ONE-STEP DISTILLATION OF DISCRETE DIFFUSION IMAGE GENERATORS VIA FIXED-POINT ITERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Discrete diffusion models excel at visual synthesis but rely on slow, iterative decoding. Existing single-step distillation methods attempt to bypass this bottleneck, either by training auxiliary score networks that effectively double compute, or by introducing specialized parameterizations and multi-stage pipelines that fragment optimization. In this paper, we introduce Fixed-Point Distillation (FPD), an end-to-end framework that constructs local correction targets by partially corrupting the student's one-step draft and refining it with a single teacher step. To compute the training objective in a semantically meaningful space, we lift discrete tokens into continuous features and apply a multi-bandwidth drift loss that pulls the student toward these corrections. To backpropagate through the discrete bottleneck, we employ a straight-through estimator that decodes the exact hard-sampled tokens in the forward pass, ensuring that training and inference operate on the same codebook manifold, while routing continuous gradients back to the student logits. This fully differentiable pathway additionally accommodates an optional unconditional adversarial objective to enhance perceptual realism. Evaluations on both class-conditional and text-to-image generation validate the effectiveness of our framework. FPD achieves competitive visual fidelity and structural alignment against existing discrete distillation methods within only a single inference step.

1 INTRODUCTION

Generative models based on iterative refinement have emerged as the leading paradigm for high-fidelity visual synthesis. By breaking down the generative process into a sequence of transitional steps, these models accurately capture complex data distributions, driving rapid progress across text-to-image generation (Ramesh et al., 2021; 2022; Rombach et al., 2022; Podell et al., 2023; Saharia et al., 2022; Nichol et al., 2021), video synthesis (Singer et al., 2022; Ho et al., 2022; Yan et al., 2021; Hong et al., 2022), image editing (Brooks et al., 2023; Meng et al., 2021), 3D asset creation (Poole et al., 2022), and world modeling (Agarwal et al., 2025; Bruce et al., 2024). Among them, discrete diffusion models that operate on quantized token sequences have attracted growing attention for their natural compatibility with language-based architectures and scalable training. However, these methods typically share a fundamental computational bottleneck: the requirement for tens or hundreds of sequential network evaluations per sample. This high inference latency severely restricts their utility in latency-sensitive environments, motivating an intensive search for efficient acceleration methods.

An effective remedy is *distillation*, which trains a student model to reproduce the output of a pre-trained teacher in far fewer steps. In the continuous domain, the literature has largely converged around three successful paradigms: trajectory imitation (Salimans & Ho, 2022; Song et al., 2023; Luo et al., 2023a; Lu & Song, 2024), distribution matching (Yin et al., 2024b;a; Luo et al., 2023b; Zhou et al., 2024b), and adversarial post-training (Sauer et al., 2024b;a). These approaches now routinely yield robust one- to four-step generators. In contrast, the distillation of discrete masked diffusion models has advanced far more slowly. Existing discrete distillers generally follow one of two suboptimal paths. They either adapt the continuous score-distillation recipe by relying on an online auxiliary network (Zhu et al., 2025a;b), which drastically inflates training compute and memory, or introduce correlation-aware objectives through bespoke designs — a mixture-model student parameterization (Hayakawa et al., 2024) or multi-stage coupling rectification (Yoo et al., 2025) — that bypass the auxiliary network but inflate per-step computation or fragment training into iterative offline stages. Neither route yields a single-step discrete distiller that is simultaneously efficient

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.



Figure 1: Text-to-image samples from MaskGen-L distilled with our Fixed-Point Distillation (FPD) framework, where each image is synthesized in a single forward pass.

to train and free of architectural or pipeline modifications. To overcome these limitations, we step away from artificial trajectory tracking and auxiliary surrogate score networks altogether, proposing a fundamentally different distillation paradigm.

In this work, we propose a principled, end-to-end distillation framework that fundamentally casts single-step generation as a fixed-point matching problem. **Our core motivation is to repurpose the frozen teacher as a state-dependent local refinement operator, rather than regressing toward pre-collected multi-step teacher rollouts.** Given a pretrained multi-step teacher, our method trains the student exclusively on its own generated drafts. Specifically, starting from a nearly fully masked token sequence with random codebook entries, the student produces a complete draft in a single forward pass. This draft is then partially re-masked and refined by the teacher through a single denoising correction, yielding a target tightly coupled to the student's current output. Ideally, the student's output distribution becomes invariant under this corruption-and-refinement cycle at convergence, meaning that single-step generation constitutes a fixed point of the distillation process. To compute the training objective in a meaningful metric space, we lift both the student's draft and the teacher's refinement into a shared continuous feature space via a frozen encoder. Since discrete tokens admit no well-defined distance, this lifted space allows us to apply a multi-bandwidth drift loss that provides stable gradient signals without auxiliary score networks. As training progresses, these local corrections accumulate, gradually steering the student's generation toward the data manifold.

However, backpropagating the drift loss through the discrete sampling step is nontrivial. Conventional continuous relaxations, such as soft probability mixtures over codebook embeddings, produce off-manifold inputs to the decoder that never occur at inference, so the loss is evaluated on images that differ from those decoded from the actual hard-sampled draft. To resolve this, we employ a straight-through estimator (STE) that feeds exact hard-sampled tokens in the forward pass, ensuring that the decoder always receives valid discrete inputs, while routing continuous gradients from the drift loss back to the student logits during the backward pass. This ensures that training and inference operate on identical discrete inputs, eliminating the train-test mismatch inherent in soft relaxations. Moreover, the fully differentiable pathway from pixel space to logits additionally accommodates an optional unconditional adversarial loss on the decoded student images, which enhances perceptual realism without sacrificing the semantic alignment enforced by the drift objective.

Our contributions are summarized as follows: **1)** We propose Fixed-Point Distillation, a framework that formulates single-step discrete distillation as a fixed-point matching problem, bypassing complex trajectory tracking, pre-collected multi-step teacher samples, auxiliary score networks, or fragmented training pipelines. **2)** We introduce a lifted drift objective paired with a straight-through estimator, enabling fully end-to-end training through the discrete bottleneck with consistent train-

ing and inference behavior. **3)** On class-conditional and text-to-image generation, FPD achieves a one-step FID of 6.75 on ImageNet and a GenEval score of 0.45, competitive with existing one-step discrete distillation methods while requiring no auxiliary network.

2 Related Work

2.1 Discrete Diffusion Models

Diffusion models learn to generate data by progressively corrupting it with a forward noising process and training a network to reverse the corruption. The continuous-domain lineage, spanning from early denoising models (Ho et al., 2020; Song et al., 2020) to modern flow matching architectures, has produced the dominant large-scale visual generators (Esser et al., 2024; Rombach et al., 2022; Podell et al., 2023). A parallel discrete branch works on quantized token sequences and forms the primary focus of this work. In the language domain, masked diffusion and score-entropy formulations (Sahoo et al., 2024; Lou et al., 2023) implement absorbing- and uniform-state corruption processes, demonstrating generation capabilities that rival those of traditional autoregressive (AR) decoders at comparable scales. For visual synthesis, a growing line of work extends discrete diffusion to image generation (Austin et al., 2021; Campbell et al., 2022; Shi et al., 2024; Bond-Taylor et al., 2022; Gu et al., 2022; Hu et al., 2022; Sun et al., 2022). Among them, masked generative transformers such as MaskGIT (Chang et al., 2022) and MaskGen (Kim et al., 2025) iteratively predict and reveal tokens over a fixed schedule, differing primarily in their tokenization strategies and conditioning mechanisms. Regardless of these design choices, discrete diffusion models typically require tens to hundreds of sequential forward passes per sample, severely limiting their deployment in latency-sensitive settings.

2.2 Discrete Diffusion Distillation

While continuous-domain distillation has matured rapidly through trajectory imitation (Salimans & Ho, 2022; Song et al., 2023; Lu & Song, 2024), distribution matching (Yin et al., 2024b;a), and adversarial frameworks (Sauer et al., 2024b;a), the literature on discrete distillation remains relatively sparse. Progress has largely relied on porting continuous primitives into categorical token spaces. This introduces severe structural challenges: the non-differentiability of token sampling, the lack of deterministic latent trajectories under masked corruption, and the factorization error inherent in single-step predictions.

To navigate these challenges, recent methods in both visual and language domains explore diverse adaptations. One prominent direction adapts distribution matching to discrete outputs by minimizing token-level divergences. DiMO performs on-policy token-level distribution matching with an auxiliary model initialized from the teacher, which approximates the student's intermediate conditional distributions and provides gradients directly on generator logits, avoiding backpropagation through discrete sampling (Zhu et al., 2025a). Soft-DiMO further relaxes discrete outputs into soft embeddings, enabling end-to-end differentiable refinements such as adversarial fine-tuning (Zhu et al., 2025b). Alternative approaches mitigate product-factorization errors by parameterizing student transitions as mixtures of product models to capture dimensional correlations (Hayakawa et al., 2024), or by iteratively rectifying source–target couplings to reduce conditional total correlation in discrete flows (Yoo et al., 2025). In the language domain, SDTT distills masked discrete diffusion language models by matching student predictions to teacher-generated multi-step denoising targets (Deschenaux & Gulcehre, 2024), while D-MMD generalizes continuous moment-matching distillation to discrete/categorical diffusion models (Hoogeboom et al., 2026).

Crucially, existing discrete distillation methods typically incur substantial computational overhead, either through online learnable auxiliary networks that effectively double training cost, or through multi-stage pipelines that fragment optimization into separate offline phases. Our framework sidesteps both issues by formulating distillation as a fixed-point matching problem, replacing trajectory-level imitation with local corrections derived from the student's own drafts, without auxiliary networks or multi-stage training.

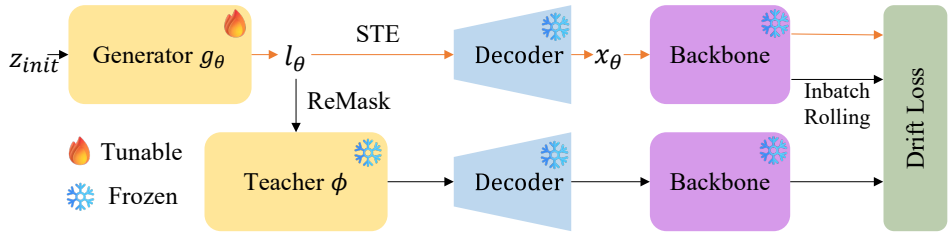


Figure 2: **Overview of the proposed distillation framework.** The student model processes a masked initialization z_{init} to output logits ℓ_θ. Via a Straight-Through Estimator (STE), these discrete outputs are decoded into a continuous image x_θ and mapped to a feature space by a frozen backbone Φ. Concurrently, the student's token draft is partially re-masked ($\mathcal{M}_r$) and refined by the frozen teacher. The decoded teacher features serve as positive targets, while cyclically shifted in-batch student features act as negatives to compute the drift loss $\mathcal{L}_{\text{drift}}$. Black lines denote forward computation paths, while red lines highlight the differentiable gradient flow that safely bypasses the discrete bottleneck. The decoder and backbone are shared by the student and teacher branches.

3 METHOD

We first review the discrete diffusion framework in Sec. 3.1, then present our fixed-point formulation that constructs training targets from the student's own re-masked drafts in Sec. 3.2. The resulting residual is computed via a multi-bandwidth drift loss in a lifted continuous feature space, as detailed in Sec. 3.3. Finally, Sec. 3.4 describes the straight-through estimator that bridges the discrete bottleneck and enables an optional adversarial objective.

3.1 PRELIMINARIES: DISCRETE DIFFUSION FOR IMAGE GENERATION

Let $x \in \mathbb{R}^{3\times H\times W}$ be an image and let $(\mathcal{E}, \mathcal{D})$ denote a frozen VQ encoder-decoder pair with codebook $E \in \mathbb{R}^{K\times d}$, where K is the codebook size and d the embedding dimension. The encoder maps an image to a length-L sequence of discrete codes $z = (z_1, \ldots, z_L) \in \mathcal{V}^L$ with $\mathcal{V} = \{1, \ldots, K\}$, and the decoder reconstructs an image via $\mathcal{D}(E[z])$. We write $E[z] \in \mathbb{R}^{L\times d}$ for the row-wise codebook lookup. Text conditioning c is produced by a separate frozen text encoder.

Forward masking process. Discrete diffusion augments the token vocabulary with a special mask symbol $[\text{M}] \notin \mathcal{V}$, yielding $\bar{\mathcal{V}} = \mathcal{V} \cup \{[\text{M}]\}$. For a continuous timestep $t \in [0, 1]$ and a monotonically increasing noise schedule $\gamma : [0, 1] \to [0, 1]$, each position is independently replaced with $[\text{M}]$ with probability $\gamma(t)$:

$$q(z_t \mid z) = \prod_{i=1}^{L} \left[(1 - \gamma(t))\, \delta_{z_{t,i}, z_i} + \gamma(t)\, \delta_{z_{t,i}, [\text{M}]}\right]. \tag{1}$$

Reverse unmasking process. A teacher network $f_\phi : \bar{\mathcal{V}}^L \times \mathcal{C} \times [0, 1] \to \mathbb{R}^{L\times K}$ outputs per-position logits $\ell = f_\phi(z_t, c, t)$, which define a categorical distribution $p_\phi(z_{0,i} = v \mid z_t, c, t) \propto \exp(\ell_{i,v})$, where $\mathcal{C}$ denotes the conditioning space. The training objective minimizes the cross-entropy loss on the masked positions,

$$\mathcal{L}_\phi = \mathbb{E}_{t,\, z\sim p_{\text{data}},\, z_t\sim q(\cdot|z)} \left[- \sum_{i:\, z_{t,i}=[\text{M}]} \log p_\phi(z_i \mid z_t, c, t) \right]. \tag{2}$$

During inference, generation proceeds iteratively from a fully masked sequence $z_T = ([\text{M}], \ldots, [\text{M}])$. At each of the T steps, the model predicts all masked positions, retains a top-confidence subset, and re-masks the remainder according to γ. We denote a single such teacher refinement step at mask ratio r by $\mathcal{T}_\phi^r$. After T steps, the final tokens $\hat{z}$ are decoded to pixels via $\hat{x} = \mathcal{D}(E[\hat{z}];\, c)$.

Distillation. To eliminate this iterative overhead, our distillation framework trains a single-step student g_θ that directly maps a masked initialization to a high-quality sample, bypassing the teacher's multi-step inference process. The following section details our approach.

3.2 Fixed-Point Iteration via Local Correction

The teacher's reverse process is a sequence of refinements $\mathcal{T}_\phi^{r_1} \circ \mathcal{T}_\phi^{r_2} \circ \cdots$ that progressively reduce the masking level. A correctly distilled single-step student g_θ should produce, in one forward pass, a sample $\hat{z}$ that is already a *fixed point* of the teacher's refinement operator: re-masking $\hat{z}$ at an intermediate level r and applying $\mathcal{T}_\phi^r$ should leave its distribution invariant. Writing p_θ for the student's marginal, this requirement can be stated as

$$p_\theta = \mathbb{E}_{r\sim\pi}\Big[\big(\mathcal{T}_\phi^r \circ \mathcal{M}_r\big)_{\#}\, p_\theta\Big], \tag{3}$$

where $\#$ denotes the pushforward measure and π is the distribution of re-masking ratios used in training.

Enforcing Eq. 3 directly as a global distributional constraint is intractable. Instead, we minimize the local correction residual. Although the teacher suffers from severe truncation error when generating in a single step from a fully masked sequence, its local refinement from intermediate noise levels remains faithful to its original training objective. Evaluating the teacher on the student's partially re-masked draft therefore serves as a local improvement target toward the data manifold.

As shown in Fig. 2, we construct an end-to-end training pipeline. For each prompt c, we instantiate a nearly fully masked initialization z_{init} at a high mask ratio r_{init}, with the unmasked positions filled by random codebook entries. The student generates a full-sequence draft in one shot:

$$\ell_\theta = g_\theta(z_{\text{init}}, c), \qquad \hat{z} \sim p_\theta(\cdot \mid z_{\text{init}}, c). \tag{4}$$

To compute the local correction signal, a single teacher refinement is performed at a randomly sampled intermediate mask ratio $r \sim \pi$:

$$\hat{z}_T = \mathcal{T}_\phi^r\big(\mathcal{M}_r(\hat{z}),\, c\big). \tag{5}$$

The discrepancy between pair $(\hat{z}, \hat{z}_T)$ constitutes the empirical fixed-point residual. The training loss in Sec. 3.3 penalizes this discrepancy in a continuous feature space.

Since both training and inference start from the same masked initialization z_{init}, the train- and test-time operating points coincide by construction.

3.3 Lifted Feature Optimization

To compute the fixed-point residual in a meaningful metric space, we lift the discrete tokens into a continuous feature space. We bridge this gap by sequentially passing tokens through the codebook embedding E, the decoder $\mathcal{D}$, and a frozen feature backbone Φ. For a student draft $\hat{z}_i$, we obtain P teacher refinements $\{\hat{z}_{T,i,p}\}_{p=1}^P$ by applying Eq. 5 with independently sampled r and masks. The lifted spatial features are $X_i = \Phi\big(\mathcal{D}(E[\hat{z}_i]; c_i)\big)$ and $Y_{i,p} = \Phi\big(\mathcal{D}(E[\hat{z}_{T,i,p}]; c_i)\big)$.

Following Deng et al. (2026), we employ a multi-bandwidth drift objective. Within a mini-batch, the empirical drift vector $V_h(X_i^f)$ at spatial position f and kernel bandwidth h is defined by an attractive force toward the teacher targets and a repulsive force from in-batch negative samples:

$$V_h(X_i^f) = \sum_{p=1}^{P} a_{ip}^{+}(h)\,\big(Y_{i,p}^f - X_i^f\big) \;-\; \sum_{s=1}^{S} a_{is}^{-}(h)\,\big(X_{\sigma_s(i)}^f - X_i^f\big). \tag{6}$$

Here, $\sigma_s(i) = (i+s) \bmod B$ cyclically shifts the batch of size B, so the S negatives are drafts of other prompts and represent the student marginal, and $a^{\pm} \propto \exp(-\|X_i^f - \cdot\|/h)$ are doubly-normalized softmax affinities derived from a Laplace kernel. The student is optimized to match the displaced target via a stop-gradient regression:

$$\mathcal{L}_{\text{drift}} = \sum_{h\in\mathcal{H}} \frac{1}{Z_h}\,\mathbb{E}_{i,f}\,\big\|X_i^f - \text{sg}\big(X_i^f + V_h(X_i^f)\big)\big\|^2, \tag{7}$$

where $\mathcal{H}$ is a set of bandwidths and Z_h is an RMS normalizer that equalizes gradient scales across bandwidths. The repulsive term is essential. Eq. 3 also admits degenerate solutions, e.g., a student that produces nearly identical drafts for different prompts, which the teacher merely completes consistently. When drafts of different prompts collapse toward each other, they fall within each other's kernel bandwidth and receive the largest repulsive affinities, which pushes them apart. Without this term, training collapses (Fig. 5).

3.4 Differentiable Routing via the Discrete Bottleneck

The drift objective operates in continuous space, yet the student outputs discrete tokens $\hat{z} \sim p_\theta$. Standard continuous relaxations, such as probability-weighted soft embeddings $\tilde{e}_i = \sum_v p_{\theta,iv} E_v$, produce off-manifold vectors. Decoding such off-manifold vectors evaluates the loss on images that no hard token sequence produces, whereas the teacher target $\hat{z}_T$ is always computed from the hard-sampled draft.

To maintain differentiability while preserving manifold consistency, we employ a hard-forward, soft-backward straight-through estimator (STE) (Bengio et al., 2013):

$$e_i^{\text{STE}} = E[\hat{z}_i] + \tilde{e}_i - \text{sg}(\tilde{e}_i). \tag{8}$$

In the forward pass, $e_i^{\text{STE}} \equiv E[\hat{z}_i]$, so the student branch decodes exactly the hard draft that the teacher refines. In the backward pass, gradients route through the soft embedding, yielding $\partial e_i^{\text{STE}}/\partial \ell_{i,v} = \partial \tilde{e}_i/\partial \ell_{i,v}$, which translates the continuous drift force from the feature space back to the student logits.

This differentiable pixel-to-logit pathway allows us to optionally add an unconditional GAN objective $\mathcal{L}_{\text{gan}}$ applied to the decoded student images. Because the drift loss already enforces conditional and structural correspondence with the teacher targets, the unconditional discriminator mainly serves to enhance high-frequency details and perceptual realism. The final training objective is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{drift}} + \lambda \mathcal{L}_{\text{gan}}$, where λ balances the adversarial penalty.

4 Experiments

We evaluate FPD on class-conditional and text-to-image generation. Sec. 4.2 compares it with multi-step teachers and existing distillation methods, and Sec. 4.3 ablates its three components: the fixed-point target, the lifted drift space, and the gradient routing through the discrete bottleneck.

4.1 Experimental Settings

Datasets and Metrics. For class-conditional generation, we use ImageNet (Deng et al., 2009) and report Fréchet Inception Distance (FID) (Heusel et al., 2017), Inception Score (IS) (Salimans et al., 2016), and Precision and Recall (Kynkäänniemi et al., 2019) over 50,000 samples. For text-to-image generation, we train on a 400K-image subset of LAION (Schuhmann & Beaumont, 2022) and evaluate on GenEval (Ghosh et al., 2023). The drift loss only needs class labels or captions to query the teacher; real images are used solely by the optional unconditional discriminator.

Implementation Details. The student is initialized from the teacher, MaskGIT (Besnier & Chen, 2023) for ImageNet and MaskGen-L (Kim et al., 2025) for text-to-image. We further distill the larger Lumina-DiMOO (Xin et al., 2025) with LoRA to assess training cost. The drift loss uses DINOv3 ViT-B/16 (Siméoni et al., 2025) features from blocks $\{2, 5, 8, 11\}$ on a 4×4 grid with bandwidths $\mathcal{H} = \{0.02, 0.05, 0.2\}$. The teacher, VQ decoder, and feature backbone stay frozen. The optional discriminator is an unconditional StyleGAN-XL (Sauer et al., 2022) discriminator trained on real images, so all conditional information comes from the drift loss. Full configurations are listed in Appendix.

Baselines. On ImageNet, we compare with the MaskGIT teacher at different step counts and with the discrete distillation methods SDTT (Deschenaux & Gulcehre, 2024), Di4C (Hayakawa et al., 2024), ReDi (Yoo et al., 2025), and DiMO (Zhu et al., 2025a). For text-to-image generation, we compare with the MaskGen teacher, the discrete distillation method DiMO (Zhu et al., 2025a), and the continuous distillation methods InstaFlow (Liu et al., 2024), SiD-LSG (Zhou et al., 2024a), RG-LCM (Li et al., 2024), LCM (Luo et al., 2023a), TDM (Luo et al., 2025), ADD (Sauer et al., 2024b), and DMD2 (Yin et al., 2024a).

4.2 Main Results

Tab. 1 and Tab. 2 report text-to-image and class-conditional results. The *Aux Net* column marks methods that train an auxiliary score network jointly with the student; such a network is typically as large as the teacher and roughly doubles training compute.

Table 1: Comparison of various text-to-image generation models on the GenEval benchmark. The symbols ↑ and ↓ indicate that higher and lower scores are preferable, respectively.

Methods	Steps ↓	Params	Aux Net	GenEval ↑						
				Single	Two	Count.	Colors	Pos.	Attr.	Overall
Foundation Models										
LDM (Rombach et al., 2022)	50	1.4B	-	0.92	0.29	0.23	0.70	0.02	0.05	0.37
DALLE 2 (Ramesh et al., 2022)	-	4.2B	-	0.94	0.66	0.49	0.77	0.10	0.19	0.52
SDXL (Podell et al., 2023)	50	2.6B	-	0.98	0.74	0.39	0.85	0.15	0.23	0.55
MaskGen (Kim et al., 2025)	16	0.6B	-	0.97	0.55	0.38	0.80	0.08	0.14	0.48
Continuous Distillation										
InstaFlow (Liu et al., 2024)	1	0.9B	✗	0.88	0.21	0.20	0.66	0.03	0.03	0.33
SiD-LSG (Zhou et al., 2024a)	1	0.9B	✓	0.93	0.37	0.21	0.57	0.03	0.03	0.36
RG-LCM (HPS) (Li et al., 2024)	2	0.9B	✗	0.97	0.54	0.35	0.82	0.07	0.14	0.48
TDM (Luo et al., 2025)	4	0.9B	✓	0.99	0.57	0.49	0.78	0.09	0.09	0.50
SDXL-LCM (Luo et al., 2023a)	1	2.6B	✗	0.75	0.11	0.14	0.59	0.01	0.03	0.27
SDXL-LCM (Luo et al., 2023a)	4	2.6B	✗	0.99	0.57	0.39	0.86	0.09	0.18	0.51
SDXL-Turbo (Sauer et al., 2024b)	1	2.6B	✗	0.99	0.65	0.52	0.87	0.12	0.19	0.55
SDXL-DMD2 (Yin et al., 2024a)	1	2.6B	✓	0.99	0.68	0.48	0.90	0.08	0.19	0.55
Discrete Distillation										
MaskGen-DiMO (Zhu et al., 2025a)	1	0.6B	✓	0.93	0.39	0.35	0.74	0.07	0.08	0.42
MaskGen-FPD (Ours)	1	0.6B	✗	0.96	0.36	0.43	0.76	0.09	0.09	0.45

Table 2: Quantitative results on class-conditional ImageNet-256 with MaskGIT teacher.

Method	Steps ↓	Aux Net	FID ↓	IS ↑	Prec. ↑	Rec. ↑
Teacher						
MaskGIT (Besnier & Chen, 2023)	16	-	6.60	224	0.83	0.40
MaskGIT (Besnier & Chen, 2023)	8	-	6.66	222	0.83	0.40
MaskGIT (Besnier & Chen, 2023)	4	-	10.73	192	0.75	0.31
MaskGIT (Besnier & Chen, 2023)	2	-	91.35	13	0.18	0.16
Discrete Distillation						
SDTT (Deschenaux & Gulcehre, 2024)	4	✗	8.97	205	0.88	0.41
SDTT (Deschenaux & Gulcehre, 2024)	1	✗	90.40	14	0.31	0.13
Di4C (Hayakawa et al., 2024)	4	✗	6.79	209	-	-
Di4C-d (Hayakawa et al., 2024)	4	✗	6.57	214	-	-
ReDi[1] (Yoo et al., 2025)	4	✗	7.58	228	0.87	0.46
ReDi[2] (Yoo et al., 2025)	4	✗	7.86	240	0.87	0.44
ReDi[3]-distill (Yoo et al., 2025)	1	✗	11.68	182	0.83	0.44
DiMO (Zhu et al., 2025a)	1	✓	6.91	214	0.83	0.38
FPD (Ours)	1	✗	6.75	209	0.81	0.36

Text-to-Image Generation. With a single step, MaskGen-FPD reaches a GenEval score of 0.45, outperforming the one-step discrete baseline DiMO (0.42) without an auxiliary network and retaining 94% of the 16-step teacher (0.48). The largest gain is on Counting (0.43 vs. 0.35). As shown in Fig. 1 and Fig. 3, FPD produces sharp one-step samples, whereas one-step sampling from the teacher collapses.

Class-Conditional Generation. On ImageNet, FPD achieves a one-step FID of 6.75, better than DiMO (6.91) and close to the 16-step MaskGIT teacher (6.60). Other distillation methods degrade sharply at one step (SDTT: 90.40, ReDi[3]-distill: 11.68), and FPD also outperforms the 4-step ReDi[1] (7.58).

Training Efficiency. Tab. 3 compares training cost on Lumina-DiMOO. FPD requires 24.05 TFLOPs per iteration, about 41% less than the analytic estimate for DiMO (40.54), which runs out of memory in our setting. Without adversarial loss, FPD raises the teacher's one-step GenEval

Table 3: Training cost and one-step quality on Lumina-DiMOO. TFLOPs are per training iteration; the DiMO value is an analytic estimate since it runs out of memory. FPD is trained without $\mathcal{L}_{\text{gan}}$.

Method	GenEval ↑	VRAM (GB) ↓	TFLOPs / iter ↓
Teacher (1 step)	0.05	-	-
Teacher (64 steps)	0.88	-	-
DiMO	-	OOM	40.54
FPD (Ours)	0.77	56.23	24.05

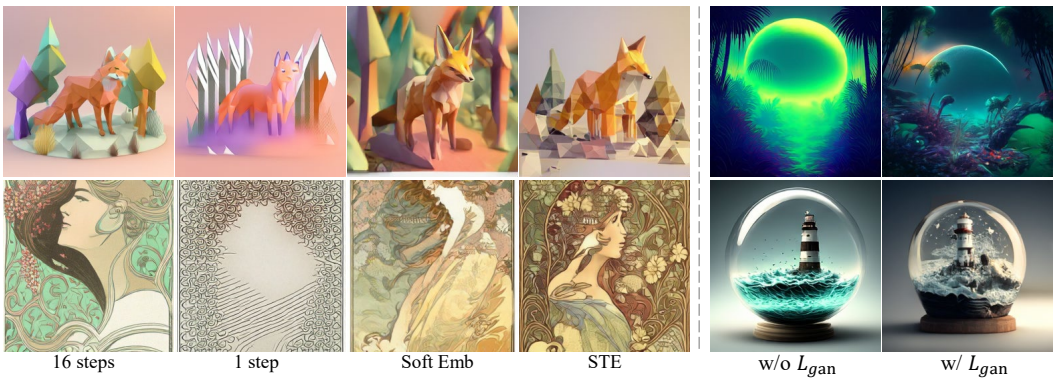


Figure 3: Training strategy comparisons. We compare the multi-step teacher baselines with our single-step FPD variants, showing that the straight-through estimator (STE) preserves structural integrity and that the optional GAN loss enhances perceptual details.

score from 0.05 to 0.77, compared with 0.88 for 64-step sampling. The cost analysis is detailed in App. B.2.

4.3 Ablation Studies

Fixed-Point Target. The teacher is effective as a local corrector of the student's own draft rather than as a standalone generator. When it refines a random token sequence instead of the re-masked student draft, GenEval drops from 0.43 to 0.22 (Tab. 4b): the teacher then has to generate from scratch, and its targets no longer relate to the student's current output. Fig. 4 further analyzes the refinement operator. Refining student samples k times in succession improves the CLIP score, with most of the gain at small k (a). The student–teacher DINO distance drops quickly during training and then plateaus, indicating that the student approaches a fixed point, and a larger re-masking ratio r yields a larger correction (b). However, overly aggressive re-masking ($r = 0.9$) lowers the CLIP score of the refined samples (c), revealing a trade-off between correction strength and locality.

Lifted Drift Space. Computing the drift loss in pixel space yields only 0.21 (Tab. 4c), since pixel distances are dominated by low-level variations; lifting to DINOv3 features raises it to 0.43. Multiple bandwidths outperform a single one (0.43 vs. 0.41), as large h enforces coarse alignment and small h fine-grained discrimination. Both the spatial layout and the depth of the features matter (Tab. 4d): global pooling discards spatial correspondence (0.34), shallow blocks alone lack semantics (0.35), and combining shallow and deep blocks on a 4×4 grid performs best. Keeping only the attraction toward the teacher targets, without the repulsive term, makes the samples collapse during training (Fig. 5).

Gradient Routing and Adversarial Loss. Replacing soft embeddings with the STE improves GenEval from 0.38 to 0.43 (Tab. 4a). Soft embeddings feed off-manifold inputs to the student-side decoder, which produces artifacts and structural distortions (Fig. 3), whereas the STE keeps training and inference on the same discrete inputs. The drift loss alone is thus already effective; adding the unconditional $\mathcal{L}_{\text{gan}}$ improves both estimators and gives the best score of 0.45, and visibly sharpens details (Fig. 3).

Table 4: Ablations on GenEval (MaskGen-L). (a) Gradient estimator through the discrete bottleneck, with and without $\mathcal{L}_{\text{gan}}$: *Soft* feeds the probability-weighted codebook mixture, *STE* feeds hard-sampled tokens. (b) Source of the draft refined by the teacher. (c) Space and kernel bandwidths of the drift loss. (d) DINOv3-B blocks used for the drift loss; *Grid* keeps a 4×4 patch grid instead of global average pooling. (b)–(d) are trained without $\mathcal{L}_{\text{gan}}$.

(a) Gradient routing

Estimator	$\mathcal{L}_{\text{gan}}$	Overall
Soft		0.38
Soft	✓	0.42
STE		0.43
STE	✓	**0.45**

(b) Draft source

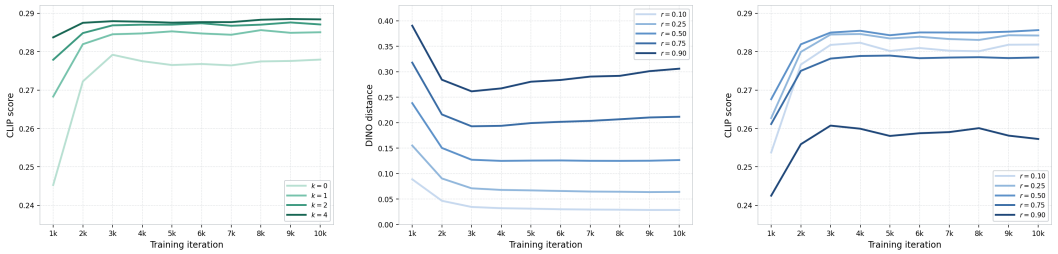
Draft	Overall
Random	0.22
Student	**0.43**

(c) Drift space and bandwidths

Space	$\mathcal{H}$	Overall
Pixel	{.02, .05, .2}	0.21
Feature	{.05}	0.41
Feature	{.02, .05, .2}	**0.43**

(d) Backbone layers

Blocks	Grid	Overall
{2,5,8,11}		0.34
{2,5}	✓	0.35
{8,11}	✓	0.42
{2,5,8,11}	✓	**0.43**



(a) Effect of iteration depth. (b) Student–teacher distance. (c) Effect of the re-masking ratio.

Figure 4: **Behavior of the refinement operator during training.** At each checkpoint, we draw 1K student samples and apply $\mathcal{T}_\phi^r \circ \mathcal{M}_r$ to them. (a) CLIP score after k successive refinements. (b) DINO distance between the student draft and its one-step refinement under different re-masking ratios r. (c) CLIP score of the one-step refinement under different r.



Figure 5: **Effect of the repulsive term** on Lumina-DiMOO (Xin et al., 2025), trained without $\mathcal{L}_{\text{gan}}$. The leftmost image is a one-step sample of the teacher; the remaining columns show checkpoints from 1k to 10k iterations. (a) With only the attraction toward the teacher targets, the samples collapse. (b) With the repulsive term, they steadily improve.

5 Conclusion

We presented Fixed-Point Distillation (FPD), a framework for single-step generation with discrete diffusion models. FPD formulates distillation as a fixed-point problem: the student's draft is partially re-masked and refined by a single teacher step, which yields a local correction target. The student is trained toward this target with a drift loss in a lifted continuous feature space, without any auxiliary score network. A straight-through estimator decodes the hard-sampled tokens in the forward pass while routing continuous gradients back to the student logits, and this differentiable pathway also accommodates an optional unconditional adversarial loss. On class-conditional and text-to-image generation, FPD achieves competitive single-step results compared with existing discrete distillation methods, approaching the quality of multi-step teachers.

AI USE STATEMENT

Large language models (LLMs) were used to assist with writing and editing the manuscript, including proofreading, correcting errors, and polishing the text. AI tools were also used to support parts of the code implementation. The authors reviewed the AI-generated content and are responsible for the final manuscript.

REPRODUCIBILITY STATEMENT

We provide details throughout the paper to facilitate reproducibility of our results. In particular, the main paper describes the proposed method and experimental setup, while additional implementation details, training configurations, hyperparameters, and experimental settings are provided in the appendix.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Victor Besnier and Mickael Chen. A pytorch reproduction of masked generative image transformer. *arXiv preprint arXiv:2310.14400*, 2023.

Sam Bond-Taylor, Peter Hessey, Hiroshi Sasaki, Toby P Breckon, and Chris G Willcocks. Unleashing transformers: Parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes. In *European Conference on Computer Vision*, pp. 170–188. Springer, 2022.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 35:28266–28279, 2022.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11315–11325, 2022.

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.

Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative modeling via drifting. *arXiv preprint arXiv:2602.04770*, 2026.

Justin Deschenaux and Caglar Gulcehre. Beyond autoregression: Fast llms via self-distillation through time. *arXiv preprint arXiv:2410.21035*, 2024.

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

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10696–10706, 2022.

Satoshi Hayakawa, Yuhta Takida, Masaaki Imaizumi, Hiromi Wakaki, and Yuki Mitsufuji. Distillation of discrete diffusion through dimensional correlations. *arXiv preprint arXiv:2410.08709*, 2024.

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

Jonathan Ho, 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.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Emiel Hoogeboom, David Ruhe, Jonathan Heek, Thomas Mensink, and Tim Salimans. Beyond single tokens: Distilling discrete diffusion models via discrete mmd. *arXiv preprint arXiv:2603.20155*, 2026.

Minghui Hu, Chuanxia Zheng, Heliang Zheng, Tat-Jen Cham, Chaoyue Wang, Zuopeng Yang, Dacheng Tao, and Ponnuthurai N Suganthan. Unified discrete diffusion for simultaneous vision-language generation. *arXiv preprint arXiv:2211.14842*, 2022.

Dongwon Kim, Ju He, Qihang Yu, Chenglin Yang, Xiaohui Shen, Suha Kwak, and Liang-Chieh Chen. Democratizing text-to-image masked generative models with compact text-aware one-dimensional tokens. In *ICCV*, 2025.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

Jiachen Li, Weixi Feng, Wenhu Chen, and William Yang Wang. Reward guided latent consistency distillation. *arXiv preprint arXiv:2403.11027*, 2024.

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, et al. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *The Twelfth International Conference on Learning Representations*, 2024.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*, 2023.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. *arXiv preprint arXiv:2410.11081*, 2024.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023a.

Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. *Advances in Neural Information Processing Systems*, 36:76525–76546, 2023b.

Yihong Luo, Tianyang Hu, Jiacheng Sun, Yujun Cai, and Jing Tang. Learning few-step diffusion models by trajectory distribution matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17719–17728, 2025.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

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

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

Subham S Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184, 2024.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10, 2022.

Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024a.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024b.

Christoph Schuhmann and Romain Beaumont. Laion-aesthetics. *LAION. AI*, 2022.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. Simplified and generalized masked diffusion for discrete data. *Advances in neural information processing systems*, 37: 103131–103167, 2024.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, pp. 32211–32252, 2023.

Haoran Sun, Lijun Yu, Bo Dai, Dale Schuurmans, and Hanjun Dai. Score-based continuous-time discrete diffusion models. *arXiv preprint arXiv:2211.16750*, 2022.

Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, et al. Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. *arXiv preprint arXiv:2510.06308*, 2025.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6613–6623, 2024b.

Jaehoon Yoo, Wonjung Kim, and Seunghoon Hong. Redi: Rectified discrete flow. *arXiv preprint arXiv:2507.15897*, 2025.

Mingyuan Zhou, Zhendong Wang, Huangjie Zheng, and Hai Huang. Long and short guidance in score identity distillation for one-step text-to-image generation. *arXiv preprint arXiv:2406.01561*, 2024a.

Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In *Forty-first International Conference on Machine Learning*, 2024b.

Yuanzhi Zhu, Xi Wang, Stéphane Lathuilière, and Vicky Kalogeiton. Di [m] o: Distilling masked diffusion models into one-step generator. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18606–18618, 2025a.

Yuanzhi Zhu, Xi Wang, Stéphane Lathuilière, and Vicky Kalogeiton. Soft-di [m] o: Improving one-step discrete image generation with soft embeddings. *arXiv preprint arXiv:2509.22925*, 2025b.

Overview. In the appendix, we present more results and details:

- **A.** More details on the method.
- **B.** More empirical studies and visual results.
- **C.** Discussion of the broader impacts, safeguards and limitations.

A Implementation Details

A.1 Stochastic Initialization and One-Step Student Generation

We detail the construction of z_{init} and $\hat{z}$ used in Sec. 3.2. The token vocabulary is $\mathcal{V} = \{1, \dots, K\}$, the sequence length is L, and $[\text{M}] \notin \mathcal{V}$ is the mask symbol. We retain the condition c of the main paper; in the class-conditional experiments, it is a class label. Below, i indexes token positions and v indexes vocabulary entries.

Random masked initialization. We first draw random codebook indices $z^{\text{rnd}} \sim \text{Uniform}(\mathcal{V}^L)$. For a fixed, high initialization mask ratio r_{init}, we select a subset $\mathcal{I}_{\text{init}} \subseteq \{1, \dots, L\}$ uniformly among all subsets of size $n_{\text{init}} = \text{round}(r_{\text{init}} L)$, independently of z^{rnd}. The student input is

$$z_{\text{init},i} = \begin{cases} [\text{M}], & i \in \mathcal{I}_{\text{init}}, \\ z_i^{\text{rnd}}, & i \notin \mathcal{I}_{\text{init}}. \end{cases} \tag{9}$$

The random indices and mask locations are resampled for each input, so the visible entries provide an additional source of randomness beyond the categorical sampling below. The ratio r_{init} is distinct from the teacher re-masking ratio r: the former specifies the student's starting distribution, whereas the latter controls the corruption of an already generated draft.

One-step draft and gradient routing. The student predicts logits for the complete sequence in one forward pass:

$$\ell_\theta = g_\theta(z_{\text{init}}, c) \in \mathbb{R}^{L \times K}. \tag{10}$$

For a sampling temperature $\tau > 0$, define

$$p_{\theta,iv} = \frac{\exp(\ell_{\theta,iv}/\tau)}{\sum_{v'=1}^{K} \exp(\ell_{\theta,iv'}/\tau)}, \qquad \hat{z}_i \sim \text{Categorical}(p_{\theta,i1}, \dots, p_{\theta,iK}). \tag{11}$$

Collecting these samples gives the hard draft $\hat{z} \sim p_\theta(\cdot \mid z_{\text{init}}, c)$ used in the main paper. All positions are predicted, including those initially occupied by random visible tokens.

The teacher branch receives this discrete draft after re-masking. On the student branch, the soft codebook expectation $\tilde{e}_i = \sum_{v=1}^{K} p_{\theta,iv} E_v$ supplies the backward path of the STE in Eq. 8. The forward value of the STE is exactly $E[\hat{z}_i]$, so the decoder is evaluated on hard codebook embeddings, and gradients through the frozen decoder and feature backbone reach ℓ_θ through the soft expectation. Training and inference use the same initialization distribution; the teacher correction described next is used only for training.

A.2 Student-Conditioned Re-Masking and Teacher Refinement

This section specifies how the frozen teacher constructs $\hat{z}_T = \mathcal{T}_\phi^r(\mathcal{M}_r(\hat{z}), c)$ from the current student draft, with the condition c unchanged.

Sampling the corruption level. We sample a re-masking ratio $r \sim \pi$, as in Sec. 3.2, using an arccosine rule:

$$u \sim \text{Uniform}(0, 1), \qquad r = \frac{\arccos(u)}{\pi/2}. \tag{12}$$

The variable u is an auxiliary random draw, not the diffusion timestep t in Sec. 3.1, so this rule does not redefine the teacher's noise schedule $\gamma(t)$. For the sampled r, we choose a subset $\mathcal{I}_r \subseteq$

$\{1, \ldots, L\}$ uniformly among all subsets of size $n_r = \max\{1, \mathrm{round}(rL)\}$, which avoids an empty mask, and form

$$z_{\text{mask}} = \mathcal{M}_r(\hat{z}), \qquad z_{\text{mask},i} = \begin{cases} [\text{M}], & i \in \mathcal{I}_r, \\ \hat{z}_i, & i \notin \mathcal{I}_r. \end{cases} \tag{13}$$

The visible context thus consists of the student's own predictions, rather than the random indices used for initialization.

A single teacher correction. The frozen teacher f_ϕ evaluates z_{mask} at the corresponding noise level. For a teacher with explicit timestep conditioning, let t_r satisfy $\gamma(t_r) = r$; for architectures without such an input, the timestep argument below is omitted. With classifier-free guidance (CFG), the logits are

$$\begin{aligned} \ell_\phi^{\text{cond}} &= f_\phi(z_{\text{mask}}, c, t_r), \\ \ell_\phi^{\text{uncond}} &= f_\phi(z_{\text{mask}}, \varnothing, t_r), \\ \ell_\phi^{\text{cfg}} &= \ell_\phi^{\text{uncond}} + w\big(\ell_\phi^{\text{cond}} - \ell_\phi^{\text{uncond}}\big), \end{aligned} \tag{14}$$

where $\varnothing$ denotes the null condition and w is the guidance scale; at $w = 1$, only the conditional evaluation is needed. The guided logits are then converted to a complete hard token target $\hat{z}_T \in \mathcal{V}^L$ by greedy decoding or temperature sampling. The teacher predicts all positions, with the visible draft tokens serving as context. The guidance and decoding choices are included in the operator $\mathcal{T}_\phi^r$, and producing $\hat{z}_T$ requires no multi-step teacher rollout.

Connection to the fixed-point objective. Decoding $\hat{z}$ and $\hat{z}_T$ and applying the frozen backbone Φ yields the features X and Y defined in Sec. 3.3. The teacher correction and its target features are computed without gradient tracking: although the target changes with the student's current draft, it is held fixed during each student update.

A.3 Training Configurations

Tab. 5 summarizes the main training configurations. The discriminator uses two frozen feature backbones, `deit_base_distilled_patch16_224` and `tf_efficientnet_lite0`, followed by trainable multi-scale heads, and is optimized with a learning rate of 2×10^{-4}.

MaskGIT and MaskGen are fine-tuned with all parameters, and the main results are obtained with the adversarial loss. Unless marked otherwise, the ablations and Lumina-DiMOO (LoRA rank 64) are trained without it.

Table 5: Training configurations for main experimental pipelines.

Setting	MaskGIT / ImageNet	MaskGen / LAION	Lumina-DiMOO / LAION
Resolution	256×256	256×256	256×256
Condition	Class label	Text caption	Text caption
Tokenizer / decoder	VQGAN	TATiTok	Lumina VQ-VAE
Codebook size	1024	8192	8192
Sequence length	16×16	128	16×16
Total batch size	32	64	8
Optimizer	AdamW	AdamW	AdamW
Learning rate	1×10^{-6}	1×10^{-5}	5×10^{-5}
CFG	3.0	3.0	2.0
DINO backbone	DINOv3-B	DINOv3-B	DINOv3-B
DINO layers	$\{2, 5, 8, 11\}$	$\{2, 5, 8, 11\}$	$\{2, 5, 8, 11\}$
DINO pool	4×4 per layer	4×4 per layer	4×4 per layer
Kernel	Laplace	Laplace	Laplace
Kernel bandwidths	$\{0.02, 0.05, 0.2\}$	$\{0.02, 0.05, 0.2\}$	$\{0.02, 0.05, 0.2\}$
Negatives / sample	4	2	1
Positive / samples	1	1	1

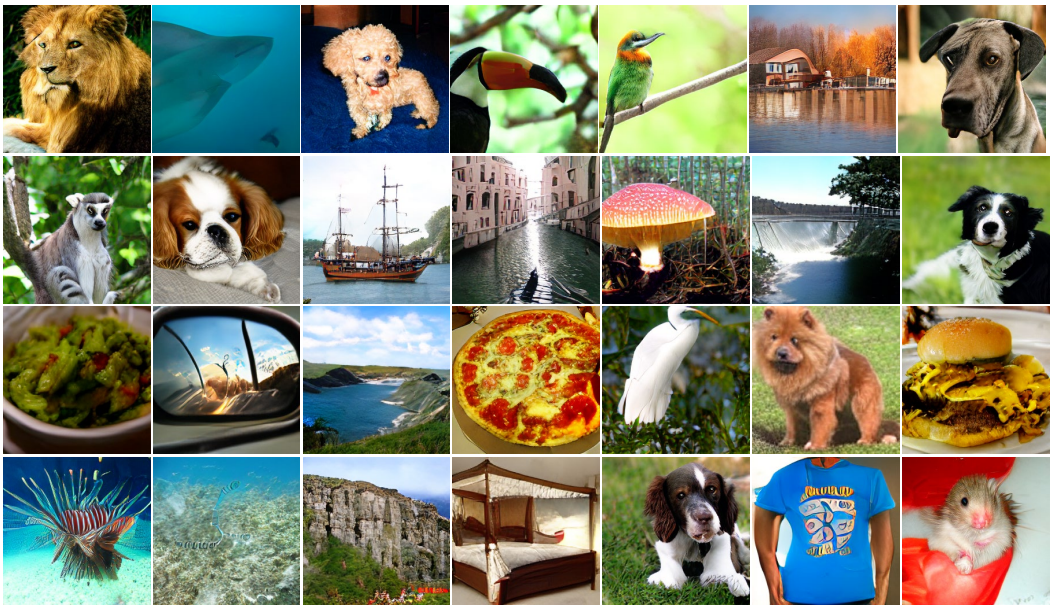


Figure 6: Single-step generation results of the MaskGIT model distilled via our proposed FPD framework on the ImageNet dataset.

A.4 Gradient of the Straight-Through Estimator

Let $\ell \in \mathbb{R}^V$ be the student logits, $p_v = \mathrm{softmax}(\ell/\tau)_v$ the categorical probabilities, and E_k the codebook vectors. The soft embedding is $\tilde{e} = \sum_{k=1}^{V} p_k E_k$.

The Jacobian of $\tilde{e}$ with respect to logit ℓ_v is given by:

$$\begin{aligned}\frac{\partial \tilde{e}}{\partial \ell_v} &= \sum_{k=1}^{V} \frac{\partial p_k}{\partial \ell_v} E_k \\ &= \sum_{k=1}^{V} \frac{1}{\tau} p_v(\delta_{kv} - p_k) E_k \\ &= \frac{1}{\tau} p_v(E_v - \tilde{e}).\end{aligned} \tag{15}$$

Applying the chain rule for the loss $\mathcal{L}$, the backward pass executes the exact vector-Jacobian product:

$$\frac{\partial \mathcal{L}}{\partial \ell_v} = \frac{1}{\tau} p_v (E_v - \tilde{e})^\top \nabla_{\tilde{e}} \mathcal{L}. \tag{16}$$

Intuitively, $E_v - \tilde{e}$ is the displacement from the soft embedding to the v-th codebook entry. The logit ℓ_v therefore increases whenever moving $\tilde{e}$ toward E_v decreases the loss, at a rate proportional to p_v/τ.

B Empirical Results

B.1 Visualizations on ImageNet

Fig. 6 provides qualitative visualizations of the MaskGIT model distilled via our Fixed-Point Distillation framework, evaluated on the ImageNet dataset. These samples demonstrate the capability of our approach to maintain high visual fidelity and strict class conditioning within a single inference step.

Table 6: GenEval performance with different numbers of positive samples.

Positive Samples	1	2	4
GenEval	0.43	0.41	0.40

Table 7: Comparison of FPD with relevant baseline methods. - means optional.

Method	Staged Training	Offline Dataset	Modified Arch	Aux Net	Discriminator	External Reward	One-step
SDTT	✓	×	×	×	×	×	×
Di4C	✓	×	✓	×	×	×	×
ReDi	✓	✓	×	×	×	×	✓
DiMO	×	×	×	✓	×	×	✓
Soft-DiMO	✓	×	×	✓	✓	✓	✓
D-MMD	×	×	✓	✓	×	×	×
FPD (Ours)	×	×	×	×	-	×	✓

B.2 Training Cost Analysis

In terms of distillation methods, the computational overhead of FPD (w/o GAN) and DiMO mainly depends on the teacher model. Specifically, the FLOPs per student update can be estimated as follows (with classifier-free guidance (CFG) enabled by default, requiring two teacher calls for both methods):

$$\begin{aligned} \text{FPD:} \quad & 5M + 3D + 3C, \\ \text{DiMO:} \quad & 6M + 3M \times P, \end{aligned} \tag{17}$$

where M and D denote the teacher model and decoder, respectively. P (1 by default) represents the update frequency of the auxiliary model, and C denotes the feature extractor.

For modern foundation models, $M \gg D$ and $M \gg C$ typically hold. Thus, the computational costs of FPD and DiMO can be approximated as $5M$ and $(6{+}3P)M$, respectively. With the default setting $P = 1$, these reduce to $5M$ and $9M$, respectively, demonstrating the computational efficiency of FPD. Furthermore, FPD achieves lower peak VRAM usage than DiMO.

B.3 Effect of Multiple Positive Samples

We further explore the use of multiple positive examples with MaskGen. Specifically, we corrupt each student's draft using different remasking ratios and obtain the corresponding teacher outputs as positive samples. The results are presented in Table 6. We observe that increasing the number of positive samples leads to a slight degradation in GenEval performance. We hypothesize that too many positive samples introduce an averaging effect, thereby blurring the teacher's guiding signal.

B.4 Method Comparisons

We compare FPD with relevant baseline methods. FPD offers the following characteristics, which previous methods do not achieve simultaneously:

- **Simplicity.** FPD does not require an auxiliary score network, multi-round stage distillation, offline sampling datasets, or RL post-training.
- **Effectiveness.** FPD works on its own and successfully supports one-step distillation. Incorporating GANs can further enhance FPD.
- **Generality.** FPD does not require modifications to the student architecture to accommodate specific theoretical assumptions.

B.5 Effect of DINOv3 Backbone Size

We evaluate the impact of the feature extractor size on generation quality on MaskGen using the GenEval benchmark. We experiment with several DINOv3 backbone sizes, including S, B, and L. As shown in Tab. 8, increasing the backbone size from S to B leads to a noticeable improvement in

Table 8: Generation performance with different DINOv3 backbone sizes on MaskGen.

DINOv3	S	B	L
GenEval	0.39	0.43	0.43



Figure 7: Failure cases on Lumina-DiMOO.

GenEval performance, while further scaling to L provides little additional gain. These results suggest that DINOv3-B provides a sufficiently strong feature representation for high-quality generation, offering a favorable trade-off between feature extraction capacity and computational cost.

B.6 Failure Case Analysis

As shown in Fig. 7, the model preserves the overall scene semantics, but exhibits mild artifacts in fine-grained textures and local structures, including brick patterns, object boundaries, small text, sheet music, and reflective surfaces.

B.7 Text Prompts for Main Figures

The complete text prompts utilized to generate the visual examples presented in Fig. 1 are detailed as follows:

- Macro photograph of a chameleon's eye, intricate scale patterns, vivid green and yellow colors, extreme close-up detail
- Photorealistic image of a cozy cabin interior with a crackling fireplace, snow visible through frosted windows, warm amber glow
- A great white shark breaching the ocean surface in South Africa, explosive water spray, dramatic action, marine wildlife photography
- Long exposure photograph of a waterfall in an ancient moss-covered forest, misty atmosphere, emerald green, Pacific Northwest
- Photorealistic image of a winding road through a tunnel of autumn maple trees, fallen leaves on asphalt, Vermont countryside
- Photorealistic winter scene of a frozen lake with ice fishing huts, pink and purple sunrise, steam rising from cracks in ice
- Close-up portrait of a young woman with heterochromia, one blue eye one brown eye, freckles, natural light, striking beauty
- A classic 1960s Mustang convertible on Route 66 at sunset, desert landscape stretching to horizon, Americana nostalgia
- Photorealistic image of a single perfect red rose with water droplets on petals, dark moody background, studio floral photography

The complete text prompts utilized to generate the visual examples presented in Fig. 3 are detailed as follows:

- 3D render of a cute low-poly fox in a geometric forest, soft pastel colors, ambient occlusion lighting

- Alien jungle planet with bioluminescent plants, strange creatures, two suns setting, sci-fi exploration
- 3D render of a glass marble containing a tiny ocean scene with waves and a lighthouse inside
- Art Nouveau poster of a woman surrounded by flowing floral vines, Alphonse Mucha inspired, ornate border

The prompt for Fig. 5 is: A photo shows a home office interior with light wood furniture. A large desk with drawers, a bookcase, and a two-drawer filing cabinet are arranged on a patterned rug. A laptop sits on the desk, next to a small potted plant and a lamp. A framed botanical print hangs on the wall, and a printer sits on the filing cabinet. A large arched window provides natural light.

The prompts for Fig. 7 are:

- A blue bicycle leans against a brick wall beneath a large round clock, with a brown leather satchel on the ground and two white flowers in a pot.
- A small bakery storefront has a striped awning, three golden croissants in the window, a pink cake, and a sign reading FRESH BREAD.
- A laboratory bench holds a brass microscope, three glass test tubes with blue liquid, a red notebook, and a yellow rubber glove.
- A polished violin rests on an open black instrument case, with a wooden bow, three sheet-music pages, and a white rose beside it.

C Discussion

C.1 Broader Impacts

Our framework effectively reduces the computational barrier associated with high-fidelity visual synthesis. By condensing the iterative diffusion process into a single inference step, we lower both the latency and the energy consumption required for image generation. This computational efficiency democratizes access to advanced generative tools, enabling researchers and practitioners with limited hardware resources to deploy large-scale models. Furthermore, rapid generation capabilities facilitate real-time interactive workflows, fostering new opportunities for digital artists and content creators while actively minimizing the environmental footprint of deploying generative artificial intelligence at scale.

C.2 Safeguards

As a distillation framework, the generative capabilities, world knowledge, and potential biases of our student model are fundamentally inherited from the pretrained teacher. Consequently, the safety and semantic boundaries of the generated content are primarily determined by the data distributions originally learned by the base model. Our core fixed-point distillation pipeline operates without real images, as it is driven entirely by the state-dependent signals provided by the teacher. While the auxiliary adversarial objective indirectly introduces a distribution of real images into the training process, it operates strictly unconditionally. This unconditional constraint ensures that the discriminator serves solely to enforce natural image manifold alignment rather than injecting new semantic concepts, thereby mitigating the risk of introducing targeted biases outside the teacher's original scope.

C.3 Limitations

Despite achieving competitive performance across benchmarks, we observe specific instances where the single-step generated images exhibit minor structural artifacts or a lack of fine-grained detail. We attribute these shortcomings to two primary factors. First, any inherent representational flaws or compositional failure modes present in the foundational teacher model are inevitably propagated to the student. Second, condensing a highly complex, multi-step generative process into a single network evaluation inherently introduces a representational capacity bottleneck, causing a natural

degradation in maximum fidelity. Addressing this capacity gap and further improving the robustness of one-step discrete generation remain valuable directions for future work.