On the Closed-Form of Flow Matching: Generalization Does Not Arise from Target Stochasticity

Anonymous Author(s)
Affiliation
Address
email

Abstract

Modern deep generative models can now produce high-quality synthetic samples that are often indistinguishable from real training data. A growing body of research aims to understand why recent methods – such as diffusion and flow matching techniques – generalize so effectively. Among the proposed explanations are the inductive biases of deep learning architectures and the stochastic nature of the conditional flow matching loss. In this work, we rule out the latter – the noisy nature of the loss – as a primary contributor to generalization in flow matching. First, we empirically show that in high-dimensional settings, the stochastic and closed-form versions of the flow matching loss yield nearly equivalent losses. Then, using state-of-the-art flow matching models on standard image datasets, we demonstrate that both variants achieve comparable statistical performance, with the surprising observation that using the closed-form can even improve performance.

1 Introduction

Recent deep generative models, such as diffusion (Sohl-Dickstein et al., 2015; Ho et al., 2020a; Song et al., 2021) and flow matching models (Lipman et al., 2023; Albergo and Vanden-Eijnden, 2023; Liu et al., 2023), have achieved remarkable success in synthesizing realistic data across a wide range of domains. State-of-the-art diffusion and flow matching methods are now capable of producing multi-modal outputs that are virtually indistinguishable from human-generated content, including images (Stability AI, 2023), audio (Borsos et al., 2023), video (Villegas et al., 2022; Brooks et al., 2024), and text (Gong et al., 2023; Xu et al., 2025).

A central question in deep generative modeling concerns the generalization capabilities and underlying mechanisms of these models. Generative models generalization remains a puzzling phenomenon, raising a number of challenging and unresolved questions: whether generative models truly generalize is still the subject of active debate. On one hand, several studies (Carlini et al., 2023; Somepalli et al., 2023b,a; Dar et al., 2023) have shown that large diffusion models are capable of memorizing individual samples from the training set, including licensed photographs, trademarked logos, and sensitive medical data.

On the other hand, Kadkhodaie et al. (2024) have empirically demonstrated that while memorization can occur in low-data regimes, diffusion models trained on a *sufficiently large* dataset exhibit clear signs of generalization. Taken together, recent work points to a sharp phase transition between memorization and generalization (Yoon et al., 2023; Zhang et al., 2024). Multiple theories have been proposed to explain the puzzling generalization of diffusion and flow matching models. On the one hand, Kadkhodaie et al. (2024); Kamb and Ganguli (2024); Ross et al. (2025) suggested a geometric framework to understand the inductive bias of modern deep convolutional networks on images. On the other hand, Vastola (2025) suggested that generalization is due to the *noisy* nature of the training loss. In this work, we clearly answer the following question:

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Does training on noisy/stochastic targets improve flow matching generalization?
If not, what are the main sources of generalization?

Contributions.

- We challenge the prevailing belief that generalization in flow matching stems from an inherently noisy loss (Section 3.1). This assumption, largely supported by studies in low-dimensional settings, fails to hold in realistic high-dimensional data regimes.
- Instead, we observe that generalization in flow matching emerges precisely when the limited-capacity neural network fails to approximate the exact velocity field (Section 3.2).
- We identify two critical time intervals – at early and late times – where this approximation breaks down (Section 3.3). We show that generalization arises predominantly early along flow matching trajectories, aligning with the transition from the stochastic to the deterministic regime of the flow matching objective.
- Finally, on standard image datasets (CIFAR-10 and CelebA), we show that explicitly regressing against the exact velocity field does not impair generalization and can, in some cases, enhance it (Section 4).

The manuscript is organized as follows. Section 2 reviews the fundamentals of conditional flow matching and recalls the closed-form of the "optimal" velocity field. Leveraging the closed-form expression of the flow matching velocity field, Section 3 investigates the key sources of generalization in flow matching. In Section 4, we introduce a learning algorithm based on the closed-form formula. Related work is discussed in detail in Section 5.

2 Recalls on conditional flow matching

Let p_0 be the source distribution and p_{data} the data distribution. We are given n data points $x^{(1)}, \dots, x^{(n)} \sim p_{\text{data}}$, $x^{(i)} \in \mathbb{R}^d$. The goal of flow matching is to find a velocity field $u : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$, such that, if one solves on $[0,1]$ the ordinary differential equation

$$\begin{cases} x(0) = x_0 \in \mathbb{R}^d \\ \dot{x}(t) = u(x(t), t) \end{cases} \tag{1}$$

then the law of $x(1)$ when $x_0 \sim p_0$ is p_{data} (in which case we say that u *transports* p_0 to p_{data}). For every value of t between 0 and 1, the law of $x(t)$ defines a *probability path*, denoted $p(\cdot|t)$ that progressively goes from p_0 to p_{data}. If one knows u, new samples can then be generated by sampling x_0 from p_0, solving the ordinary differential equation, and using $x(1)$ as the generated point.

In conditional flow matching, finding such a u is achieved in the following way.

(i) First, a conditioning variable z independent of t is defined, e.g. $z = x_1 \sim p_{\text{data}}$,

(ii) Then, a conditional probability path $p(\cdot|z,t)$ is chosen, e.g.[1] $p(\cdot|z = x_1, t) = \mathcal{N}(tx_1, (1-t)^2 \cdot \text{Id})$.

Through the continuity equation (Lipman et al., 2024, Sec. 3.5), the choice of $p(\cdot|z,t)$ defines a conditional velocity field $u^{\text{cond}}(x,z,t)$. In the setting of (i) and (ii), we can for example choose $u^{\text{cond}}(x, z = x_1, t) = \frac{x_1 - x}{1-t}$. By marginalization against z, conditional flow matching defines a probability path $p(\cdot|t)$ and a "total" velocity field

$$u^\star(x,t) = \mathbb{E}_{z|x,t}\, u^{\text{cond}}(x,z,t) \ , \tag{2}$$

which turns out to transport p_0 to p_{data} (Lipman et al., 2023, Thm. 1). Hence $u^\star$ could be approximated by a neural network $u_\theta : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ with parameters θ by minimizing

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{\substack{t \sim \mathcal{U}([0,1]) \\ x_t \sim p(\cdot|t)}} \| u_\theta(x_t, t) - u^\star(x_t, t) \|^2 \ , \tag{3}$$

but the usual observation is that $u^\star$ is intractable. As a remedy, Lipman et al. (2023, Thm. 2) show that $\mathcal{L}_{\text{FM}}(\theta)$ is equal up to a constant to the conditional flow matching loss, which for the choices (i)

[1] this choice is only valid when $p_0 = \mathcal{N}(0, \text{Id})$, which for simplicity we will assume in the sequel; more generic choices are possible and we refer the reader to Lipman et al. (2024); Gagneux et al. (2025) for deeper introductions to flow matching

and (ii) made above reads

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{\substack{x_0\sim p_0 \\ x_1\sim p_{\mathrm{data}} \\ t\sim\mathcal{U}([0,1])}} \|u_\theta(x_t,t) - u^{\mathrm{cond}}(x_t, z=x_1, t)\|^2, \text{ where } x_t := (1-t)x_0 + tx_1\,. \tag{4}$$

Note that, since $x_t := (1-t)x_0 + tx_1$, we have that $u^{\mathrm{cond}}(x_t, z=x_1, t)$ is both equal to $\frac{x_1 - x}{1-t}$ and to $x_1 - x_0$. The objective $\mathcal{L}_{\mathrm{CFM}}$ is easy to approximate, since it is easy to sample from p_0 and $\mathcal{U}([0,1])$; sampling from p_{data} is approximated by sampling from $\hat{p}_{\mathrm{data}} := \frac{1}{n}\sum_{i=1}^n \delta_{x^{(i)}}$. Although it seems natural, replacing p_{data} by $\hat{p}_{\mathrm{data}}$ in (4) has a very important consequence: it makes the minimizer $\hat{u}^\star$ of $\mathcal{L}_{\mathrm{FM}}$ available in closed-form, which we recall below.

Proposition 1 (Velocity Closed-form Formula)**.** *When p_{data} is replaced by $\hat{p}_{\mathrm{data}}$, with the previous choices, the velocity field in (2) has a closed-form formula:*

$$\hat{u}^\star(x,t) = \sum_{i=1}^n u^{\mathrm{cond}}(x, z=x^{(i)}, t) \cdot \frac{p(x|z=x^{(i)},t)}{\sum_{i'=1}^n p(x|z=x^{(i')},t)}\,. \tag{5}$$

The notation $\hat{u}^\star$ emphasizes the velocity field is optimal for the empirical probability distribution $\hat{p}_{\mathrm{data}}$, not the true one p_{data}. Since $u^{\mathrm{cond}}(x, z=x^{(i)}, t) \propto x^{(i)} - x$, the optimal velocity field $\hat{u}^\star$ is a weighted average of the n different directions $x^{(i)} - x$. More specifically, it writes as follows:

$$\hat{u}^\star(x,t) = \sum_{i=1}^n \lambda_i(x,t)\frac{x^{(i)} - x}{1-t}\,, \tag{6}$$

with $\lambda(x,t) = \mathrm{softmax}((-\frac{1}{2(1-t)^2}\|x - tx^{(i')}\|^2)_{i'=1,\dots,n}) \in \mathbb{R}^n$. This formula also holds for a Gaussian source distribution and $z \sim p_0 \times \hat{p}_{\mathrm{data}}$ (Appendix A.1).

This result can be found in various previous works, *e.g.,* Kamb and Ganguli (2024, Eq. 3), Biroli et al. (2024), Gao and Li (2024), Li et al. (2024) or Scarvelis et al. (2025). Note that the closed-form formula (5) is also valid for other choices of continuous distribution p_0 (*e.g.,* the uniform distribution).

From Equation (6), as $t \to 1$, the velocity field $\hat{u}^\star$ diverges at any point x that does not coincide with one of the training samples $x^{(i)}$, and it points in the direction of the nearest $x^{(i)}$. This creates a paradox: solving the ordinary differential equation (1) with the velocity field $\hat{u}^\star$ can only produce training samples $x^{(i)}$ (see Gao and Li 2024, Thm. 4.6 for a formal proof). Therefore, in practice, exactly minimizing the conditional flow matching loss would result in $u_\theta = \hat{u}^\star$, meaning the model memorizes the training data and fails to generalize. This naturally yields the following question:

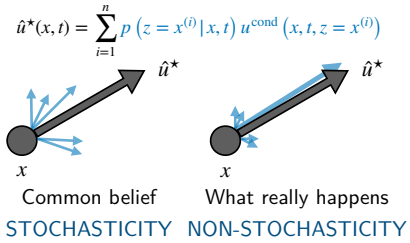
How can flow matching generalize if the optimal velocity field only generates training samples?

3 Investigating the key sources of generalization

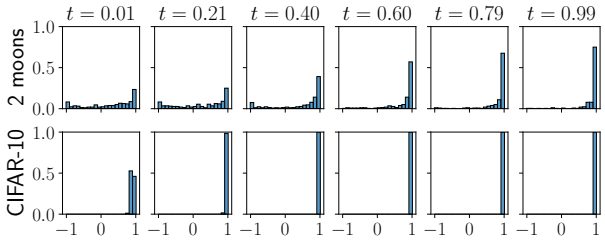
In this section, we investigate the key sources of flow matching generalization using the closed-form formula of its velocity field. First in Section 3.1 we challenge the claim that generalization stems from the stochastic approximation u^{cond} of the optimal velocity field $\hat{u}^\star$. Then, in Section 3.2 we show that generalization arises when u_θ fails to approximate the perfect velocity $\hat{u}^\star$. Interestingly, the target velocity estimation particularly fails at two critical time intervals. Section 3.3 shows that one of these critical times is particularly important for generalization.

3.1 Target stochasticity is not what you need

One recent hypothesis is that generalization arises from the fact that the regression target u^{cond} of conditional flow matching is only a stochastic estimate of $\hat{u}^\star$. The fact that the target regression objective only equals the true objective on average is referred to by Vastola (2025) as "generalization through variance". To challenge this assumption, we leverage Proposition 1, which states that the optimal velocity field $\hat{u}^\star(x,t)$ is a weighted sum of the n values of $u^{\mathrm{cond}}(x,t,z=x^{(i)}) = \frac{x^{(i)}-x}{1-t}$, for $i \in [n]$, and show that, after a *small time value* t, this average is in practice equal to a single value in the expectation (see Figures 1a and 1b).



(a) **Stochasticity vs. non-stochasticity**



(b) **Non stochasticity of $\hat{u}^\star$ for high-dimensional real data**.

Figure 1: We challenge the hypothesis that target stochasticity plays a major role in flow matching generalization. The histograms of the cosine similarities between $\hat{u}^\star((1-t)x_0+tx_1, t)$ and $u^{\text{cond}}((1-t)x_0 + tx_1, z = x_1, t) = x_1 - x_0$ are displayed for various time values t and two datasets. *For real, high-dimensional data, non-stochasticity arises very early* (before $t = 0.2$ for CIFAR-10 with dimension $(3, 32, 32)$).

Comments on Figure 1b. To produce Figure 1b, we sample 256 pairs (x_0, x_1) from $p_0 \times \hat{p}_{\text{data}}$. For each value of t, we compute the cosine similarity between the optimal velocity field $\hat{u}^\star((1-t)x_0 + tx_1, t)$ and the conditional target $u^{\text{cond}}((1-t)x_0 + tx_1, z = x_1, t) = x_1 - x_0$. The resulting similarities are aggregated and shown as histograms. The top row displays the results for the two-moons toy dataset ($d = 2$), and the bottom row displays the results for the CIFAR-10 dataset (Krizhevsky and Hinton 2009, $d = 3072$); $n = 50$k for both. As t increases, the histograms become increasingly concentrated around 1, indicating that $\hat{u}^\star$ aligns closely with a single conditional vector u^{cond}. From Equation (6), this corresponds to a collapse towards 0 of all but one of the softmax weights $\lambda_i(x_t, t)$. This time corresponds to the collapse time studied by Biroli et al. (2024) for diffusion; we discuss the connection in the related works (Section 5). On the two-moons toy dataset, this transition occurs for intermediate-to-large values of t, echoing the observations made in low-dimensional settings by Vastola (2025, Figure 1). In contrast, for high-dimensional real datasets, $\hat{u}^\star(x, t)$ aligns with a single conditional velocity field $x^{(i)} - x$, even at early time steps, suggesting that the non-stochastic regime dominates most of the generative process. This key difference between low- and high-dimensional data suggests that the transition time between the stochastic and non-stochastic regimes is strongly influenced by the dimensionality of the data.

Dimension dependence. To further illustrate the strong impact of dimensionality, Figure 2 reports the proportion of samples x_t (from a batch of 256) for which the cosine similarity between $\hat{u}^\star$ and $u^{\text{cond}} \propto x^{(i)} - x$ exceeds 0.9, as a function of time t. This analysis is performed across multiple spatial resolutions of the Imagenette dataset, obtaining $d \times d$ images by spatial subsampling. Figure 2 reveals a sharp transition: as the dimensionality increases, the proportion of high-cosine matches rapidly converges to 100%.

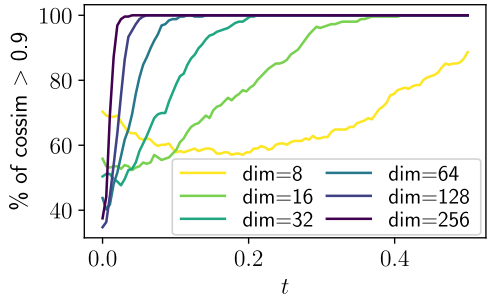


Figure 2: Alignment of $\hat{u}^\star$ and u^{cond} over time for varying image dimensions d on Imagenette.

A practical implication of this behavior is that, for sufficiently large t, if $x_0 \sim p_0$ and $x^{(i)} \sim \hat{p}_{\text{data}}$, then $\hat{u}^\star((1-t)x_0 + tx^{(i)}, t)$ is approximately proportional to $x^{(i)} - x$. Consequently, regressing on $x^{(i)}$ or on the conditional velocity $x_1 - x_0$ becomes effectively equivalent. Section 4 investigates how to learn regressing against optimal velocity field $\hat{u}^\star$, and empirically shows similar results between stochastic and non-stochastic targets.

The regime where flow matching matches stochasticity is mostly concentrated on a very short time interval, for small values of t. We hypothesize that the phenomenon observed here on the optimal velocity field $\hat{u}^\star$ has major implications on the *learned* flow matching model u_θ, which we further inspect in the next section.

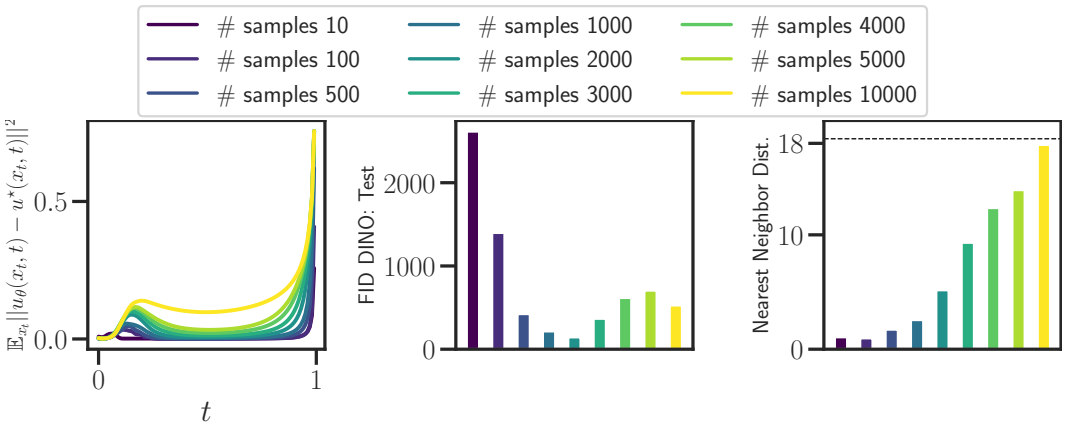


Figure 3: **Failure to learn the optimal velocity field, CIFAR-10**. *Left*: The leftmost figure represents the average error between the optimal empirical velocity field $\hat{u}^\star$ and the learned velocity u_θ for multiple values of time t. *Middle*: The middle figure displays the FID-10k computed on the test dataset, using the DINOv2 embedding. *Right*: The rightmost figure displays the average distance between the generated samples and their closest image from the training set – for reference, the horizontal dashed line indicates the mean distance between an image of CIFAR-10 train and its nearest neighbor in the dataset. All the quantities are computed/learned on a varying number of training samples (10 to 10^4) of the CIFAR-10 dataset.

3.2 Failure to learn the optimal velocity field

This subsection investigates how well the learned velocity field u_θ approximates the optimal/ideal velocity field $\hat{u}^\star$, and how the quality of this approximation correlates with generalization. To do so, we propose the following experiment.

Set up of Figure 3. To build Figure 3, we subsampled the CIFAR-10 dataset from 10 to 10^4 samples. For each size, we trained a flow matching model using a standard 34 million-parameter U-Net (see Appendix C for details). Following Kadkhodaie et al. (2024), the number of parameters of the network u_θ remains fixed across dataset sizes. Importantly, the optimal velocity field $\hat{u}^\star$ itself depends on the dataset size: as the number of samples increases, the complexity of $\hat{u}^\star$ also grows. Thus, we expect the network u_θ to accurately approximate the optimal velocity field $\hat{u}^\star$ for smaller dataset sizes.

Comments on Figure 3. The leftmost plot shows the average training error

$$\mathop{\mathbb{E}}_{\substack{x_0\sim p_0\\ x_1\sim \hat{p}_{\text{data}}}} \|u_\theta(x_t,t) - \hat{u}^\star(x_t,t)\|^2, \quad \text{where} \quad x_t := (1-t)x_0 + t x_1 \ ,$$

between the learned velocity u_θ and the optimal empirical velocity field $\hat{u}^\star$, evaluated across multiple time values t and dataset sizes. With only 10 samples (darkest curve), the network u_θ closely approximates $\hat{u}^\star$. As the dataset size increases, the complexity of $\hat{u}^\star$ grows, and the approximation by u_θ becomes less accurate. In particular, the approximation fails at two specific time intervals: around $t \approx 0.15$ and near $t = 1$. The failure near $t = 1$ is expected, as $\hat{u}^\star$ becomes non-Lipschitz at $t = 1$. Interestingly, the early-time failure at $t \approx 0.15$ corresponds to the regime where $\hat{u}^\star$ and u^{cond} start to correlate (see Figure 1b in Section 3.1). The middle plot of Figure 3 reports the FID-10k, computed on the test set in the DINOv2 embedding space (Oquab et al., 2024), for various dataset sizes. For a small dataset (*e.g.,* #samples $= 10$), u_θ approximates $\hat{u}^\star$ well but does not generalize – the test FID exceeds 10^3. As the dataset size increases ($500 \leq$ #samples ≤ 3000), the approximation u_θ becomes less accurate. Despite this, the model achieves lower FID scores on the test set but still memorizes the training data. The rightmost plot of Figure 3 illustrates this memorization by showing the average distance between each generated sample and its nearest neighbor in the training set. For larger datasets (#samples ≥ 3000), this distance increases substantially, indicating that the model generalizes better. Overall, Figure 3 also suggests that the FID metric can be misleading, even when computed on the test set. For example, the model trained with 1000 samples has a low test FID but memorizes training examples.

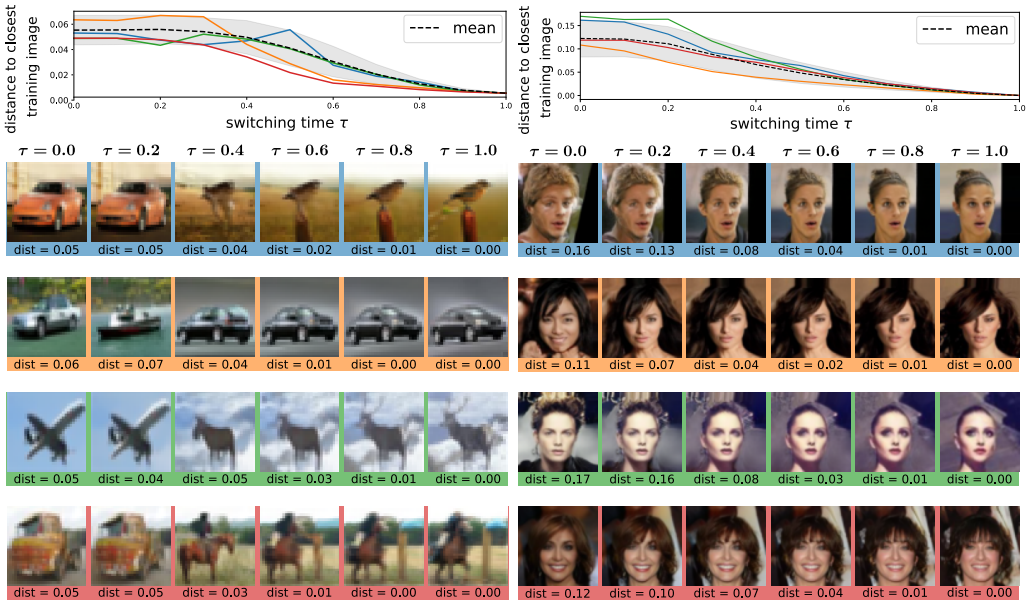


Figure 4: **Generalization occurs at small times on CIFAR-10 (left) and CelebA**64 **(right)**. *Top*: Generalization (distance between generated samples and training data) of hybrid models that follow $\hat{u}^\star$ on $[0, \tau]$, then u_θ on $[\tau, 1]$. The four colored curves correspond to four specific x_0, the black dashed curve is the mean distance over the 256 generated images. *Bottom*: visualization of generated images for the four different starting noises and various values of τ (the background color matching the curve in the top figure). Following $\hat{u}^\star$ until $\tau \geq 0.3$ yields a model that is not able to generalize.

Figure 3 confirms that generalization arises when the network u_θ fails to estimate the optimal velocity field $\hat{u}^\star$, and that this failure occurs at two specific time intervals. In Section 3.3, we investigate which of these two intervals is responsible for driving generalization.

3.3 When does generalization arise?

To investigate whether the failure to approximate $\hat{u}^\star$ matters the most at small or large values of t, we carry out the following experiment.

Set up of Figure 4. We first learn a velocity field u_θ using standard conditional flow matching (see Appendix C), then we construct a hybrid model: we define a piecewise trajectory where the flow is governed by the optimal velocity field $\hat{u}^\star$ for times $t \in [0, \tau]$, and by the learned velocity field u_θ for times $t \in [\tau, 1]$, for a given threshold parameter $\tau \in [0, 1]$. For the extreme case $\tau = 1$, the full trajectory follows $\hat{u}^\star$, and samples exactly match training data points. Conversely, when $\tau = 0$, the entire trajectory is governed by u_θ, yielding novel samples. Intermediate values of τ produce a mixture of both behaviors, which we interpret as reflecting varying degrees of generalization. To assess generalization, we measure the distance of generated samples to the dataset using the LPIPS metric (Zhang et al., 2018), which computes the feature distance between two images via some pretrained classification network. We define the distance of a generated sample x to a dataset $\mathcal{D} = \{x^{(1)}, \dots, x^{(n)}\}$ as $\text{dist}(x, \mathcal{D}) = \min_{x^{(i)} \in \mathcal{D}} \text{LPIPS}(x, x^{(i)})$.

We fix a random batch of 256 pure noise images from p_0. Then, for various threshold values τ, we generate 256 images with the hybrid model, always starting from this batch. Finally, we measure the creativity of the hybrid model as the mean of the aforementioned LPIPS distances between the 256 generated samples and the dataset. Figure 4 (top row) displays these distances as τ varies, on the CIFAR-10 (left) and CelebA - 64×64 (right) datasets. For $\tau \leq 0.2$, the hybrid model remains as creative as u_θ, despite following $\hat{u}^\star$ in the first steps. For $\tau > 0.2$, the LPIPS distance starts dropping. On the displayed generated samples (bottom rows), we in fact see that as soon as $\tau \geq 0.4$, the sample generated by the hybrid model is almost the same as the one obtained with $\hat{u}^\star$ ($\tau = 1$). This means that the final image is already determined at $t = 0.4$ and despite the generalization capacity of u_θ,

following it only after $t \geq 0.4$ is not to enough to create a new image. Hence, *generalization occurs early and cannot be explained by the failure to correctly approximate $u^\star$ at large t.*

Although we have shown that the stochastic phase was limited to small values of t in real-data settings, we have not yet definitively ruled it out as the cause of generalization. In the following section, we introduce a learning procedure designed to address this question directly.

4 Learning with the closed-form formula

In this section, to discard the impact of stochastic target on the generalization, we propose to directly regress against the closed-form formula (5).

4.1 Empirical flow matching

Regressing against the closed-form $\hat{u}^\star$ at a point (x_t, t) requires evaluating it, *i.e.,* computing a weighted sum of the conditional velocity fields over all the n training points $x^{(i)}$. For a dataset of n samples of size d, and a batch of size $|\mathcal{B}|$, computing the weights of the exact closed-form formula $\hat{u}^\star(x,t)$ of flow matching requires $\mathcal{O}(n \times |\mathcal{B}| \times d)$. These computations are prohibitive since they must be performed for each batch. One natural idea is to rely on a Monte Carlo approximation of the closed-form formula, using M samples $b^{(1)}, \dots, b^{(M)}$ to estimate the empirical mean:

$$\mathcal{L}_{\mathrm{EFM}}(\theta) = \mathbb{E}_{\substack{t\sim\mathcal{U}([0,1]) \\ x_0\sim p_0 \\ x_1\sim\hat{p}_{\mathrm{data}} \\ x_t=(1-t)x_0+tx_1 \\ b^{(1)}:=x_1\ ;\ b^{(2)},\dots,b^{(M)}\sim\hat{p}_{\mathrm{data}}}} \left\| u_\theta(x_t,t) - \hat{u}^\star_M(x_t,t) \right\|^2 , \tag{7}$$

with

$$\hat{u}^\star_M(x_t,t) = \sum_{j=1}^{M} \frac{b^{(j)} - x_t}{1-t} \cdot \left[\mathrm{softmax}\left(\left(-\frac{\|x_t - t b^{(k)}\|^2}{2(1-t)^2} \right)_{k=1,\dots,M} \right) \right]_j . \tag{8}$$

The formulation in Equation (7) may appear naive at first glance, but it hinges on a crucial trick: the Monte Carlo estimate is computed using a batch that systematically includes the point x_1, that generated the current x_t. If instead $b^{(1)}$ were sampled independently from $\hat{p}_{\mathrm{data}}$, this could introduce a sampling bias (see Ryzhakov et al. 2024, Appendix B, and the corresponding OpenReview comments[2] for an in-depth discussion). We show in Section 4.2 that Algorithm 2, aiming to solve Equation (7), brings steady improvements for high-dimensional datasets such as CIFAR-10 and CelebA. Supplementary material (Appendix B) provides further details on the unbiasedness of $\mathcal{L}_{\mathrm{EFM}}$. From a computational standpoint, despite the additional sampling of M samples, Algorithm 2 is much cheaper than multiplying the batch size by M: the M samples are simply averaged (with weights) but the backpropagation remains exactly the same as in Algorithm 1.

Algorithm 1 Vanilla Flow Matching

for k *in* $1, \dots, n_{\mathrm{iter}}$ **do**
- $t \sim \mathcal{U}([0,1])$
- $x_0 \sim \mathcal{N}(0, \mathrm{Id})$, $x_1 \sim \hat{p}_{\mathrm{data}}$,
- $x_t = (1-t)x_0 + t x_1$
- $u^{\mathrm{cond}}(x_t, t) = x_1 - x_0$
- $\mathcal{L}(\theta) = \left\| u_\theta(x_t,t) - u^{\mathrm{cond}}(x_t,t) \right\|^2$
- Compute $\nabla\mathcal{L}(\theta)$ and update θ

return u_θ

Algorithm 2 Empirical Flow Matching

param : M // Number of samples in the empirical mean

for k *in* $1, \dots, n_{\mathrm{iter}}$ **do**
- $x_0 \sim \mathcal{N}(0, \mathrm{Id})$, $x_1 \sim \hat{p}_{\mathrm{data}}$, $t \sim \mathcal{U}([0,1])$
- $x_t = (1-t)x_0 + t x_1$
- $b^{(1)} = x_1$
- $\forall j \in [\![2, M]\!]$, $b^{(j)} \sim \hat{p}_{\mathrm{data}}$ `// Samples from` $\hat{p}_{\mathrm{data}}$
- $\hat{u}^\star_M(x_t,t) = \sum_{j=1}^{M} \frac{b^{(j)} - x_t}{1-t} \cdot \left[\mathrm{softmax}\left(-\frac{\|x_t - t\cdot b\|^2}{2(1-t)^2} \right) \right]_j$
- $\mathcal{L}(\theta) = \left\| u_\theta(x_t,t) - \hat{u}^\star_M(x_t,t) \right\|^2$
- Compute $\nabla\mathcal{L}(\theta)$ and update θ

return u_θ

[2] https://openreview.net/forum?id=XYDMAckWMa

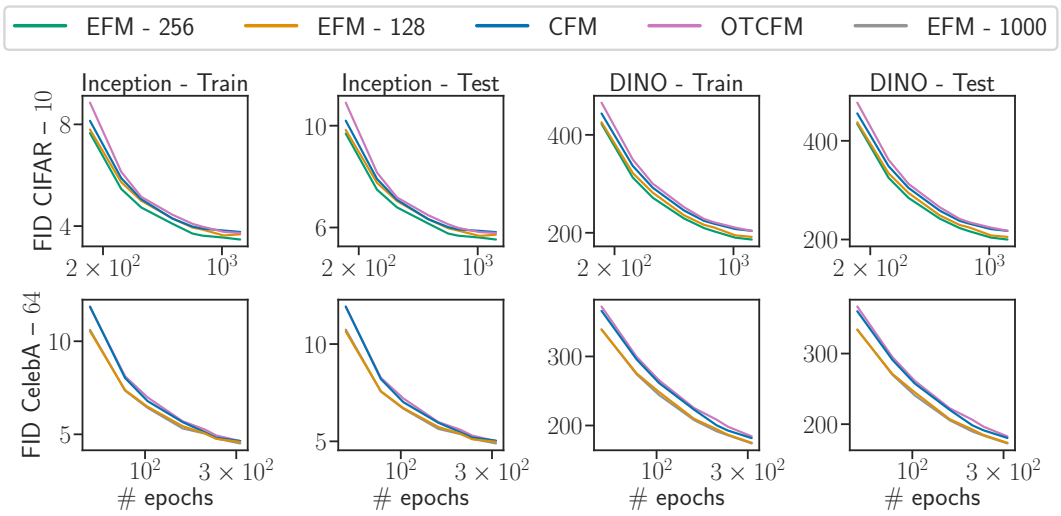


Figure 5: FID computed on the training set (50k) and the test set (10k) using multiple embeddings, Inception and DINOv2. Regressing against a more deterministic target (EFM - 128, 256, 1000) does not yield performance decreases. On the contrary, the more deterministic the target, the better the performance.

4.2 Experiments

We now learn with empirical flow matching (EFM, Equation (7) and Algorithm 2) in practical high-dimensional settings. Our goal with this empirical investigation is first to observe if regressing against a more deterministic target leads to performance improvement/degradation.

Datasets and Models. We perform experiments on the image datasets CIFAR-10 (Krizhevsky and Hinton, 2009) and CelebA 64×64 (Liu et al., 2015). For the experiments, we compare vanilla conditional flow matching (Lipman et al., 2023; Liu et al., 2023; Albergo and Vanden-Eijnden, 2023), optimal transport flow matching (Pooladian et al., 2023; Tong et al., 2024), and the empirical flow matching in Algorithm 2, for multiple numbers of samples M to estimate the empirical mean. Training details are in Appendix C.

Metrics. To assess generalization performance, we use the standard Fréchet Inception Distance (Heusel et al., 2017) with Inception-V3 (Szegedy et al., 2016) but we also follow the recommendation of Stein et al. (2023) using the DINOv2 embedding (Oquab et al., 2023), which is known to lead to a less biased evaluation. Defining and quantifying the generalization ability of generative models is overall a challenging task: train and test FID are known to be imperfect (Stein et al., 2023; Jiralerspong et al., 2023; Parmar et al., 2022) yet no superior competitor has emerged.

Comments on Figure 5. Figure 5 compares vanilla flow matching, OTCFM, and the empirical flow matching (EFM, Algorithm 2) approaches using various numbers of samples to estimate the empirical mean, $M \in \{128, 256, 1000\}$. First, we observe that learning with a more deterministic target does not degrade either training or testing performance, across both types of embeddings. On the contrary, we consistently observe modest but steady improvements as stochasticity is reduced. For both CIFAR-10 and CelebA, increasing the number of samples M used to compute the empirical mean—*i.e.,* , making the targets less stochastic—leads to more stable improvements. It is worth noting that Algorithm 2 has a computational complexity of $\mathcal{O}(M \times |\mathcal{B}| \times d)$, where $|\mathcal{B}|$ is the batch size, M is the number of samples used to estimate the empirical mean, and d is the sample dimension. In our experiments, choosing $M = |\mathcal{B}| = 128$ yielded a modest time overhead.

5 Related work

The existing literature related to our study can be roughly divided into three approaches: leveraging the closed-form, studies on the memorization vs generalization, and characterization of the different phases of the generating dynamics.

Leveraging the closed-form. Proposition 1 has been leveraged in several ways. The closest existing work is by Ryzhakov et al. (2024), who propose to regress against $\hat{u}^\star$ as we do in Section 4. Nevertheless, their motivation is that reducing the variance of the velocity field estimation makes learning more accurate: as explained in Section 3.1, we argue this claim rests on misleading 2D-based intuitions (*e.g.,* Figure 1, challenged by Section 3.1). Scarvelis et al. (2025) bypass training, and suggest using a smoothed version of $\hat{u}^\star$ to generate novel samples. In a work specific to images and convolutional neural networks, Kamb and Ganguli (2024) suggested that flow matching indeed ends up learning an optimal velocity, but that instead of memorizing training samples, the velocity memorizes a combination of all possible patches in an image and across the images. They show remarkable agreement between their theory and the trajectories followed by learned vector fields, but their work is limited to convolutional architectures.

Memorization and reasons for generalization. Kadkhodaie et al. (2024) directly relates the transition from memorization to generalization to the size of the training dataset, and proposes a geometric interpretation. We provide a complementary experiment in Section 3.2, quantifying how much the network fails to estimate the optimal velocity field. Gu et al. (2025) provide a detailed experimental investigation into the potential causes for generalization, primarily based on the characteristics of the dataset and choices for training and model. Vastola (2025) explores different factors of generalization in the case of diffusion, with a special focus on the stochasticity of the target objective in the learning problem. Through a physic-based modeling of the generative dynamics, they study the covariance matrix of the noisy estimation of the exact score. In our work, we believe that we have shown that this claim was not valid for real high-dimensional data. Niedoba et al. (2024) study the poor approximation of the exact score by the learned models: like Kamb and Ganguli (2024), they suggest that the generalization of the learned models comes from memorization of many patches in the training data.

Temporal regimes. Biroli et al. (2024) provide an analysis of the exact score, the counterpart of the exact velocity field for diffusion. For a multimodal target distribution, the authors identify three phases (we keep the convention that $t = 0$ is noise and $t = 1$ is target): for $t < t_1$, all trajectories are indistinguishable; for $t_1 < t < t_2$, trajectories converging to different modes separate; for $t > t_2$, trajectories all point to the training dataset. In the case of Gaussians mixtures target, they highlight the dependency of t_2 in the dimension and the number of samples, in $\mathcal{O}\left((\log n)/d\right)$, meaning that the first phases are observable only if the number of training points is exponential in the dimension. The methodology they adopt to validate the existence of such t_2 on real data relies on the stochasticity of the backward generative process, which does not hold in the case of flow matching. Our experiments on *learned* flow matching models allow us to take this theoretical study on memorization and temporal behaviors of generative processes a step further.

6 Conclusion, limitations and broader impact

By challenging the assumption that stochasticity in the loss function is a key driver of generalization, our findings help clarify the role of approximation of the exact velocity field in flow matching models. Beyond the different temporal phases in the generation process that we have identified, we expect further results to be obtained by uncovering new properties of the true velocity field.
As a limitation, we have not precisely characterized the velocity field that ends up being learned, in particular how it behaves outside of the trajectories defined by the optimal velocity. Leveraging existing work on the inductive biases of the architectures at hand seems like a promising venue. Though it connects with many scattered theoretical results in the literature, our methodology is mainly empirical with a focus on learned models used in practice: building on our experimental findings to develop new theoretical foundations is a next step for a broader understanding of generation.

Broader impact. We hope that identifying the key factors of generalization will lead to improved training efficiency. However, generative models also raise concerns related to misinformation (notably deepfakes), data privacy, and potential misuse in generating synthetic but realistic content.

References

M. S. Albergo and E. Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *ICLR*, 2023.

G. Biroli, T. Bonnaire, V. de Bortoli, and M. Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957, 2024.

Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang, and A. Ramesh. Video generation models as world simulators. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

G. Cardoso, S. Samsonov, A. Thin, E. Moulines, and J. Olsson. Br-snis: bias reduced self-normalized importance sampling. *NeurIPS*, 35:716–729, 2022.

N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramer, B. Balle, D. Ippolito, and E. Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023.

G. Casella and C. P. Robert. Rao-blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94, 1996.

S. U. H. Dar, A. Ghanaat, J. Kahmann, I. Ayx, T. Papavassiliu, S. O. Schoenberg, and S. Engelhardt. Investigating data memorization in 3d latent diffusion models for medical image synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 56–65. Springer, 2023.

A. Gagneux, S. Martin, R. Emonet, Q. Bertrand, and M. Massias. A visual dive into conditional flow matching. In *The Fourth Blogpost Track at ICLR*, 2025.

W. Gao and M. Li. How do flow matching models memorize and generalize in sample data subspaces? *arXiv preprint arXiv:2410.23594*, 2024.

S. Gong, M. Li, J. Feng, Z. Wu, and L. Kong. Diffuseq: Sequence to sequence text generation with diffusion models. *ICLR*, 2023.

X. Gu, C. Du, T. Pang, C. Li, M. Lin, and Y. Wang. On memorization in diffusion models. *TMLR*, 2025.

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *NeuRIPS*, 2020a.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020b.

C.-W. Huang, J. H. Lim, and A. C. Courville. A variational perspective on diffusion-based generative models and score matching. *Advances in Neural Information Processing Systems*, 34:22863–22876, 2021.

M. Jiralerspong, J. Bose, I. Gemp, C. Qin, Y. Bachrach, and G. Gidel. Feature likelihood divergence: evaluating the generalization of generative models using samples. *NeurIPS*, 2023.

Z. Kadkhodaie, F. Guth, E. P. Simoncelli, and S. Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. *ICLR*, 2024.

M. Kamb and S. Ganguli. An analytic theory of creativity in convolutional diffusion models. *arXiv preprint arXiv:2412.20292*, 2024.

A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. 2009.

S. Li, S. Chen, and Q. Li. A good score does not lead to a good generative model. *arXiv preprint arXiv:2401.04856*, 2024.

Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. *ICLR*, 2023.

Y. Lipman, M. Havasi, P. Holderrieth, N. Shaul, M. Le, B. Karrer, R. T. Chen, D. Lopez-Paz, H. Ben-Hamu, and I. Gat. Flow matching guide and code. *arXiv preprint arXiv:2412.06264*, 2024.

X. Liu, C. Gong, and Q. Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *ICLR*, 2023.

Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

S. Martin, A. Gagneux, P. Hagemann, and G. Steidl. Pnp-flow: Plug-and-play image restoration with flow matching. *ICLR*, 2025.

A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

M. Niedoba, B. Zwartsenberg, K. Murphy, and F. Wood. Towards a mechanistic explanation of diffusion model generalization. *arXiv preprint arXiv:2411.19339*, 2024.

M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. Dinov2: Learning robust visual features without supervision. *TMLR*, 2024.

A. B. Owen. *Monte Carlo theory, methods and examples*. https://artowen.su.domains/mc/, 2013.

G. Parmar, R. Zhang, and J.-Y. Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 2022.

A.-A. Pooladian, H. Ben-Hamu, C. Domingo-Enrich, B. Amos, Y. Lipman, and R. T. Chen. Multisample flow matching: Straightening flows with minibatch couplings. *ICML*, 2023.

C. P. Robert, G. Casella, and G. Casella. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

B. L. Ross, H. Kamkari, T. Wu, R. Hosseinzadeh, Z. Liu, G. Stein, J. C. Cresswell, and G. Loaiza-Ganem. A geometric framework for understanding memorization in generative models. *ICLR*, 2025.

G. Ryzhakov, S. Pavlova, E. Sevriugov, and I. Oseledets. Explicit flow matching: On the theory of flow matching algorithms with applications. *arXiv preprint arXiv:2402.03232*, 2024.

C. Scarvelis, H. S. B. de Ocáriz, and J. Solomon. Closed-form diffusion models. *TMLR*, 2025.

J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015.

G. Somepalli, V. Singla, M. Goldblum, J. Geiping, and T. Goldstein. Understanding and mitigating copying in diffusion models. *NeurIPS*, 36:47783–47803, 2023a.

G. Somepalli, V. Singla, M. Goldblum, J. Geiping, and T. Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6048–6058, 2023b.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *ICLR*, 2021.

Stability AI. https://stability.ai/stablediffusion, 2023. Accessed: 2023-09-09.

G. Stein, J. Cresswell, R. Hosseinzadeh, Y. Sui, B. Ross, V. Villecroze, Z. Liu, A. L. Caterini, E. Taylor, and G. Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *NeurIPS*, 36:3732–3784, 2023.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

A. Tong, N. Malkin, G. Huguet, Y. Zhang, J. Rector-Brooks, K. Fatras, G. Wolf, and Y. Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. In *TMLR*, 2024. URL https://openreview.net/forum?id=CD9Snc73AW.

J. J. Vastola. Generalization through variance: how noise shapes inductive biases in diffusion models. *ICLR*, 2025.

R. Villegas, M. Babaeizadeh, P.-J. Kindermans, H. Moraldo, H. Zhang, M. T. Saffar, S. Castro, J. Kunze, and D. Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *ICLR*, 2022.

M. Xu, T. Geffner, K. Kreis, W. Nie, Y. Xu, J. Leskovec, S. Ermon, and A. Vahdat. Energy-based diffusion language models for text generation. *ICLR*, 2025.

Y. Xu, S. Tong, and T. Jaakkola. Stable target field for reduced variance score estimation in diffusion models. *arXiv preprint arXiv:2302.00670*, 2023.

T. Yoon, J. Y. Choi, S. Kwon, and E. K. Ryu. Diffusion probabilistic models generalize when they fail to memorize. In *ICML 2023 workshop on structured probabilistic inference & generative modeling*, 2023.

H. Zhang, J. Zhou, Y. Lu, M. Guo, P. Wang, L. Shen, and Q. Qu. The emergence of reproducibility and consistency in diffusion models. In *ICML*, 2024.

R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: each claim of the abstract refers to a specific subsection of the paper, that provide empirical evidence of the claim.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We do have a specific section for the limitation of our work

   Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: There are no theoretical results

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provided as many details as possible in order to reproduce the results, in particular, we refer to the public implementation we used, including the specific (default) parameters used.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might

suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code will be made available along with publication

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide a specific appendix with the experimental details

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We do not report error bars, however, we do specify the number of samples used for the FID computation and highlight the strong weaknesses of the FID metric.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: we specified what type of GPU we used

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: there is a dedicated broader impact section

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [No]

Justification: We work on standard image datasets

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We properly refer the `torchcfm` and `PnPflow` codebase.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [No]

Justification: LLMs were only used for grammatical purposes.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

A Proofs of Section 2

A.1 Proof of Proposition 1

Proof. • In the case where $z \sim \hat{p}_{\text{data}}$, conditional probability writes

$$p(z = x^{(i)}|x, t) = \frac{p(x, t, z = x^{(i)})}{p(x, t)} \tag{9}$$

$$= \frac{p(x|t, z = x^{(i)})p(t, z = x^{(i)})}{p(x, t)} \tag{10}$$

$$= \frac{p(x|t, z = x^{(i)})p(t, z = x^{(i)})}{\sum_{i'=1}^{n} p(x, t, z = x^{(i')})} \tag{11}$$

$$= \frac{p(x|t, z = x^{(i)})p(t)\overbrace{p(z = x^{(i)})}^{\frac{1}{n}}}{\sum_{i'=1}^{n} p(x|t, z = x^{(i')})p(t)\underbrace{p(z = x^{(i')})}_{\frac{1}{n}}} \tag{12}$$

$$= \frac{p(x|t, z = x^{(i)})}{\sum_{i'=1}^{n} p(x|t, z = x^{(i')})} \quad . \tag{13}$$

Pluging Equation (13) in Equation (2) yields the closed-formed formula for the velocity field:

$$u^\star(x, t) = \sum_{i=1}^{n} u^{\text{cond}}(x, t, z = x^{(i)})p(z = x^{(i)}|x, t) \tag{14}$$

$$= \sum_{i=1}^{n} u^{\text{cond}}(x, t, z = x^{(i)}) \frac{p(x|t, z = x^{(i)})}{\sum_{i'=1}^{n} p(x|t, z = x^{(i')})} \quad . \tag{15}$$

which proves Equation (5); using that $x|t, z = x^{(i)} \sim \mathcal{N}(tx^{(i)}, (1-t)^2 \operatorname{Id})$ and $u^{\text{cond}}(x, t, z = x^{(i)}) = \frac{x^{(i)} - x}{1-t}$ yields Equation (6).

• For the case $z \sim p_0 \times \hat{p}_{\text{data}}$,

$$\hat{u}^\star(x, t) := \int_z u^{\text{cond}}(x, t, z)p(z|x, t) \, \mathrm{d}z \tag{16}$$

$$= \int_z u^{\text{cond}}(x, t, z) \frac{p(x, z, t)}{p(x, t)} \, \mathrm{d}z \tag{17}$$

$$= \int_z u^{\text{cond}}(x, t, z) \frac{p(x|z, t)p(z)p(t)}{\int_{z'} p(x|t, z')p(t)p(z') \, \mathrm{d}z'} \, \mathrm{d}z \tag{18}$$

$$= \int_z u^{\text{cond}}(x, t, z) \frac{p(x|z, t)p(z)}{\int_{z'} p(x|t, z')p(z') \, \mathrm{d}z'} \, \mathrm{d}z \tag{19}$$

Now for the denominator, since $z \sim p_0 \times \hat{p}_{\text{data}}$:

$$\int_{z'} p(x|t, z')p(z') \, \mathrm{d}z' = \int_{x_0} \sum_{i=1}^{n} \delta_{(1-t)x_0 + tx^{(i)}}(x) \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x_0^2) \mathrm{d}x_0 \tag{20}$$

$$= \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2(1-t)^2}(x - tx^{(i)})^2\right) \tag{21}$$

since $\delta_{(1-t)x_0 + tx^{(i)}}(x)$ imposes $x = (1-t)x_0 + tx^{(i)}$, meaning $x_0 = \frac{x - tx^{(i)}}{1-t}$.

Equivalently, the Dirac imposes $x^{(i)} - x_0 = \frac{x^{(i)}-x}{1-t}$, hence for the numerator:

$$\int_z u^{\text{cond}}(x,t,z)p(x|z,t)p(z)\,\mathrm{d}z \int_{x_0} \sum_{i=1}^{n}(x^{(i)}-x_0)\delta_{(1-t)x_0+tx^{(i)}}(x)\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}\|x_0\|^2)\mathrm{d}x_0 \tag{22}$$

$$= \sum_{i=1}^{n} \frac{x^{(i)}-x}{1-t}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2(1-t)^2}\|x-tx^{(i)})\|^2\right) \tag{23}$$

Taking the ratio of Equations (21) and (23) concludes the proof. □

B Additional details and comments on empirical flow matching

First, recalls on the optimal velocity (Equation (6)) and the empirical flow matching loss (Equations (7) and (8)) are provided in Appendix B.1. The unbiasedness of the estimator is presented in Appendix B.2, and its proof is in Appendix B.3.

B.1 Recalls

The closed-form formula of the "optimal" velocity field is:

$$\hat{u}^\star(x,t) = \sum_{l=1}^{n} \frac{x^{(l)}-x}{1-t} \cdot \left[\text{softmax}\left(\left(-\frac{\|x-tx^{(k)}\|^2}{2(1-t)^2}\right)_{k=1,\dots,n}\right)\right]_l . \tag{6}$$

The proposed loss uses mini-batches of size M (instead of all n training points) to build an estimator $\hat{u}^\star_M$ of $\hat{u}^\star$:

$$\mathcal{L}_{\text{EFM}}(\theta) = \mathbb{E}_{\substack{t\sim\mathcal{U}([0,1]) \\ x_0\sim p_0 \\ x_1\sim\hat{p}_{\text{data}} \\ x_t=(1-t)x_0+tx_1 \\ b^{(1)}:=x_1\ ;\ b^{(2)},\dots,b^{(M)}\sim\hat{p}_{\text{data}}}} \|u_\theta(x_t,t) - \hat{u}^\star_M(x_t,t)\|^2 \ , \tag{7}$$

with

$$\hat{u}^\star_M(x_t,t) = \sum_{j=1}^{M} \frac{b^{(j)}-x_t}{1-t} \cdot \left[\text{softmax}\left(\left(-\frac{\|x_t-tb^{(k)}\|^2}{2(1-t)^2}\right)_{k=1,\dots,M}\right)\right]_j . \tag{8}$$

Crucially, in Equation (7) the sample $b^{(1)}$ depends on x_t and is reused in the estimate $\hat{u}^\star_M$. This important detail yields an unbiased estimator of $\hat{u}^\star$.

B.2 Theoretical properties of the proposed estimator

Proposition 2. *Let $\hat{p}_{\text{data}}(z|x,t)$ denote $p(z|x,t)$, meaning a probability distribution on $\{x^{(i)}\}_{i=1}^n$, with expression given by Equation (13). With no constraints on the learned velocity field u_θ,*

i) The minimizer of Equation (7) writes, for all (x,t)

$$\mathbb{E}_{b^{(1)}\sim\hat{p}_{\text{data}}(\cdot|x,t)\ ;\ b^{(2)},\dots,b^{(M)}\sim\hat{p}_{\text{data}}}\left[\hat{u}^\star_M(x,t)\right] \ . \tag{24}$$

ii) In addition, for all (x,t), the minimizer of Equation (7) equals the optimal velocity field, i.e.,

$$\mathbb{E}_{b^{(1)}\sim\hat{p}_{\text{data}}(\cdot|x,t)\ ;\ b^{(2)},\dots,b^{(M)}\sim\hat{p}_{\text{data}}}\left[\hat{u}^\star_M(x,t)\right] = \hat{u}^\star(x,t) \ . \tag{25}$$

iii) The conditional variance of the estimator $\hat{u}^\star_M$ is smaller than the usual conditional variance:

$$\text{Var}_{b^{(1)}\sim\hat{p}_{\text{data}}(\cdot|x,t)\ ;\ b^{(2)},\dots,b^{(M)}\sim\hat{p}_{\text{data}}}\left[\hat{u}^\star_M(x,t)\right] \le \text{Var}_{b^{(1)}\sim\hat{p}_{\text{data}}(\cdot|x,t)}\left[u^{\text{cond}}(x,b^{(1)},t)\right] . \tag{26}$$

The proof of Proposition 2 can be found in Appendix B.3. First, we discuss below the relation between Proposition 2 and the sampling literature.

Links with importance sampling. The estimator $\hat{u}^\star$ in Equation (6) can be seen as a form of *importance sampling* (see Robert et al. 1999, Chap. 3 for an in-depth reference). In a nutshell, importance sampling is a way to estimate an expectation when one cannot easily sample from the random variable it depends on. More precisely, in the ideal case $z \sim p_{\text{data}}$ (as opposed to $z \sim \hat{p}_{\text{data}}$), the velocity field formula is the following

$$u^\star(x_t, t) = \mathbb{E}_{z|x_t,t}\left[u^{\text{cond}}(x_t, z, t)\right] \tag{27}$$

$$= \int_z u^{\text{cond}}(x_t, z, t) p(z|x_t, t) \mathrm{d}z \ . \tag{28}$$

When $z \sim p_{\text{data}}$, it is difficult to sample from $z|x_t, t$, but the latter equation can be rewritten as

$$u^\star(x_t, t) = \int_z u^{\text{cond}}(x_t, z, t) \frac{p(z|x_t, t)}{p(z)} p(z) \mathrm{d}z \tag{29}$$

and one can easily sample from $z \sim \hat{p}_{\text{data}}$ using the empirical data distribution $x^{(1)}, \dots, x^{(n)}$

$$u^\star(x_t, t) \approx \frac{1}{n} \sum_{i=1}^{n} u^{\text{cond}}(x_t, x^{(i)}, t) \frac{p(z = x^{(i)}|x_t, t)}{p(x^{(i)})} \tag{30}$$

$$= \sum_{i=1}^{n} u^{\text{cond}}(x_t, x^{(i)}, t) p(z = x^{(i)}|x_t, t) \tag{31}$$

$$:= \hat{u}^\star(x_t, t) \ . \tag{32}$$

Links with self-normalized importance sampling. The estimator $\hat{u}^\star_M$ in Equation (8) draws strong connections with self-normalized importance sampling (see Owen 2013, Chap. 9.2), and Rao-Blackwelled estimators (Casella and Robert, 1996; Cardoso et al., 2022). Very importantly, as discussed in Ryzhakov et al. (2024), self-normalized important sampling estimators are usually biased, in the sense that in general:

$$\mathbb{E}_{b^{(1)},\dots,b^{(M)} \sim \hat{p}_{\text{data}}} \hat{u}^\star_M(x_t, t) \neq u^\star(x_t, t) \ . \tag{33}$$

A crucial point is that our estimator includes $b^{(1)} \sim \hat{p}_{\text{data}}(\cdot|x_t, t)$, which yields the main result of Proposition 2

$$\mathbb{E}_{b^{(1)} \sim \hat{p}_{\text{data}}(\cdot|x_t,t), b^{(2)},\dots,b^{(M)} \sim p_{\text{data}}} \left[\hat{u}^\star_M(x_t, t)\right] = u^\star(x_t, t) \ . \tag{34}$$

Finally, this idea of regressing against a more deterministic target derived from the optimal closed-form velocity field has also been empirically explored for diffusion models (Xu et al., 2023).

B.3 Proof of Proposition 2

Proof of Item (i)). With no constraints on u_θ, the empirical flow matching loss writes:

$$\mathbb{E}_{\substack{t \sim \mathcal{U}([0,1]) \\ x_1 \sim \hat{p}_{\text{data}} \\ x_t = (1-t)x_0 + t x_1 \\ b^{(1)} := x_1 \ ; \ b^{(2)},\dots,b^{(M)} \sim \hat{p}_{\text{data}}}} \|u_\theta(x_t, t) - \hat{u}^\star_M(x_t, t)\|^2 \ , \tag{35}$$

$$= \mathbb{E}_{\substack{t \sim \mathcal{U}([0,1]) \\ x_t \sim p_t}} \mathbb{E}_{\substack{b^{(1)} \sim \hat{p}_{\text{data}}(\cdot|x_t,t) \\ b^{(2)},\dots,b^{(M)}|x_t,t}} \|u_\theta(x_t, t) - \hat{u}^\star_M(x_t, t)\|^2 \ , \tag{36}$$

$$= \mathbb{E}_{\substack{t \sim \mathcal{U}([0,1]) \\ x_t \sim p_t}} \mathbb{E}_{\substack{b^{(1)} := \hat{p}_{\text{data}}(\cdot|x_t,t) \\ b^{(2)},\dots,b^{(M)} \sim \hat{p}_{\text{data}}}} \|u_\theta(x_t, t) - \hat{u}^\star_M(x_t, t)\|^2 \ \text{ because } b^{(2)}, \dots, b^{(M)} \perp\!\!\!\perp x_t, t \ , \tag{37}$$

which is minimized when for all x_t, t

$$u_\theta(x_t, t) = \mathbb{E}_{\substack{b^{(1)} \sim \hat{p}_{\text{data}}(\cdot|x_t,t) \\ b^{(2)},\dots,b^{(M)} \sim \hat{p}_{\text{data}}}} \left[\hat{u}^\star_M(x_t, t)\right] \ . \tag{38}$$

□

Proof of Item (ii)). The minimizer for a given (x_t, t), removing these elements from the notation for conciseness and abstraction, is a weighted mean:

$$\hat{u}^\star(x_t, t) = \hat{u}^\star = \sum_{l=1}^{n} w^{(l)} u^{(l)} \ , \text{ with} \tag{39}$$

$$w^{(l)} = \hat{p}_{\text{data}}(z = x^{(l)} | t, x_t) \ , \ \sum_{l=1}^{n} w^{(l)} = 1 \tag{40}$$

$$u^{(l)} = u^{\text{cond}}(x_t, x^{(l)}, t) \tag{41}$$

We express a mini-batch as an M-valued vector of indices, $\boldsymbol{i} \in [\![1, n]\!]^M$. The mini-batch estimate from Equation (7), considering the definition of the $\mathrm{softmax}$, can be expressed as a mini-batch weighted-mean:

$$\hat{u}^\star_M(\mathbf{i}) = \frac{\sum_{j=1}^{M} w^{(\mathbf{i_j})} u^{(\mathbf{i_j})}}{\sum_{j=1}^{M} w^{(\mathbf{i_j})}} \tag{42}$$

The categorical distribution over $[\![1, n]\!]$ with probabilities following the weights w in (40) is denoted $\mathrm{Cat}(w)$ and the uniform distribution, *i.e.,* $\mathrm{Cat}(\mathbf{1}/n)$), is denoted Unif.

The main result of the following is that, in expectation over the biased-mini-batches, **where the first point is drawn according to** w and the $M-1$ other points are drawn uniformly, the mini-batch weighted-mean is an unbiased estimate of the w-weighted-mean $\hat{u}^\star$.

$$\mathbb{E}\left[\hat{u}^\star_M(\mathbf{i})\right] := \mathbb{E}_{\mathbf{i_1} \sim \mathrm{Cat}(w)} \mathbb{E}_{\boldsymbol{i}_2, \ldots, \boldsymbol{i}_M \sim \mathrm{Unif}} \left[\hat{u}^\star_M(\mathbf{i})\right] \tag{43}$$

$$= \sum_{\boldsymbol{i}_1=1}^{n} w^{(\boldsymbol{i}_1)} \mathbb{E}_{\boldsymbol{i}_2, \ldots, \boldsymbol{i}_M \sim \mathrm{Unif}} \left[\hat{u}^\star_M(\mathbf{i})\right] \tag{44}$$

$$= \sum_{\boldsymbol{i}_1=1}^{n} \mathbb{E}_{\boldsymbol{i}_2, \ldots, \boldsymbol{i}_M \sim \mathrm{Unif}} \left[w^{(\boldsymbol{i}_1)} \hat{u}^\star_M(\mathbf{i})\right] \tag{45}$$

$$= n \sum_{\boldsymbol{i}_1=1}^{n} \frac{1}{n} \mathbb{E}_{\boldsymbol{i}_2, \ldots, \boldsymbol{i}_M \sim \mathrm{Unif}} \left[w^{(\boldsymbol{i}_1)} \hat{u}^\star_M(\mathbf{i})\right] \tag{46}$$

$$= n \, \mathbb{E}_{\boldsymbol{i}_1 \sim \mathrm{Unif}} \mathbb{E}_{\boldsymbol{i}_2, \ldots, \boldsymbol{i}_M \sim \mathrm{Unif}} \left[w^{(\boldsymbol{i}_1)} \hat{u}^\star_M(\mathbf{i})\right] \tag{47}$$

$$= n \, \mathbb{E}_{\boldsymbol{i}_1, \ldots, \boldsymbol{i}_M \sim \mathrm{Unif}} \left[w^{(\boldsymbol{i}_1)} \hat{u}^\star_M(\mathbf{i})\right] \tag{48}$$

The expression in Equation (48) is invariant with respect to order of the indices $\boldsymbol{i}_1, \ldots, \boldsymbol{i}_M$: the indices in expectation in Equation (48) can be exchanged, and one thus has

$$\forall k \in [\![1, M]\!], \ \mathbb{E}\left[\hat{u}^\star_M(\mathbf{i})\right] = n \, \mathbb{E}_{\boldsymbol{i}_1, \ldots, \boldsymbol{i}_M \sim \mathrm{Unif}} \left[w^{(\boldsymbol{i}_k)} \hat{u}^\star_M(\mathbf{i})\right] \ . \tag{49}$$

Averaging Equation (49) over the indices $k \in [\![1, M]\!]$ yields the desired result

$$\frac{1}{M}\sum_{k=1}^{M}\mathbb{E}\hat{u}^{\star}_{M}(\mathbf{i}) = \frac{1}{M}\sum_{k=1}^{M} n\ \mathbb{E}_{\boldsymbol{i}_1,\ldots,\boldsymbol{i}_M\sim\mathrm{Unif}}\left[w^{(\boldsymbol{i}_k)}\hat{u}^{\star}_{M}(\mathbf{i})\right] \tag{50}$$

$$\mathbb{E}\hat{u}^{\star}_{M}(\mathbf{i}) = \frac{1}{M} n\ \mathbb{E}_{\boldsymbol{i}_1,\ldots,\boldsymbol{i}_M\sim\mathrm{Unif}}\left[\sum_{k=1}^{M} w^{(\boldsymbol{i}_k)}\hat{u}^{\star}_{M}(\mathbf{i})\right] \tag{51}$$

$$= \frac{1}{M} n\ \mathbb{E}_{\boldsymbol{i}_1,\ldots,\boldsymbol{i}_M\sim\mathrm{Unif}}\left[\sum_{k=1}^{M} w^{(\boldsymbol{i}_k)}\frac{\sum_{j=1}^{M} w^{(\mathbf{i_j})}u^{(\mathbf{i_j})}}{\sum_{j=1}^{M} w^{(\mathbf{i_j})}}\right] \tag{52}$$

$$= \frac{1}{M} n\ \mathbb{E}_{\boldsymbol{i}_1,\ldots,\boldsymbol{i}_M\sim\mathrm{Unif}}\left[\left(\sum_{k=1}^{M} w^{(\boldsymbol{i}_k)}\right)\frac{\sum_{j=1}^{M} w^{(\mathbf{i_j})}u^{(\mathbf{i_j})}}{\left(\sum_{j=1}^{M} w^{(\mathbf{i_j})}\right)}\right] \tag{53}$$

$$= \frac{1}{M} n\ \mathbb{E}_{\boldsymbol{i}_1,\ldots,\boldsymbol{i}_M\sim\mathrm{Unif}}\left[\sum_{j=1}^{M} w^{(\mathbf{i_j})}u^{(\mathbf{i_j})}\right] \tag{54}$$

$$= \frac{1}{M} n\ \sum_{j=1}^{M}\mathbb{E}_{\boldsymbol{i}_1,\ldots,\boldsymbol{i}_M\sim\mathrm{Unif}}\left[w^{(\mathbf{i_j})}u^{(\mathbf{i_j})}\right] \tag{55}$$

$$= \frac{1}{M} n\ \sum_{j=1}^{M}\mathbb{E}_{\boldsymbol{i}_j\sim\mathrm{Unif}}\left[w^{(\mathbf{i_j})}u^{(\mathbf{i_j})}\right] \tag{56}$$

$$= \frac{1}{M} n\ M\ \mathbb{E}_{l\sim\mathrm{Unif}}\left[w^{(l)}u^{(l)}\right] \tag{57}$$

$$= n\ \mathbb{E}_{l\sim\mathrm{Unif}}\left[w^{(l)}u^{(l)}\right] \tag{58}$$

$$= n\ \sum_{l=1}^{n}\frac{1}{n}\left[w^{(l)}u^{(l)}\right] \tag{59}$$

$$= \sum_{l=1}^{n}\left[w^{(l)}u^{(l)}\right] \tag{60}$$

$$= \hat{u}^{\star} \tag{61}$$

□

Proof of Item (iii)). Using the same ideas as for Item (ii)), one has

$$\mathbb{E}_{x^{(1)}\sim\hat{p}_{\text{data}}(\cdot|x_t,t)\ ;\ b^{(2)},\dots,b^{(M)}\sim\hat{p}_{\text{data}}}\left[\hat{u}_M^\star(x_t,t)^2\right] \tag{62}$$

$$= n\mathbb{E}_{\boldsymbol{i}_1,\dots,\boldsymbol{i}_M\sim\text{Unif}}\left[w^{(\boldsymbol{i}_1)}\hat{u}_M^\star(\boldsymbol{i})^2\right] \tag{63}$$

$$= n\mathbb{E}_{\boldsymbol{i}_1,\dots,\boldsymbol{i}_M\sim\text{Unif}}\left[w^{(\boldsymbol{i}_k)}\hat{u}_M^\star(\boldsymbol{i})^2\right], \forall k\in[\![1,M]\!] \tag{64}$$

$$= n\frac{1}{M}\mathbb{E}_{\boldsymbol{i}_1,\dots,\boldsymbol{i}_M\sim\text{Unif}}\left[\sum_{k=1}^M w^{(\boldsymbol{i}_k)}\hat{u}_M^\star(\boldsymbol{i})^2\right] \tag{65}$$

$$= n\frac{1}{M}\mathbb{E}_{\boldsymbol{i}_1,\dots,\boldsymbol{i}_M\sim\text{Unif}}\left[\sum_{k=1}^M w^{(\boldsymbol{i}_k)}\left(\frac{\sum_{j=1}^M w^{(\boldsymbol{i}_j)}u^{(\boldsymbol{i}_j)}}{\sum_{j=1}^M w^{(\boldsymbol{i}_j)}}\right)^2\right] \tag{66}$$

$$\leq n\frac{1}{M}\mathbb{E}_{\boldsymbol{i}_1,\dots,\boldsymbol{i}_M\sim\text{Unif}}\left[\sum_{k=1}^M w^{(\boldsymbol{i}_k)}\frac{\sum_{j=1}^M w^{(\boldsymbol{i}_j)}(u^{(\boldsymbol{i}_j)})^2}{\sum_{j=1}^M w^{(\boldsymbol{i}_j)}}\right] \text{ by convexity of } x\mapsto x^2 \tag{67}$$

$$= n\frac{1}{M}\mathbb{E}_{\boldsymbol{i}_1,\dots,\boldsymbol{i}_M\sim\text{Unif}}\left[\left(\sum_{k=1}^M w^{(\boldsymbol{i}_k)}\right)\frac{\sum_{j=1}^M w^{(\boldsymbol{i}_j)}(u^{(\boldsymbol{i}_j)})^2}{\sum_{j=1}^M w^{(\boldsymbol{i}_j)}}\right] \tag{68}$$

$$= n\frac{1}{M}\mathbb{E}_{\boldsymbol{i}_1,\dots,\boldsymbol{i}_M\sim\text{Unif}}\left[\sum_{j=1}^M w^{(\boldsymbol{i}_j)}(u^{(\boldsymbol{i}_j)})^2\right] \tag{69}$$

$$= \mathbb{E}_{\boldsymbol{i}_1\sim\text{Unif}}\left[w^{(\boldsymbol{i}_1)}(u^{(\boldsymbol{i}_1)})^2\right] \tag{70}$$

$$= \mathbb{E}_{l\sim\text{Unif}}\left[w^{(l)}(u^{(l)})^2\right] \ . \tag{71}$$

Hence

$$\mathbb{E}_{x^{(1)}\sim\hat{p}_{\text{data}}(\cdot|x_t,t)\ ;\ b^{(2)},\dots,b^{(M)}\sim\hat{p}_{\text{data}}}\left[\hat{u}_M^\star(x_t,t)^2\right] - (\hat{u}^\star)^2 \leq \mathbb{E}_{l\sim\text{Unif}}\left[w^{(l)}(u^{(l)})^2\right] - (\hat{u}^\star)^2 \ , \tag{72}$$

which is exactly

$$\text{Var}_{x^{(1)}\sim\hat{p}_{\text{data}}(\cdot|x_t,t)\ ;\ b^{(2)},\dots,b^{(M)}\sim\hat{p}_{\text{data}}}\left[\hat{u}_M^\star(x_t,t)\right] \leq \text{Var}_{x^{(1)}\sim\hat{p}_{\text{data}}(\cdot|x_t,t)}\left[u^{\text{cond}}(x_t,x^{(1)},t)\right] \ . \tag{73}$$

□

C Experiments details

For all the experiment we used all the same learning hyperparameters, the default ones form Tong et al. (2024). The hyperparameter values are summarized in Table 1. The details specific to each figure are described in Appendices C.1 to C.4

# Channels	Batch Size	Learning Rate	EMA Decay	Gradient Clipping
128	128	0.0002	0.9999	1

Table 1: Learning hyperparameters for all the CIFAR-10 and CelebA 64 experiments.

C.1 Figures 1b and 2

For Figure 1b no deep learning is involved: the datasets 2-moons and CIFAR-10 are loaded. Then, 256 points from $p_0\times\hat{p}_{\text{data}}$ are drawn, and one computes the mean of the cosine similarities between $\hat{u}^\star((1-t)x_0+tx_1,t)$ and $u^{\text{cond}}((1-t)x_0+tx_1,z=x_1,t)=x_1-x_0$, for each value of $t\in\{0,1/100,2/100,\dots,99/100\}$.

No deep learning either is involved in Figure 2: the Imagenette dataset is loaded and spatially subsampled to resolution $\mathrm{dim} = 8$, $\mathrm{dim} = 16$, $\ldots$, $\mathrm{dim} = 256$, *i.e.,* with $d = " \cdot 8^2$, $d = 3 \cdot 16^2$, $\ldots$, $d = 3 \cdot 256^2$. Then, as for Figure 1b, batches of 256 points from p_0 and p_{data} are drawn, and one computes the percentage of cosine similarities between $\hat{u}^\star((1-t)x_0 + tx_1, t)$ and $u^{\mathrm{cond}}((1-t)x_0 + tx_1, z = x_1, t) = x_1 - x_0$, that are larger than 0.9, for multiple time values t.

C.2 Figure 3

In Figure 3, networks are trained with a vanilla conditional flow matching, with the standard 34 million parameters U-Net for diffusion by Nichol and Dhariwal (2021), with default settings from the `torchfm` codebase [3] (Tong et al., 2024). Training uses the CFM loss. For this specific experiment, **we removed the usual random flip transform**, for $\hat{u}^\star$ to be simpler and easier to estimate by u_θ. For each "data" subsampling of the dataset, we trained the model for $5 \cdot 10^4$ iterations, with a batch size of 128, *i.e.,* we trained the models for 128 epochs.

C.3 Figure 4

In Figure 4, for each dataset (CIFAR-10 and CelebA 64), one network is trained using a vanilla conditional flow matching with the default parameters of Tong et al. (2024) (the most important ones are recalled in Table 1). Then images are generated first following the closed-form formula of the optimal velocity field $\hat{u}^\star$ from 0 to τ. And then following the velocity field learned with a usual conditional flow matching u_θ from τ to 1.

C.4 Figure 5

For experiments involving training on CIFAR-10 (Figures 3 and 4), we rely on the standard 34 million parameters U-Net for diffusion by Nichol and Dhariwal (2021), with default settings from the `torchfm` codebase (Tong et al., 2024). For each algorithm, the networks are trained for 500k iterations with batch size 128, *i.e.,* 1280 epochs.

For CelebA 64×64 (Figure 4), we rely on the training script of `pnpflow` library[4] (Martin et al., 2025), which uses a U-Net from Huang et al. (2021); Ho et al. (2020b).

[3] https://github.com/atong01/conditional-flow-matching
[4] https://github.com/annegnx/PnP-Flow