Di[M]O: Distilling Masked Diffusion Models into One-Step Generator

Yuanzhi Zhu[1], Xi Wang[1], Stéphane Lathuilière[2], Vicky Kalogeiton[1]
[1]LIX, École Polytechnique, CNRS, IPP [2]Inria, Univ. Grenoble Alpes, CNRS, LJK
https://yuanzhi-zhu.github.io/DiMO/

ICCV OCT 19-23, 2025 HONOLULU HAWAII

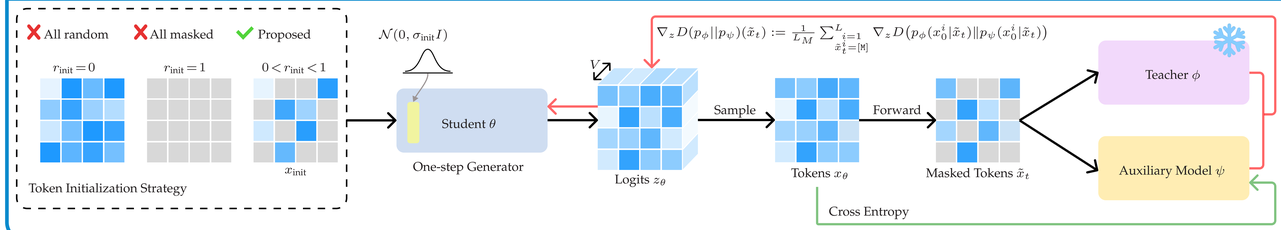
Github Code

Background

Masked Diffusion Models (MDMs) have emerged as a powerful generative modeling technique. Despite their remarkable results, they typically suffer from slow inference with several steps. In this paper, we propose Di[M]O, a novel approach that distills masked diffusion models into an ***one-step*** generator. Our contributions can be summarized as:

- We are the first to successfully achieve **one-step distillation of MDMs**.
- We propose Di[M]O, an *on policy* distillation method that enables the one-step distillation of MDMs, **with proposed efficient token initialization**.
- Our findings show that Di[M]O successfully reaches performance **close to that of MDM teachers**, while greatly enhancing the sampling efficiency.

Overview of Di[M]O Pipeline



// Generate tokens x_0
Sample $x_{\text{init}} \sim p_{\text{init}}$, $c \sim \mathcal{D}$
Compute generator logits $z_\theta(x_{\text{init}}, c) \in \mathbb{R}^{B\times h\times w\times V}$
sample token $x_\theta \in \mathbb{R}^{B\times h\times w}$ from $p_\theta(x_0|x_{\text{init}}) = \text{softmax}(z_\theta(x_{\text{init}}, c))$
// Update generator θ
Sample $t\sim\mathcal{U}[0,1]$, $\tilde{x}_t \sim q_{t|0}(\tilde{x}_t|x_\theta(x_{\text{init}}, c))$ // Forward
Calculate $p_\phi(x_0|\tilde{x}_t, c)$ and $p_\psi(x_0|\tilde{x}_t, c)$ Update θ with $\nabla_\theta \mathcal{L}_{\text{Di[M]O}}$ (Eq. (3))
// Update auxiliary model ψ
Update ψ with MDM loss to learn $x_\theta(x_{\text{init}}, c)$

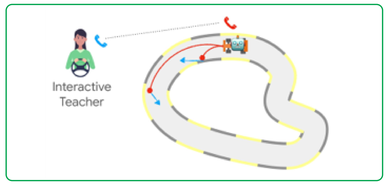
Algorithm

The core idea is *on-policy* distillation, aligning the teacher and generator across all intermediate $\tilde{x}_t$.

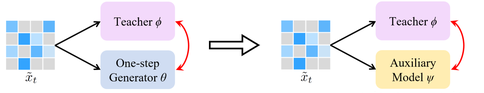
$$\mathcal{L}_{\text{Di[M]O}}(\theta) := \mathbb{E}_{x_{\text{init}},t}\left[w(t)\left(\mathbb{E}_{q_{t|0}}[D(p_\phi||p_\theta)(\tilde{x}_t)]\right)\right], \tag{1}$$

Similar to prior work (VSD, DMD, DI, SiD), we seek to approximate the following loss gradient:

$$\nabla_\theta \mathcal{L}_{\text{Di[M]O}} = \mathbb{E}_{x_{\text{init}},t}\left[w(t)\left(\mathbb{E}_{q_{t|0}}\left[\nabla_{z_\theta} D(p_\phi||p_\theta)(\tilde{x}_t)\,\frac{\mathrm{d}z_\theta(\tilde{x}_t)}{\mathrm{d}\theta}\right]\right)\right], \tag{2}$$



1. Approximation of $p_\theta(x_0^i|\tilde{x}_t)$:

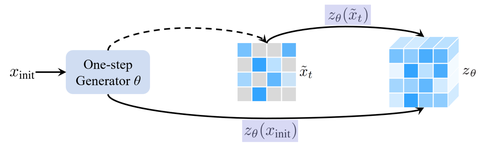


$p_\psi(x_0^i|\tilde{x}_t) \approx p_\theta(x_0^i|\tilde{x}_t)$

$\nabla_{z_\theta} D(p_\phi||p_\theta)(\tilde{x}_t) \Rightarrow \nabla_{z_\psi} D(p_\phi||p_\psi)(\tilde{x}_t)$

Given that the orange term in Eq. (2) depends on both the teacher output $p_\phi(x_0^i|\tilde{x}_t)$ and the unknown student output $p_\theta(x_0^i|\tilde{x}_t)$, we follow VSD and introduce an auxiliary model ψ to approximate $p_\theta(x_0^i|\tilde{x}_t)$.

2. Approximation of $z_\theta(\tilde{x}_t)$:



The approximation of the purple term must supply gradients for updating θ. Given this constraint, we approximate $z_\theta(\tilde{x}_t)$ with the one-step output $z_\theta(x_{\text{init}})$, assuming the consistency property.

These approximations lead to the following gradient of loss:

$$\nabla_\theta \mathcal{L}_{\text{Di[M]O}} \approx \mathbb{E}_{x_{\text{init}},t}\left[w(t)\left(\mathbb{E}_{q_{t|0}}\left[\nabla_{z_\psi} D(p_\phi||p_\psi)(\tilde{x}_t)\,\frac{\mathrm{d}z_\theta(x_{\text{init}})}{\mathrm{d}\theta}\right]\right)\right]. \tag{3}$$

In particular, we propose using Generalized Jeffrey Divergence to mitigate mode-seeking:

$$D^\beta_{\text{Jeffrey}} = (1-\beta)D_{FKL} + \beta D_{RKL}. \tag{4}$$

The gradient of FKL and RKL at each token position i are given by:

$$\frac{\partial D_{FKL_i}}{\partial {z_\psi}^i} = p_\psi(x_0^i|\tilde{x}_t) - p_\phi(x_0^i|\tilde{x}_t),$$

$$\frac{\partial D_{RKL_i}}{\partial {z_\psi}^i} = p_\psi(x_0^i|\tilde{x}_t)\left(\log\left(\frac{p_\psi(x_0^i|\tilde{x}_t)}{p_\phi(x_0^i|\tilde{x}_t)}\right) - D_{RKL_i}\right).$$

Class-conditional Results

Quantitative results on class-conditional ImageNet-256.
One-step Di[M]O matches teacher with 16× fewer steps.

	Method	Step (↓)	FID (↓)	IS (↑)	Prec. (↑)	Rec. (↑)	Den. (↑)	Cov. (↑)
Teacher	MaskGit	16	6.60	224.07	0.831	0.402	1.246	0.977
	MaskGit	8	6.66	221.57	0.827	0.397	1.233	0.974
	MaskGit	4	10.73	192.29	0.748	0.313	1.011	0.920
	MaskGit	2	91.35	13.37	0.178	0.164	0.091	0.122
Sampler	θ-trapezoidal	64	6.7	-	-	-	-	-
	θ-trapezoidal	32	7.1	-	-	-	-	-
Distill.	di4c	4	6.79	209.2	-	-	-	-
	di4c-d	4	6.57	213.6	-	-	-	-
	Di[M]O	**1**	6.91	214.0	0.828	0.377	1.255	0.967

Text-to-Image Results