

THE DIFFUSION DUALITY

Anonymous authors

Paper under double-blind review

ABSTRACT

Discrete diffusions models have been demonstrated to be surprisingly strong language models. In this work, we show that discrete diffusion language models can be further improved by adapting methods from continuous-state diffusion models. We establish a core property of uniform state diffusion: it stems from an underlying Gaussian diffusion process. This property allows us to improve both training by utilizing a curriculum learning strategy that reduces training variance and leads to $2\times$ faster convergence, as well as sampling by adapting efficient distillation methods from continuous-state diffusion models. As a result, models surpass an autoregressive model’s zero-shot perplexity on 3 out of 7 benchmarks and we manage to reduce the sampling steps by **two orders of magnitude** while preserving sample quality.

1 INTRODUCTION

An eternal theme in mathematics is that discreteness emerges from underlying continuity. From quantum mechanics, where the quantized energy states of electrons arise as solutions to continuous wave equations, to the Fourier decomposition of the Heaviside function, which results in a trigonometric series, and to the binary logic of digital circuits, fundamentally driven by smooth analog currents, discreteness has repeatedly and naturally emerged from an underlying continuum. Our work continues this tradition by demonstrating that a discrete diffusion process is, in fact, an emergent phenomenon of an underlying continuous Gaussian diffusion process. This perspective enables the design of faster training and sampling algorithms for discrete diffusion models.

Continuous diffusion models, particularly in the image domain, have had numerous advancements such as efficient parameterizations of the denoising model that improve upon mean-parameterization (Ho et al., 2020; Salimans & Ho, 2022; Zheng et al., 2023) for faster training, higher-order samplers that drastically reduce sampling steps compared to standard ancestral sampling (Karras et al., 2022), and distillation techniques enabling single-step generation (Song et al., 2023; Song & Dhariwal, 2023; Yin et al., 2024). In contrast, the design space of discrete diffusion models remains less explored. Ancestral sampling (and its equivalent Tweedie sampler (Campbell et al., 2022)) is still the standard (Austin et al., 2021; Sahoo et al., 2024a), and mean-parameterization remains the dominant approach (Sahoo et al., 2024a; Schiff et al., 2024b), with score parameterization (Lou et al., 2023) being the only alternative. While some progress has been made in distillation (Deschenaux & Gulcehre, 2024) and guidance techniques (Nisonoff et al., 2024; Schiff et al., 2024a), discrete diffusion remains largely underexplored.

The main contribution of this work is threefold: (1) We establish a theoretical connection between continuous and discrete diffusion, showing that discrete diffusion can be derived as a transformation of continuous Gaussian diffusion. This insight enables the transfer of techniques from the continuous domain to the discrete setting, unlocking new possibilities. (2) Leveraging this duality, we propose a training framework that enhances efficiency by introducing a low-variance curriculum based on the

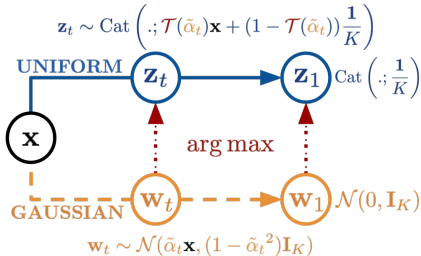


Figure 1: An illustration of uniform state discrete diffusion (top) and the underlying Gaussian diffusion (bottom). Applying $\arg \max$ maps Gaussian latents $w_t \in \mathbb{R}^n$ to discrete latents $z_t \in \mathcal{V}$, transforming their marginals from $\tilde{q}_t(\cdot|\mathbf{x}; \tilde{\alpha}_t)$ (6) to $q_t(\cdot|\mathbf{x}; \mathcal{T}(\tilde{\alpha}_t))$ (1) and adjusting diffusion parameters from $\tilde{\alpha}_t$ to $\alpha_t = \mathcal{T}(\tilde{\alpha}_t)$ (10). The ELBOs of both processes are related by (12).

underlying Gaussian diffusion. We refer to our method as **Duo**. (3) We accelerate sampling by **two orders of magnitude** by adapting efficient distillation techniques from Gaussian diffusion.

2 BACKGROUND

Notation. We represent scalar discrete random variables that can take K values as ‘one-hot’ column vectors and define $\mathcal{V} \in \{\mathbf{x} \in \{0, 1\}^K : \sum_{i=1}^K \mathbf{x}_i = 1\}$ as the set of all such vectors. Define $\text{Cat}(\cdot; \boldsymbol{\pi})$ as the categorical distribution over K classes with probabilities given by $\boldsymbol{\pi} \in \Delta^K$, where Δ^K denotes the K -simplex. Additionally, let $\mathbf{1} = \{1\}^K$ and $\langle \mathbf{a}, \mathbf{b} \rangle$ and $\mathbf{a} \odot \mathbf{b}$ respectively denote the dot and Hadamard products between two vectors \mathbf{a} and \mathbf{b} .

2.1 DISCRETE DIFFUSION MODELS

Consider the clean data $\mathbf{x} \in \mathcal{V}$ drawn from the data distribution q_{data} . In the discrete diffusion framework (Sohl-Dickstein et al., 2015; Austin et al., 2021) the complex data distribution q_{data} is mapped to a simple distribution through a sequence of markov states. Sahoo et al. (2024a) simplify this framework and propose an interpolating noise framework where the forward process $(q_t)_{t \in [0,1]}$ interpolates between the clean data distribution q_{data} and a prior distribution $\text{Cat}(\cdot; \boldsymbol{\pi})$ by defining a sequence of latents $\mathbf{z}_t \in \mathcal{V}$ whose marginals conditioned on \mathbf{x} at time t is given by:

$$q_t(\cdot | \mathbf{x}; \alpha_t) = \text{Cat}(\cdot; \alpha_t \mathbf{x} + (1 - \alpha_t) \boldsymbol{\pi}), \quad (1)$$

where the diffusion parameter $\alpha_t \in [0, 1]$ is a strictly decreasing function in t , with $\alpha_{t=0} \approx 1$ and $\alpha_{t=1} \approx 0$; see Sahoo et al. (2024a) for details. For Uniform State Diffusion Models (USDMs), $\boldsymbol{\pi} = \frac{1}{K} \mathbf{1}$, and for Masked Diffusion Models (MDMs) (Sahoo et al., 2024b), $\boldsymbol{\pi} = \mathbf{m}$ where $\mathbf{m} \in \mathcal{V}$ is a special mask token. The reverse diffusion model $p_t^\theta : \mathcal{V} \times [0, 1] \rightarrow \Delta^K$ is parameterized by a neural network with parameters θ . For a given datapoint $\mathbf{x} \sim \mathcal{D}$ sampled from the dataset \mathcal{D} , the denoising model is trained to maximize a variational lower bound (ELBO) on the log-likelihood $\log p_\theta(\mathbf{x})$. Given a number of discretization steps T , defining s as a shorthand for $s(i) = (i - 1)/T$ and t for $t(i) = i/T$, and using $D_{\text{KL}}[\cdot]$ to denote the Kullback–Leibler divergence, the ELBO ($q, p_\theta; \mathbf{x}$) for the forward process q and the corresponding reverse process p_θ equals (Sohl-Dickstein et al., 2015):

$$\mathbb{E}_q \left[\underbrace{\log p_{t=T}^\theta(\mathbf{x} | \mathbf{z}_{t(1)})}_{\mathcal{L}_{\text{recons}}} - \underbrace{\sum_{i=1}^T D_{\text{KL}}[q_{s|i}(\mathbf{z}_s | \mathbf{z}_t, \mathbf{x}) \| p_{s|i}^\theta(\mathbf{z}_s | \mathbf{z}_t)]}_{\mathcal{L}_{\text{diffusion}}} \right] - \underbrace{D_{\text{KL}}[q(\mathbf{z}_{t(T)} | \mathbf{x}) \| p_{t=T}^\theta(\mathbf{z}_{t(T)})]}_{\mathcal{L}_{\text{prior}}} \quad (2)$$

Schiff et al. (2024c) demonstrate that the true reverse posterior for USDMs is given as:

$$q_{s|t}(\cdot | \mathbf{z}_t, \mathbf{x}; \alpha_s, \alpha_t) = \text{Cat} \left(\cdot; \frac{K \alpha_t \mathbf{z}_t \odot \mathbf{x} + (\alpha_{t|s} - \alpha_t) \mathbf{z}_t}{K \alpha_t \langle \mathbf{z}_t, \mathbf{x} \rangle + 1 - \alpha_t} + \frac{(\alpha_s - \alpha_t) \mathbf{x} + (1 - \alpha_{t|s})(1 - \alpha_s) \mathbf{1}/K}{K \alpha_t \langle \mathbf{z}_t, \mathbf{x} \rangle + 1 - \alpha_t} \right) \quad (3)$$

where $\alpha_{t|s} = \alpha_t / \alpha_s$ and the approximate reverse posterior as $p_{s|t}^\theta(\cdot | \mathbf{z}_t; \alpha_s, \alpha_t) = q_{s|t}(\cdot | \mathbf{z}_t, \mathbf{x} = \mathbf{x}_\theta(\mathbf{z}_t, t); \alpha_t, \alpha_s)$. The terms $\mathcal{L}_{\text{recons}}$ and $\mathcal{L}_{\text{prior}}$ in (2) analytically reduce to 0 by choosing $\alpha_{t=0} = 1$ and $\alpha_{t=1} = 0$. Furthermore, by setting $T \rightarrow \infty$, the ELBO (2) for USDMs reduces to (Schiff et al., 2024c):

$$\text{ELBO}(q, p_\theta; \mathbf{x}) = \mathbb{E}_{t \sim \mathcal{U}[0,1], q_t(\mathbf{z}_t | \mathbf{x}; \alpha_t)} f(\mathbf{z}_t, \mathbf{x}_\theta(\mathbf{z}_t, t), \alpha_t; \mathbf{x}). \quad (4)$$

where

$$f(\mathbf{z}_t, \mathbf{x}_\theta(\mathbf{z}_t, t), \alpha_t; \mathbf{x}) = \frac{\alpha_t'}{K \alpha_t} \left[\frac{K}{\bar{\mathbf{x}}_i} - \frac{K}{(\bar{\mathbf{x}}_\theta)_i} - \sum_{j \text{ s.t. } (\mathbf{z}_t)_j=0} \left(\frac{\bar{\mathbf{x}}_j}{\bar{\mathbf{x}}_i} \right) \log \left(\frac{(\bar{\mathbf{x}}_\theta)_i \cdot \bar{\mathbf{x}}_j}{(\bar{\mathbf{x}}_\theta)_j \cdot \bar{\mathbf{x}}_i} \right) \right]. \quad (5)$$

where \mathbf{x}_i denotes the i^{th} index of a vector \mathbf{x} , $\bar{\mathbf{x}} = K \alpha_t \mathbf{x} + (1 - \alpha_t) \mathbf{1}$, $\bar{\mathbf{x}}_\theta = K \alpha_t \mathbf{x}_\theta(\mathbf{z}_t, t) + (1 - \alpha_t) \mathbf{1}$, α_t' denotes the time-derivative of the α_t , and $i = \arg \max_{j \in [K]} (\mathbf{z}_t)_j$ is the non-zero entry of \mathbf{z}_t .

2.2 GAUSSIAN DIFFUSION MODELS

Similar to USDMs, Gaussian diffusion maps a data distribution q_{data} to a simple prior distribution through a sequence of noisy latents $\mathbf{w}_t \sim \tilde{q}_t(\cdot|\mathbf{x})$, whose marginal distribution is given by:

$$\tilde{q}_t(\cdot|\mathbf{x}; \tilde{\alpha}_t) = \mathcal{N}(\tilde{\alpha}_t \mathbf{x}, (1 - \tilde{\alpha}_t^2) \mathbf{I}_n), \quad (6)$$

where the diffusion parameter $\tilde{\alpha}_t \in [0, 1]$ is a monotonically decreasing function in t . Since, by design, $\tilde{\alpha}_{t=0} \approx 1$ and $\tilde{\alpha}_{t=1} \approx 0$, the $\mathcal{L}_{\text{recons}} \approx 0$ and $\mathcal{L}_{\text{prior}} \approx 0$ and the ELBO is given as:

$$\text{ELBO}(\tilde{q}, p_\theta; \mathbf{x}) = \mathbb{E}_{t \sim \mathcal{U}[0,1], \tilde{q}_t(\mathbf{w}_t|\mathbf{x}; \tilde{\alpha}_t)} \nu'(t) \|\mathbf{x} - \mathbf{x}_\theta(\mathbf{w}_t, t)\|_2^2 \quad (7)$$

where $\nu'(t)$ is the time derivative of the signal-to-noise ratio $\nu(t) = \tilde{\alpha}_t^2 / (1 - \tilde{\alpha}_t^2)$.

2.3 CONSISTENCY DISTILLATION

Consistency models (Song et al., 2023; Song & Dhariwal, 2023) are a class of generative models that can be initialized from pre-trained diffusion models and generate high-quality samples in just a few steps. They build upon deterministic samplers for Gaussian diffusion (Song et al., 2020; 2021), specifically leveraging the Probability-Flow ODE (PF-ODE). To train a consistency model, we first generate a noisy sample \mathbf{w}_t from the forward process $\tilde{q}_t(\cdot|\mathbf{x})$ (6) and obtain a less noisy sample \mathbf{w}_s by numerically solving the PF-ODE for one step using the denoising model \mathbf{x}_θ . The consistency model then trains a student model \mathbf{x}_θ (with parameters θ) to match the teacher model’s estimate of the clean sample, given the noisy sample \mathbf{w}_t . The teacher model \mathbf{x}_{θ^-} (with parameters θ^-) provides the less noisy sample \mathbf{w}_s , and the student model is optimized to minimize the loss:

$$\mathcal{L}(\theta, \theta^-) = \lambda(t) d(\mathbf{x}_\theta(\mathbf{w}_t, t), \mathbf{x}_{\theta^-}(\mathbf{w}_s, s)), \quad (8)$$

where $d: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^+$ denotes the error between the teacher model’s reconstruction $\mathbf{x}_{\theta^-}(\mathbf{w}_t, t)$ and the student model’s reconstruction $\mathbf{x}_\theta(\mathbf{w}_t, t)$ of the original sample and $\lambda: [0, 1] \rightarrow \mathbb{R}^+$ is a weighting function that scales the loss based on the diffusion time-step t . Typically, the teacher model is set as the Exponentially Moving Average (EMA) of the student model’s parameters during training.

3 DIFFUSION DUALITY

Gaussian diffusion is well-studied. There have been many theoretically grounded advances in improving training (Ho et al., 2020; Salimans & Ho, 2022; Zheng et al., 2023) and sampling (Karras et al., 2022; Song et al., 2023; Song & Dhariwal, 2023; Yin et al., 2024). Our goal in this section is to establish a theoretical connection between discrete-state diffusion and continuous-state diffusion which will allow us to leverage tools from continuous diffusion models to improve discrete diffusion.

There are two main questions that we must answer in order to apply methods for efficient training and sampling in continuous diffusion to discrete diffusion. The first question is how to map the continuous valued samples from a Gaussian diffusion process to discrete valued samples from discrete diffusion process. We answer this with a simple construction: we map Gaussian vectors to discrete space with an **arg max** operation. The second question is how to relate their marginal distributions. We derive a closed-form expression for that the diffusion parameters of Gaussian and discrete diffusion processes which allows us to match their marginal distributions.

We begin by defining a Gaussian diffusion process on $\mathbf{x} \in \mathcal{V}$ as per (6), with $\tilde{q}_{t=0} \approx q_{\text{data}}$ and $\tilde{q}_{t=1} = \mathcal{N}(0, \mathbf{I}_K)$. Let $\mathbf{w}_t \sim \tilde{q}_t(\cdot|\mathbf{x})$ be an intermediate latent at time t . Next, define the operation **arg max**: $\mathbb{R}^K \rightarrow \mathcal{V}$ as the transformation that maps a continuous vector $\mathbf{w} \in \mathbb{R}^K$ to a one-hot vector corresponding to the index of the largest entry in \mathbf{w} . We define **arg max**(\mathbf{w}) as $\mathbf{arg max}_{z \in \mathcal{V}} \mathbf{z}^\top \mathbf{w}$. Let $\mathbf{z}_t = \mathbf{arg max}(\mathbf{w}_t)$. In Suppl. B, we demonstrate that \mathbf{z}_t is distributed as:

$$\mathbf{z}_t \sim \text{Cat}\left(\cdot; \mathcal{T}(\tilde{\alpha}_t) \mathbf{x} + (1 - \mathcal{T}(\tilde{\alpha}_t)) \frac{\mathbf{1}}{K}\right), \quad (9)$$

where the function $\mathcal{T}: [0, 1] \rightarrow [0, 1]$ is given as:

$$\mathcal{T}(\tilde{\alpha}_t) = \frac{K}{K-1} \left[\int_{-\infty}^{\infty} \phi\left(z - \frac{\tilde{\alpha}_t}{\sqrt{1 - \tilde{\alpha}_t^2}}\right) \Phi^{K-1}(z) dz - \frac{1}{K} \right], \quad (10)$$

where $\phi(z) = \exp(-z^2)/\sqrt{2\pi}$ is the standard Normal distribution and $\Phi(z) = \int_{-\infty}^z \exp(-t^2/2)dz/\sqrt{2\pi}$ is the cumulative distribution function of the normal distribution.

The implications of (9) are quite profound. The discretized latent \mathbf{z}_t follows the same distribution as the marginal distribution of an intermediate latent in (1) when \mathbf{x} undergoes uniform state discrete diffusion with a diffusion parameter $\mathcal{T}(\tilde{\alpha}_t)$. This reveals a fundamental connection between uniform state discrete diffusion and Gaussian diffusion:

*The **arg max** operation transforms Gaussian diffusion into uniform state diffusion, with the diffusion coefficients related by (10).*

More formally, this can be expressed as:

$$q_t(\mathbf{z}_t|\mathbf{x}; \mathcal{T}(\tilde{\alpha}_t)) = [\text{arg max}]_\star \tilde{q}_t(\mathbf{w}_t|\mathbf{x}; \tilde{\alpha}_t) \quad (11)$$

where \star operator denotes the application of the **arg max** operation on a Gaussian probability density function transforming it to a categorical probability mass function. Hence, as \mathbf{x} undergoes a uniform-state diffusion in the discrete space, there is an underlying representation in which \mathbf{x} undergoes Gaussian diffusion in the continuous space, as depicted in Fig. 1.

Note that these two processes are separate Markov chains with no transitions between them, and they induce different bounds on the log-likelihood of the data (as will be discussed later in this section). In this work, we exploit this equivalence to design a low variance training algorithm that leads to faster training (Sec. 4.1) and a distillation scheme that speeds up the sampling process from discrete diffusion models by two orders of magnitude (Sec. 4.2).

Evidence Lower Bound These two processes have 2 different ELBOs– (4) for the discrete diffusion process and (7) for the Gausssian diffusion process.

Theorem 3.1. *ELBO for the uniform state discrete diffusion process is tighter than the underlying Gaussian diffusion process.*

We provide a detailed proof in Suppl. A. In brief, we derive the following relationship:

$$\log p_\theta(\mathbf{x}) \geq \text{ELBO}(q, p_\theta; \mathbf{x}) \geq \text{ELBO}(\tilde{q}, p_\theta; \mathbf{x}), \quad (12)$$

where the equality holds for an optimal denoiser p_θ . Because the ELBO is naturally tighter in the discrete space, it is advantageous to operate in the discrete space.

4 APPLICATIONS

We now present two applications where discrete diffusion models benefit from the underlying Gaussian diffusion. In Sec. 4.1, we introduce a curriculum learning strategy that reduces training variance and speeds up convergence. Then, in Sec. 4.2, we propose a distillation algorithm that cuts the number of sampling steps by two orders of magnitude with minimal impact on sample quality.

4.1 FASTER TRAINING USING CURRICULUM LEARNING

Curriculum learning (Bengio et al., 2009) gradually exposes models to increasingly complex data, starting with simpler, easier-to-denoise noise patterns and progressing to more challenging ones. Here, we design a curriculum for USDMs by exploiting the underlying Gaussian diffusion.

Similar to relaxation methods in discrete gradient estimation (Jang et al., 2017; Maddison et al., 2017), our curriculum is centered around annealing the temperature parameter of a smooth approximation of **arg max**. We reformulate the ELBO for discrete diffusion in terms of **arg max** over Gaussian latents (Sec. 4.1.1). We provide a lower-variance but biased estimator of the ELBO by relaxing the argmax operator with tempered softmax (Sec. 4.1.2). At the beginning of this curriculum, training resembles a simple Gaussian diffusion process – with the ELBO for discrete diffusion – and gradually transitions towards the standard discrete diffusion process.

4.1.1 DISCRETE ELBO WITH GAUSSIAN LATENTS

Consider the discrete diffusion ELBO, $\mathcal{L}_{\text{diffusion}}$, from (2), which marginalizes $f(\mathbf{x}, \mathbf{z}_t, \alpha_t, \mathbf{x}_\theta(\mathbf{z}_t))$ over $\mathbf{x} \sim q_{\text{data}}$ and $\mathbf{z}_t \sim q(\cdot|\mathbf{x})$ for $t \in [0, 1]$. We aim to re-express this in terms of Gaussian latents $\mathbf{w}_t \sim \tilde{q}_t(\cdot|\mathbf{x})$ such that marginalizing over \mathbf{w}_t yields the same value. In Suppl. C, we show

$$\begin{aligned} \mathcal{L}_{\text{diffusion}} &= \mathbb{E}_{t \sim \mathcal{U}[0,1], q_t(\mathbf{z}_t|\mathbf{x}; \alpha_t)} f(\mathbf{z}_t, \mathbf{x}_\theta(\mathbf{z}_t), \alpha_t; \mathbf{x}) \\ &= \mathbb{E}_{t \sim \mathcal{U}[0,1], \tilde{q}_t(\mathbf{w}_t|\mathbf{x}; \tilde{\alpha}_t)} f(\mathbf{z}_t := \arg \max(\mathbf{w}_t), \mathbf{x}_\theta(\arg \max(\mathbf{w}_t)), \alpha_t := \mathcal{T}(\tilde{\alpha}_t); \mathbf{x}), \end{aligned} \quad (13)$$

where $\alpha_t = \mathcal{T}(\tilde{\alpha}_t)$ is obtained via (10) from the Gaussian diffusion coefficient $\tilde{\alpha}_t$. Importantly, this reparameterization **does not imply that we are defining a markov process between a Gaussian latent and a discrete latent**. As mentioned earlier in Sec. 3, these two are separate markov processes whose marginals are merely connected by (11). This reparameterization underpins our curriculum learning strategy which we present in the next section.

4.1.2 LOW VARIANCE TRAINING LOSS

To control training variance, we replace $\arg \max(\mathbf{w}_t)$ with a tempered softmax. First, note that $\arg \max$ can be expressed as a limiting case of softmax (Jang et al., 2017; Maddison et al., 2017):

$$\arg \max(\mathbf{w}_t) = \lim_{\tau \rightarrow 0^+} \text{softmax}\left(\frac{\mathbf{w}_t}{\tau}\right). \quad (14)$$

We relax the $\arg \max$ operation by introducing a temperature parameter $\tau > 0$. In the extreme case where $\tau \gg 0$, the values of $\mathbf{w}_t \in \mathbb{R}$ are mapped near the center of the simplex defined by \mathcal{V} , thereby significantly reducing variance in the input to the denoising model. Conversely, as $\tau \rightarrow 0$, the softmax output moves closer to the simplex vertices, introducing more variability.

Unlike previous discrete diffusion approaches (Sahoo et al., 2024a; Austin et al., 2021; Lou et al., 2023), we design the denoising model $p_t^\theta : \Delta^K \cup \mathcal{V} \rightarrow \Delta^K$ to handle both continuous latents and discrete latents; see Suppl. D.1 for more details. We set τ as a function of the training iteration n , annealing it from an initial value $\tau(n=0) = \tau_{\text{max}}$ to a final value $\tau(n=N) = \tau_{\text{min}} \approx 0$, where N is the total number of training steps. This ensures that $\text{softmax}(\mathbf{w}_t/\tau(n))$ produces values closer to the vertices of the simplex as $\tau(n)$ decreases and resembles the samples from a discrete diffusion process. As shown in Figure 3, this approach results in lower training variance compared to previous Absorbing and Uniform State discrete diffusion models, ultimately leading to an improved ELBO.

Thus, we define the following training loss:

$$\mathcal{L}_{\text{train}} = \mathbb{E}_{t \sim [0,1], \tilde{q}_t(\mathbf{w}_t|\mathbf{x}; \tilde{\alpha}_t)} f\left(\mathbf{z}_t := \arg \max(\mathbf{w}_t), \mathbf{x}_\theta(\text{softmax}(\mathbf{w}_t/\tau), t), \alpha_t := \mathcal{T}(\tilde{\alpha}_t); \mathbf{x}\right). \quad (15)$$

This loss doesn't correspond to a valid ELBO because the denoising model operates on a continuous-time random variable, while the ELBO is defined for a discrete diffusion process. **It only becomes a valid ELBO in the limiting case $\lim_{\tau \rightarrow 0^+}$** . During evaluation, we evaluate the model as a discrete diffusion model using (5).

4.2 DUAL CONSISTENCY DISTILLATION

In this section, we introduce a method to accelerate sampling by leveraging the dual Gaussian diffusion. Specifically, we adapt Consistency Distillation, a widely used technique for distilling Gaussian diffusion models. This approach relies on deterministic trajectories from the teacher model, generated via PF-ODEs. However, discrete diffusion lacks PF-ODEs, making direct application infeasible. To address this, we propose DCD (Dual Consistency Distillation), which overcomes this limitation by utilizing the PF-ODE of the underlying Gaussian diffusion to generate deterministic trajectories. In Gaussian diffusion, the trained score model enables deterministic sampling via PF-ODE simulation. In contrast, in discrete diffusion, the denoiser p_t^θ cannot be applied to Gaussian latents at inference since it is exclusively trained on discrete latents (as τ in (14) is annealed to 0).

Deterministic Discrete Trajectories (DDT) Given a clean data sample $\mathbf{x} \sim q_{\text{data}}$ and a noise sample $\epsilon \sim \mathcal{N}(0, \mathbf{I}_K)$, we generate a sequence of latents that follow a deterministic trajectory by reversing

the PF-ODE using the DDIM sampler under an optimal denoiser (Song et al., 2021); see Suppl. F for a detailed discussion. Let $\mathcal{P}_{\text{DDIM}}(\mathbf{x}, \epsilon) = \{\tilde{\alpha}_t \mathbf{x} + \sqrt{1 - \tilde{\alpha}_t^2} \epsilon\}_{t \in [0,1]}$ denote such a trajectory. Next, we map this sequence of Gaussian latents to the discrete space using `arg max` operator in the following manner: $\mathcal{P}_{\text{DDT}}(\mathbf{x}, \epsilon) = \{\mathbf{z}_t = \arg \max (\tilde{\alpha}_t \mathbf{x} + \sqrt{1 - \tilde{\alpha}_t^2} \epsilon)\}_{t \in [0,1]}$, where \mathcal{P}_{DDT} represents the mapping of $\mathcal{P}_{\text{DDIM}}$ to the discrete space. In the later part of this section, we discuss how these trajectories are used to specify the inputs to the student and the teacher model while performing distillation.

Distillation Given a teacher model \mathbf{x}_{θ^-} , our goal is to distill it into a student model \mathbf{x}_{θ} that can generate samples of similar quality but in fewer steps. To perform distillation, we sample an adjacent pair of latents $(\mathbf{z}_s, \mathbf{z}_t) \sim \{(\mathbf{z}_{j-\Delta}, \mathbf{z}_j) | \mathbf{z}_{\cdot} \in \mathcal{P}_{\text{DDT}}(\mathbf{x}, \epsilon), j \in [\Delta, 1]\}$ for a given step size $\Delta \in [0, 1]$. Here, \mathbf{z}_s is a less noisy sample than \mathbf{z}_t , and \mathbf{z}_t serves as the input to the student model. Let $\mathbf{x}_{\theta^-}(\mathbf{z}_s)$ and $\mathbf{x}_{\theta}(\mathbf{z}_t)$ represent the distributions of samples generated by the teacher and student models, respectively. We train the student model to match the teacher model’s distribution by minimizing the loss $D_{\text{KL}}(\mathbf{x}_{\theta}(\mathbf{z}_t), \mathbf{x}_{\theta^-}(\mathbf{z}_s))$ as proposed in Deschenaux & Gulcehre (2024). Thus the final distillation loss is $\mathcal{L}_{\text{DCD}}(\theta, \theta^-) = D_{\text{KL}}(\mathbf{x}_{\theta}(\mathbf{z}_t, t), \mathbf{x}_{\theta^-}(\mathbf{z}_s, s))$. Following (Song et al., 2023), we optimize this objective via stochastic gradient descent (SGD) on the student model parameters θ , while updating the teacher model parameters θ^- using an exponential moving average (EMA). The distillation process consists of K rounds, each with M training steps. The full algorithm is summarized in Algorithm 1.

5 EXPERIMENTS

We evaluate Duo on standard language modeling benchmarks, training on LM1B (Chelba et al., 2014) and OpenWebText (OWT)(Gokaslan et al., 2019) with sequence packing(Raffel et al., 2020). We train our models for 1M steps with a batch size of 512 on both datasets. For LM1B, we use a context length of 128 with the `bert-base-uncased` tokenizer (Devlin et al., 2018). Unlike Austin et al. (2021), prior works (Sahoo et al., 2024a; Lou et al., 2023; He et al., 2022) did not use sequence packing, making our setup more challenging and leading to higher perplexity in retrained models (Table 3). For OWT, we use a context length of 1024 with the GPT-2 tokenizer (Radford et al., 2019). Following Sahoo et al. (2024a), we reserve the last 100K documents for validation. Our model is a 170M-parameter modified diffusion transformer (DiT)(Peebles & Xie, 2023) with rotary positional encoding(Su et al., 2023) and adaptive layer norm for conditioning on diffusion time, consistent with prior work (Lou et al., 2023; Sahoo et al., 2024a). Training is conducted on 8xA100s or 8xH100s with `bfloat16` precision. We train Duo using (15), which requires computing the integral in (10). To reduce computation overhead, we precompute and cache 100K $(\tilde{\alpha}_t, \mathcal{T}(\tilde{\alpha}_t))$ tuples, significantly smaller than the denoising network. The Gaussian diffusion parameter $\tilde{\alpha}_t$ is parameterized using a linear schedule i.e., $(\tilde{\alpha}_t = 1 - t)_{t \in [0,1]}$.

5.1 IMPROVED TRAINING

Our experiments show that (1) the proposed curriculum learning strategy (Sec. 4.1.2) **accelerates training by 2× and achieves a new state-of-the-art among USDMs** (Tables 1, 3), and (2) Duo performs competitively with Absorbing State diffusion across major language modeling benchmarks, even **surpassing AR models on 3/7 zero-shot PPL benchmarks** (Table 2).

Experimental setup The primary baselines for Duo are the leading USDMs (SEDD Uniform (Lou et al., 2023) and UDLM (Schiff et al., 2024b)) and the SOTA Gaussian diffusion method, PLAID (Gulrajani & Hashimoto, 2024). For PLAID on LM1B, we retrained it without self-conditioning (Chen et al., 2023) to match our denoising model’s parameter count. Due to their inefficient open-source codebase¹, we report PLAID results for LM1B at 100K steps, as further training was infeasible. For OWT, we report results from Lou et al. (2023), where PLAID was trained with self-conditioning for 1.3M steps, favoring the baseline. Additionally, we compare Duo with autoregressive models and absorbing state discrete diffusion models, including MDLM (Sahoo et al., 2024a) (SOTA), SEDD Absorb (Lou et al., 2023), and D3PM Absorb (Austin et al., 2021).

¹<https://github.com/igul222/plaid>

Table 2: Zero-shot perplexities (\downarrow) of models trained for 1M steps on OWT. All perplexities for diffusion models are upper bounds. [†] Taken from Sahoo et al. (2024a). [¶] Taken from (Lou et al., 2023) models were trained for 1.3Msteps as opposed to the baselines that were trained for 1Msteps. All perplexities for diffusion models are upper bounds. Best uniform / Gaussian diffusion values are **bolded** and diffusion values better than AR are underlined. [‡] denotes retrained model.

	PTB	Wikitext	LM1B	Lambda	AG News	Pubmed	Arxiv
<i>Autoregressive</i>							
Transformer [†]	82.05	25.75	51.25	51.28	52.09	49.01	41.73
<i>Diffusion (absorbing state)</i>							
SEDD Absorb [†]	100.09	34.28	68.20	<u>49.86</u>	62.09	<u>44.53</u>	<u>38.48</u>
D3PM Absorb [¶]	200.82	50.86	138.92	93.47	-	-	-
MDLM [†]	95.26	32.83	67.01	<u>47.52</u>	61.15	<u>41.89</u>	<u>37.37</u>
<i>Diffusion (uniform state / Gaussian)</i>							
SEDD Uniform [‡]	105.51	41.10	82.62	57.29	82.64	55.89	50.86
Plaid [¶]	142.60	50.86	91.12	57.28	-	-	-
Duo (Ours)	89.35	33.57	73.86	<u>49.78</u>	67.81	<u>44.48</u>	<u>40.39</u>

Faster training We analyze the effect of τ on training loss. Figure 6 visualizes the loss on the LM1B dataset over 150K steps for $\tau \in \{0, 0.001, 0.01, 0.1\}$. Here, $\tau = 0$ (blue) corresponds to (5), meaning no curriculum learning. A larger τ reduces training variance but introduces bias. Ideally, τ should balance both. When $\tau = 0.1$ (red), excessive bias causes a sharp loss drop, making it suboptimal. Lowering τ to 0.01 (orange) and 0.001 (purple) stabilizes the loss, with $\tau = 0.001$ being optimal—it closely follows the blue curve (low bias) while maintaining significantly lower variance. Thus, we set $\tau(n < 500K) = 0.001$ for the first 500K steps, then switch to $\tau(n > 500K) = 0$ (as in (5)) until 1M steps to eliminate bias. After just 10K fine-tuning steps as a discrete diffusion model (510K total steps), Duo achieves a likelihood of 35.20—almost 1.5 perplexity points better than UDLM trained for 1M steps— **curriculum learning accelerates convergence by at least 2 \times** .

Likelihood Evaluation To compute ppl for Duo, we use (5) with $\alpha_t = 1 - t$ (Schiff et al., 2024b). On LM1B (Table 3) and OWT (Table 1), Duo outperforms previous USDMs and Gaussian diffusion models notably SEDD Uniform and UDLM and shrinks the gap with absorbing diffusion below 2 perplexity points. On LM1B, We retrained Plaid which attained a perplexity of 89.91 in 100K steps while DUO achieves a perplexity of 43.01 in the same number of steps. We don’t report this number in the table due to incomplete training.

Zero-Shot Likelihood Evaluation We measure the zero-shot generalization of the models trained on OWT by evaluating their PPL on 7 other datasets. Following Sahoo et al. (2024a), our zero-shot datasets include the validation splits of Penn Tree Bank (PTB; Marcus et al. (1993)), WikiText (Merity et al., 2016), LM1B, Lambda (Paperno et al., 2016), AG News (Zhang et al., 2015), and Scientific papers from ArXiv and Pubmed (Cohan et al., 2018). We observe that Duo outperforms SEDD Uniform and Plaid across all benchmarks. More notably, it achieves a better PPL score than SEDD Absorbing on 4 / 7 datasets, MDLM on 1 / 7, most notably, **ourperforming an autoregressive transformer on 3 / 7 datasets**.

Table 1: Test perplexities (PPL; \downarrow) on OWT for models trained for 262B tokens. [†] Reported in Sahoo et al. (2024a). We report bounds for diffusion models. Best diffusion value is underlined. [‡] Denotes retrained model.

	PPL (\downarrow)
<i>Autoregressive</i>	
Transformer [†]	17.54
<i>Diffusion (absorbing state)</i>	
SEDD Absorb [†] (Lou et al., 2023)	24.10
MDLM [†] (Sahoo et al., 2024a)	<u>23.21</u>
<i>Diffusion (uniform state)</i>	
SEDD Uniform [‡] (Lou et al., 2023)	29.69
Duo (Ours)	<u>25.20</u>

5.2 IMPROVED SAMPLING

Our sampling experiments show that for undistilled models, (1) **Duo generates higher-quality samples than all previous diffusion models** (Table 7). (2) After distillation, **Duo dominates in the low-sampling-step regime** (Fig. 2)

Experimental Setup We distill Duo on OWT, using MDLM distilled with SDTT (Deschenaux & Gulcehre, 2024) as our primary baseline. Similar to SDTT, we perform $K = 5$ rounds of distillation, setting the discretization step $\Delta = 1/512$ in Algorithm 1 and doubling it every $M = 10k$ steps. To evaluate sample quality, we use GPT-2 Large generative perplexity (Gen PPL) as a quality metric and average sequence entropy as a diversity metric. As noted by Zheng et al. (2024), low-precision sampling can be problematic in masked diffusion models, leading to reduced diversity and potentially misleading Gen PPL scores. To mitigate this, we use `float64` precision for all sampling experiments.

Results We observe that DUO surpasses all previous diffusion models in terms of Gen PPL across all sampling steps $T \in \{8, \dots, 1024\}$ (Fig. 7). Notably, the entropy of MDLM, SEDD Absorb, and SEDD Uniform closely resembles that of the autoregressive model without nucleus sampling. Meanwhile, for $T \in \{32, \dots, 1024\}$, Duo’s entropy aligns with that of the AR model using nucleus sampling with $p = 0.9$. The entropies of MDLM, SEDD Absorb, and SEDD uniform and the autoregressive model w/o nucleus sampling are surprisingly similar. Duo’s entropy for $T \in \{32, \dots, 1024\}$ is similar to that of the AR model with nucleus sampling $p = 0.9$. However, for $T \in \{8, 16\}$, Duo’s entropy drops to 4.9 and 5.1, indicating lower diversity in samples. Later, we show that distillation mitigates this issue by increasing entropy and reducing Gen PPL for smaller T values.

Distillation In Fig. 2, we compare Duo (DCD-distilled) with MDLM (SDTT-distilled), where darker shades indicate more distillation rounds. Duo outperforms for all T values up to round 2. After five rounds, DUO dominates in the low NFE region ($T \leq 64$), while MDLM excels in the high NFE region ($T \geq 64$). Notably, each distillation round increases Duo’s sample entropy and reduces Gen PPL, improving diversity and quality (Fig. 8).

6 RELATED WORK AND CONCLUSION

Related Work Previous work attempted to use Gaussian diffusion for language modeling. Plaid (Gulrajani & Hashimoto, 2024), DiffusionLM (Li et al., 2022) and CDCD (Dieleman et al., 2022) inject Gaussian noise in continuous embedding vectors. However, these achieve poorer performance than recent discrete diffusion models (Lou et al., 2023; Sahoo et al., 2024a; Shi et al., 2025). We prove that the discrete ELBO is tighter, and therefore results in a better model. Prior work in discrete diffusion language models adhere strictly to discrete space: D3PM (Austin et al., 2021) introduces a discrete time, discrete space framework using Markov corruption processes, Masked diffusion models Sahoo et al. (2024a); Ou et al. (2024); Shi et al. (2025); He et al. (2022) advance the absorbing corruption process, Lou et al. (2023) defines the forward process in terms of a continuous time Markov process. While previous work in this paragraph studies discrete diffusion in isolation from Gaussian diffusion processes, our work shows that uniform discrete space diffusion emerges as the arg max of an underlying Gaussian process, and we use this connection to improve training and sampling. Deschenaux & Gulcehre (2024) introduce a distillation scheme for absorbing diffusion models that doesn’t rely on deterministic samplers unlike distillation schemes in Gaussian diffusion (Salimans & Ho, 2022; Luhman & Luhman, 2021). Our proposed DCD algorithm is a form of consistency distillation (Song et al., 2023; Song & Dhariwal, 2023).

Conclusion In this work, we formulated a theoretical connection between continuous, Gaussian diffusion models and discrete, uniform-state diffusion models. We exploited this connection to achieve a 2x speed-up in training convergence, as well as two-orders of magnitude improvement in sampling speed. We hope that our theoretical foundation inspires further connections between efficient methods for continuous and discrete diffusion models.

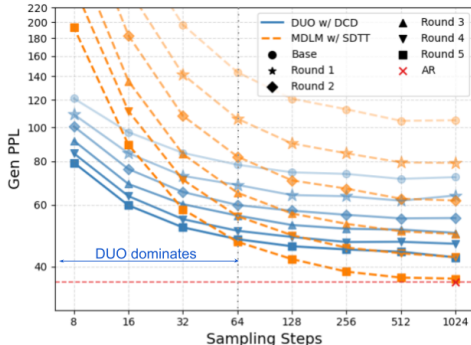


Figure 2: Distillation results for DUO when distilled using DCD (1) and MDLM distilled using SDTT. After distillation, Duo dominates in the low sampling steps regime ≤ 64 .

REFERENCES

- 432
433
434 Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured
435 denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing*
436 *Systems*, 34:17981–17993, 2021.
- 437
438 Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Inter-*
439 *national Conference on Machine Learning*, 2009. URL <https://api.semanticscholar.org/CorpusID:873046>.
- 440
441 Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, George Deligiannidis, and
442 Arnaud Doucet. A continuous time framework for discrete denoising models. *Advances in Neural*
443 *Information Processing Systems*, 35:28266–28279, 2022.
- 444
445 Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony
446 Robinson. One billion word benchmark for measuring progress in statistical language modeling,
447 2014.
- 448
449 Ting Chen, Ruixiang ZHANG, and Geoffrey Hinton. Analog bits: Generating discrete data using
450 diffusion models with self-conditioning. In *The Eleventh International Conference on Learning*
451 *Representations*, 2023. URL <https://openreview.net/forum?id=3itjR9QxFw>.
- 452
453 Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang,
454 and Nazli Goharian. A discourse-aware attention model for abstractive summarization of long
455 documents. *Proceedings of the 2018 Conference of the North American Chapter of the Association*
456 *for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018.
doi: 10.18653/v1/n18-2097. URL <http://dx.doi.org/10.18653/v1/n18-2097>.
- 457
458 Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov.
459 Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint*
arXiv:1901.02860, 2019.
- 460
461 Justin Deschenaux and Caglar Gulcehre. Beyond autoregression: Fast llms via self-distillation
462 through time. *arXiv preprint arXiv:2410.21035*, 2024.
- 463
464 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep
465 bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- 466
467 Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H
468 Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, et al. Continuous diffusion
for categorical data. *arXiv preprint arXiv:2211.15089*, 2022.
- 469
470 Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. Openwebtext corpus. [http://](http://Skyllion007.github.io/OpenWebTextCorpus)
471 Skyllion007.github.io/OpenWebTextCorpus, 2019.
- 472
473 Ishaan Gulrajani and Tatsunori B Hashimoto. Likelihood-based diffusion language models. *Advances*
in Neural Information Processing Systems, 36, 2024.
- 474
475 Zhengfu He, Tianxiang Sun, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. Diffusion-
476 bert: Improving generative masked language models with diffusion models. *arXiv preprint*
arXiv:2211.15029, 2022.
- 477
478 Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in*
479 *neural information processing systems*, 33:6840–6851, 2020.
- 480
481 Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In
482 *International Conference on Learning Representations*, 2017. URL <https://openreview.net/forum?id=rkE3y85ee>.
- 483
484 Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-
485 based generative models. *Advances in neural information processing systems*, 35:26565–26577,
2022.

- 486 Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances*
487 *in neural information processing systems*, 34:21696–21707, 2021.
- 488
- 489 Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm
490 improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:
491 4328–4343, 2022.
- 492 Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion language modeling by estimating
493 the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*, 2023.
- 494
- 495 Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved
496 sampling speed, 2021. URL <https://arxiv.org/abs/2101.02388>.
- 497
- 498 Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous
499 relaxation of discrete random variables. In *International Conference on Learning Representations*,
500 2017. URL <https://openreview.net/forum?id=SljE5L5gl>.
- 501 Mitch Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus
502 of english: The penn treebank. *Computational linguistics*, 19(2):313–330, 1993.
- 503
- 504 Gonzalo Mena, David Belanger, Scott Linderman, and Jasper Snoek. Learning latent permutations
505 with gumbel-sinkhorn networks. In *International Conference on Learning Representations*, 2018.
506 URL <https://openreview.net/forum?id=Byt3oJ-0W>.
- 507 Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture
508 models, 2016.
- 509
- 510 Hunter Nisonoff, Junhao Xiong, Stephan Allenspach, and Jennifer Listgarten. Unlocking guidance
511 for discrete state-space diffusion and flow models. *arXiv preprint arXiv:2406.01572*, 2024.
- 512
- 513 Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li.
514 Your absorbing discrete diffusion secretly models the conditional distributions of clean data, 2024.
515 URL <https://arxiv.org/abs/2406.03736>.
- 516 Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi,
517 Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernandez. The LAMBADA dataset:
518 Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting*
519 *of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1525–1534,
520 Berlin, Germany, August 2016. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/P16-1144>.
- 521
- 522 William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of*
523 *the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.
- 524
- 525 Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language
526 models are unsupervised multitask learners. 2019.
- 527
- 528 Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi
529 Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text
530 transformer. *Journal of machine learning research*, 21(140):1–67, 2020.
- 531
- 532 Subham Sekhar Sahoo, Marianne Arriola, Aaron Gokaslan, Edgar Mariano Marroquin, Alexander M
533 Rush, Yair Schiff, Justin T Chiu, and Volodymyr Kuleshov. Simple and effective masked diffusion
534 language models. In *The Thirty-eighth Annual Conference on Neural Information Processing*
Systems, 2024a. URL <https://openreview.net/forum?id=L4uaAR4ArM>.
- 535
- 536 Subham Sekhar Sahoo, Aaron Gokaslan, Christopher De Sa, and Volodymyr Kuleshov. Diffusion
537 models with learned adaptive noise. In *The Thirty-eighth Annual Conference on Neural Information*
Processing Systems, 2024b. URL <https://openreview.net/forum?id=loMa99A4p8>.
- 538
- 539 Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models, 2022.
URL <https://arxiv.org/abs/2202.00512>.

- 540 Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov.
541 Caduceus: Bi-directional equivariant long-range dna sequence modeling. *arXiv preprint*
542 *arXiv:2403.03234*, 2024a.
- 543 Yair Schiff, Subham Sekhar Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dalla-torre,
544 Bernardo P de Almeida, Alexander Rush, Thomas Pierrot, and Volodymyr Kuleshov. Simple
545 guidance mechanisms for discrete diffusion models. *arXiv preprint arXiv:2412.10193*, 2024b.
546
- 547 Yair Schiff, Subham Sekhar Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dalla-torre,
548 Bernardo P. de Almeida, Alexander Rush, Thomas Pierrot, and Volodymyr Kuleshov. Simple
549 guidance mechanisms for discrete diffusion models, 2024c. URL [https://arxiv.org/abs/
550 2412.10193](https://arxiv.org/abs/2412.10193).
- 551 Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K. Titsias. Simplified and
552 generalized masked diffusion for discrete data, 2025. URL [https://arxiv.org/abs/
553 2406.04329](https://arxiv.org/abs/2406.04329).
- 554 Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised
555 learning using nonequilibrium thermodynamics. In *International conference on machine learning*,
556 pp. 2256–2265. PMLR, 2015.
- 558 Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Inter-
559 national Conference on Learning Representations*, 2021. URL [https://openreview.net/
560 forum?id=St1gjarCHLP](https://openreview.net/forum?id=St1gjarCHLP).
- 561 Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models, 2023. URL
562 <https://arxiv.org/abs/2310.14189>.
- 564 Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben
565 Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint*
566 *arXiv:2011.13456*, 2020.
- 567 Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models, 2023. URL
568 <https://arxiv.org/abs/2303.01469>.
- 569
- 570 Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced
571 transformer with rotary position embedding, 2023.
- 572 Yi Tay, Mostafa Dehghani, Vamsi Aribandi, Jai Gupta, Philip M Pham, Zhen Qin, Dara Bahri,
573 Da-Cheng Juan, and Donald Metzler. Omninet: Omnidirectional representations from transformers.
574 In *International Conference on Machine Learning*, pp. 10193–10202. PMLR, 2021.
- 575
- 576 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz
577 Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing*
578 *systems*, 30, 2017.
- 579 Alex Wang and Kyunghyun Cho. Bert has a mouth, and it must speak: Bert as a markov random field
580 language model. *arXiv preprint arXiv:1902.04094*, 2019.
- 581
- 582 Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T Freeman,
583 and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, 2024.
- 584 Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text
585 classification. In *NIPS*, 2015.
- 586
- 587 Kaiwen Zheng, Cheng Lu, Jianfei Chen, and Jun Zhu. Improved techniques for maximum likelihood
588 estimation for diffusion odes. In *International Conference on Machine Learning*, pp. 42363–42389.
589 PMLR, 2023.
- 590 Kaiwen Zheng, Yongxin Chen, Hanzi Mao, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Masked
591 diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical
592 sampling. *arXiv preprint arXiv:2409.02908*, 2024.
- 593

594	CONTENTS	
595		
596	1 Introduction	1
597		
598	2 Background	2
599		
600	2.1 Discrete Diffusion Models	2
601		
602	2.2 Gaussian Diffusion Models	3
603		
604	2.3 Consistency Distillation	3
605		
606	3 Diffusion Duality	3
607		
608	4 Applications	4
609		
610	4.1 Faster Training using curriculum learning	4
611		
612	4.2 Dual Consistency Distillation	5
613		
614	5 Experiments	6
615		
616	5.1 Improved Training	6
617		
618	5.2 Improved sampling	7
619		
620	6 Related Work and Conclusion	8
621		
622	Appendices	13
623		
624	Appendix A ELBO comparision: Continuous vs Discrete space	13
625		
626	Appendix B Equivalence Derivation	13
627		
628	Appendix C ELBO Equivalence	14
629		
630	Appendix D Training details	14
631		
632	D.1 Denoising model	14
633		
634	Appendix E Curriculum Learning	15
635		
636	E.1 Clipping diffusion time	15
637		
638	Appendix F Optimal DDIM trajectories	15
639		
640	Appendix G Additional Experiments	17
641		
642	G.1 LM1B	17
643		
644	G.2 Tau ablations	17
645		
646	G.3 Sample Quality	18
647		
	G.4 Generative perplexity and Entropy of DUO and SDTT	19
	G.5 Samples	19

Appendices

APPENDIX A ELBO COMPARISON: CONTINUOUS VS DISCRETE SPACE

Let \mathbf{w}_t be a Gaussian random variable, and \mathbf{z}_t be a categorical random variable obtained from (6) and (1) respectively. Let $q(\mathbf{w}_t, \mathbf{z}_t)$ denote their joint distribution for the forward process and $p(\mathbf{w}_t, \mathbf{z}_t)$ denote the joint distribution in the reverse process. Then,

$$\begin{aligned}
 & D_{\text{KL}}(q(\mathbf{w}_t, \mathbf{z}_t) \| p(\mathbf{w}_t, \mathbf{z}_t)) \\
 &= \sum_{\mathbf{z}_t} \int_{\mathbf{w}_t} q(\mathbf{w}_t, \mathbf{z}_t) \log \frac{q(\mathbf{w}_t, \mathbf{z}_t)}{p(\mathbf{w}_t, \mathbf{z}_t)} d\mathbf{w}_t \\
 &= \sum_{\mathbf{z}_t} \int_{\mathbf{w}_t} q(\mathbf{w}_t, \mathbf{z}_t) \log \frac{q(\mathbf{w}_t, \mathbf{z}_t)}{p(\mathbf{w}_t, \mathbf{z}_t)} d\mathbf{w}_t \\
 &= \sum_{\mathbf{z}_t} \int_{\mathbf{w}_t} q(\mathbf{w}_t | \mathbf{z}_t) q(\mathbf{z}_t) \log \frac{q(\mathbf{w}_t | \mathbf{z}_t) q(\mathbf{z}_t)}{p(\mathbf{w}_t | \mathbf{z}_t) p(\mathbf{z}_t)} d\mathbf{w}_t \\
 &= \mathbb{E}_{\mathbf{z}_t} D_{\text{KL}}(q(\mathbf{w}_t | \mathbf{z}_t) \| p(\mathbf{w}_t | \mathbf{z}_t)) + D_{\text{KL}}(q(\mathbf{z}_t) \| p(\mathbf{z}_t))
 \end{aligned} \tag{16}$$

Also,

$$\begin{aligned}
 & D_{\text{KL}}(q(\mathbf{w}_t, \mathbf{z}_t) \| p(\mathbf{w}_t, \mathbf{z}_t)) \\
 &= \int_{\mathbf{w}_t} \sum_{\mathbf{z}_t} q(\mathbf{w}_t, \mathbf{z}_t) \log \frac{q(\mathbf{w}_t, \mathbf{z}_t)}{p(\mathbf{w}_t, \mathbf{z}_t)} d\mathbf{w}_t \\
 &= \int_{\mathbf{w}_t} \sum_{\mathbf{z}_t} q(\mathbf{w}_t) q(\mathbf{z}_t | \mathbf{w}_t) \log \frac{q(\mathbf{w}_t) q(\mathbf{z}_t | \mathbf{w}_t)}{p(\mathbf{w}_t) p(\mathbf{z}_t | \mathbf{w}_t)} d\mathbf{w}_t \\
 &= \mathbb{E}_{\mathbf{w}_t} \underbrace{D_{\text{KL}}(q(\mathbf{z}_t | \mathbf{w}_t) \| p(\mathbf{z}_t | \mathbf{w}_t))}_{=0} + D_{\text{KL}}(q(\mathbf{w}_t) \| p(\mathbf{w}_t)) \\
 &= D_{\text{KL}}(q(\mathbf{w}_t) \| p(\mathbf{w}_t))
 \end{aligned} \tag{17}$$

From (16) and (17) we get

$$\underbrace{-D_{\text{KL}}(q(\mathbf{w}_t) \| p(\mathbf{w}_t))}_{\equiv (7)} = \underbrace{\mathbb{E}_{\mathbf{z}_t} - \left[D_{\text{KL}}(q(\mathbf{w}_t | \mathbf{z}_t) \| p(\mathbf{w}_t | \mathbf{z}_t)) \right]}_{<0} + \underbrace{(-D_{\text{KL}}(q(\mathbf{z}_t) \| p(\mathbf{z}_t)))}_{\equiv (4)} \tag{18}$$

$$\implies \text{ELBO}(\tilde{q}, p) > \text{ELBO}(q, p) \tag{19}$$

Thus, ELBO in the Gaussian space is “looser” or lower than the ELBO in the discrete space. This proof is inspired by [Mena et al. \(2018\)](#).

APPENDIX B EQUIVALENCE DERIVATION

Let $\mathbf{x} \in \mathcal{V}$ s.t. $\mathbf{x}_k = 1$ i.e., \mathbf{x} contains 1 at the k^{th} index. Consider a random variable $\mathbf{y} = \alpha_t \mathbf{x} + \sigma \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}_K)$ and $\sigma = \sqrt{1 - \alpha^2}$. Let the random variable $Z = \max\{\mathbf{y}_i : i \neq k\}$. Let $P(\arg \max(\mathbf{y}) = k)$ denote the probability that index of the maximum entry in \mathbf{y} is k .

Next, let’s compute $P(Z < l)$ for $l \in \mathbb{R}$.

$$P(Z < l) = \prod_{i \neq k} P(\mathbf{y}_i < l) = \Phi^{K-1} \left(\frac{l}{\sigma} \right) \quad \because \mathbf{y}_i \sim \mathcal{N}(0, \sigma^2) \tag{20}$$

where $\Phi(z) = \int_{-\infty}^z \exp(-t^2/2) dz / \sqrt{2\pi}$ is the cumulative distribution function of the normal distribution.

$$P(\arg \max(\mathbf{y}) = k) = P(Z < \mathbf{y}_k)$$

$$\begin{aligned}
&= \int_{-\infty}^{\infty} P(\mathbf{y}_k = l)P(Z < l)dl \\
&= \int_{-\infty}^{\infty} \phi\left(\frac{l-\alpha}{\sigma}\right)P(Z < l)dl \\
&= \int_{-\infty}^{\infty} \phi\left(\frac{l-\alpha}{\sigma}\right)\Phi^{K-1}\left(\frac{l}{\sigma}\right)dl && \text{From (20)} \\
&= \int_{-\infty}^{\infty} \phi\left(\tilde{l}-\frac{\alpha}{\sigma}\right)\Phi^{K-1}(\tilde{l})d\tilde{l} && \text{Substituting } \tilde{l} = l/\sigma \\
&= \int_{-\infty}^{\infty} \phi\left(\tilde{l}-\frac{\alpha}{\sqrt{1-\alpha^2}}\right)\Phi^{K-1}(\tilde{l})d\tilde{l} \tag{21}
\end{aligned}$$

where $\phi(z) = \exp(-z^2)/\sqrt{2\pi}$ is the standard Normal distribution.

In the context of diffusion, given $\mathbf{x} \sim p_{\text{data}}$, let's define a discrete diffusion process on \mathbf{x} s.t. the marginal of the intermediate latent \mathbf{z}_t is given by (1). From (1),

$$P(\mathbf{z}_t = \mathbf{x}) = \alpha_t + \frac{1-\alpha_t}{K} \tag{22}$$

Now, let's define a separate Gaussian diffusion on \mathbf{x} . The marginal of the intermediate latent \mathbf{w}_t is given by 6. From (21), we have $P(\arg \max(\mathbf{w}_t) = \mathbf{x}) = \int_{-\infty}^{\infty} \phi\left(\tilde{l}-\frac{\alpha}{\sqrt{1-\alpha^2}}\right)\Phi^{K-1}(\tilde{l})d\tilde{l}$. Since, \mathbf{z}_t is the discrete equivalent of \mathbf{w}_t , from (22) and (21) we get,

$$\begin{aligned}
\alpha_t + \frac{1-\alpha_t}{K} &= \int_{-\infty}^{\infty} \phi\left(\tilde{l}-\frac{\alpha}{\sqrt{1-\alpha^2}}\right)\Phi^{K-1}(\tilde{l})d\tilde{l} \\
\implies \alpha_t &= \frac{K}{K-1} \left[\int_{-\infty}^{\infty} \phi\left(\tilde{l}-\frac{\alpha}{\sqrt{1-\alpha^2}}\right)\Phi^{K-1}(\tilde{l})d\tilde{l} - \frac{1}{K} \right] \tag{23}
\end{aligned}$$

APPENDIX C ELBO EQUIVALENCE

Let $\mathbf{w}_t \sim \tilde{q}_t(\cdot|\mathbf{x}) = \mathcal{N}(\tilde{\alpha}_t, (1-\tilde{\alpha}_t^2)\mathbf{I}_K)$, where $\tilde{\alpha}_t = g(t)$, and $g: [0, 1] \rightarrow [0, 1]$ is a monotonically decreasing function satisfying $g(0) = 1$ and $g(1) = 0$. Define $\tilde{\mathbf{z}}_t = \arg \max(\mathbf{w}_t)$. From (10), it follows that $\tilde{\mathbf{z}}_t \sim (q_t)'$, where

$$(q_t)' = \text{Cat}(\cdot|\mathcal{T}(\tilde{\alpha}_t)\mathbf{x} + (1-\mathcal{T}(\tilde{\alpha}_t))\mathbf{1}/K).$$

Let $(\alpha_t)'$ be the discrete diffusion parameter for this process, given by $(\alpha_t)' = \mathcal{T}(\tilde{\alpha}_t) = \mathcal{T}(g(t))$. The ELBO for this uniform-state discrete diffusion process remains the same as that for any other uniform-state diffusion, regardless of the noise schedule α_t . This invariance follows from (4), which does not depend on the functional form of α_t (Schiff et al., 2024c). Notably, this property also holds for absorbing state diffusion (Sahoo et al., 2024a) and Gaussian diffusion (Kingma et al., 2021).

Thus, without loss of generality,

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{\mathbf{x}, t \sim \mathcal{U}[0, 1], \mathbf{z}_t \sim q_t(\cdot|\mathbf{x})} f(\mathbf{x}, \mathbf{z}_t, \alpha_t, \mathbf{x}_\theta(\mathbf{z}_t)) \tag{24}$$

$$= \mathbb{E}_{\mathbf{x}, t \sim \mathcal{U}[0, 1], \mathbf{z}_t \sim (q_t)'(\cdot|\mathbf{x})} f(\mathbf{x}, \mathbf{z}_t, \alpha_t = \mathcal{T}(g(t)), \mathbf{x}_\theta(\mathbf{z}_t)) \tag{25}$$

$$= \mathbb{E}_{\mathbf{x}, t \sim \mathcal{U}[0, 1], \mathbf{w}_t \sim \tilde{q}_t(\cdot|\mathbf{x})} f(\mathbf{x}, \mathbf{z}_t = \arg \max(\mathbf{w}_t), \alpha_t = \mathcal{T}(g(t)), \mathbf{x}_\theta(\mathbf{z}_t = \arg \max(\mathbf{w}_t))) \tag{26}$$

The equality between (25) and (26) holds because, when $\mathbf{w}_t \sim \tilde{q}_t$ in (26), $\arg \max(\mathbf{w}_t)$ follows the distribution $(q_t)'$, as initially established. Consequently, the discrete vector $\arg \max(\mathbf{w}_t)$ in (26) and the discrete vector \mathbf{z}_t in (25) share the same distribution $(q_t)'$.

As a sanity check, we evaluated UDLM (Schiff et al., 2024c) trained on LM1B (Table 3) with the ELBO (24) with a

APPENDIX D TRAINING DETAILS

D.1 DENOISING MODEL

Unlike prior discrete diffusion approaches, we design the denoising model $p_t^\theta: \Delta^K \cup \mathcal{V} \rightarrow \Delta^K$ to handle both a continuous latent $\tilde{\mathbf{w}} \in \Delta^K$ and a discrete latent $\mathbf{z} \in \mathcal{V}$. We implement p_t^θ as

756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809

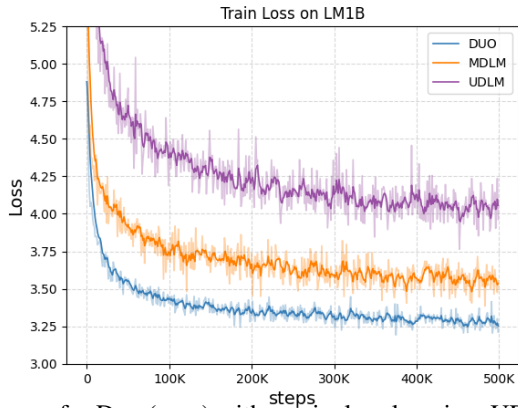


Figure 3: Training loss curves for Duo (ours) with curriculum learning, UDLM, and MDLM. We see observe that curriculum learning leads to low variance training.

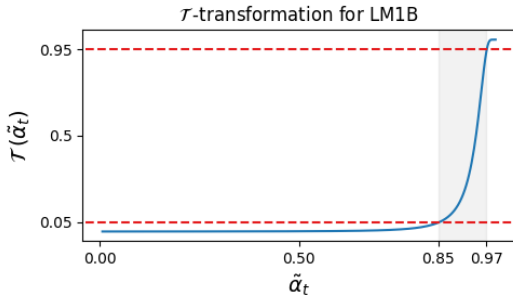


Figure 4: Caption

a transformer (Vaswani et al., 2017), with the token-to-embedding mapping defined by matrix multiplication $\mathbf{y}^\top \text{vocab_embeddings}$ in the first layer, where $\text{vocab_embeddings} \in \mathbb{R}^{K \times m}$ and m is the dimensionality of the vocabulary embeddings. Discrete inputs $\mathbf{y} \in \mathcal{V}$ correspond to standard embedding lookups, while continuous inputs $\mathbf{y} \in \Delta^K$ act as “soft lookups.”

APPENDIX E CURRICULUM LEARNING

E.1 CLIPPING DIFFUSION TIME

In our approach, $\alpha_t = \mathcal{T}(\tilde{\alpha}_t)$ is derived from the Gaussian diffusion parameter $\tilde{\alpha}_t$ as shown in Fig. 4. It’s important to note that the diffusion ELBO $\mathcal{L}_{\text{diffusion}}$ is weighted by α_t' in (5), so when $\alpha_t' \approx 0$, the contribution of the diffusion time step t to the ELBO is negligible and hence provides little learning signal. Prior work (Sahoo et al., 2024a; Lou et al., 2023) used a linear schedule for α_t and did not face this issue. Hence, while training on Gaussian latents, we restrict the training window $t \sim [t_{\min}, t_{\max}]$ to exclude the region where $\alpha_t' \approx 0$. Although this yields a slightly biased estimate of the ELBO, it effectively reduces training variance. In Fig. 4, we observe that for $t \in [t_{\min}, t_{\max}]$, the Gaussian latent has a higher signal level compared to its discrete counterpart, making it easier for the denoising model to recover the clean signal from the Gaussian latent. Consequently, the task of denoising is easier for the denoising model for $\tau > 0$ than in the limiting case $\lim \tau \rightarrow 0^+$ which corresponds to the discrete setting. Consequently, the loss term

As discussed earlier, we set the diffusion time range such that $\alpha_t = \mathcal{T}(\tilde{\alpha}_t) \in [0.05, 0.95]$ in the discrete diffusion process. While this range depends on vocabulary size, we found it to be similar for both the gpt-2 and bert-base-uncased tokenizers, with $[t_{\min}, t_{\max}] = [0.03, 0.15]$.

APPENDIX F OPTIMAL DDIM TRAJECTORIES

Algorithm 1 Dual Consistency Distillation (DCD)

```

810 Input: data  $\mathbf{x} \sim q_{\text{data}}$ , learning rate  $\eta$ , number of distillation rounds  $K$ , number of training iterations
811 per round  $M$ , ema  $\mu$ , discretization step  $\Delta$ .
812 for  $i = 1$  to  $K$  do
813    $\theta \leftarrow \text{stopgrad}(\theta^-)$ 
814   for  $i = 1$  to  $M$  do
815     Sample  $\mathbf{x} \sim q_{\text{data}}$ ,  $t \sim \mathcal{U}[0, 1]$ , and  $\epsilon \sim \mathcal{N}(0, I_K)$ .
816      $s \leftarrow \max(t - i \cdot \Delta, 0)$ 
817      $\mathbf{z}_s \leftarrow \arg \max(\tilde{\alpha}_s \mathbf{x} + \sqrt{1 - \tilde{\alpha}_s^2} \epsilon)$ 
818      $\mathbf{z}_t \leftarrow \arg \max(\tilde{\alpha}_t \mathbf{x} + \sqrt{1 - \tilde{\alpha}_t^2} \epsilon)$ 
819      $\mathcal{L}_{\text{DCD}}(\theta, \theta^-) \leftarrow \text{D}_{\text{KL}}(\mathbf{x}_\theta(\mathbf{z}_t, t), \mathbf{x}_{\theta^-}(\mathbf{z}_s, s))$ 
820      $\theta \leftarrow \theta - \eta \nabla_{\theta} \mathcal{L}_{\text{DCD}}(\theta, \theta^-)$ 
821      $\theta^- \leftarrow \text{stopgrad}(\mu \theta^- + (1 - \mu) \theta)$ 
822   end for
823 end for

```

Table 3: Test perplexities (PPL; \downarrow) on LM1B. *Reported in He et al. (2022). Best uniform/Gaussian diffusion value is bolded. [¶]Denotes the dataset didn’t incorporate sentence packing. For diffusion models, we report the bound on the likelihood. Best diffusion value is underlined. [‡]Denotes retrained models.

	PPL (\downarrow)
<i>Autoregressive</i>	
Transformer-X Base (Dai et al., 2019)	23.5
OmniNet _T (Tay et al., 2021)	21.5
Transformer [‡]	22.83
<i>Diffusion (absorbing state)</i>	
BERT-Mouth* (Wang & Cho, 2019)	142.89
D3PM Absorb (Austin et al., 2021)	76.90
D3PM Uniform (Austin et al., 2021)	137.90
DiffusionBert [¶] (He et al., 2022)	63.78
SEDD Absorb [‡] (Lou et al., 2023)	32.71
MDLM [‡] (Sahoo et al., 2024a)	<u>32.03</u>
<i>Diffusion (uniform state / Gaussian)</i>	
Diffusion-LM ^{¶*} (Li et al., 2022)	118.62
SEDD Uniform [¶] (Lou et al., 2023)	40.25
UDLM [‡] (Schiff et al., 2024b)	36.71
Duo (Ours)	33.68

For the forward diffusion process 6 and a denoising model $\mathbf{x}_\theta : \mathbb{R}^K \rightarrow \Delta$, the DDIM (Song et al., 2021) update step is given as where

$$\mathbf{z}_s = \tilde{\alpha}_s \mathbf{x}_\theta(\mathbf{z}_t) + \sqrt{1 - \tilde{\alpha}_s^2} \epsilon_\theta(\mathbf{z}_t) \quad (27)$$

where $\epsilon_\theta(\mathbf{z}_t) = (\mathbf{z}_t - \alpha_t \mathbf{x}_\theta(\mathbf{z}_t)) / \sqrt{1 - \alpha_t^2}$.

For an optimal denoiser, we assume $\mathbf{x}_\theta(\mathbf{z}_t) = \mathbf{x}$. Given $\mathbf{z}_{t=1} = \tilde{\epsilon} \sim \mathcal{N}(0, \mathbf{I}_K)$ and $\mathbf{x} \sim q_{\text{data}}$, it can be easily seen that (27) reduces to $\mathbf{z}_s = \tilde{\alpha}_s \mathbf{x} + \sqrt{1 - \tilde{\alpha}_s^2} \tilde{\epsilon}$. This holds $\forall s \in [0, 1]$. Hence the optimal DDIM trajectory $\mathcal{P}_{\text{DDT}}(\mathbf{x}, \epsilon)$ is given as $\mathcal{P}_{\text{DDIM}}(\mathbf{x}, \epsilon) = \{\tilde{\alpha}_t \mathbf{x} + \sqrt{1 - \tilde{\alpha}_t^2} \tilde{\epsilon}\}_{t \in [0, 1]}$

APPENDIX G ADDITIONAL EXPERIMENTS

G.1 LM1B

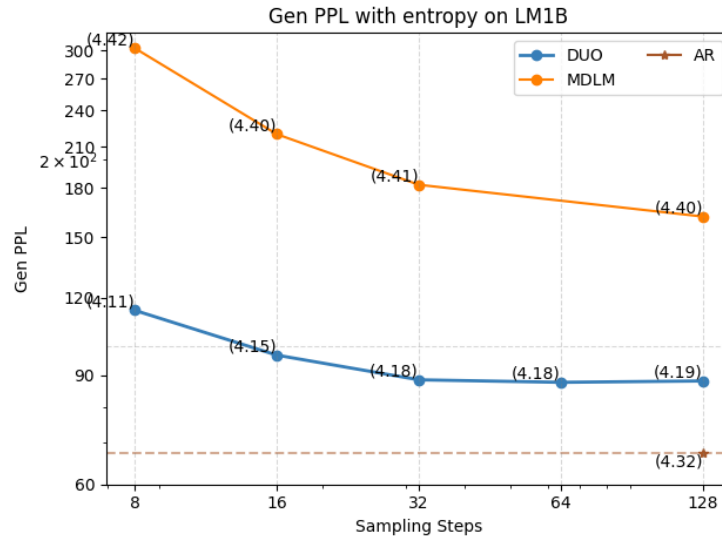
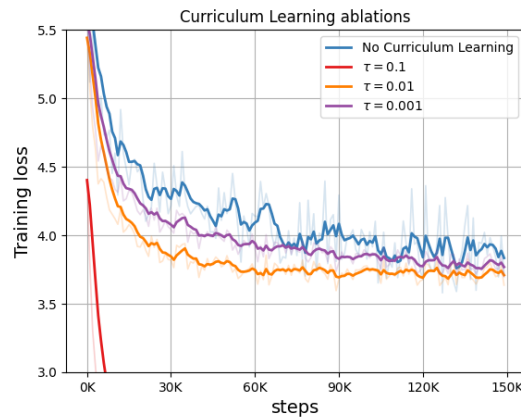


Figure 5: Gen PPL for Duo and MDLM.

G.2 TAU ABLATIONS

Figure 6: Train loss for curriculum learning by varying τ . Models were trained on the LM1B dataset.

G.3 SAMPLE QUALITY

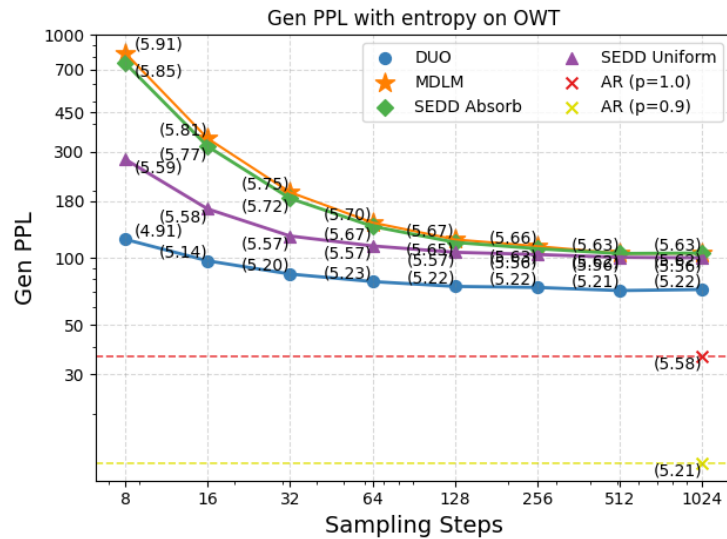


Figure 7: Sample quality comparison between Duo (ours), MDLM, SEDD (Absorb / Uniform), and AR. The numbers in brackets denote the entropy of the samples. Note that the entropy of samples for DUO is quite similar to that of the AR model with nucleus sampling ($p=0.9$) while the entropy of the samples for other methods is similar to the AR model without nucleus sampling.

G.4 GENERATIVE PERPLEXITY AND ENTROPY OF DUO AND SDTT

Table 4: Generative perplexity and entropy for Duo distilled using Dual Distillation (DD) (1) and MDLM distilled SDTT.

	MDLM w/ SDTT		DUO w/ DD	
	GenPPL	entropy	GenPPL	entropy
<i>Base Model</i>				
1024	104.85	5.63	72.05	5.22
512	104.43	5.63	71.24	5.21
256	112.70	5.66	73.59	5.22
128	120.77	5.67	74.37	5.22
64	143.88	5.70	78.19	5.23
32	196.79	5.75	84.52	5.20
16	343.33	5.81	96.89	5.14
8	830.82	5.91	121.02	4.91
<i>Round 1</i>				
1024	79.12	5.59	63.85	5.26
512	79.40	5.59	61.69	5.26
256	84.28	5.61	63.54	5.25
128	89.97	5.62	64.01	5.26
64	105.90	5.65	68.46	5.28
32	141.78	5.69	72.65	5.26
16	249.15	5.76	84.35	5.21
8	618.15	5.85	108.88	5.02
<i>Round 2</i>				
1024	61.75	5.53	55.03	5.27
512	62.52	5.53	54.91	5.27
256	66.80	5.56	56.20	5.29
128	70.52	5.57	57.76	5.28
64	82.51	5.60	59.95	5.30
32	107.93	5.65	65.35	5.30
16	183.41	5.71	76.00	5.25
8	458.83	5.80	100.61	5.10
<i>Round 3</i>				
1024	49.53	5.48	49.89	5.27
512	50.42	5.49	50.99	5.28
256	52.96	5.50	51.28	5.28
128	56.70	5.52	52.55	5.30
64	65.02	5.55	55.92	5.32
32	83.85	5.59	60.30	5.32
16	135.75	5.64	68.87	5.28
8	323.56	5.71	91.56	5.15
<i>Round 4</i>				
1024	42.53	5.44	46.44	5.27
512	43.61	5.44	47.06	5.27
256	45.27	5.46	46.98	5.28
128	49.14	5.48	48.57	5.32
64	55.72	5.50	50.60	5.33
32	70.82	5.54	54.61	5.35
16	111.40	5.59	63.59	5.32
8	253.59	5.65	84.23	5.22
<i>Round 5</i>				
1024	36.89	5.39	42.46	5.25
512	37.16	5.40	44.05	5.25
256	38.65	5.41	44.73	5.28
128	41.98	5.43	45.69	5.31
64	47.04	5.45	47.87	5.34
32	58.29	5.49	51.74	5.36
16	89.17	5.53	59.83	5.34
8	193.05	5.58	79.24	5.25

G.5 SAMPLES

Samples from a distilled Duo.

G.5.1 T = 1024

</endoftext|> the funds to give them by April.
n

1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079

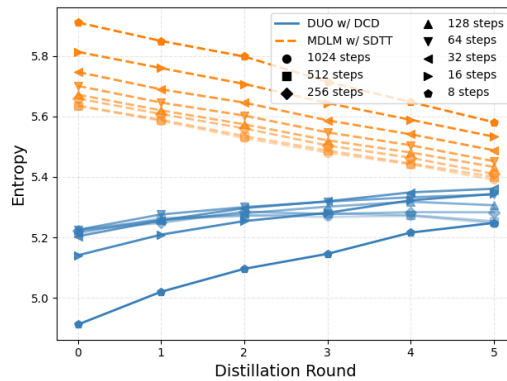


Figure 8: Entropy of MDLM distilled using SDTT, and of DUO distilled using CDC. The entropy of the SDTT-distilled MDLM decreases with distillation, while the entropy of the CDC-distilled DUO model increases. The curves corresponding to a higher number of sampling steps are displayed with lighter colors to emphasize the low sampling step regimes.

nIt will will take the community 30 days to be ready for a camp, on Dec. 2, and the out of the group’s headquarters.

n

nThe community is coming out of a fairly cautious, old and un-vigilant approach not coming too months in advance. He is willing to create a new opportunity for the community to hold meetings in share a meeting in person.

n

nIt is great to be a community, but it is important to be encouraged by the community to find the best leaders, he said in the letter.

n

nParqua City Responders also deploy a first-class command and response vehicle, leaving the Centennial Township Police Department in central Pennsylvania with the equipment it will provide.

n

nOn Friday, Diencio in Congress approved federal funds to support operations in Washington, U.S., and he will use the funds to help global relief efforts and climate recovery efforts.

n

nFIRST REQUITMENT

n

nForbes, published last year information tech companies, review the costs of many class claims, such as mobile insurance and mortgage claims, based on individual claims, and found that costouts and cost delays could lead to a lower claim.

n

n

nTech giants Alphabet (GOOG) and Facebook (FB.O) introduced a bill earlier this week to increase more court requirements for such claims between public and private companies in the United States.

n

nUniversity of Minnesota (UMN) students gather at Metcalf University in Minneapolis, Minnesota in this March 18, file photo. REUTERS/Kathy J. Toner

n

nAdditional reporting by Rosalee Warrington<|endoftext|>NEW YORK (Reuters) - Public-interest group PPC Capital has agreed to re-

1080 duce royalty payments for the residential customers owned by the
1081 well to 10 percent in 2012, according to a Los Angeles-based trade
1082 union source familiar with the federal antitrust talks.
1083 n
1084 nThe new group, 201cVibrant Public Electric201d (PGP), would,
1085 still still include some customers in the Pacific Parcel, other
1086 parts of California owned by the, and a part-owner and owner of of
1087 intellectual utilities.
1088 n
1089 nThe prospect of a reorganization, PPC and the company at least \$1
1090 billion over the option of operating without it. The 201cParties
1091 Reduction Review201d agreed Tuesday with state regulators to put a
1092 stop to the utility from operating this year.
1093 n
1094 nFILE - In this this June 24, 2013 file photo, PG&T Chairman
1095 Jerry Hayward speaks in front of Reuters Reuters reporters dur-
1096 ing a deadline to announce proposed lawsuits with federal regula-
1097 tors. WSDA officials met several times about the possibility of
1098 an agreement, according to the sources.
1099 n
1100 n201cWe will the outstanding issues including water and servicing
1101 payments, residential rates, the property service fees, customer
1102 services, and debt issues,201d said John McAlpine, the CPUC agency
1103 spokesman.
1104 n
1105 nPG&T faces a \$2 billion operating loss, and despite the deal,
1106 PG&T faces uncertainty on a combination either paying the utility
1107 a reduction in operating revenue and losing back to shareholders
1108 and a a partial reorganization.
1109 n
1110 nAt the same time, the prospects also look grim for an industry
1111 enjoying explosive growth this year, with growing demand more
1112 sources of electricity and with more federal government competi-
1113 tion.
1114 n
1115 nThe move also put pressure on General Electric, the maker Wall
1116 Edison and Reliant,, cut down power output by 20 percent in 2011.
1117 n
1118 n201cBASED CUMMARY201d
1119 n
1120 nPG&T2019s shares fell \$1.50 per share, after a loss of \$2.6 bil-
1121 lion in the fourth quarter ended Q3.
1122 n
1123 nThe company in the price-per-share market have seen higher compe-
1124 tition, higher operating costs and a cost per diluted rate that it
1125 the the rates offered by PPC Securities.
1126 n
1127 n201cWith revenue customers, shareholders, such higher increased
1128 rates, and better customer service, demand a higher diluted
1129 share price,201d said David Smith, president of communications
1130 in PG&T2019s public relations statement Tuesday.
1131 n
1132 nTheT filed suit in November against a \$2 proposed combined util-
1133 ity company and put the suit on Monday with the Federal Trade Com-
mission.

1134 cording to a Wall Street Journal report earlier this month that
 1135 said.
 1136 n
 1137 nAmong other questions about utilities’ potentially toxic as-
 1138 sets, the timing of dividend payments will be under
 1139 scrutiny and the way due distribution of payments to its work-
 1140 ers and retirees will be affected by Congress under an overhaul
 1141 passed Section 18 U.S.C. 111.
 1142 n
 1143 n[Bottom: of Reforms]
 1144 n
 1145 nPG&T owns or percent of the company, make up about 40 percent of
 1146 the, <|endoftext|>

1147 G.5.2 T = 8

1149 <|endoftext|> at the time, also critical of that the authors used
 1150 the site promote Mideastism, the to ’coexist.’ He blamed the
 1151 Saudi authorities, for carrying out ’ca Censorship Convention
 1152 on the website, which being known for censorship on this [[.’
 1153 ’The site’ are not officially known but, according the gov-
 1154 ernment officials, have provoked international criticism,, in a
 1155 number of some criticism of the university’s activities the
 1156 members of the South Asia of the Gulf Cooperation Council, said
 1157 that the university had no legal right to to restrict access to
 1158 students.
 1159 n
 1159 nAh Jafar, No. correspondent at The Guardian, has noted that the
 1160 attacks further proof that Yemen is a country and can consider any
 1161 plot, or a potential terrorist plot of serious consequences, to
 1162 its own people. and, indeed, though this opinion article has se-
 1163 vere outrage in some European countries. As Ash Khan, also not
 1164 the removed by the non-The Guardian Times for this piece, a piece
 1165 purports to be in solidarity with the terrorists’s massacre
 1166 of satirical magazine Charlie Hebdo because it features a cari-
 1167 cature of a woman in a, high-color, according Mr the Times’s
 1168 investigative,, conctions.
 1169 n
 1169 nMs. She describes, if the piece of Charlie Hebdo that pro-
 1170 vokes readers’s outrage, of her pet children as ’bruits of
 1171 pedophiles.’ But these things are not acceptable in France.
 1172 In fact, not in the context of Orthodox Jews (and I don’t’t get
 1173 a muchness at the thought of dealing to the Lat’mod and) Stamp-
 1174 sia), is know as much as I do about the film’s final act). But
 1175 The Guardian doesn’t need a correction, I guess, the page I put it
 1176 on, for condemning a cartoon and depicts her a terrorist. And she
 1177 has human speech.<|endoftext|>The Hunger Games has already been
 1178 criticized the sexism among many ways, and-g andes’s following the
 1179 example of actor Akji Aoki, a wting comment the Oct. 18, 2008 is-
 1180 sue of The Hollywood Reporter in which he complained about video
 1181 games led to negative discussion of gender gender people on women,
 1182 bizarrely contending to say that games are social games: There
 1183 very many games that played, I would play them, but they were not
 1184 very good.Äs he said, religion in games only religious ideas of
 1185 games, and then video games came out, game was declined. While
 1186 the existence of games is,, but quite disturbing because especially
 1187 in a democratic culture shows how the extent to which it uses in
 its power to maintain the respect of freedom of thought. Without
 that process, the concepts of thought and speech are violated.

1188 n
1189 nTurn investigation of the down issues is problematic not only be-
1190 cause it are so able to be honest and honest about religious not
1191 not cults are presstitutes in video popular world; it is also the
1192 general tendency of the media to tolerate such comments of suppos-
1193 edly the characters.
1194 n
1195 n[1]. Critic of the following:
1196 n
1197 nŹ2022 Punishment because of Ź01cincidentity an offensive and
1198 equivocal views of a Western people.Ź01d
1199 n
1200 nŹ2022 The tendency of people to get involved in violence. For ex-
1201 ample, Mr. Miri vilified factory father, a a Buddhist follower,
1202 in a speech he is an atheist act for being attacked by a proper
1203 audience, which is, well, that bad.
1204 n
1205 nŹ2022 Violation of orator of speech, about love, and thedom. The
1206 wasler, too, for of the '-do-violence was punished because the
1207 violence was fostered by close support the victim had. Just of
1208 this, that is, the diatoms, lazy peopleand dystopism.
1209 n
1210 nThe The Book Translation is a collection of translations of En-
1211 glish published on this author here. The Images are taken from
1212 the translation.<|endoftext|>Timeing Reassignment, Love and Free-
1213 dom is a Ph.D. film. U. began to in the Filipinos in 2001 and had
1214 crosstexts in a film. It brought back out in 2008 and added back
1215 back in 2010.
1216 n
1217 nIn the days, Latin, they scanned the country about 30,, a sec-
1218 ond per minute. In,, the First and Vietnam war, and around the
1219 world during the Second World War War The U U's Center of Bi His-
1220 tory says the footage of the Central American television shows was
1221 short but not as continuous. The result.
1222 n
1223 nŹ thought I was in it like ina mediocrat. When I saw it that
1224 I was working journalist, a situation that I was in, with half
1225 a second of my time off. The person to give a what being out
1226 but no one stood up for me. I would have say, Ź018You are do-
1227 ing your wrong. You are good it.Ź019Ź01d ThatŹ019s an the di-
1228 rector John Kershaw said the first time he was making U.A.A. War.
1229 I.<|endoftext|>,"
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241