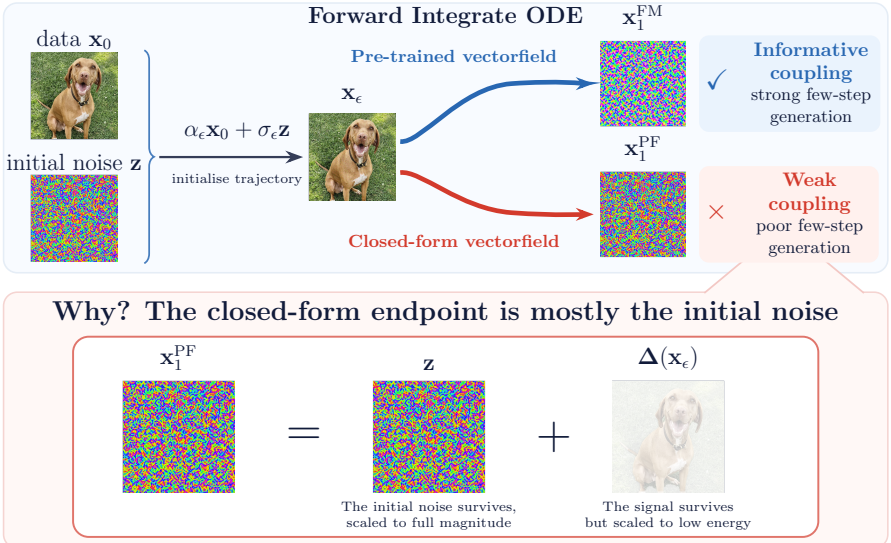
THE LIMITS OF CLOSED-FORM COUPLINGS FOR FEW-STEP GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-step generation remains an open challenge for flow models. Rectified flows address it by distilling a pretrained teacher into a few-step student using the noise–data couplings induced by the teacher. Obtaining the pretrained model and sampling the couplings is expensive. However, for a finite dataset and a Gaussian probability path, the probability-flow vector field of the empirical distribution is available in closed form, so one could use this field to construct the couplings directly and skip pretraining altogether. Surprisingly, this fails, even though the pretrained coupling is almost always compatible with the data-point assignment induced by the closed-form flow. Students trained on closed-form couplings perform much closer to those trained on random noise–data pairings than to those trained on teacher-generated couplings. We trace this failure to an exact endpoint identity for the closed-form flow. The integration returns the noise vector drawn at initialisation, plus a drift that is empirically a faint copy of the data point, almost independent of the noise. The coupling therefore differs from an independent pairing only through this drift. By bounding the drift and combining the bound with the known collapse time of empirical diffusion dynamics, we show that the informative share of the coupling grows with the dataset size N and shrinks with the data dimension d, scaling as $\frac{\log N}{d}$. Growing the dataset therefore helps only logarithmically, while lowering the dimension helps linearly. We confirm this empirically by decreasing the dimension of the data itself, downsampling images to a range of resolutions. The drift share grows as $\frac{1}{d}$, and students trained on closed-form couplings recover an increasing fraction of the gap between random and teacher couplings. However, even at the lowest resolutions the recovery remains incomplete, suggesting that closed-form couplings become competitive teachers only at values of $\frac{\log N}{d}$ far beyond those seen in natural image data.



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.

1 INTRODUCTION

Diffusion and flow-based generative models learn a time-dependent vector field that transports samples from a simple base distribution to the data distribution by integrating an ODE (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021; Lipman et al., 2023). Flow Matching (FM) trains a parametric field by regressing to the conditional velocity of a chosen probability path, enabling simulation-free training. Sampling, however, still requires integrating the ODE over many steps, and reducing this cost remains an active area of research.

One line of work learns a direct map between different points along the generative trajectory, enabling large jumps in time with a single model evaluation, either by distilling a pretrained model or by training from scratch (Salimans & Ho, 2022; Song et al., 2023; Albergo et al., 2025; Boffi et al., 2025). While distillation requires a pretrained teacher, direct training from scratch can be optimisation-sensitive and therefore requires substantial stabilisation and tuning. An alternative approach keeps the FM objective but changes the coupling between noise and data on which it trains. The regression target depends on how the endpoints are paired, and the choice of couplings can yield straighter trajectories that can be integrated in fewer steps. Several approaches construct such couplings using optimal transport (Pooladian et al., 2023; Tong et al., 2024; Mousavi-Hosseini et al., 2026). These couplings require no pretrained model, but the improvement they yield is limited. The most prominent approach, rectified flows (Liu et al., 2023), instead obtains couplings from a pretrained teacher. In this case, each noise sample is integrated through the teacher's probability-flow ODE to generate a datapoint, and a student is trained with FM on the resulting pairs. Recent improvements have shown that a single rectification step can be enough for state-of-the-art few-step generation (Lee et al., 2024; Kim et al., 2025). However, this procedure remains expensive, as it requires pretraining the teacher.

This raises a natural question: if rectification benefits from "good" couplings, could we avoid pretrained couplings entirely by using closed-form couplings? Concretely, given a Gaussian probability path and a dataset defining the target marginal, one can derive an "empirical" optimal regression target (vector field) for FM/diffusion models (Bertrand et al., 2025; Gao & Li, 2026; Scarvelis et al., 2025). Park et al. (2026) successfully exploit this idea for discrete flow models using closed-form inversion, suggesting that a similar shortcut might be possible for rectification in the continuous case.

However, the empirical vector field does not specify a one-to-one pairing between data points and noise. The induced ODE dynamics are strongly contracting as $t \to 0$, so entire regions of the prior are driven into the same data point when integrating towards $t = 0$, forming semi-discrete couplings (Biroli et al., 2024; Pierret et al., 2026). As a result, the closed-form field admits many possible noise–data pairings and does not yield a canonical choice of $(\mathbf{x}_0, \mathbf{x}_1)$ training pairs (Figure 4). In this work, we show that the natural way of resolving this ambiguity, by sampling a noisy initial condition near each data point and integrating the closed-form field forward, produces couplings that perform close to random couplings and therefore fail to train a good student. Surprisingly, the pretrained and closed-form constructions largely associate the same data points with the same regions of prior space, but select very different endpoints within those regions. We explain this failure theoretically, identify the two quantities that control it, the dataset size N and the dimension d, and show empirically that by moving the whole construction into lower-dimensional spaces, the closed-form couplings become better teachers.

Our contributions are as follows:

- **Closed-form couplings are poor teachers** (Section 4). Students trained on closed-form probability-flow couplings perform far worse than students trained on pretrained couplings, and only marginally better than students trained on random pairings. We show that this happens, even though the pretrained couplings themselves correspond to one of the pairings admitted by the closed-form field, with more than 99.9% matching accuracy.
- **Closed-form couplings are dominated by a linear noise map** (Section 5). We prove an endpoint decomposition for the closed-form flow (Proposition 5.1). This shows that the terminal coupling point $\mathbf{x}_1$ decomposes into the injected initialisation noise and a trajectory-dependent drift. We further derive an exact cutoff decomposition that bounds this drift in terms of the trajectory's tracking error (Proposition 5.2). Empirically, the drift is small and largely independent of the injected noise, so each datapoint is paired essentially with the

fresh Gaussian sample used to initialise its integration, plus a small datapoint-dependent correction, which explains the poor performance of closed-form couplings.

- **Closed-form couplings require exponentially more data in higher dimensions** (Sections 5.2 and 6). Combining our drift bound with the collapse time of empirical diffusion dynamics predicts that the informative share of the coupling scales as $\frac{\log N}{d}$. Therefore, maintaining a fixed drift share as the dimension grows requires the dataset size to scale exponentially with d. We test the complementary prediction by progressively reducing the image resolution. The drift share increases as $\frac{1}{d}$ and student performance improves, although the gap to pretrained couplings remains substantial even at the lowest resolutions we study.

2 Related Work

Couplings and rectified flows. The choice of couplings used for training can significantly affect the straightness of Flow Matching trajectories. In minibatch optimal-transport methods, non-independent couplings are constructed to obtain straighter flows and therefore faster sampling (Pooladian et al., 2023; Tong et al., 2024), with recent improvements by developing scalable semi-discrete variants (Mousavi-Hosseini et al., 2026). Instead of relying on semi-discrete couplings, rectified flows use couplings obtained from a pre-trained teacher instead (Liu et al., 2023). This idea has enabled high-quality few-step generation (Liu et al., 2024; Lee et al., 2024; Yan et al., 2024; Kim et al., 2025). Based on these advancements, recent theoretical studies investigate when rectification produces straight or optimal couplings (Bansal et al., 2026; Chambolle & Hertrich, 2026). Since the promising works on rectified flows require a pretrained teacher, we ask whether the pretrained model can be removed altogether by obtaining endpoint pairs from the closed-form vector field.

Closed-form empirical flows. For a finite empirical distribution and Gaussian probability path, the optimal diffusion score or Flow Matching velocity can be evaluated analytically. Prior work has used these closed-form dynamics to study memorisation and generalisation (Scarvelis et al., 2025; Bertrand et al., 2025; Gao & Li, 2026; Liu et al., 2026), including the temporal phases of the dynamics and the transition from averaging over many data points to the collapse on individual training examples (Biroli et al., 2024; Gagneux et al., 2025). Recent work further characterises the geometry produced by the exact semi-discrete Flow Matching dynamics (Pierret et al., 2026). In contrast, we ask whether the endpoint coupling obtained from the empirical probability flow can itself serve as a teacher for few-step generation. We show that its datapoint-dependent information is confined to a small drift term and characterise when this signal becomes large enough to improve over random pairings.

3 Background

Let $\mathbf{x}_0 \sim p_0$ denote the data distribution on $\mathbb{R}^d$. We consider the Gaussian conditional path

$$p_t(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\big(\mathbf{x}_t \mid \alpha_t \mathbf{x}_0, \sigma_t^2 \mathbf{I}\big) \qquad t \in (0, 1] \,, \tag{1}$$

for differentiable schedules $\alpha_t, \sigma_t \in \mathbb{R}^*$ which induce marginals $p_t(\mathbf{x}) = \int p_t(\mathbf{x} \mid \mathbf{x}_0) p_0(\mathbf{x}_0) \mathrm{d}\mathbf{x}_0$. A velocity field transporting these marginals can be obtained from a conditional velocity field that, together with $p_t(\mathbf{x}_t \mid \mathbf{x}_0)$, satisfies the continuity equation. For Equation (1), the conditional velocity field is

$$\mathbf{v}_t(\mathbf{x}_t \mid \mathbf{x}_0) = \frac{\mathrm{d}\alpha_t}{\mathrm{d}t}\, \mathbf{x}_0 + \frac{\mathrm{d}\log \sigma_t}{\mathrm{d}t}(\mathbf{x}_t - \alpha_t \mathbf{x}_0) \,. \tag{2}$$

The marginal probability-flow vector field is given by the posterior average $\mathbf{v}_t(\mathbf{x}) = \int \mathbf{v}_t(\mathbf{x} \mid \mathbf{x}_0)\, p_t(\mathbf{x}_0 \mid \mathbf{x})\, \mathrm{d}\mathbf{x}_0$. For an empirical distribution $\hat{p}_0(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N \delta(\mathbf{x} - \mathbf{x}_0^i)$, this integral reduces to the finite sum

$$\mathbf{v}_t^{PF}(\mathbf{x}) = \sum_{i=1}^N w_t^i(\mathbf{x}) \Big(\frac{\mathrm{d}\alpha_t}{\mathrm{d}t} \mathbf{x}_0^i + \frac{\mathrm{d}\log \sigma_t}{\mathrm{d}t} \big(\mathbf{x} - \alpha_t \mathbf{x}_0^i\big) \Big) \,, \qquad w_t^i(\mathbf{x}) = \frac{\mathcal{N}(\mathbf{x} \mid \alpha_t \mathbf{x}_0^i, \sigma_t^2 \mathbf{I})}{\sum_{j=1}^N \mathcal{N}(\mathbf{x} \mid \alpha_t \mathbf{x}_0^j, \sigma_t^2 \mathbf{I})} \,. \tag{3}$$

*We assume $\sigma_t > 0$ for $t \in (0, 1]$, with $\alpha_0 = 1$, $\sigma_0 = 0$ and $\alpha_1 = 0$, $\sigma_1 = 1$.

Since $\mathbf{v}_t^{PF}$ is induced by the empirical data distribution, its reverse dynamics are strongly contracting as $t \to 0$. In particular, starting from a point in the Gaussian prior at $t = 1$ and integrating the exact field backwards to $t = 0$ sends it to one of the data points in the support of $\hat{p}_0$. Hence, whole regions of prior space can contract onto the same training example (Figure 4). In practice, Flow Matching approximates $\mathbf{v}_t^{PF}$ with a parametric model $\mathbf{v}^{FM}(\mathbf{x}, t)$ by least-squares regression to a Monte Carlo draw of the conditional velocity. Specifically, sample $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, set $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \mathbf{z}$, and use the target $\frac{d\mathbf{x}_t}{dt} = \frac{\mathrm{d}\alpha_t}{\mathrm{d}t}\mathbf{x}_0 + \frac{\mathrm{d}\sigma_t}{\mathrm{d}t}\mathbf{z}$. Theoretically, a model with sufficient capacity would eventually converge to Equation (3) (Yi et al., 2023; Kamb & Ganguli, 2025).

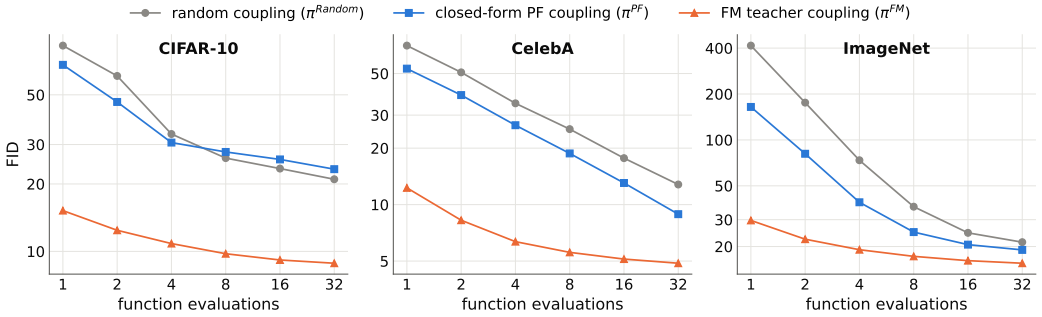


Figure 1: **FID against sampling steps for students trained on π^{PF}, π^{FM} and π^{Random}.** For a fair comparison, each model is trained exactly on the N forward couplings permitted by the datasets. This limits the FID value, but enables us to provide a clean comparison between couplings. Although the gap narrows as the number of integration steps increases, it is substantial in the few-step regime.

4 Closed-form *PF-ODE* couplings are poor teachers

A natural way to avoid pretraining a Flow Matching teacher is to construct couplings directly from the closed-form probability-flow field introduced in Equation (3). Surprisingly, this construction fails in practice. As shown in Figure 1, students trained on these couplings perform much closer to students trained on random noise-data pairings than to those trained on pretrained-model couplings.

To construct a coupling for each datapoint $\mathbf{x}_0^i$, we sample from the conditional path in Equation (1) at a small time $\epsilon > 0$, avoiding the singularity at $t = 0$:

$$\mathbf{x}_\epsilon = \alpha_\epsilon \mathbf{x}_0 + \sigma_\epsilon \mathbf{z}, \quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \, . \tag{4}$$

Starting from $\mathbf{x}_\epsilon^i$, we integrate either $\mathbf{v}^{PF}$ or $\mathbf{v}^{FM}$ forward from $t = \epsilon$ to $t = 1$, obtaining the endpoint $\mathbf{x}_1^i$. We use $\epsilon = 10^{-3}$ unless stated otherwise. This gives $\pi^{(\cdot)} = \{(\mathbf{x}_0^i, \mathbf{x}_1^i)\}_{i=1}^N$, with marginals $\pi_0^{(\cdot)} = \{\mathbf{x}_0^i\}_{i=1}^N$ and $\pi_1^{(\cdot)} = \{\mathbf{x}_1^i\}_{i=1}^N$. Here, π^{PF} denotes couplings generated with the closed-form field and π^{FM} those generated with the pretrained Flow Matching model. Random couplings π^{Random} are obtained by independently pairing datapoints with Gaussian noise samples. We use exaclty N forward couplings in all cases, such that the data marginal $\pi_0^{(\cdot)}$ is identical across constructions and differences can be attributed to the induced pairing. Further details are given in Appendix E.

Figure 1 compares rectified-flow students trained with the Flow Matching objective on the three coupling sets π^{PF}, π^{FM}, and π^{Random}. Within each dataset, all students are trained on the same number of couplings, with the same architecture and optimisation settings. The only difference is that, on CIFAR-10 and CelebA, students on π^{FM} are trained for fewer steps, since their FID stops improving earlier (Appendix E). The difference is most pronounced at a single sampling step, where students trained on π^{PF} have roughly 4–5$\times$ higher FID than those trained on π^{FM} across datasets. This gap is particularly striking because $\mathbf{v}^{PF}$ is the exact probability-flow field of the empirical distribution, whereas $\mathbf{v}^{FM}$ is only learned as an approximation to it. We therefore ask whether the pretrained endpoints nevertheless fall within the same datapoint-associated regions as those induced by the closed-form *PF-ODE*.

The large performance gap could suggest that the pretrained model learns a fundamentally different correspondence between data and prior space. To test this, we take the endpoints produced by

$\mathbf{v}^{FM}$ and integrate them backwards with the closed-form field $\mathbf{v}^{PF}$. On all three datasets the trajectory returns to the data point from which the pretrained coupling originated in more than 99.9% of the cases (Appendix B). Thus, although $\mathbf{v}^{FM}$ and $\mathbf{v}^{PF}$ define different forward maps, the pretrained endpoints almost always lie within the closed-form PF-ODE region associated with the same datapoint. Combined with the observation above, this suggests that the pretrained model selects a much narrower subset within the broader regions admitted by the closed-form flow. While $\mathbf{v}^{PF}$ is highly sensitive to the initial perturbation, $\mathbf{v}^{FM}$ maps different perturbations of the same datapoint to much more similar endpoints (Table 1).

The results from this section isolate the source of the failure. Although the closed-form and pretrained models induce different forward maps, their endpoints are largely associated with the same datapoint regions in prior space. The difference seems to lie in how they select endpoints within these regions. The closed-form construction is highly sensitive to the initial perturbation and appears to produce a particularly uninformative pairing. In the next section, we characterise this selection explicitly and show that the endpoint is dominated by the noise used to initialise the integration.

5 An exact endpoint decomposition: the flow returns the initialisation noise

Following the empirical results, we now show that the endpoint of the closed-form flow admits a simple decomposition. The integration returns the injected noise $\mathbf{z}$ (Equation (4)) plus a nonlinear data point dependent drift $\mathbf{\Delta}(\mathbf{x}_\epsilon)$, which captures how the trajectory deviates from $\mathbf{x}_0$ over time. This drift term is empirically small, carrying less than 0.6% of the endpoints' squared norm in all datasets we investigate, and preserves the structure of the originating data point. Thus, the closed-form coupling is close to an independent noise-data pairing, with the drift providing the datapoint-dependent correction.

We now derive this decomposition exactly. First, we split the closed-form field into a part that is linear in $\mathbf{x}$ and a nonlinear remainder. Using $\sum_i w_t^i(\mathbf{x}) = 1$, Equation (3) can be rewritten as

$$\mathbf{v}_t^{PF}(\mathbf{x}) = \frac{\mathrm{d}\log\sigma_t}{\mathrm{d}t}\mathbf{x} + \gamma_t m_t(\mathbf{x}) \;, \quad \gamma_t := \frac{\mathrm{d}\alpha_t}{\mathrm{d}t} - \alpha_t\frac{\mathrm{d}\log\sigma_t}{\mathrm{d}t} \;, \quad m_t(\mathbf{x}) := \sum_{i=1}^{N} w_t^i(\mathbf{x})\,\mathbf{x}_0^i \;, \tag{5}$$

where $m_t(\mathbf{x})$ is the posterior mean of the data given $\mathbf{x}_t = \mathbf{x}$. The first term is linear in $\mathbf{x}$ and depends on the data only through the schedule. All nonlinearity of the field is carried by the posterior mean. The linear part can be integrated exactly, which is what produces the following identity.

Proposition 5.1 (informal)**.** *Let $(\mathbf{x}_t)_{t\in[\epsilon,1]}$ solve $\frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t} = \mathbf{v}_t^{PF}(\mathbf{x}_t)$ with initial condition $\mathbf{x}_\epsilon = \alpha_\epsilon\mathbf{x}_0 + \lambda\,\sigma_\epsilon\mathbf{z}$, $\lambda \in [0,1]$. Then, with γ_t and m_t as in Equation (5),*

$$\mathbf{x}_1 = \lambda\mathbf{z} + \mathbf{\Delta}(\mathbf{x}_\epsilon) \;, \qquad \mathbf{\Delta}(\mathbf{x}_\epsilon) := \int_\epsilon^1 \frac{\gamma_t}{\sigma_t}\left(m_t(\mathbf{x}_t) - \mathbf{x}_0\right)\mathrm{d}t \;. \tag{6}$$

The formal proposition with full proof is given in Appendix C. The injected noise therefore survives exactly as an additive term at the endpoint. Although it enters the initial condition only through the small perturbation $\lambda\sigma_\epsilon\mathbf{z}$, its endpoint contribution is $\lambda\mathbf{z}$, corresponding to an amplification by $\frac{1}{\sigma_\epsilon}$. For our choice $\sigma_t = t$ and $\epsilon = 10^{-3}$, this factor is 10^3, explaining the strong sensitivity to even small amounts of initialisation noise. This exact recovery of the injected noise leaves the drift as the only component which also depends on the data point. We therefore turn to $\mathbf{\Delta}(\mathbf{x}_\epsilon)$ and ask what information it carries. Visually, it resembles a faint copy of the corresponding data point (Figure 2, with further examples in Appendix A), suggesting that it is largely determined by $\mathbf{x}_0$ rather than by the particular noise realisation. We first isolate this component by setting $\lambda = 0$ and removing the noise from the integration, then test its dependence on the injected noise, and finally measure its magnitude relative to the full endpoint.

The case $\lambda = 0$. Setting $\lambda = 0$ removes the injected noise and gives $\mathbf{x}_1 = \mathbf{\Delta}(\alpha_\epsilon\mathbf{x}_0)$, the drift evaluated along the mean-initialised trajectory. This provides a direct way to visualise the data point dependent part of the endpoint. As shown in Figures 2 and 5 and Table 2(b), starting from $\mathbf{x}_\epsilon = \alpha_\epsilon\mathbf{x}_0$

produces structure-preserving endpoints rather than Gaussian noise. Therefore, while setting $\lambda = 0$ removes the injected noise that appears to make the coupling difficult to learn from, the resulting endpoints no longer have the required Gaussian marginal and cannot be used as training couplings.

Empirically, the drift depends only weakly on the noise. We test this by integrating the flow twice from every data point, with two independent noise samples $\mathbf{z}, \mathbf{z}' \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and subtracting the injected noise from each endpoint. This gives the two drifts $\boldsymbol{\Delta}(\mathbf{x}_\epsilon) = \mathbf{x}_1 - \mathbf{z}$ and $\boldsymbol{\Delta}(\mathbf{x}'_\epsilon) = \mathbf{x}'_1 - \mathbf{z}'$. A third, noise-free integration from $\mathbf{x}_\epsilon = \alpha_\epsilon \mathbf{x}_0$ gives the *PF-ODE* ($\lambda = 0$) endpoint $\boldsymbol{\Delta}(\alpha_\epsilon \mathbf{x}_0)$. If the drift depended strongly on the noise, the two drifts would differ, and neither would resemble $\boldsymbol{\Delta}(\alpha_\epsilon \mathbf{x}_0)$. However, the drift is closely aligned with the *PF-ODE* ($\lambda = 0$) endpoint, $\cos(\boldsymbol{\Delta}(\mathbf{x}_\epsilon), \boldsymbol{\Delta}(\alpha_\epsilon \mathbf{x}_0)) = 0.94$–$0.95$, and is well described by an amplified copy of it. We found that $\boldsymbol{\Delta}(\mathbf{x}_\epsilon) \approx g\, \boldsymbol{\Delta}(\alpha_\epsilon \mathbf{x}_0)$ with $g \approx 1.2$ on all datasets, which leaves only 12–14% of the drift energy unexplained. The drifts of the two noises agree almost as well, $\cos(\boldsymbol{\Delta}(\mathbf{x}_\epsilon), \boldsymbol{\Delta}(\mathbf{x}'_\epsilon)) = 0.89$–$0.91$. They share the component $g\, \boldsymbol{\Delta}(\alpha_\epsilon \mathbf{x}_0)$ and differ only by a small noise-dependent remainder (Table 2; see Appendix D.1 for a detailed discussion). Empirically, we therefore have the approximation

$$\boldsymbol{\Delta}(\mathbf{x}_\epsilon) \approx g \boldsymbol{\Delta}(\alpha_\epsilon \mathbf{x}_0) \qquad \Longleftrightarrow \qquad \mathbf{x}_1(\lambda, \epsilon) \approx g \boldsymbol{\Delta}(\alpha_\epsilon \mathbf{x}_0) + \lambda \mathbf{z}\,. \tag{7}$$

The flow therefore returns the injected noise unchanged, on top of a drift that is almost the same for every noise (Figure 2).

Empirically, the drift is small. Relative to the Gaussian endpoint, for which $\mathbb{E}\|\mathbf{x}_1\|^2 = d$, it accounts for only 0.52%, 0.14%, and 0.18% of the squared norm on CIFAR-10, CelebA and ImageNet-64, respectively (Table 2)(a). Interestingly, the absolute drift norm remains similar across datasets despite their different dimensionalities (Appendix D.1).

Together with Equation (7), this gives a concrete picture of the closed-form coupling. Each data point is paired with the Gaussian sample drawn to initialise its own integration, plus a weak copy of itself, leaving only a weak datapoint-dependent signal in $\mathbf{x}_1$. This explains why training on π^{PF} performs closer to training on π^{Random} than on π^{FM}.

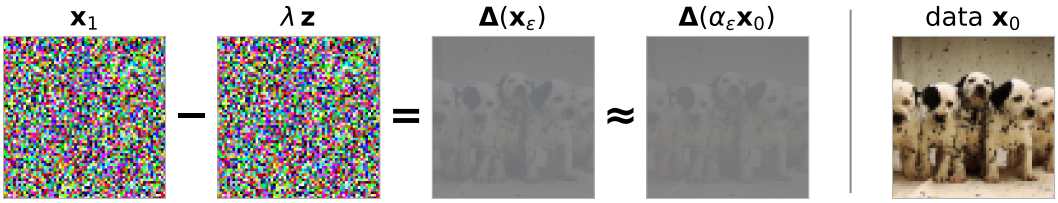


Figure 2: **Endpoint decomposition of an ImageNet-64 integration** ($\lambda = 1$, $\epsilon = 10^{-3}$, closed-form velocity against all 1.28M training images). The endpoint $\mathbf{x}_1$ looks like pure Gaussian noise. Subtracting the injected noise $\lambda \mathbf{z}$ reveals the drift $\boldsymbol{\Delta}(\mathbf{x}_\epsilon)$, a faint copy of the datapoint that keeps most of its structure. The drift is well aligned with the data, with $\cos(\boldsymbol{\Delta}(\mathbf{x}_\epsilon), \mathbf{x}_0) = 0.88$ for this image (0.86 on average). The drift of the noise-free *PF-ODE* ($\lambda = 0$) integration, $\boldsymbol{\Delta}(\alpha_\epsilon \mathbf{x}_0)$, is almost identical to it, with $\cos(\boldsymbol{\Delta}(\mathbf{x}_\epsilon), \boldsymbol{\Delta}(\alpha_\epsilon \mathbf{x}_0)) = 0.96$ (0.95 on average). The injected noise therefore barely changes the information the drift carries about $\mathbf{x}_0$. The datapoint $\mathbf{x}_0$ on the right is shown for reference. All panels use the same colour scale.

5.1 A GENERAL BOUND ON THE DRIFT

Having established the endpoint decomposition and the empirical properties of the drift, we now ask why $\boldsymbol{\Delta}(\mathbf{x}_\epsilon)$ remains small. We derive a general bound showing that the drift is determined by how long the *PF-ODE* trajectory remains close to the conditional path of its datapoint. If it remains close to this path until the signal-to-noise ratio (SNR) is already low, the drift must remain small. Empirically, this is precisely the regime we observe across all datasets, in line with the collapse behaviour identified in previous work (Achilli et al., 2024; Biroli et al., 2024; Gagneux et al., 2025; Liu et al., 2026).

Proposition 5.2 (informal). *Let $(\mathbf{x}_t)_{t\in[\epsilon,1]}$ solve $\frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t} = \mathbf{v}_t^{PF}(\mathbf{x}_t)$ with initial condition $\mathbf{x}_\epsilon = \alpha_\epsilon \mathbf{x}_0 + \sigma_\epsilon \mathbf{z}$ and $\mathbf{r}_t := \frac{1}{\sigma_t}(\mathbf{x}_t - (\alpha_t \mathbf{x}_0 + \sigma_t \mathbf{z}))$ defines the normalised devitation between the conditional and unconditional trajectories. Then, for every $\tau \in (\epsilon, 1)$, the endpoint drift from Equation (6) satisfies*

$$\|\mathbf{\Delta}(\mathbf{x}_\epsilon)\|_{L^2} \leq \|\mathbf{r}_\tau\|_{L^2} + 2\frac{\alpha_\tau}{\sigma_\tau}\sqrt{V}\,, \tag{8}$$

with V being the total variance of the empirical data distribution and $\|\cdot\|_{L^2}$ the root-mean-square norm over $\mathbf{x}_0 \sim \hat{p}_0$ and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

The full proposition statement including proof can be found in Appendix C. The bound separates the drift into a tracking-error term and an SNR-dependent term. In Figure 7 we evaluate these for all datasets on different values of τ (see Appendix D.2). We observe that the bound is dominated by the term $2\frac{\alpha_\tau}{\sigma_\tau}\sqrt{V}$ for most of the trajectory, while the normalised tracking error $\|\mathbf{r}_\tau\|$ stays extremely small. This observation is consistent with previous work, which identified a regime in which the dynamics are dominated by, or even collapse onto, an individual training sample (Biroli et al., 2024; Gagneux et al., 2025). As τ increases, the trajectory starts to deviate from its conditional path while the schedule ratio $\frac{\alpha_\tau}{\sigma_\tau} \to 0$, causing the bound to become increasingly tight.

Together, the bound and our empirical results show that the drift is governed primarily by the SNR at which the trajectory stops tracking its originating data point, scaled by the overall spread of the data. The tracking error itself remains small until this transition. Understanding the drift magnitude therefore reduces to understanding when the trajectory loses track of its originating data point. In the next section, we show how this transition depends on the dataset size and dimension.

5.2 Scaling the drift

The drift is the only component of the endpoint that depends on the data point, so its significance is set by its size relative to the Gaussian noise. As the drift becomes larger relative to the Gaussian component, the datapoint-dependent signal becomes more prominent in the coupling and may therefore be easier for a student to exploit. We now ask how this ratio scales with N and d.

Proposition 5.2 reduces the size of the drift to the signal-to-noise ratio $\frac{\alpha_t}{\sigma_t}$ at which the trajectory stops to track its generating datapoint. Let τ denote the timestep of this transition. As long as the trajectory remains close to its conditional path, $\mathbf{x}_t \approx \alpha_t \mathbf{x}_0^i + \sigma_t \mathbf{z}$, the relative posterior weight of a competing datapoint $\mathbf{x}_0^j$ satisfies

$$\log \frac{w_t^j(\mathbf{x}_t)}{w_t^i(\mathbf{x}_t)} \approx \underbrace{\frac{\alpha_t}{\sigma_t}\mathbf{z}^\top(\mathbf{x}_0^j - \mathbf{x}_0^i)}_{\text{random advantage}} - \underbrace{\frac{\frac{\alpha_t}{\sigma_t}^2}{2}\|\mathbf{x}_0^j - \mathbf{x}_0^i\|^2}_{\text{distance penalty}}. \tag{9}$$

In high dimensions, if pairwise distances concentrate, a typical squared distance satisfies $\|\mathbf{x}_0^j - \mathbf{x}_0^i\|^2 = \Theta(V)$. Hence, conditional on the data points,

$$\mathbf{z}^\top(\mathbf{x}_0^j - \mathbf{x}_0^i) \sim \mathcal{N}\left(0, \|\mathbf{x}_0^j - \mathbf{x}_0^i\|^2\right) \approx \mathcal{N}(0, V)\,, \tag{10}$$

so a typical fluctuation has scale $\sqrt{V}$. Across N competing data points, the largest such fluctuation is of order $\sqrt{V \log N}$. The largest random advantage therefore has scale $\frac{\alpha_t}{\sigma_t}\sqrt{V \log N}$, whereas the distance penalty has scale $\frac{\alpha_t}{\sigma_t}^2 V$. The posterior stops concentrating on the generating data point when these become close, i.e.

$$\frac{\alpha_\tau}{\sigma_\tau}^2 V = \Theta(\log N)\,. \tag{11}$$

Equation (11) predicts the SNR at which the posterior loses track of the generating datapoint. We test this prediction in Figure 8 (left) by varying the data dimension, which in turn changes the total variance V, and measuring the corresponding collapse time. Across CIFAR-10, CelebA and ImageNet-64, the predicted transition closely tracks the measured one. This is the same high-dimensional collapse transition studied by Biroli et al. (2024). Their analysis identifies $\frac{\log N}{d}$ as the parameter

controlling the time when the trajectory collapses to a data point and hence the start of memorisation in the reverse generative dynamics. Related analyses of the closed-form vector field observed the same temporal transition from averaging over multiple training points to dynamics dominated by a single data point (Bertrand et al., 2025; Gagneux et al., 2025; Liu et al., 2026).

Choosing τ at the beginning of this transition, the normalized tracking error $\|\mathbf{r}_\tau\|_{L^2}$ is empirically small (Figure 7) compared with the SNR term $\frac{\alpha_\tau}{\sigma_\tau}\sqrt{V}$. Hence, by Proposition 5.2 and Equation (11) we have

$$\|\mathbf{\Delta}(\mathbf{x}_\epsilon)\|_{L^2} \lesssim \frac{\alpha_\tau}{\sigma_\tau}\sqrt{V} = O(\sqrt{\log N})\ . \tag{12}$$

This characterisation implies that the drift grows only as $\sqrt{\log N}$. Figure 8 (right) tests this prediction directly by varying the size of the empirical distribution (details in Appendix D.3). We found that across all three datasets, the measured drift closely follows $\mathbb{E}\|\mathbf{\Delta}(\mathbf{x}_\epsilon)\| \approx 1.3\sqrt{\log N}$ up to a constant offset.

Finally, the exact probability-flow ODE preserves the prescribed marginals. Since $\alpha_1 = 0$ and $\sigma_1 = 1$, its endpoint satisfies $\mathbf{x}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, and therefore $\mathbb{E}\|\mathbf{x}_1\|^2 = d$. The fraction of endpoint energy carried by the drift is consequently

$$\frac{\mathbb{E}\|\mathbf{\Delta}(\mathbf{x}_\epsilon)\|^2}{\mathbb{E}\|\mathbf{x}_1\|^2} = O\left(\frac{\log N}{d}\right), \tag{13}$$

Equation (13) identifies the lever we could use to increase the informative fraction, as it scales linear in $\frac{1}{d}$ but only logarithmic in N. This motivates our next section, in which we study closed-form couplings on the same datasets for different resolutions, hence controlling d.

6 Reducing the dimension improves closed-form couplings

Equation (13) predicts that the informative share of the closed-form coupling grows linearly in $\frac{1}{d}$ but only logarithmically in N. We test this by downsampling the images and repeating the coupling construction of Section 4, leaving all other parts of the setup unchanged. Concretely, we vary $d = 3r^2$ through the resolution r of the same images, using $r \in \{16, 32, 64\}$ for CelebA and ImageNet and $r \in \{16, 24, 32\}$ for CIFAR-10. Since the dataset size N is fixed across resolutions, it does not confound the comparison. Our goal is not to reach the best possible FID, but to track how closed-form couplings perform relative to random and pretrained couplings as d decreases. Full experimental details are given in Appendix E.

Lower dimension makes the coupling more informative. We first test the prediction of Equation (13) that the drift share grows linearly in $\frac{1}{d}$. As shown in Figure 3 (left), the drift norm barely changes across resolutions, so its share of the endpoint energy grows almost exactly as $\frac{1}{d}$, from $\sim 0.15\%$ at $d = 12288$ to $\sim$ 2–2.5% at $d = 768$. The same trend appears in the coupling statistics. The cosine similarity between each data point and its endpoint rises almost fourfold from 64×64 to 16×16 on CelebA and ImageNet, and doubles from 32×32 to 16×16 on CIFAR-10 (Table 5). This matches the endpoint decomposition. Since $\mathbf{z}$ is independent of $\mathbf{x}_0$, we have $\cos(\mathbf{x}_0, \mathbf{x}_1) \approx \frac{\|\mathbf{\Delta}(\mathbf{x}_\epsilon)\|}{\sqrt{d}} \cos(\mathbf{\Delta}(\mathbf{x}_\epsilon), \mathbf{x}_0)$, which scales as $1/\sqrt{d}$ at fixed drift norm. Meanwhile, the alignment of the drift with its data point, $\cos(\mathbf{\Delta}(\mathbf{x}_\epsilon), \mathbf{x}_0)$, stays nearly constant for each dataset. The drift therefore carries similar information at every resolution, but this information becomes easier to extract as the noise shrinks relative to it. A linear readout from $\mathbf{x}_1$ recovers 39% of the variance of held-out ImageNet images at 64×64 and 56% at 16×16, with the same trend on CelebA and CIFAR-10 (Table 6). Teacher couplings, in contrast, are strongly aligned and linearly decodable at every resolution, and even at 16×16 the closed-form couplings remain below them (Tables 5 and 6).

More informative couplings produce better students. We next test whether the more informative coupling translates into better students. Since FID values are not directly comparable across resolutions, we measure the gap of the closed-form student relative to the random and FM students at the same resolution,

$$G = \frac{\mathrm{FID}_{\mathrm{rnd}} - \mathrm{FID}_{\mathrm{PF}}}{\mathrm{FID}_{\mathrm{rnd}} - \mathrm{FID}_{\mathrm{FM}}}\ , \tag{14}$$

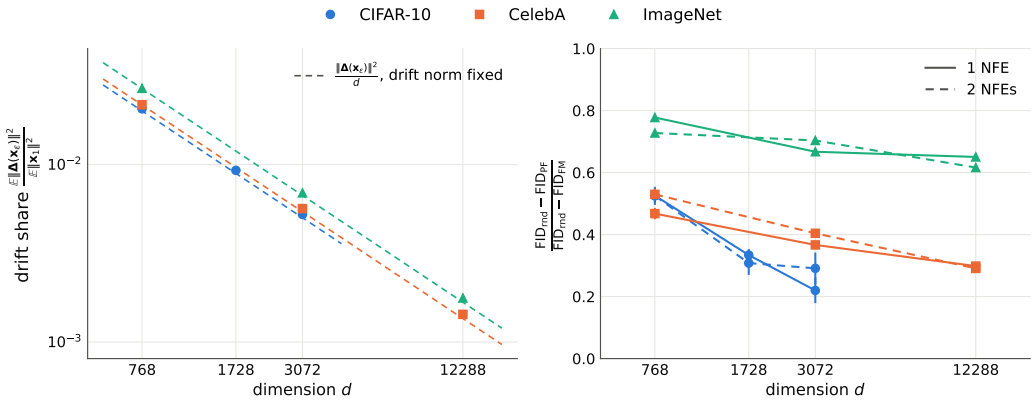


Figure 3: **Closed-form couplings against the dimension. Left:** Share of the endpoint's squared norm carried by the drift, $\mathbb{E}\|\boldsymbol{\Delta}(\mathbf{x}_\epsilon)\|^2/\mathbb{E}\|\mathbf{x}_1\|^2$ with $\boldsymbol{\Delta}(\mathbf{x}_\epsilon) = \mathbf{x}_1 - \mathbf{z}$. The dashed lines show $\frac{\|\boldsymbol{\Delta}(\mathbf{x}_\epsilon)\|^2}{d}$ with the drift norm held fixed at its measured value. Each dataset follows this scaling closely. **Right:** Gap score G (Equation (14)) of the closed-form students at 1 and 2 Euler steps. $G = 0$ means the closed-form student is no better than the random one, and $G = 1$ means it matches the student trained on teacher couplings. π^{PF} improves as the dimension shrinks on all datasets, but does not reach π^{FM}. Note that G is not comparable across datasets. On ImageNet, the random student cannot fit its 1.28M random training pairs, which inflates G (Appendix G).

so that $G = 0$ corresponds to the random student and $G = 1$ to the student trained on teacher couplings. As shown in Figure 3 (right), G increases consistently as d decreases (FIDs are in Table 4). At a single step, it more than doubles on CIFAR-10, rises by roughly 60% on CelebA, and by 20% on ImageNet. The absolute level of G should only be compared within a dataset, since it depends on the random baseline (discussion in Appendix G). The student also uses the drift beyond its training pairs. On coupling endpoints constructed from held-out images, π^{PF} students reconstruct 33%–44% of the image variance on CIFAR-10, 47%–61% on CelebA and 40%–51% on ImageNet, again improving as d decreases (Table 7). However, the gap to the teacher couplings does not close. π^{FM} students reconstruct more than 92% of the held-out variance at every resolution, and their FID approaches that of the teacher at low resolution (Tables 3 and 4). Hence, teacher endpoints retain most of the information about their data point in a form that is easy to extract. Lowering the dimension thus makes closed-form couplings better teachers, but does not make them competitive with pretrained ones.

7 Conclusion

We asked whether the couplings used for rectified flows, usually obtained from a pretrained teacher, can instead be constructed directly from the closed-form probability-flow field of the empirical data distribution. In our experiments, they cannot: students trained on these closed-form pairs perform much closer to those trained on random pairings than to those trained on teacher-generated pairs. The problem is not that the two constructions associate noise with different data points. They largely agree on which regions of noise space correspond to which data points, but choose very different endpoints within those regions. Our analytical endpoint decomposition explains why: the endpoint produced by the closed-form flow is dominated by the Gaussian noise used to initialise the trajectory, with only a small data-dependent drift. The relative contribution of this drift scales as $\frac{\log N}{d}$, so the coupling becomes less informative as the dimension grows unless the dataset size increases exponentially. We confirm empirically that closed-form couplings improve in lower-dimensional spaces. Even there, however, they remain substantially worse teachers than pretrained couplings.

This suggests that the key advantage of a learned teacher is not the coarse assignment of noise-space regions to data points, but how it selects an endpoint within each region. Understanding what determines that selection, and whether the same structure can be recovered without training a teacher, is the main open question left by our results.

DISCLOSURE OF LLM USAGE

We used LLMs to assist with writing and editing the manuscript, to improve clarity, precision, wording, and flow, and to help present and summarise experimental results, in particular in Appendix F and Section 6. We also used them to assist with coding and figure generation and to identify potentially relevant related work. In Section 5 and Appendix C, LLMs were additionally used to provide feedback on formulations and theoretical arguments, including checks for correctness. All cited work, mathematical statements, proofs, experimental results, code, and figures were reviewed and verified by the authors.

REPRODUCIBILITY STATEMENT

All datasets used in this work are publicly available, and their preprocessing is described in detail in Appendix E.1. Our FM teacher and student training procedures build on existing work with publicly available code. We provide complete training details in Appendices E.2 and E.4, including all deviations from the original settings. The construction of each coupling is described in Appendix E.3. Finally, all theoretical results stated in the paper, are provided with full proofs in Appendix C.

REFERENCES

Beatrice Achilli, Enrico Ventura, Gianluigi Silvestri, Bao Pham, Gabriel Raya, Dmitry Krotov, Carlo Lucibello, and Luca Ambrogioni. Losing dimensions: Geometric memorization in generative diffusion. *arXiv preprint arXiv:2410.08727*, 2024.

Michael 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.

Vansh Bansal, Saptarshi Roy, Alessandro Rinaldo, and Purnamrita Sarkar. On the convergence and straightness of rectified flow. In *International Conference on Artificial Intelligence and Statistics*, 2026.

Quentin Bertrand, Anne Gagneux, Mathurin Massias, and Rémi Emonet. On the closed-form of flow matching: Generalization does not arise from target stochasticity. *arXiv preprint arXiv:2506.03719*, 2025.

Giulio Biroli, Tony Bonnaire, Valentin De Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957, 2024.

Nicholas Matthew Boffi, Michael Samuel Albergo, and Eric Vanden-Eijnden. Flow map matching with stochastic interpolants: A mathematical framework for consistency models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Antonin Chambolle and Johannes Hertrich. Limit points of reflow with minibatch optimal transport. *arXiv preprint arXiv:2608.07042*, 2026.

Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Anne Gagneux, Ségolène Martin, Rémi Gribonval, and Mathurin Massias. The generation phases of flow matching: A denoising perspective. *arXiv preprint arXiv:2510.24830*, 2025.

Weiguo Gao and Ming Li. How do flow matching models memorize and generalize in sample data subspaces. *npj Artificial Intelligence*, 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Mason Kamb and Surya Ganguli. An analytic theory of creativity in convolutional diffusion models. In *International Conference on Machine Learning*, 2025.

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

Beomsu Kim, Yu-Guan Hsieh, Michal Klein, Jong Chul Ye, Bahjat Kawar, James Thornton, et al. Simple reflow: Improved techniques for fast flow models. In *International Conference on Learning Representations*, 2025.

Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

Sangyun Lee, Zinan Lin, and Giulia Fanti. Improving the training of rectified flows. *Advances in neural information processing systems*, 37:63082–63109, 2024.

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

Haoming Liu, Jinnuo Liu, Yanhao Li, Liuyang Bai, Yunkai Ji, Yuanhe Guo, Shenji Wan, and Hongyi Wen. From navigation to refinement: Revealing the two-stage nature of flow-based diffusion models through oracle velocity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2649–2658, 2026.

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

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and qiang liu. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *International Conference on Learning Representations*, 2024.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015.

Alireza Mousavi-Hosseini, Stephen Zhang, Michal Klein, et al. Flow matching with semidiscrete couplings. In *International Conference on Learning Representations*, 2026.

Mingue Park, Jisung Hwang, Seungwoo Yoo, Kyeongmin Yeo, and Minhyuk Sung. Pairflow: Closed-form source-target coupling for few-step generation in discrete flow models. In *International Conference on Learning Representations*, 2026.

Emile Pierret, Johannes Hertrich, Samuel Hurault, and Julie Delon. Tessellations of semi-discrete flow matching. *arXiv preprint arXiv:2605.07513*, 2026.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky T Chen. Multisample flow matching: Straightening flows with minibatch couplings. *International Conference on Machine Learning*, 2023.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3), 2015.

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

Christopher Scarvelis, Haitz Sáez de Ocáriz Borde, and Justin Solomon. Closed-form diffusion models. *Transactions on Machine Learning Research*, 2025.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2015.

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

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

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, pp. 1–34, 2024.

Hanshu Yan, Xingchao Liu, Jiachun Pan, Jun Hao Liew, Qiang Liu, and Jiashi Feng. Perflow: Piecewise rectified flow as universal plug-and-play accelerator. *Advances in Neural Information Processing Systems*, 37, 2024.

Mingyang Yi, Jiacheng Sun, and Zhenguo Li. On the generalization of diffusion model. *arXiv preprint arXiv:2305.14712*, 2023.

A VISUALISING $\mathbf{v}^{PF}$

In this section we build intuition for the closed-form field $\mathbf{v}^{PF}$ and the couplings it induces. We start with a two-dimensional toy dataset, where the partition of prior space can be drawn directly and we can visualise the entire distribution. We then show that the same picture carries over to ImageNet-64.

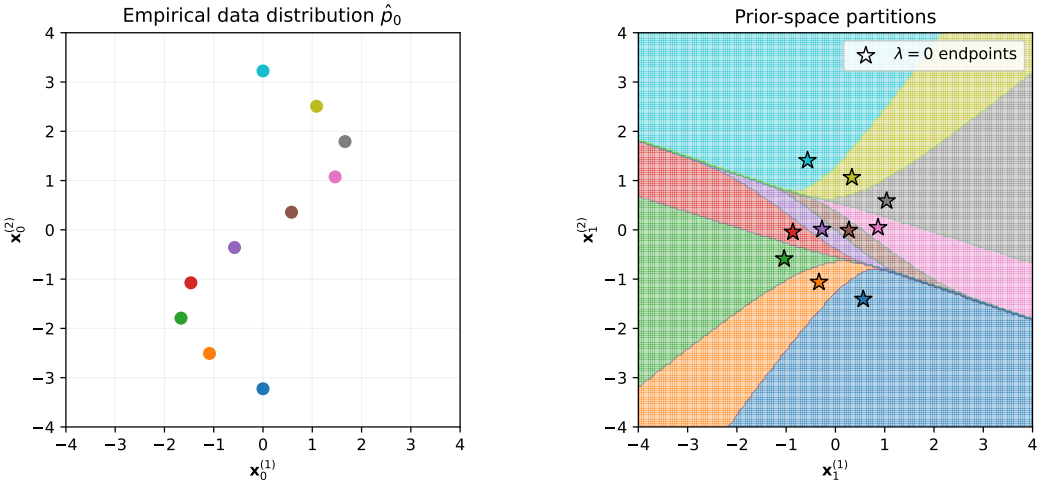


Figure 4: **Many-to-one structure of the closed-form probability-flow ODE. Left:** An empirical two-dimensional dataset $\hat{p}_0$, with each colour denoting one data point. **Right:** The corresponding partition of the Gaussian prior space p_1 induced by the closed-form *PF-ODE*. Each coloured region contains all prior points that are mapped back to the same data point when the flow is integrated in reverse. The matching colours indicate the associated data point and region. Stars mark the endpoints obtained by forward integration from the conditional mean, $\mathbf{x}_\epsilon = \alpha_\epsilon \mathbf{x}_0$ ($\lambda = 0$). Thus, a single data point corresponds to an entire region of prior space rather than a unique endpoint.

Many-to-one contraction. Figure 4 colours every point of the prior by the data point it converges to when $\mathbf{v}^{PF}$ is integrated backwards from $t = 1$ to $t = 0$. Because the empirical dynamics contract as $t \to 0$, each data point attracts not a single point but an entire region of prior space, and together these regions tile the plane. The regions differ considerably in shape and size. Datapoints at the two ends of the arc own large regions that extend far into the tails, while those near its centre are confined to thin bands through the origin. Their probability mass under the prior, however, is identical. The exact flow transports $\mathcal{N}(\mathbf{0}, \mathbf{I})$ onto the uniform empirical distribution, so every region carries mass $\frac{1}{N}$.

The partition also shows why the closed-form coupling is ambiguous. Every point in the region of $\mathbf{x}_0^i$ is a valid endpoint for $\mathbf{x}_0^i$, and which one we obtain is decided entirely by the initial condition. Starting from $\mathbf{x}_\epsilon = \alpha_\epsilon \mathbf{x}_0 + \sigma_\epsilon \mathbf{z}$, the endpoint can land anywhere in the region, depending on $\mathbf{z}$. Removing the noise, i.e. starting from $\mathbf{x}_\epsilon = \alpha_\epsilon \mathbf{x}_0$ ($\lambda = 0$), instead yields a single deterministic endpoint per data point. These endpoints trace out a contracted version of the data arc and preserve the order of the data points along it, but they cluster around the origin and no longer follow the Gaussian prior.

Noisy versus noise-free initialisation. Figure 5 shows that the two initialisations behave analogously on ImageNet-64. We integrate $\mathbf{v}^{PF}$ forward either from the noisy initial condition $\mathbf{x}_\epsilon = \alpha_\epsilon \mathbf{x}_0 + \sigma_\epsilon \mathbf{z}$ used to construct π^{PF}, or from the noise-free initial condition $\mathbf{x}_\epsilon = \alpha_\epsilon \mathbf{x}_0$ used by *PF-ODE* ($\lambda = 0$). Note that the latter is the mean of the conditional path in Equation (1) at time ϵ. Although the two initial conditions are visually indistinguishable, their endpoints differ enormously. The noisy initialisation produces endpoints that look like pure Gaussian noise, with no visible trace of the image. The noise-free initialisation produces a scaled down version of the image that is easily matched to its data point. This is the high-dimensional counterpart of Figure 4. The injected noise decides where within its region the endpoint lands and, in doing so, hides the data point, while the



(a) **Integrating $\mathbf{v}^{PF}$ from $\alpha_\epsilon \mathbf{x}_0 + \sigma_\epsilon \mathbf{z}$.** (b) **Integrating $\mathbf{v}^{PF}$ from $\alpha_\epsilon \mathbf{x}_0$.**

Figure 5: **Noisy versus noise-free initialisation on ImageNet-64** ($\epsilon = 10^{-3}$). The two initial conditions differ only by $\sigma_\epsilon \mathbf{z}$ and are visually indistinguishable. Their endpoints, however, differ completely. Starting from the noisy initialisation **(a)** produces endpoints that resemble Gaussian noise, whereas starting from the conditional mean **(b)** produces faint, low-contrast copies of the data points, from which each pair can be identified by eye.

noise-free endpoint retains the structure of the data point but is far too small in norm to be a sample from the prior.

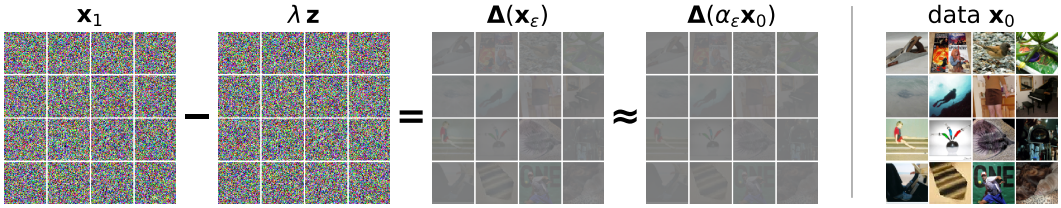


Figure 6: **Visualising the endpoint decomposition on ImageNet-64** ($\lambda = 1$, $\epsilon = 10^{-3}$). The endpoints $\mathbf{x}_1$ resemble pure Gaussian noise, yet subtracting the injected noise $\lambda\mathbf{z}$ reveals the drift $\boldsymbol{\Delta}(\mathbf{x}_\epsilon)$, a faint, structure-preserving copy of each data point that closely matches the *PF-ODE* ($\lambda = 0$) endpoint $\boldsymbol{\Delta}(\alpha_\epsilon \mathbf{x}_0)$ of a noise-free integration. The two drifts differ mainly in magnitude, with $\boldsymbol{\Delta}(\mathbf{x}_\epsilon) \approx 1.2\,\boldsymbol{\Delta}(\alpha_\epsilon \mathbf{x}_0)$, while their directions agree closely, with an average cosine similarity of 0.95 (Table 2).

Recovering the drift. Finally, Figure 6 connects the two initialisations through the endpoint decomposition of Proposition 5.1. Since $\mathbf{x}_1 = \lambda\mathbf{z} + \boldsymbol{\Delta}(\mathbf{x}_\epsilon)$ holds exactly, subtracting the injected noise from the Gaussian-looking endpoints removes precisely the component that hides the image. What remains is the drift $\boldsymbol{\Delta}(\mathbf{x}_\epsilon)$, again a faint copy of the data point, and nearly indistinguishable from the *PF-ODE* ($\lambda = 0$) endpoint $\boldsymbol{\Delta}(\alpha_\epsilon \mathbf{x}_0)$. The information about $\mathbf{x}_0$ is therefore not lost during the noisy integration. It is merely buried under the injected noise, which carries more than 99% of the endpoint energy. Moreover, the drift looks almost the same whether or not noise was injected, which is the approximation stated in Equation (7). This is a high-dimensional effect. In two dimensions, the drift must carry every endpoint back into the region of its data point (Figure 4) and therefore depends more strongly on the noise, and even across our image resolutions, $\cos(\boldsymbol{\Delta}(\mathbf{x}_\epsilon), \mathbf{x}_0)$ drops slightly as the dimension decreases (Table 5). We quantify these observations in Appendix D.1.

B Comparing teacher and closed-form couplings

In this section, we provide the details for the two observations from Section 4 that compare teacher and closed-form couplings. First, the teacher couplings are consistent with the closed-form field, since teacher endpoints lie in the closed-form region of their own data point. Second, within this region, the teacher selects a much narrower set of endpoints than the closed-form flow.

We take 10000 random teacher couplings $(\mathbf{x}_0^i, \mathbf{x}_1^i)$ per dataset and integrate each endpoint $\mathbf{x}_1^i$ backwards through the closed-form field of the full training set, from $t = 1$ to $t = 10^{-3}$. A coupling counts as a match if the nearest training image of the result is $\mathbf{x}_0^i$, which by chance happens with probability $\frac{1}{N}$. The matching accuracy is 99.95% on CIFAR-10, 99.94% on CelebA and 100%

on ImageNet. Almost every teacher endpoint therefore lies in the region that the closed-form flow assigns to its own data point.

Table 1: **Teacher couplings select a narrow set of endpoints within each region.** For the first 2000 training images on each data point, the endpoints from three independent noise draws ($\lambda = 1$, $\epsilon = 10^{-3}$) are integrated with 250 Euler steps using $\mathbf{v}^{FM}$ and $\mathbf{v}^{PF}$. Then, we measure the noise share, i.e. fraction of the endpoint variance explained by the injected noise, the cosine similarity between two endpoints of the same data point and fraction of endpoints whose nearest neighbour among the endpoints of another noise draw belongs to the same data point (chance level 0.05%).

dataset	coupling	noise share	$\cos(\mathbf{x}_1, \mathbf{x}_1')$	NN match
CIFAR-10	teacher	6.1%	0.94	100%
	closed-form	99.6%	0.004	0.05%
CelebA	teacher	8.2%	0.92	100%
	closed-form	99.9%	0.001	0.05%
ImageNet	teacher	4.3%	0.96	100%
	closed-form	99.8%	0.002	0.03%

To compare how the two couplings select endpoints within these regions, we integrate the first 2000 training images of each dataset from three independent noise draws through each teacher, using its best checkpoint and 250 Euler steps, and compare them with closed-form couplings of the same images (Table 1). The two couplings behave in opposite ways. For the teacher, the injected noise explains only 4–8% of the endpoint variance. Two endpoints of the same data point are nearly identical, with a cosine similarity of 0.92–0.96, and the nearest neighbour of every teacher endpoint among the endpoints of another noise draw belongs to the same data point. For the closed-form flow, the injected noise explains more than 99% of the endpoint variance. Two endpoints of the same data point are almost orthogonal, like two independent Gaussian samples, and the nearest neighbour identifies the data point only at chance level.

C PROOFS

In this section, we provide the proofs of our propositions.

Proposition C.1 (Endpoint decomposition)**.** *Let $(\mathbf{x}_t)_{t\in[\epsilon,1]}$ solve $\frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t} = \mathbf{v}_t^{PF}(\mathbf{x}_t)$ with initial condition $\mathbf{x}_\epsilon = \alpha_\epsilon \mathbf{x}_0 + \lambda\,\sigma_\epsilon \mathbf{z}$, for arbitrary $\mathbf{x}_0, \mathbf{z} \in \mathbb{R}^d$, $\lambda \in \mathbb{R}$ and $\epsilon \in (0,1)$. Then, with γ_t and m_t as in Equation* (5)*,*

$$\mathbf{x}_1 = \alpha_1 \mathbf{x}_0 + \lambda\,\sigma_1 \mathbf{z} + \sigma_1 \boldsymbol{\Delta}(\mathbf{x}_\epsilon), \qquad \boldsymbol{\Delta}(\mathbf{x}_\epsilon) := \int_\epsilon^1 \frac{\gamma_t}{\sigma_t}\big(m_t(\mathbf{x}_t) - \mathbf{x}_0\big)\,\mathrm{d}t. \tag{15}$$

For the schedules of this paper ($\alpha_t = 1 - t$, $\sigma_t = t$) this reads $\mathbf{x}_1 = \lambda\mathbf{z} + \boldsymbol{\Delta}(\mathbf{x}_\epsilon)$ with $\boldsymbol{\Delta}(\mathbf{x}_\epsilon) = -\int_\epsilon^1 t^{-2}\big(m_t(\mathbf{x}_t) - \mathbf{x}_0\big)\,\mathrm{d}t$. In particular: if the posterior remains concentrated on the generating datapoint, i.e. $m_t(\mathbf{x}_t) = \mathbf{x}_0$ for all $t \in [\epsilon, 1]$ (e.g. for $N = 1$), then $\boldsymbol{\Delta}(\mathbf{x}_\epsilon) = \mathbf{0}$ and the flow returns the injected noise exactly, $\mathbf{x}_1 = \lambda\mathbf{z}$.

Proof. Since $\sigma_t > 0$ on $[\epsilon, 1]$ and the weights w_t^i are smooth there, $\mathbf{v}^{PF}$ is smooth on $[\epsilon, 1] \times \mathbb{R}^d$ and the trajectory $(\mathbf{x}_t)_{t\in[\epsilon,1]}$ is well defined. Recall the split of the field into its linear and nonlinear parts, Equation (5),

$$\mathbf{v}_t^{PF}(\mathbf{x}) = \frac{\mathrm{d}\log\sigma_t}{\mathrm{d}t}\,\mathbf{x} + \gamma_t\, m_t(\mathbf{x}), \qquad \gamma_t = \frac{\mathrm{d}\alpha_t}{\mathrm{d}t} - \alpha_t \frac{\mathrm{d}\log\sigma_t}{\mathrm{d}t}. \tag{16}$$

Define the conditional trajectory $\mathbf{y}_t := \alpha_t \mathbf{x}_0 + \lambda\sigma_t \mathbf{z}$. Direct substitution gives

$$\frac{\mathrm{d}\mathbf{y}_t}{\mathrm{d}t} = \frac{\mathrm{d}\alpha_t}{\mathrm{d}t}\,\mathbf{x}_0 + \lambda\frac{\mathrm{d}\sigma_t}{\mathrm{d}t}\,\mathbf{z} = \frac{\mathrm{d}\log\sigma_t}{\mathrm{d}t}\,\mathbf{y}_t + \gamma_t\,\mathbf{x}_0, \tag{17}$$

i.e. $\mathbf{y}_t$ solves the same ODE as $\mathbf{x}_t$ with $m_t(\cdot)$ replaced by the constant $\mathbf{x}_0$. We will derive an ODE for the deviation of the true trajectory from this reference and integrate it, this will give us the endpoint decomposition.

Let $\mathbf{e}_t := \mathbf{x}_t - \mathbf{y}_t$ denote the deviation. Subtracting the two equations,

$$\frac{\mathrm{d}\mathbf{e}_t}{\mathrm{d}t} = \frac{\mathrm{d}\log\sigma_t}{\mathrm{d}t}\,\mathbf{e}_t + \gamma_t\big(m_t(\mathbf{x}_t) - \mathbf{x}_0\big), \qquad \mathbf{e}_\epsilon = \mathbf{x}_\epsilon - \mathbf{y}_\epsilon = \mathbf{0}. \tag{18}$$

Using the integrating factor method (with the factor $\exp\big(-\int^t \frac{\mathrm{d}\log\sigma_s}{\mathrm{d}s}\,\mathrm{d}s\big) = \frac{1}{\sigma_t}$), we get

$$\frac{\mathrm{d}\big(\frac{\mathbf{e}_t}{\sigma_t}\big)}{\mathrm{d}t} = \frac{\gamma_t}{\sigma_t}\big(m_t(\mathbf{x}_t) - \mathbf{x}_0\big). \tag{19}$$

Integrating from ϵ to 1 with $\mathbf{e}_\epsilon = \mathbf{0}$ yields $\mathbf{e}_1 = \sigma_1 \int_\epsilon^1 \frac{\gamma_t}{\sigma_t}\big(m_t(\mathbf{x}_t) - \mathbf{x}_0\big)\,\mathrm{d}t = \sigma_1\boldsymbol{\Delta}(\mathbf{x}_\epsilon)$. Hence $\mathbf{x}_1 = \mathbf{y}_1 + \mathbf{e}_1 = \alpha_1\mathbf{x}_0 + \lambda\sigma_1\mathbf{z} + \sigma_1\boldsymbol{\Delta}(\mathbf{x}_\epsilon)$.

If $m_t(\mathbf{x}_t) = \mathbf{x}_0$ for all $t \in [\epsilon, 1]$, the integrand vanishes, so $\boldsymbol{\Delta}(\mathbf{x}_\epsilon) = \mathbf{0}$ and $\mathbf{x}_1 = \lambda\mathbf{z}$. For $N = 1$ we have $m_t(\mathbf{x}) = \mathbf{x}_0$ identically, so this case applies with no further assumption . □

Proposition C.2 (Exact cutoff decomposition)**.** *Let $\mathbf{x}_0 \sim \widehat{p}_0$ and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ independently, and let $(\mathbf{x}_t)_{t\in[\epsilon,1]}$ solve the empirical probability-flow ODE from $\mathbf{x}_\epsilon = \alpha_\epsilon\, x_0 + \sigma_\epsilon\mathbf{z}$. Assume that α_t and σ_t are continuously differentiable, $\sigma_t > 0$ on $[\epsilon, 1]$, $\alpha_1 = 0$, $\sigma_1 = 1$, and that $u_t := \frac{\alpha_t}{\sigma_t} \geq 0$ is nonincreasing, and set $\omega_t := -\frac{\mathrm{d}u_t}{\mathrm{d}t} \geq 0$. Let $\bar{\mathbf{x}} := \frac{1}{N}\sum_{j=1}^N \mathbf{x}_0^j$ and $V := \frac{1}{N}\sum_{j=1}^N \|\mathbf{x}_0^j - \bar{\mathbf{x}}\|^2$ be the mean and total variance of the dataset, and write $\|Y\|_{L^2} := (\mathbb{E}_{\mathbf{x}_0,\mathbf{z}}[\|Y\|^2])^{\frac{1}{2}}$ for a random vector $Y = Y(\mathbf{x}_0, \mathbf{z})$. For a cutoff $\tau \in (\epsilon, 1)$, let $\mathbf{r}_\tau := \frac{1}{\sigma_\tau}\big(\mathbf{x}_\tau - (\alpha_\tau\mathbf{x}_0 + \sigma_\tau\mathbf{z})\big)$ denote the normalised deviation of the trajectory from its conditional path at time τ. Then, for every $\tau \in (\epsilon, 1)$, the endpoint drift from Equation* (15) *satisfies the pathwise identity*

$$\boldsymbol{\Delta}(\mathbf{x}_\epsilon) = u_\tau(\mathbf{x}_0 - \bar{\mathbf{x}}) + \mathbf{r}_\tau - \int_\tau^1 \omega_t\big(m_t(\mathbf{x}_t) - \bar{\mathbf{x}}\big)\,\mathrm{d}t, \tag{20}$$

and consequently

$$\|\boldsymbol{\Delta}(\mathbf{x}_\epsilon)\|_{L^2} \leq \|\mathbf{r}_\tau\|_{L^2} + 2u_\tau\sqrt{V}. \tag{21}$$

Proof. Let $\mathbf{y}_t = \alpha_t\mathbf{x}_0 + \sigma_t\mathbf{z}$ and $\mathbf{e}_t := \mathbf{x}_t - \mathbf{y}_t$ be the reference path and the deviation from the proof of Proposition C.1 with $\lambda = 1$. Using

$$\frac{\gamma_t}{\sigma_t} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\alpha_t}{\sigma_t}\right) = -\omega_t, \tag{22}$$

we can rewrite the ODE for the relative deviation, Equation (19), as

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\mathbf{e}_t}{\sigma_t}\right) = -\omega_t\big(m_t(\mathbf{x}_t) - \mathbf{x}_0\big). \tag{23}$$

At $t = 1$, the reference path equals $\mathbf{y}_1 = \mathbf{z}$, because $\alpha_1 = 0$ and $\sigma_1 = 1$. Hence $\frac{\mathbf{e}_1}{\sigma_1} = \mathbf{x}_1 - \mathbf{z} = \boldsymbol{\Delta}(\mathbf{x}_\epsilon)$. At $t = \tau$, we have $\frac{\mathbf{e}_\tau}{\sigma_\tau} = \mathbf{r}_\tau$ by definition of $\mathbf{r}_\tau$. Integrating the previous identity from τ to 1 therefore gives

$$\boldsymbol{\Delta}(\mathbf{x}_\epsilon) - \mathbf{r}_\tau = \frac{\mathbf{e}_1}{\sigma_1} - \frac{\mathbf{e}_\tau}{\sigma_\tau} = -\int_\tau^1 \omega_t\big(m_t(\mathbf{x}_t) - \mathbf{x}_0\big)\,\mathrm{d}t. \tag{24}$$

To isolate the dependence on $\mathbf{x}_0$, we write $m_t(\mathbf{x}_t) - \mathbf{x}_0 = \big(m_t(\mathbf{x}_t) - \bar{\mathbf{x}}\big) - (\mathbf{x}_0 - \bar{\mathbf{x}})$ inside the integral. The second term does not depend on t, and the multiplier ω_t integrates to

$$\int_\tau^1 \omega_t\,\mathrm{d}t = -\int_\tau^1 \frac{\mathrm{d}u_t}{\mathrm{d}t}\,\mathrm{d}t = u_\tau - u_1 = u_\tau, \tag{25}$$

since $u_1 = \frac{\alpha_1}{\sigma_1} = 0$. Solving for $\boldsymbol{\Delta}(\mathbf{x}_\epsilon)$ yields the exact decomposition

$$\boldsymbol{\Delta}(\mathbf{x}_\epsilon) = u_\tau(\mathbf{x}_0 - \bar{\mathbf{x}}) + \mathbf{r}_\tau - \int_\tau^1 \omega_t\big(m_t(\mathbf{x}_t) - \bar{\mathbf{x}}\big)\,\mathrm{d}t. \tag{26}$$

It remains to bound the three terms in L^2. Since $\mathbf{x}_\epsilon = \alpha_\epsilon \mathbf{x}_0 + \sigma_\epsilon \mathbf{z}$, the initial state has distribution p_ϵ. The probability-flow ODE therefore has marginal distribution p_t at every time, so $\mathbf{x}_t \sim p_t$. Moreover, m_t is the posterior mean under the Gaussian path. A posterior mean cannot have more variance than the variable it estimates, and the law of total variance gives

$$\mathbb{E}_{X\sim p_t}\left[\|m_t(X) - \bar{\mathbf{x}}\|^2\right] \le V. \tag{27}$$

Applying the L^2 triangle inequality, also to the time integral,

$$\begin{aligned}\|\boldsymbol{\Delta}(\mathbf{x}_\epsilon)\|_{L^2} &\le u_\tau \|\mathbf{x}_0 - \bar{\mathbf{x}}\|_{L^2} + \|\mathbf{r}_\tau\|_{L^2} + \int_\tau^1 \omega_t \left(\mathbb{E}_{X\sim p_t}\left[\|m_t(X) - \bar{\mathbf{x}}\|^2\right]\right)^{\frac{1}{2}} \mathrm{d}t \\ &\le u_\tau \sqrt{V} + \|\mathbf{r}_\tau\|_{L^2} + \sqrt{V}\int_\tau^1 \omega_t \,\mathrm{d}t \\ &= \|\mathbf{r}_\tau\|_{L^2} + 2u_\tau\sqrt{V}.\end{aligned} \tag{28}$$

This proves the claim. □

D EMPIRICAL INVESTIGATION OF THE DRIFT

This section contains the empirical measurements to support Section 5. Appendix D.1 quantifies the two empirical properties of the drift on which the endpoint decomposition builds. First, the drift is small, and second, it depends only weakly on the injected noise. Appendix D.2 evaluates both terms of the bound in Proposition 5.2 along the trajectory, and Appendix D.3 tests the two predictions of the scaling argument in Section 5.2. Unless stated otherwise, the empirical distribution is the full training set of each dataset, and we use $\epsilon = 10^{-3}$.

D.1 THE PROPERTIES OF THE DRIFT

Table 2: **Empirical verification of Equation (7).** Every data point is integrated twice with independent noises $\mathbf{z}, \mathbf{z}' \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ($\lambda = 1$) and once from $\alpha_\epsilon \mathbf{x}_0$ (*PF-ODE* ($\lambda = 0$)), for 25000 data points per dataset, with the full training set as the empirical distribution and $\epsilon = 10^{-3}$. **(a)** Share of the endpoint energy carried by the drift (first column), energy of the differences between the drifts $\boldsymbol{\Delta}(\mathbf{x}_\epsilon) = \mathbf{x}_1 - \mathbf{z}$, $\boldsymbol{\Delta}(\mathbf{x}'_\epsilon) = \mathbf{x}'_1 - \mathbf{z}'$ and $\boldsymbol{\Delta}(\alpha_\epsilon \mathbf{x}_0)$ relative to the drift energy $\mathbb{E}\|\boldsymbol{\Delta}(\mathbf{x}_\epsilon)\|^2$ (second and third columns), and a least-squares fit $\boldsymbol{\Delta}(\mathbf{x}_\epsilon) \approx g\,\boldsymbol{\Delta}(\alpha_\epsilon \mathbf{x}_0)$ with its gain g and the energy share of the remainder (last two columns). **(b)** Cosine similarities, averaged over data points.

(a) Energy of the differences

Dataset	$\frac{\mathbb{E}\|\boldsymbol{\Delta}(\mathbf{x}_\epsilon)\|^2}{\mathbb{E}\|\mathbf{x}_1\|^2}$	$\frac{\mathbb{E}\|\boldsymbol{\Delta}(\mathbf{x}_\epsilon)-\boldsymbol{\Delta}(\mathbf{x}'_\epsilon)\|^2}{\mathbb{E}\|\boldsymbol{\Delta}(\mathbf{x}_\epsilon)\|^2}$	$\frac{\mathbb{E}\|\boldsymbol{\Delta}(\mathbf{x}_\epsilon)-\boldsymbol{\Delta}(\alpha_\epsilon\mathbf{x}_0)\|^2}{\mathbb{E}\|\boldsymbol{\Delta}(\mathbf{x}_\epsilon)\|^2}$	fit $\boldsymbol{\Delta}(\mathbf{x}_\epsilon) \approx g\,\boldsymbol{\Delta}(\alpha_\epsilon\mathbf{x}_0)$	
				g	remainder
CIFAR-10	0.52%	25.5%	15.5%	1.19	13.3%
CelebA	0.14%	27.3%	16.8%	1.21	14.2%
ImageNet	0.18%	22.0%	14.0%	1.18	12.0%

(b) Cosine similarities

Dataset	$\cos(\boldsymbol{\Delta}(\mathbf{x}_\epsilon), \boldsymbol{\Delta}(\mathbf{x}'_\epsilon))$	$\cos(\boldsymbol{\Delta}(\mathbf{x}_\epsilon), \boldsymbol{\Delta}(\alpha_\epsilon\mathbf{x}_0))$	$\cos(\boldsymbol{\Delta}(\mathbf{x}_\epsilon), \mathbf{x}_0)$	$\cos(\boldsymbol{\Delta}(\alpha_\epsilon\mathbf{x}_0), \mathbf{x}_0)$
CIFAR-10	0.899	0.945	0.867	0.932
CelebA	0.888	0.939	0.746	0.796
ImageNet	0.911	0.952	0.864	0.923

Table 2 is based on three integrations per data point. Two start from independent noisy initial conditions ($\lambda = 1$) and give the drifts $\boldsymbol{\Delta}(\mathbf{x}_\epsilon) = \mathbf{x}_1 - \mathbf{z}$ and $\boldsymbol{\Delta}(\mathbf{x}'_\epsilon) = \mathbf{x}'_1 - \mathbf{z}'$. The third starts from the noise-free initial condition $\alpha_\epsilon \mathbf{x}_0$ and gives the *PF-ODE* ($\lambda = 0$) endpoint $\boldsymbol{\Delta}(\alpha_\epsilon \mathbf{x}_0)$.

The drift is small. On all three datasets, the drift carries well below 1% of the endpoint energy (Table 2)(a). In absolute terms, its norm is nearly the same across datasets, between 4.0 and 4.7

(Figure 7), although the dimension of CelebA and ImageNet-64 is four times that of CIFAR-10. The smaller share on these two datasets therefore reflects the larger dimension rather than a smaller drift. This is the $\frac{1}{d}$ dependence that Section 5.2 predicts and Section 6 exploits.

The drift depends only weakly on the noise. The least-squares fit in Table 2(a) shows that, on every dataset, the drift is an amplified copy of the noise-free drift, with gain $g \approx 1.2$, and the remainder carries only 12–14% of the drift energy. The cosine similarities in Table 2(b) show that this remainder is the part that changes with the noise. Each noisy drift is closely aligned with the noise-free drift, with $\cos(\boldsymbol{\Delta}(\mathbf{x}_\epsilon), \boldsymbol{\Delta}(\alpha_\epsilon \mathbf{x}_0)) = 0.94$–$0.95$. If every noisy drift consists of the noise-free drift plus an independent noise-specific part, $\boldsymbol{\Delta}(\mathbf{x}_\epsilon)$ and $\boldsymbol{\Delta}(\mathbf{x}'_\epsilon)$ would share only the noise-free drift, and their cosine similarity should be approximately the square of this value, 0.88–0.91. The measured $\cos(\boldsymbol{\Delta}(\mathbf{x}_\epsilon), \boldsymbol{\Delta}(\mathbf{x}'_\epsilon))$ of 0.89–0.91 matches this closely. Therefore, $\boldsymbol{\Delta}(\alpha_\epsilon \mathbf{x}_0)$ captures what the noisy drifts have in common, and only the small remainder varies with the noise. This supports the approximation in Equation (7), up to the gain g.

The drift carries the data point. Finally, both drifts are well aligned with the data point itself (Table 2(b)). The noise-free drift is consistently the better aligned of the two, as expected if the noise-specific remainder carries little information about $\mathbf{x}_0$. The alignment is lowest on CelebA, but even there the drift retains much of the structure of its data point. Figures 2 and 6 show what these numbers look like on ImageNet-64.

D.2 The bound on the drift investigated

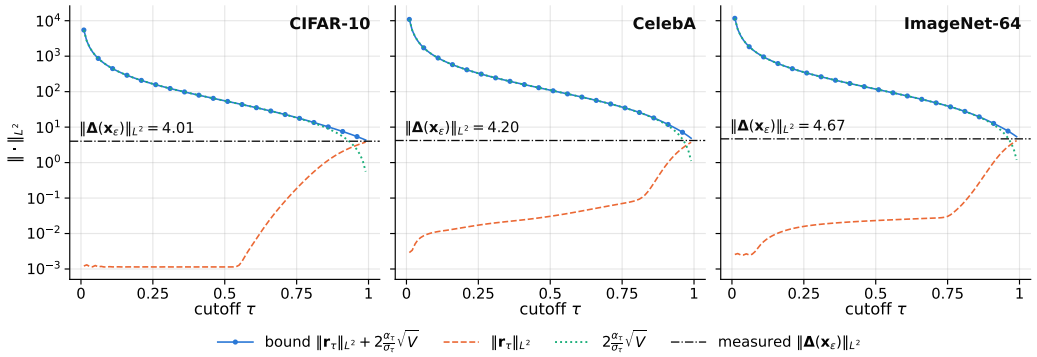


Figure 7: **Evaluation of the drift bound.** We evaluate both terms of Proposition 5.2 on all three datasets for different values of the cutoff τ, using the entire training set. The bound becomes increasingly tight as τ increases, showing that Proposition 5.2 provides a practical bound on the drift of the *PF-ODE* trajectory. Furthermore, the tracking error $\|\mathbf{r}_\tau\|_{L^2}$ remains negligible for small τ, supporting the assumption in Equation (12).

Figure 7 evaluates the two terms of Proposition 5.2 as a function of the cutoff τ, together with the measured drift.

The two terms trade off along the trajectory. For small τ, the trajectory tracks the conditional path of its data point almost perfectly. The tracking error $\|\mathbf{r}_\tau\|_{L^2}$ stays between 10^{-3} and 10^{-1}, and the bound is dominated by the SNR term $2\frac{\alpha_\tau}{\sigma_\tau}\sqrt{V}$, which is still large because the SNR is high. Only late in the trajectory, from $\tau \approx 0.55$ on CIFAR-10, $\tau \approx 0.8$ on CelebA and $\tau \approx 0.75$ on ImageNet-64, does the tracking error rise sharply as the trajectory leaves the conditional path. By then, the SNR term has decayed, and the bound approaches the measured drift. At $\tau = 1$ the bound is exact, since $\alpha_1 = 0$ and $\mathbf{r}_1 = \mathbf{x}_1 - \mathbf{z} = \boldsymbol{\Delta}(\mathbf{x}_\epsilon)$.

This supports the step from Proposition 5.2 to Equation (12) in Section 5.2. Before the trajectory stops tracking its data point, the tracking error is negligible, and the drift is controlled by the SNR term alone. At such τ, the bound exceeds the measured drift by roughly an order of magnitude, so it captures how the drift scales rather than its exact value.

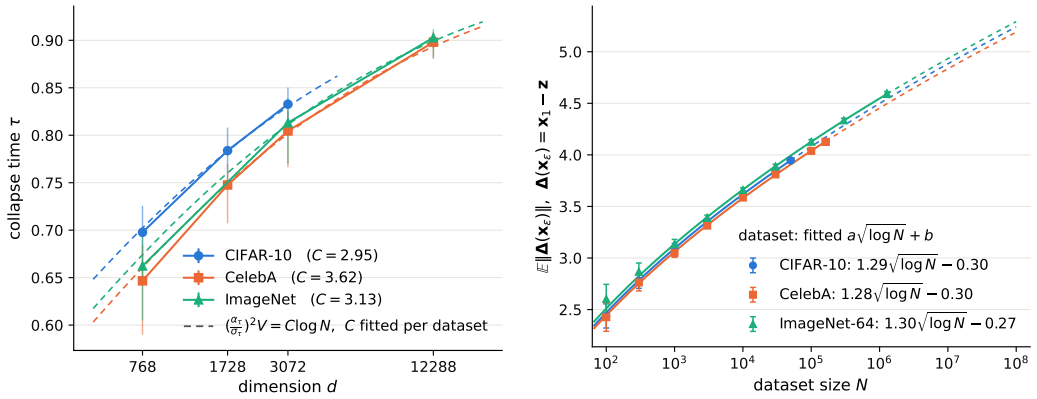


Figure 8: **Empirical validation of the scaling argument. Left:** The posterior collapse time τ as the data dimension is varied by downsampling. Dashed lines show the prediction $(\frac{\alpha_\tau}{\sigma_\tau})^2 V = C \log N$ from Equation (11), with a single constant C fitted per dataset, which closely tracks the measured transition across CIFAR-10, CelebA and ImageNet-64. **Right:** Drift magnitude as a function of the dataset size N. Solid lines show fits of the form $\mathbb{E}\|\mathbf{\Delta}(\mathbf{x}_\epsilon)\| = a\sqrt{\log N} + b$, confirming the $\sqrt{\log N}$ scaling predicted by Equation (12). Dashed lines extrapolate the fits beyond the full training set, and error bars are 95% confidence intervals.

D.3 Scaling of the drift contribution

Figure 8 tests the two predictions of Section 5.2, specifically the collapse criterion in Equation (11) and the resulting $\sqrt{\log N}$ growth of the drift in Equation (12).

Collapse time. To measure the collapse time, we follow the conditional path $\mathbf{x}_t = \alpha_t \mathbf{x}_0^i + \sigma_t \mathbf{z}$ for 10000 data points and evaluate the posterior weights $w_t^j(\mathbf{x}_t)$ of Equation (3) against the full training set. The collapse time τ is the first t at which the mean maximum weight $\max_j w_t^j(\mathbf{x}_t)$ falls below $\frac{1}{2}$. For the prediction, we solve Equation (11) for τ, using the measured data variance V at each resolution and fitting a single constant C per dataset. For our linear schedule, this gives $\tau = \left(1 + \sqrt{C \log N / V}\right)^{-1}$. Lowering the dimension reduces V and with it the distance penalty that keeps the posterior on its data point, so the posterior collapses earlier. The collapse time drops from $\tau \approx 0.9$ at $d = 12288$ to $\tau \approx 0.65$–0.7 at $d = 768$, and the prediction follows this trend on every dataset with C between 2.95 and 3.62 (Figure 8, left).

Drift magnitude. To vary N, we restrict the empirical distribution to random subsets of the training set, from $N = 10^2$ up to the full dataset. Over four orders of magnitude in N, the mean drift norm grows only from about 2.5 to 4.6 and is well described by $a\sqrt{\log N} + b$ with $a \approx 1.3$ on all three datasets (Figure 8, right). The three curves nearly coincide, even though the datasets differ in content and CIFAR-10 has a quarter of the dimension of the other two. This is consistent with Equation (12), in which the drift depends on N but not on d, and with the nearly constant drift norm across resolutions in Figure 3. Extrapolating the fits shows how little more data would help. Even at $N = 10^8$, the drift norm would reach only about 5.3, which at $d = 12288$ is still only about 0.2% of the endpoint energy.

E Experimental Details

This section describes our experimental setup. We first describe the datasets and how we obtain the different resolutions, then the flow-matching teachers, the construction of the three coupling sets, and finally the training and evaluation of the rectified-flow students.

E.1 DATASETS AND RESOLUTIONS

We use the training sets of CIFAR-10 (Krizhevsky et al., 2009), CelebA (Liu et al., 2015) (the standard training partition of the aligned-and-cropped release) and ImageNet-1k (Russakovsky et al., 2015), all unconditional, without augmentation, with pixel values scaled to $[-1, 1]$. The dimension $d = 3\,r^2$ is varied by changing the resolution r. All resolutions of a dataset are generated from the same source images, so neighbouring resolutions differ in nothing but resolution.

Downsampling. For CelebA, we take the aligned 178×218 images, crop the central 140×140 pixels and resize them with bilinear interpolation to 16, 32 and 64 px. ImageNet images are centre-cropped to a square and box-downsampled to 64 px, following Chrabaszcz et al. (2017). The 32 and 16 px versions are 2×2 and 4×4 box averages of these 64 px images. CIFAR-10 is only available at 32 px. We obtain the 16 px version as a 2×2 box average, and the 24 px version, a non-integer factor of $4/3$, with PIL's area filter, which generalises the box average to non-integer factors. Across all resolutions, the per-pixel standard deviation stays between 0.48 and 0.55, so we use the same preconditioning constant $\sigma_{\mathrm{data}} = 0.5$ throughout.

E.2 FLOW-MATCHING TEACHERS

All teachers are trained with the flow-matching objective on the linear path $\alpha_t = 1 - t$, $\sigma_t = t$, with the EDM preconditioning of Karras et al. (2022) around the network ($\sigma_{\mathrm{data}} = 0.5$), the EDM loss weight, and timesteps t drawn from a logit-normal distribution in log-SNR (mean 2.5, standard deviation 2.5). Optimisation uses Adam, an exponential moving average of the weights with rate 0.9999, and a linear learning-rate warm-up over the first 20000 steps. Model selection uses the checkpoint with the lowest FID-50k at 32 Euler steps. Teacher quality (FID-50k, best checkpoint) is reported in Table 3 at 64 and 128 Euler steps. To have a fair comparison between the teacher model and the *PF-ODE*, we train the model without data augmentation.

CIFAR-10 and CelebA. The network is the DDPM++ UNet of Song et al. (2021) following the exact configuration used in Karras et al. (2022) for CIFAR-10. At 16, 24 and 32 px, this network is used unchanged (only the spatial size of its feature maps changes). However, at 64 px for CelebA we use EDM's FFHQ-64 variant with channel multipliers $[1, 2, 2, 2]$ and attention at 16×16. Teachers train for 300000 steps at global batch size 512 with learning rate 10^{-3} (2×10^{-4} with gradient clipping at norm 1 for the 64 px teacher, following EDM's 64-px setting).

ImageNet. The network is the ADM UNet of Dhariwal & Nichol (2021) following Karras et al. (2022) ImageNet-64 configuration. The identical network is used at 16, 32 and 64 px, so the three ImageNet rungs share their capacity exactly. Only the attention resolutions scale with the image. Teachers train with learning rate 10^{-4} at global batch size 1024 in bf16 mixed precision for 400000 steps (600000 at 64 px).

E.3 COUPLINGS

Every coupling set pairs each of the N training images $\mathbf{x}_0^i$ with one noise endpoint $\mathbf{x}_1^i$ and the students are then trained on these fixed pairs only.

Random couplings. To construct $\pi^{\textit{Random}}$ we draw $\mathbf{x}_1^i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and couple them independently with $\mathbf{x}_0^i$.

Closed-form couplings. For π^{PF} we start from $\mathbf{x}_\epsilon^i \sim \mathcal{N}(\alpha_\epsilon \mathbf{x}_0^i, \sigma_\epsilon^2 \mathbf{I})$ with $\epsilon = 10^{-3}$ and $\lambda = 1$ and integrate the closed-form probability-flow field of the empirical distribution of all N training images from $t = \epsilon$ to $t = 1$ with an adaptive Dormand–Prince solver (relative tolerance 10^{-8}, absolute tolerance 10^{-10}). The field in Equation (3) is evaluated exactly against all N training images, without the nearest-neighbour approximations used in Scarvelis et al. (2025); Bertrand et al. (2025). The injected noise $\mathbf{z}$ is then used to measure the drift $\mathbf{\Delta} = \mathbf{x}_1 - \mathbf{z}$. Each field evaluation touches the whole training set, which makes it memory intensive. Therefore, for CelebA and ImageNet the N trajectories are integrated in chunks against the full set of N guiding points held on the GPU.

Teacher couplings. To obtain π^{FM} we integrate the same initial states $\mathbf{x}_\epsilon^i$ forward through the flow-matching teacher's EMA weights with Euler steps on a uniform time grid to $t = 1$. For each dataset we use 128 Euler steps from $\epsilon = 10^{-3}$ and the teacher's best checkpoint.

E.4 Rectified-flow students

For every coupling set one rectified-flow student is trained on the fixed pairs $(\mathbf{x}_0^i, \mathbf{x}_1^i)$ with the flow-matching loss along the straight interpolant. The timesteps t are drawn from the tilted-uniform distribution of Lee et al. (2024) with tilt $a = 4$ (antithetic, symmetric). We use Adam with 5000 warm-up steps and an EMA of rate 0.9999. The student has the same architecture as the teacher of its resolution. On CIFAR-10 and CelebA the students use learning rate 2×10^{-4}, global batch 512 and the EDM loss weight. The students on random and closed-form couplings train for 200000 steps and students on teacher couplings for 100000 steps, after which their FID no longer improves. On ImageNet all students use learning rate 10^{-4}, global batch 512, no loss weighting, bf16 mixed precision and 200000 steps. Students never see augmented data, since the coupling fixes the pairing.

Repeats. To estimate the robustness of our findings, we repeat experiments on CIFAR-10 and CelebA for three seeds. Each repeat $k \in \{1, 2, 3\}$ draws fresh random couplings and fresh injected noise for the closed-form couplings with seed k and trains its students with seed $41 + k$. The teacher couplings are the same file in all three repeats, so their repeats differ only in the student's initialisation and data order.

Evaluation. All FIDs are estimated using 50000 samples against the Inception-v3 pool-3 statistics of the full training set at the same resolution. Students are sampled with the Euler method on a uniform time grid with $k \in \{1, 2, 4, 8, 16, 32\}$ steps from $t = 1$ to $t = 10^{-3}$, using the EMA weights, and every saved checkpoint (every 25000 steps) is evaluated separately at each k. Teachers are sampled the same way with 64 and 128 Euler steps.

The gap score G in Equation (14) compares the three students of a setting at equal k. $G = 0$ means the closed-form student is no better than the random one, and $G = 1$ that it matches the student trained on the teacher's couplings. Since the three students of a setting share the architecture, optimisation settings and data, G is comparable across the rungs of one dataset, however across datasets it is not, because the score depends heavily on the random baseline. For further discussion see Appendix G.

F Additional results on reducing the dimension

In this section we provide the results from Section 6 in more detail. We first confirm that the teachers used to generate π^{FM} are strong, and then report the FIDs of the three students, from which the gap score is computed. We then show that lowering the dimension makes closed-form couplings more informative, both in how closely each endpoint aligns with its data point and in how much of the data point can be recovered from it, whereas teacher couplings are highly informative at every resolution. Finally, we show that students trained on closed-form couplings use this information beyond their training pairs.

The teachers are strong. All teachers reach low FIDs (Table 3), between 1.2 and 3.4 at 128 Euler steps in every setting except ImageNet-64, the hardest one, where they reach 7.2. These numbers are competitive given that we have not used data augmentation or class-conditional generation.

FIDs behind the gap score. Table 4 shows the one- and two-step FIDs from which the gap score in Figure 3 is computed. The ordering is the same in every resolution. In each setting, the π^{FM} student is best, the π^{Random} student worst, and the π^{PF} student in between. Absolute FIDs, however, are not comparable across resolutions, which is why Section 6 compares the students through G. At the lowest resolution, the π^{FM} students come close to their teachers, with one-step FIDs of 2.0–4.6 against 1.2–2.1 for the teachers at 128 steps. On ImageNet, the random students reach FIDs above 400 at one step. They cannot fit their 1.28M random training pairs and simply predict the dataset mean, which inflates G on this (Appendix G).

Table 3: **FID of the flow-matching teachers used to generate π^{FM}.** FID of the best checkpoint at 64 and 128 Euler steps.

dataset	px	d	64 steps	128 steps
CIFAR-10	16	768	2.6	2.1
CIFAR-10	24	1728	3.4	3.1
CIFAR-10	32	3072	3.6	3.1
CelebA	16	768	1.5	1.2
CelebA	32	3072	2.6	2.7
CelebA	64	12288	4.4	3.4
ImageNet	16	768	1.8	1.3
ImageNet	32	3072	3.8	3.2
ImageNet	64	12288	8.1	7.2

Table 4: **FID of the students presented in Figure 3.** One- and two-step FID of the students trained on π^{Random}, π^{PF} and π^{FM}, from which the gap score G (Equation (14)) is computed. For CIFAR-10 and CelebA we report mean $\pm$ standard deviation over three seeds.

		π^{Random}		π^{PF}		π^{FM}	
dataset	px	1 step	2 steps	1 step	2 steps	1 step	2 steps
CIFAR-10	16	61.0 ± 1.9	38.1 ± 0.7	31.3 ± 1.2	20.2 ± 1.1	4.6 ± 0.0	3.9 ± 0.0
CIFAR-10	24	72.7 ± 0.9	49.1 ± 0.6	51.5 ± 1.5	36.5 ± 1.1	9.3 ± 0.2	8.3 ± 0.3
CIFAR-10	32	82.9 ± 0.7	60.7 ± 6.1	68.0 ± 2.3	46.5 ± 2.5	15.2 ± 0.5	12.4 ± 0.3
CelebA	16	38.3 ± 1.0	29.2 ± 0.8	21.3 ± 0.1	14.7 ± 0.1	2.0 ± 0.1	2.0 ± 0.0
CelebA	32	43.1 ± 2.1	29.6 ± 1.1	29.1 ± 1.8	20.2 ± 0.7	4.8 ± 0.0	6.4 ± 0.2
CelebA	64	70.4 ± 0.4	50.7 ± 0.4	53.1 ± 0.2	38.4 ± 0.0	12.3 ± 0.6	8.3 ± 0.3
ImageNet	16	430.1	158.0	97.5	44.2	2.3	1.6
ImageNet	32	417.8	184.1	145.1	59.0	9.1	6.3
ImageNet	64	415.5	175.6	164.6	81.2	29.7	22.3

Lower dimension makes closed-form couplings more informative. Table 5 measures how closely each endpoint aligns with its data point. For random couplings, the cosine similarity is zero, as expected for independent high-dimensional vectors. For closed-form couplings, it is small but grows as the dimension decreases, almost fourfold from 64 to 16 px on CelebA and ImageNet and twofold from 32 to 16 px on CIFAR-10, in line with the $\frac{1}{\sqrt{d}}$ scaling derived in Section 6. The last column shows why. The alignment of the drift with its data point stays roughly constant across resolutions, so the drift carries similar information at every resolution, and only its weight relative to the injected noise changes. Teacher couplings, in contrast, are strongly aligned at every resolution, roughly four to six times more than closed-form couplings at the native resolution. Even at 16 px, closed-form couplings remain less aligned than teacher couplings at the highest resolution, except on CelebA, where the two are about equal.

The information becomes easier to extract. Table 6 asks how much of the data point a linear map can recover from the endpoint. Random endpoints carry no information, and the held-out R^2 is zero. For closed-form couplings, the held-out R^2 rises as the dimension decreases, from 0.45 to 0.52 on CIFAR-10, from 0.45 to 0.60 on CelebA and from 0.39 to 0.56 on ImageNet. Since the drift keeps its alignment with the data point, this gain comes from the noise masking less of it. Teacher couplings reach 0.74–0.85 at every resolution, above the best closed-form value in any setting. At the highest resolutions, the training R^2 noticeably exceeds the held-out one, e.g. 0.52 against 0.39 for closed-form couplings on ImageNet-64. The ridge regression starts to overfit, as the non-zero training R^2 of random couplings in the same setting shows, so the held-out values are the relevant ones.

Students use the drift beyond their training pairs. Finally, Table 7 tests whether the students exploit this information, or merely memorise their training pairs. We build couplings for held-out

Table 5: **Alignment between data points and their coupled endpoints.** Cosine similarity $\cos(\mathbf{x}_0, \mathbf{x}_1)$ for the three couplings, averaged over data points. The last column is the alignment of the closed-form drift with its data point, $\cos(\boldsymbol{\Delta}(\mathbf{x}_\epsilon), \mathbf{x}_0)$.

dataset	px	d	π^{Random}	π^{PF}	π^{FM}	$\cos(\boldsymbol{\Delta}(\mathbf{x}_\epsilon), \mathbf{x}_0)$
CIFAR-10	16	768	0.000	0.120	0.335	0.852
CIFAR-10	24	1728	0.000	0.082	0.284	0.861
CIFAR-10	32	3072	0.000	0.062	0.230	0.867
CelebA	16	768	0.000	0.103	0.242	0.703
CelebA	32	3072	0.000	0.054	0.171	0.730
CelebA	64	12288	0.000	0.028	0.108	0.746
ImageNet	16	768	0.000	0.135	0.360	0.826
ImageNet	32	3072	0.000	0.070	0.271	0.850
ImageNet	64	12288	0.000	0.036	0.204	0.864

Table 6: **Linear read-out of the data point from the coupled endpoint.** Held-out and training R^2 of a ridge regression $\mathbf{x}_0 - \bar{\mathbf{x}} \approx \mathbf{W}^\top \mathbf{x}_1$, i.e. the fraction of the data variance a linear map recovers from the endpoint $\mathbf{x}_1$ (0: predicts the mean image, 1: exact recovery). Fitted on the first 40000 (CIFAR-10) or 100000 (CelebA, ImageNet) pairs, tested on the next 10000.

			π^{Random}		π^{PF}		π^{FM}	
dataset	px	d	train	test	train	test	train	test
CIFAR-10	16	768	0.01	0.00	0.54	0.52	0.86	0.85
CIFAR-10	24	1728	0.03	-0.01	0.52	0.48	0.84	0.82
CIFAR-10	32	3072	0.05	-0.01	0.52	0.45	0.81	0.78
CelebA	16	768	0.00	0.00	0.60	0.60	0.84	0.84
CelebA	32	3072	0.02	-0.01	0.55	0.53	0.79	0.78
CelebA	64	12288	0.09	-0.03	0.56	0.45	0.80	0.74
ImageNet	16	768	0.00	0.00	0.56	0.56	0.85	0.85
ImageNet	32	3072	0.02	0.00	0.51	0.48	0.82	0.81
ImageNet	64	12288	0.09	-0.02	0.52	0.39	0.81	0.75

images in the same way and compare the one-step output of each student with the paired image. Students trained on closed-form couplings reconstruct 33–61% of the held-out variance, and again more at lower dimensions. This is similar to what the linear read-out recovers from the same kind of endpoints, whereas students trained on teacher couplings reconstruct at least 92% at every resolution. Random students behave differently. On CIFAR-10 and CelebA, they reproduce their training pairs with $R^2 \geq 0.85$, but reach a negative R^2 on held-out images, because they output plausible images that are unrelated to the paired data point. On ImageNet, they fit their training pairs only partially at 16 px and hardly at all at 64 px, which we discuss in Appendix G.

G Fitting the training pairs and the random baseline

On ImageNet, the gap score G is considerably higher than on CIFAR-10 and CelebA (Figure 3), although the closed-form couplings are no more informative there (Tables 5 and 6). In this section, we show that the higher score is caused by the random baseline, which cannot fit its 1.28M training pairs.

Every student is trained on a fixed set of N pairs, and a student that learns its coupling well will also fit these pairs, i.e. map each training endpoint $\mathbf{x}_1^i$ close to its paired image $\mathbf{x}_0^i$. Fitting the training pairs is therefore not a sign of overfitting. For the teacher coupling, which is a smooth deterministic map from noise to data, fitting the pairs and generalising coincide, as the π^{FM} students reconstruct 92%–99% of the variance of held-out images from unseen endpoints (Table 7). The situation is different for the random coupling. Since $\mathbf{x}_1$ is independent of $\mathbf{x}_0$, the optimal velocity at $t = 1$ is $\mathbb{E}[\mathbf{x}_1 - \mathbf{x}_0 \mid \mathbf{x}_1] = \mathbf{x}_1 - \bar{\mathbf{x}}$, so the optimal one-step student outputs the mean image $\bar{\mathbf{x}}$. A random

Table 7: **Reconstruction of held-out images.** $R^2 = 1 - \mathbb{E}\|\hat{\mathbf{x}}_0 - \mathbf{x}_0\|^2/V$ of the one-step student output $\hat{\mathbf{x}}_0$ against the paired image, on the student's own training couplings (train) and on couplings built the same way for held-out images (test). V is the variance of the training images, so $R^2 = 0$ is the accuracy of predicting the mean image. Test couplings of π^{PF} integrate the closed-form field of the training images together with the held-out images.

			π^{Random}		π^{PF}		π^{FM}	
dataset	px	d	train	test	train	test	train	test
CIFAR-10	16	768	0.96	−0.54	0.97	0.44	1.00	0.98
CIFAR-10	24	1728	0.95	−0.51	0.96	0.40	1.00	0.96
CIFAR-10	32	3072	0.92	−0.59	0.92	0.33	0.99	0.92
CelebA	16	768	0.95	−0.52	0.97	0.61	1.00	0.99
CelebA	32	3072	0.89	−0.50	0.92	0.52	1.00	0.98
CelebA	64	12288	0.85	−0.47	0.92	0.47	0.99	0.98
ImageNet	16	768	0.62	−0.35	0.82	0.51	1.00	0.99
ImageNet	32	3072	0.17	−0.06	0.78	0.43	0.99	0.98
ImageNet	64	12288	0.02	−0.01	0.65	0.40	0.99	0.97

student can only improve on the mean image by fitting its finite set of pairs, and its quality therefore depends on whether N pairs can be fitted within the training budget.

Probing pair fitting. We feed each student its own training endpoints and compare the one-step output with the paired image, which gives the training R^2 in Table 7. On CIFAR-10 and CelebA, both the random and the closed-form students fit their pairs, with R^2 between 0.85 and 0.97. On ImageNet, with 80 epochs over 1.28M pairs, the random student fits its pairs only partially at 16 px ($R^2 = 0.62$) and hardly at all at 32 and 64 px (0.17 and 0.02), and it has no information in its coupling to fall back on. At 64 px it is effectively a mean predictor, with an FID of 415, close to the FID of the mean image (446). The closed-form student does not fit its pairs perfectly either, but it still recovers 65–82% of the paired image's variance and 40–51% on held-out images, consistent with it decoding the drift.

Control with 50k pairs. To isolate the effect of N, we train the ImageNet-16 students on the first 50k pairs of each coupling, with the same network, batch size and number of steps (Table 8). The random student improves eightfold at one step and reaches the level of the CIFAR-10 random student, and G drops to the CIFAR-10 level. The π^{FM} student becomes worse. Its coupling already contains the full noise-to-data map, so the student's quality is set by how densely the pairs sample this map, and 50k pairs sample it 25 times more sparsely than 1.28M.

In conclusion, the high G on ImageNet reflects a random baseline that cannot fit its pairs, not closed-form couplings that are better than on the other datasets. Since the random students are similarly poor at every ImageNet resolution, with one-step FIDs between 416 and 430 (Table 4), the increase of G as d decreases remains meaningful within ImageNet.

Table 8: **Controlling for the number of pairs.** ImageNet-16 students trained on all pairs (80 epochs) and on the first 50k pairs (2048 epochs), with the same network, batch size and number of steps (best checkpoint).

		π^{Random}		π^{PF}		π^{FM}		G	
dataset	pairs	1 step	2 steps	1 step	2 steps	1 step	2 steps	1 step	2 steps
ImageNet-16	1.28M	430.1	158.0	97.5	44.2	2.3	1.6	0.78	0.73
ImageNet-16	50k	53.8	28.2	35.3	25.1	12.1	11.6	0.44	0.18