GLOBAL TRANSPORT COUPLINGS FOR CLASSIFIER-FREE GUIDED FLOWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Optimal-transport couplings have been shown to reduce training variance in unconditional flow models, but their role in conditional generation remains unclear. A natural approach constructs separate couplings for each condition, but this is impractical for large or continuous conditioning spaces found in modern image foundation models. We introduce **Global Transport** (**GT**), a global class-agnostic optimal-transport coupling, computed without class labels. **GT** can associate different conditions with different regions of the source noise, and consequently worsens performance without guidance. However, when combined with classifier-free guidance (CFG), **GT** consistently improves generation across domains, model scales, and sampling budgets. This reversal suggests that couplings for conditional flows should be evaluated both empirically and theoretically under the guided flow used at inference, rather than on unguided generation. We evaluate **GT** over both discrete class and continuous text conditioned image generation across model scales, and investigate how coupling choice alters guided trajectories. These results identify coupling design in the guided flow setting as a simple training time axis to improve performance without modifying existing architectures, samplers, or guidance mechanisms.

1 INTRODUCTION

Flow matching (Lipman et al., 2023; Albergo & Vanden-Eijnden, 2023; Peluchetti, 2023) and diffusion models (Ho et al., 2020; Song et al., 2021) have emerged as powerful paradigms for generative modeling, achieving state-of-the-art generation results across a wide range of domains including images (Rombach et al., 2022; Esser et al., 2024; Ma et al., 2024), videos (Ho et al., 2022; Blattmann et al., 2023; Polyak et al., 2024), and biological data (Morehead et al., 2026; Li et al., 2026). In the simplest construction, source and data samples are paired independently. A widely adopted strategy for improving generation quality is by incorporating optimal transport (OT) *couplings* between noise samples and data points (Pooladian et al., 2023; Tong et al., 2024a). OT pairs noise and data to minimize transport cost, reducing the training variance and enabling more efficient numerical integration and sampling efficiency of flow matching models (Lu et al., 2025; Berthelot et al., 2026; Malnick et al., 2026). These benefits suggest that coupling design could also improve conditional generation. It is less obvious, however, what an appropriate coupling should be when every data sample is associated with a class, text prompt, or other condition.

The natural extension of this paradigm is to match source and data separately for each condition. Such class-conditional OT respects the common source distribution used to sample each class, but requires sufficiently large batches of examples sharing a condition (Chemseddine et al., 2025; Kerrigan et al., 2024; Mousavi-Hosseini et al., 2025; Kong et al., 2026; Cheng & Schwing, 2025). This becomes impractical when there are many classes and is infeasible when conditions, such as text-embeddings, are continuous. As a result, class-conditional OT has achieved limited adoption in the conditional generative setting.

There is, however, an additional confounder in the case of conditional generation. Conditional flow models are almost ubiquitously sampled with classifier-free guidance (CFG) (Ho & Salimans, 2022), which combines conditional and unconditional velocities and is used to steer towards a condition and away from the unconditional density. Class-conditional OT has been shown to improve both the unguided conditional flow, and the guided flow (Cheng & Schwing, 2025). The natural extension of this paradigm to the *conditional* generative setting is via *class-conditional* OT, where one seeks an optimal coupling for each *class*. However, this procedure necessitates the computation of couplings

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.

for class-conditioned batches, which becomes impractical for large conditioning spaces frequently found in modern generative models, and infeasible in continuous conditioning spaces used in text-to-image models (Esser et al., 2024). As a result, class-conditional OT has achieved limited adoption in the conditional generative setting, and yields modest gains for the added computational cost.

CFG trains a flow model to perform both conditional and unconditional generation, dropping the conditioning signal on some fraction of the data points; at inference time, guided samples are drawn by linearly composing the two velocity fields, modulated by a guidance scale. As a result, CFG guided samples follow trajectories defined jointly by both conditional and unconditional velocity fields, rather than by the conditional model alone. Class conditional OT improves both the unconditional and conditional fields separately. A natural question then arises:

> Can a coupling that considers both fields together produce a better *guided* generator?

We answer in the affirmative with **Global Transport** (**GT**), which computes a single *global* OT assignment across source and target samples without considering their conditions. **GT** is straightforward to apply to both class labels and text conditioning, and requires no change to the model, sampler, or CFG rule. However, **GT** comes with a tradeoff: Although the coupling preserves the overall source marginal, the conditional source distribution do not need to be consistent with the unconditional source distribution. Following from this mismatch, **GT** performs worse in unguided conditional sampling in our experiments. Under CFG, however, the ordering reverses with **GT** improving guided generation across domains, model scales, and sampling budgets.

To investigate the cause of this reversal, we study the coupling choice from the perspective of the *guided* flow. We find **GT** reduces the *prediction gap* (Wang et al., 2025a; Cai et al., 2026), the difference between the conditional and unconditional flows. This reduces a pattern we term the *yo-yo effect*, where independent flows exhibit pronounced contraction, re-expansion, and overshooting along guided trajectories with increasing guidance strength. Our contributions are:

- We propose **GT**, a practical, low-overhead coupling strategy for conditional generation that requires no change to architectures, samplers, or guidance mechanisms.
- We demonstrate that unguided conditional performance does not always correlate with guided performance, and analyze how coupling choice affects guided trajectories and the prediction gap.
- We show that **GT** consistently yields competitive generation performance on class-conditioned image generation (e.g. FID 1.91 on ImageNet-256), text-conditioned image generation, and single-cell gene expression data, across model scales and sampling budgets, and outperforms both independent and class-conditional couplings.

2 Background and Preliminaries

Flow Matching. Flow matching (FM) (Lipman et al., 2023; Albergo et al., 2023; Liu et al., 2023; Peluchetti, 2023) allows continuous time transport between distributions: the algorithm trains a velocity field that evolves samples from a source distribution p_0 to a target data distribution $p_1 = p_{\text{data}}$, where p_0 is typically chosen to be an easy-to-sample distribution such as $\mathcal{N}(0, I_d)$.

Posit a continuous-time transport, i.e. a *probability path*, $(p_t)_{t\in[0,1]}$ with the prescribed boundary conditions (p_0, p_1). Any such p_t is generated by a marginal velocity field u_t such that $\partial_t p_t + \nabla \cdot (p_t u_t) = 0$. FM seeks to approximate u_t with a model u_t^θ by least-squares regression, minimizing $\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t\sim U[0,1],\, x_t\sim p_t}\left[\left\|u_t^\theta(x_t) - u_t(x_t)\right\|_2^2\right]$. This objective is intractable, as in practice u_t is unavailable in closed form. Conditional FM (Tong et al., 2024a; Albergo et al., 2023) instead introduces a coupling $\pi \in \Pi(p_0, p_1)$ and a conditional probability path $p_t(\cdot \mid x_0, x_1)$ generated by tractable conditional velocity field $u_t(\cdot \mid x_0, x_1)$. Taking expectations yields the CFM objective, which trains a neural velocity field u^θ by regression onto the closed-form conditional velocity $u_t(x_t \mid x_0, x_1)$:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t,\,(x_0,x_1)\sim\pi,\, x_t\sim p_t(\cdot|x_0,x_1)}\left[\left\|u_t^\theta(x_t) - u_t(x_t \mid x_0, x_1)\right\|_2^2\right]. \tag{1}$$

Once trained, samples from p_1 can be generated by first drawing from the source $x_0 \sim p_0$, then numerically integrating the ODE $\dot{x}_t = u_t(x_t)$.

Classifier-free Guidance. Conditional generation, present in many applications such as text-to-image or text-to-video synthesis, requires sampling given a prompt $c \in \mathcal{C}$, achieved by learning a prompt-conditioned velocity field $u_t(x_t \mid c)$ via the analogue of equation 1,

$$\mathcal{L}^c_{\mathrm{CFM}}(\theta) = \mathbb{E}_{t,\,(x_0,x_1,c)\sim\pi,\,x_t\sim p_t(\cdot|x_0,x_1)}\left[\left\|u^\theta_t(x_t,c) - u_t(x_t \mid x_0,x_1,c)\right\|_2^2\right]. \tag{2}$$

Classifier-free guidance (CFG) (Ho & Salimans, 2022) improves sample quality and prompt alignment by sampling from a linear combination of the marginal and conditional fields,

$$\widetilde{u}^w_t(x_t \mid c) = u_t(x_t) + w\, g(x_t,t,c), \quad g(x_t,t,c) = u_t(x_t \mid c) - u_t(x_t) \tag{3}$$

where $w \geq 1$ and g are the guidance weight and vector respectively. In practice, a single model is trained across all conditions, with c replaced by a null token $\varnothing$ with some probability, so that $u^\theta_t(x_t,\varnothing)$ estimates the marginal field. Appendix E.1 reviews the theory underlying CFG.

Optimal Transport. Optimal transport conditional flow matching (OT-CFM) (Tong et al., 2024a; Pooladian et al., 2023; Mousavi-Hosseini et al., 2025; Kong et al., 2026) replaces the independent coupling $\pi = p_0 \otimes p_1$ with approximations of the Euclidean OT coupling π_{OT}, defined as the solution to

$$\pi_{\mathrm{OT}} = \mathrm{argmin}_{\pi\in\Pi(p_0,p_1)}\mathbb{E}_{(x_0,x_1)\sim\pi}\|x_0 - x_1\|^2. \tag{4}$$

The resulting trajectories are straighter in practice and therefore easier to integrate in the few-step sampling regime, shown to achieve modest but consistent gains in the unconditional setting.

Class-conditional Optimal Transport. Class-conditional generation involves sampling from data distribution $p_1(x \mid c)$ conditioned on selected class $c \in \mathcal{C}$. A naïve adaptation of minibatch OT-CFM to the class-conditional setting involves solving $|\mathcal{C}|$ independent OT problems, one for each class $c \in \mathcal{C}$. However, this quickly becomes infeasible when the number of classes $|\mathcal{C}|$ becomes even moderately large (for instance, ImageNet contains 1000 classes), and is not possible in settings where the condition itself is a continuous latent state, as is the case for text-to-image generation for example. (Kerrigan et al., 2024; Chemseddine et al., 2025; Cheng & Schwing, 2025) proposed to approximate this by minimizing

$$\pi_{\mathrm{C^2OT}} = \mathrm{argmin}_{\pi\in\Pi(p_0,p_1)}\mathbb{E}_{((x_0,c_0),(x_1,c_1))\sim\pi}\left[\|x_0 - x_1\|^2 + \beta\|c_0 - c_1\|^2\right] \tag{5}$$

where β is adjusted to fit the scale and $c_{0,1}$ denotes the class associated with $x_{0,1}$. This allows the use of much smaller batches in practice, but with unknown degradation, and thus far has not been widely adopted in conditional generation settings, where independent couplings are almost universal.

3 Global Transport

We next introduce **Global Transport** (**GT**) a practical and low overhead application of optimal transport which improves performance in the class-conditioned setting. **GT** computes an approximate optimal transport coupling without encouraging or constraining assignments to preserve class labels. The method does not change the architecture nor inference algorithm and adds only a training-time assignment step. We first describe the overall objective, then two approximations either using (1) a mini-batch approximation (Fatras et al., 2021; Tong et al., 2024a; Pooladian et al., 2023) or (2) a semi-discrete approximation (Mousavi-Hosseini et al., 2025; Kong et al., 2026).

Algorithm 1 **GT** training step

$x_0 \sim p_0,\ (x_1,c) \sim p_1,\ t \sim \mathcal{U}(0,1)$
$\tilde{c} \leftarrow \varnothing$ w.p. p_{uncond}, else $\tilde{c} \leftarrow c$
$\hat{x}_0 \leftarrow \pi_{\mathtt{GT}}(x_0); \quad x_t \leftarrow t x_1 + (1-t)\,\hat{x}_0$
$\theta \leftarrow \mathrm{Update}\big(\theta, \nabla_\theta \|u^\theta_t(x_t,\tilde{c}) - (x_1 - \hat{x}_0)\|^2\big)$

General objective. Given noise samples $x_0 \sim p_0$ and class conditional data samples $(x_1,c) \sim p_1$, for the optimal transport coupling π_{OT} as defined in equation 4, and a standard linear flow matching path $x_t = tx_1 + (1-t)x_0$ where the noise is rearranged to match the closest data points, ignoring the labels c. This results in the conditional flow matching objective:

$$\mathcal{L}_{\mathrm{GT}}(\theta) = \mathbb{E}_{t\sim U[0,1],(x_0,x_1)\sim\pi_{\mathrm{OT}}}\|u^\theta_t(x_t,\tilde{c}) - (x_1 - x_0)\|^2 \tag{6}$$

where $\tilde{c}$ is set to $\varnothing$ with probability p_{uncond} and c otherwise. The full algorithm appears in Algorithm 1.

Minibatch-OT Implementation. We compute the assignment using the Hungarian algorithm with squared Euclidean cost in the model input space. For a batch of size B, the cost matrix requires $O(B^2)$

memory and the exact assignment $O(B^3)$ time. This is generally dwarfed by network evaluation time on large systems and is negligible (usually $< 2\%$ overhead) for modern workflows. However, it can be expensive for large batch sizes, if this setting is desired a regularized (Tong et al., 2024b; Cuturi, 2013; Zhang et al., 2025), or semi-discrete approach may be more appropriate.

Semi-discrete OT Implementation. Building off of recent work on Flow matching with semi-discrete optimal transport couplings, we can also use an approximate semi-discrete OT implementation. This approach first performs an expensive preprocessing step to optimize a semi-dual potential function over all discrete datapoints, which can then be used to calculate couplings with any member of the continuous noise measure.

The conditional source distribution mismatch. Ignoring conditions while constructing the coupling changes the distribution of (x_0, x_1) pairs the conditional model learns from. To illustrate this, we consider the full batch, idealized setting using the true OT coupling π_{OT}, with associated transport map $T : p_0 \to p_1$. Draw $x_0 \sim p_0$ and set $x_1 = T(x_0)$ with condition C. Then the source points that are paired with points with condition c do not need to be distributed like p_0 as different conditions may be paired with source points from different regions of the prior. In other words, although $x_0 \sim p_0$, the conditional distribution $x_0 \mid C = c$ need not equal p_0. This tells us that under the unguided setting ($w = 1$), the standard class conditioned inference procedure starting with samples from p_0 will not necessarily land at $p_1|c$, and is only guaranteed to if the initial samples are drawn from the intractable distribution $x_0 \mid C = c$.

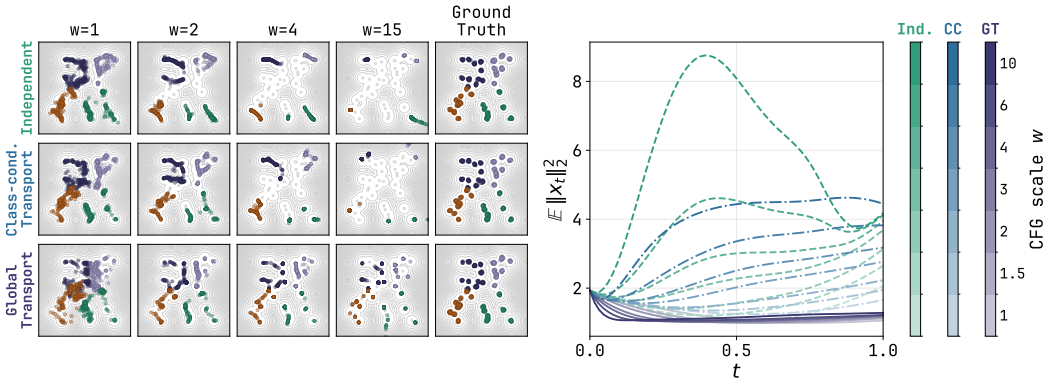

Figure 1: **2-dimensional depiction of the "yo-yo" effect.** This effect pushes samples away from the center at high guidance scales for the independent and class-conditional OT couplings, degrading generative performance. Global Transport fixes this and is stable under increasing guidance w.

4 Coupling Choice With Classifier-Free Guidance

The conditional source distribution mismatch described above suggests that **GT** may be a poor choice for unguided conditional generation ($w = 1$), which we observe in a 2-dimensional 40-Gaussian mixture model example with four classes (Figures 1 and 6 with details in Appendix A). However, the result changes under guidance ($w > 1$). This guidance dependent effect motivates studying the unconditional and conditional fields together as they are used during inference, rather than judging a coupling by its unguided conditional generation performance.

4.1 Geometry of Guided Trajectories

Let $X_t^{(w)}$ denote a sample generated by integrating $\dot{X}_t^{(w)} = \widetilde{u}_t^w(X_t^{(w)} \mid c)$ from $X_0^{(w)} \sim p_0$. To track its radial position over time, we measure

$$Y_w(t) = \mathbb{E}\|X_t^{(w)}\|_2. \tag{7}$$

Unlike the distribution of linear interpolants used in training, $Y_w(t)$ describes the radial position of trajectories produced by the guided sampler.

In our GMM example, trajectories trained with independent pairing contract before expanding toward their class modes (Figure 2). As w increases, the re-expansion becomes more pronounced and can

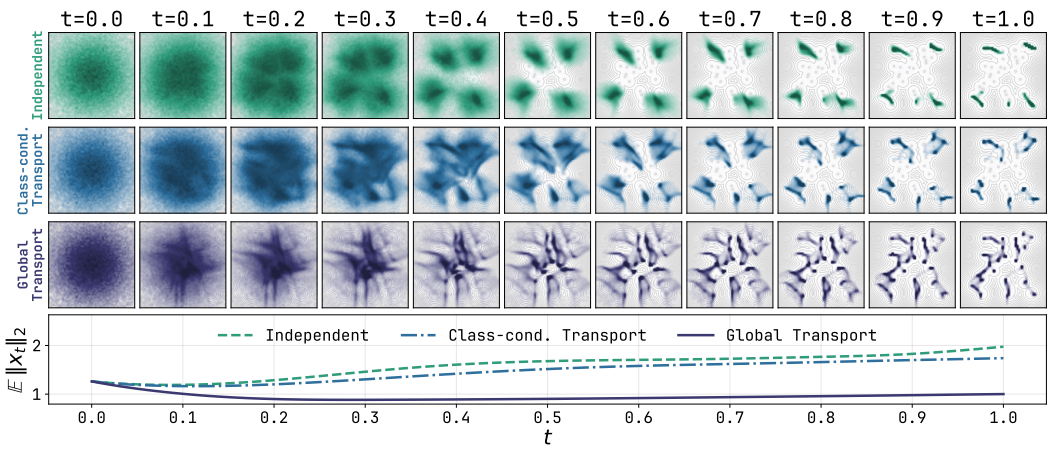

Figure 2: Density over time for Independent, Class conditional, and **GT** couplings. Independent "yoyo"s, contracting towards the class origin then expanding outwards, whereas **GT** is stable.

overshoot the target regions (Figure 1). We refer to this contraction followed by re-expansion as the *yo-yo effect*. Class-conditional OT reduces some aspects of this behavior, while **GT** exhibits less pronounced contraction and overshoot over the guidance scales we test (Figures 1 and 3).

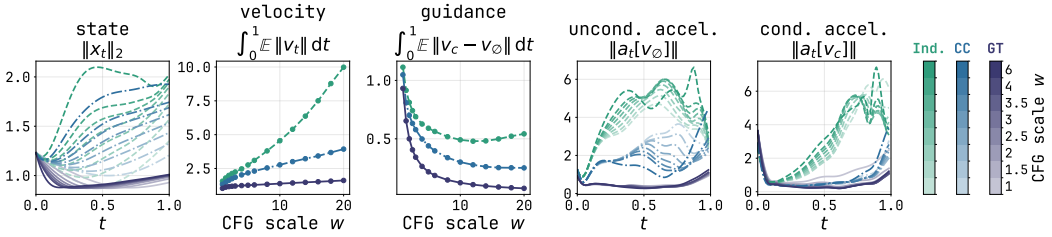

Figure 3: **Guided dynamics on the 40-GMM** across guidance scales w. From left: state norm along guided trajectories; integrated guided velocity norm; integrated prediction gap $\|v_c - v_\varnothing\|$; unconditional and conditional material acceleration $\|a_t[v]\|$.

4.2 Coupling controls the prediction gap

Recall from equation 3 the guidance vector $g(x_t, t, c) = u_t(x_t|c) - u_t(x_t)$ measures the gap between unconditional and conditional velocity fields. Following Wang et al. (2025a) we call $\|g(x, t, c)\|^2$ the *prediction gap* and ask how the training coupling constrains its size.

Let π be the joint distribution of matched source points x_0, data points x_1, and conditions C. Its (x_0, x_1) marginal couples p_0 and p_1. Define the training interpolant $x_t = (1-t)x_0 + tx_1$ and its coupling cost $\mathrm{Cost}(\pi) = \mathbb{E}_\pi\|x_1 - x_0\|^2$.

Proposition 4.1 (Transport cost bounds the on-path prediction gap)**.** *Suppose $u_t(x)$ and $u_t(x \mid c)$ are the population squared-error minimizers for the common flow-matching target $x_1 - x_0$ under π. Then*

$$\mathbb{E}_{t,\pi}\left[\|g(X_t, t, C)\|^2\right] \le \mathrm{Cost}(\pi) - \mathcal{W}_2^2(p_0, p_1). \tag{8}$$

The proof in Appendix B.1 first expresses the expected prediction gap as the difference between the optimal unconditional and conditional flow-matching regression errors. It then bounds the integrated unconditional error by the excess quadratic transport cost. A lower-cost coupling can therefore give a tighter upper bound on the average prediction gap along its training interpolants. We note that this bound holds strictly along the linear training interpolants and does not govern the distribution of

states visited during guided inference ($w > 1$), which we verify empirically. We also note that at the optimal transport limit, the bound contracts to zero because deterministic transport renders class conditioning redundant on the training support. In practice, finite mini-batch OT retains a non-zero prediction gap while regularizing the velocity fields. Further, we can show a hierarchy of coupling costs between different methods compared in this paper, which we do in the following proposition.

Proposition 4.2. *For all β the coupling costs satisfy*

$$\mathrm{Cost}(\pi^{\star}_{\texttt{GT}}) \leq \mathrm{Cost}(\pi^{\star}_{C^2OT(\beta)}) \leq \mathrm{Cost}(\pi_{\mathrm{ind}}). \tag{9}$$

with the lower equality achieved at $\beta = 0$.

which establishes that the **GT** algorithm has a tighter upperbound on the prediction gap than C^2OT (for $\beta > 0$) which has a tighter upper bound on the prediction gap than independent couplings.

While these results establish tighter upperbounds on the prediction gap, this does not establish causality between lower prediction gaps and improved performance. We use the bound to motivate the prediction-gap measurements in Figure 3, not as an explanation of generation performance on its own.

5 Experiments

We now demonstrate the effectiveness of **GT** on conditional generation in several different settings including image generation and single-cell experiments. We first show that **GT** improves the performance of classifier-free guided class-conditional generation on ImageNet (Deng et al., 2009) (256×256) when training flow matching models and distilling them into flow maps. Then we move to the case of continuous conditioned models in the text-to-image setting, before finally demonstrating the impact on single-cell data generation guided by cell type labels across multiple single-cell datasets. We denote the minibatch version of **GT**-MB as **GT**, and the semi-discrete OT version of **GT** as **GT** + SD-OT.

Table 1: **Left:** FID↓ comparison against baselines. †denotes results obtained with guidance interval (Kynkäänniemi et al., 2024). **Right:** Curated class-conditional samples from SiT-XL/2 + **GT** on ImageNet-256.

Method	NFE	CFG	Params	FID ↓
GANs / Normalizing Flows / Autoregressive models				
StyleGAN-XL (Sauer et al., 2022)	1	✗	166M	2.30
STARFlow (Gu et al., 2025)	1	✗	1.4B	2.40
VAR-$d30$ (Tian et al., 2024)	20	✓	2B	1.92
MAR-H/2 (Li et al., 2024)	64	✓	943M	1.55
Diffusion / Flow models				
ADM (Dhariwal & Nichol, 2021)	500	✓	554M	10.94
LDM (Rombach et al., 2022)	500	✓	400M	3.60
RIN (Jabri et al., 2023)	1000	✗	410M	3.42
SimDiff (Hoogeboom et al., 2023)	1024	✓	2B	2.77
U-ViT-H/2 (Bao et al., 2023)	100	✓	501M	2.29
DiT-XL/2 (Peebles & Xie, 2023)	500	✓	675M	2.27
SiT-XL/2† (Ma et al., 2024)	500	✓	675M	2.06
SiT-XL/2 + **GT** † **(ours)**	500	✓	675M	**1.91**



5.1 Class- and text- conditioned image generation on ImageNet-256

We follow the hyperparameter set-up of Ma et al. (2024) to train a SiT model on conditional generation across B/2, L/2 and XL/2 model scales using independent, class-conditional and **GT** transport. We report the Fréchet Inception Distance (FID) (Heusel et al., 2017) and $\mathrm{FD}_{\mathrm{DINOv2}}$ (Stein et al., 2023) metrics to measure the distributional distance between the generated and real distributions. We highlight our best configuration in table 1 showing an improvement from the standard SiT-XL/2 trained with independent couplings to ours trained with the **GT** strategy, improving from 2.06 to 1.91 in FID. This one result, however, lays in front of a more interesting story.

Namely, with **GT** the guidance scale w can be pushed to much larger values before performance begins to degrade, enabling us to push far more aggressively into high guidance strength regimes. To illustrate, consider fig. 4 where we compare the performance at various guidance scales w along with applying the guidance interval technique (Kynkäänniemi et al., 2024). We notice that for the model trained with **GT** couplings we can use markedly higher guidance scales than the model trained with independent couplings; along with having an overall better minimum w.r.t. FID. Observe that interval tuning improves all the couplings and shifts their optima to larger w, whilst **GT** obtains the best global optima among the coupling strategies. Notably, the gap between **GT** and independent couplings grows as w increases. Further, observe that semi-discrete **GT** which is closer to the exact OT map is the most robust at higher guidance scales, however does not reach the overall global minima FID. To complement fig. 4 we report the best results across the different coupling strategies under several model sizes, swept over different guidance scales (per metric) in table 2. Observe that across all NFE and model sizes the **GT** couplings obtain the strongest performance yielding a noticeable improvement. For model and training configurations please refer to appendix C.

Table 2: Comparison of different couplings across various NFEs and model sizes reported in FID and $\text{FD}_{\text{DINOv2}}$. We swept for the optimal guidance strength w^* for each setting and report the best result.

Model	Coupling	FID ↓			$\text{FD}_{\text{DINOv2}}$ ↓		
		16	32	64	16	32	64
SiT-B/2 (130M)	Independent	5.88	5.33	5.16	137.3	132.9	132.3
	Class-cond. OT	5.79	5.20	5.08	136.4	132.2	131.4
	GT + SD-OT	**5.11**	4.90	4.85	199.6	194.6	184.7
	GT	5.15	**4.35**	**4.15**	**123.6**	**118.0**	**116.4**
SiT-L/2 (459M)	Independent	5.50	5.74	2.95	91.1	84.6	83.3
	Class-cond. OT	5.47	5.68	2.98	102.0	99.2	95.0
	GT	**3.94**	**3.84**	**2.69**	**83.5**	**78.7**	**73.5**
SiT-XL/2 (675M)	Independent	3.95	3.00	2.78	84.2	78.7	77.8
	GT	**3.22**	**2.69**	**2.59**	**72.7**	**69.1**	**68.4**

Distillation into a flow map. We next assess whether **GT** couplings can serve as a distillation strategy for few-step generators. Following Lee et al. (2026), we distill a SiT-B/2 and SiT-L/2 flow matching teacher into a flow map with using the meanflow distillation objective (Geng et al., 2025). The teacher is pretrained with either independent or **GT** couplings and the student is distilled with **GT** in both cases. On SiT-B/2, observe that applying **GT** only at distillation (i.e., from an independently pretrained teacher) matches pretraining with **GT** throughout, indicating that **GT** is effective purely as a distillation strategy for stronger few-step generators outperforming other coupling choices.

Adaptive step sampling. We compare coupling plans under the adaptive `dopri5` solver as shown in Table 4. **GT** attains the lowest FID and $\text{FD}_{\text{DINOv2}}$ overall, while semi-discrete **GT** requires the fewest function evaluations and outperforms the other couplings at high guidance scales.

Continuous Conditions. We next demonstrate that **GT** improves text-conditioned generation under classifier-free guidance. Following the text-conditioned ImageNet-256 setup of Cheng & Schwing (2025), we condition on image captions from an enriched version of ImageNet (Visual Layer). Captions are encoded by a frozen pretrained CLIP text encoder followed by an MLP that maps them to the conditioning signal. The `null` condition used for guidance dropout is a zero vector. As shown in Figure 5, **GT** achieves lower FID and $\text{FD}_{\text{DINOv2}}$ at higher guidance scales.



Figure 4: **Left:** Guidance scale w sweep for SiT-B/2 (Euler-64 steps) FID ↓ under a tuned guidance interval $[0, 0.7]$. **Right:** Class-conditional samples for class 339 (`sorrel`) from SiT-L/2 (Euler-64 steps, $w = 2.5$) under independent, class-conditional, and **GT** transport.

Table 3: **Left**: FID↓ on ImageNet-256 for DMF distilled with independent, class-conditional and GT couplings, at B/2 and L/2 scale. **Right**: Curated samples from GT DMF-B/2 at 4-NFE.

Model	Coupling	Teacher (NFE) 16	32	64	Student (NFE) 1	2	4
DMF-B/2	Independent	5.88	5.33	5.16	5.82	5.19	5.19
	Class-cond. OT	5.79	5.20	5.08	5.83	5.18	5.25
	GT (from GT teacher)	**5.15**	**4.35**	**4.15**	**5.66**	**4.17**	**4.14**
	GT (from ind. teacher)	5.88	5.32	5.22	5.88	4.21	4.18
DMF-L/2	Independent	5.50	5.74	2.95	**3.31**	2.82	2.68
	GT	**3.94**	**3.84**	**2.69**	4.03	**2.46**	**2.37**



Table 4: FID ↓, $\text{FD}_{\text{DINOv2}}$ ↓, and NFE ↓ for SiT-B/2 over coupling plans and guidance scales w.

	Independent			Class-cond. OT			GT			GT +SD-OT		
w	FID	$\text{FD}_{\text{DINOv2}}$	NFE	FID	$\text{FD}_{\text{DINOv2}}$	NFE	FID	$\text{FD}_{\text{DINOv2}}$	NFE	FID	$\text{FD}_{\text{DINOv2}}$	NFE
1.0	**25.01**	**586.47**	50.83	25.23	589.00	47.73	29.65	618.75	49.17	42.39	772.55	**38.00**
2.0	5.17	**206.41**	60.99	5.08	208.96	60.56	**4.06**	220.08	61.54	7.56	341.00	**50.00**
3.0	11.13	138.30	76.39	10.98	137.98	75.84	8.77	130.56	73.45	**4.79**	226.95	**62.80**
4.0	15.77	132.94	90.04	15.56	132.06	89.86	13.37	**116.34**	85.63	**5.47**	194.00	**77.09**

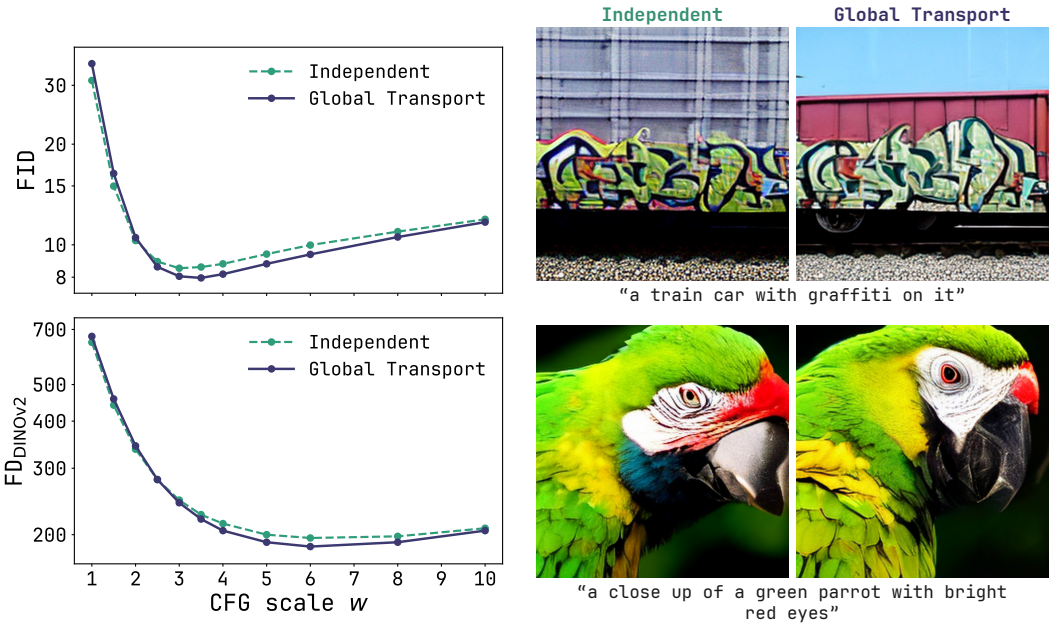

Figure 5: **Left:** FID↓ and $\text{FD}_{\text{DINOv2}}$↓ for text-conditioned ImageNet-256 across guidance scales w. **Right:** Text-conditioned ImageNet samples from independent and GT coupling models at $w = 4$.

5.2 Single-cell Experiments

We evaluate GT on conditional single-cell generation following Palma et al. (2025), conditioning on cell type for PBMC3K[1], Dentate gyrus (La Manno et al., 2018) and HLCA (Sikkema et al., 2023). Against c-CFGen (Palma et al., 2025), its linear interpolant variant, scDiffusion (Luo et al., 2024) and scVI (Gayoso et al., 2021), GT is best on three of six metrics and second best on three out of six (Table 5, details in Appendix D).

[1] https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

Table 5: Comparison of **GT** with single-cell generative models on distribution-matching metrics (RBF-kernel MMD and 2-Wasserstein distance), averaged over three seeds.

	PBMC3K		Dentate gyrus		HLCA	
	MMD (↓)	WD (↓)	MMD (↓)	WD (↓)	MMD (↓)	WD (↓)
c-CFGen	0.45 ± 0.00	11.17 ± 0.02	0.06 ± 0.00	7.26 ± 0.03	$\mathbf{0.06 \pm 0.00}$	5.06 ± 0.00
c-CFGen-linear	0.41 ± 0.00	10.27 ± 0.08	0.06 ± 0.00	6.69 ± 0.01	0.07 ± 0.00	$\mathbf{4.89 \pm 0.01}$
scDiffusion	0.67 ± 0.06	12.11 ± 0.16	0.06 ± 0.00	$\mathbf{5.89 \pm 0.01}$	0.12 ± 0.00	5.42 ± 0.01
scVI	0.58 ± 0.01	13.39 ± 0.16	0.11 ± 0.00	7.34 ± 0.03	0.13 ± 0.00	6.39 ± 0.01
GT (ours)	$\mathbf{0.39 \pm 0.00}$	$\mathbf{9.80 \pm 0.02}$	$\mathbf{0.05 \pm 0.00}$	6.61 ± 0.01	0.07 ± 0.00	4.98 ± 0.00

6 Related Work

Optimal Transport for Generative Models. Optimal transport (Benamou & Brenier, 2000) is widely used to improve unconditional (Tong et al., 2024a;b; Pooladian et al., 2023; Calvo-Ordonez et al., 2026) and class-conditional (Cheng & Schwing, 2025; Chemseddine et al., 2025) generation, at scale via semi-discrete potentials (Mousavi-Hosseini et al., 2025; Kong et al., 2026) and "re-flow"-style strategies that exploit flow invertibility (Kim et al., 2024; Berthelot et al., 2026). In biology, it has been applied to single-cell trajectory inference (Schiebinger et al., 2019; Kapuśniak et al., 2024; Petrović et al., 2025) and measure-to-measure transport (Haviv et al., 2025; Vandergrift et al., 2026).

Flow Maps. Flow maps (Frans et al., 2025; Geng et al., 2025; Boffi et al., 2025) have recently emerged as an efficient route to one- and few-step generators, either distilled (Sabour et al., 2026; Lee et al., 2026) or trained from scratch (Geng et al., 2025; Boffi et al., 2025; Geng et al., 2026), with applications to image (Lu et al., 2026; Wang et al., 2025b) and video (Gu et al., 2026; Shaul et al., 2026) generation.

Classifier-Free Guidance for Flow Models. Classifier-free guidance (CFG) (Ho & Salimans, 2022) is a widely adopted strategy for improving conditional generation with diffusion and flow matching models (Zheng et al., 2023). Subsequent work includes guidance interval tuning (Kynkäänniemi et al., 2024), velocity field projection (Fan et al., 2025; Cai et al., 2026), guidance weight schedules (Chung et al., 2025; Wang et al., 2024; Galashov et al., 2025) and guiding with a weaker checkpoint of the same model (Karras et al., 2024).

7 Conclusion

In this work, we studied how coupling choice affects conditional flow models sampled with classifier-free guidance. We introduced **Global Transport** (**GT**), which constructs *condition-agnostic* OT couplings and requires no change to the model architecture, sampler, or guidance rule. Although this choice may lead to a mismatch between the source distribution seen by each condition during training and degrades performance in the unguided conditional generation setting, in our experiments **GT** improves guided generation across all settings we study. This contrast shows that unguided performance is not a reliable basis for choosing a coupling when CFG is used during inference, and opens up a new direction of inquiry in designing couplings for more exotic inference strategies.

Limitations and future work Our prediction gap bound does not control the learned fields along CFG trajectories or guarantee a resulting hierarchy of generation quality. Understanding when each construction is preferable and how coupling choice interacts with other guidance and post-training methods are useful directions for future work. As CFG is primarily used in image and video generation, transferring these gains to domains outside of cells in the life sciences, where other factors dominate, remains open.

AI use statement

In this work, we used generative AI tools for improving parts of the manuscript, related work discovery, implementation and coding support, mathematical claims, and internal review of the paper. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

This work proposes a training-time coupling strategy for conditional flow models and uses only publicly available datasets. No new data were collected and no human subjects were involved. As with other image generative models, our models could in principle be misused, and they may reproduce biases present in their training data. We do not think this work raises ethical concerns beyond those generally common to generative modeling research.

Reproducibility statement

We describe **GT** in the main text and provide implementation details, hyperparameters, and proofs in the appendix. All datasets used are publicly available. We will release our code in a public GitHub repository upon publication.

References

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *International Conference on Learning Representations (ICLR)*, 2023.

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint 2303.08797*, 2023.

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models, 2023. URL https://arxiv.org/abs/2209.12152.

Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

David Berthelot, Tianrong Chen, Jiatao Gu, Marco Cuturi, Laurent Dinh, Bhavik Chandna, Michal Klein, Josh Susskind, and Shuangfei Zhai. The coupling within: Flow matching via distilled normalizing flows, 2026. URL https://arxiv.org/abs/2603.09014.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Nicholas M. Boffi, Michael S. Albergo, and Eric Vanden-Eijnden. How to build a consistency model: Learning flow maps via self-distillation, 2025. URL https://arxiv.org/abs/2505.18825.

Samuel Boïté, Julie Delon, and Kimia Nadjahi. Expected batch optimal transport plans and consequences for flow matching. *arXiv preprint arXiv:2605.12174*, 2026.

Jian-Feng Cai, Haixia Liu, Zhengyi Su, and Chao Wang. Improving classifier-free guidance of flow matching via manifold projection. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=IPp3LD6u16.

Sergio Calvo-Ordonez, Matthieu Meunier, Alvaro Cartea, Christoph Reisinger, Yarin Gal, and Jose Miguel Hernandez-Lobato. Weighted conditional flow matching, 2026. URL https://arxiv.org/abs/2507.22270.

Jannis Chemseddine, Paul Hagemann, Gabriele Steidl, and Christian Wald. Conditional wasserstein distances with applications in bayesian ot flow matching. *Journal of Machine Learning Research*, 26(141):1–47, 2025.

Ho Kei Cheng and Alexander Schwing. The curse of conditions: Analyzing and improving optimal transport for conditional flow-based generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15875–15884. IEEE, 2025.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. Cfg++: Manifold-constrained classifier free guidance for diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 30824–30850, 2025.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Neural Information Processing Systems (NIPS)*, 2013.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. URL https://arxiv.org/abs/2105.05233.

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 *International Conference on Machine Learning (ICML)*, 2024.

Weichen Fan, Amber Yijia Zheng, Raymond A Yeh, and Ziwei Liu. Cfg-zero*: Improved classifier-free guidance for flow matching models. *arXiv preprint arXiv:2503.18886*, 2025.

Kilian Fatras, Younes Zine, Szymon Majewski, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Minibatch optimal transport distances; analysis and applications. *arXiv preprint 2101.01792*, 2021.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *International Conference on Learning Representations*, volume 2025, pp. 34668–34684, 2025.

Alexandre Galashov, Ashwini Pokle, Arnaud Doucet, Arthur Gretton, Mauricio Delbracio, and Valentin De Bortoli. Learn to guide your diffusion model, 2025. URL https://arxiv.org/abs/2510.00815.

Adam Gayoso, Zoë Steier, Romain Lopez, Jeffrey Regier, Kristopher L Nazor, Aaron Streets, and Nir Yosef. Joint probabilistic modeling of single-cell multi-omic data with totalvi. *Nature methods*, 18 (3):272–282, 2021.

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

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

Jiatao Gu, Tianrong Chen, David Berthelot, Huangjie Zheng, Yuyang Wang, Ruixiang Zhang, Laurent Dinh, Miguel Angel Bautista, Josh Susskind, and Shuangfei Zhai. Starflow: Scaling latent normalizing flows for high-resolution image synthesis, 2025. URL https://arxiv.org/abs/2506.06276.

Yuchao Gu, Guian Fang, Yuxin Jiang, Weijia Mao, Song Han, Han Cai, and Mike Zheng Shou. Anyflow: Any-step video diffusion model with on-policy flow map distillation, 2026. URL https://arxiv.org/abs/2605.13724.

Doron Haviv, Aram-Alexandre Pooladian, Dana Pe'er, and Brandon Amos. Wasserstein flow matching: Generative modeling over families of distributions. In *International Conference on Machine Learning (ICML)*, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

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

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

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. Simple diffusion: End-to-end diffusion for high resolution images, 2023. URL https://arxiv.org/abs/2301.11093.

Allan Jabri, David Fleet, and Ting Chen. Scalable adaptive computation for iterative generation, 2023. URL https://arxiv.org/abs/2212.11972.

Kacper Kapuśniak, Peter Potaptchik, Teodora Reu, Leo Zhang, Alexander Tong, Michael Bronstein, Avishek Joey Bose, and Francesco Di Giovanni. Metric flow matching for smooth interpolations on the data manifold. *Neural Information Processing Systems (NeurIPS)*, 2024.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself, 2024. URL https://arxiv.org/abs/2406.02507.

Gavin Kerrigan, Giosue Migliorini, and Padhraic Smyth. Dynamic conditional optimal transport through simulation-free flows. *Advances in Neural Information Processing Systems*, 37:93602–93642, 2024.

Beomsu Kim, Yu-Guan Hsieh, Michal Klein, Marco Cuturi, Jong Chul Ye, Bahjat Kawar, and James Thornton. Simple reflow: Improved techniques for fast flow models, 2024. URL https://arxiv.org/abs/2410.07815.

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. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=nTCF3QNsIN.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *Advances in Neural Information Processing Systems*, 37:122458–122483, 2024.

Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E Kastriti, Peter Lönnerberg, Alessandro Furlan, et al. Rna velocity of single cells. *Nature*, 560(7719):494–498, 2018.

Kyungmin Lee, Sihyun Yu, and Jinwoo Shin. Decoupled meanflow: Turning flow models into flow maps for accelerated sampling. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=P4VsQexEPe.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization, 2024. URL https://arxiv.org/abs/2406.11838.

Zihao Li, Zhichen Zeng, Xiao Lin, Feihao Fang, Yanru Qu, Zhe Xu, Zhining Liu, Xuying Ning, Tianxin Wei, Ge Liu, et al. Flow matching meets biology and life science: a survey. *npj Artificial Intelligence*, 2(1):17, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *International Conference on Learning Representations (ICLR)*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *International Conference on Learning Representations (ICLR)*, 2023.

Yiyang Lu, Qiao Sun, Xianbang Wang, Zhicheng Jiang, Hanhong Zhao, and Kaiming He. Bidirectional normalizing flow: From data to noise and back, 2025. URL https://arxiv.org/abs/2512.10953.

Yiyang Lu, Susie Lu, Qiao Sun, Hanhong Zhao, Zhicheng Jiang, Xianbang Wang, Tianhong Li, Zhengyang Geng, and Kaiming He. One-step latent-free image generation with pixel mean flows, 2026. URL https://arxiv.org/abs/2601.22158.

Erpai Luo, Minsheng Hao, Lei Wei, and Xuegong Zhang. scdiffusion: conditional generation of high-quality single-cell data using diffusion model. *Bioinformatics*, 40(9):btae518, 2024.

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, 2024. URL https://arxiv.org/abs/2401.08740.

Shimon Malnick, Matan Rusanovsky, Ohad Fried, and Shai Avidan. Optimal transport flow matching by design, 2026. URL https://arxiv.org/abs/2606.04092.

Laurence Illing Midgley, Vincent Stimper, Gregor N. C. Simm, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Flow annealed importance sampling bootstrap, 2023. URL https://arxiv.org/abs/2208.01893.

Alex Morehead, Lazar Atanackovic, Akshata Hegde, Yanli Wang, Frimpong Boadu, Joel Selvaraj, Alexander Tong, Aditi Krishnapriyan, and Jianlin Cheng. Flow matching for generative modelling in bioinformatics and computational biology. *Nature Machine Intelligence*, pp. 1–18, 2026.

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

Alessandro Palma, Till Richter, Hanyi Zhang, Manuel Lubetzki, Alexander Tong, Andrea Dittadi, and Fabian Theis. Multi-modal and multi-attribute generation of single cells with cfgen, 2025. URL https://arxiv.org/abs/2407.11734.

William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

Stefano Peluchetti. Non-denoising forward-time diffusions. *arXiv preprint arXiv:2312.14589*, 2023.

Katarina Petrović, Lazar Atanackovic, Viggo Moro, Kacper Kapuśniak, İsmail İlkan Ceylan, Michael Bronstein, Avishek Joey Bose, and Alexander Tong. Curly flow matching for learning non-gradient field dynamics, 2025. URL https://arxiv.org/abs/2510.26645.

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 T.Q. Chen. Multisample flow matching: Straightening flows with minibatch couplings. *International Conference on Learning Representations (ICLR)*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Amirmojtaba Sabour, Sanja Fidler, and Karsten Kreis. Align your flow: Scaling continuous-time flow map distillation. *Advances in Neural Information Processing Systems*, 38:146459–146512, 2026.

Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets, 2022. URL https://arxiv.org/abs/2202.00273.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, Lia Lee, Jenny Chen, Justin Brumbaugh, Philippe Rigollet, Konrad Hochedlinger, Rudolf Jaenisch, Aviv Regev, and Eric S. Lander. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943.e22, 2019.

Neta Shaul, Chao Liu, Arash Vahdat, and Julius Berner. Parallel decoding distillation for fast image and video generation, 2026. URL https://arxiv.org/abs/2607.26004.

Lisa Sikkema, Ciro Ramírez-Suástegui, Daniel C Strobl, Tessa E Gillett, Luke Zappia, Elo Madissoon, Nikolay S Markov, Laure-Emmanuelle Zaragosi, Yuge Ji, Meshal Ansari, et al. An integrated cell atlas of the lung in health and disease. *Nature medicine*, 29(6):1563–1577, 2023.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=08zf7kTOoh.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction, 2024. URL https://arxiv.org/abs/2404.02905.

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 (TMLR)*, 2024a.

Alexander Tong, Nikolay Malkin, Kilian Fatras, Lazar Atanackovic, Yanlei Zhang, Guillaume Huguet, Guy Wolf, and Yoshua Bengio. Simulation-free schrödinger bridges via score and flow matching. *AISTATS*, 2024b.

Matthew Vandergrift, Martha White, Yury Polyanskiy, Philippe Rigollet, and Lazar Atanackovic. Measure-to-measure regression with transformers. *arXiv preprint arXiv:2605.28075*, 2026.

Visual Layer. ImageNet-1K-VL-Enriched. Hugging Face dataset. https://huggingface.co/datasets/visual-layer/imagenet-1k-vl-enriched.

Kaibo Wang, Jianda Mao, Tong Wu, and Yang Xiang. Towards a golden classifier-free guidance path via foresight fixed point iterations, 2025a. URL https://arxiv.org/abs/2510.21512.

Xi Wang, Nicolas Dufour, Nefeli Andreou, Marie-Paule Cani, Victoria Fernandez Abrevaya, David Picard, and Vicky Kalogeiton. Analysis of classifier-free guidance weight schedulers, 2024. URL https://arxiv.org/abs/2404.13040.

Zidong Wang, Yiyuan Zhang, Xiaoyu Yue, Xiangyu Yue, Yangguang Li, Wanli Ouyang, and Lei Bai. Transition models: Rethinking the generative learning objective, 2025b. URL https://arxiv.org/abs/2509.04394.

Stephen Zhang, Alireza Mousavi-Hosseini, Michal Klein, and Marco Cuturi. On fitting flow models with large sinkhorn couplings, 2025. URL https://arxiv.org/abs/2506.05526.

Qinqing Zheng, Matt Le, Neta Shaul, Yaron Lipman, Aditya Grover, and Ricky T. Q. Chen. Guided flows for generative modeling and decision making, 2023. URL https://arxiv.org/abs/2311.13443.

A SYNTHETIC EXAMPLE

Dataset construction. We construct a 40-Gaussian-Mixture with 4 classes. Following Midgley et al. (2023), we use a two-dimensional 40-component Gaussian mixture (40-GMM). All components have equal weight and share the isotropic covariance

$$\Sigma = \begin{pmatrix} 40 & 0 \\ 0 & 40 \end{pmatrix}, \tag{10}$$

with means μ_i drawn uniformly from the box $[-40, 40]^2$, i.e. $\mu_i \sim \mathcal{U}(-40, 40)^2$, giving the density

$$p_{\text{gmm}}(x) = \frac{1}{40} \sum_{i=1}^{40} \mathcal{N}(x;\, \mu_i, \Sigma). \tag{11}$$

The ground truth is shown in figure 6.

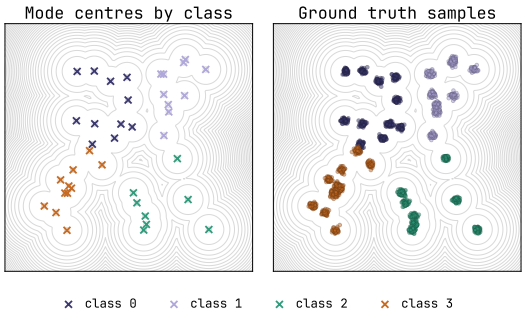

Figure 6: 40-GMM ground truth data. Left: Mode centers by class, Right: Ground truth samples.

Generated trajectories. We provide generated trajectories across a wide range of guidance scales for Independent, Class-conditional and **GT** transport. We observe that "yo-yo" is present in Independent with trajectories curving inwards and outwards close to the conditioned class.

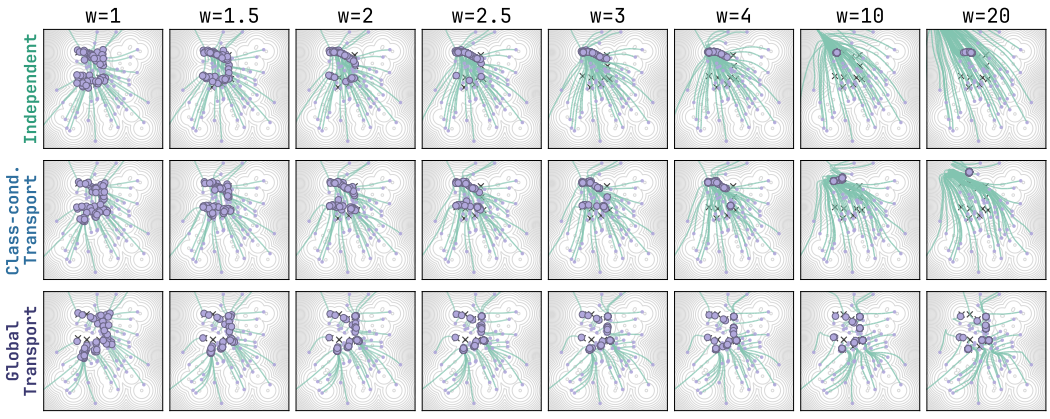

Figure 7: Generated trajectories for Independent, Class-conditional and **GT** on 40-GMM.

Velocity Field Heatmaps. We compute heatmaps for unconditional and conditional velocity fields comparing different coupling plans. Under the independent coupling, both unconditional and conditional field start as a sink located around class origin, reorganizing rapidly at $t = 0.5$. Under **GT** we observe that unconditional velocity field forms already at $t = 0$ and remains stable across the trajectory.

Additional samples. We provide additional generated samples to accompany figure 1 demonstrating **GT** does not drop off the modes even in cases of extremely high guidance scale $w = 20$.

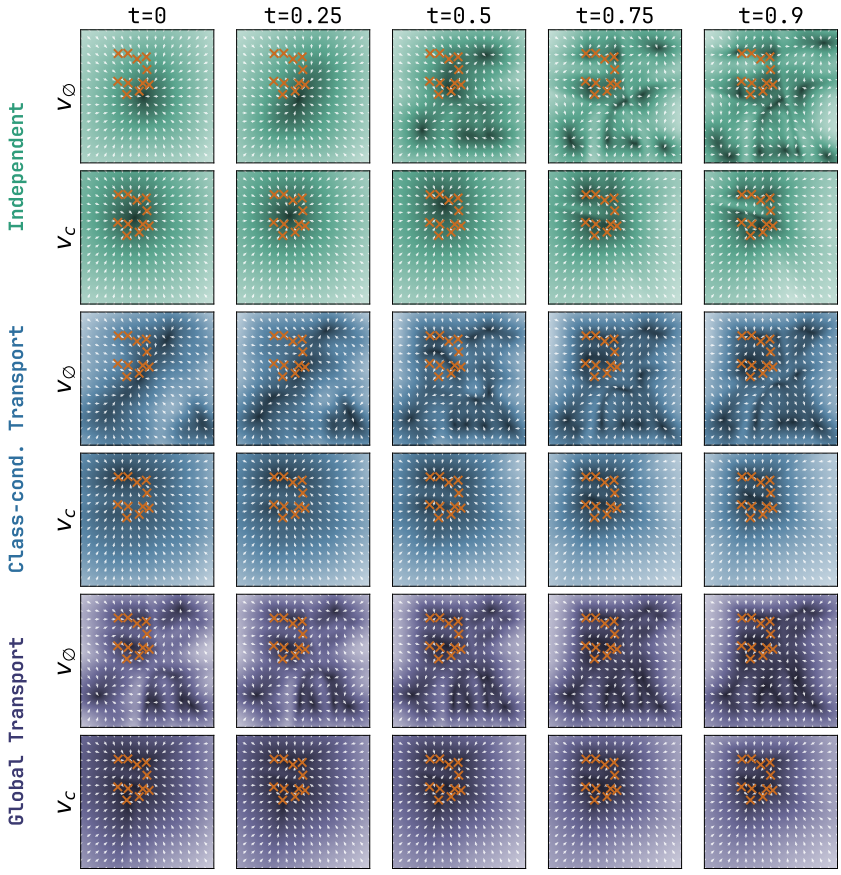

Figure 8: Velocity field heat maps for conditional v_c and unconditional $v_\varnothing$ for Independent, Class-conditional and **GT** transport.

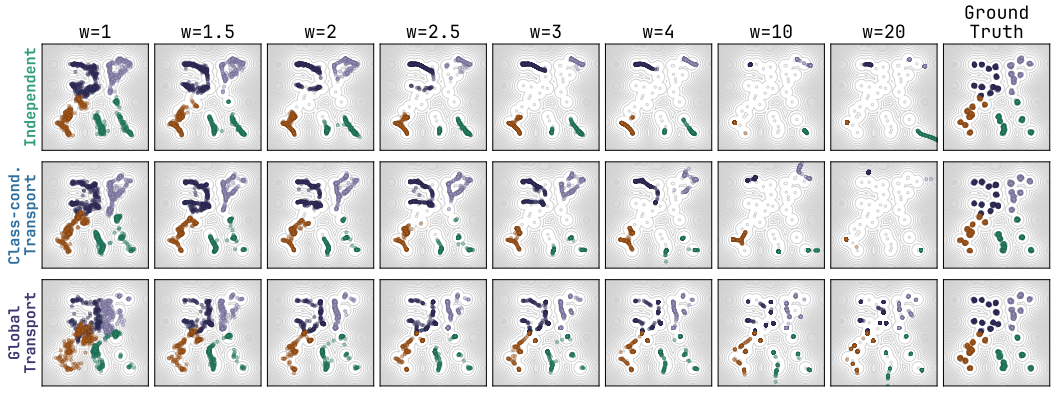

Figure 9: Additional 40-GMM samples across a range of guidance scales.

B Proofs and Derivations

B.1 Relationship between coupling and prediction gap

For the ease of notation we will write $u_t(x_t)$ as v_u and $u_t(x_t \mid c)$ as v_c.

Proposition B.1. *For any coupling π,* $\mathbb{E}_\pi \left[\|u_t(x_t \mid c) - u_t(x_t)\|^2 \right] = L_u^\pi(t) - L_c^\pi(t)$.

Proof. We can write the following by adding and subtracting u to $g = v_c - v_u$

$$u - v_u = (u - v_c) + (v_c - v_u) \tag{12}$$

By squaring and taking the norm we get

$$\begin{aligned} \|u - v_u\|^2 = \|u - v_c\|^2 + \|v_c - v_u\|^2 \\ + 2\langle u - v_c, v_c - v_u \rangle. \end{aligned} \tag{13}$$

We now take expectation over $(x_0, x_1, c) \sim \pi$

$$\begin{aligned} \mathbb{E}_\pi \|u - v_u\|^2 = \mathbb{E}_\pi \|u - v_c\|^2 + \mathbb{E}_\pi \|v_c - v_u\|^2 \\ + 2\mathbb{E}_\pi \langle u - v_c, v_c - v_u \rangle. \end{aligned} \tag{14}$$

The expectation of the final term is zero:

$$\begin{aligned} &\mathbb{E}_\pi \left[\langle u - v_c, v_c - v_u \rangle \right] \\ &= \mathbb{E}_\pi \left[\mathbb{E}_\pi \left[\langle u - v_c, v_c - v_u \rangle \mid x_t, c \right] \right] \\ &= \mathbb{E}_\pi \left[\langle \mathbb{E}_\pi [u - v_c \mid x_t, c], v_c - v_u \rangle \right]. \end{aligned} \tag{15}$$

By the definition of the conditional field,

$$v_c(x_t, t, c) = \mathbb{E}_\pi [u \mid x_t, c], \tag{16}$$

and therefore

$$\mathbb{E}_\pi [u - v_c \mid x_t, c] = 0. \tag{17}$$

Hence

$$\mathbb{E}_\pi \left[\langle u - v_c, v_c - v_u \rangle \right] = 0. \tag{18}$$

Taking expectations in equation 13 now gives

$$\mathbb{E}_\pi \|u - v_u\|^2 = \mathbb{E}_\pi \|u - v_c\|^2 + \mathbb{E}_\pi \|v_c - v_u\|^2. \tag{19}$$

Using the definitions of $L_u^\pi(t)$ and $L_c^\pi(t)$, we obtain

$$\boxed{\mathbb{E}_\pi \|v_c - v_u\|^2 = L_u^\pi(t) - L_c^\pi(t)} \tag{20}$$

which proves the result. □

Proposition B.2. *The error term $L_u^\pi(t)$ can be upper bounded as*

$$\int_0^1 L_u^\pi(t) \, \mathrm{d}t \le c(\pi) - \mathcal{W}_2^2(p_0, p_1) =: d(\pi), \tag{21}$$

where $c(\pi)$ is the coupling cost and $d(\pi) \ge 0$ denotes the excess quadratic transport cost, i.e., the suboptimality of π relative to the population-optimal coupling (Boïté et al., 2026).

Proof. Let us consider coupling cost $c(\pi) = \mathbb{E}_\pi \left[\|x_1 - x_0\|^2 \right]$ and ideal minimizer of the unconditional field $v_u(x, t) = \mathbb{E}_\pi [u \,|\, x = x_t]$, where $u = x_1 - x_0$. We assume that the joint distribution $\pi(x_0, x_1)$ induces a probability path $x_t \sim p_t$. We can then express time-dependent error $L_u^\pi(t)$ as

$$L_u^\pi(t) = \mathbb{E}_\pi \left[\|u - v_u(x_t, t)\|^2 \right] \tag{22}$$

$$= \mathbb{E}_\pi \left[\|u\|^2 \right] - 2 \underbrace{\mathbb{E}_\pi \langle u, v_u(x_t, t) \rangle}_{\star} + \mathbb{E}_{x_t \sim p_t} \left[\|v_u(x_t, t)\|^2 \right] \tag{23}$$

We further express $\star$ as

$$\begin{aligned}\mathbb{E}_\pi\big[\langle u,\, v_u(x_t,t)\rangle\big] &= \mathbb{E}_{x_t\sim p_t}\Big[\big\langle \underbrace{\mathbb{E}_\pi[u \mid x_t]}_{=\, v_u(x_t,t)},\; v_u(x_t,t)\big\rangle\Big] \\ &= \mathbb{E}_{x_t\sim p_t}\big[\langle v_u(x_t,t),\, v_u(x_t,t)\rangle\big] \\ &= \mathbb{E}_{x_t\sim p_t}\big[\,\|v_u(x_t,t)\|^2\,\big]. \end{aligned} \tag{24}$$

From here it follows

$$L_u^\pi(t) = \underbrace{\mathbb{E}_\pi\left[\|u\|^2\right]}_{c(\pi)} - \mathbb{E}_{x_t\sim p_t}\left[\|v_u(x_t,t)\|^2\right] \tag{25}$$

or equivalently

$$L_u^\pi(t) = c(\pi) - \mathbb{E}_{x_t\sim p_t}\left[\|v_u(x_t,t)\|^2\right] \tag{26}$$

If we integrate both sides with respect to t

$$\int_0^1 L_u^\pi(t)\mathrm{d}t = c(\pi) - \int_0^1 \mathbb{E}_{x_t\sim p_t}\left[\|v_u(x_t,t)\|^2\right]\mathrm{d}t \tag{27}$$

Using Benamou & Brenier (2000) we can express

$$\int_0^1 \mathbb{E}_{x_t\sim p_t}\left[\|v_u(x_t,t)\|^2\right]\mathrm{d}t \geq \mathcal{W}_2^2(p_0,p_1) \tag{28}$$

From which identity follows

$$\boxed{\int_0^1 L_u^\pi(t)\mathrm{d}t \leq \underbrace{c(\pi) - \mathcal{W}_2^2(p_0,p_1)}_{d(\pi)}}. \tag{29}$$

□

B.2 Proof of Proposition 4.2: Hierarchy of Coupling Costs

Proposition B.3 (Coupling Cost Ordering). *For all $\beta \geq 0$, the quadratic transport costs of the global transport coupling $\pi^\star_{\texttt{GT}}$, the class-conditional optimal transport coupling $\pi^\star_{\mathrm{C^2OT}(\beta)}$, and the independent coupling $\pi_{\mathrm{ind}} = p_0 \otimes p_1$ satisfy:*

$$\mathrm{Cost}(\pi^\star_{\texttt{GT}}) \leq \mathrm{Cost}(\pi^\star_{\mathrm{C^2OT}(\beta)}) \leq \mathrm{Cost}(\pi_{\mathrm{ind}}), \tag{30}$$

with the lower equality achieved when $\beta = 0$.

Proof. Let $\mathcal{X} = \mathbb{R}^d$ denote the state space and $\mathcal{C}$ denote the conditioning space (endowed with metric $d_{\mathcal{C}}$ or squared norm $\|c_0 - c_1\|^2$). Let p_0 be the source noise distribution on $\mathcal{X}$, and let $p_1(x_1, c_1)$ be the joint data-conditioning distribution on $\mathcal{X} \times \mathcal{C}$ with spatial marginal $p_1(x_1) = \int_{\mathcal{C}} p_1(x_1, c_1)\,\mathrm{d}c_1$. Recall that the quadratic transport cost of any coupling π with marginals p_0 and $p_1(x_1)$ is defined by:

$$\mathrm{Cost}(\pi) := \mathbb{E}_{(x_0,x_1)\sim\pi}\left[\|x_1 - x_0\|^2\right]. \tag{31}$$

Part 1: $\mathrm{Cost}(\pi^\star_{\texttt{GT}}) \leq \mathrm{Cost}(\pi^\star_{\mathrm{C^2OT}(\beta)})$**.** By definition, the global transport coupling $\pi^\star_{\texttt{GT}}$ is the minimizer of the unconstrained Kantorovich optimal transport problem between p_0 and the spatial data marginal $p_1(x_1)$:

$$\pi^\star_{\texttt{GT}} \in \operatorname*{argmin}_{\pi\in\Pi(p_0,p_1)} \mathbb{E}_\pi\left[\|x_1 - x_0\|^2\right]. \tag{32}$$

Therefore, for *any* admissible joint distribution $\pi \in \Pi(p_0, p_1)$, we have by definition of the infimum:

$$\mathrm{Cost}(\pi^\star_{\texttt{GT}}) \leq \mathrm{Cost}(\pi). \tag{33}$$

Now consider the class-conditional optimal transport coupling $\pi^\star_{\mathrm{C^2OT}(\beta)}$. For any regularization strength $\beta \geq 0$, $\pi^\star_{\mathrm{C^2OT}(\beta)}$ is defined as the minimizer of the joint spatial and condition transport problem:

$$\pi^\star_{\mathrm{C^2OT}(\beta)} \in \operatorname*{argmin}_{\pi \in \Pi(p_0,p_1)} \mathbb{E}_{((x_0,c_0),(x_1,c_1))\sim\pi} \left[\|x_1 - x_0\|^2 + \beta\|c_1 - c_0\|^2\right]. \tag{34}$$

Since the spatial marginals of any candidate coupling in equation 34 are constrained to be p_0 and $p_1(x_1)$, the resulting optimal coupling $\pi^\star_{\mathrm{C^2OT}(\beta)}$ is itself an admissible coupling in $\Pi(p_0, p_1)$.

Applying the optimality property of $\pi^\star_{\mathtt{GT}}$ from equation 33 directly to $\pi = \pi^\star_{\mathrm{C^2OT}(\beta)}$ yields:

$$\mathrm{Cost}(\pi^\star_{\mathtt{GT}}) \leq \mathrm{Cost}(\pi^\star_{\mathrm{C^2OT}(\beta)}), \quad \forall \beta \geq 0. \tag{35}$$

When $\beta = 0$, the penalty on the condition labels vanishes, and the objective in equation 34 reduces identically to the objective in equation 32. Thus, $\pi^\star_{\mathrm{C^2OT}(0)} = \pi^\star_{\mathtt{GT}}$, achieving equality.

Part 2: $\mathrm{Cost}(\pi^\star_{\mathrm{C^2OT}(\beta)}) \leq \mathrm{Cost}(\pi_{\mathrm{ind}})$. Next, we show that the independent product coupling $\pi_{\mathrm{ind}} = p_0 \otimes p_1$ serves as an upper bound on $\mathrm{Cost}(\pi^\star_{\mathrm{C^2OT}(\beta)})$. Let $(x_0, c_0) \sim p_0 \times p_{\mathcal{C}}$ and $(x_1, c_1) \sim p_1$ be drawn independently, where $p_{\mathcal{C}}$ is the marginal condition distribution under p_1 assigned independently to noise points x_0. Because $\pi_{\mathrm{ind}} \in \Pi(p_0, p_1)$ is an admissible joint distribution, the minimality of $\pi^\star_{\mathrm{C^2OT}(\beta)}$ for the joint objective in equation 34 implies:

$$\begin{aligned}\mathbb{E}_{\pi^\star_{\mathrm{C^2OT}(\beta)}} \left[\|x_1 - x_0\|^2\right] &\leq \mathbb{E}_{\pi^\star_{\mathrm{C^2OT}(\beta)}} \left[\|x_1 - x_0\|^2\right] + \beta\, \mathbb{E}_{\pi^\star_{\mathrm{C^2OT}(\beta)}} \left[\|c_1 - c_0\|^2\right] \\ &\leq \mathbb{E}_{\pi_{\mathrm{ind}}} \left[\|x_1 - x_0\|^2 + \beta\|c_1 - c_0\|^2\right]. \end{aligned} \tag{36}$$

In the independent coupling π_{ind}, the conditioning assignments c_0 and c_1 are mutually independent identically distributed samples from $p_{\mathcal{C}}$.

In the hard class-matching limit ($\beta \to \infty$ or exact class-conditional coupling where $c_0 = c_1$), the transport problem decomposes into $|\mathcal{C}|$ independent within-class sub-problems. For each class $c \in \mathcal{C}$ with prevalence $p(c)$:

$$\pi^\star_{\mathrm{C^2OT}(\infty)} = \sum_{c\in\mathcal{C}} p(c)\, \pi^\star_c, \quad \text{where } \pi^\star_c \in \operatorname*{argmin}_{\pi \in \Pi(p_0, p_1(\cdot|c))} \mathbb{E}_\pi \left[\|x_1 - x_0\|^2\right]. \tag{37}$$

For every class c, the product measure $\pi_{\mathrm{ind},c} = p_0 \otimes p_1(\cdot \mid c)$ is an admissible coupling for the c-th sub-problem. By the optimality of $\pi^\star_c$ over $\Pi(p_0, p_1(\cdot \mid c))$:

$$\mathbb{E}_{\pi^\star_c} \left[\|x_1 - x_0\|^2\right] \leq \mathbb{E}_{\pi_{\mathrm{ind},c}} \left[\|x_1 - x_0\|^2\right]. \tag{38}$$

Taking the expectation over the class distribution $p(c)$ yields:

$$\begin{aligned}\mathrm{Cost}(\pi^\star_{\mathrm{C^2OT}(\infty)}) &= \sum_{c\in\mathcal{C}} p(c)\, \mathbb{E}_{\pi^\star_c} \left[\|x_1 - x_0\|^2\right] \\ &\leq \sum_{c\in\mathcal{C}} p(c)\, \mathbb{E}_{\pi_{\mathrm{ind},c}} \left[\|x_1 - x_0\|^2\right] \\ &= \mathbb{E}_{\pi_{\mathrm{ind}}} \left[\|x_1 - x_0\|^2\right] = \mathrm{Cost}(\pi_{\mathrm{ind}}). \end{aligned} \tag{39}$$

By monotonicity of the optimal transport objective under relaxation of the condition constraint, $\mathrm{Cost}(\pi^\star_{\mathrm{C^2OT}(\beta)}) \leq \mathrm{Cost}(\pi_{\mathrm{ind}})$ holds for all $\beta \geq 0$.

Combining Part 1 and Part 2 gives the complete chain of inequalities:

$$\mathrm{Cost}(\pi^\star_{\mathtt{GT}}) \leq \mathrm{Cost}(\pi^\star_{\mathrm{C^2OT}(\beta)}) \leq \mathrm{Cost}(\pi_{\mathrm{ind}}), \tag{40}$$

which concludes the proof. □

B.3 Coupling controls flow curvature

We next compare the acceleration in the flow induced by the coupling in the standardized Gaussian setting. We use a Gaussian plug-in model of mini-batch OT which provides an analytically tractable proxy for the intractable discrete mini-batch coupling used in our experiments. For a velocity field v_t, let $a_t[v](x) := \partial_t v_t(x) + D_x v_t(x) v_t(x)$ be the acceleration experienced by a particle $\dot{x}_t = v_t(x_t)$ following the flow, also known as the *material acceleration*.

Proposition B.4 (Material acceleration under Gaussian couplings). *Let $p_0 = p_1 = \mathcal{N}(0, \mathrm{Id}_d)$, and let v_t^{ind}, $v_{t,n}$, and $v_t^\star$ denote respectively the vector fields constructed using (i) the independent coupling, (ii) mini-batch plug-in Gaussian OT coupling with batch size n, and (iii) the population OT coupling. Fixing d, as $n \to \infty$, the following holds pointwise in (t, x):*

$$a_t[v^{\mathrm{ind}}](x) = \tfrac{x}{((1-t)^2+t^2)^2}, \qquad a_t[v_n](x) = \tfrac{2\|x\|^2-(d-1)}{2n} x + o\left(\tfrac{1}{n}\right), \quad a_t[v^\star](x) = 0. \tag{41}$$

Define the total material acceleration $\mathfrak{A}(v)^2 := \mathbb{E}_{X_0} \int_0^1 \|a_t[v](X_t)\|^2 \,\mathrm{d}t$. Then,

$$\mathfrak{A}(v^{\mathrm{ind}}) = \sqrt{\left(2 + \tfrac{3\pi}{4}\right) d}, \qquad \mathfrak{A}(v_n) = \tfrac{\sqrt{d(d^2+18d+41)}}{2n} + o(n^{-1}), \qquad \mathfrak{A}(v^\star) = 0. \tag{42}$$

Thus both pointwise and total material acceleration scale as $O(n^{-1})$ for Gaussian plug-in mini-batch OT, remain $O(1)$ for independent FM, and vanish for population OT.

Gaussian plug-in minibatch model We study minibatch OT flow matching using an analytically tractable Gaussian model. Let $p_0 = \mathcal{N}(0, \mathrm{Id})$ and $p_1 = \mathcal{N}(0, \Sigma)$ be the source and target distributions. For independent minibatches of size n from p_0 and p_1, we fit the source and target means and covariances, couple the fitted Gaussians by their Gaussian Monge map, and marginalize the resulting conditional flow-matching fields over the randomly drawn minibatches. We emphasize that this is a *Gaussian plug-in model* of mini-batch transport that we use for its analytical tractability, since exact analysis of the discrete Hungarian coupling used in our experiments would likely be significantly more involved.

Writing $\widehat{p}_{t,n}(x)$ and $\widehat{v}_{t,n}(x)$ for the path density and velocity induced by a batch pair, the marginal flow field learned by flow matching is

$$v_{t,n}(x) = \frac{\mathbb{E}[\widehat{p}_{t,n}(x)\widehat{v}_{t,n}(x)]}{\mathbb{E}[\widehat{p}_{t,n}(x)]}. \tag{43}$$

Here and throughout this section, these expectations are understood to be taken over the independently sampled source and target minibatches. This posterior density weighting creates a nonlinear correction even though every batch-conditional field is affine.

Since p_0 is isotropic, we may assume that the target has diagonal covariance $\Sigma = \mathrm{diag}(\sigma_1^2, \dots, \sigma_d^2)$, $\sigma_i > 0$. Define $m_i(t) = (1-t) + t\sigma_i$, and the coefficients

$$\theta_{ij}(t) := \frac{\sigma_i \sigma_j}{\sigma_i + \sigma_j} \left[\frac{2t\sigma_i\sigma_j}{\sigma_i + \sigma_j}\big(m_i(t) + m_j(t)\big) - m_i(t)m_j(t) \right], \tag{44}$$

$$\kappa_i(t) := -\frac{\sigma_i^2}{m_i(t)^2} \sum_{j \neq i} \frac{\sigma_j(2t\sigma_j - m_j(t))}{(\sigma_i + \sigma_j)^2 m_j(t)}. \tag{45}$$

Proposition B.5 (Gaussian plug-in minibatch field). *For fixed d, positive-definite diagonal Σ as above, and fixed (x, t), the marginalized field (equation 43) satisfies, as $n \to \infty$,*

$$[v_{t,n}(x)]_i = \frac{\sigma_i - 1}{m_i(t)} x_i + \frac{1}{n} \left[\kappa_i(t) x_i + \frac{x_i}{m_i(t)^3} \sum_{j=1}^d \theta_{ij}(t) \frac{x_j^2}{m_j(t)^3} \right] + o(n^{-1}). \tag{46}$$

Corollary B.6 (Standardized endpoints). *If $\Sigma = \mathrm{Id}$, then*

$$v_{t,n}(x) = \frac{2t-1}{4n} \left(2\|x\|^2 - (d-1)\right) x + o(n^{-1}), \tag{47}$$

$$\partial_t v_{t,n}(x) = \frac{1}{2n} \left(2\|x\|^2 - (d-1)\right) x + o(n^{-1}). \tag{48}$$

Proposition B.7 (Independent Gaussian flow-matching field)**.** *For independent flow matching with* $p_0 = \mathcal{N}(0, \mathrm{Id})$ *and* $p_1 = \mathcal{N}(0, \mathrm{diag}(\sigma_1^2, \dots, \sigma_d^2))$*, define* $q_i(t) = (1-t)^2 + t^2\sigma_i^2$, $r_i(t) = t\sigma_i^2 - (1-t)$*. Then*

$$[v_t^{\mathrm{ind}}(x)]_i = \frac{r_i(t)}{q_i(t)} x_i, \tag{49}$$

$$[\partial_t v_t^{\mathrm{ind}}(x)]_i = \left[\frac{1+\sigma_i^2}{q_i(t)} - \frac{2r_i(t)^2}{q_i(t)^2}\right] x_i. \tag{50}$$

If $\Sigma = \mathrm{Id}$*, these expressions specialize to*

$$v_t^{\mathrm{ind}}(x) = \frac{2t-1}{(1-t)^2+t^2} x, \qquad \partial_t v_t^{\mathrm{ind}}(x) = \frac{4t(1-t)}{((1-t)^2+t^2)^2} x. \tag{51}$$

The population Gaussian OT field is $[v_t^\star(x)]_i = (\sigma_i - 1)x_i/m_i(t)$. Unlike the plug-in correction, the independent field in Proposition B.7 has no batch-size dependence. When $\Sigma = \mathrm{Id}$, population OT is the identity and has zero velocity and time derivative, whereas independent FM exhibits an order-one contraction–expansion (referred to in the main paper as the "yo-yo" effect) despite having identical endpoint distributions. Gaussian plug-in minibatching retains a residual version of this motion, but Corollary B.6 shows that its magnitude and Eulerian time derivative decay as $1/n$.

We next prove Proposition B.5 and its corollary; Proposition B.7 follows from the joint Gaussian conditioning calculation given afterward.

Setup. We keep d and Σ fixed as $n \to \infty$ and assume $n > d$. With $\varepsilon = n^{-1/2}$ and unbiased sample covariances, write the source and target parameters as

$$\widehat{\mu}_n^0 = \varepsilon h, \qquad \widehat{\Sigma}_n^0 = \mathrm{Id} + \varepsilon H, \tag{52}$$

$$\widehat{\mu}_n^1 = \varepsilon \Sigma^{1/2} h', \qquad \widehat{\Sigma}_n^1 = \Sigma + \varepsilon \Sigma^{1/2} H' \Sigma^{1/2}. \tag{53}$$

Here h, h' are independent standard Gaussian vectors, while H, H' are independent centered Wishart fluctuations and are independent of the sample means. Their leading second moments are

$$\mathbb{E}[H_{ij} H_{kl}] = \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk} + O(n^{-1}), \tag{54}$$

and likewise for H'. The $O(n^{-1})$ term in equation 54 contributes only beyond the order retained below.

The Monge map between the two fitted Gaussians is

$$T_n(x) = \widehat{\mu}_n^1 + A_n(x - \widehat{\mu}_n^0), \qquad A_n \widehat{\Sigma}_n^0 A_n = \widehat{\Sigma}_n^1, \qquad A_n \succ 0. \tag{55}$$

Expanding $A_n = \Sigma^{1/2} + \varepsilon K + \varepsilon^2 L + O_p(\varepsilon^3)$ and matching powers in the covariance identity gives

$$\Sigma^{1/2} K + K \Sigma^{1/2} = \Sigma^{1/2}(H' - H)\Sigma^{1/2}, \tag{56}$$

$$\Sigma^{1/2} L + L \Sigma^{1/2} = -K^2 - \Sigma^{1/2} H K - K H \Sigma^{1/2}. \tag{57}$$

Since Σ is diagonal, the first equation has the entrywise solution

$$K_{ij} = c_{ij}(H'_{ij} - H_{ij}), \qquad c_{ij} := \frac{\sigma_i \sigma_j}{\sigma_i + \sigma_j}. \tag{58}$$

For later use, define

$$M_t = (1-t)\mathrm{Id} + t\Sigma^{1/2}, \qquad W_t = M_t^{-1}. \tag{59}$$

The diagonal entries of M_t are the scalars $m_i(t)$. The population Gaussian OT field is

$$v_t^\star(x) = (\Sigma^{1/2} - \mathrm{Id}) W_t x, \qquad [v_t^\star(x)]_i = \frac{\sigma_i - 1}{m_i(t)} x_i. \tag{60}$$

Batch-conditional and marginalized fields. For a fixed pair of batch fits, set

$$X_t = (1-t)X_0 + tT_n(X_0), \qquad X_0 \sim \mathcal{N}(\widehat{\mu}_n^0, \widehat{\Sigma}_n^0). \tag{61}$$

Writing

$$B_{t,n} = (1-t)\mathrm{Id} + tA_n, \qquad \widehat{m}_{t,n} = (1-t)\widehat{\mu}_n^0 + t\widehat{\mu}_n^1, \tag{62}$$

the conditional path is Gaussian and its FM field is the affine map

$$\widehat{v}_{t,n}(x) = \widehat{\mu}_n^1 - \widehat{\mu}_n^0 + (A_n - \mathrm{Id})B_{t,n}^{-1}(x - \widehat{m}_{t,n}). \tag{63}$$

If

$$\alpha_t = (1-t)h + t\Sigma^{1/2}h', \qquad e_t = \Sigma^{1/2}W_t(h' - h), \tag{64}$$

then expansion of equation 63 yields

$$\widehat{v}_{t,n}(x) = v_t^\star(x) + \varepsilon v_t^{(1)}(x) + \varepsilon^2 v_t^{(2)}(x) + O_p(\varepsilon^3), \tag{65}$$

where

$$v_t^{(1)}(x) = W_t K W_t x + e_t, \tag{66}$$

$$v_t^{(2)}(x) = (W_t L W_t - tW_t K W_t K W_t)\, x - W_t K W_t \alpha_t. \tag{67}$$

The marginalization in equation 43 weights each batch-conditional field by $\widehat{p}_{t,n}(x)$. We therefore expand this conditional path density about $p_t^\star = \mathcal{N}(0, M_t^2)$.

Conditional on the fitted batches, the covariance of X_t has the expansion

$$B_{t,n}\widehat{\Sigma}_n^0 B_{t,n} = M_t^2 + \varepsilon G_t + O_p(\varepsilon^2), \qquad G_t = M_t H M_t + t(KM_t + M_t K). \tag{68}$$

Consequently,

$$\widehat{p}_{t,n}(x) = p_t^\star(x)\left[1 + \varepsilon \ell_t(x) + O_p(\varepsilon^2)\right], \tag{69}$$

with

$$\ell_t(x) = \alpha_t^\mathsf{T} W_t^2 x + \frac{1}{2}x^\mathsf{T} W_t^2 G_t W_t^2 x - \frac{1}{2}\operatorname{tr}(W_t^2 G_t). \tag{70}$$

Since α_t and G_t are centered, $\mathbb{E}[\ell_t(x)] = 0$; similarly, the centered fluctuations K and e_t give $\mathbb{E}[v_t^{(1)}(x)] = 0$. Expanding the density-weighted ratio equation 43 gives

$$v_{t,n}(x) = v_t^\star(x) + \frac{1}{n}\left\{\mathbb{E}[v_t^{(2)}(x)] + \mathbb{E}[\ell_t(x)v_t^{(1)}(x)]\right\} + o(n^{-1}). \tag{71}$$

The second-order density fluctuation cancels between the numerator and denominator. The second expectation in equation 71 is precisely the posterior density-weighting term omitted by an unweighted batch average.

Closed-form coefficient. We now evaluate the coefficient of the order-$1/n$ term in equation 71. In the posterior-weighting term $\mathbb{E}[\ell_t(x)v_t^{(1)}(x)]$, the covariance-dependent part of ℓ_t is linear in G_t, while the matrix-dependent part of $v_t^{(1)}$ is linear in K. Thus, using equation 58 and the Wishart moments in equation 54, their required covariance contraction is

$$\mathbb{E}[[G_t]_{ij}K_{kl}] = \theta_{ij}(t)(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}) + O(n^{-1}). \tag{72}$$

Combining this contraction with the second Sylvester equation in equation 57, which determines the mean of $v_t^{(2)}$, gives

$$\mathbb{E}[v_t^{(2)}(x)]_i = \beta_i(t)x_i + O(n^{-1}), \tag{73}$$

$$\mathbb{E}[\ell_t(x)v_t^{(1)}(x)]_i = \eta_i(t)x_i + \frac{x_i}{m_i(t)^3}\sum_{j=1}^d \theta_{ij}(t)\frac{x_j^2}{m_j(t)^3} - \frac{\theta_{ii}(t)}{m_i(t)^4}x_i + O(n^{-1}), \tag{74}$$

where

$$\beta_i(t) = -\frac{\sigma_i^2}{m_i(t)^2}\sum_{j=1}^d (1+\delta_{ij})\frac{\sigma_j(2t\sigma_j - m_j(t))}{(\sigma_i+\sigma_j)^2 m_j(t)}, \tag{75}$$

$$\eta_i(t) = \frac{\sigma_i(2t\sigma_i - m_i(t))}{m_i(t)^3}. \tag{76}$$

The diagonal identities

$$\beta_i(t) = \kappa_i(t) - \frac{1}{2}\eta_i(t), \qquad \frac{\theta_{ii}(t)}{m_i(t)^4} = \frac{1}{2}\eta_i(t) \tag{77}$$

cancel all diagonal linear contributions. Inserting the remaining terms into equation 71 proves equation 46.

Standardized specialization. When $\Sigma = \mathrm{Id}$, we have $\sigma_i = m_i(t) = 1$ and $2t\sigma_i - m_i(t) = 2t - 1$. Equations equation 44 and equation 45 reduce to

$$\theta_{ij}(t) = \frac{2t-1}{2}, \qquad \kappa_i(t) = -\frac{(d-1)(2t-1)}{4}. \tag{78}$$

Substituting these expressions into equation 46 gives equation 47; differentiating at fixed x gives equation 48, proving Corollary B.6. At $t = 1/2$ the field is in fact exactly zero for every n: exchanging the i.i.d. fitted batches preserves the midpoint and reverses the displacement.

In one dimension, the off-diagonal sum in equation 45 is empty. Writing $m(t) = (1-t) + t\sigma$, the general result becomes

$$v_{t,n}(x) = \frac{\sigma - 1}{m(t)}x + \frac{1}{n}\frac{\sigma(2t\sigma - m(t))}{2m(t)^5}x^3 + o(n^{-1}). \tag{79}$$

Hence the entire linear order-$1/n$ correction cancels in one dimension.

Comparison with the independent coupling.

Proof of Proposition B.7. For independent $X_0 \sim \mathcal{N}(0, \mathrm{Id})$ and $X_1 \sim \mathcal{N}(0, \Sigma)$, let $X_t = (1-t)X_0 + tX_1$ and $U = X_1 - X_0$. Joint Gaussian conditioning yields

$$v_t^{\mathrm{ind}}(x) = \mathbb{E}[U \mid X_t = x] = (t\Sigma - (1-t)\mathrm{Id})\left((1-t)^2\mathrm{Id} + t^2\Sigma\right)^{-1} x. \tag{80}$$

Diagonalizing this expression and differentiating at fixed x gives equation 49 and equation 50; setting $\Sigma = \mathrm{Id}$ recovers equation 51. □

The independent field is also exactly the field induced by singleton discrete minibatch OT, because the sole source and target observations must be paired. This does not identify it with Gaussian plug-in OT at $n = 1$, for which an unbiased sample covariance is undefined.

Acceleration along flow trajectories. For a time-dependent velocity field v_t, define its material acceleration by

$$a_t[v](x) := \partial_t v_t(x) + D_x v_t(x) v_t(x). \tag{81}$$

This is the acceleration of a trajectory satisfying $\dot{x}_t = v_t(x_t)$. In particular, the population Gaussian OT field follows straight displacement trajectories and therefore satisfies $a_t[v^\star](x) = 0$.

For the plug-in field, let

$$b_i(t) := \frac{\sigma_i - 1}{m_i(t)} \tag{82}$$

and write overdots for time derivatives. Expanding equation 81 using equation 46 gives

$$\begin{aligned}[a_t[v_n](x)]_i = \frac{x_i}{n}\Bigg[&\dot{\kappa}_i(t) + 2b_i(t)\kappa_i(t) \\ &+ \sum_{j=1}^{d} \frac{\dot{\theta}_{ij}(t) - (b_i(t) + b_j(t))\theta_{ij}(t)}{m_i(t)^3 m_j(t)^3} x_j^2\Bigg] + o(n^{-1}).\end{aligned} \tag{83}$$

For independent flow matching, the corresponding expression is exact:

$$[a_t[v^{\mathrm{ind}}](x)]_i = \frac{\sigma_i^2}{q_i(t)^2}x_i. \tag{84}$$

When $\Sigma = \mathrm{Id}$, write $q(t) = (1-t)^2 + t^2$. The three acceleration fields simplify to

$$a_t[v^\star](x) = 0, \qquad a_t[v_n](x) = \frac{2\|x\|^2 - (d-1)}{2n}x + o(n^{-1}), \qquad a_t[v^{\mathrm{ind}}](x) = \frac{x}{q(t)^2}. \tag{85}$$

The plug-in material acceleration agrees with the Eulerian derivative in equation 48 to order $1/n$, because the convective term is of order $1/n^2$ in the standardized case.

Finally, let x_t denote the flow trajectory initialized at x_0, and define its integrated squared acceleration by

$$\mathcal{A}[v; x_0] := \int_0^1 \|a_t[v](x_t)\|^2 \, \mathrm{d}t. \tag{86}$$

For standardized endpoints, $x_t^\star = x_0$, $x_t^{\mathrm{ind}} = \sqrt{q(t)}\, x_0$, and $x_{t,n}^{\mathrm{plug}} = x_0 + O(n^{-1})$. Consequently,

$$\mathcal{A}[v^\star; x_0] = 0, \tag{87}$$

$$\mathcal{A}[v_n; x_0] = \frac{\|x_0\|^2 \left(2\|x_0\|^2 - (d-1)\right)^2}{4n^2} + o(n^{-2}), \tag{88}$$

$$\mathcal{A}[v^{\mathrm{ind}}; x_0] = \left(2 + \frac{3\pi}{4}\right) \|x_0\|^2. \tag{89}$$

Thus the leading plug-in acceleration energy is nonincreasing in n and decays as n^{-2}, while independent FM incurs order-one acceleration and exact OT incurs none. For every fixed $x_0 \neq 0$, independent FM therefore has the largest of these three acceleration energies for all sufficiently large n. This comparison concerns the controlled large-n expansion and does not assert exact monotonicity of the finite-n Gaussian plug-in field.

If the initial condition is itself random, $X_0 \sim \mathcal{N}(0, \mathrm{Id})$, then $R := \|X_0\|^2 \sim \chi_d^2$, with

$$\mathbb{E}[R] = d, \qquad \mathbb{E}[R^2] = d(d+2), \qquad \mathbb{E}[R^3] = d(d+2)(d+4). \tag{90}$$

Consequently, assuming the plug-in expansion above holds in L^1 with respect to X_0,

$$\mathbb{E}[\mathcal{A}[v^\star; X_0]] = 0, \tag{91}$$

$$\mathbb{E}[\mathcal{A}[v_n; X_0]] = \frac{d\left(d^2 + 18d + 41\right)}{4n^2} + o(n^{-2}), \tag{92}$$

$$\mathbb{E}\big[\mathcal{A}[v^{\mathrm{ind}}; X_0]\big] = \left(2 + \frac{3\pi}{4}\right) d. \tag{93}$$

Thus, at fixed dimension, the expected plug-in acceleration energy again decays as n^{-2}, whereas the independent-coupling energy remains order one in n.

C IMAGENET-256

C.1 IMPLEMENTATION DETAILS

Training details. In tables 6 and 7, we show training configurations across our SiT (Ma et al., 2024) and DMF (Lee et al., 2026) experiments for B/2, L/2 and XL/2 model scales. All ImageNet-256 experiments use mini-batch setting of 256 batch size. DMF (Lee et al., 2026) conditions the encoder on time t and adds additional time r to condition its decoder, turning a flow matching model into a flow map. Following their set-up, we also select `logit`-normal distribution to sample (t, r) pairs using the time proposal parameters in table 6. DMF is trained by finetuning a pretrained SiT model via self-distillation meanflow objective (Geng et al., 2025) with a diagonal split of 0.5.

Table 6: ImageNet-256 implementation details across model scales.

	B/2	L/2	XL/2
Backbone			
Resolution	256×256	256×256	256×256
Params (M)	130	458	675
FLOPS (G)	23.1	80.7	118.6
Hidden dim.	768	1024	1152
Heads	12	16	16
Patch size	2×2	2×2	2×2
Sequence length	256	256	256
Layers	12	24	28
Flow Matching (SiT) (Ma et al., 2024)			
Training iterations	800K	800K	800K
Epochs	160	160	160
Class dropout probability	0.1	0.1	0.1
Flow Map (DMF) (Lee et al., 2026)			
DMF depth	8	18	20
Training iterations	400K	400K	400K
Epochs	80	80	80
Class dropout probability		0.1	
Time proposal μ_{FM}		0.0	
Time proposal $(\mu^{(1)}_{\text{MF}}, \mu^{(2)}_{\text{MF}})$		$(0.4, -1.2)$	
Model guidance scale ω	0.5	0.6	0.6
Guidance interval	$[0.0, 1.0]$	$[0.0, 0.7]$	$[0.0, 0.7]$

Table 7: Optimization hyperparameters, shared across all model scales and both training stages.

Optimizer	AdamW
Batch size	256
Learning rate	1e-4
Adam (β_1, β_2)	$(0.9, 0.95)$
Adam ϵ	1e-8
Weight decay	0.0
EMA decay rate	0.9999

Compute budget. In table 8, we show cost per optimizer step for each of the coupling plans. We observe that class-conditional OT and **GT** add a modest cost per optimizer step, with **GT** being slightly higher than class-conditional due to not being restricted per class.

Choice of β for class-conditional OT. We choose β such that it prevents any cross-class coupling, thereby computing transport plan within each class following (Cheng & Schwing, 2025). In practice, we define $\beta = \beta_m \cdot \frac{1}{B^2} \sum_{i=1}^{B} \sum_{j=1}^{B} \|x_i - x_j\|_2^2$. We sweep β_m and show analysis in figure 10, choosing $\beta_m = 10$.

Table 8: Compute cost per optimizer step of SiT-B/2 training on ImageNet-256 for the three coupling plans. The second column is the increase over the independent coupling.

Coupling	ms / step	↑ cost
Independent	155.5	–
Class-cond. OT	157.9	↑ 1.6%
GT	160.4	↑ 3.1%

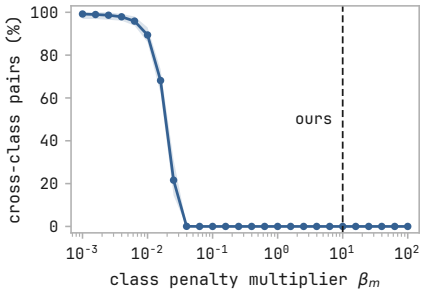

Figure 10: Sweeping β_m for class-conditional OT.

Text-conditioning for ImageNet. Every ImageNet-1k image is paired with the captions taken from Visual Layer dataset. Each caption is encoded once with the frozen DFN5B CLIP ViT-H/14 text tower into its pooled, ℓ_2-normalised 1024-d text embedding $c \in \mathbb{R}^{1024}$.

C.2 Additional results on ImageNet-256

GT preserves mode stability. Figure 11 sweeps w for fixed noise and class. Under the independent coupling, increasing guidance results in abrupt change in composition and pose, e.g. from a portrait to a full-body view. Under **GT** the object remains stable for the full range of w.

Guidance interval tuning. We show that **GT** achieves strongest results overall with applied guidance interval tuning across B/2, L/2 and XL/2 scales in table 10.

Table 9: FID↓ and FD_{DINOv2}↓ vs. guidance scale w for independent, class-conditional OT and **GT** couplings across SiT scales (64 Euler steps), under the full guidance interval $[0, 1]$.

SiT-B/2												
Coupling	Metric	w=1.0	1.5	2.0	2.5	3.0	3.5	4.0	5.0	6.0	8.0	10.0
Independent	FID	26.44	6.68	5.16	7.93	11.04	13.65	15.76	18.63	20.42	22.40	23.28
	FD_{DINOv2}	605.11	337.16	212.66	159.87	139.10	132.46	132.27	139.64	149.79	169.51	189.02
Class-cond. OT	FID	26.72	6.70	5.08	7.81	10.90	13.47	15.59	18.44	20.28	22.25	23.12
	FD_{DINOv2}	607.86	340.46	215.04	160.77	139.17	131.64	131.41	138.27	148.37	168.91	189.17
GT	FID	31.16	8.13	4.15	5.88	8.72	11.32	13.41	16.51	18.62	20.81	21.62
	FD_{DINOv2}	636.79	365.26	226.75	161.94	132.62	120.25	116.36	119.84	128.33	148.82	171.99

SiT-L/2									
Coupling	Metric	w=1.0	1.5	1.75	2.0	2.5	3.0	3.5	4.0
Independent	FID	13.85	2.95	3.86	5.85	10.10	13.53	16.03	17.89
	FD_{DINOv2}	360.44	154.77	115.56	95.16	83.28	86.46	94.58	103.47
Class-cond. OT	FID	13.96	2.98	3.84	5.81	10.02	13.42	15.93	17.79
	FD_{DINOv2}	360.38	154.83	115.16	95.01	83.41	86.38	94.01	102.64
GT	FID	17.72	2.96	2.69	3.88	7.38	10.66	13.25	15.17
	FD_{DINOv2}	392.20	170.86	123.70	97.58	76.33	73.48	77.32	83.28

SiT-XL/2										
Coupling	Metric	w=1.0	1.5	1.75	2.0	2.25	2.5	3.0	3.5	4.0
Independent	FID	12.58	2.78	3.98	6.10	8.32	10.41	13.76	16.14	17.87
	FD_{DINOv2}	328.83	135.91	101.78	85.35	78.40	77.76	83.29	92.02	101.67
GT	FID	16.11	2.63	2.59	3.87	5.65	7.47	10.70	13.16	15.09
	FD_{DINOv2}	357.67	150.28	108.56	85.80	73.95	69.01	68.44	73.27	80.05

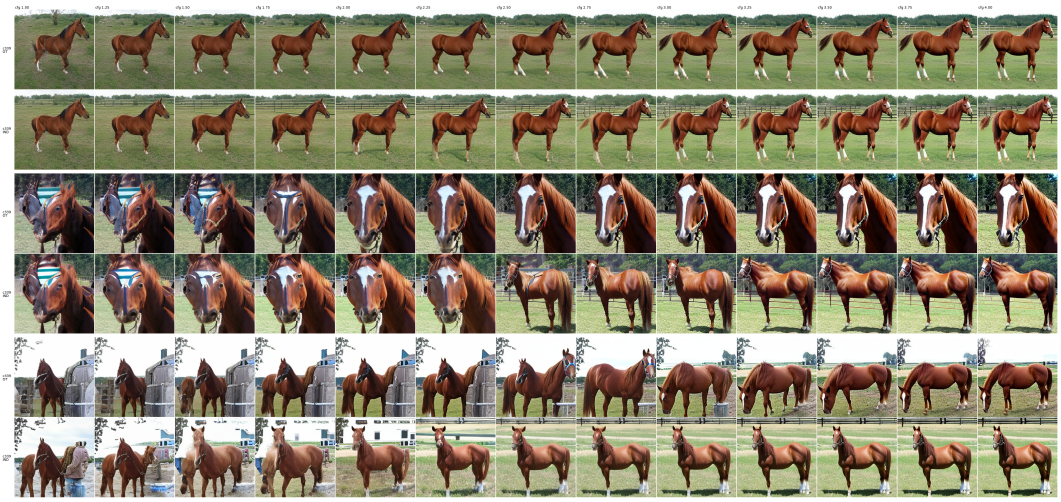

Figure 11: SiT-XL/2 curated samples for class 339 `sorrel` across guidance scales w and coupling choices. Under **GT** the sample object characteristics remain stable, while for independent and class-conditional object changes its composition, pose, background at higher w.

D SINGLE-CELL DATASETS

PBMC3K. 2,638 peripheral blood mononuclear cells from a healthy donor across 8 cell types.

Dentate gyrus. 18,213 cells from the developing mouse hippocampus (La Manno et al., 2018), annotated with 14 cell types.

HLCA. 584,944 human lung cells from 486 individuals across 49 datasets (Sikkema et al., 2023), annotated with 50 cell types.

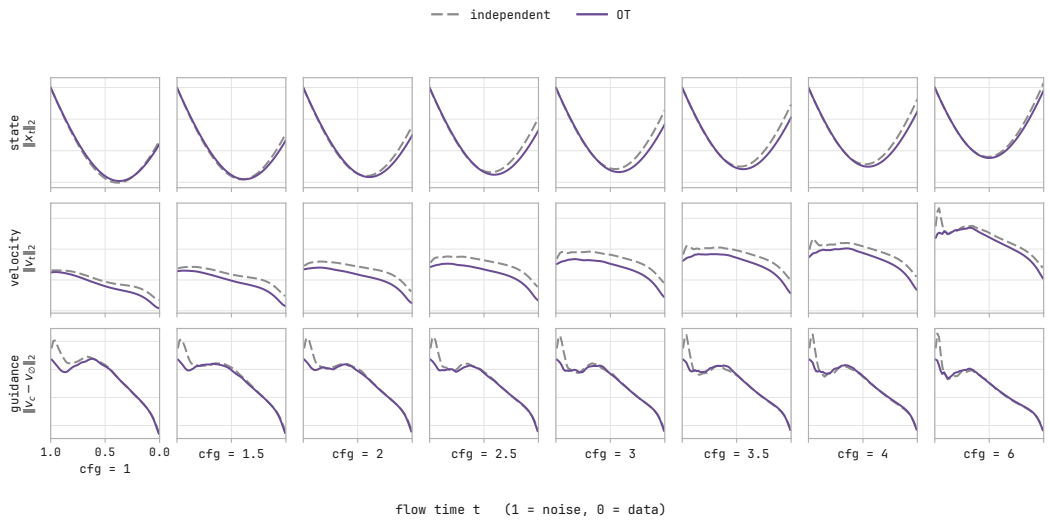

Figure 12: Measuring yo-yo effect and prediction gap across trajectory using SiT-XL/2 (64 steps).

Table 10: FID↓ across model scales and couplings, under the full guidance interval and tuned interval $[0, 0.7]$.

			Full					GI [0, 0.7]				
Model	Coupling	w=1.0	1.5	2.0	2.5	3.0	4.0	1.5	2.0	2.5	3.0	4.0
SiT-B/2	Independent	**26.44**	**6.68**	5.16	7.93	11.04	15.76	**11.93**	**5.95**	3.95	3.72	5.07
	Class-cond. OT	26.72	6.70	5.08	7.81	10.90	15.59	12.13	6.00	**3.92**	3.68	5.00
	GT	31.16	8.13	**4.15**	**5.88**	**8.72**	**13.41**	15.63	8.03	4.71	**3.52**	**3.70**
SiT-L/2	Independent	**13.85**	**2.95**	5.85	10.10	13.53	17.89	**4.53**	**2.35**	2.62	3.65	6.00
	GT	17.72	2.96	**3.88**	**7.38**	**10.66**	**15.17**	6.64	2.96	**2.17**	**2.48**	**4.02**
SiT-XL/2	Independent	**12.58**	2.78	6.10	10.41	13.76	17.87	**3.96**	**2.16**	2.63	3.73	6.04
	GT	16.11	**2.63**	**3.87**	**7.47**	**10.70**	**15.09**	5.89	2.61	**2.01**	**2.40**	**3.91**

Implementation Details. We add couplings on top of the uni-modal generation set-up described in Palma et al. (2025). Within the single-cell dataset, CFGen splits data into 90% training and 10% validation. Evaluation is performed on the corresponding held-out test data. Since CFGen (Palma et al., 2025) considers interpolant $x_t = tx_1 + \sigma_t\epsilon$, $\epsilon \sim \mathcal{N}(0, I_d)$, we implement baseline with linear interpolant used in our image experiments (CFGen-linear) which trains CFGen with independent coupling. We then compare both baselines to **GT** transport. Our evaluation script uses $0, 1, 2$ seeds.

CFGen splits its training into two stages: 1) encoding single-cell data with an RNA autoencoder and 2) training a conditional flow matching model in latent space. We show hyperparameter choices for training and backbone model in tables 11 and 12. Note that PBMC3k and Dentate gyrus use `resnet small` and HLCA uses `resnet big`.

Table 11: Training configuration for uni-modal generation with CFGen.

Parameter	Autoencoder	Latent Flow Matching
Batch size	256	256 (64 for PBMC3K)
Epochs	300	1500
Optimizer	AdamW	AdamW
Learning rate	1×10^{-3}	1×10^{-4}
Weight decay	1×10^{-5}	1×10^{-6}
Gradient clipping	1.0	1.0
Train/validation split	90% / 10%	90% / 10%

Table 12: CFGen ResNet Backbone.

Hyperparameter	ResNet Small	ResNet Big
Hidden dimension	32	64
Residual blocks	3	3
Embedding dimension	20	100
Dropout probability	0.0	0.0
Condition dropout probability	0.2	0.2

E Additional Background

E.1 Classifier and classifier-free guidance

Classifier Guidance. By applying Bayes' Rule we can write

$$p_t(x \mid c) = \frac{p_t(c \mid x) p_t(x)}{p_t(c)} \tag{94}$$

Taking a logarithm of each side we get

$$\log p_t(x \mid c) = \log p_t(c \mid x) + \log p_t(x) - \log p_t(c) \tag{95}$$

By applying ∇_x to both sides we get

$$\nabla_x \log p_t(x \mid c) = \nabla_x \log p_t(c \mid x) + \nabla_x \log p_t(x) - \cancel{\nabla_x \log p_t(c)} \tag{96}$$

leading to

$$\nabla_x \log p_t(x \mid c) = \nabla_x \log p_t(c \mid x) + \nabla_x \log p_t(x) \tag{97}$$

Using the conversion formula for score we can further write this as

$$v_c(x, t, c) = v_u(x_t, t) + a_t \nabla_x \log p_t(c \mid x) \tag{98}$$

Classifier guidance constructs a velocity field by enhancing classifier $p_t(c \mid x)$ with scaling factor w which we refer to as guidance scale, yielding

$$\boxed{v_c^{CF}(x, t, c) = v_u(x_t, t) + w a_t \nabla_x \log p_t(c \mid x)} \tag{99}$$

where $a_t = \frac{1-t}{t}$ and $b_t = \frac{1}{t}$.

The main issue with using classifier guidance is the need to train $p_t(c \mid x)$ classifier, which drastically increases training compute. In the following section, we derive classifier-free guidance.

Classifier-free Guidance. Classifier-free guidance takes a step further to construct a guided field that does not require additional training of classifier $p_t(c \mid x)$.

We can add $b_t x_t$ and subtract from the RHS of equation 99

$$\begin{aligned}
v^{\mathrm{CFG}}(x_t, t, c) &= v_u(x_t, t) + w\, a_t \nabla_x \log p_t(c \mid x_t) \\
&= v_u(x_t, t) + w\, a_t \nabla_x \log \frac{p_t(x_t \mid c)\, p_t(c)}{p_t(x_t)} \\
&= v_u(x_t, t) + w \big[a_t \nabla_x \log p_t(x_t \mid c) + \cancel{a_t \nabla_x \log p_t(c)} - a_t \nabla_x \log p_t(x_t) \big] \\
&= v_u(x_t, t) + w \big[\underbrace{a_t \nabla_x \log p_t(x_t \mid c) + b_t x_t}_{v_c(x_t, t, c)} - \underbrace{\big(a_t \nabla_x \log p_t(x_t) + b_t x_t \big)}_{v_u(x_t, t)} \big]
\end{aligned} \tag{100}$$

which leads to

$$\boxed{v_c^{CFG}(x, t, c) = v_u(x_t, t) + w \underbrace{\big(v_c(x, t, c) - v_u(x, t) \big)}_{g(x, t, c)}} \tag{101}$$

F GT UNCURATED SAMPLES

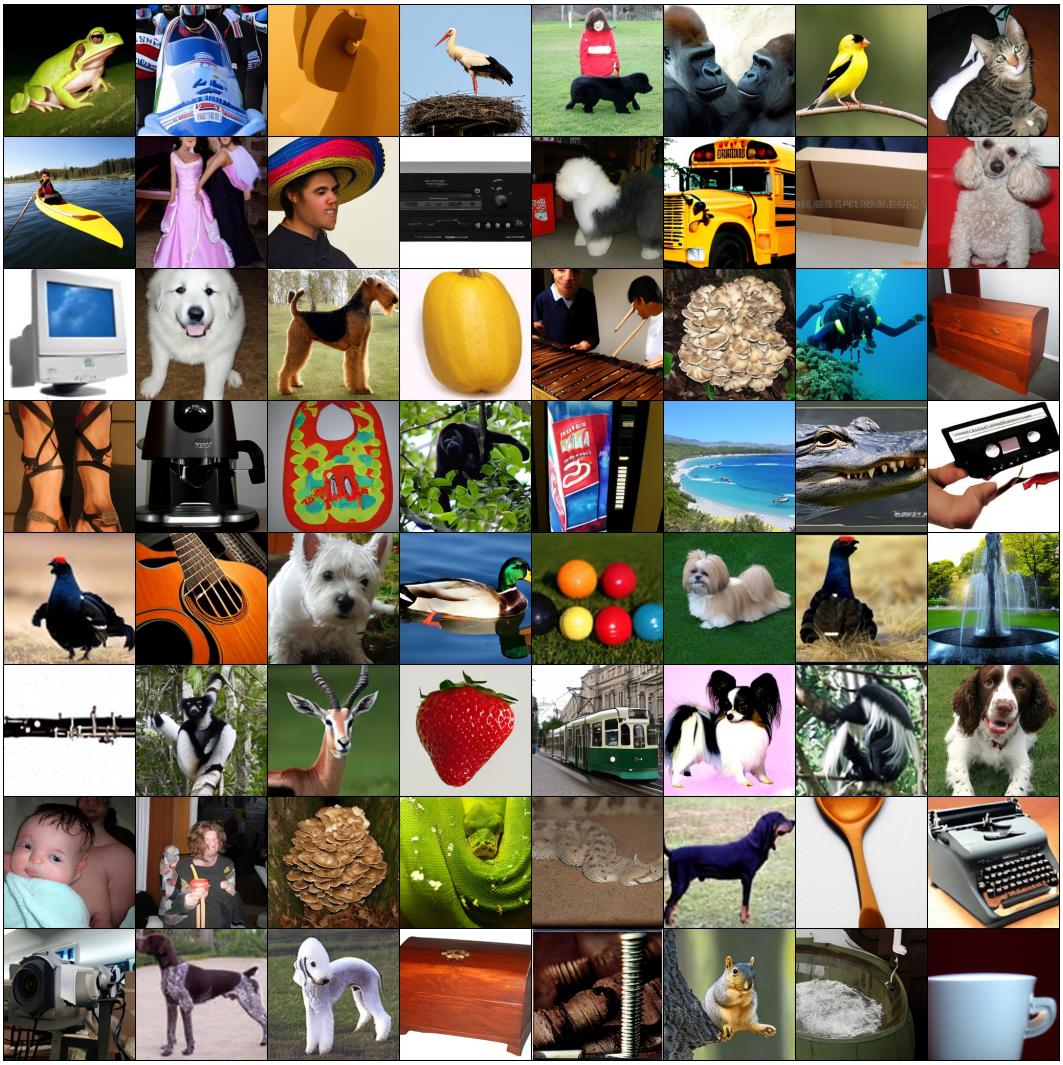

Figure 13: Samples generated with SiT-L/2, Euler (16 steps), $w = 4.0$

G CLASS-CONDITIONAL OT UNCURATED SAMPLES

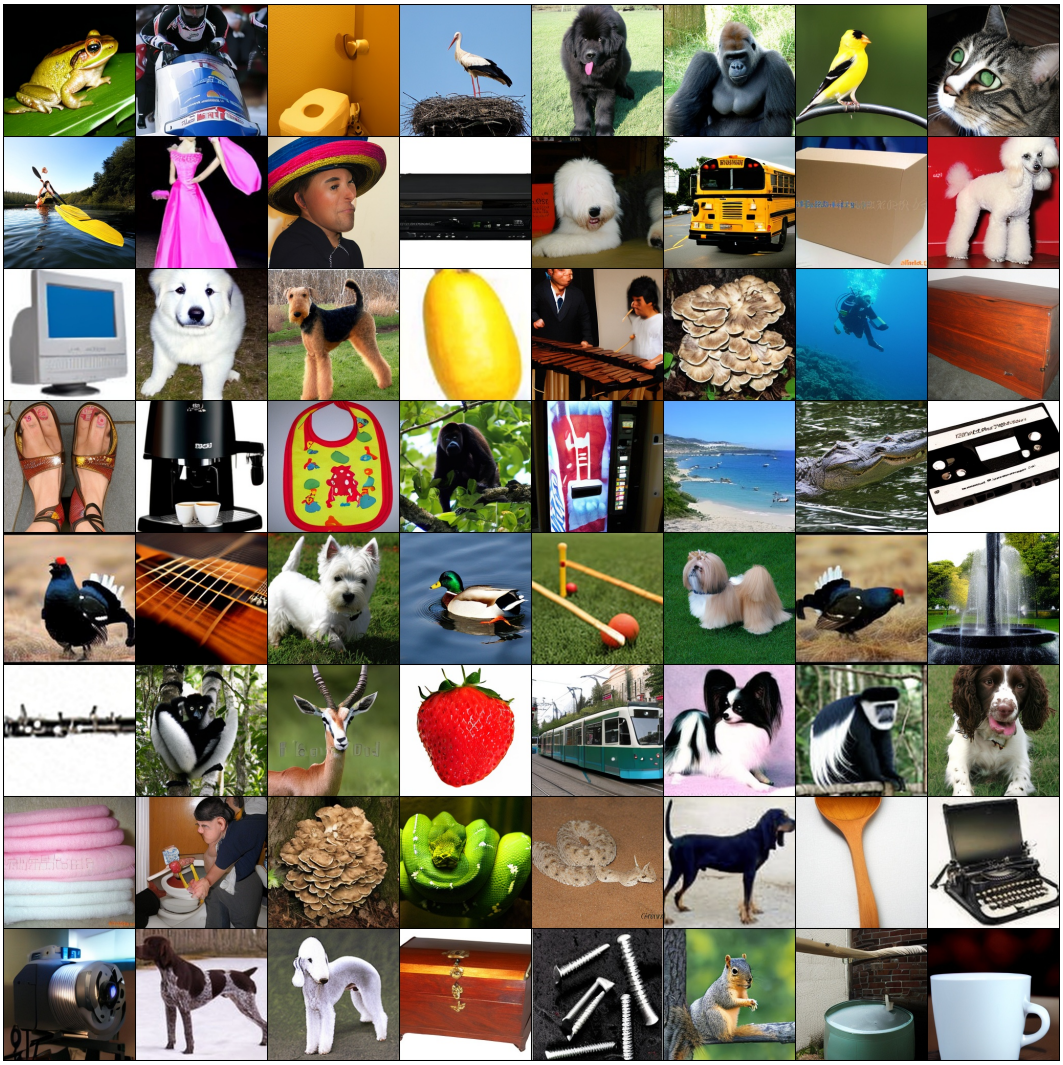

Figure 14: Samples generated with SiT-L/2, Euler (16 steps), $w = 4.0$

H INDEPENDENT COUPLING UNCURATED SAMPLES

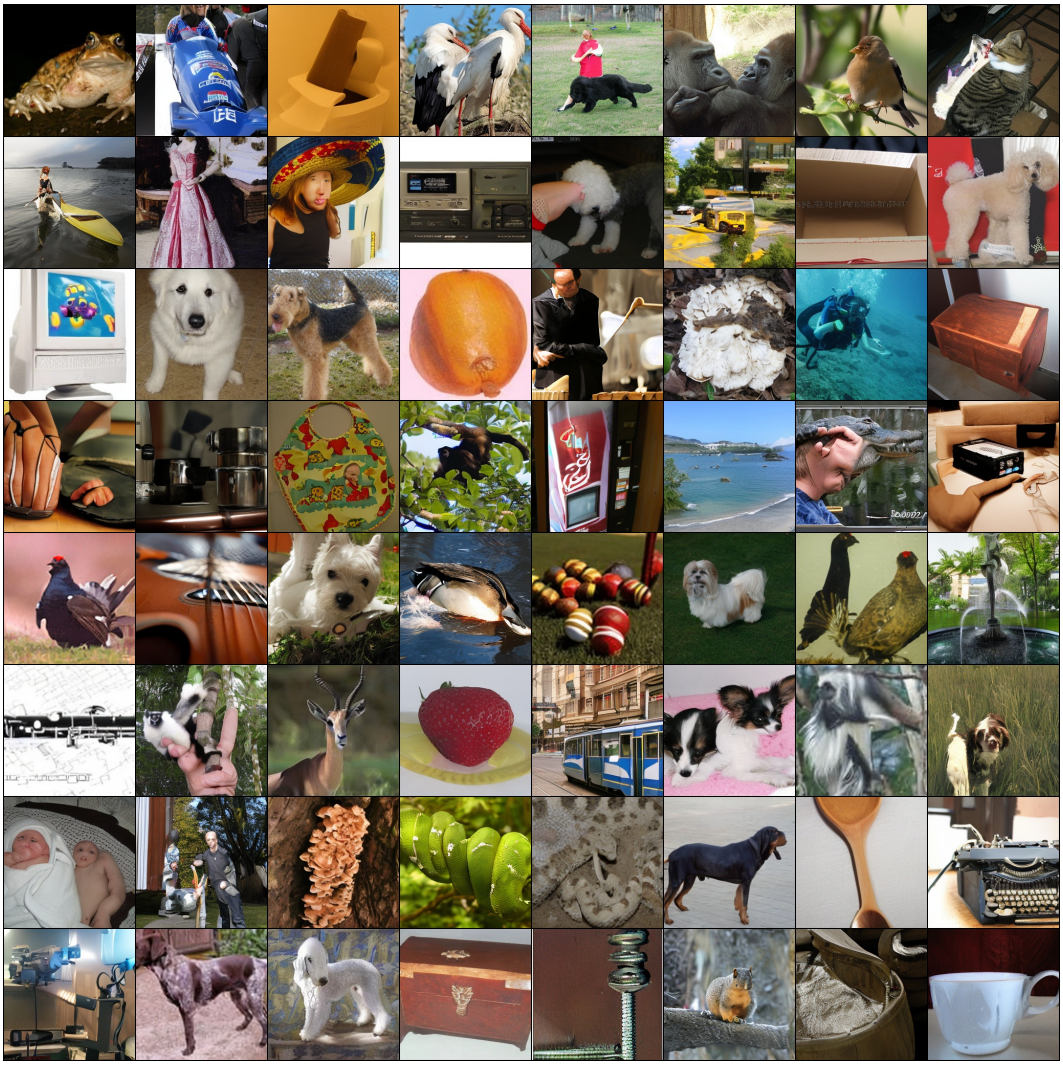

Figure 15: Samples generated with SiT-L/2, Euler (16 steps), $w = 4.0$

I TEXT-CONDITIONED IMAGENET-256 UNCURATED SAMPLES

We show samples for the following prompts read left to right:{two mittens with colorful yarn on a bed, a white keyboard with a small white keypad, a sink with a white marble top and a black base, a green and purple flower with a black center, three monkeys sitting on a wooden bench, a black and white photo of a stethoscope, a dog is looking at ducks in a pond, a woman feeding her baby with a bottle, a bowl of food, a dog sitting on the grass with a red building in the background, a rocky cliff, a snake is laying on top of hay in a cage, a plate of food on a table, a white monkey hanging on a wooden pole, casio dg-2000 boombox, a dog laying in the grass, a bathroom with a toilet and a bathtub with gummy bears, a chameleon is sitting on a branch with green leaves, a group of masks with different colors and designs, a monkey sitting in a tree with leaves, a person walking along a road near a body of water, a wolf laying on the ground, a black and white photo of a wheel, a large dam with a large waterfall in front of it, a lab coat with a picture of a man in a lab coat, a female swimmer in the pool with a yellow cap, a brown and white dog with a collar on, a plate with shrimp and bacon, a pair of scissors on a wooden table, a bird with a long beak sitting on a branch, a spider sits on its web in the sun, a tree with a branch, a group of people standing in a line, a pool table in an empty building with a green light, panasonic pd-wg-g1, a meerkat standing on a rock in a zoo, a small hamster sleeping in a person's hand, a woman looking at a dinosaur in a museum, a knitting dish cloth and a knitting needle, a stone wall in the middle of a yard, a white and yellow sea slug on a coral reef, a large cicada sitting on a person's hand, two women in kimono, a vintage sewing machine sitting in the grass, a couple standing in front of a yurt, a man lifting a barbell on a competition stage, a circular clock with a white circle in the middle, a military vehicle with a gun mounted on top, three graduates pose for a photo in blue graduation gowns, two cars driving on a race track, a woman holding a large fruit, a large black and white whale with its tail out in the water, a wall of banjos hanging on a wall, a bowl of guacamole with a tortilla chip on top, a close up of a dog with a collar, two dogs are standing in the dirt, a dock with a concrete wall and palm trees, a plate of mashed potatoes, a pair of sunglasses on the ground, a large building with many people walking around it, a small orange fish in a bowl, a close up of a metal fan with a metal cover, three monkeys sitting on a log, a skillet with food in it}

I.1 GT TEXT-CONDITIONED IMAGENET-256

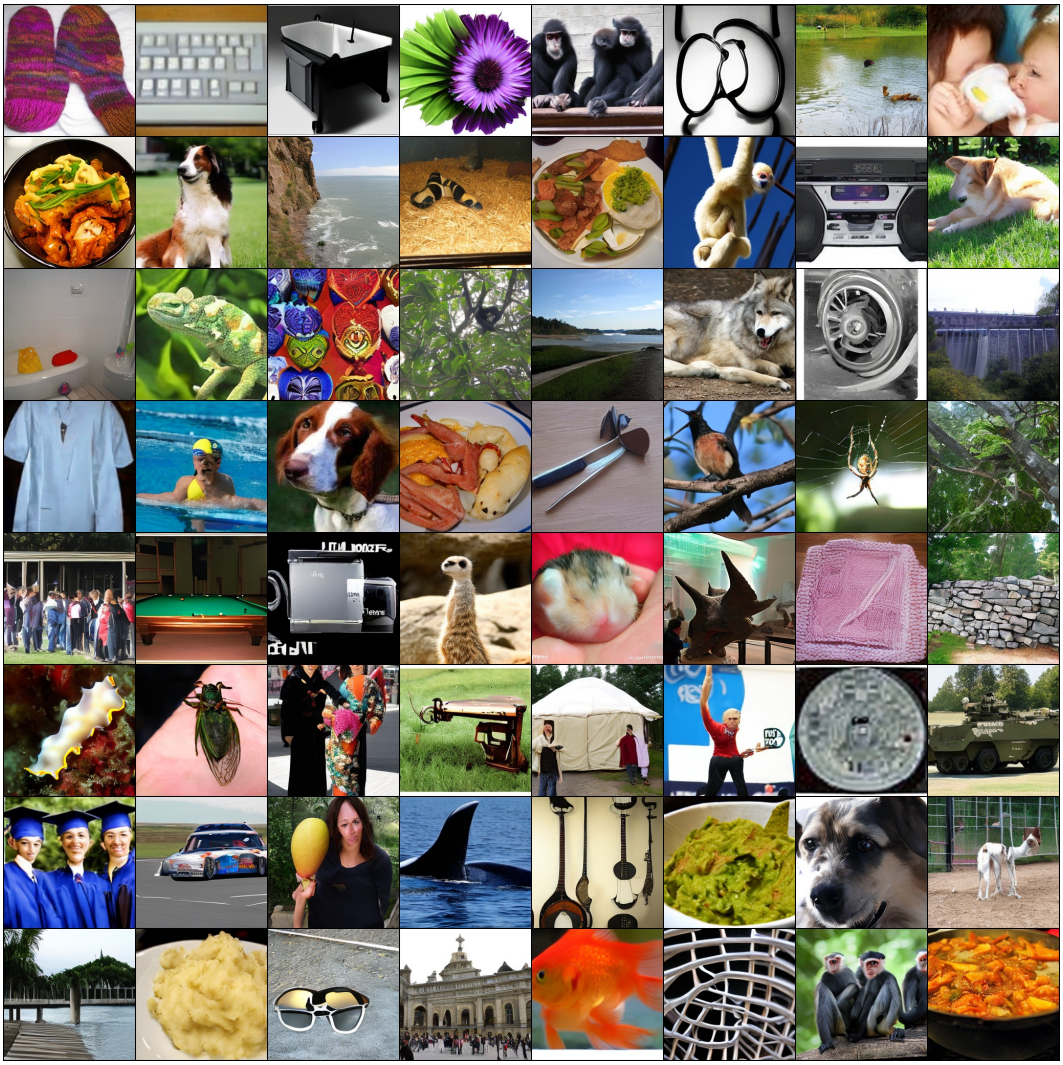

Figure 16: Samples generated with SiT-B/2, Euler (16 steps), $w = 4.0$

I.2 INDEPENDENT TEXT-CONDITIONED IMAGENET-256



Figure 17: Samples generated with SiT-B/2, Euler (16 steps), $w = 4.0$