{(n)} - m_k\|^2 \le \frac{1}{k+1}\int_0^1 \tau^k \mathbb{E}\|R_k^{(n)}(r + h\tau)\|^2\, d\tau \tag{58}$$

$$\le \frac{\varepsilon_{k,n}^2}{(k+1)^2}. \tag{59}$$

□

The factor $(k+1)^{-1}$ reflects normalization of the moment, not a claim that higher heads are easier to train. The assumption is on the marginal hierarchy residual, which is distinct from the sampled, adaptively weighted training loss.

E.2 Transport consistency and a distributional bound

Proposition 4 (Convergence with a vanishing readout perturbation)**.** *Fix K. Couple exact and learned one-step outputs by the same terminal noise X_1. Suppose the exact characteristic transports its terminal law to p_{data}, all required second moments are finite, and $\|\widehat{m}_k^{(n)} - m_k\|_{L^2} \to 0$ for every $k \le K$. If $q_n \to 0$, then the law $\widehat{p}_n$ of $X_1 - \sum_k a_k(q_n)\widehat{m}_k^{(n)}$ satisfies*

$$W_2(\widehat{p}_n, p_{\mathrm{data}}) \longrightarrow 0. \tag{60}$$

Proof. Since $d_K(0) = 1$, the coefficients are continuous near zero and $a_k(q_n) \to \mathbf{1}_{\{k=0\}}$. Write $\beta_k(q) = a_k(q) - \mathbf{1}_{\{k=0\}}$. The shared-noise coupling and Minkowski's inequality yield

$$W_2(\widehat{p}_n, p_{\mathrm{data}}) \le \left\| \sum_k a_k(q_n)\widehat{m}_k^{(n)} - m_0 \right\|_{L^2} \tag{61}$$

$$\le \sum_k |a_k(q_n)| \|\widehat{m}_k^{(n)} - m_k\|_{L^2} + \sum_k |\beta_k(q_n)| \|m_k\|_{L^2} \longrightarrow 0. \tag{62}$$

Here W_2 is the infimum, over couplings, of the root mean-square distance between paired samples; our coupling is an admissible upper bound. □

This establishes a consistent limit within the readout family because that family contains the exact average at $q = 0$. For the fixed nonzero parameter used in the experiments, the second term need not vanish. Increasing K or eliminating moment-estimation error alone does not remove that bias. The experiments keep q fixed; their error decomposition is given in Appendix E.5.

E.3 Stability under repeated short-interval updates

Let $\Phi_{r,t}$ be the exact backward transport map. Assume it is spatially Lipschitz with constant at most $e^{L(t-r)}$, $L \ge 0$, on a region containing both exact and approximate states. Consider uniform steps of length $h = T/N$ and a learned update Ψ_j. Suppose uniformly on that region

$$\|\Psi_j(z) - \Phi_{t_{j+1},t_j}(z)\| \le h\delta_h + Ch^2, \qquad t_j = T - jh. \tag{63}$$

The first term accounts for moment-estimation error after readout; the second is the smooth-path reweighting bias from Eq. (36). For fixed stable coefficients, bounded material derivatives, and uniform head error η_h, one may take $\delta_h \le \sum_k |a_k|\eta_h$.

Starting both trajectories from the same state, their errors satisfy $E_{j+1} \le e^{Lh}E_j + h\delta_h + Ch^2$. Iterating this recurrence proves

$$E_N \le (h\delta_h + Ch^2)\sum_{j=0}^{N-1} e^{Lhj} \le \begin{cases} \dfrac{e^{LT}-1}{L}(\delta_h + Ch), & L > 0, \\ T(\delta_h + Ch), & L = 0. \end{cases} \tag{64}$$

Thus uniform approximation error $\delta_h \to 0$ gives convergent repeated updates. A nonzero error floor remains when δ_h does not vanish. The theorem requires accuracy at every queried state and interval; increasing NFE for a fixed network does not establish that assumption. The finite-NFE diagnostics in the supplementary results therefore need not improve monotonically.

E.4 Why additional heads can improve a learned readout

Proposition 5 (Variance reduction under constant-path calibration)**.** *Suppose a constant path has velocity v and the predicted heads satisfy $\widehat{m}_k = c_k v + \eta_k$, $c_k = 1/(k+1)$, with $\mathbb{E}\eta_k = 0$. Let $G_{k\ell} = \mathbb{E}\langle \eta_k, \eta_\ell \rangle$ be positive definite. Among all real linear readouts $a^\top \widehat{m}$ that preserve constant velocity, the minimum mean-squared error is*

$$a^\star = \frac{G^{-1}c}{c^\top G^{-1}c}, \qquad \min_{c^\top a = 1} \mathbb{E}\|a^\top \widehat{m} - v\|^2 = \frac{1}{c^\top G^{-1}c} \le G_{00}. \tag{65}$$

The inequality is strict unless $a^\star$ equals the center coefficient vector.

Proof. Under $c^\top a = 1$, the error is $\sum_k a_k \eta_k$ and its squared expectation is $a^\top G a$. The Lagrange condition for minimizing this strictly convex quadratic is $Ga = \lambda c$. Enforcing the constraint gives the stated solution and value. The center vector $(1, 0, \ldots, 0)^\top$ is feasible, so the optimum cannot be worse and is strictly better unless the center is itself the unique optimum. □

For example, with two independent heads of equal error energy σ^2, $c = (1, 1/2)^\top$ and $G = \sigma^2 I$, the optimum is $a^\star = (4/5, 2/5)^\top$ with error $4\sigma^2/5$, versus σ^2 for the center. This is an illustrative noise model, not a measurement of our trained heads.

The proposition establishes a concrete benefit available to joint predictions: redundant unbiased estimates can reduce variance while preserving the transport target. The generating-function readout restricts the admissible coefficients to $a(q)$ and need not attain the unrestricted optimum. Although $q = 0$ makes the center available, a particular selected nonzero q is not automatically better. For nonconstant paths and biased predictions, the full criterion is Eq. (43). Joint training, useful error diversity, and reliable readout selection remain empirical requirements.

E.5 How local supervision, cocycle consistency, and readout interact

The three components constrain different sources of error. Local velocity supervision anchors the target transport, cocycle consistency couples predictions across intervals, and the readout combines the remaining moment errors. The following bound applies at a fixed nonzero q, including the experimental choice $q = -1 + 0.5i$; it does not require $q \to 0$.

Fix $r < s < t$, $a = s - r$, $b = t - s$, and $h = a + b$. Evaluate the full and subinterval predictions on the same exact characteristic. Write $e_k^{r,t} = \widehat{m}_k(X_t, r, t) - m_k(X_t, r, t)$ and let C_k be the coefficient residual in Appendix D.8. Subtracting the exact composition identity yields

$$\begin{aligned} e_k^{r,t} = C_k &+ \left(\frac{a}{h}\right)^{k+1} e_k^{r,s} \\ &+ \sum_{j=0}^{k} \binom{k}{j} \left(\frac{a}{h}\right)^{k-j} \left(\frac{b}{h}\right)^{j+1} e_j^{s,t}. \end{aligned} \tag{66}$$

Consequently, if $\|C_k\| \le \delta_k$ and subinterval errors are bounded by $\varepsilon_k^{r,s}$ and $\varepsilon_j^{s,t}$, then

$$\|e_k^{r,t}\| \le B_k := \delta_k + \left(\frac{a}{h}\right)^{k+1} \varepsilon_k^{r,s} + \sum_{j=0}^{k} \binom{k}{j} \left(\frac{a}{h}\right)^{k-j} \left(\frac{b}{h}\right)^{j+1} \varepsilon_j^{s,t}. \tag{67}$$

If the left prediction uses an approximate state $\widetilde{X}_s$, add $(a/h)^{k+1}L_k\|\widetilde{X}_s - X_s\|$ to B_k, assuming the left predictor is L_k-Lipschitz. For an EMA-target residual, teacher prediction errors replace the corresponding subinterval errors. Thus a small consistency residual controls full-interval error only together with accurate subinterval predictions and intermediate states.

Let $b_K(q) = \sum_{k=0}^{K} a_k(q)m_k - m_0$ be the exact-moment readout bias. The endpoint error of the active update obeys

$$\|X_t - h\widehat{u}_K(q) - X_r\| \le h\sum_{k=0}^{K}|a_k(q)|B_k + h\|b_K(q)\|. \tag{68}$$

When local supervision yields accurate subinterval predictions, cocycle consistency controls their compatibility with the full-interval prediction. The readout propagates the resulting errors through explicit coefficients, without requiring q to vanish. Cocycle consistency alone cannot identify the desired transport, because zero predictions satisfy the composition identities. Conversely, an accurate local target need not ensure compatibility of finite-network predictions across intervals.

The bound controls endpoint error through subinterval accuracy, composition residuals, and temporal bias. Equation (43) further characterizes the roles of cross-head correlations and bias in readout improvement. The cocycle ablation and paired readout comparisons evaluate the training and sampling components of this mechanism, respectively; endpoint-error control does not imply an FID ordering.