Is your Flow Matching Model Really Generalising? A Path-Length Diagnostic

Anonymous Authors[1]

Abstract

Flow-matching models learn a time-dependent velocity field that transports a Gaussian source to the data distribution along ODE trajectories. We propose a simple geometric diagnostic for assessing whether such models have truly learnt to generate from the data manifold: the path length, defined as the integrated velocity norm along a trajectory. Studying three flow-matching models trained on MNIST under different computational budgets, we find a striking pattern. Path-length distributions on training and test data, computed via the reverse ODE, are nearly indistinguishable and shrink monotonically as training progresses. Path-length distributions on freshly generated samples, by contrast, remain essentially constant across training budgets. As a result, three regimes emerge: at high FID (approximately 130), generated paths are shorter than data paths; at moderate FID (approximately 13), the two distributions match; and at low FID (approximately 8), generated paths exceed data paths. This pattern is not explained by memorisation, nor by numerical integration error, nor by velocity-magnitude artefacts at the trajectory endpoints. We interpret matching path-length distributions as a signal that generation has landed on the data manifold, and we relate the observed asymmetry to recent results on the implicit regularisation and trajectory stability of flow matching. Path length thus provides a cheap, model-internal probe that complements FID and memorisation metrics.

1. Introduction

Generative modelling via normalising flows (Papamakarios et al., 2021) has long relied on learning explicit transformations between a simple prior and a data distribution. Recent work on flow matching (Lipman et al., 2022; Liu et al., 2022; Albergo & Vanden-Eijnden, 2022) has shifted the paradigm towards learning local dynamics: instead of specifying a full bijective map $T : \mathcal{X} \to \mathcal{Z}$, flow-matching models learn a time-dependent velocity field $\mathbf{v}_\theta(\mathbf{x}, t)$ whose integral over the time interval $[0, 1]$ produces global transport from a Gaussian prior p_0 to the data distribution p_{data}.

Samples are generated by solving the ODE

$$\frac{d\mathbf{x}_t}{dt} = \mathbf{v}_\theta(\mathbf{x}_t, t), \quad \mathbf{x}_0 \sim \mathcal{N}(0, \mathbb{I}), \tag{1}$$

where the velocity field is parameterised by a neural network. This local-to-global perspective avoids invertibility constraints, scales naturally to high-dimensional data, and provides a flexible framework for understanding transport geometry.

A natural question arises: how can we assess whether a flow-matching model has truly learnt to sample from the data distribution, rather than memorising the training set or settling on a poor mapping from prior to data? Standard metrics such as Fréchet Inception Distance (FID) measure sample quality but conflate model capacity, mode coverage, and true generalisation. Memorisation metrics based on nearest-neighbour distances detect direct copying but miss subtler failure modes.

In this work, we propose a simple model-internal diagnostic based on the geometry of the learned trajectories. For a sample $\mathbf{x}$ integrated along the velocity field, we define the path length

$$L(\mathbf{x}) = \int_0^1 \|\mathbf{v}_\theta(\mathbf{x}_t, t)\| \, dt, \tag{2}$$

and compare its distribution between the data and the generated samples. We argue that, if the model has genuinely learnt the data manifold, the forward trajectory from noise to a generated sample and the reverse trajectory from a real data point to noise should cover similar geometric distances. Matching path-length distributions, therefore, signals that the generation lands on the same manifold that the data inhabits.

Empirically, this diagnostic uncovers a striking and reproducible pattern. Path-length distributions on generated samples are remarkably stable across training budgets, whilst

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the FoGen Workshop at ICML 2026. Do not distribute.

path-length distributions on data shrink monotonically as the model improves. The two distributions match only at an intermediate point in training, and diverge in opposite directions before and after that point. We probe this phenomenon through targeted ablations and connect it to recent results on the implicit regularisation (Bertrand et al., 2026) and trajectory stability (Briq et al.; English & Suzuki) of flow matching.

Our contributions are:

1. We propose path length as a cheap, model-internal diagnostic for flow-matching models, complementary to FID and memorisation metrics.

2. Across three training budgets on MNIST, we identify three regimes (generated paths shorter than, matched to, and longer than data paths). We show that this pattern is produced by data path lengths shrinking around an essentially fixed generated distribution of the path lengths, rather than by changes in the generated distribution itself.

3. We perform ablations ruling out numerical integration error and velocity-magnitude artefacts at the trajectory endpoints as explanations for the path-length discrepancy.

4. We connect the observed asymmetry between forward and reverse dynamics to recent results, arguing that the stability of generated path lengths is a quantitative signature of the implicit regularisation that underpins flow matching.

Note that we do not claim path-length matching is a complete generalisation criterion. At 100 epochs, our diagnostic disagrees with FID and NNDR, and we view this disagreement as a feature: path length captures geometric information orthogonal to standard quality metrics, and we examine the resulting tension in Section 6.

2. Related Work

Flow matching and continuous-time generative models. Flow matching (Lipman et al., 2022; Albergo & Vanden-Eijnden, 2022; Liu et al., 2022) provides a simulation-free objective for training continuous-time generative models that transport a Gaussian source to data via a learned velocity field. The framework subsumes score-based diffusion as a special case (Song et al., 2020; Karras et al., 2022) and admits a natural optimal-transport interpretation through the Benamou–Brenier formulation. Subsequent work has emphasised straighter trajectories: Rectified Flow (Liu et al., 2022) introduces a reflow procedure that reduces trajectory curvature, and OT-CFM (Tong et al., 2023) couples source and target via minibatch optimal transport. These works establish straightness as a desirable property of trained flow models, motivated by both sampling efficiency and transport optimality.

Trajectory geometry and transport quality. A separate line of work examines the geometric properties of generative trajectories. Straightness, defined as the deviation of a trajectory from its endpoint chord (Liu et al., 2022), is a training-time signal optimised by Rectified Flow and consistency-style training (Song et al., 2023). EDM (Karras et al., 2022) reframes diffusion as a noise-schedule design problem in which trajectory geometry directly affects sampling efficiency. Related analyses probe the manifold structure encoded by score-based models (Pidstrigach, 2022; Stanczuk et al., 2024). These approaches engage with trajectory geometry as a training objective or theoretical object; they are rarely used as a post-hoc, model-internal diagnostic.

Stability and regularisation in flow matching. Recent work has highlighted two related phenomena specific to flow matching. First, the trajectories produced by flow-matching models are remarkably stable: similar noise vectors map to similar generated samples even across architectures and conditional path formulations (Briq et al.; English & Suzuki). Second, Bertrand et al. (2026) derived a closed-form expression for the empirically optimal velocity field under conditional flow matching with a Gaussian source and showed that practical training does not fully minimise this objective. The resulting deviation acts as an implicit regulariser, preventing memorisation. Our diagnostic provides a quantitative geometric view of these phenomena: we observe that generated path lengths are essentially fixed across training budgets, which we interpret as a signature of the same implicit regularisation from a network.

Kinetic energy and sparsity. Li et al. (2026) report that higher kinetic energy in flow-matching trajectories is associated with sparsity of the data distribution. We frame our diagnosis in terms of path length rather than kinetic energy and argue (Section 4) that the geometric reading provides a more transparent interpretation of such observations: trajectories are simply longer when the data lies further from the prior or is more spread out.

3. Background

3.1. Flow Matching

Given a Gaussian prior $p_0(\mathbf{x}) = \mathcal{N}(0, \mathbb{I})$ and a data distribution $p_{\text{data}}(\mathbf{x})$, flow matching defines a time-dependent transport map $\Phi_t : \mathcal{X} \to \mathcal{X}$ parametrised by a velocity field:

$$\frac{d\mathbf{x}_t}{dt} = \mathbf{v}_\theta(\mathbf{x}_t, t), \quad \mathbf{x}_0 \sim p_0, \quad \mathbf{x}_1 \sim p_{\text{data}}. \tag{3}$$

The velocity field is trained to minimise a regression objective:

$$\mathcal{L}(\theta) = \mathbb{E}_{t,\mathbf{x}_t}\left[\|\mathbf{v}_\theta(\mathbf{x}_t, t) - \mathbf{u}_t(\mathbf{x}_t)\|^2\right], \tag{4}$$

where $\mathbf{u}_t$ is a target velocity field, often derived from optimal transport or diffusion theory. Generation proceeds by sampling $\mathbf{x}_0 \sim p_0$ and integrating the ODE forward, $\mathbf{x}_1 = \mathbf{x}_0 + \int_0^1 \mathbf{v}_\theta(\mathbf{x}_t, t)\, dt$. Compared with normalising flows (Papamakarios et al., 2021), this local-to-global perspective avoids invertibility constraints and scales naturally to high-dimensional data.

3.2. Closed-Form Optimal Velocity

Bertrand et al. (2026) showed that under conditional flow matching with a Gaussian source $p(x \mid x_1 = x^{(i)}, t) = \mathcal{N}(tx^{(i)}, (1-t)^2 I_d)$, the empirically optimal velocity field admits a closed-form expression:

$$\hat{u}^\star(x,t) = \sum_{i=1}^{n} \lambda_i(x,t)\frac{x^{(i)} - x}{1-t}, \tag{5}$$

with

$$\lambda_i(x,t) = \frac{\exp\left(-\frac{\|x - tx^{(i)}\|^2}{2(1-t)^2}\right)}{\sum_{j=1}^{n}\exp\left(-\frac{\|x - tx^{(j)}\|^2}{2(1-t)^2}\right)}. \tag{6}$$

A key observation in their work is that practical training does not fully minimise the conditional flow-matching objective; the deviations from $\hat{u}^\star$ are penalised differently by the network's implicit regularisation that helps generalisation. We return to this implicit-regularisation interpretation in Section 6, where it informs our reading of why generated path lengths remain stable across training.

3.3. Evaluating Flow-Matching Models

The standard evaluation toolkit for generative models combines distributional metrics such as FID with proximity-based memorisation checks (e.g., nearest-neighbour distance ratios in pixel or feature space). Both have known limitations. FID is sensitive to the choice of feature extractor and can mask failure modes in coverage or in the prior-to-data mapping. Nearest-neighbour metrics detect copying but say little about whether the model has learnt the underlying manifold geometry. There is growing interest in geometric and trajectory-based probes of generative models, but these are typically used as training objectives rather than as post-hoc diagnostics.

4. Method

4.1. Path Length as a Geometric Diagnostic

For a trajectory $(\mathbf{x}_t)_{t\in[0,1]}$ governed by the ODE $\frac{d\mathbf{x}_t}{dt} = \mathbf{v}_\theta(\mathbf{x}_t, t)$, the path length is the arc length:

$$L(\mathbf{x}) = \int_0^1 \|\mathbf{v}_\theta(\mathbf{x}_t, t)\|\, dt. \tag{7}$$

This is the geometric distance traversed along the integral curves of the learned velocity field. The Benamou–Brenier formulation links this quantity to optimal transport: the squared 2-Wasserstein distance is the minimum kinetic energy over all transport plans, so per-sample path lengths are bounded below by the corresponding optimal transport costs.

Why path length rather than kinetic energy? A common alternative is to study the kinetic energy along a trajectory,

$$E(\mathbf{x}) = \frac{1}{2}\int_0^1 \|\mathbf{v}_\theta(\mathbf{x}_t, t)\|^2\, dt. \tag{8}$$

The two quantities are related by the Cauchy–Schwarz inequality and shift in the same direction, so qualitative comparisons are largely preserved. We nonetheless prefer the path-length framing for two reasons. First, the kinetic-energy interpretation imports a particle-mechanics picture in which a unit-mass particle travels through the velocity field. The flow-matching sampling process does not strictly satisfy this assumption, since the local density along the trajectory varies with time. The geometric reading via path length sidesteps this idealisation. Second, the geometric reading offers a clearer interpretation of recent findings: Li et al. (2026) report that higher kinetic energy is associated with sparsity in the data distribution, an observation that is naturally explained as longer geometric paths when data points are far from the prior or are more spread out, as their toy examples illustrate.

4.2. Forward and Reverse Path-Length Distributions

We compute path-length distributions in two complementary ways:

1. **Reverse path length** (data $\to$ noise): for samples $\mathbf{x} \sim p_{\text{data}}$, we integrate the velocity field from $t = 1$ back to $t = 0$ and record $L(\mathbf{x})$.

2. **Forward path length** (noise $\to$ data): for samples $\mathbf{z} \sim \mathcal{N}(0, \mathbb{I})$, we integrate forward from $t = 0$ to $t = 1$ to obtain $\mathbf{x}_1$ and record $L(\mathbf{z})$.

For a fixed pair $(\mathbf{z}, \mathbf{x}_1)$ on the same ODE trajectory, forward and reverse integrations yield identical path lengths up to

numerical error. Aggregated over many samples, however, the two distributions need not coincide: the forward distribution starts from $\mathcal{N}(0, \mathbb{I})$ and lands on whatever the model produces, whilst the reverse distribution starts from real data and lands on whatever the model maps it to. We use the comparison between these two aggregate distributions, together with their evolution across training, as our central diagnostic.

Numerical integration error is a natural confound: if forward and reverse ODE solves accumulate different errors, any divergence between the two distributions could be an artefact of the solver rather than the learned dynamics. We control for this by chaining solvers on the same trajectory in two directions. In one chain, we generate a sample by integrating forward from noise, then immediately reverse-integrating the generated sample back to noise; if numerical error were dominant, the two halves of this round trip would yield noticeably different path lengths. In the second chain, we reverse-integrate a test sample to obtain a noise vector, then forward-integrate from that noise vector. We compare the path-length distributions of the two halves in each chain to assess the magnitude of solver-induced discrepancy.

4.3. Velocity-Norm Profiles

Beyond the scalar path length, we examine the velocity-norm profile across discretisation steps:

$$v(t) = \mathbb{E}_{\mathbf{x}} \left[\|\mathbf{v}_\theta(\mathbf{x}_t, t)\| \right]. \tag{9}$$

This shows where along the trajectory any path-length discrepancy is concentrated, which we use to test whether the gap arises from localised effects (e.g., velocity magnitudes spiking near the data) or whether it is distributed over the full integration interval.

5. Experiments

5.1. Setup

We trained flow-matching models on MNIST, comprising 60,000 training images and 10,000 test images. To study how path-length distributions evolve with model quality, we trained three models under different computational budgets: 30 epochs (limited training), 80 epochs (intermediate), and 100 epochs (extended). The corresponding FID scores (Table 1) span FID $\approx$ 130, 13, and 8 respectively, covering a wide range of generative quality.

For each model, we computed path-length distributions over (i) training samples mapped to noise via the reverse ODE, (ii) test samples mapped to noise via the reverse ODE, and (iii) freshly generated samples obtained by integrating the forward ODE from $\mathcal{N}(0, \mathbb{I})$. We additionally measured velocity-norm profiles across discretisation steps and computed nearest-neighbour distance ratios (NNDR) using both DINOv2 features and pixel-level distance to assess memorisation.

5.2. Path-Length Distributions Across Training Budgets

Figure 1 and Table 1 reveal a clear three-regime pattern.

30 epochs (FID ≈ 130): generated paths are shorter than data paths. Generated samples have mean path length 3.6824 ± 0.0915, well below the train (4.6047 ± 0.1206) and test (4.6247 ± 0.1168) baselines. The under-trained velocity field can not traverse the geometric distance required to bridge the data manifold and the noise distribution; new samples are pushed to nearby low-density regions rather than to high-quality data points.

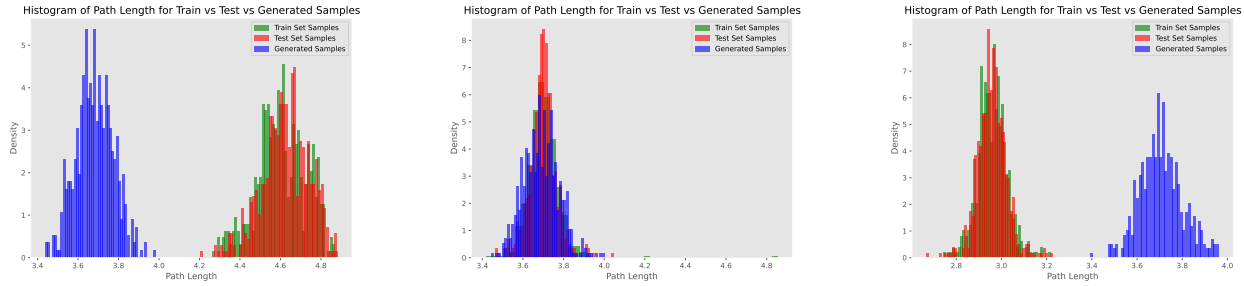
80 epochs (FID ≈ 13): generated and data path lengths match. The three distributions are nearly indistinguishable: generated 3.6892 ± 0.0874, train 3.7001 ± 0.0878, test 3.6987 ± 0.0672. Forward and reverse trajectories cover the same geometric distance, consistent with generation having landed on the same manifold the data inhabits.

100 epochs (FID ≈ 8): generated paths exceed data paths. Although FID continues to improve, the generated path-length distribution (3.7070 ± 0.0941) now exceeds train (2.9585 ± 0.0647) and test (2.9600 ± 0.0657) by a substantial margin. Crucially, this is because the data path-length distribution has shrunk further whilst the generated distribution remained unchanged. If the true data on the manifold can be mapped to noise via shorter paths, the longer generated paths suggest that generation no longer fully latches onto the same manifold.

5.3. Stability of the Generated Path-Length Distribution

A noteworthy feature of Table 1 is that the generated path length is essentially constant across the three budgets (≈ 3.68 to 3.71), even as FID improves by more than an order of magnitude. The training and test path lengths, in contrast, decrease monotonically from ≈ 4.60 at 30 epochs to ≈ 2.96 at 100 epochs. The crossover between the two occurs around 80 epochs, the regime where path-length matching is observed.

The stability of the generated distribution has two implications. First, the three regimes we identified are produced not by a moving generated distribution but by a moving data distribution that crosses a near-static generated one. Second, the geometric structure of forward dynamics, from noise to data, appears to settle early in training and to remain effectively fixed thereafter. We return to this observation in the discussion.



(a) 30 Epochs *(b)* 80 Epochs *(c)* 100 Epochs

Figure 1. Comparison of path length distributions for Train, Test, and Generated samples across different computational budgets. Note that for the early-stage model (30 epochs), the data path lengths are significantly higher than the generated distribution's path length, while for the fully trained model (100 epochs), they tend to be lower.

5.4. Visual Quality of Generated Samples

Figure 2 shows samples from each model. The 30-epoch model produces visibly degraded samples, consistent with its high FID. However, samples from the 80-epoch and 100-epoch models are of comparable visual quality. Path length captures geometric information about the learned transport that is not directly visible in sample appearance, and path-length matching may serve as a more nuanced indicator of generalisation than visual inspection.

We also observe that across the three models, the same noise vector tends to produce visually similar generated images. This is consistent with prior reports that flow-matching trajectories are largely determined by initialisation and that learned velocity fields exhibit substantial agreement across models, even under different conditional path formulations (Briq et al.; English & Suzuki). Our path-length results refine this picture: the geometric structure of forward dynamics is stable across training budgets, whilst the reverse dynamics on real data continue to evolve.

5.5. Memorisation Control

To rule out memorisation as an explanation, we computed nearest-neighbour distance ratios (NNDR) using DINOv2 features and pixels as features (Table 1). The mean DINOv2 NNDR changes only marginally across budgets, from 0.9482 ± 0.0006 at 30 epochs to 0.9392 ± 0.0003 at 100 epochs. The minimum ratio (capturing potential exact copies) remains stable, and pixel-level NNDR shows a comparable pattern around 0.95. None of the three models exhibits substantial memorisation, and the path-length differences across regimes cannot be attributed to copying. Path length thus tracks something orthogonal to memorisation: the geometric properties of the learned transport rather than its proximity to training data.

5.6. Ablation: Numerical Integration Error

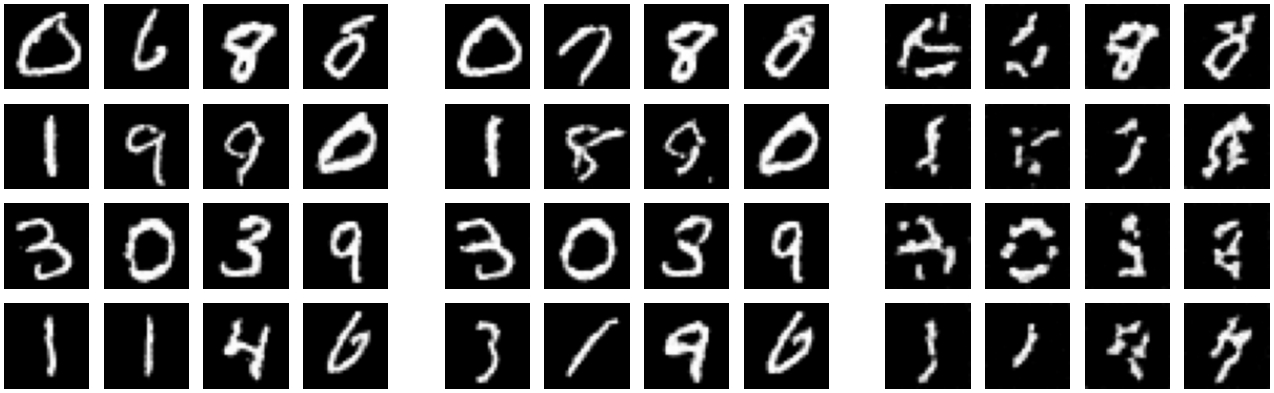
A natural concern is whether the path-length discrepancy at 100 epochs arises from numerical errors in ODE integration. We performed two complementary controls (Figure 3).

In the first, we integrated the forward ODE from noise to obtain generated samples, then applied the reverse ODE to those generated samples and compared the path-length distributions of the two halves of the round trip. We observed no meaningful difference between forward and reverse path lengths when starting from generated samples. In the second, we took test samples, applied the reverse ODE to obtain corresponding noise vectors, and then ran the forward ODE from these noise vectors; the two halves again produced nearly identical distributions, overlapping so closely that they appear as a single distribution in Figure 3b. This pattern persists across all three computational budgets. Numerical integration error therefore cannot explain the path-length discrepancy observed at 100 epochs; the discrepancy reflects a genuine difference in the learned geometric structure of forward (noise-to-data) and reverse (data-to-noise) dynamics on freshly generated samples.

5.7. Velocity-Norm Profiles

To localise the source of the path-length discrepancy, we examined the velocity-norm profiles across discretisation steps (Figures 4, 6, 5).

A natural hypothesis is that relatively longer generated paths at 100 epochs result from elevated velocity magnitudes near the data end of the trajectory, as the model commits to a final sample. The profiles do not support this localised explanation: for the 100-epoch model, the generated velocity magnitude remains high throughout the integration interval, not just near the data. The same broad pattern holds across all three budgets, mirroring the stability of the scalar-generated path length in Table 1.



(a) Model trained until 100 Epochs *(b)* Model trained until 80 Epochs *(c)* Model trained until 30 Epochs

Figure 2. Samples from three generative models, each trained under different training budgets.

A separate observation concerns the structure of the true-data velocity profiles. For both training and test samples, the velocity magnitude is substantially more variable near the noise end of the reverse ODE than near the data end. Generated velocity profiles never display this variability near noise. This asymmetry suggests that the model encodes different geometric structures at the two ends of the trajectory: the reverse direction from data permits diverse paths into noise, whilst the forward direction from noise follows a more constrained, deterministic route into data.

Together with the stability of generated path lengths across budgets, these observations suggest that the geometric properties of flow-matching generation are largely determined early in training and remain fixed thereafter. Improvements in FID at later training stages are driven by refinements in the reverse dynamics on real data rather than by changes in forward dynamics.

6. Discussion

6.1. What Does Path-Length Behaviour Reveal?

Our findings are driven by a single asymmetry: data path lengths shrink monotonically as training progresses, whilst generated path lengths remain essentially fixed. The three regimes we identified (short, matched, long paths) are produced not by a moving generated distribution but by the data distribution sweeping past a near-static generated one.

This raises a question that our diagnosis alone cannot resolve. At 80 epochs, the generated and data path lengths match, which is consistent with generation having landed on the same geometric manifold the data inhabits. At 100 epochs, FID continues to improve, and NNDR shows no increase in memorisation, yet the data path lengths have shrunk further, whilst generated path lengths have not. One reading is that generation now fails to reach the refined data manifold; another is that the shrinking data path lengths represent a form of geometric overfitting that FID and nearest-neighbour metrics do not detect. We do not adjudicate between these readings here, but we note that the disagreement between path-length matching and FID at 100 epochs is itself informative: the diagnostic captures something orthogonal to standard quality metrics.

6.2. Are Shrinking Data Path Lengths a Geometric Form of Overfitting?

A separate reading of our results foregrounds the monotonic shrinkage of train and test path lengths as training continues, from ≈ 4.60 at 30 epochs to ≈ 2.96 at 100 epochs. The model is progressively learning shorter trajectories from data to noise, tightening the geometry of the reverse dynamics around the empirical distribution. In a classical setting, this would suggest overfitting. However, FID continues to improve from ≈ 13 to ≈ 8 over the same interval, and NNDR remains essentially flat, so neither generation quality nor memorisation gives evidence of overfitting in the conventional sense.

This suggests one of two interpretations. Either the path-length shrinkage is a benign sign of an improving fit, in which case the matched regime at 80 epochs is incidental rather than diagnostic, or path length detects a geometric form of overfitting that proximity-based memorisation metrics and feature-space distances do not see. Distinguishing these will require studying path-length behaviour as a function of model capacity, dataset size, and training duration. In particular, we conjecture that for larger or smaller models, the crossover point between data and generated path lengths may shift, and the rate of data path-length shrinkage may correlate with capacity. We leave a systematic study to future work.

Table 1. Quantitative comparison of model performance across training stages (30, 80, and 100 epochs). We report FID scores, Nearest Neighbor Distance Ratio (NNDR) using DINOv2 features and pixel-level fitting, and Path Length statistics.

METRIC	30 EPOCHS	80 EPOCHS)	100 EPOCHS
FID SCORE ↓	129.9912	13.2540	8.2292
AVG. NNDR (FITTING)	0.9753	0.9539	0.9507
DINOV2 METRICS			
MEAN NNDR	0.9482 ± 0.0006	0.9446 ± 0.0008	0.9392 ± 0.0003
MIN RATIO (EXACT COPY)	0.6022 ± 0.0228	0.6157 ± 0.0026	0.5954 ± 0.0375
PATH LENGTH			
TRAIN SET BASELINE	4.6047 ± 0.1206	3.7001 ± 0.0878	2.9585 ± 0.0647
TEST SET BASELINE	4.6247 ± 0.1168	3.6987 ± 0.0672	2.9600 ± 0.0657
GENERATED SAMPLES	3.6824 ± 0.0915	3.6892 ± 0.0874	3.7070 ± 0.0941

6.3. Connection to Trajectory Stability

Prior work has reported that flow-matching models tend to produce similar samples from the same noise initialisation, even across architectures, training procedures, and conditional path formulations (Briq et al.; English & Suzuki). This stability has been documented at the level of generated outputs. Our findings extend this observation to the geometric structure of the trajectories themselves: the generated path-length distribution is essentially constant across training budgets, even as FID improves substantially and the data path lengths shrink considerably. We hypothesise that this stability reflects inductive biases of the neural network architecture, with the velocity field settling into a particular geometric structure for the forward direction early in training and subsequent training refining the reverse direction without meaningfully altering the forward one.

The velocity-norm profiles in Section 5.7 sharpen this picture: the reverse direction from data permits diverse paths closer to noise, whilst the forward direction from noise follows a more constrained, near-deterministic route into data. The stability documented at the level of generated outputs in prior work may therefore be a downstream consequence of a tighter geometric structure on the forward dynamics, with the freedom of the reverse dynamics largely invisible to output-level comparisons.

6.4. Connection to Implicit Regularisation

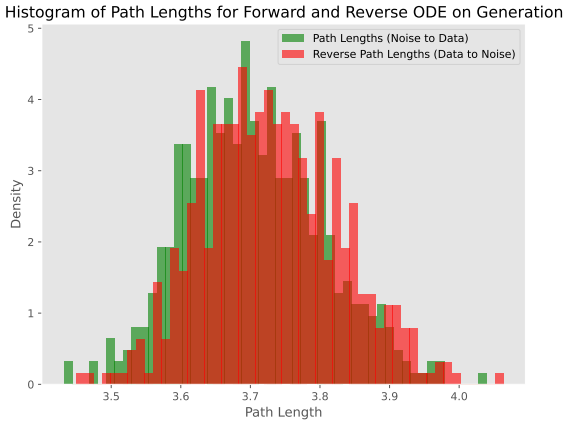
A natural mechanistic question is what produces this stability. Bertrand et al. (2026) showed that practical training does not fully minimise the conditional flow-matching objective and that the resulting deviation from the closed-form optimum $\hat{u}^\star$ in Equation 5 is implicitly regularised differently at different time points, preventing memorisation. We conjecture that this same implicit regularisation is responsible for the path-length stability we observe. A model that fully minimises the closed-form objective would produce trajectories tightly constrained by the empirical distribution, and these trajectories would shorten as the data path lengths refine. The fact that generated path lengths remain stable whilst data path lengths shrink suggests that the optimisation never reaches the regime where forward and reverse dynamics are forced to be geometrically symmetric. The implicit regularisation that grants flow matching its generalisation properties may therefore also produce the asymmetry that our diagnostic reveals.

6.5. Limitations

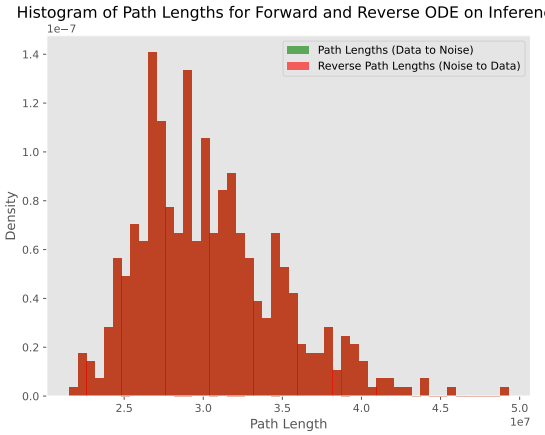
We acknowledge several limitations. Our experiments are conducted on MNIST with a single U-NET-based architecture; whether the path-length pattern persists on natural image datasets, on higher-resolution data, or across other conditional path formulations is an open question. We do not provide a formal analysis of when path-length matching can be expected, although the connection to closed-form flow matching offers a starting point. Finally, we do not yet know whether the relatively longer/shorter generated paths at 100 epochs correspond to specific qualitative features of the generated samples; a careful per-sample analysis correlating path length with sample quality, novelty, or classifier confidence is left for future work.

6.6. Conclusion

We have introduced path length as a cheap, model-internal diagnostic for flow-matching models. On MNIST, three regimes emerge across training budgets: long data paths at high FID, matched paths at moderate FID, and short data paths at low FID. The pattern is driven by data path lengths shrinking around an essentially fixed generated distribution, and it is not explained by memorisation, numerical integration error, or localised velocity-magnitude effects. We connect the stability of generated path lengths to recent results on trajectory stability and implicit regularisation in flow matching. Path length is not a replacement for FID or memorisation metrics, but a complementary geometric signal that is straightforward to compute and interpret. We



(a) Generation: Histogram of path lengths for forward and reverse ODE.



(b) Inference: Histogram of path lengths for forward and reverse ODE.

Figure 3. Analysis of path length distributions. Top: Distributions during the generation phase. Bottom: Distributions during the inference phase.

see particular promise in its use for early stopping, model diagnosis, and as motivation for path-length-aware regularisation strategies.

References

Albergo, M. S. and Vanden-Eijnden, E. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

Bertrand, Q., Gagneux, A., Massias, M., and Emonet, R. On the closed-form of flow matching: Generalization does not arise from target stochasticity. *Advances in neural information processing systems*, 38:8522–8549, 2026.

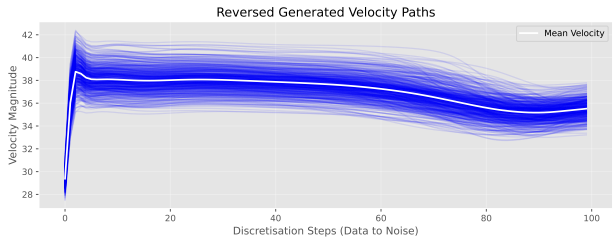
Briq, R., Kamp, M., Fried, O., Cohen, S., and Kesselheim, S. The amazing stability of flow matching. In *EurIPS 2025 Workshop on Principles of Generative Modeling (PriGM)*.

English, E. and Suzuki, T. Path invariance and the robustness of flow matching: Beyond architectural and data perturbations. In *ICLR 2026 2nd Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy*.

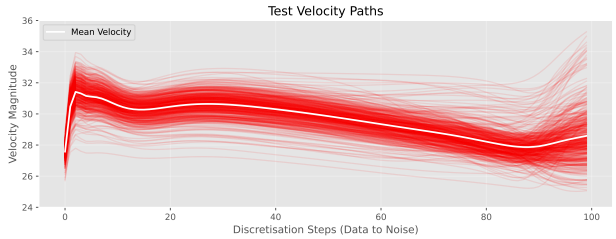
Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35: 26565–26577, 2022.

Li, Z., Hu, H., Lim, S. H., Li, X., Gao, F., Diao, E., Ding, Z., Vazirgiannis, M., and Bostrom, H. A kinetic-energy perspective of flow matching. *arXiv preprint arXiv:2602.07928*, 2026.

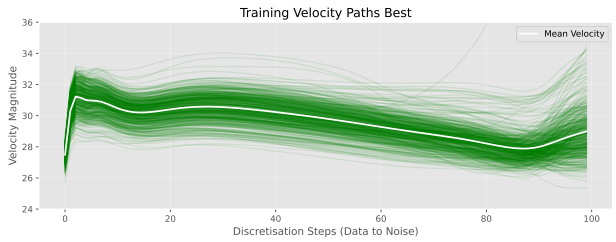
Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.



(a) Generated Velocity Paths



(b) Test Velocity Paths



(c) Training Velocity Paths

Figure 4. Comparison of 100 epoch trained model's velocity magnitude paths across discretisation steps ($0 \rightarrow 100$) for generated, test, and training data. The white line represents the mean velocity profile.

Liu, X., Gong, C., and Liu, Q. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

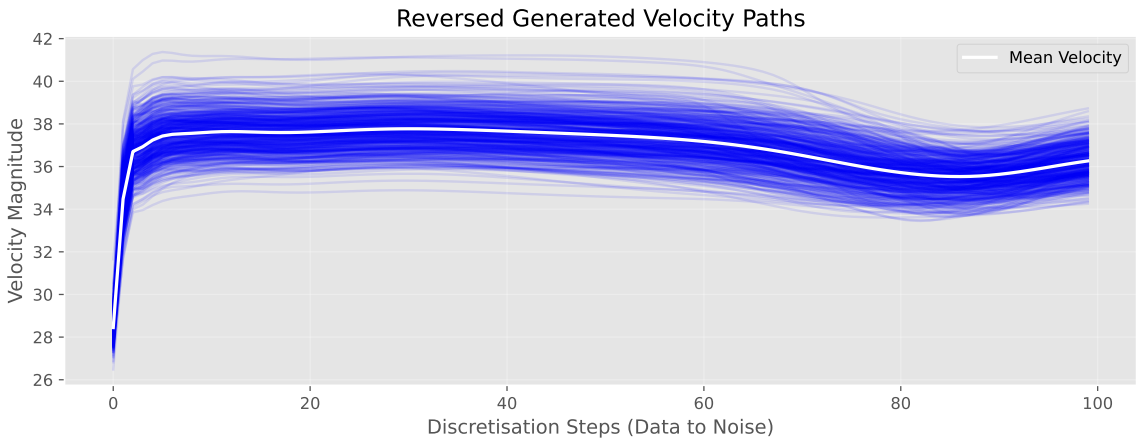
Pidstrigach, J. Score-based generative models detect manifolds. *Advances in Neural Information Processing Systems*, 35:35852–35865, 2022.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

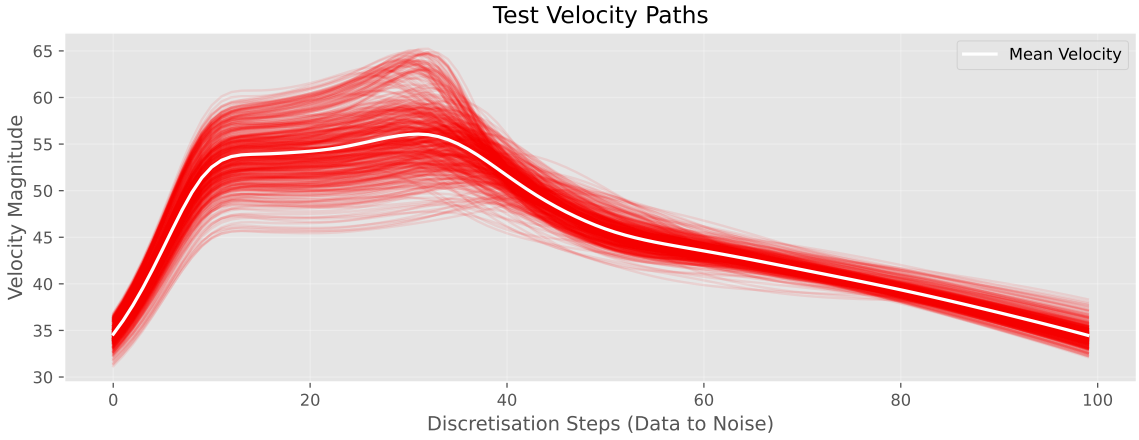
Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. Consistency models. 2023.

Stanczuk, J. P., Batzolis, G., Deveney, T., and Schönlieb, C.-B. Diffusion models encode the intrinsic dimension of data manifolds. In *Forty-first International Conference on Machine Learning*, 2024.

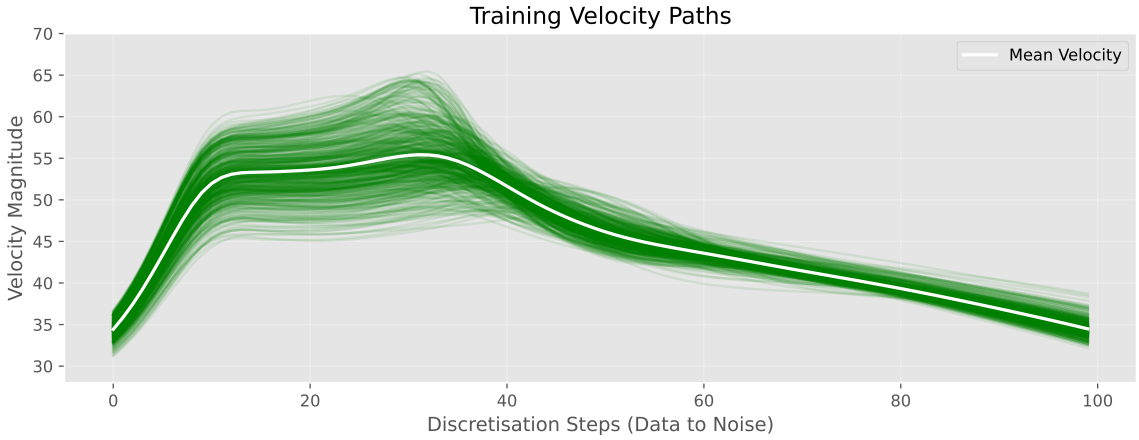
Tong, A., Fatras, K., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Wolf, G., and Bengio, Y. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.



(a) Generated Velocity Paths

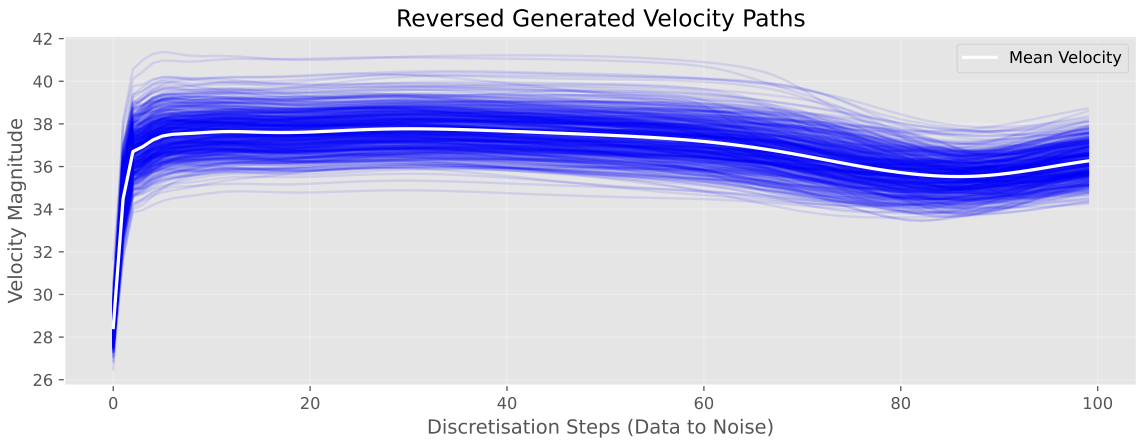


(b) Test Velocity Paths

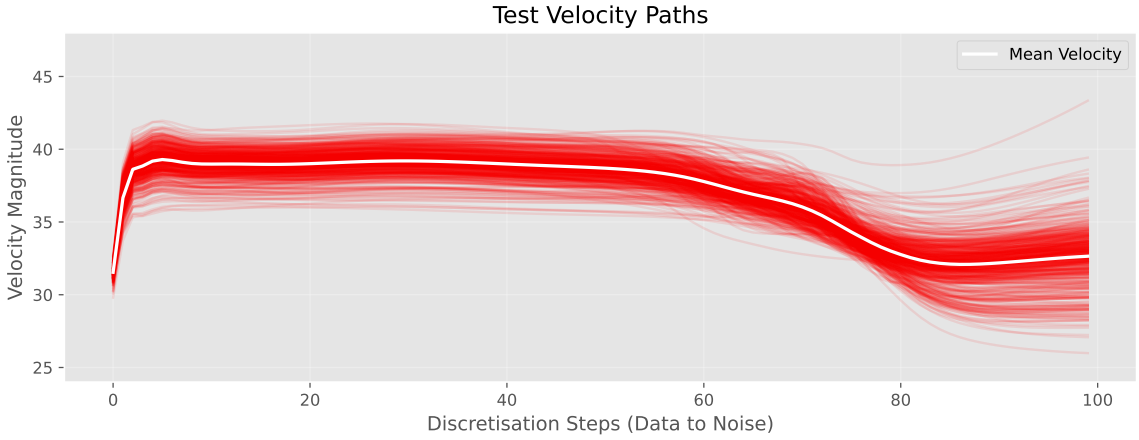


(c) Training Velocity Paths

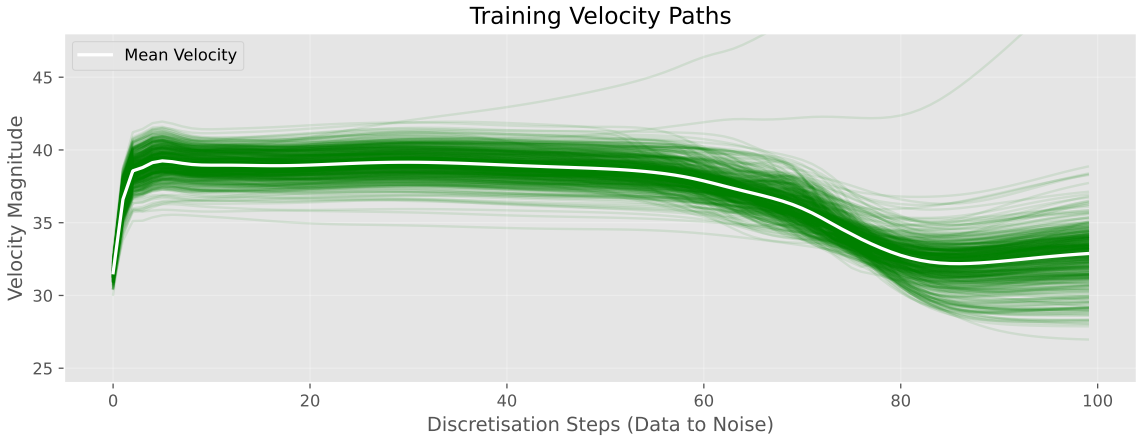
Figure 5. Comparison of 20 epoch trained model's velocity magnitude paths across discretisation steps ($0 \rightarrow 100$) for generated, test, and training data. The white line represents the mean velocity profile.



(a) Generated Velocity Paths



(b) Test Velocity Paths



(c) Training Velocity Paths

Figure 6. Comparison of 80 epoch trained model's velocity magnitude paths across discretisation steps ($0 \rightarrow 100$) for generated, test, and training data. The white line represents the mean velocity profile.