INFERENCE-TIME ATTRIBUTE DISTRIBUTION ALIGNMENT FOR UNCONDITIONAL DIFFUSION

Hao Luan[1] **See-Kiong Ng**[1,2] **Chun Kai Ling**[1]
[1]School of Computing, National University of Singapore
[2]Institute of Data Science, National University of Singapore
haoluan@comp.nus.edu.sg, {seekiong, chunkail}@nus.edu.sg

ABSTRACT

Controllable generation is crucial to real-world applications of diffusion models. However, many applications require control beyond individual samples: the distribution of semantic attributes (e.g., proportions of styles, objects, or demographics) over generated samples should match user-specified targets that may change at test time. We formalize this setting as the inference-time attribute distributional alignment problem for pretrained diffusion models. To address this, we cast the attribute distributional alignment as an optimal control problem over the reverse diffusion process, viewing the process as the rollout of a dynamical system and augmenting it with an additive, time-dependent perturbation as control. We solve for the perturbations using an optimal-control-based algorithm to optimize a differentiable distribution-matching objective while penalizing control effort to preserve data fidelity. Experiment results in image generation demonstrate that our proposed plug-and-play approach can better align attribute distributions to diverse test-time targets compared to baselines, without retraining or fine-tuning the pretrained diffusion model.[1]

1 INTRODUCTION

Diffusion models are powerful generative models that have achieved remarkable success across a wide range of domains, including images (Ho et al., 2020; Rombach et al., 2022), video (Ho et al., 2022; Bar-Tal et al., 2024; Hayakawa et al., 2025), graphs (Niu et al., 2020; Madeira et al., 2024; Luan et al., 2025b), and robotics (Janner et al., 2022; Chi et al., 2023; Feng et al., 2025; Carvalho et al., 2025), *etc.*. A key advantage of diffusion-based generative models is their amenability to controllable generation through guided generation (Dhariwal & Nichol, 2021; Ye et al., 2024), constrained generation (Fishman et al., 2023; Feng et al., 2024; Zampini et al., 2025; Liang et al., 2025), or joint generation (Ruan et al., 2023; Zeng et al., 2024; Luan et al., 2025a; Hao et al., 2025), *etc.*

In many controllable generation applications, however, the goal is not merely that each *individual* sample satisfies a condition, but that the *population* exhibits a desired *distribution* over some *abstract attributes*. Such attributes may be styles or semantic categories for images, or for robotic systems, higher-level behaviors exhibited by a trajectory. Controlling the *distribution* of such attributes is essential in many real-world scenarios. For example, fairness objectives in human face generation naturally require balancing demographic attributes (*e.g.*, gender, race, age) toward a uniform distribution (Choi et al., 2020; Parihar et al., 2024; Kang et al., 2025). In many scenarios, however, the desired attribute distribution is not known before deployment and can change over time. This makes retraining or fine-tuning the model for each new target impractical, and motivates inference-time methods that adapt fixed pretrained models to user-specified distributions.

We refer to this goal as the test-time diffusion *attribute distributional alignment* (DADA) problem: given a pretrained diffusion model and a target *attribute distribution* specified *at test time*, generate samples whose attribute distribution matches the target. This problem is distinct from most common diffusion alignment and conditional generation settings (Wallace et al., 2024; Liu et al., 2024; Li et al., 2024; Uehara et al., 2025), because distributional alignment is a *population-level* objective

[1]Code is available at https://github.com/EdmundLuan/diffusion_attribute_distribution_alignment.

and thus generally cannot be evaluated by a reward function $r(\mathbf{x})$ defined on a *single sample*. The most conceptually related topics include fairness-aware generation and sample diversity promotion in diffusion- and flow-based models. However, existing approaches (Kang et al., 2025; He et al., 2024; Choi et al., 2024; Friedrich et al., 2025; Jiang et al., 2025; Li et al., 2025; Morshed & Boddeti, 2025) are often tailored to text-to-image (T2I) diffusion models, require retraining or fine-tuning for new targets, or lack flexibility when the target attribute distribution changes at test time. To our knowledge, there are few generic methods that can align pretrained diffusion models to different test-time target attribute distributions in a plug-and-play fashion, *without retraining or finetuning the model weights or training extra components*.

In this work, we show that attribute distributional alignment can be formulated as an optimal control (OC) problem over the reverse diffusion process. Concretely, we view sampling from a pretrained diffusion model as rolling out a dynamical system that defines a prior over realistic samples, and augment the learned dynamics with an additive, time-dependent perturbation. We leverage optimal-control algorithms to solve for perturbations that optimize a differentiable distribution-alignment objective. Taking this control-theoretic perspective is appealing for three reasons: (i) it provides a principled formulation that can *explicitly* balance the trade-off between the alignment objective and data fidelity by penalizing control effort, ; (ii) it is inherently *target-flexible* and naturally an *inference-time* approach, because changing the target attribute distribution only changes the objective, not the pretrained model; (iii) the step-wise perturbations are computed by algorithms grounded in optimal control theory rather than relying on handcrafted heuristic weighting schedules in many guidance methods, ensuring that the required distributional shifts are achieved via minimal deviations to the original sampling dynamics.

We make the following contributions in this paper: (i) We formulate the attribute distributional alignment problem for diffusion models as an OC problem; (ii) we propose a practical inference-time, training-free algorithm for pretrained unconditional diffusion to align the *attribute distribution* with flexible target distributions; (iii) we demonstrate the empirical effectiveness of our method in aligning attribute distributions with test-time targets in image generation.

2 Related Work

Diffusion Models with Conditional Generation Diffusion models define the generative process as iterative denoising (Ho et al., 2020) or, equivalently, as score-based Langevin dynamics (Song & Ermon, 2019; Song et al., 2021b). While DDIM (Song et al., 2021a) and the EDM framework (Karras et al., 2022) significantly improved sampling efficiency and design clarity, controllable generation often relies on guidance mechanisms. Dhariwal & Nichol (2021) introduced classifier guidance to steer pretrained models, while Ho & Salimans (2022) proposed classifier-free guidance for joint training with conditioning signals. Chung et al. (2023) proposed posterior sampling for inverse problems, upon which many more general training-free guidance techniques were built (Guo et al., 2024; Ye et al., 2024; Feng et al., 2024).

Fairness and Diversity in Diffusion Fair generation with diffusion models is an instance of the DADA problem. Shen et al. (2024) propose a supervised finetuning method using distributional alignment loss and adjusted direct finetuning to mitigate demographic biases in T2I diffusion models. Miao et al. (2024) take a reinforcement learning approach for fine-tuning with policy gradient methods and a diversity reward. As for test-time methods, Friedrich et al. (2025) conduct a preliminary model audit to generate a lookup table of biased prompts and attributes, and then employs Semantic Guidance to promote fair attribute classes and suppress biased terms. Jiang et al. (2025) train extra attribute-specific adapters and guide diffusion generation in a plug-and-play fashion; Fair Mapping (Li et al., 2025) adopts a linear network to map input embeddings into a debiased representation space. FairGen (Kang et al., 2025) uses an additionally-trained latent module and an extra memory module to steer generations toward user-specified fair distributions. Distribution-focused techniques such as IDA (He et al., 2024) optimize the weights of multi-directional text descriptions, while Parihar et al. (2024) trains a predictor to map h-space features to attribute distributions during the denoising diffusion process. However, most of the above methods are specifically designed for T2I conditional models and leverage text conditioning mechanisms that are baked inside the T2I models, thus not generalizable to unconditional models nor to other domains.

Optimal Control in Diffusion- and Flow-Based Models Early theoretical connections between Stochastic Optimal Control (SOC) and diffusion were established in (Berner et al., 2024). Domingo-Enrich et al. (2024; 2025) propose Adjoint Matching and Stochastic Optimal Control Matching for fine-tuning diffusion- and flow-based models by learning optimal control fields via regression objectives, as other approaches follow (Han et al., 2025; Liu et al., 2025). Park et al. (2024); Zhu et al. (2025) study diffusion bridges from the SOC perspective. In contrast, training-free methods such as OC-Flow (Wang et al., 2025) and DTM (Pandey et al., 2025) steer pre-trained diffusion or flow models by differentiating through the generative ODE or SDE to minimize control costs without updating model parameters. The OC formulation also extends to solving inverse problems (Li & Pereira, 2024), adaptive guidance strength scheduling (Azangulov et al., 2024), stylization (Rout et al., 2025) and multi-subject generation (Bill et al., 2025) by modulating trajectories or maximizing likelihoods directly during the sampling process. While our method resonates with these works in terms of methodology, none of them are addressing the attribute distribution alignment problem.

3 Diffusion Attribute Distribution Alignment

Notations Let $n, m, o \in \mathbb{Z}^+$ be positive integers and $t \in \mathbb{R}_{\geq 0}$ be the time variable. Let $\mathbf{x}, \mathbf{z}, \in \mathbb{R}^n$, $\mathbf{u} \in \mathbb{R}^m$, $\mathbf{y} \in \mathbb{R}^o$ be vector random variables. Vector constants are in bold symbol as $\boldsymbol{x} \in \mathbb{R}^n$. $\boldsymbol{I}_n$ stands for an $n \times n$ identity matrix. The Euclidean norm is denoted by $\|\cdot\|$ and $\|\cdot\|_F$ is the Frobenius norm. Let $\langle \boldsymbol{A}, \boldsymbol{B} \rangle_F := \operatorname{tr}(\boldsymbol{A}^\top \boldsymbol{B})$ denote the Frobenius inner product. We note the probability density function of $\mathbf{x}$ (or probability mass function for discrete variables) by $p_{\mathbf{x}}(\mathbf{x})$, and may, for notational brevity, omit the subscript of random variable when it is clear from the context. $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ stands for a Gaussian distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^n$ and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$. Let $\mathbb{D}_{\mathrm{KL}}\left[\cdot \,\|\, \cdot\right]$ denote the Kullback-Leibler (KL) divergence.

3.1 Review of the Dynamics of Diffusion Models

Song et al. (2021b) describe a forward diffusion process with stochastic differential equation (SDE):

$$\mathrm{d}\mathbf{x}_t = \boldsymbol{f}(\mathbf{x}_t, t)\,\mathrm{d}t + g(t)\,\mathrm{d}\mathbf{w}_t, \quad t \in [0, T],$$

where $\boldsymbol{f}(\cdot, t)$ is the drift coefficient, $g(t)$ is the diffusion coefficient, and $\mathbf{w}_t$ is the standard Wiener process. The reverse diffusion process, in which noise are transformed into data, is given by Anderson (1982); Song et al. (2021a) as

$$\mathrm{d}\mathbf{x}_t = \left[\boldsymbol{f}(\mathbf{x}_t, t) - g(t)^2 \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)\right] \mathrm{d}t + g(t)\,\mathrm{d}\bar{\mathbf{w}}_t, \tag{1}$$

wherein $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$ is the *(Stein) score function* of the marginal distribution over $\mathbf{x}_t$ at time t, and $\bar{\mathbf{w}}_t$ is the standard Wiener process in reverse time. Song et al. (2021a) also identify the probability flow ordinary differential equations (PF-ODE) that yields the same marginal distribution over $\mathbf{x}_t$ at t as SDE Eq. (1):

$$\mathrm{d}\mathbf{x}_t = \boldsymbol{f}(\mathbf{x}_t, t) - \frac{1}{2} g(t)^2 \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)\,\mathrm{d}t. \tag{2}$$

For complex, high-dimensional distributions, the score function is analytically intractable in general, so it is often approximated by a neural network $\mathbf{s}^\theta(\mathbf{x}, t)$ parameterized by θ and learned via score matching (Song & Ermon, 2019). Equivalent to learning the score, the DDPM (Ho et al., 2020) and DDIM (Song et al., 2021a) formulations, in which the forward diffusion process is constructed as sequentially adding noise to the initial data distribution, learn a noise prediction model $\boldsymbol{\epsilon}^\theta(\mathbf{x}_t, t)$ for removing the noise during the reverse process to reconstruct the data distribution.

In this work, we focus on the *PF-ODE of the reverse diffusion process*, which appears in various forms across literature. We identify two instances here. With the formulation of EDM (Karras et al., 2022), the PF-ODE of a reverse diffusion process reads: $\dot{\mathbf{x}}_t = -t\,\mathbf{s}^\theta(\mathbf{x}_t, t)$. The associated PF-ODE of DDIM is as follows (Song et al., 2021a):

$$\mathrm{d}\tilde{\mathbf{x}}_t = \boldsymbol{\epsilon}^\theta\left(\frac{\tilde{\mathbf{x}}_t}{\sqrt{1 + \sigma(t)^2}}\right) \mathrm{d}\sigma(t) \quad \text{with} \quad \tilde{\mathbf{x}}_t := \frac{\mathbf{x}_t}{\sqrt{\alpha_t}} \quad \text{and} \quad \sigma(t) := \sqrt{(1 - \alpha_t)/\alpha_t}. \tag{3}$$

with α_t being a decreasing sequence related to the diffusion noising schedule[2], and ϵ^θ a learned noise prediction model parameterized by θ. Without loss of generality, we denote both of them with

$$\dot{\mathbf{x}}_t = \mathbf{f}^\theta(\mathbf{x}_t, t), \quad \mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n), \quad t \in [0, T]. \tag{4}$$

3.2 Attribute Distribution Alignment Formulation

Suppose that there is a pretrained diffusion model $\mathbf{f}^\theta(\mathbf{x}, t)$ and the reverse diffusion process is as Eq. (4). Following Domingo-Enrich et al. (2024; 2025), we treat the diffusion PF-ODE as a dynamical system and introduce a time-dependent, additive perturbation $\mathbf{u}_t \in \mathbb{R}^m$ called control:

$$\dot{\mathbf{x}}_t := \mathbf{f}^\theta(\mathbf{x}_t, t) + \mathbf{g}(\mathbf{x}_t, t)\mathbf{u}_t, \tag{5}$$

for $t \in [0, T]$, with $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$ and $\mathbf{g} : \mathbb{R}^n \times \mathbb{R}_{\geq 0} \to \mathbb{R}^{n \times m}$ as an actuation field. We further let $p^u_{\mathbf{x}_T}$ denote the distribution of $\mathbf{x}_T$ sampled from process Eq. (5).

Remark 1 (Time Direction). From this point on, we follow the convention in dynamical systems and take the time direction as from 0 to T to describe *reverse* diffusion.

Let the attribute of interest, $\mathbf{y} \in \mathbb{R}^o$, associated with a sample $\mathbf{x}_T$, be determined by a continuously differentiable function $\Psi : \mathbb{R}^n \to \mathbb{R}^o$: $\mathbf{y} = \Psi(\mathbf{x}_T)$. The distribution of $\mathbf{y}$ for samples generated by the perturbed process is then $p^u_{\mathbf{y}}(\mathbf{y}) = \int_{\mathbf{y}=\Psi(\mathbf{x})} p^u_{\mathbf{x}_T}(\mathbf{x})\, \mathrm{d}\mathbf{x}$.

Inference-time Attribute Distribution Alignment. Given a *test-time* target attribute *distribution* $p^{\text{tar}}_{\mathbf{y}}$, our goal is to find control $\mathbf{u}_t$ for $t \in [0, T]$ such that $p^u_{\mathbf{y}}(\mathbf{y})$ aligns with the target *without retraining or fine-tuning* the pretrained generative model $\mathbf{f}^\theta(\mathbf{x}, t)$, *i.e.*,

$$\min_{\mathbf{u}_t, t \in [0,T]} \mathbb{D}[p^u_{\mathbf{y}} || p^{\text{tar}}_{\mathbf{y}}] \tag{6}$$

where $\mathbb{D}[\cdot||\cdot]$ is a statistical distance or divergence between two distributions with the same support.

Perturbed PF-ODE Instances In this work, we instantiate the perturbed PF-ODE Eq. (5) under the EDM (Karras et al., 2022) and DDIM (Song et al., 2021a) formulations. For EDM, we take the following *perturbed* ODE:

$$\mathbf{f}^\theta_{\text{EDM}}(\mathbf{x}_t, \mathbf{u}_t, t) := \dot{\mathbf{x}}_t = (T - t)\mathbf{s}^\theta(\mathbf{x}_t, T - t) + \mathbf{u}_t, \tag{7}$$

setting $\mathbf{g}(\mathbf{x}_t, t) = \boldsymbol{I}_n$. For DDIM, we adopt

$$\mathbf{f}^\theta_{\text{DDIM}}(\tilde{\mathbf{x}}_\sigma, \mathbf{u}_\sigma, \sigma) := \frac{\mathrm{d}\tilde{\mathbf{x}}_\sigma}{\mathrm{d}\sigma} = \epsilon^\theta\left(\frac{\tilde{\mathbf{x}}_\sigma}{\sqrt{1 + \sigma^2}}\right) + \mathbf{u}_\sigma, \tag{8}$$

and perform control in the spaces of $\tilde{\mathbf{x}}_\sigma$ and $\sigma(T - t)$.

Connection with diffusion guidance. The perturbed PF-ODE Eq. (5) can conceptually encompass classifier guidance: Let $\mathbf{g}(\mathbf{x}_t, t) \propto \nabla_{\mathbf{x}_t} \log p(c|\mathbf{x}_t)$ for a condition c and let the control $\mathbf{u}_t := w_t$ be a scheduled scalar weight.

4 Optimal Control for Distribution Alignment

4.1 Review of Sample-wise Optimal Control

Let $\mathbf{x}$ denotes the state and $\mathbf{u}$ the *control input* for a dynamical system. Let $L : \mathbb{R}^n \to \mathbb{R}$ and $\ell : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ be continuously differentiable functions. Let $\mathbf{f} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}_{\geq 0}$ be a continuous vector function with continuous first partial derivatives with respect to the first argument. An OC problem is formulated as follows:

$$\min_{\mathbf{u}_t, t \in [0,T]} J(\mathbf{x}, \mathbf{u}) = L(\mathbf{x}_T) + \int_0^T \ell(\mathbf{x}_t, \mathbf{u}_t, t)\, \mathrm{d}t \tag{9a}$$

$$\text{s.t.} \quad \dot{\mathbf{x}}_t = \mathbf{f}(\mathbf{x}_t, \mathbf{u}_t, t) \tag{9b}$$

$$\mathbf{x}_0 = \boldsymbol{x}_{\text{init}}, \quad \mathbf{u}_t \in \mathcal{U}, \quad t \in [0, T], \tag{9c}$$

[2]There is a notional mismatch in α_t across different work. The α_t in (Song et al., 2021a) corresponds to the $\bar{\alpha}_t$ in (Ho et al., 2020). We refer readers to (Ho et al., 2020; Song et al., 2021a) for the details. In the above, we adopt the notations by Song et al. (2021a).

where $\mathcal{U} \subseteq \mathbb{R}^m$ is a closed set we call the admissible control set. L is called the terminal cost function, ℓ is the running cost or transient cost, and J is the cost functional. The constraint Eq. (9b) is the equation of motion of a dynamical system. See (Fleming & Rishel, 2012) for more rigorous definitions. The formulation Eq. (9) is explicitly minimizing a scalar total cost that encodes some desired effects at the terminal time and the control efforts exerted to the system, while simultaneously respecting the system dynamics, initial state conditions, and admissible control constraints.

4.2 Optimal Control Formulation for DADA

Finite-sample Batched OC. Different from previous work that applies OC to diffusion generation to achieve sample-wise objectives, our alignment objective in Eq. (6) inherently depends on multiple samples. As such, we first introduce the following batched notations. Let $\mathbf{X} \in \mathbb{R}^{M\times n}$ stack M states $\mathbf{x}^{[i]}$ row-wise, and $\mathbf{U} \in \mathbb{R}^{M\times m}$ stack controls $\mathbf{u}^{[i]}$ row-wise, for $i \in \{1, \cdots, M\}$. Define the row-wise dynamics operator $\mathbf{F}^\theta(\mathbf{X}, \mathbf{U}, t) \in \mathbb{R}^{M\times n}$ by $[\mathbf{F}^\theta(\mathbf{X}, \mathbf{U}, t)]_{i,:} = \mathbf{f}^\theta(\mathbf{x}^{[i]}, \mathbf{u}^{[i]}, t)$. We stack the i.i.d. initial states $\boldsymbol{x}_{\text{init}}^{[i]} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$ as $\boldsymbol{X}_{\text{init}} \in \mathbb{R}^{M\times n}$. We also write $\mathcal{U}^M := \mathcal{U} \times \cdots \times \mathcal{U}$.

With the above notations, we formulate the DADA problem as a finite-sample batched OC problem:

$$\min_{\mathbf{U}_t \in \mathcal{U}^M,\ t\in[0,T]} \Phi\left(\hat{p}_{\mathbf{y}}^u(\mathbf{X}_T)\right) + \frac{\rho}{2}\int_0^T \|\mathbf{U}_t\|_F^2 \,\mathrm{d}t \tag{10a}$$

$$\text{s.t.} \qquad \dot{\mathbf{X}}_t = \mathbf{F}^\theta(\mathbf{X}_t, \mathbf{U}_t, t), \quad \mathbf{X}_0 = \boldsymbol{X}_{\text{init}}, \quad \mathbf{U}_t \in \mathcal{U}^M \tag{10b}$$

where $\hat{p}_{\mathbf{y}}^u(\mathbf{X}_T)$ denotes attribute distribution estimated by the batched samples $\mathbf{X}_T$, and we adopt the reverse KL divergence against the target as the terminal cost:

$$\Phi\big(\hat{p}_{\mathbf{y}}^u(\mathbf{X}_T)\big) := \mathbb{D}_{\text{KL}}\left[\hat{p}_{\mathbf{y}}^u(\mathbf{X}_T) \,\|\, p_{\mathbf{y}}^{\text{tar}}\right]. \tag{11}$$

Note that while the dynamics Eq. (10b) are sample-wise, the terminal cost Eq. (11) couples the batch through $\hat{p}_{\mathbf{y}}^u(\mathbf{X}_T)$. See Section 5 for a differentiable approximation of $\hat{p}_{\mathbf{y}}^u(\mathbf{X}_T)$.

4.3 Optimal Controller for DADA OC

By Pontryagin's Maximum Principle (PMP) (Levine, 1972; Fleming & Rishel, 2012), we may obtain the *necessary* conditions for solving problem Eq. (10). Define the Hamiltonian of Eq. (10) as

$$\widetilde{H}(\mathbf{X}_t, \mathbf{N}_t, \mathbf{U}_t, t) := \frac{\rho}{2}\|\mathbf{U}_t\|_F^2 + \langle \mathbf{N}_t, \mathbf{F}^\theta(\mathbf{X}_t, \mathbf{U}_t), t)\rangle_F$$

where $\mathbf{N}_t \in \mathbb{R}^{M\times n} = \left[\boldsymbol{\nu}_t^{[i]}\right]$ denotes the adjoint states stacked row-wise. Let $(\tilde{\mathbf{X}}_t^*, \tilde{\mathbf{N}}_t^*, \tilde{\mathbf{U}}_t^*)$ denote an optimal trajectory for the problem Eq. (10). The *necessary* conditions by PMP are as follows:

$$\dot{\tilde{\mathbf{X}}}_t^* = \mathbf{F}^\theta(\tilde{\mathbf{X}}_t^*, \tilde{\mathbf{U}}_t^*, t), \tag{12a}$$

$$\dot{\tilde{\mathbf{N}}}_t^* = -\nabla_{\mathbf{X}}\mathbf{F}^\theta(\tilde{\mathbf{X}}_t^*, \tilde{\mathbf{U}}_t^*, t)^\top \tilde{\mathbf{N}}_t^*, \tag{12b}$$

$$\tilde{\mathbf{N}}_T^* = \nabla_{\mathbf{X}}\Phi\left(\hat{p}_{\mathbf{y}}^u(\tilde{\mathbf{X}}_T^*)\right), \tag{12c}$$

$$\tilde{\mathbf{U}}_t^* \in \operatorname*{arg\,min}_{\mathbf{U}\in\mathcal{U}^M} \widetilde{H}(\tilde{\mathbf{X}}_t^*, \tilde{\mathbf{N}}_t^*, \mathbf{U}, t). \tag{12d}$$

where all gradient operators with respect to batched variables above are *applied row-wise*. The adjoint dynamics Eq. (12b) also independently apply row-wise: $\dot{\boldsymbol{\nu}}_t^{[i]} = -\big(\nabla_{\mathbf{x}}\mathbf{f}^\theta(\mathbf{x}_t^{[i]}, \mathbf{u}_t^{[i]}, t)\big)^\top \boldsymbol{\nu}_t^{[i]}$.

However, jointly solving all conditions in Eq. (12) is essentially a two-point boundary value problem, and for general nonlinear, nonconvex dynamics, it is challenging to solve. Rather than directly solving them in joint, we follow Wang et al. (2025) and employ the Extended Method of Successive Approximations (E-MSA) (Li et al., 2018) to solve a *proximalized* OC subproblem in an iterative fashion for stability. Specifically, given a reference control $\boldsymbol{U}_t^{\text{ref}}$, we consider a subproblem

$$\min_{\mathbf{U}_t\in\mathcal{U}^M} \Phi\big(\mathbf{X}_T\big) + \frac{1}{2}\int_0^T \left(\rho\|\mathbf{U}_t\|_F^2 + \gamma\|\mathbf{U}_t - \boldsymbol{U}_t^{\text{ref}}\|_F^2\right)\mathrm{d}t, \tag{13}$$

subject to Eq. (10b) . Applying PMP to this *proximalized* subproblem yields the *necessary* conditions as in Eq. (12) but in terms of the extended Hamiltonian

$$H(\mathbf{X}_t, \mathbf{N}_t, \mathbf{U}_t, t) := \widetilde{H}(\mathbf{X}_t, \mathbf{N}_t, \mathbf{U}_t, t) + \frac{\gamma}{2}\|\mathbf{U}_t - \boldsymbol{U}_t^{\mathrm{ref}}\|_F^2$$

and the optimal trajectories $(\mathbf{X}_t^*, \mathbf{N}_t^*, \mathbf{U}_t^*)$ that solves the proximalized problem. Accordingly,

$$\mathbf{U}_t^* \in \underset{\mathbf{U}\in\mathcal{U}^M}{\arg\min}\ H(\mathbf{X}_t^*, \mathbf{N}_t^*, \mathbf{U}, t). \tag{14}$$

Closed-form Control Update A quadratic running cost with control-affine dynamics gives a closed-form minimizer for Eq. (14):

$$\mathbf{U}_t^* = \Pi_{\mathcal{U}^M}\left(\xi \boldsymbol{U}_t^{\mathrm{ref}} - \eta\, \mathbf{G}\left(\mathbf{X}_t^*, t\right)^\top \mathbf{N}_t^*\right), \tag{15}$$

where $\xi := \frac{\gamma}{\rho+\gamma}$, $\eta := \frac{1}{\rho+\gamma}$, $\Pi_{\mathcal{U}^M}$ denotes row-wise projection, and $\mathbf{G}(\mathbf{X}, t)^\top \mathbf{N} \in \mathbb{R}^{M\times m}$ is also a row-wise operator: $[\mathbf{G}(\mathbf{X}, t)^\top \mathbf{N}]_{i,:} = \mathbf{g}\left(\mathbf{x}^{[i]}, t\right)^\top \boldsymbol{\nu}^{[i]}$.

Proposition 4.1. *For any fixed $t \in [0, T]$ and fixed $(\boldsymbol{x}_t^{[i]}, \boldsymbol{\nu}_t^{[i]})$, with $\rho + \gamma > 0$, the sample-wise optimal control $\mathbf{u}_t^{*,[i]}$ for Eq. (14) is given by*

$$\mathbf{u}_t^{*,[i]} \in \Pi_{\mathcal{U}}\left(\bar{\boldsymbol{u}}_t^{[i]}\right) := \underset{\mathbf{u}\in\mathcal{U}}{\arg\min} \left\|\mathbf{u} - \bar{\boldsymbol{u}}_t^{[i]}\right\|^2, \quad \textit{where } \bar{\boldsymbol{u}}_t^{[i]} := \frac{1}{\rho+\gamma}\left(\gamma\, \boldsymbol{u}_t^{\mathrm{ref,[i]}} - \mathbf{g}\left(\boldsymbol{x}_t^{[i]}, t\right)^\top \boldsymbol{\nu}_t^{[i]}\right),$$

and $\Pi_{\mathcal{U}}\left(\bar{\boldsymbol{u}}_t^{[i]}\right)$ is nonempty. Further, $\mathbf{u}_t^{*,[i]}$ is unique if $\mathcal{U}$ is convex.

Proof. See Appendix **A**. □

Proposition 4.1 indicates that the optimal control at each time is a function of the state and adjoint at that time. The resulting algorithm performs the following three steps iteratively: (i) Simulate Eq. (10b) in forward time; (ii) Evaluate the distribution-alignment cost, take the gradients, and simulate the adjoint dynamics in backward time. (iii) Update controls with Eq. (15).

In practice, both steps (i) and (ii) are performed with Euler discretizations, and we leverage the vector-Jacobian product (VJP) when simulating the adjoints without materializing the entire gigantic dynamics Jacobian. The resulting algorithm is in Appendix **B.1** Algorithm 1.

5 Experiments

We address the following research questions through experiments in image generation:

RQ 1: How effective is our method at aligning the attribute distribution, compared to baselines?

RQ 2: Can our method better preserve sample quality while achieving distributional alignment, compared to baseline methods?

5.1 CIFAR-100 with Hierarchical Semantic Classes

As a proof of concept, we first demonstrate our method by generating low-resolution images across semantic classes. We treat the semantic class of an image as an attribute and consider flexible test-time target distributions with different support sizes.

Setup: We adopt an unconditional EDM model (Karras et al., 2022) trained on the CIFAR-100 dataset (Krizhevsky et al., 2009) as the base diffusion model and use ResNet (He et al., 2016) image classifiers as the attribute model Ψ. We test our method on three different levels of class labels with a coarse-to-fine hierarchy: `meta5`, `coarse`, and `fine`, with 5, 20, and 100 classes, respectively. We choose three different attribute distributions as test-time targets for each class level: Uniform, ZigZag, and Gaussian. Figure 1 (top row) shows three instances of the target distributions. See implementation details in Appendix **B.2**.

Table 1: Quantitative evaluation metrics on `CIFAR-100`. Best in **bold**, second-best underlined.

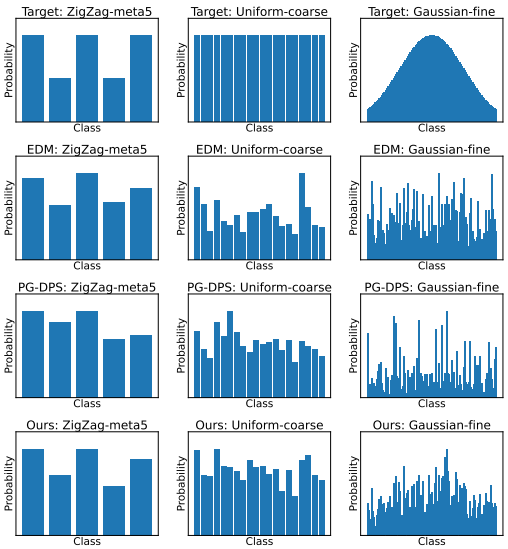
METHOD	meta5 TV↓	meta5 JS↓	meta5 χ^2↓	meta5 FID↓	coarse TV↓	coarse JS↓	coarse χ^2↓	coarse FID↓	fine TV↓	fine JS↓	fine χ^2↓	fine FID↓
Gaussian												
OURS	**0.136**	**0.117**	**0.0272**	**16.5**	**0.176**	**0.146**	**0.0421**	**15.7**	**0.184**	**0.173**	**0.0573**	**13.7**
EDM	0.252	0.196	0.0749	17.4	0.274	0.227	0.0988	17.4	0.248	0.231	0.101	16.1
PG-DPS	0.154	0.133	0.0350	21.3	0.195	0.165	0.0528	16	0.348	0.299	0.165	33.1
Zigzag												
OURS	**0.0544**	**0.0495**	**0.00489**	**14**	**0.103**	**0.0927**	**0.0171**	**14.8**	**0.171**	**0.15**	**0.0442**	**12.6**
EDM	0.067	0.0574	0.00658	15.3	0.185	0.148	0.0433	16.1	0.24	0.205	0.0812	16.1
PG-DPS	0.113	0.1	0.0199	22	0.133	0.116	0.0267	15.7	0.341	0.288	0.156	33.8
Uniform												
OURS	**0.0281**	**0.0292**	**0.00171**	**13.1**	**0.069**	**0.0655**	**0.00853**	**12.6**	**0.141**	**0.12**	**0.0284**	**12.6**
EDM	0.083	0.0651	0.00845	15.5	0.132	0.114	0.0255	15.5	0.186	0.159	0.0495	15.5
PG-DPS	0.0692	0.0529	0.00559	22.4	0.0805	0.0742	0.0109	15.2	0.287	0.255	0.123	30.7

Differentiable Approximations for Terminal Cost: For discrete attributes, the attribute model is a continuously differentiable neural network classifier that maps an input $\mathbf{x}$ to an attribute-logit vector. To maintain differentiability, we estimate the empirical distribution $\hat{p}^u_{\mathbf{y}}$ over a batch of M generated samples by averaging their softmax probabilities across the batch: $\hat{p}^u_{\mathbf{y}} = \frac{1}{M}\sum_{i=1}^{M} \text{softmax}\left(\Psi\left(\mathbf{x}_T^{[i]}\right)\right)$. The terminal cost is then the element-wise KL divergence:

$$\Phi\left(\hat{p}^u_{\mathbf{y}}\right) = \mathbb{D}_{\text{KL}}\left[\hat{p}^u_{\mathbf{y}} \,\|\, p^{\text{tar}}_{\mathbf{y}}\right] = \sum_{j=1}^{m} \hat{p}^{u,(j)}_{\mathbf{y}} \log \frac{\hat{p}^{u,(j)}_{\mathbf{y}}}{p^{\text{tar},(j)}_{\mathbf{y}}}, \tag{16}$$

where m is the total number of attribute classes, and the superscript (j) indexes the probability mass associated with the j-th class.

Baseline: We compare our method with vanilla EDM, which corresponds to i.i.d. sampling from the learned distribution, and a guidance-based method, Particle Guidance (PG) (Corso et al., 2024). For PG, the guidance potential uses the same terminal cost as our method, but it is applied to the estimated terminal sample $\hat{\mathbf{x}}_T$ obtained via Tweedie's formula (Efron, 2011), akin to DPS (Chung et al., 2023). See Appendix **B.2** for details.



(a) ZigZag-`meta5` (b) Uniform-`coarse` (c) Gaussian-`fine`

Figure 1: **Top row**: Test-time *target* attribute distributions (CIFAR-100). **2nd/3rd rows:** Generated distributions of baselines. **Bottom row:** Generated distributions of our method.

Evaluation Metrics: We use several statistical distances/divergences to quantify how well the semantic-class distribution of generated samples matches the target: Total Variation (TV), Jensen–Shannon divergence (JS), and the χ^2 distance[3]. We use the Fréchet Inception Distance (FID) to evaluate image quality.[4]

Results: Figure 1 (except the top row) shows the attribute distributions of samples generated by different methods, with the targets in the first row. Table 1 presents comprehensive quantitative comparisons across methods and target attribute distributions with different support sizes. Across all settings, our method achieves the best performance in terms of both attribute-distribution align-

[3]We adopt the symmetric χ^2 distance (Markatou et al., 2018): $d_{\chi^2}(P,Q) = \frac{1}{2}\sum_{i=1}^{K}(P_i - Q_i)^2/(P_i + Q_i)$ for two discrete distributions P and Q with the same finite support.

[4]The FID references are computed via sampling from the training set by each target attribute distribution.

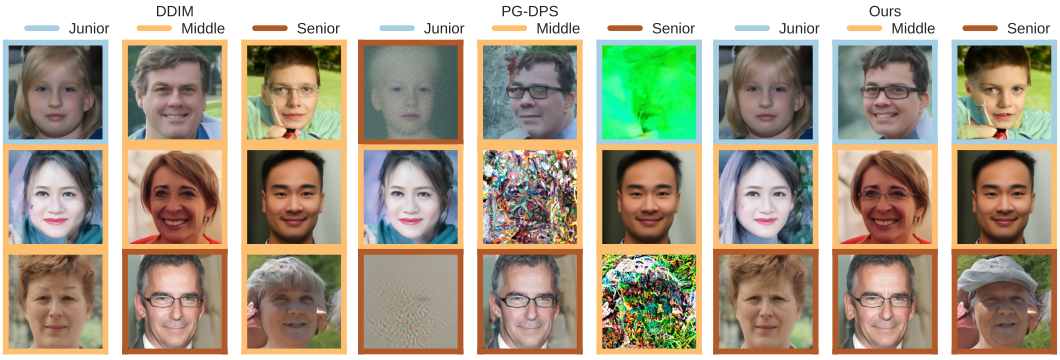


Figure 2: Qualitative samples *from a single batch* of our method (right), compared to vanilla DDIM (left) and PG (mid) for generating human faces with a *fair* target distribution over age groups (Uniform-`age`), where the ratio of faces in three age groups should be 1:1:1. The pretrained diffusion model learned a highly imbalanced age distribution from the FFHQ dataset, where most faces are classified as `Middle`. Our approach aligns the generated attribute distribution with the target by minimally editing facial details such as wrinkles and face shapes.

ment and image quality. Surprisingly, PG performs even worse than vanilla EDM in some cases (*e.g.*, ZigZag-`meta5` and Gaussian-`fine`). This suggests that naively applying guidance-like approaches with a batch-wise distributional loss may be insufficient for attribute distribution alignment. Additional results including ablation studies in batch size M for the empirical distribution estimation are in Appendix **B.2**.

5.2 Human Face Generation with DDIM

We aim to achieve human-face image generation with *fairness* across genders, races, and ages, mitigating potential biases introduced by pretrained diffusion models *without retraining or fine-tuning them*. Beyond fairness, we also demonstrate our method for achieving distributional alignment to target attribute distributions specified at test time, *e.g.*, a race distribution that faithfully reflects the demographics of a region during a given time period.

Setup: We use a DDIM model pretrained by Choi et al. (2022) on the FFHQ dataset (Karras et al., 2019). We consider three attributes in this task: gender, race, and age. For gender, we consider {`Female`, `Male`}; for race, we adopt the 4-way classification by Karkkainen & Joo (2021): {`Asian`, `Black`, `Indian`, `WMELH`}[5]; for age, we partition ages into three groups: {`Junior`, `Middle`, `Senior`}[6]. We conduct generation with both fairness targets and customized targets for each single attribute and *joint* attributes. We also use an image classifier as the attribute model. Implementation details are in Appendix **B.3**.

Aligning Joint Attribute Distribution: For *joint* attributes, the alignment target is the factorized joint distribution $p^{\text{tar}}_{\mathbf{y}}(\mathbf{y}) \propto \prod_{i=1}^{N} p^{\text{tar}}_{\mathbf{y}_i}(\mathbf{y}_i)$, where we assume *independence among all attributes*. Note that this independence assumption generally *does not* hold for the generated attribute distribution $p^{u}_{\mathbf{y}}(\mathbf{y})$. The terminal cost then decomposes as $\mathbb{D}_{\text{KL}}\left[\hat{p}^{u}_{\mathbf{y}} \,\|\, p^{\text{tar}}_{\mathbf{y}}\right] = \sum_{i=1}^{N} \mathbb{D}_{\text{KL}}\left[\hat{p}^{u}_{\mathbf{y}_i} \,\|\, p^{\text{tar}}_{\mathbf{y}_i}\right] + \sum_{i=1}^{N} \text{H}(\hat{p}^{u}_{\mathbf{y}_i}) - \text{H}(p^{\text{tar}}_{\mathbf{y}_i})$ where $\hat{p}^{u}_{\mathbf{y}_i}$ is the empirical marginal, and $\text{H}(\cdot)$ is entropy.

Baselines: We compare our method with vanilla DDIM (Song et al., 2021a) and PG (Corso et al., 2024). Setup of PG is similar to that in the previous experiment.

Evaluation Metrics: As in the previous experiment, we measure attribute-distribution alignment using JS, TV, and χ^2, and we use FID to evaluate image quality. We also measure the Fairness Discrepancy (FD), following prior work on fair generation (Choi et al., 2020; Parihar et al., 2024): $\text{FD} := \|p^{\text{tar}}_{\mathbf{y}} - \mathbb{E}_{\mathbf{x}\sim p^{\theta,\mathbf{u}}_{x}(\mathbf{x})} \hat{p}_{\mathbf{y}}(\mathbf{y} \mid \mathbf{x})\|$, where $\hat{p}_{\mathbf{y}}(\mathbf{y} \mid \mathbf{x})$ is the *softmax output* of $\Psi(\mathbf{x})$.

[5] `WMELH` stands for the merge of White, Middle Eastern, and Latino Hispanic.
[6] `Junior`: 0–19 years old; `Middle`: 20–50; `Senior`: 50–120.

Table 2: Quantitative evaluation metrics (all lower the better) on face generation with controlling only one of the three attributes. Best in **bold**, second-best underlined.

METHOD	age TV↓	age JS↓	age χ^2↓	age FD↓	age FID↓	gender TV↓	gender JS↓	gender χ^2↓	gender FD↓	gender FID↓	race TV↓	race JS↓	race χ^2↓	race FD↓	race FID↓
Uniform															
DDIM	.21	.15	.044	.25	50.70	.042	.029	.0017	.052	50.70	.56	.45	.36	.65	50.70
PG-DPS	.15	.13	.033	.18	114.8	.043	.030	.0018	.047	77.99	.37	.32	.19	.45	104.4
OURS	**.023**	**.018**	**6.4e-4**	**.028**	**48.59**	**.0052**	**.0037**	**2.7e-5**	**.0087**	**48.37**	**.028**	**.025**	**.0012**	**.038**	**45.57**
Custom-1															
	$[4:1:3]$					$[2:8]$					$[4:3:2:1]$				
DDIM	.42	.32	.20	.51	49.65	.26	.20	.076	37	48.58	.41	.35	.22	.50	48.71
PG-DPS	.38	.31	.18	.46	123.1	.24	.18	.065	.33	129.6	.27	.27	.13	.30	165.3
OURS	**.049**	**.035**	**.0025**	**.063**	**46.14**	**.0094**	**.0082**	**1.4e-4**	**.017**	**46.52**	**.043**	**.040**	**.0032**	**.062**	**43.03**
Custom-2															
	$[2:3:4]$					$[7:3]$					$[1:1:4:4]$				
DDIM	.23	.18	.066	.32	51.39	.24	.17	.060	.33	55.74	.70	.56	.53	.84	51.45
PG-DPS	.30	.24	.11	.36	90.79	.043	.032	.0021	.084	156.9	.31	.28	.14	.36	253.2
OURS	**.013**	**.012**	**2.8e-4**	**.018**	**46.25**	**1.2e-8**	**1.0e-4**	**1.7e-16**	**.0011**	**48.69**	**.053**	**.041**	**.0034**	**.075**	**46.53**

Results: Table 2 presents quantitative comparisons of our method against baselines in aligning the target distributions for *each individual* attribute. Table 3 shows results for aligning the *joint* distribution of *all attributes*, See Figure 2 for a qualitative comparison of samples. In the case of Uniform targets, i.i.d. sampling from the pretrained DDIM exhibits significant age and race biases inherited from the training dataset. While the baseline PG can mitigate these biases, as evidenced by improved distributional metrics, it significantly compromises image quality, as reflected by much higher FID scores. In contrast, our method more effectively aligns attribute distributions toward the uniform fairness target across all metrics, while preserving image quality due to the OC-based formulation. When targets are customized to drastically different distributions (*e.g.*, `gender` and `race` in Custom-1 and Custom-2 are heavily skewed toward different modes), our method consistently aligns the generated distributions with the target while maintaining data fidelity. The performance of PG, however, is inconsistent across different targets. Even for the simple `gender` attribute, PG is effective for one target but has negligible effect on the other, while yielding the worst sample quality among the three methods. Furthermore, for `age` with the Custom-2 target, PG performs worse than vanilla DDIM generation. Similar patterns appear in joint attribution distribution alignment: PG tends to make a substantial trade-off between sample quality and alignment, yet does not consistently outperform vanilla DDIM in alignment. Conversely, our method demonstrates substantially better alignment and preserves sample quality. Additional results are available in Appendix **C.2**.

Table 3: Quantitative results for face generation with *joint attribute control*. Best in **bold**, second-best underlined.

METHOD	age×gender×race TV↓	JS↓	χ^2↓	FD↓	FID↓
UniformJoint					
DDIM	0.566	0.477	0.393	0.308	46.26
PG-DPS	0.444	0.404	0.281	0.238	127.0
OURS	**0.112**	**0.107**	**0.0225**	**0.0606**	**41.28**
CustomJoint					
$[5:2:3] \times [3:7] \times [4:3:2:1]$					
DDIM	0.493	0.444	0.337	0.305	44.28
PG-DPS	0.531	0.442	0.335	0.287	74.77
OURS	**0.142**	**0.131**	**0.034**	**0.090**	**42.19**

6 DISCUSSIONS AND CONCLUSION

In this work, we study the attribute distribution alignment problem for pretrained unconditional diffusion models. We formulate the problem as an optimal control problem, and propose an inference-time, plug-and-play method that does not require any extra training or fine-tuning. Our results show that the proposed method is effective in aligning the generation attribute distribution to flexible test-time targets, while better preserving sample quality compared to training-free baselines. Discussion of limitations of our method are in Appendix **D**. Future work includes extensions to finding optimal guidance weights in conditional diffusion models and applying our approach in other domains such as robot trajectory generation, graph generation, *etc.*

REPRODUCIBILITY

Our implementation code is available at https://github.com/EdmundLuan/diffusion_attribute_distribution_alignment.

IMPACT STATEMENT

This work advances inference-time control of unconditional diffusion generative models by enabling alignment to user-specified attribute distributions without training or fine-tuning. This may benefit applications such as fairness-aware data generation, controllable simulation, and robotic autonomous systems that require calibrated mixtures of behaviors. At the same time, the ability to steer attribute distributions could also be misused by malicious parties to amplify societal biases, manipulate demographic representation, or generate content that appears balanced in some attributes while hiding other harmful properties, depending on how attributes are defined and measured. We encourage practitioners to use validated models as attribute model, audit distributional outcomes across relevant subgroups and contexts, and apply standard safeguards (dataset governance, filtering, and usage policies) upon deployment of the proposed method. Overall, we view this paper as providing a general technical tool for distribution-level controllability whose societal impact depends on careful choice of attributes and responsible deployment.

ACKNOWLEDGMENT

This research/project is supported by the Ministry of Education, Singapore, under the Academic Research Fund Tier 1 (FY2025) and by the National University of Singapore, under the Start-Up Grant Scheme and the HPC Grant (Grant No: NUSREC-HPC-00001). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Education, Singapore.

REFERENCES

Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Iskander Azangulov, Peter Potaptchik, Qinyu Li, Eddie Aamari, George Deligiannidis, and Judith Rousseau. Adaptive diffusion guidance via stochastic optimal control. *arXiv preprint arXiv:2410.21245*, 2024.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia*, pp. 94:1–94:11, 2024. URL https://doi.org/10.1145/3680528.3687614.

Julius Berner, Lorenz Richter, and Karen Ullrich. An optimal control perspective on diffusion-based generative modeling. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=oYIjw37pTP.

Eric Tillmann Bill, Enis Simsar, and Thomas Hofmann. Optimal control meets flow matching: A principled route to multi-subject fidelity. *arXiv preprint arXiv:2510.02315*, 2025.

João Carvalho, An Thai Le, Piotr Kicki, Dorothea Koert, and Jan Peters. Motion planning diffusion: Learning and adapting robot motion planning with diffusion models. *IEEE Transactions on Robotics*, 41:4881–4901, 2025. doi: 10.1109/TRO.2025.3593109.

Yaofo Chen. Pytorch cifar models. https://github.com/chenyaofo/pytorch-cifar-models, 2025. Accessed: 2025-5-17.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, pp. 02783649241273668, 2023.

Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11472–11481, 2022.

Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon. Fair generative modeling via weak supervision. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1887–1898. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/choi20a.html.

Yujin Choi, Jinseong Park, Hoki Kim, Jaewook Lee, and Saerom Park. Fair sampling in diffusion models through switching mechanism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21995–22003, 2024.

Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=OnD9zGAGT0k.

Gabriele Corso, Yilun Xu, Valentin De Bortoli, Regina Barzilay, and Tommi S. Jaakkola. Particle guidance: non-i.i.d. diverse sampling with diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KqbCvIFBY7.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf.

Carles Domingo-Enrich, Jiequn Han, Brandon Amos, Joan Bruna, and Ricky T. Q. Chen. Stochastic optimal control matching. In *Advances in Neural Information Processing Systems*, volume 37, pp. 112459–112504, 2024. doi: 10.52202/079017-3573.

Carles Domingo-Enrich, Michal Drozdzal, Brian Karrer, and Ricky T. Q. Chen. Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xQBRrtQM8u.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Zeyu Feng, Hao Luan, Pranav Goyal, and Harold Soh. LTLDoG: Satisfying temporally-extended symbolic constraints for safe diffusion-based planning. *IEEE Robotics and Automation Letters*, 9 (10):8571–8578, 2024. doi: 10.1109/LRA.2024.3443501.

Zeyu Feng, Hao Luan, Kevin Yuchen Ma, and Harold Soh. Diffusion meets options: Hierarchical generative skill composition for temporally-extended tasks. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10854–10860, 2025. doi: 10.1109/ICRA55743.2025.11127641.

Nic Fishman, Leo Klarner, Emile Mathieu, Michael Hutchinson, and Valentin De Bortoli. Metropolis sampling for constrained diffusion models. *Advances in Neural Information Processing Systems*, 36:62296–62331, 2023.

Wendell H Fleming and Raymond W Rishel. *Deterministic and stochastic optimal control*, volume 1. Springer Science & Business Media, 2012.

Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. Auditing and instructing text-to-image generation models on fairness. *AI and Ethics*, 5(3):2103–2123, Jun 2025. ISSN 2730-5961. doi: 10.1007/s43681-024-00531-5.

Yingqing Guo, Hui Yuan, Yukang Yang, Minshuo Chen, and Mengdi Wang. Gradient guidance for diffusion models: An optimization perspective. *Advances in Neural Information Processing Systems*, 37:90736–90770, 2024.

Yinbin Han, Meisam Razaviyayn, and Renyuan Xu. Stochastic control for fine-tuning diffusion models: Optimality, regularity, and convergence. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=DnNV3Ea09e.

Ce Hao, Anxing Xiao, Zhiwei Xue, and Harold Soh. CHD: Coupled hierarchical diffusion for long-horizon tasks. In *9th Annual Conference on Robot Learning*, 2025. URL https://openreview.net/forum?id=tXY6VQlXfA.

Akio Hayakawa, Masato Ishii, Takashi Shibuya, and Yuki Mitsufuji. MMDisco: Multi-modal discriminator-guided cooperative diffusion for joint audio and video generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=agbiPPuSeQ.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Ruifei He, Chuhui Xue, Haoru Tan, Wenqing Zhang, Yingchen Yu, Song Bai, and Xiaojuan Qi. Debiasing text-to-image diffusion models. In *Proceedings of the 1st ACM Multimedia Workshop on Multi-Modal Misinformation Governance in the Era of Foundation Models*, MIS '24, pp. 29–36, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400712012. doi: 10.1145/3689090.3689387. URL https://doi.org/10.1145/3689090.3689387.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

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.

Michael Janner, Yilun Du, Joshua B. Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning*, 2022.

Yilei Jiang, Wei-Hong Li, Yiyuan Zhang, Minghong Cai, and Xiangyu Yue. Fairgen: Enhancing fairness in text-to-image diffusion models via self-discovering latent directions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18411–18420, October 2025.

Mintong Kang, Vinayshekhar Bannihatti Kumar, Shamik Roy, Abhishek Kumar, Sopan Khosla, Balakrishnan Murali Narayanaswamy, and Rashmi Gangadharaiah. FairGen: Controlling sensitive attributes for fair generations in diffusion models via adaptive latent guidance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25336–25350. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1287.

Kimmo Karkkainen and Jungseock Joo. FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1548–1558, 2021.

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

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=k7FuTOWMOc7.

Alex Krizhevsky et al. Learning multiple layers of features from tiny images. *Technical Report*, 2009.

W. Levine. Optimal control theory: An introduction. *IEEE Transactions on Automatic Control*, 17 (3):423–423, 1972. doi: 10.1109/TAC.1972.1100008.

Henry Li and Marcus Aloysius Pereira. Solving inverse problems via diffusion optimal control. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=wqLC4G1GN3.

Jia Li, Lijie Hu, Jingfeng Zhang, Tianhang Zheng, Hua Zhang, and Di Wang. Fair text-to-image diffusion via fair mapping. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39 (25):26256–26264, 2025. doi: 10.1609/aaai.v39i25.34823. URL https://ojs.aaai.org/index.php/AAAI/article/view/34823.

Qianxiao Li, Long Chen, Cheng Tai, and Weinan E. Maximum principle based algorithms for deep learning. *Journal of Machine Learning Research*, 18(165):1–29, 2018. URL http://jmlr.org/papers/v18/17-653.html.

Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Yusuke Kato, and Kazuki Kozuka. Aligning diffusion models by optimizing human utility. *Advances in Neural Information Processing Systems*, 37:24897–24925, 2024.

Jinhao Liang, Jacob K Christopher, Sven Koenig, and Ferdinando Fioretto. Simultaneous multi-robot motion planning with projected diffusion models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Sp7jclUwkV.

Buhua Liu, Shitong Shao, Bao Li, Lichen Bai, Zhiqiang Xu, Haoyi Xiong, James Kwok, Sumi Helal, and Zeke Xie. Alignment of diffusion models: Fundamentals, challenges, and future. *arXiv preprint arXiv:2409.07253*, 2024.

Zhen Liu, Tim Z. Xiao, Carles Domingo-Enrich, Weiyang Liu, and Dinghuai Zhang. Value gradient guidance for flow matching alignment. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=6MmOy2Ji8V.

Hao Luan, Yi Xian Goh, See-Kiong Ng, and Chun Kai Ling. Projected coupled diffusion for test-time constrained joint generation. *arXiv preprint arXiv:2508.10531*, 2025a.

Hao Luan, See-Kiong Ng, and Chun Kai Ling. DDPS: Discrete diffusion posterior sampling for paths in layered graphs. In *ICLR 2025 Frontiers in Probabilistic Inference: Learning Meets Sampling Workshop*, 2025b. URL https://openreview.net/forum?id=DBdkU0Ikzy.

Manuel Madeira, Clément Vignac, Dorina Thanou, and Pascal Frossard. Generative modelling of structurally constrained graphs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 137218–137262, 2024.

Marianthi Markatou, Yang Chen, Georgios Afendras, and Bruce G Lindsay. Statistical distances and their role in robustness. In *New advances in statistics and data science*, pp. 3–26. Springer, 2018.

Zichen Miao, Jiang Wang, Ze Wang, Zhengyuan Yang, Lijuan Wang, Qiang Qiu, and Zicheng Liu. Training diffusion models towards diverse image generation with reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10844–10853, 2024.

Mashrur M. Morshed and Vishnu Boddeti. Diverseflow: Sample-efficient diverse mode coverage in flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23303–23312, June 2025.

Chenhao Niu, Yang Song, Jiaming Song, Shengjia Zhao, Aditya Grover, and Stefano Ermon. Permutation invariant graph generation via score-based generative modeling. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 4474–4484. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/niu20a.html.

Kushagra Pandey, Farrin Marouf Sofian, Felix Draxler, Theofanis Karaletsos, and Stephan Mandt. Variational control for guidance in diffusion models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2025.

Rishubh Parihar, Abhijnya Bhat, Abhipsa Basu, Saswat Mallick, Jogendra Nath Kundu, and R. Venkatesh Babu. Balancing act: Distribution-guided debiasing in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6668–6678, June 2024.

Byoungwoo Park, Jungwon Choi, Sungbin Lim, and Juho Lee. Stochastic optimal control for diffusion bridges in function spaces. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=WyQW4G57Zd.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Litu Rout, Yujia Chen, Nataniel Ruiz, Abhishek Kumar, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. RB-modulation: Training-free stylization using reference-based modulation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=bnINPG5A32.

Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10219–10228, 2023.

Xudong Shen, Chao Du, Tianyu Pang, Min Lin, Yongkang Wong, and Mohan Kankanhalli. Finetuning text-to-image diffusion models for fairness. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hnrB5YHoYu.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, pp. 11895–11907, 2019.

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*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

Masatoshi Uehara, Yulai Zhao, Chenyu Wang, Xiner Li, Aviv Regev, Sergey Levine, and Tommaso Biancalani. Inference-time alignment in diffusion models with reward-guided generation: Tutorial and review. *arXiv preprint arXiv:2501.09685*, 2025.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.

Luran Wang, Chaoran Cheng, Yizhen Liao, Yanru Qu, and Ge Liu. Training free guided flow-matching with optimal control. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=61ss5RA1MM.

Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Zou, and Stefano Ermon. TFG: Unified training-free guidance for diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 22370–22417, 2024.

Stefano Zampini, Jacob K Christopher, Luca Oneto, Davide Anguita, and Ferdinando Fioretto. Training-free constrained generation with stable diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=TrNB08KuHK.

Yu Zeng, Vishal M Patel, Haochen Wang, Xun Huang, Ting-Chun Wang, Ming-Yu Liu, and Yogesh Balaji. Jedi: Joint-image diffusion models for finetuning-free personalized text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6786–6795, 2024.

Kaizhen Zhu, Mokai Pan, Yuexin Ma, Yanwei Fu, Jingyi Yu, Jingya Wang, and Ye Shi. UniDB: A unified diffusion bridge framework via stochastic optimal control. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=uqCfoVXb67.

A THEORETICAL DERIVATIONS

Proposition A.1 (Proposition 4.1)**.** *For any fixed $t \in [0, T]$ and fixed $(\boldsymbol{x}_t, \boldsymbol{\nu}_t)$, with $\rho + \gamma > 0$, the optimal control $\mathbf{u}_t^*$ specified in Eq. (14) is given by*

$$\mathbf{u}_t^* \in \Pi_{\mathcal{U}}(\bar{\boldsymbol{u}}_t) := \arg\min_{\mathbf{u} \in \mathcal{U}} \|\mathbf{u} - \bar{\boldsymbol{u}}_t\|^2,$$

where

$$\bar{\boldsymbol{u}}_t := \frac{1}{\rho + \gamma} \left(\gamma \, \boldsymbol{u}_t^{\text{ref}} - \mathbf{g}(\boldsymbol{x}_t, t)^\top \boldsymbol{\nu}_t \right),$$

and $\Pi_{\mathcal{U}}(\bar{\boldsymbol{u}}_t)$ is nonempty. Further, $\mathbf{u}_t^*$ is unique if $\mathcal{U}$ is convex.

Proof. For any fixed $t, (\boldsymbol{x}_t, \boldsymbol{\nu}_t)$, extract the part of $\widetilde{H}$ that depends on $\mathbf{u}$ as

$$\Lambda_t(\mathbf{u}) := \frac{\rho}{2}\|\mathbf{u}\|^2 + \frac{\gamma}{2}\|\mathbf{u} - \boldsymbol{u}_t^{\text{ref}}\|^2 + \boldsymbol{\nu}_t^\top \mathbf{g}(\boldsymbol{x}_t, t)\mathbf{u}.$$

Since $\rho + \gamma > 0$, completing squares yields

$$\Lambda_t(\mathbf{u}) = \frac{\rho + \gamma}{2}\|\mathbf{u} - \bar{\boldsymbol{u}}_t\|^2 + c_t,$$

with

$$\bar{\boldsymbol{u}}_t = \frac{1}{\rho + \gamma} \left(\gamma \, \boldsymbol{u}_t^{\text{ref}} - \mathbf{g}(\boldsymbol{x}_t, t)^\top \boldsymbol{\nu}_t \right)$$

and

$$c_t = \frac{\gamma}{2}\|\boldsymbol{u}_t^{\text{ref}}\|^2 - \frac{\rho + \gamma}{2}\|\bar{\boldsymbol{u}}_t\|^2.$$

Therefore,

$$\widetilde{H}(\boldsymbol{x}_t, \boldsymbol{\nu}_t, \mathbf{u}, t) = \frac{\rho + \gamma}{2}\|\mathbf{u} - \bar{\boldsymbol{u}}_t\|^2 + \boldsymbol{\nu}_t^\top \mathbf{f}^\theta(\boldsymbol{x}_t, t) + c_t.$$

The last two terms do not depend on $\mathbf{u}$, hence

$$\arg\min_{\mathbf{u} \in \mathcal{U}} \widetilde{H}(\boldsymbol{x}_t, \boldsymbol{\nu}_t, \mathbf{u}, t) = \arg\min_{\mathbf{u} \in \mathcal{U}} \|\mathbf{u} - \bar{\boldsymbol{u}}_t\|^2 = \Pi_{\mathcal{U}}(\bar{\boldsymbol{u}}_t).$$

If $\mathcal{U}$ is bounded, since function $f(\mathbf{u}) = \|\mathbf{u} - \bar{\boldsymbol{u}}_t\|^2$ is continuous, the minimum can be attained; if $\mathcal{U}$ is unbounded, the minimum can also be attained because $\|\mathbf{u} - \bar{\boldsymbol{u}}_t\|^2$ is coercive. Hence, $\Pi_{\mathcal{U}}(\bar{\boldsymbol{u}}_t)$ is nonempty. Further, if $\mathcal{U}$ is convex, since $\|\mathbf{u} - \bar{\boldsymbol{u}}_t\|^2$ is strictly convex in $\mathbf{u}$, the minimizer is unique. □

B EXPERIMENT DETAILS

Software and Codebase All experiments were run using PyTorch (Paszke et al., 2019). For the CIFAR experiment, our implementation builds upon (Karras et al., 2022) and the implementation of the image classifier used as the attribute model therein is from (Chen, 2025). For the human face generation experiment, implementations are based on (Choi et al., 2022) and image classifier implementation is from (Luan et al., 2025a).

Computational Hardware The training of the unconditional EDM model in the CIFAR experiment and *all* evaluations (including computing all metrics) were run on a workstation with 1 AMD Ryzen Threadripper PRO 5995WX 64-Core CPU, 504 GB RAM, and 2 NVIDIA RTX A6000 GPUs each with 48GB VRAM. Inference of all methods was run on a high-performance-computing (HPC) cluster with NVIDIA H200 GPUs. For each experiment run, only 1 GPU was utilized.

B.1 ALGORITHM

The algorithmic description of our method is in Algorithm 1.

Algorithm 1 DADA via E-MSA with Euler discretization

Require: Dynamics $\mathbf{F}^\theta$; terminal cost Φ; batch init $\boldsymbol{X}_{\text{init}}$; time grid $\{t_k\}_{k=0}^K$ with $t_0 = 0, t_K = T$; $\rho > 0, \xi \geq 0$; max iteration MaxIter; optional $\mathcal{U}$.

$\eta \leftarrow (1-\xi)/\rho$
$h_k \leftarrow t_{k+1} - t_k$ for $k = 0, \ldots, K-1$
$\mathbf{U}_k \leftarrow \mathbf{0}$ for $k = 0, \ldots, K-1$
for $j = 1$ to MaxIter **do**
  ▷ *(1) Forward pass with control* $\mathbf{U}$ ◁
  $\mathbf{X}_0 \leftarrow \mathbf{X}_{\text{init}}$
  for $k = 0, \ldots, K-1$ **do**
    $\mathbf{X}_{k+1} \leftarrow \mathbf{X}_k + h_k\, \mathbf{F}^\theta(\mathbf{X}_k, \mathbf{U}_k, t_k)$
  ▷ *(2) Cost evaluation and backward pass* ◁
  $\mathbf{N}_K \leftarrow \nabla_{\mathbf{X}} \Phi\big(\hat{p}^u_{\mathbf{Y}}(\mathbf{X}_K)\big)$
  for $k = K-1, \ldots, 0$ **do**
    ▷ *Computed via VJP* ◁
    $\mathbf{V}_k \leftarrow \big(\nabla_{\mathbf{X}} \mathbf{F}^\theta(\mathbf{X}_k, \mathbf{U}_k, t_k)\big)^\top \mathbf{N}_{k+1}$
    $\mathbf{N}_k \leftarrow \mathbf{N}_{k+1} + h_k \mathbf{V}_k$
  ▷ *(3) Closed-form control update* ◁
  for $k = 0, \ldots, K-1$ **do**
    $\mathbf{U}_k^{\text{new}} \leftarrow \xi \mathbf{U}_k - \eta\, \mathbf{G}(X_k, t_k)^\top \mathbf{N}_{k+1}$
    $\mathbf{U}_k \leftarrow \Pi_{\mathcal{U}^M}(\mathbf{U}_k^{\text{new}})$ if $\mathcal{U}$ specified else $\mathbf{U}_k^{\text{new}}$
  if Converged **then break**
▷ *Forward simulation with final control* ◁
for $k = 0, \ldots, K-1$ **do**
  $\mathbf{X}_{k+1} \leftarrow \mathbf{X}_k + h_k\, \mathbf{F}^\theta(\mathbf{X}_k, \mathbf{U}_k, t_k)$
return $\mathbf{X}_K$

B.2 CIFAR Experiment

Hierarchical Attribute Distributions We construct three different levels of class labels with a coarse-to-fine hierarchy: `meta5`, `coarse`, and `fine`, with numbers of classes of 5, 20, and 100, respectively. The `coarse` and `fine` levels of class labels are native in CIFAR-100. The `meta5` level is obtained by merging every 4 `coarse` classes. For each class level, we choose three different attribute distributions as test-time targets : Uniform, ZigZag, and Gaussian. ZigZag has an alternating "high-low" pattern where every even-numbered class is exactly twice as likely to be selected as any odd-numbered class. For Gaussian resembles a smooth bell curve centered in the middle-numbered class, with the standard deviation as $1/4$ of the support size.

Base Diffusion and Attribute Classifiers We trained a base EDM model on CIFAR-100 dataset (Krizhevsky et al., 2009) with the default training split and the network backbone is the UNet in (Song & Ermon, 2019). The EDM training followed default hyperparameters disclosed in (Karras et al., 2022). All attribute models are ResNet56 trained on CIFAR-100 training set, we adopt the implementations and training hyperparameters in (Chen, 2025).

Diffusion Inference and Evaluation For diffusion inference for all methods, we adopt the default settings provided in (Karras et al., 2022), with $K = 18$ sampling steps, but all with Euler discretization. For each method, we sample 10240 images and evaluate all metrics based on the empirical distribution of attributes.

Particle Guidance Implementation We implement the Particle Guidance (PG) (Corso et al., 2024) method by taking the same cost function used in our method as the potential field for PG. PG operating on the PF-ODE of diffusion is as follows:

$$\frac{\mathrm{d}\mathbf{x}_t^{[i]}}{\mathrm{d}t} = -\boldsymbol{f}(\mathbf{x}_t^{[i]}, t) + \frac{1}{2} g(t)^2 \left(\mathbf{s}^\theta \left(\mathbf{x}_t^{[i]}, t \right) + \nabla_{\mathbf{x}_t^{[i]}} \log \mathcal{L} \left(\mathbf{x}_t^{[1]}, \ldots, \mathbf{x}_t^{[M]} \right) \right), \tag{17}$$

where $\log \mathcal{L}$ is a potential field defined over M samples. Concretely, we set

$$\log \mathcal{L}\left(\mathbf{x}_t^{[1]}, \ldots, \mathbf{x}_t^{[M]}\right) := -\, w\, \mathbb{D}_{\mathrm{KL}}\left[\hat{p}_{\mathbf{y}}^t \,\|\, p_{\mathbf{y}}^{\mathrm{tar}}\right], \quad \hat{p}_{\mathbf{y}}^t := \frac{1}{M}\sum_{i=1}^{M} \mathrm{softmax}\left(\Psi\left(\hat{\mathbf{x}}_T^{[i]}(\mathbf{x}_t^{[i]})\right)\right) \tag{18}$$

where $w > 0$ is a hyperparameter, and $\hat{\mathbf{x}}_T^{[i]}(\mathbf{x}_t^{[i]})$ is a "predicted clean sample" given a noisy sample at time t, obtained via Tweedie's formula (Efron, 2011). Under EDM (Karras et al., 2022) with a learned score model $\mathbf{s}^\theta$, it takes the form of

$$\hat{\mathbf{x}}_T^{[i]}(\mathbf{x}_t^{[i]}) = \mathbf{x}_t^{[i]} + t^2 \mathbf{s}^\theta(\mathbf{x}_t^{[i]}, t). \tag{19}$$

Under DDIM(Song et al., 2021a) with a learned noisd prediction model $\boldsymbol{\epsilon}^\theta$, we take the PF-ODE in $\tilde{\mathbf{x}}_t$ (see Eq. (3)) and this term reads

$$\hat{\tilde{\mathbf{x}}}_T^{[i]}(\tilde{\mathbf{x}}_t^{[i]}) = \tilde{\mathbf{x}}_t^{[i]} - \sigma(t)\boldsymbol{\epsilon}^\theta\left(\tilde{\mathbf{x}}_t^{[i]}/\sqrt{1+\sigma(t)^2},\, t\right). \tag{20}$$

Hyperparameters The hyperparameters used to obtain the results in Table 1 are reported in Table 4 and Table 5.

Table 4: Hyperparameters used for our method in CIFAR.

Target	MaxIter	ρ	M	ξ
Uniform-`meta5`	10	0.1	32	0.99
Uniform-`coarse`	10	0.1	64	0.99
Uniform-`fine`	10	0.1	256	0.99
ZigZag-`meta5`	10	0.05	32	0.99
ZigZag-`coarse`	10	0.05	64	0.99
ZigZag-`fine`	10	0.05	256	0.99
Gaussian-`meta5`	10	0.1	64	0.95
Gaussian-`coarse`	10	0.1	128	0.95
Guassian-`fine`	10	0.1	256	0.95

Table 5: Hyperparameters used for PG-DPS in CIFAR.

Target	w	M
Uniform-`meta5`	4.0	32
Uniform-`coarse`	4.0	64
Uniform-`fine`	4.0	256
ZigZag-`meta5`	4.0	32
ZigZag-`coarse`	4.0	64
ZigZag-`fine`	4.0	256
Gaussian-`meta5`	4.0	64
Gaussian-`coarse`	4.0	128
Guassian-`fine`	4.0	256

B.3 Fair Human Face Generation Experiment

Implementation details We use the pretrained DDIM weights from (Choi et al., 2022) for the base diffusion diffusion model. For the `race` attribute, since the FFHQ-Aging dataset does not have a label for it, we leverage a pretrained classifier from (Karkkainen & Joo, 2021) to label all images in FFHQ-Aging and use them as ground truth. The attribute model is a lightweight ResNet image classifier implemented by Luan et al. (2025a), and we train the classifier on FFHQ-Aging (with the added labels for `race`). For all methods, we take $K = 25$ sampling steps for during diffusion inference with FP16 precision, with other inference hyperparameters following the defaults set in (Choi et al., 2022). We sample 960 images for each method in the single-attribute distribution alignment experiments (Table 2) for evaluation, and sample 1080 images for each method in the joint attribute distribution alignment cases (Table 3).

Hyperparameters The hyperparameters specific to each method used to obtain the results in Table 2 and Table 3 are reported in Table 6 and Table 7. The selection of hyperparameters is based on each method's best performance on the TV metric.

Table 6: Hyperparameters used for our method in human face generation.

Target	MaxIter	ρ	M	ξ
Custom1-age	12	0.0015	24	0.9500
Custom1-gender	10	0.00125	20	0.9500
Custom1-race	10	0.0005	20	0.9500
Custom2-age	10	0.0010	24	0.9500
Custom2-gender	10	0.00075	20	0.9500
Custom2-race	10	0.0010	20	0.9500
CustomJoint	10	0.0002	72	0.9500
Uniform-age	10	0.002	24	0.9500
Uniform-gender	10	0.001	20	0.9500
Uniform-race	10	0.0005	20	0.9500
UniformJoint	10	0.0002	72	0.9500

Table 7: Hyperparameters used for PG-DPS in human face generation.

Target	w	M
Custom1-age	10	24
Custom1-gender	50	20
Custom1-race	50	20
Custom2-age	10	24
Custom2-gender	50	20
Custom2-race	100	20
CustomJoint	4	72
Uniform-age	15	24
Uniform-gender	10	20
Uniform-race	10	20
UniformJoint	7	72

C Additional Results

C.1 Additional Results for CIFAR Experiment

Ablation Study We perform ablation study in the batch size M used for empirical distribution calculation. We set $M \in \{8, 16, 32, 64, 128, 256\}$ and test all methods with all of the three types of target distributions (Uniform, Guassian, ZigZag) with all three levels of attributes (`meta5`, `coarse`, `fine`). The resulting metrics are plotted in Figure 3, Figure 4, and Figure 5. The figures show that across all target distributions with different support sizes, the performance (in terms of distributional metrics) appears to first improve as the batch size M increases, and then start to slowly degrade if M keeps growing (note the log scale of the x-axis in all figures). A similar pattern is also reported by Parihar et al. (2024) when using sample batches to perform distribution guidance. The turning point appears to vary across different targets: ~ 5 times of the support size in Uniform, while up to 10 times of the support size in Guassian and ZigZag. For baseline PG-DPS, the performance change appears to be minor as the batch size M changes. With a reasonable batch size M for estimating the empirical distribution (considering both attribute distribution support size and the target), our method can achieve better performance than the baselines.

C.2 Additional Results for Face Generation Experiment

Additional Samples. We present additional qualitative samples in Figure 6, Figure 8, Figure 9.

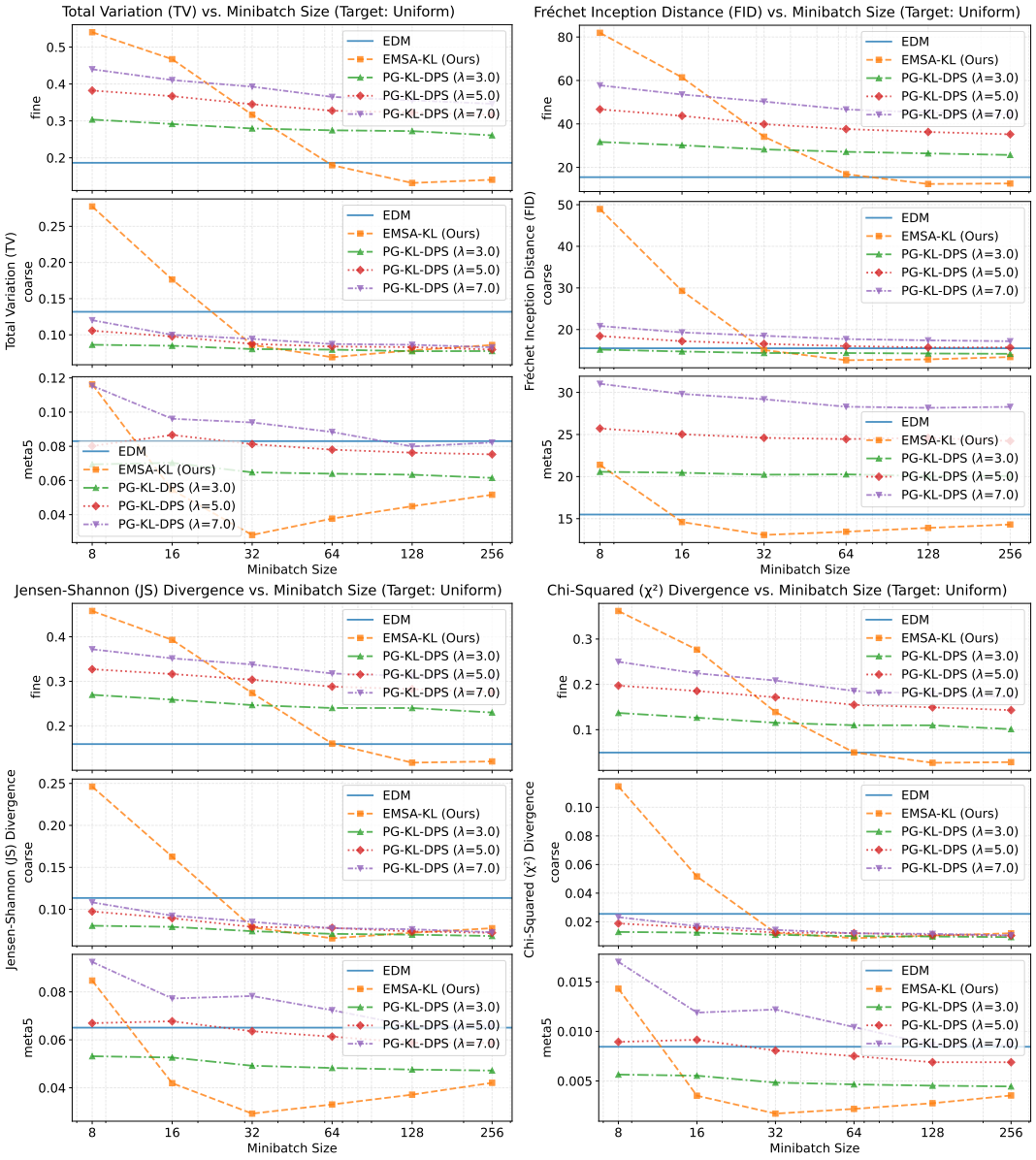


Figure 3: Total Variation (top left), FID (top right), Jensen-Shannon divergence, and χ^2 divergence metrics with different batch size M for target Uniform.

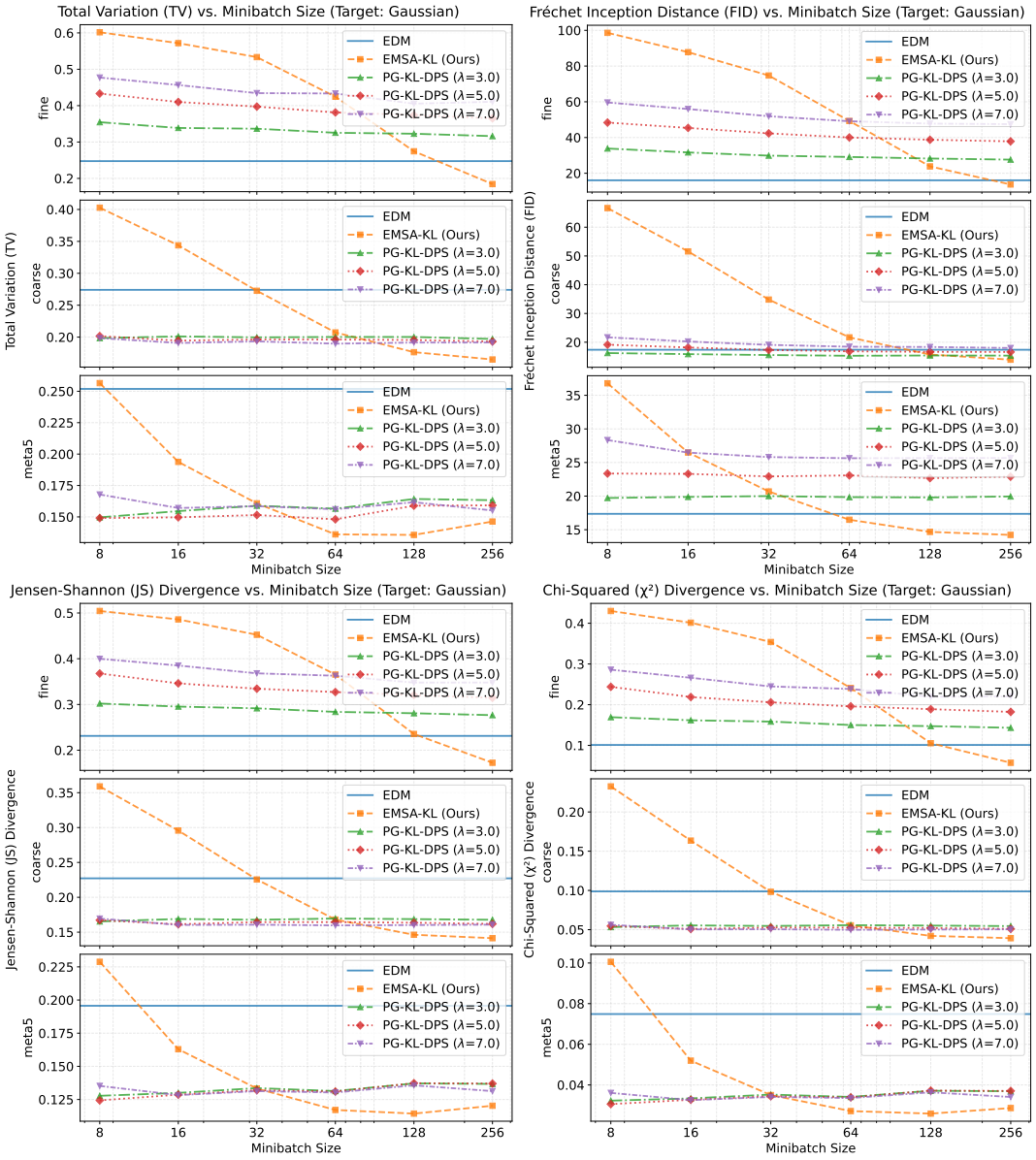


Figure 4: Total Variation (top left), FID (top right), Jensen-Shannon divergence, and χ^2 divergence metrics with different batch size M when target is Guassian.

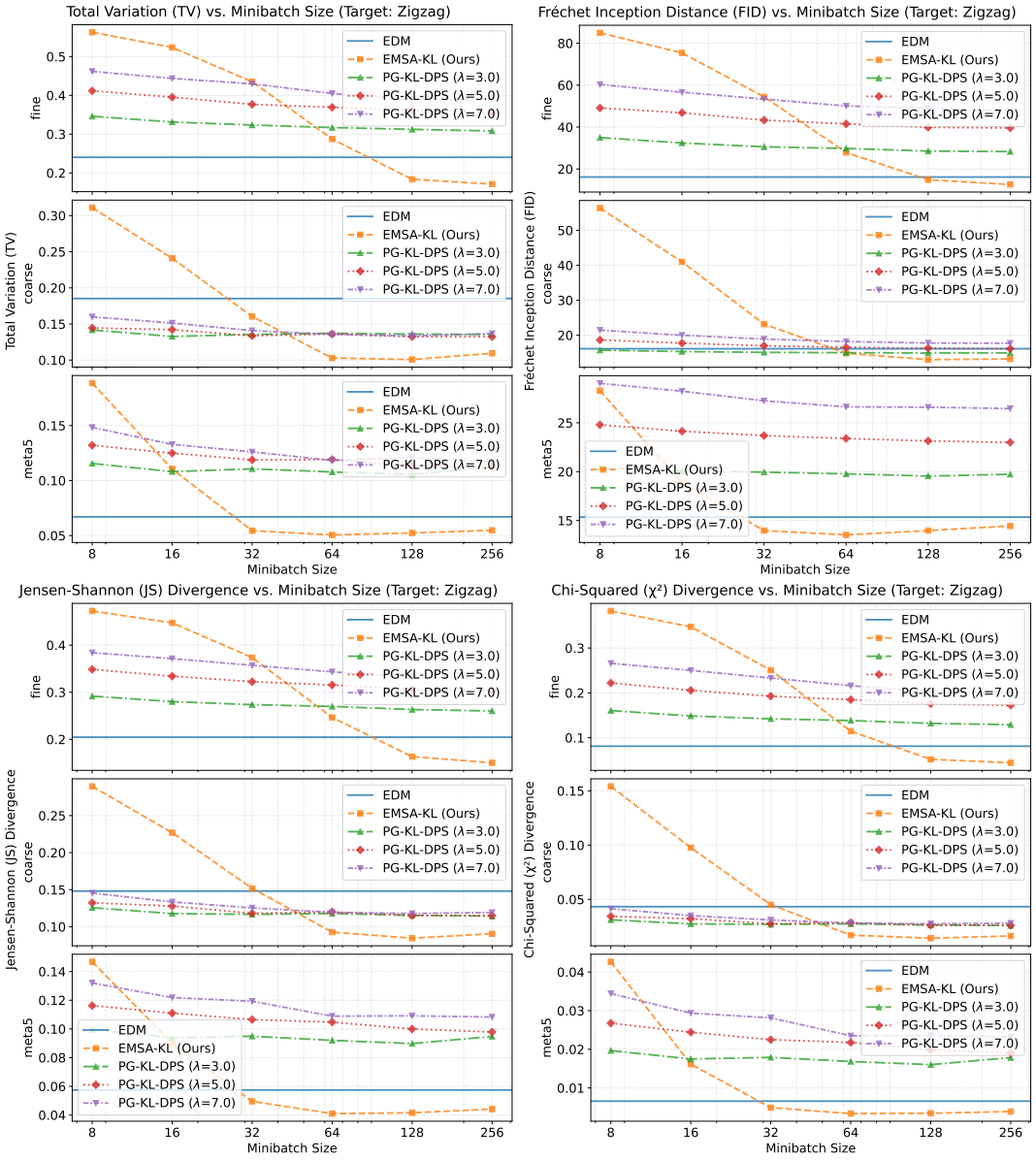


Figure 5: Total Variation (top left), FID (top right), Jensen-Shannon divergence, and χ^2 divergence metrics with different batch size M when target is Guassian.

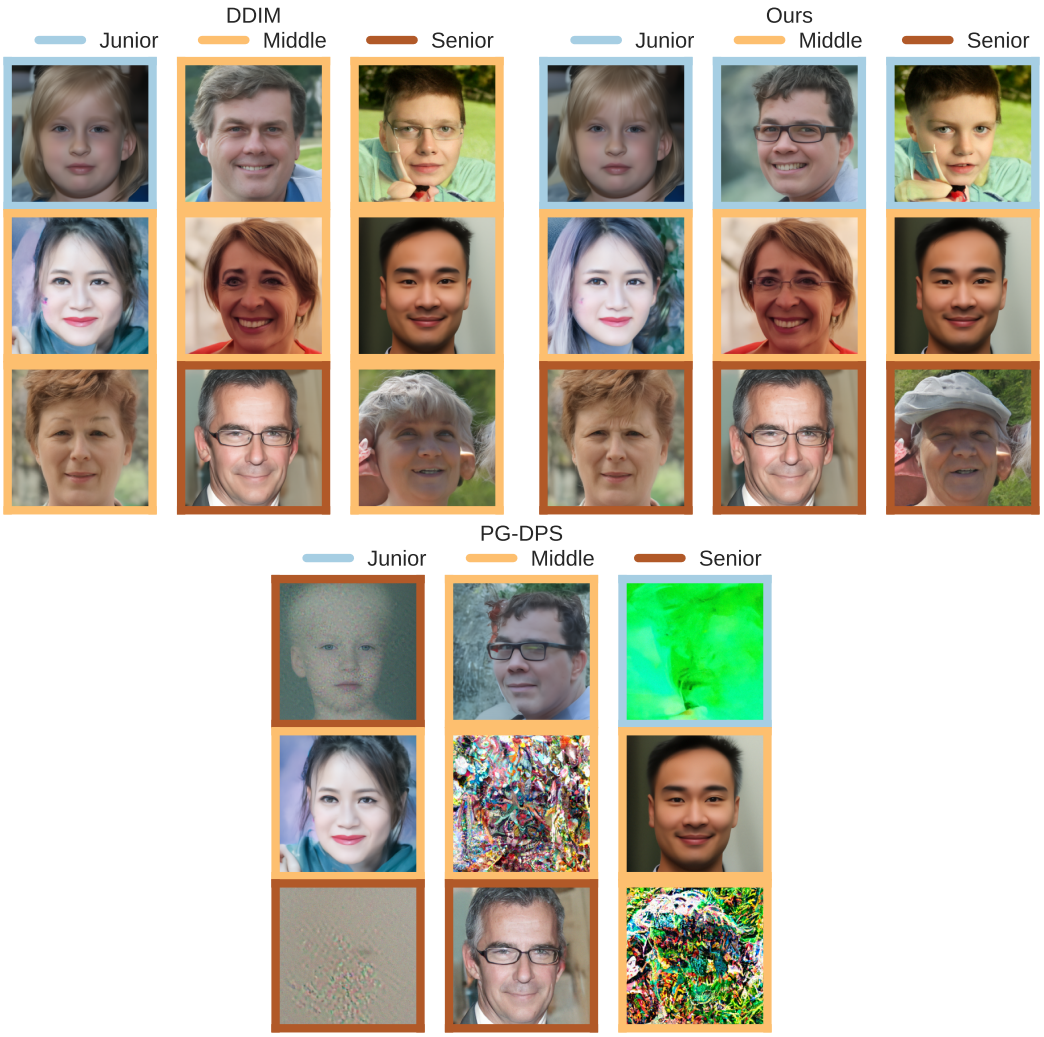


Figure 6: Additional samples from all methods in the face generation experiment.

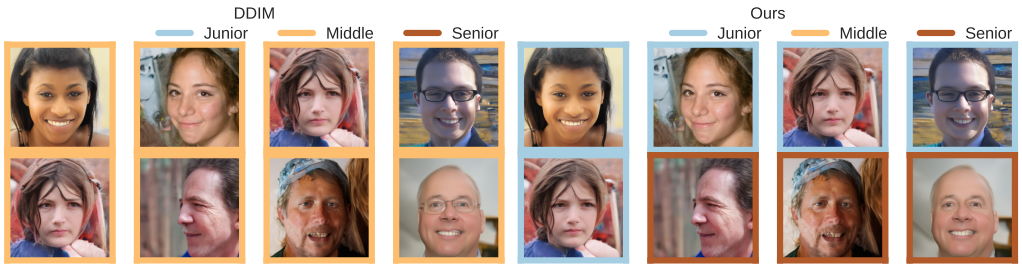


Figure 7: Our method could exploit a potentially vulnerable attribute model by introducing unnoticeable perturbations into samples.

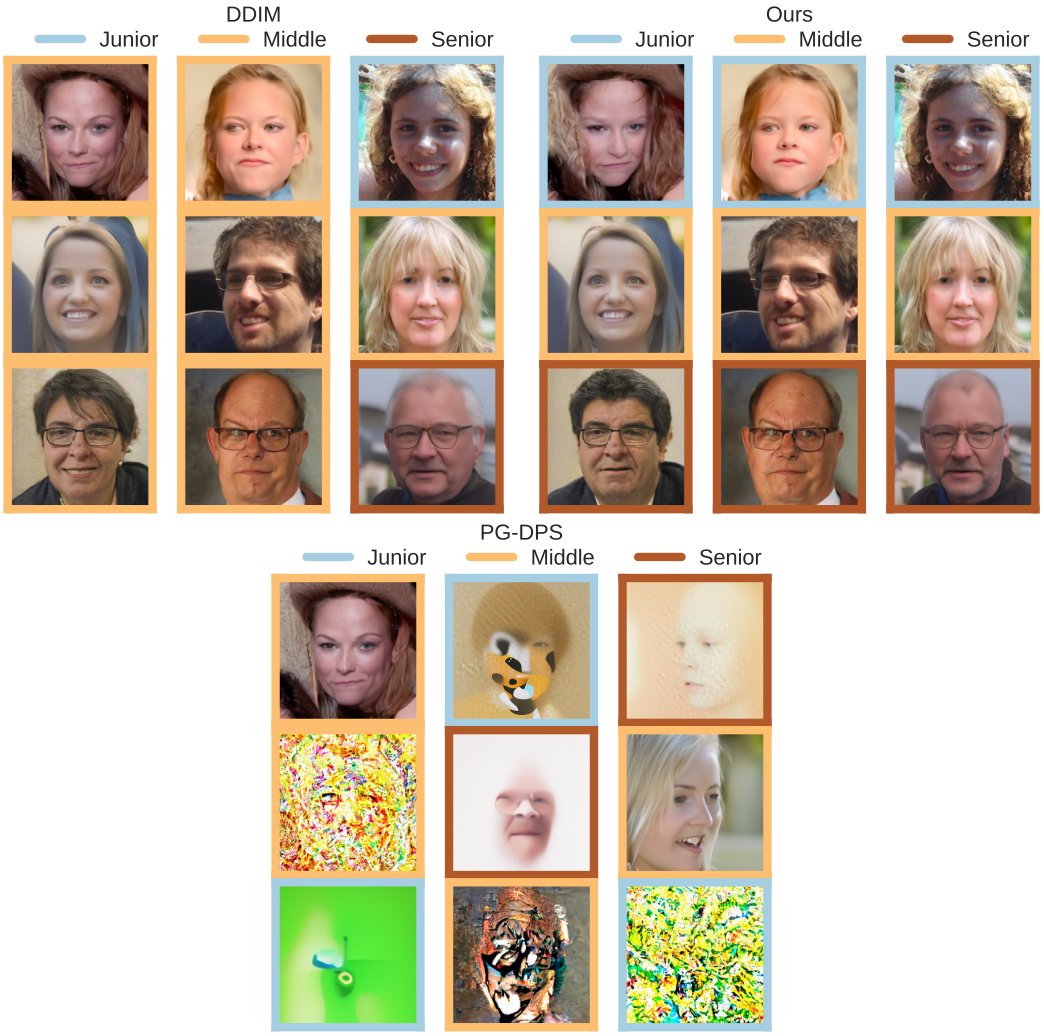


Figure 8: Additional samples from all methods in the face generation experiment.

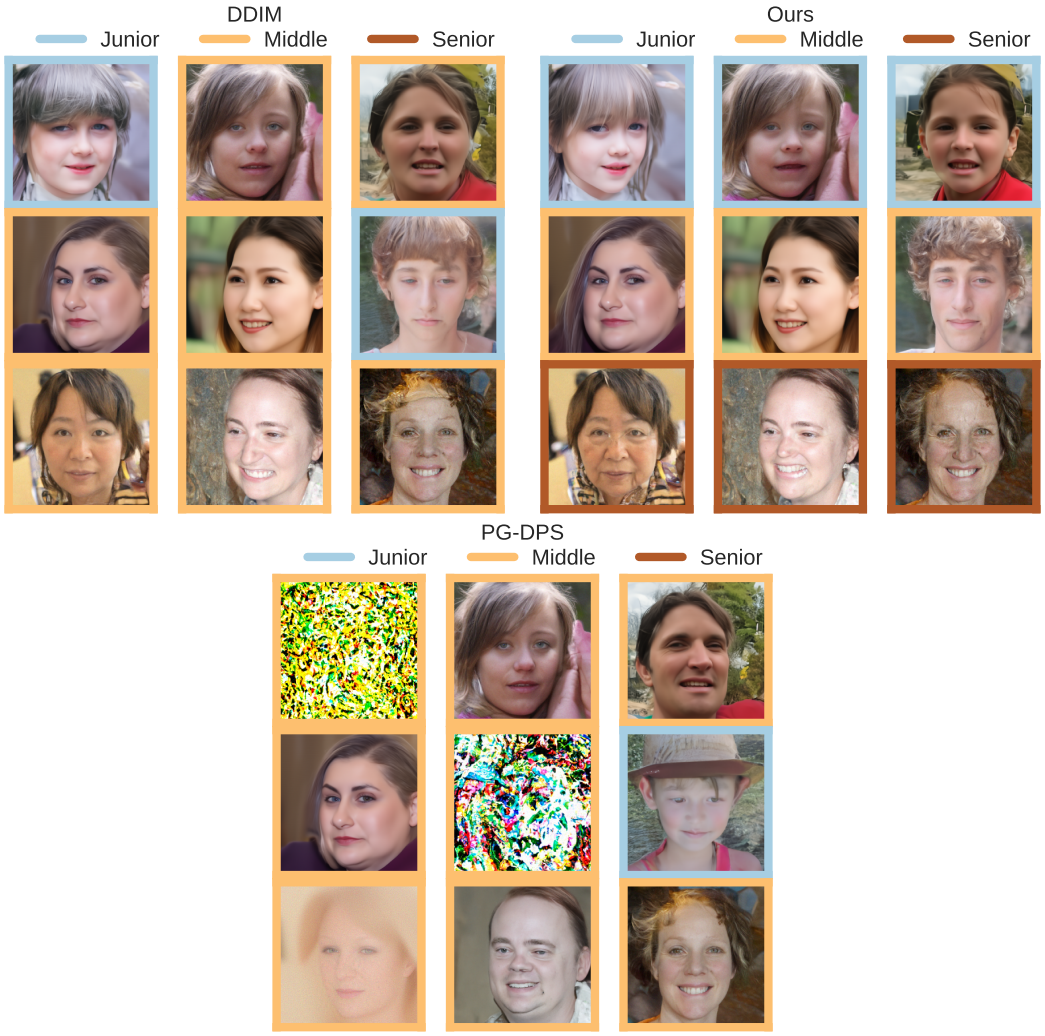


Figure 9: Additional samples from all methods in the face generation experiment.

D Discussions and Limitations

Computational Cost The computational complexity of our method scales linearly with the number of optimization iterations MaxIter, diffusion PF-ODE sampling steps K, and batch size M. The primary bottleneck is the per-step VJP computation, which necessitates differentiating through the diffusion model. Empirically, we find MaxIter $\approx 6-12$ suffices, and early stopping can further reduce cost when the distributional objective plateaus. While our approach incurs a higher inference cost than vanilla sampling, it effectively eliminates the need for expensive model fine-tuning or retraining and yields stronger distribution-level alignment than per-step guidance baselines in our experiments. We view this computational overhead as a necessary trade-off for high-fidelity distributional matching without parameter updates.

Batch Size for Empirical Distribution Estimation Our method relies on a reasonable batch size M for empirical distribution estimation. As shown in the ablation study, when the batch size M is too small to provide a reliable empirical distribution for an attribute distribution with a potentially large support size, the distribution aligning performance would degrade. However, the same issue exists for other approaches including training-based methods such as (Parihar et al., 2024).

Potential Attack for Attribute Model Interestingly, our method might yield adversarial samples against an imperfect attribute model. The iterative optimization process in our method can introduce tiny perturbations that are barely noticeable to humans in image samples and "trick" the classifier into producing different attribute outputs. See Figure 7 for some examples. Exploiting this vulnerability is consistent with our problem formulation, so the method can find such a "shortcut" to achieve the alignment objective with minimal control effort. We view this issue as a limitation of the attribute model rather than of our method.