# EXPLOITING LOW FREQUENCIES AS A PRIOR FOR OPTIMAL TRANSPORT FLOWS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Optimal transport (OT) couplings between prior and data straighten flow matching trajectories and enable fast, even single-step, generation. But computing OT is cubic in the number of samples and existing methods solve the problem approximately at the cost of persistent bias or significant overhead. Instead of treating the prior distribution (typically, a Gaussian) as a given, we treat it as a design choice. The question now is what prior distribution simplifies OT coupling? There might be many different priors that achieve this, leaving us free to choose one that is also tractable to sample. We identify low-frequency projection of natural images as such a choice, and show that the identity coupling between data and its low-frequency representation is OT-optimal for nearly all sampled pairs. The prior is structured enough to be sampled by a lightweight model at inference, and the remaining flow-matching task reduces to synthesizing high-frequency detail. The criterion admits more than the deterministic projection. Interpolating the prior with Gaussian noise improves generation quality and the identity coupling remains optimal. The approach requires no modifications to the flow model itself, and integrates naturally with latent-space models, classifier-free guidance, and one-step generation frameworks. Our method reduces trajectory curvature by more than $2\times$ relative to the strongest baseline on each benchmark. [1]

## 1 INTRODUCTION

Generative models have become one of the most transformative tools in modern machine learning. In just a few years, they have scaled to produce photorealistic images at high resolution (Esser et al., 2024; Labs, 2024; Ma et al., 2024; Rombach et al., 2022), generate temporally coherent video (Jin et al., 2025; HaCohen et al., 2024; Polyak et al., 2024; Ho et al., 2022b), and are rapidly expanding to new modalities including audio (Le et al., 2023; Mehta et al., 2024; Borsos et al., 2023), 3D (Hui et al., 2025; LAN et al., 2025), and beyond (Jing et al., 2023; Song et al., 2023b). They are the backbone of both open-source and commercial generation systems (Esser et al., 2024; Labs, 2024; Polyak et al., 2024), powering creative tools used by millions.

At the heart of every generative model lies a single conceptual task, learn to transform a simple, easy-to-sample distribution into a complex, unknown target distribution from which we only observe finite samples. If one could compute the optimal transport (OT) map between a simple distribution and the true underlying data distribution, generation would reduce to a single map evaluation, transforming any noise sample directly into a data sample. This would be the perfect generator, a direct and cost-minimizing transformation from noise to data, and the ideal we aim to approach, high-quality generation in as few steps as possible, ideally just one.

But this map is out of reach. Computing the optimal transport coupling between a high-dimensional data distribution and a standard Gaussian prior is computationally prohibitive at modern scales. Exact solvers scale cubically in the number of samples (Kuhn, 1955), while entropic relaxations (Cuturi, 2013) remain prohibitive in the dimensionalities of modern data. Faced with this intractability, the field turned to flow matching (Lipman et al., 2022; Liu et al., 2022; Albergo & Vanden-Eijnden, 2022). Rather than computing the transport map directly, flow matching learns a velocity field that continuously deforms the simple distribution into the data distribution over a unit time interval. In

[1] Our code will be released publicly.



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.

standard practice, this velocity field is trained on independently sampled noise-data pairs, forgoing any global coupling structure. This has proven remarkably effective, and flow-based architectures now underpin the leading generation systems (Esser et al., 2024; Labs, 2024; HaCohen et al., 2024; Jin et al., 2025).

Yet flow matching inherits a fundamental tension from the intractability it circumvents. Since the true OT coupling is unknown, standard training pairs noise and data independently at random. This forces the learned velocity field to average over conflicting transport directions, producing curved, crossing trajectories that demand many integration steps at inference (Pooladian et al., 2023; Tong et al., 2024). This is the central obstacle to efficient few-step and single-step generation, since straighter trajectories require fewer steps to traverse.

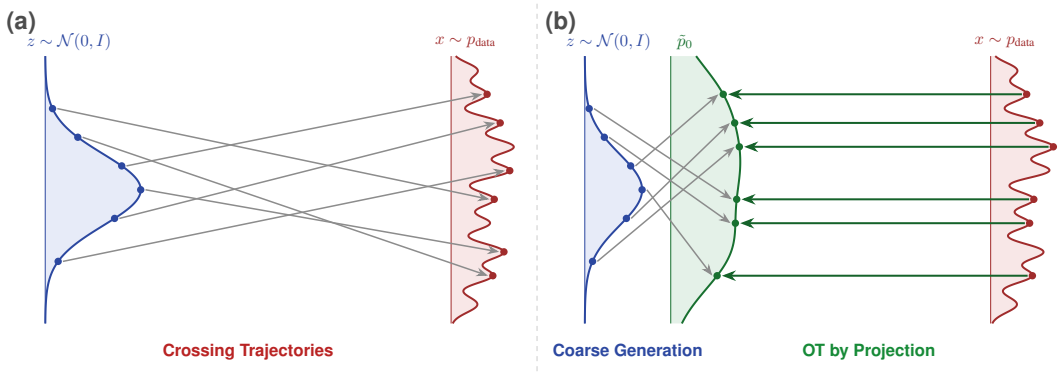


Figure 1: **Designing the Prior, Not the Coupling.** **(a)** Random noise-data pairing in standard flow matching (Lipman et al., 2022) induces crossing trajectories, motivating an OT coupling. **(b)** Projecting samples onto their low-frequency representation defines an intermediate distribution $\tilde{p}_0$ paired with the data by the identity coupling, while the mapping from Gaussian noise to $\tilde{p}_0$ reduces to coarse generation.

A number of works have sought to recover what random pairing leaves on the table. Minibatch OT methods (Pooladian et al., 2023; Tong et al., 2024) solve small transport problems within each training batch, but these local approximations introduce persistent bias that does not vanish with training (Kornilov et al., 2024). To move beyond single-batch scope, some methods accumulate and refine couplings across minibatches throughout training (Davtyan et al., 2025), while semi-discrete OT approaches (Kong et al., 2025; Mousavi-Hosseini et al., 2025a) solve a global transport problem over the full dataset as a preprocessing step, adding significant overhead. Distillation (Liu et al., 2022; seong Kim et al., 2025) and trajectory optimization (Yue et al., 2025) straighten paths post-training, adding further stages on top of a pretrained model. All of these methods treat the OT problem as something to be solved.

We argue it can be reframed instead, treating the prior itself as a variable to be chosen rather than a fixed input. As illustrated in Fig. 1a, standard flow matching pairs a Gaussian distribution with a complex target through random coupling, producing crossing trajectories that curve the learned velocity field and demand many integration steps at inference. Rather than attempting to solve or approximate the OT coupling for a fixed Gaussian prior, we hold the coupling fixed at the identity and ask what this requires of the *prior*. OT-optimality then becomes a criterion against which a candidate prior is verified, rather than a coupling to be computed. Concretely, we seek a *new prior distribution*, defined by a transformation of the data, such that the identity coupling to the data meets this criterion, the new prior can be sampled efficiently and generation from the target distribution remains faithful. Fig. 1b illustrates this idea, where a structured intermediate distribution $\tilde{p}_0$ replaces the Gaussian. Data and prior are paired by the identity, while the remaining mapping from Gaussian noise to $\tilde{p}_0$ only needs to generate coarse image structure in a lower-dimensional space.

Such a prior can be obtained using a simple, fixed transformation that requires no data-dependent learning or optimization. Leveraging the spectral concentration of natural images (Ruderman, 1994), we project each image onto its low-frequency representation. Most of the $\ell_2$ energy of a natural image lives in its low frequencies, so the projection retains coarse structure while discarding only fine detail. This yields a prior that is informative enough to keep distinct images well separated,

and simple enough to be sampled by a lightweight generator. The identity coupling to this prior is empirically OT-optimal.

Optimality alone, however, does not make a prior easy to generate from. A deterministic projection leaves the prior on a low-dimensional subspace, where velocity fields are hard to learn (Song & Ermon, 2019), and sample quality suffers. Following the remedy adopted there, we interpolate the low-frequency projection with Gaussian noise, a choice the coupling criterion leaves free, and obtain a prior far better suited to generation. This decomposes generation into a low-dimensional problem and an OT-structured one. The first stage generates only coarse image structure at a fraction of the original dimensionality, and the second synthesizes the remaining high-frequency detail at full resolution.

Our contributions are as follows: **(1)** we present a new perspective on optimal transport in flow matching. Rather than solving an OT problem for a fixed prior, we design the prior so that its identity coupling to the data is empirically OT-optimal, rather than approximate at significant cost as existing methods do. **(2)** We construct this prior using low-frequency projections of training images, verify the resulting identity coupling is OT-optimal with the Hungarian algorithm, interpolate the prior with Gaussian noise, a choice the coupling criterion leaves free, which improves generation quality while preserving the coupling, and show that OT-optimality alone is not sufficient without low-frequency structure. **(3)** We show that this design consistently reduces trajectory curvature by more than $2\times$ relative to the strongest baseline on each benchmark, yields strong few-step and single-step generation across CIFAR-10, FFHQ, and ImageNet, and integrates naturally with latent-space models, classifier-free guidance, and one-step frameworks such as MeanFlow (Geng et al., 2025).

## 2 Related work

Flow matching (Lipman et al., 2022; Albergo & Vanden-Eijnden, 2022; Liu et al., 2022) trains continuous normalizing flows (Chen et al., 2018), enabling simulation-free optimization on linear interpolation paths, and now powers state-of-the-art image synthesis (Labs, 2024; Esser et al., 2024). Pairing noise and data independently during training produces curved generation trajectories (Tong et al., 2024; Pooladian et al., 2023) that demand many Ordinary Differential Equation (ODE) solver steps at inference. Two complementary lines of work address this, improving the noise-data coupling (Sec. 2.1), and replacing the pure-noise prior with more informative starting distributions (Sec. 2.2). Our work connects them, showing that a simple choice of prior can simultaneously yield OT couplings and straight trajectories.

### 2.1 Straightening generation trajectories

**Minibatch OT coupling.** These methods (Pooladian et al., 2023; Tong et al., 2024) solve an OT problem within each training batch. The resulting coupling is locally straighter but biased (Kornilov et al., 2024; Mousavi-Hosseini et al., 2025b), and exact per-batch assignment is cubic in the batch size and scales poorly with dimension (Zhang et al., 2025; Mousavi-Hosseini et al., 2025b). A recent variant (Lin et al., 2025) selects pairs by model prediction error rather than geometric cost, but remains within single batches.

**Global OT coupling.** Several methods compute a coupling across the full dataset. One propagates assignments across overlapping batches during training (Davtyan et al., 2025), while AlignFlow (Kong et al., 2025) and SD-FM (Mousavi-Hosseini et al., 2025a) solve a semi-discrete OT problem offline. These yield higher-quality couplings than minibatch OT, but require dataset-wide precomputation and tie assignments to specific training samples rather than learning a source distribution.

**Post-training straightening.** A separate line of work straightens trajectories after the base model has been trained, through iterative self-distillation (Liu et al., 2022), guided distillation (seong Kim et al., 2025), or action minimization (Yue et al., 2025). These methods are complementary to the choice of coupling and can be applied on top of any training-time method, including ours, to further refine trajectories.

**Few-step generation via modified objectives.** Consistency models (Song et al., 2023a; Yang et al., 2024), shortcut models (Frans et al., 2025), and MeanFlow (Geng et al., 2025) modify the training objective itself for few-step or single-step generation. Because they leave the noise-data

coupling unchanged, they are complementary to our prior design and can be composed with it, as we demonstrate by integrating our prior into MeanFlow in Sec. 5.3.

### 2.2 Alternative priors for generation

The standard isotropic Gaussian prior discards all data structure. A growing body of work replaces or augments this pure-noise prior with degradations that retain partial information about the target. Cold Diffusion (Bansal et al., 2023) showed that purely deterministic degradations such as blur, downsampling, and masking suffice for generation when inverted by a learned model. Subsequent work formalized frequency-aware destruction (Hoogeboom & Salimans, 2023), interpolated between blur and noise (Daras et al., 2023), and corrupted Fourier magnitude while preserving spatial phase (Zeng et al., 2025). Anisotropic Gaussians capturing local manifold geometry (Bamberger et al., 2025) and time-varying blue noise (Huang et al., 2024) have similarly been shown to improve flow matching and diffusion respectively. In the conditional setting, priors reflecting class or prompt structure shorten transport paths (Issachar et al., 2025; Kim et al., 2026).

Table 1: **Trade-Offs in Trajectory Straightening.** *From Scratch*: no pre-trained model required. *OT-Based*: coupling derived from optimal transport. *Scales w/ $B$*: no per-batch OT solve. *Scales w/ $N$*: no dataset-wide precomputation. *Dataset-agnostic*: no assignment tied to training samples.

| | From Scratch | OT-Based | Scales w/ batch $B$ | Scales w/ data $N$ | Dataset-agnostic |
|---|---|---|---|---|---|
| OT-FM / Multisample FM (Tong et al., 2024; Pooladian et al., 2023) | ✓ | ✓ | ✗ | ✓ | ✓ |
| OFM (Kornilov et al., 2024) | ✓ | ✓ | ✗ | ✗ | ✗ |
| LOOM-CFM (Davtyan et al., 2025) | ✓ | ✓ | ✗ | ✓ | ✗ |
| AlignFlow (Kong et al., 2025) | ✓ | ✓ | ✓ | ✗ | ✗ |
| SD-FM (Mousavi-Hosseini et al., 2025a) | ✓ | ✓ | ✓ | ✗ | ✗ |
| Lee *et al.* (Lee et al., 2023) | ✓ | ✗ | ✓ | ✓ | ✓ |
| Reflow (Liu et al., 2022) | ✗ | ✗ | ✓ | ✓ | ✓ |
| Conic Reflow (seong Kim et al., 2025) | ✗ | ✗ | ✓ | ✓ | ✓ |
| OAT-FM (Yue et al., 2025) | ✗ | ✓ | ✓ | ✓ | ✓ |
| **Ours** | ✓ | ✓ | ✓ | ✓ | ✓ |

Our work shares this motivation, but selects the prior by a different criterion: not the corruption process that produces it, but the coupling it induces. The identity coupling between our prior and the data is verified to be OT-optimal, which straightens trajectories without an OT solver. Tab. 1 summarizes these trade-offs.

## 3 Background

### 3.1 Flow matching

Let $p_0$ and $p_1$ denote continuous distributions over $\mathbb{R}^d$, serving as the base (noise) and data distributions, respectively. Flow matching (Lipman et al., 2022; Albergo & Vanden-Eijnden, 2022; Liu et al., 2022) learns a time-dependent flow $\phi_t : \mathbb{R}^d \to \mathbb{R}^d$ for $t \in [0, 1]$, which maps each sample $x_0 \sim p_0$ along a continuous path such that $\phi_0(x_0) = x_0$ and $\phi_1(x_0) \sim p_1$, *i.e.*, $\phi_1$ pushes $p_0$ forward to $p_1$. For brevity, we denote $x_t = \phi_t(x_0)$. The evolution of $x_t$ is governed by an Ordinary Differential Equation (ODE):

$$\frac{d}{dt}x_t = u_t(x_t), \tag{1}$$

where $u_t : \mathbb{R}^d \to \mathbb{R}^d$, for $t \in [0, 1]$, is the velocity field that generates the flow. The goal of flow matching is to learn a neural velocity field $v_\theta(x_t, t)$ that approximates $u_t(x_t)$.

Directly regressing onto $u_t$ is intractable, but an equivalent *conditional* objective can be optimized instead (Lipman et al., 2022). Under a linear interpolation path:

$$\phi_t(x_0|x_1) = x_t = tx_1 + (1 - t)x_0, \tag{2}$$

where $x_0 \sim p_0, x_1 \sim p_1$ are sampled independently. The corresponding conditional velocity is $u_t(x_t|x_1) = x_1 - x_0$, and the training objective reduces to:

$$\mathcal{L}(\theta) = \mathop{\mathbb{E}}_{t,\, x_0,\, x_1} \left[ \|v_\theta(x_t, t) - (x_1 - x_0)\|_2^2 \right], \tag{3}$$

where $t \sim \mathcal{U}[0, 1], x_0 \sim p_0, x_1 \sim p_1$. At inference, new samples are generated by solving the ODE in Eq. (1) from $t = 0$ to $t = 1$, initialized from $x_0 \sim p_0$.

While each conditional velocity $u_t(x_t|x_1) = x_1 - x_0$ defines a straight path between a specific pair $(x_0, x_1)$, the independent sampling of $x_0$ and $x_1$ in Eq. (3) means that these paths may cross one

another (Fig. 1a) (Geng et al., 2025; Pooladian et al., 2023). The learned velocity field $v_\theta$ must average over such conflicting directions, resulting in curved trajectories at inference time. Straightening these trajectories is desirable, as it reduces the number of ODE solver steps required for generation.

### 3.2 Optimal transport for straightening trajectories

Given a map $T\colon \mathbb{R}^d \to \mathbb{R}^d$, we denote the push-forward of $p_1$ by $T$ as $T_\#(p_1)$, *i.e.*, the distribution of $T(x)$ for $x \sim p_1$. The Monge optimal transport problem seeks a map that transports $p_1$ to $p_0$ with minimal squared Euclidean cost:

$$\inf_{T\colon T_\#(p_1)=p_0} \mathbb{E}_{x_1\sim p_1}\left[\|T(x_1)-x_1\|^2\right]. \tag{4}$$

In the context of flow matching, replacing the independent coupling in Eq. (3) with the OT-optimal pairing eliminates path crossings, since any crossing could be "uncrossed" to reduce the total cost, contradicting optimality. The resulting velocity field produces significantly straighter trajectories (Pooladian et al., 2023; Tong et al., 2024), reducing the number of ODE steps needed at inference.

However, computing the OT coupling between a high-dimensional data distribution and a standard Gaussian prior is intractable. While the optimal map is guaranteed to exist (Brenier, 1991), obtaining it requires $O(n^3)$ time and $O(n^2)$ memory in the number of samples (Kuhn, 1955; Cuturi, 2013), rendering it infeasible at modern data scales. As discussed in Secs. 1 and 2, existing methods attempt to solve or approximate this coupling at significant cost. Each of these keeps the prior *fixed* and searches for a coupling. We reverse the two roles, holding the coupling at the identity and *designing* the prior instead.

## 4 Method

Our approach replaces the standard Gaussian prior with a prior of our own design: the low-frequency projection of each training image, interpolated with Gaussian noise. Fig. 2 illustrates the full pipeline. We first describe the prior design (Sec. 4.1), then the training and inference procedure (Sec. 4.2).

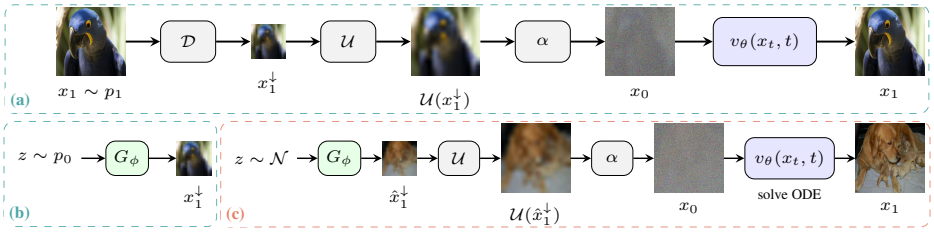


Figure 2: **Full Pipeline.** **(a)** During training, data samples are projected to low frequencies ($\mathcal{D}$), upsampled ($\mathcal{U}$), and perturbed with noise (Eq. (5)) to form $x_0$, paired with $x_1$ to train $v_\theta$. **(b)** $G_\phi$ is trained to map Gaussian noise to the low-frequency distribution. **(c)** At inference, $G_\phi$ produces a low-frequency sample , from which $x_0$ is constructed as in (a), and $v_\theta$ solves the ODE.

### 4.1 Designing the OT prior

Rather than solving the OT problem for a fixed prior, we treat the prior itself as a design choice. Given the data distribution $p_1$, we seek a transformation $T\colon \mathbb{R}^d \to \mathbb{R}^d$ that defines the prior as the distribution it induces, $\tilde{p}_0 = T_\#(p_1)$, with two design goals. (i) The identity coupling between data and prior should be OT-optimal, yielding straight, non-crossing trajectories for the flow model. Under the squared-cost OT objective of Eq. (4), the identity coupling pays the mean squared reconstruction error of $T$, so this goal favors $T$ with small reconstruction error on $p_1$. (ii) The prior should be expressive enough to support generation of $p_1$, yet structured enough to be sampled by a lightweight model at inference. The two goals are not equivalent, and (i) alone is not sufficient, as we show below.

Goal (i) alone does not determine the prior. The criterion is met by construction for any orthogonal projection (App. D), including degenerate ones that leave the original generation problem untouched. Among candidates that keep the same fraction of the input dimensionality, all equally optimal under (i), only those retaining low-frequency structure yield straight trajectories and good sample quality (Tab. C.3). Optimality selects a family of priors; it does not rank the members.

We argue for low-frequency projections . The power spectrum of natural images follows a $1/f^2$ decay (Field, 1987; Ruderman, 1994), so the low frequencies carry most of each image's $\ell_2$ energy and projecting onto them incurs a small reconstruction error, addressing goal (i). The resulting prior consists of low-resolution images capturing coarse structure, simple enough for a lightweight model to sample, addressing (ii).

The resulting design relates to a recurring principle in generative modeling. Cascaded diffusion (Ho et al., 2022a; Saharia et al., 2022), latent diffusion (Rombach et al., 2022), and pyramid flow (Jin et al., 2025) all decompose generation across resolutions or feature scales, generating coarse structure first and refining it with high-frequency detail. Our contribution is to make this decomposition explicit at the level of the flow's coupling. A cascade supplies the coarse signal as a conditioning input to the model, whose prior, and hence whose coupling to the data, is unchanged; we supply it as the starting sample $x_0$ itself. The flow model never has to invent coarse structure, its starting point already contains it, and what remains is to synthesize high-frequency detail on top of it.

The deterministic coupling defined by $T$ pairs each $x_1$ with a single fixed $T(x_1)$. While this is the exact coupling we sought, the resulting prior concentrates on a low-dimensional subspace, where nearby prior samples may correspond to distant data points. Such a prior is a difficult starting point for a velocity field, in line with the difficulty reported for score-based models on data concentrated near a low-dimensional manifold (Song & Ermon, 2019). Deterministic coupling also exposes the model to the same pair $(T(x_1), x_1)$ every epoch, risking overfitting. The optimality criterion constrains $T$, but says nothing about how the prior is spread around $T(x_1)$. Prior design leaves us this second degree of freedom, and we use it, interpolating the low-frequency representation and Gaussian noise:

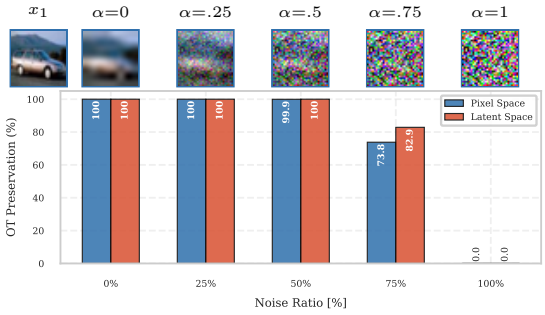


Figure 3: **OT Preservation Under Noise.** Fraction of pairs whose optimal assignment matches the identity coupling, in *pixel space* (CIFAR-10) and *latent space* (FFHQ). Above the bars: a sample at each noise level.

$$x_0 = (1-\alpha)\,T(x_1) + \alpha\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{5}$$

where $\alpha \in [0, 1]$ controls the noise ratio. $\alpha = 0$ recovers the deterministic coupling, and $\alpha = 1$ reduces to standard flow matching with random pairings.

The expected squared distance between two noisy prior samples decomposes as

$$\mathbb{E}\left[\|x_0^{(i)} - x_0^{(j)}\|^2\right] = (1-\alpha)^2\,\mathbb{E}\left[\|T(x_1^{(i)}) - T(x_1^{(j)})\|^2\right] + \alpha^2\,\mathbb{E}\left[\|\epsilon_i - \epsilon_j\|^2\right]. \tag{6}$$

The second term evaluates to $2\alpha^2 d$, growing linearly with the data dimensionality, so even moderate noise levels provide substantial separation between prior samples (Fig. B.3b).

This separation is only useful if it does not cost us the coupling, so we ask how much noise the coupling tolerates. At $\alpha = 0$, Hungarian assignment between 10K deterministic prior samples and their data samples recovers the identity coupling for all pairs. Under noise interpolation the assignment is unaffected as $\alpha$ grows, and at $\alpha = 0.5$ at most six of the 10K pairs depart from the identity coupling, which breaks down only beyond that point (Fig. 3). App. B.1 extends this analysis with a finer $\alpha$ sweep and the full derivation of Eq. (6).

The coupling constraint leaves $\alpha$ free, so generation quality selects it. FID improves steadily from $\alpha = 0$ to $\alpha = 0.5$, then degrades as noise overwhelms the low-frequency structure (Fig. B.3a), and we adopt $\alpha = 0.5$ for all our experiments. The $\alpha = 0$ endpoint is itself a meaningful baseline,

a pure coarse-to-fine pipeline in which $v_\theta$ refines a deterministic projection prior with no noise interpolation, so the FID gap to $\alpha = 0.5$ isolates what the second degree of freedom contributes beyond a coarse-to-fine substrate alone.

### 4.2 Pipeline and Training

We instantiate $T = \mathcal{U} \circ \mathcal{D}$, where the downsampling $\mathcal{D}\colon \mathbb{R}^d \to \mathbb{R}^{d'}$ produces $x_1^{\downarrow} = \mathcal{D}(x_1) \in \mathbb{R}^{d'}$, and $\mathcal{U} : \mathbb{R}^{d'} \to \mathbb{R}^d$ is its corresponding upsampling, so that $T$ maps $x_1$ to its low-frequency approximation in $\mathbb{R}^d$. During training (Fig. 2a), $T(x_1)$ is computed directly from each data sample and perturbed to form the starting point $x_0$, paired with $x_1$ for standard flow matching training of $v_\theta$ (see Alg. A.1). At inference (Fig. 2c), a lightweight generator $G_\phi$ maps Gaussian noise to $\mathbb{R}^{d'}$. In practice, $G_\phi$ is itself a flow matching model operating at a small fraction of the original dimensionality , trained separately (Fig. 2b) to produce $\hat{x}_1^{\downarrow}$, which is deterministically upsampled to $\mathcal{U}(\hat{x}_1^{\downarrow})$ and perturbed with noise. The flow model $v_\theta$, under the OT-identity coupling, synthesizes the remaining high-frequency details, with a coupling that costs nothing to obtain and no architectural changes.

## 5 Experiments

### 5.1 Experimental Setup

**Datasets.** We evaluate on three benchmarks of increasing complexity: CIFAR-10 (Krizhevsky et al., 2009) ($32 \times 32$), FFHQ (Karras et al., 2019) ($256 \times 256$), and ImageNet (Russakovsky et al., 2015) ($256 \times 256$). CIFAR-10 is trained in pixel space, following the architecture and training recipe of Tong *et al.* (Tong et al., 2024), while for the $256 \times 256$ datasets, we follow standard practice and train using the SiT architecture (Geng et al., 2025; Ma et al., 2024), and operate in the latent space of a pretrained autoencoder (Rombach et al., 2022), yielding $32 \times 32 \times 4$ latents. We downsample by $4\times$ per spatial dimension, producing an $8 \times 8$ low-frequency prior. $G_\phi$ is trained on a smaller variant of the same architecture family for each experiment, and operates on these representations. Full architectural details, training procedures, low-frequency projection specifications, and ablations are provided in Apps. A and C.

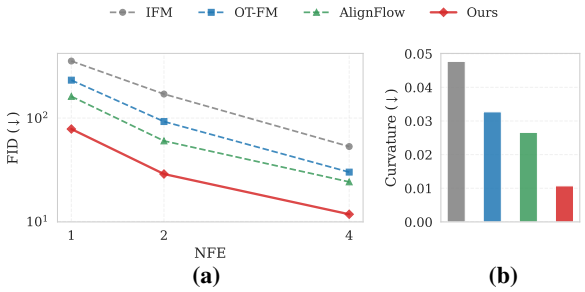


Figure 4: **CIFAR-10 Results.** Our method outperforms all baselines in FID **(a)** and trajectory curvature **(b)**.

**Baselines.** We compare against IFM (Lipman et al., 2022) (independent coupling), OT-FM (Tong et al., 2024) (minibatch OT), and AlignFlow (Kong et al., 2025) (semi-discrete OT). All baselines are single-network models, and our flow model $v_\theta$ uses the same architecture and parameter count as theirs (Tabs. A.1 and A.2). On ImageNet we evaluate with classifier-free guidance (CFG) (Ho & Salimans, 2022) and integrate our prior into MeanFlow (Geng et al., 2025), a one-step generation framework.

**Evaluation.** We report FID (Heusel et al., 2017) for generation quality (log scale in all plots), and curvature (Pooladian et al., 2023; Lee et al., 2023), which quantifies how close the learned trajectories are to straight lines. NFE counts evaluations of $v_\theta$ only, equally for all methods. The two-stage pipeline adds the lightweight generator $G_\phi$, which is reflowed (Liu et al., 2022) and run for a single step in all experiments, at $\approx 0.09\times$ the wall-clock time of one $v_\theta$ step (Fig. A.1). To bound how much of the remaining gap is attributable to the prior generator, App. B.2 replaces $\hat{x}_1^{\downarrow}$ with the true $x_1^{\downarrow}$, isolating the prior generator's contribution. We use $\alpha = 0.5$ Sec. 4.1, raised slightly at inference (App. A.4).

### 5.2 PIXEL SPACE (CIFAR-10)

We begin with CIFAR-10 in pixel space. Fig. 4a plots FID vs NFE for IFM (Lipman et al., 2022), OT-FM (Tong et al., 2024), AlignFlow (Kong et al., 2025), and our method. Our method achieves the best FID at every step count. Fig. 4b explains this improvement through trajectory curvature. Our method achieves substantially lower curvature than all baselines, $2.5\times$ below AlignFlow, the second-best method, and $4.5\times$ lower than IFM.

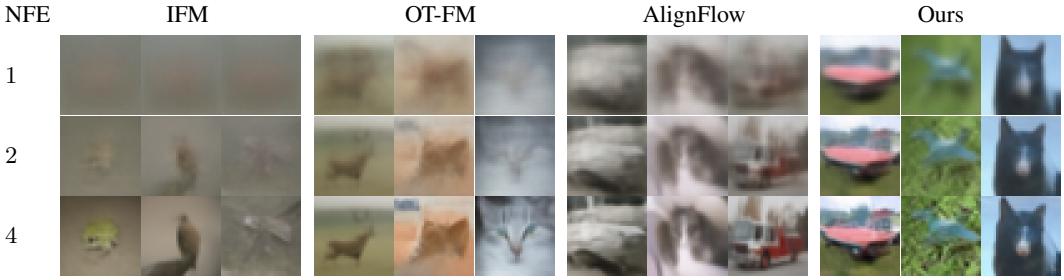


Figure 5: **CIFAR-10 few-step generation.** Samples from IFM (Lipman et al., 2022), OT-FM (Tong et al., 2024), AlignFlow (Kong et al., 2025) and our method at NFE $\in \{1, 2, 4\}$. Within a column the sample is fixed, so each column shows one image as the step budget grows. At a single step our method already produces recognizable content, while the baselines remain blurred.

Fig. 5 provides qualitative support. At 1 NFE the baselines produce blurred outputs, as a single step must absorb an entire curved trajectory. Our method begins from a prior that already carries the coarse layout of the target (Fig. 2c), so one step is enough to synthesize recognizable content. The visual margin narrows as the step budget grows, while the quantitative gap holds at every budget (Fig. 4a).

### 5.3 LATENT SPACE (FFHQ & IMAGENET)

Fig. 6a reports FID versus NFE on FFHQ $256 \times 256$. Our method achieves the best FID at every step count, with the largest margin at a single step, where its FID is less than half that of the strongest baseline. Fig. 6b explains this improvement through trajectory curvature. Our method achieves substantially lower curvature than all baselines, $2.1\times$ below OT-FM, the second-best method, and $2.4\times$ lower than IFM. The identity coupling is verified to hold in this space as well (Fig. 3).

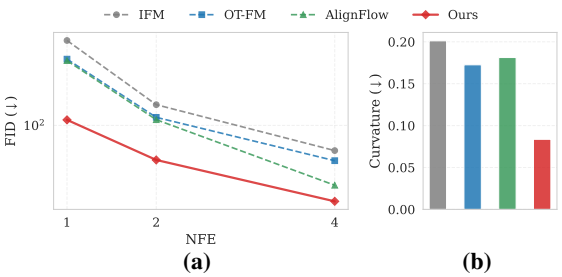


Figure 6: **FFHQ Results.** Our method outperforms all baselines in FID **(a)** and trajectory curvature **(b)**.

Fig. 7 compares IFM, OT-FM, and our method at matched NFE. At a single step the baselines produce blurred faces with incomplete structure, whereas our samples already exhibit coherent global layout and facial detail, and the ordering persists across the grid. The qualitative gap tracks the curvature measurement: a prior that already carries the coarse layout leaves less for $v_\theta$ to traverse.

**ImageNet and Classifier-Free Guidance.** To evaluate scaling to larger datasets, we train the same SiT backbone on class-conditional ImageNet $256 \times 256$. Due to compute constraints, we train on a shorter schedule than the one used by fully trained reference models (Tab. A.2), so absolute FID values are high. Both methods share the same architecture and training budget, so the comparison remains controlled. Tab. 2 reports FID and curvature at 1 NFE . Our method lowers FID from 263.82 to 145.36 with CFG and from 321.87 to 188.52 without CFG. Curvature is reduced by more than $2.2\times$ compared to IFM in both configurations, suggesting that the straighter trajectories observed on FFHQ extend to a large scale dataset such as ImageNet. The improvement holds with and without guidance, so the prior complements CFG rather than substituting for it.

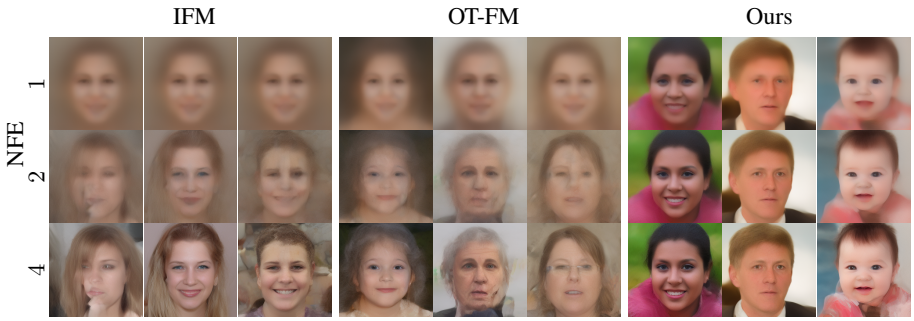


Figure 7: **FFHQ few-step generation.** Our method produces coherent faces already at low NFE, while the baselines remain blurry.

**One-Step Generation with MeanFlow.** MeanFlow (Geng et al., 2025) is a one-step generation framework that trains from scratch by modeling average rather than instantaneous velocity, requiring no distillation or curriculum learning. We integrate our prior into MeanFlow and evaluate on ImageNet $256 \times 256$. We use the pretrained checkpoint from a public PyTorch reproduction (Zhu, 2025) as our baseline, which achieves better FID than the original paper, and train our variant from scratch using the same codebase. As shown in Tab. 3, our prior reduces curvature by $2.3\times$ at a single step, with a small FID improvement alongside it, showing that straighter trajectories are available even in a framework already optimized for one-step generation. This integration illustrates how our prior complements training-objective changes such as consistency models (Song et al., 2023a; Yang et al., 2024) and distillation (Liu et al., 2022; seong Kim et al., 2025), which modify the learning objective while our method modifies the coupling the velocity field is trained on.

**Ablations.** We ablate the projection operator (App. C.1) and the downsampling factor (App. C.2), the first showing that low-frequency structure is necessary beyond OT-optimality alone.

Table 2: **ImageNet CFG Results.** FID and curvature at 1 NFE, with and without classifier-free guidance.

| | Method | FID (↓) | Curv. (↓) |
|---|---|---|---|
| w/o CFG | IFM | 321.87 | 0.233 |
| | Ours | **188.52** | **0.099** |
| w/ CFG | IFM | 263.82 | 0.239 |
| | Ours | **145.36** | **0.104** |

Table 3: **MeanFlow on ImageNet.** FID and trajectory curvature at 1 NFE, without classifier-free guidance.

| Method | FID (↓) | Curv. (↓) |
|---|---|---|
| MeanFlow | 59.12 | 0.221 |
| MeanFlow + Ours | **57.02** | **0.097** |

## 6 Limitations and Conclusion

**Prior generation.** Our empirical OT-identity coupling applies to the flow from $\tilde{p}_0$ to the data, but not to generating $\tilde{p}_0$ itself, as $G_\phi$ is trained under independent coupling, so the entire pipeline is not fully OT-coupled. The prior is nonetheless a cheaper target to model: $G_\phi$ operates at reduced resolution, so it is a small model trained on a fraction of the dimensions of the full generation problem. Training $G_\phi$ is an additional stage, separate from training the flow model. Since $G_\phi$ is decoupled from $v_\theta$, improving it should directly improve the full pipeline, and an oracle study bounds the headroom this leaves (see App. B.2).

**Domain specificity.** Our prior uses low-frequency projections, which empirically preserve the OT-identity coupling for natural images because their spectrum is concentrated at low frequencies (Ruderman, 1994), keeping projected samples well separated. The same holds in VAE latent space (Sec. 5.3). For other modalities, a projection must both preserve the identity coupling and keep the reconstruction error small, which depends on the data and should be examined per domain.

Overall, our results show that the prior in flow matching is a design variable, not a fixed input. Replacing the Gaussian prior with low-frequency projections, interpolated with Gaussian noise, yields an identity coupling verified to be OT-optimal, straighter trajectories, and improved few-step generation quality, while integrating naturally with latent-space models, classifier-free guidance, and one-step generation frameworks.

## AI USAGE DISCLOSURE

In this work, we used generative AI tools to implement methods and run experiments, and to assist in drafting and refining the proofs in App. D. We have not used generative AI tools to generate synthetic data, to formulate the method or its mathematical claims, or to interpret results, and the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to edit the manuscript for readability and to write and debug software code. We have reviewed all AI-assisted work. All AI-generated code was verified and tested for correctness by the authors, and all proofs were independently checked by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

## REFERENCES

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

Jacob Bamberger, Iolo Jones, Dennis Duncan, Michael M Bronstein, Pierre Vandergheynst, and Adam Gosztolai. Carr\'e du champ flow matching: better quality-generalisation tradeoff in generative models. *arXiv preprint arXiv:2510.05930*, 2025.

Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie S. Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=XH3ArccntI.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533, 2023.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

Giannis Daras, Mauricio Delbracio, Hossein Talebi, Alex Dimakis, and Peyman Milanfar. Soft diffusion: Score matching with general corruptions. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=W98rebBxlQ.

Aram Davtyan, Leello Tadesse Dadi, Volkan Cevher, and Paolo Favaro. Faster inference of flow-based generative models via improved data-noise coupling. In *The Thirteenth International Conference on Learning Representations*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

David J Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12):2379–2394, 1987.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=OlzB6LnXcS.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=uWj4s7rMnR.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022a.

Jonathan Ho, Tim Salimans, Alexey A. Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022b. URL https://openreview.net/forum?id=f3zNgKga_ep.

Emiel Hoogeboom and Tim Salimans. Blurring diffusion models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=OjDkC57x5sz.

Xingchang Huang, Corentin Salaun, Cristina Vasconcelos, Christian Theobalt, Cengiz Oztireli, and Gurprit Singh. Blue noise for diffusion models. In *ACM SIGGRAPH 2024 conference papers*, pp. 1–11, 2024.

Ka-Hei Hui, Chao Liu, Xiaohui Zeng, Chi-Wing Fu, and Arash Vahdat. Not-so-optimal transport flows for 3d point cloud generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=62Ff8LDAJZ.

Noam Issachar, Mohammad Salama, Raanan Fattal, and Sagie Benaim. Designing a conditional prior distribution for flow-based generative models. *arXiv preprint arXiv:2502.09611*, 2025.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong MU, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=66NzcRQuOq.

Bowen Jing, Bonnie Berger, and Tommi Jaakkola. Alphafold meets flow matching for generating protein ensembles. In *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*, 2023. URL https://openreview.net/forum?id=yQcebEgQfH.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Junwan Kim, Jiho Park, Seonghu Jeon, and Seungryong Kim. Better source, better flow: Learning condition-dependent source distribution for flow matching. *arXiv preprint arXiv:2602.05951*, 2026.

Lingkai Kong, Molei Tao, Yang Liu, Bryan Wang, Jinmiao Fu, Chien-Chih Wang, and Huidong Liu. Alignflow: Improving flow-based generative models with semi-discrete optimal transport. *arXiv preprint arXiv:2510.15038*, 2025.

Nikita Kornilov, Petr Mokrov, Alexander Gasnikov, and Aleksandr Korotin. Optimal flow matching: Learning straight trajectories in just one step. *Advances in Neural Information Processing Systems*, 37:104180–104204, 2024.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Yushi LAN, Shangchen Zhou, Zhaoyang Lyu, Fangzhou Hong, Shuai Yang, Bo Dai, Xingang Pan, and Chen Change Loy. Gaussiananything: Interactive point cloud flow matching for 3d generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=P4DbTSDQFu.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. Voicebox: Text-guided multilingual universal speech generation at scale. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 14005–14034. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/2d8911db9ecedf866015091b28946e15-Paper-Conference.pdf.

Sangyun Lee, Beomsu Kim, and Jong Chul Ye. Minimizing trajectory curvature of ode-based generative models. In *International Conference on Machine Learning*, pp. 18957–18973. PMLR, 2023.

Yexiong Lin, Yu Yao, and Tongliang Liu. Beyond optimal transport: Model-aligned coupling for flow matching, 2025. URL https://arxiv.org/abs/2505.23346.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. Matcha-tts: A fast tts architecture with conditional flow matching. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11341–11345. IEEE, 2024.

Alireza Mousavi-Hosseini, Stephen Y Zhang, Michal Klein, and Marco Cuturi. Flow matching with semidiscrete couplings. *arXiv preprint arXiv:2509.25519*, 2025a.

Alireza Mousavi-Hosseini, Stephen Y. Zhang, Michal Klein, and marco cuturi. On fitting flow models with large sinkhorn couplings. In *NeurIPS 2025 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2025b. URL https://openreview.net/forum?id=OhCgpYn1Pu.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky TQ Chen. Multisample flow matching: Straightening flows with minibatch couplings. *arXiv preprint arXiv:2304.14772*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Daniel L Ruderman. The statistics of natural images. *Network: computation in neural systems*, 5(4): 517, 1994.

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):211–252, 2015.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Shin seong Kim, Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Balanced conic rectified flow. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252. PMLR, 23–29 Jul 2023a. URL https://proceedings.mlr.press/v202/song23a.html.

Yuxuan Song, Jingjing Gong, Minkai Xu, Ziyao Cao, Yanyan Lan, Stefano Ermon, Hao Zhou, and Wei-Ying Ma. Equivariant flow matching with hybrid probability transport for 3d molecule generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b. URL https://openreview.net/forum?id=hHUZ5V9XFu.

Alexander Tong, Kilian FATRAS, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=CD9Snc73AW. Expert Certification.

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

Angxiao Yue, Anqi Dong, and Hongteng Xu. Oat-fm: Optimal acceleration transport for improved flow matching. *arXiv preprint arXiv:2509.24936*, 2025.

Yu Zeng, Charles Ochoa, Mingyuan Zhou, Vishal M Patel, Vitor Guizilini, and Rowan McAllister. Neuralremaster: Phase-preserving diffusion for structure-aligned generation. *arXiv preprint arXiv:2512.05106*, 2025.

Stephen Zhang, Alireza Mousavi-Hosseini, Michal Klein, and Marco Cuturi. On fitting flow models with large sinkhorn couplings. *arXiv preprint arXiv:2506.05526*, 2025.

Yu Zhu. Meanflow: Pytorch implementation. https://github.com/zhuyu-cs/MeanFlow, 2025. PyTorch implementation of Mean Flows for One-step Generative Modeling.

# APPENDICES

The appendices are accompanied by an interactive HTML page. To view it, extract the full supplementary archive and open `visualizations.html` in any modern browser. The page contains slider-based exploration of the noise parameter $\alpha$, extended generation grids across NFE budgets, and side-by-side comparisons for CIFAR-10, FFHQ and ImageNet.

We provide implementation details (App. A), empirical validation of the OT coupling (App. B), ablation studies (App. C), proofs for the orthogonal-projection case (App. D), and qualitative results (App. E).

## A IMPLEMENTATION DETAILS

### A.1 ARCHITECTURE AND TRAINING

Table A.1: **CIFAR-10 Architecture and Training.** Models share the UNet backbone, as prior work (Lipman et al., 2022; Tong et al., 2024), differing only in capacity (bottom block).

| | | |
|---|---|---|
| *Shared architecture* | | |
| Architecture | UNet | |
| Base channels | 128 | |
| Res. blocks / level | 2 | |
| Attn. heads | 4 (64 dim/hd) | |
| Dropout | 0.1 | |
| *Shared training* | | |
| Optimizer | Adam ($2\times10^{-4}$) | |
| LR warmup | 5k (linear) | |
| Batch size | 128 | |
| Total steps | 400k | |
| EMA decay | 0.9999 | |
| Grad. clip | 1.0 | |
| Augmentation | Horiz. flip | |
| | $v_\theta$ | $G_\phi$ |
| Input res. | $32^2\times3$ | $8^2\times3$ |
| Chan. mult. | [1,2,2,2] | [1,2] |
| Attn. res. | 16 | 4 |
| Parameters | $\sim$36 M | $\sim$10 M |

**CIFAR-10.** Tab. A.1 lists the architecture and training configuration for CIFAR-10 (Sec. 5.2). We train IFM (Lipman et al., 2022), OT-FM (Tong et al., 2024), AlignFlow (Kong et al., 2025), and our method from scratch, except that for AlignFlow we additionally use the pretrained assignment model from the official repository.

**FFHQ, ImageNet, and MeanFlow.** Tab. A.2 lists the configuration for the latent-space experiments. All methods are trained from scratch, except for MeanFlow (Geng et al., 2025) where we use the pretrained checkpoint from a public PyTorch reproduction (Zhu, 2025) as our baseline. Training is partial in all latent-space experiments. On ImageNet, the MeanFlow experiments use 80 epochs, matching the ablation budget in Tab. 1 of the original paper (Geng et al., 2025); the SiT experiments train for 240 epochs. On FFHQ, we use a comparable number of training steps. For the classifier-free guidance (CFG) setting, the same iteration count is reached with a larger batch size for efficiency, and we use a guidance scale of 5.0 at inference.

**Prior construction.** In all experiments, we set $\alpha = 0.5$ (Eq. (5)) for training $v_\theta$. The low-frequency prior is obtained by downsampling by a factor of 4 in each spatial dimension, applied to pixels for CIFAR-10 and to latents for FFHQ and ImageNet. Concretely, $\mathcal{D}$ is area downsampling and $\mathcal{U}$ is bicubic upsampling; App. C.1 compares this choice against alternative operators. Alg. A.1 provides a PyTorch-style pseudocode implementation of the training and inference procedures.

Table A.2: **256×256 Architecture and Training.** All setups share SiT (Ma et al., 2024) backbones for $v_\theta$ and $G_\phi$. MeanFlow adds dual time embedding $(r, t)$.

| *Shared training* | | | |
|---|---|---|---|
| Optimizer | AdamW ($1\times10^{-4}$, const.) | | |
| EMA decay | 0.9999 | | |
| | $v_\theta$ (B/4) | $G_\phi$ (S/2) | |
| Depth/hid./heads | 12/768/12 | 12/384/6 | |
| Patch / tokens | 4 / 64 | 2 / 16 | |
| Input | $32^2\times4$ | $8^2\times4$ | |
| Parameters | ∼130 M | ∼33 M | |
| | *FFHQ* | *ImageNet* | *MeanFlow* |
| Time emb. | Single ($t$) | | Dual ($r$,$t$) |
| Betas | (.9, .999) | | (.9, .95) |
| Time sampler | Uniform | | Logit-norm. |
| $r\neq t$ ratio | — | | 0.25 |
| Grad. clip | — | | 1.0 |
| Batch size | 256 | 1024 | 256 |
| Epochs | 1400 | 240 | 80 |
| Steps | 380k | 300k | 400k |

**Prior generator.** $G_\phi$ is a sampler for the low-frequency prior: the design fixes the distribution it draws from, not the mechanism used to draw from it, and we use a flow model for uniformity with $v_\theta$. To make single-step prior generation viable, $G_\phi$ is reflowed (Liu et al., 2022): we retrain it on pairs that it generates itself. Reflow adds 100k training steps, a quarter of $G_\phi$'s own training budget, and changes only its weights, leaving the architecture and the per-step cost unchanged.

**Inference overhead.** Our two-stage pipeline introduces a small fixed overhead from the lightweight generator $G_\phi$, which operates on $8\times8$ representations using a smaller variant of the main architecture (for SiT models, SiT-S/2 with 16 tokens, compared to SiT-B/4 with 64 tokens for $v_\theta$). Fig. A.1 reports the wall-clock time per step for both models across batch sizes. One $G_\phi$ step consistently takes $\approx 0.08$–$0.09\times$ the time of a $v_\theta$ step, and $G_\phi$ is run for a single step in all experiments, so the prior costs under $10\%$ of one $v_\theta$ step at any generation budget.

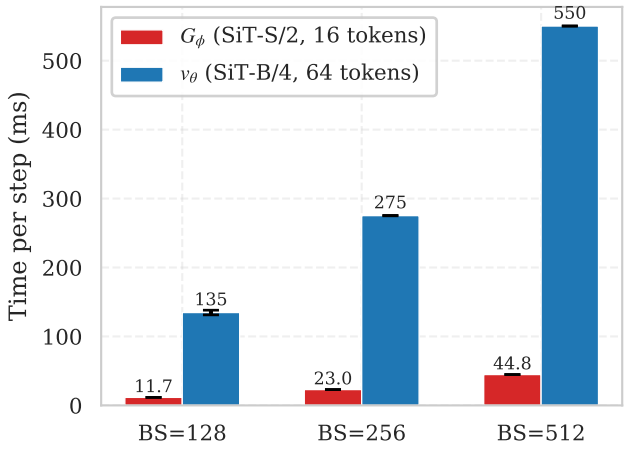


Figure A.1: $v_\theta$ **vs.** $G_\phi$ **Wall Time.** Inference time per step (ms) for the low-frequency generator $G_\phi$ (SiT-S/2, 16 tokens) and the main model $v_\theta$ (SiT-B/4, 64 tokens) across different batch sizes. One $G_\phi$ step is $\approx 0.08$–$0.09$ steps of $v_\theta$.

**Algorithm A.1 Training and Inference.**

**Hyperparameters:** noise ratio $\alpha = 0.5$, downsampling factor $k = 4$

▷ **Training: construct OT-structured pairs**

**function** BUILDCOUPLING($x_1$)
  $x_1^{\downarrow} \leftarrow$ `F.interpolate(`$x_1$`, scale_factor=1/`$k$`, mode="area")`
  $\mathcal{U}(x_1^{\downarrow}) \leftarrow$ `F.interpolate(`$x_1^{\downarrow}$`, scale_factor=`$k$`, mode="bicubic")`
  $\epsilon \leftarrow$ `torch.randn_like(`$\mathcal{U}(x_1^{\downarrow})$`)`
  $x_0 \leftarrow (1-\alpha)\,\mathcal{U}(x_1^{\downarrow}) + \alpha\,\epsilon$ ▷ Eq. (5)
  **return** $x_0$
**end function**

▷ **Training loop**

**for** $x_1 \sim p_1$ **do**
  $x_0 \leftarrow$ BUILDCOUPLING($x_1$)
  $t \leftarrow$ `torch.rand(1)`
  $x_t \leftarrow t\,x_1 + (1-t)\,x_0$ ▷ Eq. (2)
  $\mathcal{L} \leftarrow \|v_\theta(x_t, t) - (x_1 - x_0)\|^2$ ▷ Eq. (3)
  Update $\theta$
**end for**

▷ **Inference**

**function** GENERATE($G_\phi$, $v_\theta$)
  $z \leftarrow$ `torch.randn(...)` ▷ sample Gaussian noise
  $\hat{x}_1^{\downarrow} \leftarrow G_\phi(z)$ ▷ generate low-frequency sample
  $\mathcal{U}(\hat{x}_1^{\downarrow}) \leftarrow$ `F.interpolate(`$\hat{x}_1^{\downarrow}$`, scale_factor=`$k$`, mode="bicubic")`
  $\epsilon \leftarrow$ `torch.randn_like(`$\mathcal{U}(\hat{x}_1^{\downarrow})$`)`
  $\tilde{x}_0 \leftarrow (1-\alpha)\,\mathcal{U}(\hat{x}_1^{\downarrow}) + \alpha\,\epsilon$
  $x_1 \leftarrow$ ODESOLVE($v_\theta, \tilde{x}_0, t{=}0 \to 1$)
  **return** $x_1$
**end function**

### A.2 COMPUTE RESOURCES

Experiments were run on a heterogeneous mix of NVIDIA A5000, A6000, and H100 GPUs. For a uniform reporting unit, we report the wall-clock training time of a single run on one A5000-equivalent GPU (training only, excluding evaluation), separately for the main flow model $v_\theta$ and the lightweight low-frequency generator $G_\phi$, including its reflow stage: CIFAR-10: $v_\theta \approx 28$ h, $G_\phi \approx 1.9$ h; FFHQ-256: $v_\theta \approx 25$ h, $G_\phi \approx 3.1$ h; MeanFlow ImageNet-256: $v_\theta \approx 260$ h, $G_\phi \approx 30$ h. $G_\phi$ thus adds $\sim$7-13% on top of the main-model training cost.

### A.3 CURVATURE

To quantify trajectory straightness, we report the curvature metric used in prior work (Pooladian et al., 2023; Lee et al., 2023), which measures the deviation of the learned trajectory from a straight path,

$$\mathcal{C} = \mathbb{E}_t \left[ \|(x_1 - x_0) - v_\theta(x_t, t)\|^2 \right], \tag{A.1}$$

where $v_\theta(x_t, t)$ is the learned velocity and $x_1 - x_0$ is the constant velocity of the straight-line path. A perfectly straight trajectory has zero curvature. The expectation is taken over each method's own source distribution, so $x_0$ is a Gaussian sample for the baselines and a low-frequency prior sample for ours, and $\mathcal{C}$ measures the trajectory of $v_\theta$ alone. Following Lee *et al.* (Lee et al., 2023), we evaluate on 10,000 generated samples using 128 Euler steps.

### A.4 Inference-time noise calibration

At inference, the starting point is constructed from $\mathcal{U}(\hat{x}_1^{\downarrow})$ by adding Gaussian noise at level $\alpha$ (Fig. 2c), according to Eq. (5). During training, this is set to $\alpha = 0.5$. At inference, we use a slightly higher noise level $\alpha^* = \alpha + \varepsilon$, with $\varepsilon > 0$, which is most beneficial in the few-step regime. This is consistent with observations in score-based generative modeling (Song & Ermon, 2019), where slightly expanding the support of a concentrated distribution improves the conditioning of the learned vector field and reduces artifacts at inference time.

Across all benchmarks, the optimal inference noise follows $\alpha^* = \alpha_{\text{train}} + \varepsilon$, where $\varepsilon > 0$ is a small excess that decreases with the number of function evaluations. In the few-step regime (up to 4 NFE), we find $\varepsilon \approx 0.05$ in both pixel space (CIFAR-10) and latent space (FFHQ, ImageNet), consistent across datasets and resolutions. For MeanFlow, the effect is milder ($\varepsilon \approx 0.01$), consistent with its single-step formulation being less sensitive to the source distribution.

## B Empirical validation

### B.1 OT preservation under noise

The identity coupling induced by our prior is verified empirically, by exact discrete OT assignment, rather than proved (Fig. 3). This section gives the protocol behind that measurement, a finer sweep over the noise ratio $\alpha$, the FID-vs-$\alpha$ curve from which $\alpha$ is chosen, and the pairwise-distance analysis behind Eq. (6).

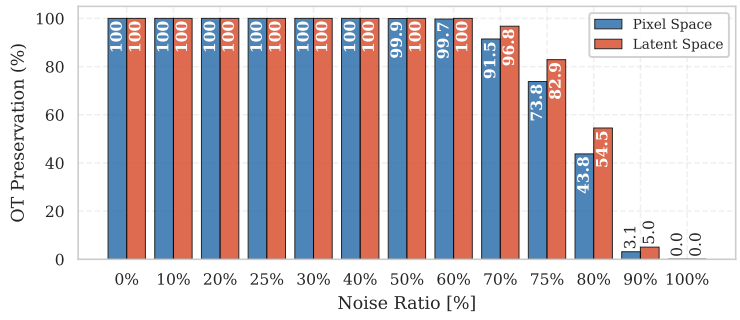


Figure B.2: **Comparison of noise ratio vs. OT preservation.** An elaborated version of Fig. 3 showing the OT preservation metric as a function of the noise ratio $\alpha$ in Eq. (5).

**OT coupling preservation.** In Fig. 3, we measure the fraction of training pairs for which the exact discrete OT assignment (computed using the POT library) coincides with the identity coupling, as the noise level $\alpha$ in Eq. (5) increases. The experiment uses 10K CIFAR-10 samples in pixel space and 10K FFHQ samples in the latent representation of a pretrained autoencoder (Rombach et al., 2022; Ma et al., 2024). Fig. B.2 extends this analysis to finer noise-ratio increments, confirming that the identity coupling is fully preserved up to $\alpha \approx 0.5$, so the optimality criterion pins the projection and leaves $\alpha$ free across the range we use.

**FID vs. noise ratio.** Fig. B.3a reports FID on CIFAR-10 as a function of $\alpha$, evaluated at 2 NFE after an early training stage of 40K steps (10% of full training). Quality improves from $\alpha = 0$ to $\alpha = 0.5$, then degrades as noise overwhelms the low-frequency structure. The $\alpha = 0$ endpoint is a purely deterministic coarse-to-fine pipeline, so the gap between the two endpoints isolates the contribution of the noise interpolation over a projection prior alone.

**Pairwise distance analysis.** Fig. B.3b (discussed in Sec. 4.1) shows the mean pairwise $\ell_2$ distance between prior samples as a function of $\alpha$ (proportional to noise ratio). At $\alpha = 0$, the mean pairwise distance between low-frequency projections on CIFAR-10 is $\approx 5.4 \times 10^{-3}$ (normalized by $d = 3{,}072$), roughly 11% below the original-image distance of $\approx 6.1 \times 10^{-3}$. This concentration is consistent with the $1/f^2$ spectral decay of natural images (Field, 1987), which places the majority of image energy in the lowest frequencies. As $\alpha$ increases, the noise term dominates and the mean pairwise

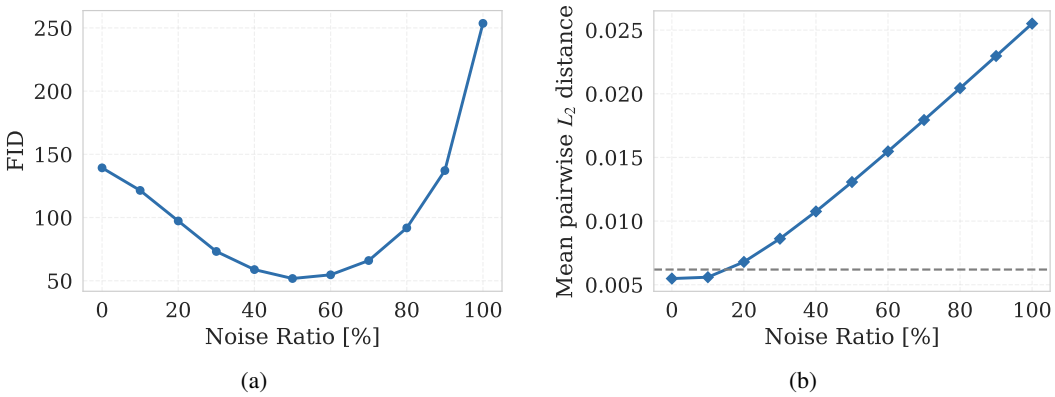


Figure B.3: **Effect of noise ratio $\alpha$ on generation quality and prior spread. (a)** FID on CIFAR-10 (2 NFE) vs. noise ratio. **(b)** Pairwise $\ell_2$ distance on CIFAR-10 vs. noise ratio; *dashed*: original images.

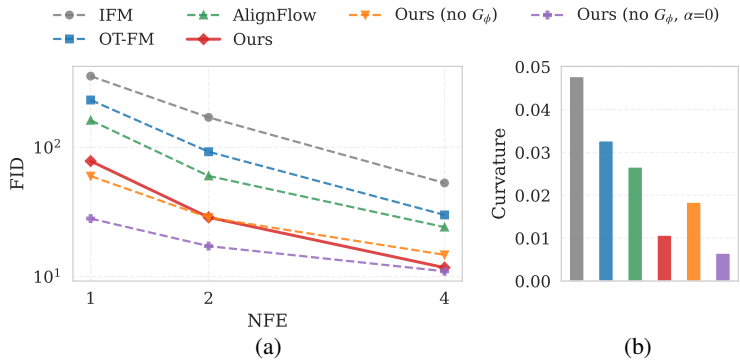


Figure B.4: **Oracle Experiment on CIFAR-10.** FID (↓) vs. NFE and trajectory curvature (↓) for baselines, our full pipeline, and two oracle conditions that bypass $G_\phi$ using true training-image projections. Panels (a) and (b) show FID and curvature, respectively. *Ours (no $G_\phi$)*: noised oracle, $\alpha = 0.5$. *Ours (no $G_\phi$, $\alpha$=0)*: deterministic oracle, $\alpha = 0$.

distance grows steadily. The deterministic prior is therefore packed more tightly than the data it is transported to, and the $2\alpha^2 d$ term of Eq. (6) is what supplies the separation, at a noise level the coupling still tolerates (Fig. B.2).

In the main text (Eq. (6)), we state that the cross term vanishes and the expected distance decomposes into a projection term and a noise term. We now show this step by step. For two noisy prior samples $x_0^{(i)}$ and $x_0^{(j)}$ constructed via Eq. (5), we expand the squared norm,

$$\|x_0^{(i)} - x_0^{(j)}\|^2 = \|(1-\alpha)\big(T(x^{(i)}) - T(x^{(j)})\big) + \alpha\big(\epsilon^{(i)} - \epsilon^{(j)}\big)\|^2. \tag{B.2}$$

Since the noise terms are independent of the data and of each other, the cross term vanishes in expectation,

$$\mathbb{E}\big[\|x_0^{(i)} - x_0^{(j)}\|^2\big] = (1-\alpha)^2\,\mathbb{E}\big[\|T(x^{(i)}) - T(x^{(j)})\|^2\big] + 2\alpha^2 d, \tag{B.3}$$

where $d$ is the data dimensionality and the last term follows from $\epsilon^{(i)} - \epsilon^{(j)} \sim \mathcal{N}(0, 2I_d)$. The first term captures the pairwise spread in the projected subspace, scaled down by $(1-\alpha)^2$, while the second grows linearly with $d$, explaining the steady increase visible in Fig. B.3b.

### B.2 Oracle Experiment

To isolate the effect of $G_\phi$ on generation quality, we evaluate two oracle conditions alongside our full pipeline on CIFAR-10. Both bypass $G_\phi$ entirely, constructing starting points directly from training images.

**Oracle conditions.** *Noised oracle* uses our standard training procedure ($\alpha = 0.5$) but replaces the low-frequency image that $G_\phi$ would generate, $\hat{x}_1^{\downarrow}$, with the true low-frequency image of a training sample, $x_1^{\downarrow}$, from which the starting point is built as in Eq. (5). *Deterministic oracle* trains with $\alpha = 0$ (fully deterministic coupling, $x_0 = T(x_1)$) and starts inference from the exact projections $T(x_i)$. Fig. B.4 reports FID and curvature for all conditions.

**Results.** The deterministic oracle achieves lower FID than the full pipeline across all step counts, confirming that improving $G_\phi$ can further close the gap to oracle-level performance, and attains the lowest curvature overall, as expected from a fully deterministic coupling, in which every starting point is paired with a single fixed data sample.

The full pipeline attains lower curvature than the noised oracle, and trails it in FID only at a single step. The two share the same $v_\theta$ and differ only in where inference starts, so the priors produced by $G_\phi$ stand in for exact low-frequency images without costing trajectory straightness.

# C Ablation studies

## C.1 Projection variants

This section accompanies the claim in Sec. 4.1 that OT-optimality of the identity coupling does not by itself select a usable prior. We compare four candidate transformations $T$, each retaining $1/16$ of the input dimensionality. Three are strict orthogonal projections, for which the identity coupling $x_i \mapsto T(x_i)$ is OT-optimal by construction (Props. D.1 and D.2). The area/bicubic operator we adopt is not strictly orthogonal, and for it the identity coupling is verified empirically below. Either way, only the low-frequency members yield a usable prior in practice.

**Operator family.** We instantiate four operators on CIFAR-10 images ($32 \times 32 \times 3$). *Area/bicubic* sets $T = \mathcal{U} \circ \mathcal{D}$ with $\mathcal{D}$ area downsampling $32 \times 32 \to 8 \times 8$ and $\mathcal{U}$ bicubic upsampling (the operator used throughout the paper). *Fourier truncation* keeps the lowest-frequency $8 \times 8$ block of DFT coefficients per channel and zeros the rest. *Random pixel mask* applies a fixed binary mask, sampled once and held constant across all images, that retains a random subset of $1/16$ of the pixels. *Random patch mask* applies a fixed binary mask over an $8 \times 8$ grid of $2 \times 2$ patches, retaining $1/16$ of the patches. A diagonal $0/1$ mask is symmetric and idempotent, so both masking operators are valid orthogonal projections. The first two preserve low-frequency structure. The last two do not.

**Setup.** For each operator we train $v_\theta$ at the full CIFAR-10 capacity (Tab. A.1) for 150,000 steps under the noisy prior of Eq. (5), once at $\alpha = 0$ and once at $\alpha = 0.5$. We evaluate in the oracle condition of App. B.2, *i.e.*, starting points are drawn from real test images projected through $T$ rather than from the prior generator $G_\phi$. This isolates the effect of the projection operator on $v_\theta$ from the quality of the lightweight prior model. Reconstruction MSE $\mathbb{E}\,\|T(x) - x\|^2$ on the training set is the squared transport cost paid by the identity coupling under the OT objective of Eq. (4). We additionally verify OT-identity preservation empirically by computing the exact discrete OT assignment on 10K pairs (Hungarian algorithm) for each operator at both noise levels. Preservation exceeds 99.9% in all eight settings, so the identity coupling is optimal for all but a negligible fraction of pairs under every operator we test, both deterministically and under the $\alpha = 0.5$ noise interpolation.

**Results.** Tab. C.3 shows that OT-optimality does not by itself select a usable prior. Reconstruction MSE splits the operators into two groups, with the low-frequency family (0.009 and 0.010) roughly $30\times$ lower than the masking family (0.271 for both). This reflects the spectral structure of natural images, under which the lowest $8 \times 8$ frequency block captures most of the $\ell_2$ energy, while $1/16$ of the pixels or $1/16$ of $2 \times 2$ patches do not.

At $\alpha = 0$, curvature already separates the two families: the low-frequency operators stay in the $10^{-6}$ to $10^{-5}$ range, while the masking operators sit two to three orders of magnitude higher ($\sim 1.4 \times 10^{-3}$). FID does not. Random pixel masking reaches 117.90, ahead of Fourier truncation at 119.29, and only random patch masking (144.95) falls clearly behind *Area/bicubic* (102.05). At $\alpha = 0.5$, the regime we adopt, the two families separate on FID as well. The low-frequency operators improve to $\approx 88$ while the masking operators degrade to 226.72 and 205.57, worse than they were deterministically. Noise helps only the operators that retain low-frequency structure, so the FID we rely on is reached

Table C.3: **Projection Variants on CIFAR-10.** Four projections (strictly orthogonal except for area/bicubic resampling), each retaining $1/16$ of the input dimensionality, trained for 150,000 steps under Eq. (5) at $\alpha = 0$ and $\alpha = 0.5$. FID and curvature are evaluated at 1 NFE in the oracle condition (real test images projected through $T$, $G_\phi$ bypassed). Reconstruction MSE is $\mathbb{E}\,\|T(x) - x\|^2$ on the training set. Hungarian assignment finds OT-identity preservation above $99.9\%$ for every operator at both noise levels. The shaded row is the operator used throughout the paper. Best per column in bold.

| | | $\alpha = 0$ | | $\alpha = 0.5$ | |
|---|---|---|---|---|---|
| Operator | MSE (↓) | FID (↓) | Curv. (↓) | FID (↓) | Curv. (↓) |
| Area/bicubic | **0.009** | **102.05** | **$5.0\times10^{-6}$** | **87.82** | **0.013** |
| Fourier truncation | 0.010 | 119.29 | $1.2\times10^{-5}$ | 87.99 | **0.013** |
| Random pixel mask | 0.271 | 117.90 | $1.5\times10^{-3}$ | 226.72 | 0.022 |
| Random patch mask | 0.271 | 144.95 | $1.4\times10^{-3}$ | 205.57 | 0.023 |

by the low-frequency projection and the noise interpolation together, and by no other combination in the table.

OT-optimality is therefore not sufficient. Hungarian preservation exceeds $99.9\%$ for every operator and every noise level we test, so the OT property does not distinguish them. What does is whether the projection retains low-frequency structure. Within the low-frequency family, *Area/bicubic* edges out Fourier truncation on reconstruction MSE, matches it at $\alpha = 0.5$, and is the operator we adopt.

### C.2 Downsampling Factor

The downsampling factor controls the dimensionality of the low-frequency prior and directly affects the trade-off between the two design goals of Sec. 4.1. Too little downsampling retains high-frequency content, leaving $G_\phi$ with a generation task comparable in difficulty to the original problem and negating the benefit of a lightweight first stage. Too much downsampling discards structure needed for the flow model to produce high-quality samples. Fig. C.5 compares $4\times$ and $8\times$ spatial downsampling on CIFAR-10. A $4\times$ reduction ($8 \times 8$ prior from $32 \times 32$ images) yields consistently better FID across step counts, while $8\times$ downsampling degrades quality as the prior loses too much structural detail to serve as an effective starting point. We adopt $4\times$ in all experiments.

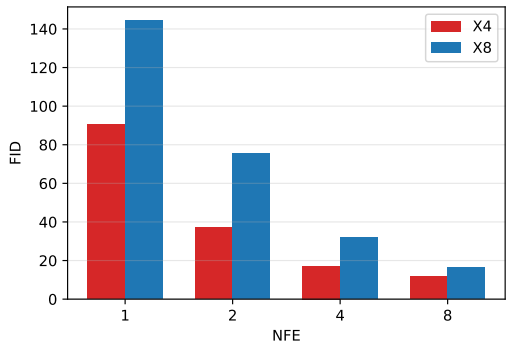


Figure C.5: **Downsample factor ablation on CIFAR-10.** FID as a function of NFE for two spatial downsample factors ($\times4$) and ($\times8$). A lower downsample factor consistently yields better FID across all NFE budgets

## D Proofs and Formal Details

**Proposition D.1** (Orthogonal projections preserve self-proximity)**.** *Let $T\colon \mathbb{R}^d \to \mathbb{R}^d$ be an orthogonal projection,* i.e.*, $T^2 = T$ (idempotent) and $T^\top = T$ (symmetric). Then for any $x, x' \in \mathbb{R}^d$,*

$$\|T(x') - x\|^2 = \|T(x) - x\|^2 + \|T(x') - T(x)\|^2. \tag{D.4}$$

*In particular,* $\|T(x) - x\| \leq \|T(x') - x\|$*, with equality iff* $T(x') = T(x)$.

*Proof.* Denote $\delta = x' - x$. By linearity of $T$,

$$T(x') - x = T(x) + T(\delta) - x = T(\delta) + (T(x) - x). \tag{D.5}$$

The first term $T(\delta) \in \mathrm{Im}(T)$. For the second, note that by idempotency

$$T(T(x) - x) = T^2(x) - T(x) = 0. \tag{D.6}$$

so $(T(x) - x) \in \ker(T)$. For an orthogonal projection, $\mathrm{Im}(T)$ and $\ker(T)$ are orthogonal subspaces, so by the Pythagorean theorem,

$$\|T(x') - x\|^2 = \|T(\delta)\|^2 + \|Tx - x\|^2 = \|T(x') - T(x)\|^2 + \|T(x) - x\|^2. \tag{D.7}$$

□ □

**Proposition D.2** (Discrete OT optimality of projections)**.** *Let* $P\colon \mathbb{R}^d \to \mathbb{R}^d$ *be an orthogonal projection, and let* $\{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ *be a finite point set with distinct projections,* i.e.*,* $Px_i \neq Px_j$ *for* $i \neq j$*. Then the identity assignment* $\sigma = \mathrm{id}$*, mapping* $x_i \mapsto Px_i$*, uniquely minimizes the total transport cost*

$$C(\sigma) = \sum_{i=1}^{n} \|Px_{\sigma(i)} - x_i\|^2 \tag{D.8}$$

*over all permutations* $\sigma$ *of* $\{1, \ldots, n\}$*, and hence over all couplings of the two empirical measures.*

*Proof.* By Prop. D.1, for any $i$ and any $j \neq i$,

$$\|Px_j - x_i\|^2 = \|Px_i - x_i\|^2 + \|Px_j - Px_i\|^2.$$

Since the projections are distinct, $\|Px_j - Px_i\|^2 > 0$, so

$$\|Px_i - x_i\|^2 < \|Px_j - x_i\|^2. \tag{D.9}$$

Every point $x_i$ is therefore strictly closer to its own projection than to the projection of any other point.

Now consider any permutation $\sigma \neq \mathrm{id}$. There exists at least one index $i$ with $\sigma(i) \neq i$. For that index, $\|Px_{\sigma(i)} - x_i\|^2 > \|Px_i - x_i\|^2$. Since $\|Px_{\sigma(j)} - x_j\|^2 \geq \|Px_j - x_j\|^2$ for all $j$, with strict inequality for at least one, we have $C(\sigma) > C(\mathrm{id})$. □

**Remark** (Low-pass filtering as orthogonal projection)**.** *The discrete Fourier transform (DFT) is a unitary transformation. Let* $F$ *denote the DFT matrix and let* $P_S$ *be the diagonal matrix that retains a subset* $S$ *of Fourier coefficients and zeros out the rest.* $P_S$ *is an orthogonal projection in the frequency domain. Since conjugation by a unitary preserves orthogonal projections,* $P = F^{-1}P_S F$ *is an orthogonal projection in pixel space. In particular, choosing* $S$ *to be the lowest-frequency coefficients yields low-pass filtering, so Props. D.1 and D.2 apply directly.*

*In our pipeline (Sec. 4.2), the transformation* $T = \mathcal{U} \circ \mathcal{D}$ *applies downsampling* $\mathcal{D}$ *followed by upsampling* $\mathcal{U}$*. When* $\mathcal{D}$ *is area downsampling, which averages each block, and* $\mathcal{U}$ *replicates each value back over its block,* $T$ *is exactly an orthogonal projection onto the subspace of block-constant images, and both propositions above apply. In practice, we find it better to use area interpolation for* $\mathcal{D}$ *and bicubic interpolation for* $\mathcal{U}$*, which is not strictly orthogonal, with the comparison against the alternatives given in App. C.1, and the OT structure is preserved empirically as confirmed in Fig. 3.*

## E Qualitative results

We visualize how the noise level $\alpha$ in Eq. (5) affects the structure of the prior $x_0$, in pixel space (CIFAR-10, App. E.1) and latent space (FFHQ, App. E.2). As $\alpha$ increases, the low-frequency structure of the prior is progressively obscured, consistent with the quantitative analysis in Fig. 3 and App. B.1. We then show class-conditional ImageNet samples in the few-step regime (App. E.3).

### E.1 PIXEL-SPACE PRIOR (CIFAR-10)

Fig. E.6 shows representative CIFAR-10 samples across all noise levels. Each column displays the original image $x_1$, its low-frequency projection $T(x_1)$, and the noised starting point $x_0$ for $\alpha \in \{0, 0.1, \dots 1\}$. At low noise levels, the dominant color, spatial layout, and coarse object structure of the original image remain clearly visible, since the projection discards little of the image. Around $\alpha \approx 0.5$, noise begins to dominate and the visual correspondence to the original gradually fades.

### E.2 LATENT-SPACE PRIOR (FFHQ)

Fig. E.7 shows the same visualization for FFHQ in the latent space of a pretrained autoencoder (Rombach et al., 2022). Each column displays the original image, its encoded latent, the downsampled latent $x_1^{\downarrow} = \mathcal{D}(x_1)$, and $x_0$ at increasing $\alpha$ values. Unlike pixel space, the structural correspondence is harder to assess visually in latent representations, particularly at moderate noise levels. The coupling is measured rather than inspected, and Fig. 3 reports that the identity coupling is preserved up to $\alpha = 0.5$ in this space.

### E.3 IMAGENET FEW-STEP GENERATION

Fig. E.8 shows class-conditional ImageNet $256\times256$ samples in the few-step regime, comparing the Gaussian-prior baseline against our low-frequency prior at $1$, $2$, and $4$ NFE. Because both methods are driven by the same noise within each column, the visual differences isolate the effect of the prior. At a single NFE, our samples already exhibit coherent global structure and color, whereas the baseline remains largely noise; the gap narrows as the NFE budget grows.

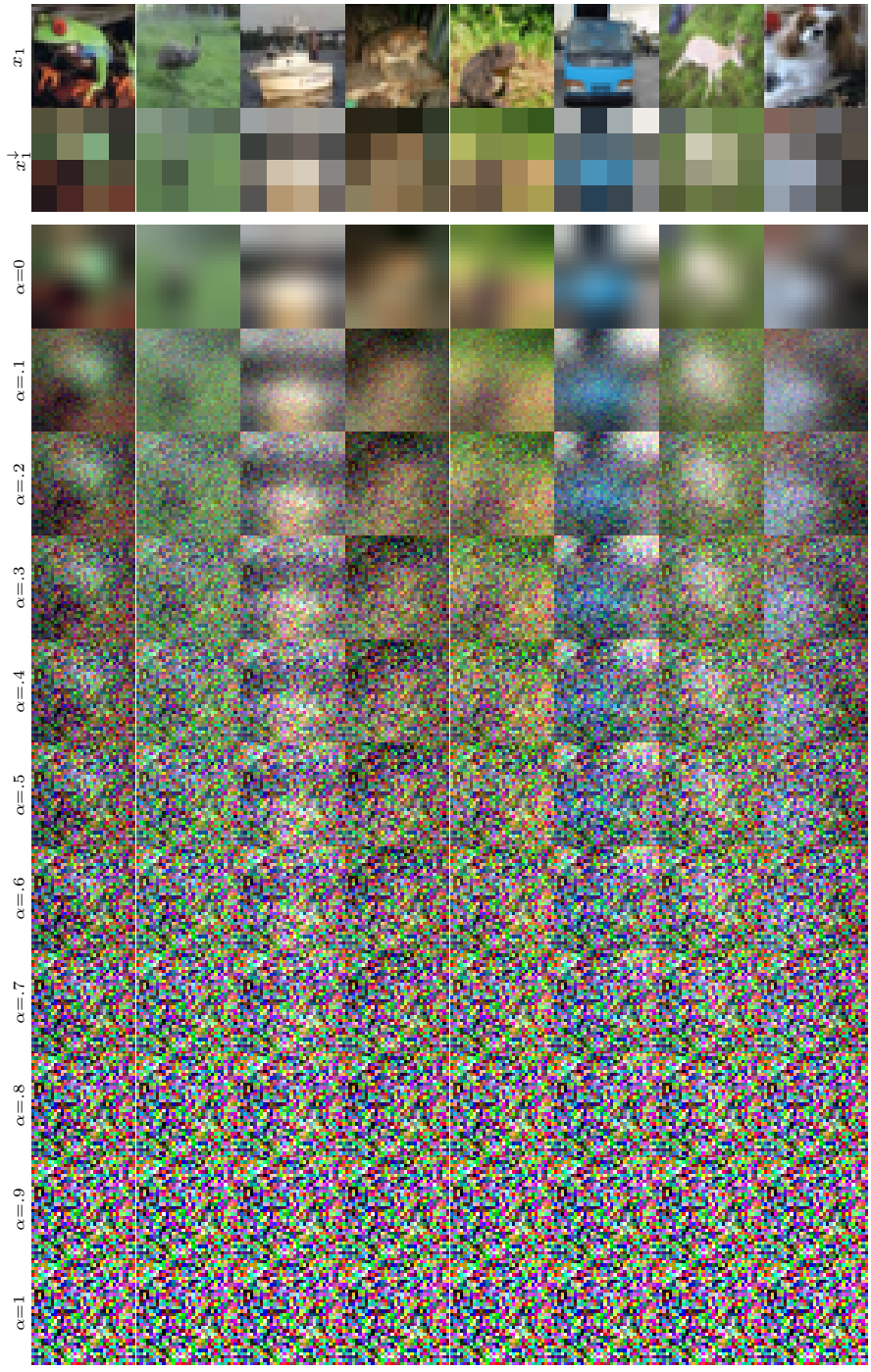


Figure E.6: **Prior Construction on CIFAR-10.** Each column shows one training image $x_1$ (top), its low-frequency representation $x_1^{\downarrow} = \mathcal{D}(x_1)$, and the noised starting point $x_0 = (1-\alpha)\,\mathcal{U}(x_1^{\downarrow}) + \alpha\,\epsilon$ (Eq. (5)) at increasing noise levels, ranging from the pure upsampled projection ($\alpha$=0) through the operating point ($\alpha$=0.5), where coarse structure remains visible, to pure Gaussian noise ($\alpha$=1).

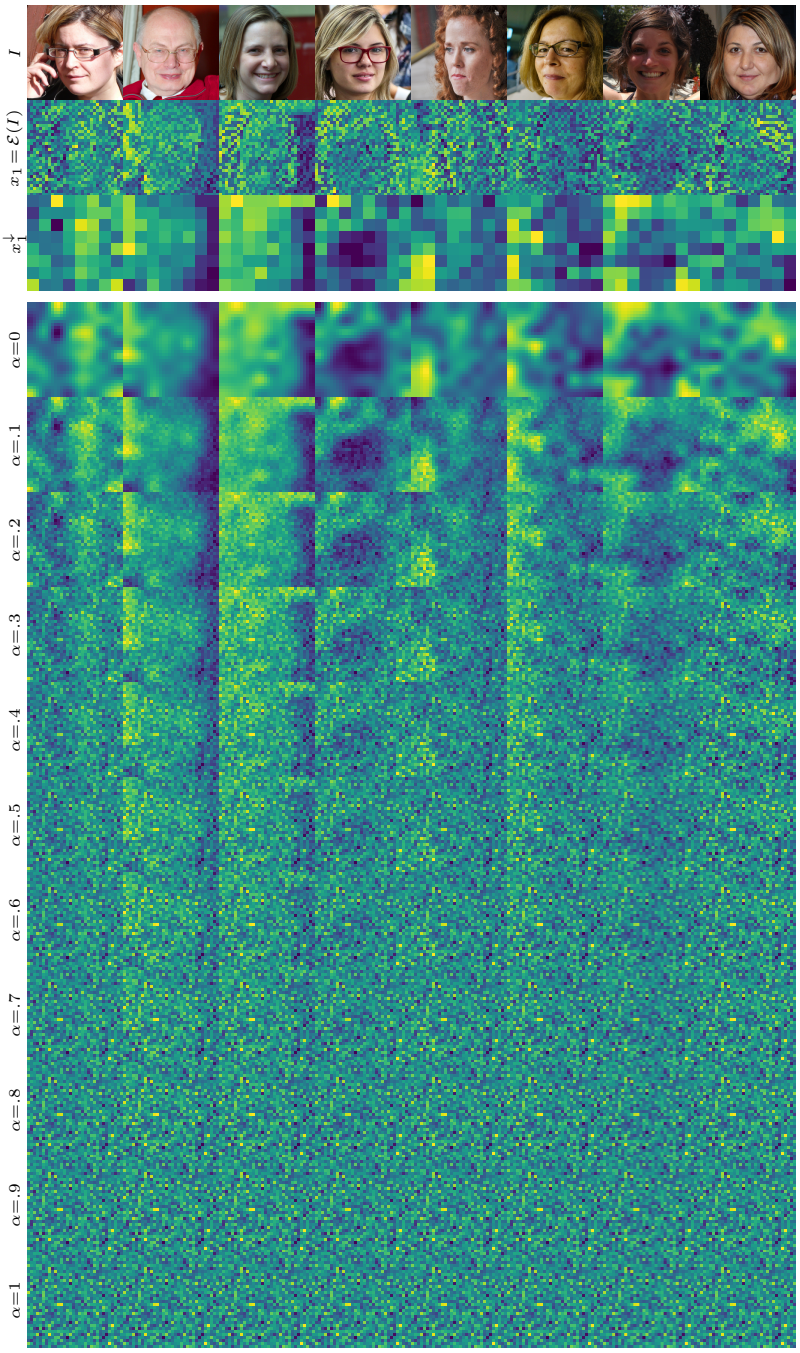


Figure E.7: **Prior Construction on FFHQ (Latent Space).** The first row shows original images $x_1$ for reference. The second row visualizes the corresponding latent encoding, and the third row its low-frequency representation $x_1^{\downarrow} = \mathcal{D}(x_1)$. Subsequent rows show the noised starting point $x_0 = (1-\alpha)\,\mathcal{U}(x_1^{\downarrow}) + \alpha\,\epsilon$ (Eq. (5)) at increasing noise levels, ranging from the pure upsampled projection ($\alpha$=0) through the operating point ($\alpha$=0.5) to pure noise ($\alpha$=1). All operations from the third row onward are applied in the latent space of a pretrained autoencoder (Rombach et al., 2022).

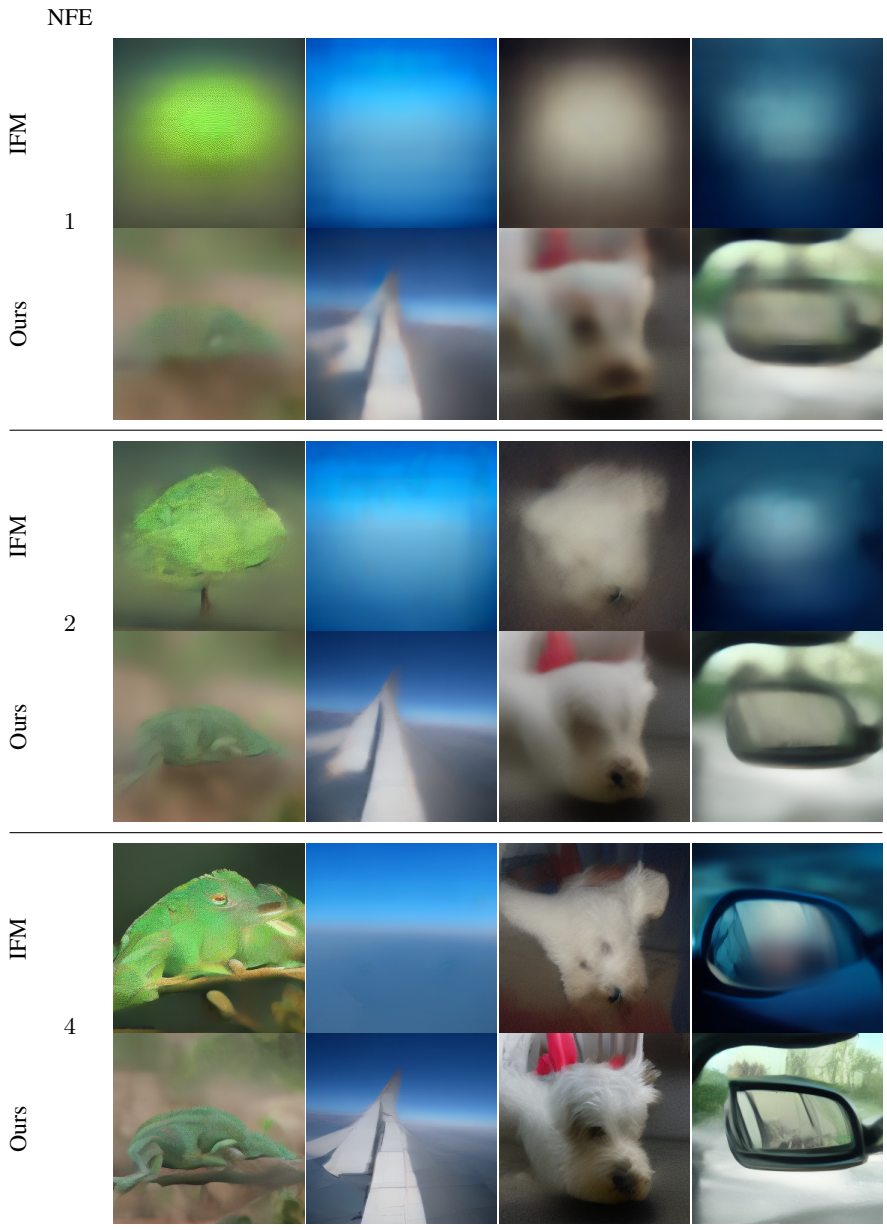


Figure E.8: **ImageNet few-step generation.** Class-conditional 256×256 samples at 1, 2, and 4 NFE, of IFM (Lipman et al., 2022) and ours. Each column shows a different class: *African chameleon*, *wing*, *West Highland white terrier*, and *car mirror* (left to right).