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, NUS

haoluan@comp.nus.edu.sg, {seekiong,chunkail}@nus.edu.sg

Paper Code

TL;DR

We formulate **attribute distributional alignment** of diffusion models as an **optimal control (OC) problem** and solve it at **inference time** without retraining.

- **Training-free:** Works with any pretrained unconditional diffusion model
- **Flexible targets:** Change target attribute distribution at test time
- **Alignment + quality:** Align attribute distributions and preserve sample quality.

Motivation & Key Insights

Problem: Real applications need *population-level* distributional control (fairness, diversity, customization) over sample *attributes*, not only per-sample conditioning. Target attribute distributions may change frequently at test-time, so retraining per target is impractical.

Research question: Given a pretrained unconditional diffusion model, an attribute model, and an arbitrary target distribution, how to align the *attribute distribution* of generated samples to the target *at test time*?

Why optimal control?

1. Principled trade-off between alignment and data fidelity via control cost penalty.
2. Inherently test-time: changing target only changes the objective, not the model.
3. Step-wise perturbations are grounded in OC theory, not ad-hoc heuristics.

Method: OC for Attribute Distribution Alignment

Controlled reverse diffusion with additive perturbation $\mathbf{U}_t$ over a sample *batch*:

$$\dot{\mathbf{X}}_t = \mathbf{F}^\theta(\mathbf{X}_t, t) + \mathbf{G}(\mathbf{X}_t, t)^\top \mathbf{U}_t, \quad \mathbf{X}_0 = \mathbf{X}_{\text{init}} \tag{1}$$

Optimal control problem with a regularized distributional objective:

$$\min_{\mathbf{U}_t \in \mathcal{U}^M} \underbrace{\mathbb{D}_{\mathsf{KL}}\big[\,\hat{p}^u_{\mathbf{y}}(\mathbf{X}_T) \parallel p^{\text{tar}}_{\mathbf{y}}\,\big]}_{\text{alignment (terminal cost)}} + \underbrace{\frac{\rho}{2}\int_0^T \|\mathbf{U}_t\|_F^2 \,\mathrm{d}t}_{\text{data fidelity (running cost)}}$$

$$\text{s.t.} \quad \text{Dynamics } (1)$$

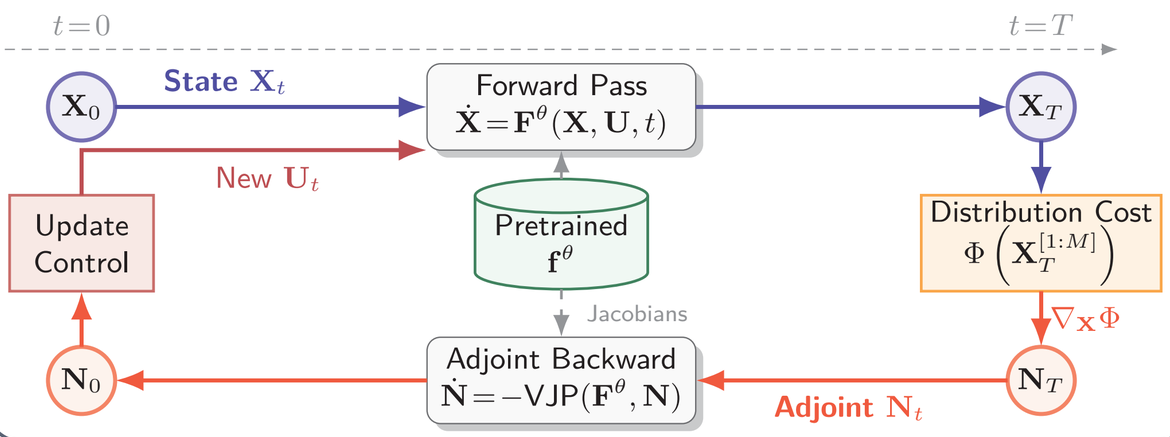
- **Terminal cost** couples all M samples via the empirical distribution $\hat{p}^u_{\mathbf{y}}$;
- **Running cost** penalizes deviations from the pretrained distribution;
- ρ controls the alignment–fidelity trade-off.

Closed-form control update rule derived by Pontryagin's Maximum Principle:

$$\mathbf{U}^*_t = \Pi_{\mathcal{U}^M}\Big(\xi\,\mathbf{U}^{\text{ref}}_t - \eta\,\mathbf{G}(\mathbf{X}_t, t)^\top \mathbf{N}_t\Big)$$

where ξ, η are hyperparameters related to ρ, $\mathbf{U}^{\text{ref}}_t$ is a reference control, $\mathbf{N}_t$ is the adjoint (costate), and $\Pi_{\mathcal{U}^M}$ denotes (optional) projection onto feasible control set.

Method Pipeline



Algorithm: E-MSA for Distribution Alignment

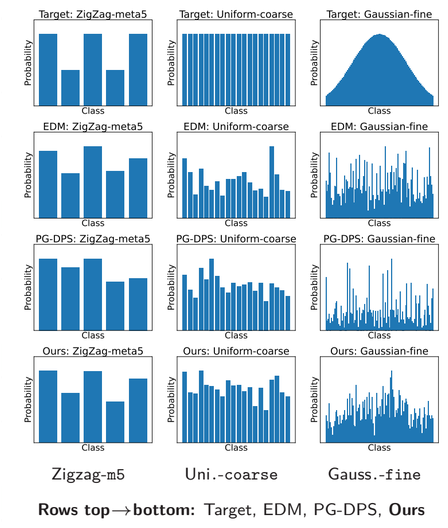
Require: Pretrained dynamics $\mathbf{F}^\theta$; terminal cost Φ; batch init $\mathbf{X}_{\text{init}}$; time grid $\{t_k\}_{k=0}^K$; $\rho > 0$; $\xi \geq 0$; MaxIter

$\eta \leftarrow (1-\xi)/\rho$; $\quad \mathbf{U}_k \leftarrow \mathbf{0}$ for all k
for $j = 1$ to MaxIter **do**
  //**Forward:**
  $\mathbf{X}_0 \leftarrow \mathbf{X}_{\text{init}}$
  for $k = 0, \ldots, K-1$ **do**
    $\mathbf{X}_{k+1} \leftarrow \mathbf{X}_k + h_k\,\mathbf{F}^\theta(\mathbf{X}_k, \mathbf{U}_k, t_k)$
  //**Backward:**
  $\mathbf{N}_K \leftarrow \nabla_{\mathbf{X}}\Phi\big(\hat{p}^u_{\mathbf{y}}(\mathbf{X}_K)\big)$
  for $k = K-1, \ldots, 0$ **do**
    $\mathbf{N}_k \leftarrow \mathbf{N}_{k+1} + h_k\big(\nabla_{\mathbf{X}}\mathbf{F}^\theta\big)^\top \mathbf{N}_{k+1}$ (VJP)
    Update: $\mathbf{U}_k \leftarrow \Pi_{\mathcal{U}^M}\big(\xi\,\mathbf{U}_k - \eta\,\mathbf{G}^\top \mathbf{N}_{k+1}\big)$
return $\mathbf{X}_K$ (aligned samples)

Proof of Concept: CIFAR-100

Setup: Unconditional EDM on CIFAR-100; semantic *class* as attribute; 3 hierarchy levels (`meta5`/`coarse`/`fine`: 5/20/100 classes) × 3 target types (Uniform, Zigzag, Gaussian); ResNet classifier as attribute model Ψ. Metrics: TV, JS, χ^2, FID ↓.

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												
EDM	.252	.196	.075	17.4	.274	.227	.099	17.4	.248	.231	.101	16.1
PG-DPS	.154	.133	.035	21.3	.195	.165	.053	16.0	.348	.299	.165	33.1
Ours	**.136**	**.117**	**.027**	**16.5**	**.176**	**.146**	**.042**	**15.7**	**.184**	**.173**	**.057**	**13.7**
Zigzag												
EDM	.067	.057	.007	15.3	.185	.148	.043	16.1	.240	.205	.081	16.1
PG-DPS	.113	.100	.020	22.0	.133	.116	.027	15.7	.341	.288	.156	33.8
Ours	**.054**	**.050**	**.005**	**14.0**	**.103**	**.093**	**.017**	**14.8**	**.171**	**.150**	**.044**	**12.6**
Uniform												
EDM	.083	.065	.008	15.5	.132	.114	.026	15.5	.186	.159	.050	15.5
PG-DPS	.069	.053	.006	22.4	.081	.074	.011	15.2	.287	.255	.123	30.7
Ours	**.028**	**.029**	**.002**	**13.1**	**.069**	**.066**	**.009**	**12.6**	**.141**	**.120**	**.028**	**12.6**



Rows top→bottom: Target, EDM, PG-DPS, **Ours**

Results: Our method achieves the best alignment *and* quality across all settings. PG-DPS can even fall below vanilla EDM (e.g., Zigzag-`meta5`, Gaussian-`fine`), showing that naive batch-level guidance is insufficient. Our method consistently maintains competitive FID while significantly reducing all distributional metrics.

Face Generation with Demographic Attributes

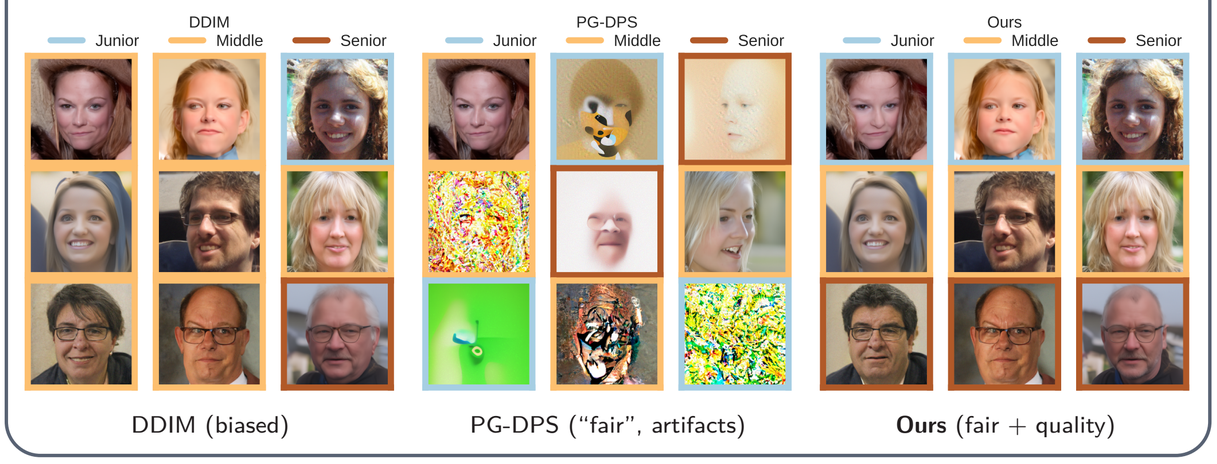
Setup: DDIM on FFHQ; 3 attributes (age, gender, race). We target both Uniform (fairness) and customized distributions over a single attribute or the *joint*.

Joint factorization: The target joint decomposes as $p^{\text{tar}}_{\mathbf{y}} \propto \prod_i p^{\text{tar}}_{\mathbf{y}_i}$ assuming attribute independence. The terminal cost then decomposes into per-attribute marginal KL divergences plus entropy correction terms:

$$\mathbb{D}_{\mathsf{KL}}\big[\,\hat{p}^u_{\mathbf{y}} \parallel p^{\text{tar}}_{\mathbf{y}}\,\big] = \sum_i \mathbb{D}_{\mathsf{KL}}\big[\,\hat{p}^u_{\mathbf{y}_i} \parallel p^{\text{tar}}_{\mathbf{y}_i}\,\big] + \sum_i \mathrm{H}(\hat{p}^u_{\mathbf{y}_i}) - \mathrm{H}(\hat{p}^u_{\mathbf{y}}).$$

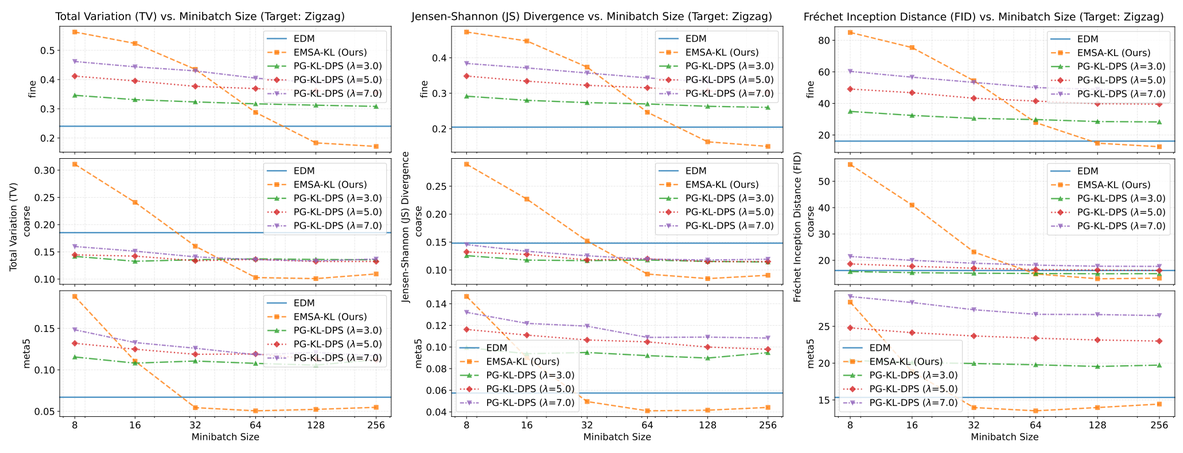
Method	age×gender×race TV↓	JS↓	χ^2↓	FD↓	FID↓
Uniform					
DDIM	0.566	0.477	0.393	0.308	46.26
PG-DPS	0.444	0.404	0.281	0.238	127.0
Ours	**0.112**	**0.107**	**0.023**	**0.061**	**41.28**
CustomJoint $[5:2:3] \times [3:7] \times [4:3:2:1]$					
DDIM	0.493	0.444	0.337	0.305	44.28
PG-DPS	0.531	0.442	0.335	0.287	74.77
Ours	**0.142**	**0.131**	**0.034**	**0.090**	**42.19**

Results: Our method achieves the best alignment across all metrics while preserving image quality. Vanilla DDIM inherits significant age and race biases from the FFHQ training set. PG-DPS aggressively trades quality for alignment yet does not consistently outperform vanilla DDIM. Our OC-based method prevents excessive deviation from the pretrained distribution, enabling simultaneous distributional alignment and sample quality.



DDIM (biased) PG-DPS ("fair", artifacts) **Ours** (fair + quality)

Ablations & Discussion



TV vs. Batch Size M JS vs. Batch Size M FID vs. Batch Size M

- **Batch size:** A reasonable M is needed for reliable empirical distribution estimation, depending on attribute distribution's support size and skewness of target.
- **Empirical Convergence:** 6–12 E-MSA iterations generally suffice in our experiments, depending on difference between the target and the pretrained model's generative distribution; early stopping is viable when the cost plateaus.
- **Runtime Cost:** Runtime scales linearly with E-MSA iterations and diffusion steps. E-MSA is $\sim 15\times$ slower than PG-DPS, but provides significantly better alignment and sample quality, and eliminates expensive retraining.