THE DIFFUSION DUALITY

Subham Sekhar Sahoo[†] [*] **Justin Deschenaux**[§] **Aaron Gokaslan**[†]

Guanghan Wang[†] **Justin Chiu**[‡] **Volodymyr Kuleshov**[†]

ABSTRACT

Discrete diffusion models have been shown to be surprisingly effective as language models. In this work, we uncover a fundamental property of uniform state diffusion (a specific class of discrete diffusion processes) that it emerges from an underlying Gaussian diffusion process. This insight enables us to transfer techniques from Gaussian diffusion to improve discrete diffusion models. Leveraging this property, we improve both training and sampling efficiency. We introduce a curriculum learning strategy that reduces training variance, leading to **2**× faster convergence, and adapt efficient distillation methods from continuous-state diffusion models to accelerate sampling. As a result, our models surpass autoregressive models in zero-shot perplexity on 3 out of 7 benchmarks while reducing the number of sampling steps by **two orders** of magnitude without compromising 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.

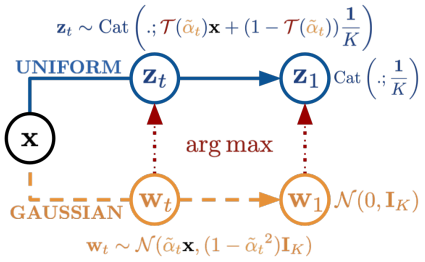

Figure 1: An illustration of uniform state discrete diffusion (top) and the underlying Gaussian diffusion (bottom). Applying $\arg\max$ maps Gaussian latents $\mathbf{w}_t \in \mathbb{R}^n$ to discrete latents $\mathbf{z}_t \in \mathcal{V}$, transforming their marginals from $\tilde{q}_t(.|\mathbf{x}; \tilde{\alpha}_t)$ (6) to $q_t(.|\mathbf{x}; \mathcal{T}(\tilde{\alpha}_t))$ (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).

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., 2025), 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., 2024), discrete diffusion remains largely underexplored.

[*]Correspondence to Subham Sekhar Sahoo: ssahoo@cs.cornell.edu
[†]Cornell Tech, NY, USA. [§]EPFL, Lausanne, Switzerland. [‡] Cohere, NY, USA.

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 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 $\text{Cat}(.; \boldsymbol{\pi})$ by defining a sequence of latents $\mathbf{z}_t \in \mathcal{V}$ whose marginals conditioned on $\mathbf{x}$ at time t is given by:

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

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 $p_t^\theta : \mathcal{V} \times [0,1] \to \Delta^K$ 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):

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

Schiff et al. (2025) demonstrate that the true reverse posterior for USDMs is given as:

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

where $\alpha_{t|s} = \alpha_t/\alpha_s$ and the approximate reverse posterior as $p^\theta_{s|t}(.|\mathbf{z}_t; \alpha_s, \alpha_t) = q_{s|t}(.\mid \mathbf{z}_t, \mathbf{x} = \mathbf{x}_\theta(\mathbf{z}_t, t); \alpha_t, \alpha_s)$. 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 $T \to \infty$, the ELBO (2) for USDMs reduces to (Schiff et al., 2025):

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

where

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

where $\mathbf{x}_i$ denotes the i^{th} index of a vector $\mathbf{x}$, $\bar{\mathbf{x}} = K\alpha_t \mathbf{x} + (1 - \alpha_t)\mathbf{1}$, $\bar{\mathbf{x}}_\theta = K\alpha_t \mathbf{x}_\theta(\mathbf{z}_t, t) + (1 - \alpha_t)\mathbf{1}$, $\alpha_t{}'$ 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 $\mathbf{w}_t \sim \tilde{q}_t(.|\mathbf{x})$, whose marginal distribution is given by:

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

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:

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

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 $\tilde{q}_t(.|\mathbf{x})$ (6) and obtain a less noisy sample $\mathbf{w}_s$ by numerically solving the PF-ODE for one step using the denoising model $\mathbf{x}_\theta$. The consistency model then trains a student model $\mathbf{x}_\theta$ (with parameters $\boldsymbol{\theta}$) to match the teacher model's estimate of the clean sample, given the noisy sample $\mathbf{w}_t$. The teacher model $\mathbf{x}_{\boldsymbol{\theta}^-}$ (with parameters $\boldsymbol{\theta}^-$) provides the less noisy sample $\mathbf{w}_s$, and the student model is optimized to minimize the loss:

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

where $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$ denotes the error between the teacher model's reconstruction $\mathbf{x}_{\boldsymbol{\theta}^-}(\mathbf{w}_t, t)$ and the student model's reconstruction $\mathbf{x}_\theta(\mathbf{w}_t, t)$ of the original sample and $\lambda : [0, 1] \to \mathbb{R}^+$ 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 $\mathbf{w}_t \sim \tilde{q}_t(.|\mathbf{x})$ be an intermediate latent at time t. Next, define the operation $\arg\max : \mathbb{R}^K \to \mathcal{V}$ 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 $\arg\max_{\mathbf{z}\in\mathcal{V}} \mathbf{z}^\top \mathbf{w}$. Let $\mathbf{z}_t = \arg\max(\mathbf{w}_t)$. In Suppl. A, we demonstrate that $\mathbf{z}_t$ is distributed as:

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

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

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

where $\phi(z) = \exp(-z^2)/\sqrt{2\pi}$ is the standard Normal distribution and $\Phi(z) = \int_{-\infty}^{z} \exp(-t^2/2)\mathrm{d}z/\sqrt{2\pi}$ 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:

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

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.3. In brief, we derive the following relationship:

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

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 $\mathbf{z}_t \sim q(.|\mathbf{x})$ for $t \sim [0, 1]$. We aim to re-express this in terms of Gaussian latents $\mathbf{w}_t \sim \tilde{q}_t(.|\mathbf{x})$ such that marginalizing over $\mathbf{w}_t$ yields the same value. In Suppl. B.1, we show

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

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):

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

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 $\tau \to 0$, 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 $p_t^\theta : \Delta^K \cup \mathcal{V} \to \Delta^K$ to handle both continuous latents and discrete latents ; see Suppl. C.1 for more details. We set τ as a function of the training iteration n, annealing it from an initial value $\tau(n = 0) = \tau_{\max}$ to a final value $\tau(n = N) = \tau_{\min} \approx 0$, 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:

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

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** $\lim_{\tau\to 0^+}$. 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 $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}_K)$, 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. B.2 for a detailed discussion. Let $\mathcal{P}_{\text{DDIM}}(\mathbf{x}, \boldsymbol{\epsilon}) = \left\{\tilde{\alpha}_t \mathbf{x} + \sqrt{1-\tilde{\alpha}_t{}^2}\boldsymbol{\epsilon}\right\}_{t\in[0,1]}$ denote such a trajectory. Next, we map this sequence of Gaussian latents to the discrete space using arg max operator in the following manner: $\mathcal{P}_{\text{DDT}}(\mathbf{x}, \boldsymbol{\epsilon}) = \left\{\mathbf{z}_t = \arg\max\left(\tilde{\alpha}_t \mathbf{x} + \sqrt{1-\tilde{\alpha}_t{}^2}\boldsymbol{\epsilon}\right)\right\}_{t\in[0,1]}$, where $\mathcal{P}_{\text{DDT}}$ 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 $\mathbf{x}_{\boldsymbol{\theta}^-}$, our goal is to distill it into a student model $\mathbf{x}_{\boldsymbol{\theta}}$ that can generate samples of similar quality but in fewer steps. To perform distillation, we sample an adjacent pair of latents $(\mathbf{z}_s, \mathbf{z}_t) \sim \{(\mathbf{z}_{j-\Delta}, \mathbf{z}_j) | \mathbf{z}_{\{.\}} \in \mathcal{P}_{\text{DDT}}(\mathbf{x}, \boldsymbol{\epsilon}), j \in [\Delta, 1]\}$ 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 $\mathbf{x}_{\boldsymbol{\theta}^-}(\mathbf{z}_s)$ and $\mathbf{x}_{\boldsymbol{\theta}}(\mathbf{z}_t)$ 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 $\mathrm{D}_{\text{KL}}(\mathbf{x}_{\boldsymbol{\theta}}(\mathbf{z}_t), \mathbf{x}_{\boldsymbol{\theta}^-}(\mathbf{z}_s))$ as proposed in Deschenaux & Gulcehre (2024). Thus the final distillation loss is $\mathcal{L}_{\text{DCD}}(\boldsymbol{\theta}, \boldsymbol{\theta}^-) = \mathrm{D}_{\text{KL}}(\mathbf{x}_{\boldsymbol{\theta}}(\mathbf{z}_t, t), \mathbf{x}_{\boldsymbol{\theta}^-}(\mathbf{z}_s, s))$. Following (Song et al., 2023), we optimize this objective via stochastic gradient descent (SGD) on the student model parameters $\boldsymbol{\theta}$, while updating the teacher model parameters $\boldsymbol{\theta}^-$ 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 8×A100s or 8×H100s 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., 2025)) 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 codebase[1], 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 (↓) 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	Lambada	AG News	Pubmed	Arxiv
Autoregressive							
Transformer†	82.05	25.75	51.25	51.28	52.09	49.01	41.73
Diffusion (absorbing state)							
SEDD Absorb†	100.09	34.28	68.20	49.86	62.09	44.53	38.48
D3PM Absorb¶	200.82	50.86	138.92	93.47	-	-	-
MDLM†	95.26	32.83	67.01	47.52	61.15	41.89	37.37
Diffusion (uniform state / Gaussian)							
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	-	-	-
UDLM‡	112.82	39.42	77.59	53.57	80.96	50.98	44.08
Duo (Ours)	**89.35**	**33.57**	**73.86**	**49.78**	**67.81**	**44.48**	**40.39**

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 $\tau(n < 500\text{K}) = 0.001$ for the first 500K steps, then switch to $\tau(n > 500\text{K}) = 0$ (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×**.

Likelihood Evaluation To compute ppl for Duo, we use (5) with $\alpha_t = 1 - t$ (Schiff et al., 2025). 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.

Table 1: Test perplexities (PPL; ↓) 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 (↓)
Autoregressive	
Transformer†	17.54
Diffusion (absorbing state)	
SEDD Absorb† (Lou et al., 2023)	24.10
MDLM† (Sahoo et al., 2024a)	23.21
Diffusion (uniform state)	
SEDD Uniform‡ (Lou et al., 2023)	29.69
UDLM‡ (Schiff et al., 2025)	27.43
Duo (Ours)	**25.20**

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, Lambada (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**.

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)

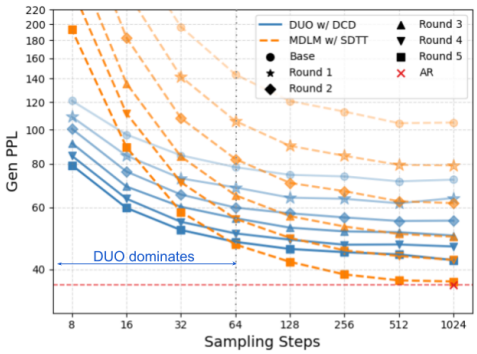

Figure 2: Distillation results for DUO when distilled using DCD (1) and MDLM distilled using SDTT. After distillation, Duo dominates in the low sampling steps regime ≤ 64.

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.

REFERENCES

Marianne Arriola, Subham Sekhar Sahoo, Aaron Gokaslan, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Justin T Chiu, and Volodymyr Kuleshov. Block diffusion: Interpolating between autoregressive and diffusion language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tyEyYT267x.

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, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *International Conference on Machine Learning*, 2009. URL https://api.semanticscholar.org/CorpusID:873046.

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.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling, 2014.

Ting Chen, Ruixiang ZHANG, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=3itjR9QxFw.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. A discourse-aware attention model for abstractive summarization of long documents. *Proceedings of the 2018 Conference of the North American Chapter of the Association 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.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, et al. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*, 2022.

Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

Ishaan Gulrajani and Tatsunori B Hashimoto. Likelihood-based diffusion language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhengfu He, Tianxiang Sun, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. Diffusionbert: Improving generative masked language models with diffusion models. *arXiv preprint arXiv:2211.15029*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=rkE3y85ee.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35: 4328–4343, 2022.

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

Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed, 2021. URL https://arxiv.org/abs/2101.02388.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=S1jE5L5gl.

Mitch Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330, 1993.

Gonzalo Mena, David Belanger, Scott Linderman, and Jasper Snoek. Learning latent permutations with gumbel-sinkhorn networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Byt3oJ-0W.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016.

Hunter Nisonoff, Junhao Xiong, Stephan Allenspach, and Jennifer Listgarten. Unlocking guidance for discrete state-space diffusion and flow models. *arXiv preprint arXiv:2406.01572*, 2024.

Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. Your absorbing discrete diffusion secretly models the conditional distributions of clean data, 2024. URL https://arxiv.org/abs/2406.03736.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernandez. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1525–1534, Berlin, Germany, August 2016. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P16-1144.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Subham Sekhar Sahoo, Marianne Arriola, Aaron Gokaslan, Edgar Mariano Marroquin, Alexander M Rush, Yair Schiff, Justin T Chiu, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL https://openreview.net/forum?id=L4uaAR4ArM.

Subham Sekhar Sahoo, Aaron Gokaslan, Christopher De Sa, and Volodymyr Kuleshov. Diffusion models with learned adaptive noise. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=loMa99A4p8.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models, 2022. URL https://arxiv.org/abs/2202.00512.

Yair Schiff, Chia-Hsiang Kao, Aaron Gokaslan, Tri Dao, Albert Gu, and Volodymyr Kuleshov. Caduceus: Bi-directional equivariant long-range dna sequence modeling. *arXiv preprint arXiv:2403.03234*, 2024.

Yair Schiff, Subham Sekhar Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dalla-torre, Bernardo P de Almeida, Alexander M Rush, Thomas PIERROT, and Volodymyr Kuleshov. Simple guidance mechanisms for discrete diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=i5MrJ6g5G1.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K. Titsias. Simplified and generalized masked diffusion for discrete data, 2025. URL https://arxiv.org/abs/2406.04329.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

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

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models, 2023. URL https://arxiv.org/abs/2310.14189.

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, 2023. URL https://arxiv.org/abs/2303.01469.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Alex Wang and Kyunghyun Cho. Bert has a mouth, and it must speak: Bert as a markov random field language model. *arXiv preprint arXiv:1902.04094*, 2019.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, 2024.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NIPS*, 2015.

Kaiwen Zheng, Cheng Lu, Jianfei Chen, and Jun Zhu. Improved techniques for maximum likelihood estimation for diffusion odes. In *International Conference on Machine Learning*, pp. 42363–42389. PMLR, 2023.

Kaiwen Zheng, Yongxin Chen, Hanzi Mao, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling. *arXiv preprint arXiv:2409.02908*, 2024.

CONTENTS

Appendices

Appendix A The Diffusion Duality

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 r.v. $\mathbf{y} = \tilde{\alpha}_t \mathbf{x} + \tilde{\sigma}_t \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}_K)$ and $\tilde{\sigma}_t = \sqrt{1 - \tilde{\alpha}_t^2}$.

A.1 Discrete Marginals

Our goal in this section is to derive the pmf of the r.v. $\arg\max(\mathbf{y})$. The proof has three parts. In **part 1**, we derive pdf of the the random variables $\mathbf{y}_k$ and $\mathbf{y}_{i \neq k}$. Next in **part 2**, we derive the pdf of the random variable $Z_{\neq k} = \max(\{\mathbf{y}_i : i \neq k\})$. Finally in **part 3**, we derive the distribution of $\max(Z_{\neq k}, \mathbf{y}_k)$ which is the key to constructing the pmf of the r.v. $\arg\max(\mathbf{y})$.

Part 1 It can be easily seen that every entry in $\mathbf{y}$ is a Gaussian r.v. with

$$\mathbf{y}_k \sim \mathcal{N}(\tilde{\alpha}_t, \tilde{\sigma}_t^2) \tag{16}$$

$$\mathbf{y}_{i \neq k} \sim \mathcal{N}(0, \tilde{\sigma}_t^2). \tag{17}$$

Part 2 Since, $\mathbf{y}_{i \neq k}$ follows a Gaussian distribution with 0 mean and $\tilde{\sigma}_t$ standard deviation, the probability of $\mathbf{y}_{i \neq k} < l$ where $l \in \mathbb{R}$ is

$$P(\mathbf{y}_{i \neq k} < l) = \Phi\left(\frac{l}{\tilde{\sigma}_t}\right) \tag{18}$$

where $\Phi(z) = \int_{-\infty}^{z} \exp(-t^2/2) \mathrm{d}z / \sqrt{2\pi}$ is the cumulative distribution function of the Gaussian distribution. This allows us to compute the pdf of the r.v. $Z_{\neq k} = \max(\{\mathbf{y}_i : i \neq k\})$ in the following manner:

$$P(Z_{\neq k} < l) = \Pi_{i \neq k} P(\mathbf{y}_i < l) = \Phi^{K-1}\left(\frac{l}{\tilde{\sigma}_t}\right), \tag{19}$$

where $P(Z_{\neq k} < l)$ is the probability that $Z_{\neq k} < l$ for $l \in \mathbb{R}$.

Part 3 Let $P(\arg\max(\mathbf{y})_k = 1)$ denote the probability that the index k is the index of the maximum entry in $\mathbf{y}$. This is equal to the probability of every other entry $\mathbf{y}_{i \neq k} < \mathbf{y}_k$. Let $\phi(z) = \exp(-z^2)/\sqrt{2\pi}$ denote the standard Normal distribution. Hence,

$$\begin{aligned}
P(\arg\max(\mathbf{y})_k = 1) &= P(Z_{\neq k} < \mathbf{y}_k) \\
&= \int_{-\infty}^{\infty} P(Z_{\neq k} < l) P(\mathbf{y}_k = l) \mathrm{d}l \\
&= \int_{-\infty}^{\infty} P(Z_{\neq k} < l) \left[\frac{1}{\tilde{\sigma}_t} \phi\left(\frac{l - \tilde{\alpha}_t}{\tilde{\sigma}_t}\right)\right] \mathrm{d}l && \text{From (16)} \\
&= \int_{-\infty}^{\infty} \Phi^{K-1}\left(\frac{l}{\tilde{\sigma}_t}\right) \left[\frac{1}{\tilde{\sigma}_t} \phi\left(\frac{l - \tilde{\alpha}_t}{\tilde{\sigma}_t}\right)\right] \mathrm{d}l && \text{From (19)} \\
&= \int_{-\infty}^{\infty} \Phi^{K-1}\left(\tilde{l}\right) \phi\left(\tilde{l} - \frac{\tilde{\alpha}_t}{\tilde{\sigma}_t}\right) \mathrm{d}\tilde{l} && \text{Substituting } \tilde{l} = l/\tilde{\sigma}_t \\
&= \int_{-\infty}^{\infty} \phi\left(\tilde{l} - \frac{\tilde{\alpha}_t}{\sqrt{1 - \tilde{\alpha}_t^2}}\right) \Phi^{K-1}\left(\tilde{l}\right) \mathrm{d}\tilde{l}.
\end{aligned} \tag{20}$$

Note that the indices $i \neq k$ and $j \neq k$ have the same probability of being the indices of maximum entry in $\mathbf{y}$ because both r.v.s $\mathbf{y}_{i \neq k}$ and $\mathbf{y}_{j \neq k}$ have the same pmf specified by (17). Thus,

$$P(\arg\max(\mathbf{y})_{i \neq k} = 1) = P(\arg\max(\mathbf{y})_{j \neq k} = 1) \quad \forall 0 \leq i \neq k < K, 0 \leq j \neq k < K. \tag{21}$$

Thus we can compute $P(\arg\max(\mathbf{y})_{i\neq k}=1)$ in the following manner:

$$\begin{aligned}
&\sum_i P(\arg\max(\mathbf{y})_i=1)=1\\
&\implies P(\arg\max(\mathbf{y})_k=1)+\sum_{i\neq k}P(\arg\max(\mathbf{y})_i=1)=1\\
&\implies P(\arg\max(\mathbf{y})_k=1)+(K-1)P(\arg\max(\mathbf{y})_{i\neq k}=1)=1 && \text{From (21)}\\
&\implies P(\arg\max(\mathbf{y})_{i\neq k}=1)=\frac{1}{K-1}\left[1-P(\arg\max(\mathbf{y})_k=1)\right]\\
&\implies P(\arg\max(\mathbf{y})_{i\neq k}=1)=\frac{1}{K-1}\left[1-\int_{-\infty}^{\infty}\phi\left(\tilde{l}-\frac{\tilde{\alpha}_t}{\sqrt{1-\tilde{\alpha}_t{}^2}}\right)\Phi^{K-1}\left(\tilde{l}\right)\mathrm{d}\tilde{l}\right] && \text{From (20)} \quad (22)
\end{aligned}$$

Let $\beta_t = P(\arg\max(\mathbf{y})_{i\neq k}=1)$. Then, from (20) and (22) we have $P(\arg\max(\mathbf{y})_{i=k}=1) = \beta_t + (1-K)\beta_t$. Thus,

$$P(\arg\max(\mathbf{y})_i=1)=\begin{cases}\beta_t, & i\neq k\\ \beta_t+(1-K)\beta_t. & i=k\end{cases} \tag{23}$$

(23) can be written in vectorized form in the following manner:

$$P(\arg\max(\mathbf{y}))=\mathrm{Cat}(.;\beta_t\mathbf{1}+(1-K\beta_t)\mathbf{x}). \tag{24}$$

A.2 Time change of Discrete Marginals

Let p_t denote $P(\arg\max(\mathbf{y}))$ in (24). It's time-derivative $\frac{\mathrm{d}}{\mathrm{d}t}p_t$ is as follows:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}p_t &= \beta_t'\mathbf{1}-K\beta_t'\mathbf{x}\\
&=\beta_t'(\mathbf{1}-K\mathbf{x})\\
&=\frac{\beta_t'}{1-K\beta_t}(1-K\beta_t)(\mathbf{1}-K\mathbf{x})\\
&=\frac{\beta_t'}{1-K\beta_t}(\beta_tK\mathbf{1}-\beta_tK\mathbf{1}+(1-K\beta_t)(\mathbf{1}-K\mathbf{x}))\\
&=\frac{\beta_t'}{1-K\beta_t}(\beta_t[\mathbf{1}\mathbf{1}^\top]\mathbf{1}-\beta_tK\mathbf{1}+(1-K\beta_t)(\mathbf{1}-K\mathbf{x}))\\
&=\frac{\beta_t'}{1-K\beta_t}(\beta_t([\mathbf{1}\mathbf{1}^\top]\mathbf{1}-K\mathbf{1})+(1-K\beta_t)(\mathbf{1}-K\mathbf{x}))\\
&=\frac{\beta_t'}{1-K\beta_t}(\beta_t[\mathbf{1}\mathbf{1}^\top-K\mathbf{I}]\mathbf{1}+(1-K\beta_t)(\mathbf{1}-K\mathbf{x}))\\
&=\frac{\beta_t'}{1-K\beta_t}(\beta_t[\mathbf{1}\mathbf{1}^\top-K\mathbf{I}]\mathbf{1}+(1-K\beta_t)(\mathbf{1}\mathbf{1}^\top\mathbf{x}-K\mathbf{x}))\\
&=\frac{\beta_t'}{1-K\beta_t}(\beta_t[\mathbf{1}\mathbf{1}^\top-K\mathbf{I}]\mathbf{1}+(1-K\beta_t)[\mathbf{1}\mathbf{1}^\top-K\mathbf{I}]\mathbf{x})\\
&=\frac{\beta_t'}{1-K\beta_t}[\mathbf{1}\mathbf{1}^\top-K\mathbf{I}][\beta_t\mathbf{1}+(1-K\beta_t)\mathbf{x}]\\
&=\frac{\beta_t'}{1-K\beta_t}[\mathbf{1}\mathbf{1}^\top-K\mathbf{I}]p_t
\end{aligned} \tag{25}$$

Let $\alpha_t = 1-K\beta_t$. The functional form of α_t is given as:

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

$$= \frac{K}{K-1}\int_{-\infty}^{\infty}\phi\left(\tilde{l}-\frac{\tilde{\alpha}_t}{\sqrt{1-\tilde{\alpha}_t{}^2}}\right)\Phi^{K-1}\left(\tilde{l}\right)\mathrm{d}\tilde{l}-\frac{1}{K-1}$$
$$= \frac{K}{K-1}\left[\int_{-\infty}^{\infty}\phi\left(\tilde{l}-\frac{\tilde{\alpha}_t}{\sqrt{1-\tilde{\alpha}_t{}^2}}\right)\Phi^{K-1}\left(\tilde{l}\right)\mathrm{d}\tilde{l}-\frac{1}{K}\right] \tag{26}$$

Substituting $\beta_t = (1-\alpha_t)/K$ in (24) and (25), we get:

$$p_t = \mathrm{Cat}(.; \alpha_t \mathbf{x} + (1-\alpha_t)\pi) \tag{27}$$
$$\frac{\mathrm{d}}{\mathrm{d}t}p_t = -\frac{\alpha_t{}'}{K\alpha_t}[\mathbf{1}\mathbf{1}^\top - K\mathbf{I}]p_t \tag{28}$$

where $\alpha_t{}'$ denotes the time-derivative of α_t. Let $\mathbf{z}_t = \arg\max(\mathbf{y})$. The pmf of $\mathbf{z}_t$ is specified in (27) which evolves according to an ordinary differential equation (ODE) (28). This pmf and the ODE are the unique signatures of a uniform state discrete diffusion process (Lou et al., 2023; Schiff et al., 2025). This concludes our proof.

A.3 Gaussian ELBO vs Discrete ELBO

Let $\mathbf{w}_s \sim \tilde{q}_t(.|\mathbf{x})$ and $\mathbf{w}_t \sim \tilde{q}_t(.|\mathbf{x})$ be two intermediate latents for the Gaussian diffusion process defined on $\mathbf{x}$. Let $\mathbf{z}_s = \arg\max(\mathbf{w}_s)$ and $\mathbf{z}_t = \arg\max(\mathbf{w}_t)$. Let $q(\mathbf{w}_s, \mathbf{z}_s|\mathbf{w}_t, \mathbf{z}_t, \mathbf{x})$ denote the true joint reverse posterior and $p_\theta(\mathbf{w}_s, \mathbf{z}_s|\mathbf{w}_t, \mathbf{z}_t)$ denote the approximate reverse joint posterior.

To derive the relationship we require the following properties:

Since, $\mathbf{z}_t$ is a deterministic transformation of $\mathbf{w}_t$,

$$q(\mathbf{w}_s|\mathbf{z}_s, \mathbf{w}_t, \mathbf{z}_t, \mathbf{x}) = q(\mathbf{w}_s|\mathbf{z}_s, \mathbf{w}_t, \mathbf{x}) \tag{29}$$

Since, the transition $\mathbf{z}_t \to \mathbf{z}_s$ is Markov,

$$q(\mathbf{z}_s|\mathbf{w}_t, \mathbf{z}_t, \mathbf{x}) = q(\mathbf{z}_s|\mathbf{z}_t, \mathbf{x}) \tag{30}$$
$$p_\theta(\mathbf{z}_s|\mathbf{w}_t, \mathbf{z}_t) = p_\theta(\mathbf{z}_s|\mathbf{z}_t) \tag{31}$$

Since, the transition $\mathbf{w}_t \to \mathbf{w}_s$ is Markov,

$$q(\mathbf{w}_s|\mathbf{w}_t, \mathbf{z}_t, \mathbf{x}) = q(\mathbf{w}_s|\mathbf{w}_t, \mathbf{x}) \tag{32}$$

Since, $\mathbf{z}_s = \arg\max(\mathbf{w}_s)$,

$$q(\mathbf{z}_s|\mathbf{w}_s, \mathbf{w}_t, \mathbf{z}_t, \mathbf{x}) = q(\mathbf{z}_s|\mathbf{w}_s) = \mathrm{Cat}(.; \arg\max(\mathbf{w}_s)) \tag{33}$$

$$\begin{aligned}
&\mathrm{D}_{\mathrm{KL}}(q(\mathbf{w}_s, \mathbf{z}_s|\mathbf{w}_t, \mathbf{z}_t, \mathbf{x})\|p_\theta(\mathbf{w}_s, \mathbf{z}_s|\mathbf{w}_t, \mathbf{z}_t)) \\
&= \sum_{\mathbf{z}_s}\int_{\mathbf{w}_s} q(\mathbf{w}_s, \mathbf{z}_s|\mathbf{w}_t, \mathbf{z}_t, \mathbf{x}) \log \frac{q(\mathbf{w}_s, \mathbf{z}_s|\mathbf{w}_t, \mathbf{z}_t, \mathbf{x})}{p_\theta(\mathbf{w}_s, \mathbf{z}_s|\mathbf{w}_t, \mathbf{z}_t)}\mathrm{d}\mathbf{w}_s \\
&= \sum_{\mathbf{z}_s}\int_{\mathbf{w}_s} q(\mathbf{w}_s, \mathbf{z}_s|\mathbf{w}_t, \mathbf{z}_t, \mathbf{x}) \log \frac{q(\mathbf{w}_s, \mathbf{z}_s|\mathbf{w}_t, \mathbf{z}_t, \mathbf{x})}{p_\theta(\mathbf{w}_s, \mathbf{z}_s|\mathbf{w}_t, \mathbf{z}_t)}\mathrm{d}\mathbf{w}_s \\
&= \sum_{\mathbf{z}_s}\int_{\mathbf{w}_s} q(\mathbf{w}_s|\mathbf{z}_s, \mathbf{w}_t, \mathbf{z}_t, \mathbf{x})q(\mathbf{z}_s|\mathbf{w}_t, \mathbf{z}_t, \mathbf{x}) \log \frac{q(\mathbf{w}_s|\mathbf{z}_s, \mathbf{w}_t, \mathbf{z}_t, \mathbf{x})q(\mathbf{z}_s|\mathbf{w}_t, \mathbf{z}_t, \mathbf{x})}{p_\theta(\mathbf{w}_s|\mathbf{z}_s, \mathbf{w}_t, \mathbf{z}_t)p_\theta(\mathbf{z}_s|\mathbf{w}_t, \mathbf{z}_t)}\mathrm{d}\mathbf{w}_s \\
&= \sum_{\mathbf{z}_s}\int_{\mathbf{w}_s} q(\mathbf{w}_s|\mathbf{z}_s, \mathbf{w}_t, \mathbf{x})q(\mathbf{z}_s|\mathbf{z}_t, \mathbf{x}) \log \frac{q(\mathbf{w}_s|\mathbf{z}_s, \mathbf{w}_t, \mathbf{x})q(\mathbf{z}_s|\mathbf{z}_t, \mathbf{x})}{p_\theta(\mathbf{w}_s|\mathbf{z}_s, \mathbf{w}_t)p_\theta(\mathbf{z}_s|\mathbf{z}_t)}\mathrm{d}\mathbf{w}_s \\
&= \mathbb{E}_{\mathbf{z}_s}\mathrm{D}_{\mathrm{KL}}(q(\mathbf{w}_s|\mathbf{z}_s, \mathbf{w}_t, \mathbf{x})\|p_\theta(\mathbf{w}_s|\mathbf{z}_s, \mathbf{w}_t)) + \mathrm{D}_{\mathrm{KL}}(q(\mathbf{z}_s|\mathbf{z}_t, \mathbf{x})\|p_\theta(\mathbf{z}_s|\mathbf{z}_t))
\end{aligned} \tag{34}$$

Thus, we have Also,

$$\begin{aligned}
&\mathrm{D}_{\mathrm{KL}}(q(\mathbf{w}_s, \mathbf{z}_s|\mathbf{w}_t, \mathbf{z}_t, \mathbf{x})\|p_\theta(\mathbf{w}_s, \mathbf{z}_s|\mathbf{w}_t, \mathbf{z}_t)) \\
&= \int_{\mathbf{w}_s}\sum_{\mathbf{z}_s} q(\mathbf{w}_s, \mathbf{z}_s|\mathbf{w}_t, \mathbf{z}_t, \mathbf{x}) \log \frac{q(\mathbf{w}_s, \mathbf{z}_s|\mathbf{w}_t, \mathbf{z}_t, \mathbf{x})}{p_\theta(\mathbf{w}_s, \mathbf{z}_s|\mathbf{w}_t, \mathbf{z}_t)}\mathrm{d}\mathbf{w}_s
\end{aligned}$$

$$= \int_{\mathbf{w}_s} \sum_{\mathbf{z}_s} q(\mathbf{z}_s|\mathbf{w}_s, \mathbf{w}_t, \mathbf{z}_t, \mathbf{x}) q(\mathbf{w}_s|\mathbf{w}_t, \mathbf{z}_t, \mathbf{x}) \log \frac{q(\mathbf{w}_s|\mathbf{w}_t, \mathbf{z}_t, \mathbf{x}) q(\mathbf{z}_s|\mathbf{w}_s, \mathbf{w}_t, \mathbf{z}_t, \mathbf{x})}{p(\mathbf{w}_s|\mathbf{w}_t, \mathbf{z}_t) p_\theta(\mathbf{z}_s|\mathbf{w}_t, \mathbf{z}_t)} \mathrm{d}\mathbf{w}_s$$

$$= \int_{\mathbf{w}_s} \sum_{\mathbf{z}_s} q(\mathbf{z}_s|\mathbf{w}_s, \mathbf{w}_t, \mathbf{z}_t, \mathbf{x}) q(\mathbf{w}_s|\mathbf{w}_t, \mathbf{x}) \log \frac{q(\mathbf{w}_s|\mathbf{w}_t, \mathbf{x}) q(\mathbf{z}_s|\mathbf{w}_s, \mathbf{w}_t, \mathbf{x})}{p_\theta(\mathbf{w}_s|\mathbf{w}_t) p_\theta(\mathbf{z}_s|\mathbf{w}_s, \mathbf{z}_t)} \mathrm{d}\mathbf{w}_s$$

$$= \int_{\mathbf{w}_s} \sum_{\mathbf{z}_s} q(\mathbf{z}_s|\mathbf{w}_s) q(\mathbf{w}_s|\mathbf{w}_t, \mathbf{x}) \log \frac{q(\mathbf{w}_s|\mathbf{w}_t, \mathbf{x}) q(\mathbf{z}_s|\mathbf{w}_s)}{p_\theta(\mathbf{w}_s|\mathbf{w}_t) p_\theta(\mathbf{z}_s|\mathbf{w}_s)} \mathrm{d}\mathbf{w}_s$$

$$= \int_{\mathbf{w}_s} q(\mathbf{z}_s = \arg\max(\mathbf{w}_s)|\mathbf{w}_s) q(\mathbf{w}_s|\mathbf{w}_t, \mathbf{x}) \log \frac{q(\mathbf{w}_s|\mathbf{w}_t, \mathbf{x}) q(\mathbf{z}_s = \arg\max(\mathbf{w}_s)|\mathbf{w}_s)}{p_\theta(\mathbf{w}_s|\mathbf{w}_t) p_\theta(\mathbf{z}_s = \arg\max(\mathbf{w}_s)|\mathbf{w}_s)} \mathrm{d}\mathbf{w}_s$$

$$= \int_{\mathbf{w}_s} q(\mathbf{w}_s|\mathbf{w}_t, \mathbf{x}) \log \frac{q(\mathbf{w}_s|\mathbf{w}_t, \mathbf{x})}{p_\theta(\mathbf{w}_s|\mathbf{w}_t)} \mathrm{d}\mathbf{w}_s$$

$$= \mathrm{D}_{\mathrm{KL}}(q(\mathbf{w}_s|\mathbf{w}_t, \mathbf{x}) \| p_\theta(\mathbf{w}_s|\mathbf{w}_t)) \tag{35}$$

From (34) and (35) we get

$$\mathrm{D}_{\mathrm{KL}}(q(\mathbf{w}_s|\mathbf{w}_t, \mathbf{x}) \| p(\mathbf{w}_s|\mathbf{w}_t)) = \underbrace{\mathbb{E}_{\mathbf{z}_t}\Big[\mathbb{E}_{\mathbf{z}_s} \mathrm{D}_{\mathrm{KL}}(q(\mathbf{w}_s|\mathbf{z}_s, \mathbf{w}_t, \mathbf{x}) \| p(\mathbf{w}_s|\mathbf{z}_s, \mathbf{w}_t))\Big]}_{\geq 0} + \mathrm{D}_{\mathrm{KL}}(q(\mathbf{z}_s|\mathbf{z}_t, \mathbf{x}) \| p(\mathbf{z}_s|\mathbf{z}_t))$$

$$\implies \mathrm{D}_{\mathrm{KL}}(q(\mathbf{w}_s|\mathbf{w}_t, \mathbf{x}) \| p(\mathbf{w}_s|\mathbf{w}_t)) \geq \mathrm{D}_{\mathrm{KL}}(q(\mathbf{z}_s|\mathbf{z}_t, \mathbf{x}) \| p(\mathbf{z}_s|\mathbf{z}_t))$$

$$\implies \underbrace{-\mathrm{D}_{\mathrm{KL}}(q(\mathbf{w}_s|\mathbf{w}_t, \mathbf{x}) \| p(\mathbf{w}_s|\mathbf{w}_t))}_{\equiv \mathrm{ELBO}(\tilde{q}_t, p_\theta)\,(7)} \leq \underbrace{-\mathrm{D}_{\mathrm{KL}}(q(\mathbf{z}_s|\mathbf{z}_t, \mathbf{x}) \| p(\mathbf{z}_s|\mathbf{z}_t))}_{\equiv \mathrm{ELBO}(q_t, p_\theta)\,(4)}$$

$$\implies \mathrm{ELBO}(\tilde{q}_t, p_\theta; \mathbf{x}) \leq \mathrm{ELBO}(q_t, p_\theta; \mathbf{x})$$

$$\implies \log p_\theta(\mathbf{x}) \geq \mathrm{ELBO}(q_t, p_\theta; \mathbf{x}) \geq \mathrm{ELBO}(\tilde{q}_t, p_\theta; \mathbf{x}) \tag{36}$$

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 Additional Proofs

B.1 ELBO Equivalence

Note that $f(\mathbf{z}_t, \mathbf{x}_\theta(\mathbf{z}_t, t), \alpha_t; \mathbf{x})$ is invariant to the functional form of the noise schedule α_t as long as $\alpha_{t=0} = 1$ and $\alpha_{t=1} = 0$ (Schiff et al., 2025).

Consider a discrete diffusion process q' with a noise schedule $\mathcal{T}(g(t))$ where $g : [0, 1] \to [0, 1]$ and $\mathcal{T}$ .... Note that at t=0 and t=1 the noise schedule evaluates to ... Thus,

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{\mathbf{x}, t\sim\mathcal{U}[0,1], \mathbf{z}_t\sim q_t(.|\mathbf{x})} f(\mathbf{z}_t, \mathbf{x}_\theta(\mathbf{z}_t, t), \alpha_t; \mathbf{x}) \tag{37}$$

$$= \mathbb{E}_{\mathbf{x}, t\sim\mathcal{U}[0,1], \mathbf{z}_t\sim q'_t(.|\mathbf{x})} f(\mathbf{x}, \mathbf{z}_t, \alpha_t = \mathcal{T}(g(t)), \mathbf{x}_\theta(\mathbf{z}_t)) \tag{38}$$

Let the Gaussian diffusion process underlying q' be $\tilde{q}_t$. This Gaussian process Thus, for $\mathbf{w}_t \sim \tilde{q}_t$, $\tilde{\mathbf{z}}_t = \arg\max(\mathbf{w}_t) \sim q'$ from (...). Thus,

$$= \mathbb{E}_{\mathbf{x}, t\sim\mathcal{U}[0,1], \mathbf{z}_t\sim q'_t(.|\mathbf{x})} f(\mathbf{x}, \mathbf{z}_t, \alpha_t = \mathcal{T}(g(t)), \mathbf{x}_\theta(\mathbf{z}_t)) \tag{39}$$

$$= \mathbb{E}_{\mathbf{x}, t\sim\mathcal{U}[0,1], \mathbf{w}_t\sim\tilde{q}_t(.|\mathbf{x})} f(\mathbf{x}, \mathbf{z}_t := \arg\max(\mathbf{w}_t), \alpha_t = \mathcal{T}(g(t)), \mathbf{x}_\theta(\mathbf{z}_t := \arg\max(\mathbf{w}_t))) \tag{40}$$

As a sanity check, we empirically verify the equivalence of (13) and (37). To do this, we trained UDLM (Schiff et al., 2025) on LM1B (Table 3) using the true ELBO from (37). We then evaluated the model using Gaussian latents and (13), and recovered the same perplexity (36.71) as when using discrete latents. For each datapoint $\mathbf{x}$, we used 1000 Monte Carlo samples for t sampled using antithetic-sampling, with a linear schedule for $\tilde{\alpha}_t = 1 - t$.

B.2 Optimal DDIM trajectories

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

$$\mathbf{z}_s = \tilde{\alpha}_s \mathbf{x}_\theta(\mathbf{z}_t) + \sqrt{1 - \tilde{\alpha}_s^2} \epsilon_\theta(\mathbf{z}_t) \tag{41}$$

Algorithm 1 Dual Consistency Distillation (DCD)

Input: data $\mathbf{x} \sim q_{\text{data}}$, learning rate η, number of distillation rounds K, number of training iterations per round M, ema μ, discretization step Δ.
for $i = 1$ **to** K **do**
  $\boldsymbol{\theta} \leftarrow \text{stopgrad}(\boldsymbol{\theta}^-)$
  for $i = 1$ **to** M **do**
    Sample $\mathbf{x} \sim q_{\text{data}}$, $t \sim \mathcal{U}[0,1]$, and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I_K)$.
    $s \leftarrow \max(t - i \cdot \Delta, 0)$
    $\mathbf{z}_s \leftarrow \arg\max(\tilde{\alpha}_s \mathbf{x} + \sqrt{1 - \tilde{\alpha}_s{}^2}\boldsymbol{\epsilon})$
    $\mathbf{z}_t \leftarrow \arg\max(\tilde{\alpha}_t \mathbf{x} + \sqrt{1 - \tilde{\alpha}_t{}^2}\boldsymbol{\epsilon})$
    $\mathcal{L}_{\text{DCD}}(\boldsymbol{\theta}, \boldsymbol{\theta}^-) \leftarrow \text{D}_{\text{KL}}(\mathbf{x}_\theta(\mathbf{z}_t, t), \mathbf{x}_{\boldsymbol{\theta}^-}(\mathbf{z}_s, s))$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{DCD}}(\boldsymbol{\theta}, \boldsymbol{\theta}^-)$
    $\boldsymbol{\theta}^- \leftarrow \text{stopgrad}(\mu \boldsymbol{\theta}^- + (1 - \mu)\boldsymbol{\theta})$
  end for
end for=0

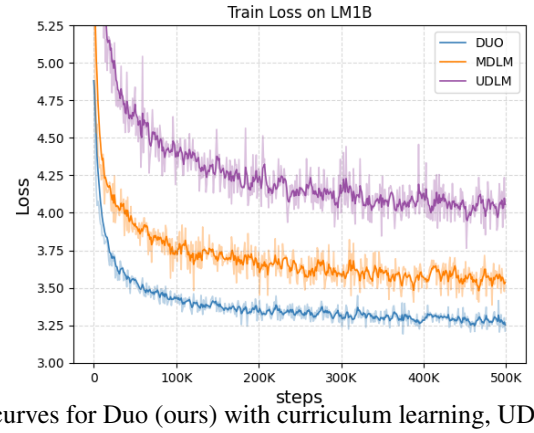

Figure 3: Training loss curves for Duo (ours) with curriculum learning, UDLM, and MDLM. We see observe that curriculum learning leads to low variance training.

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 (41) reduces to $\mathbf{z}_s = \tilde{\alpha}_t \mathbf{x} + \sqrt{1 - \tilde{\alpha}_t{}^2}\tilde{\epsilon}$. This holds $\forall s \in [0,1]$. Hence the optimal DDIM trajectory $\mathcal{P}_{\text{DDT}}(\mathbf{x}, \boldsymbol{\epsilon})$ is given as $\mathcal{P}_{\text{DDIM}}(\mathbf{x}, \boldsymbol{\epsilon}) = \left\{\tilde{\alpha}_t \mathbf{x} + \sqrt{1 - \tilde{\alpha}_t{}^2}\boldsymbol{\epsilon}\right\}_{t \in [0,1]}$

Appendix C Training Details

C.1 Denoising Model

Unlike prior discrete diffusion approaches, we design the denoising model $p_t^\theta : \Delta^K \cup \mathcal{V} \to \Delta^K$ 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 a transformer (Vaswani et al., 2017), with the token-to-embedding mapping defined by matrix multiplication $\mathbf{y}^\top \texttt{vocab\_embeddings}$ in the first layer, where $\texttt{vocab\_embeddings} \in \mathbb{R}^{K \times m}$ 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 D Curriculum Learning

D.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 $\alpha_t{}'$ in (5), so when $\alpha_t{}' \approx 0$, the

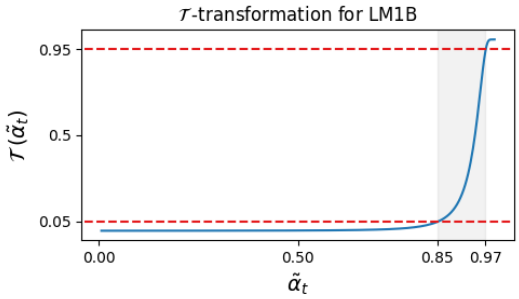

Figure 4: Caption

Table 3: Test perplexities (PPL; ↓) on LM1B. *Reported in He et al. (2022). Best uniform/Gaussian diffusion value is bolded. ¶Denotes the dataset didn't incorporate sentence packing. †Reported in Arriola et al. (2025). For diffusion models, we report the bound on the likelihood. Best diffusion value is underlined. ‡Denotes retrained models.

	PPL (↓)
Autoregressive	
Transformer‡	22.32 / 22.83¶†
Diffusion (absorbing state)	
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>27.03 / 31.78</u>¶†
Diffusion (uniform state / Gaussian)	
Diffusion-LM¶* (Li et al., 2022)	118.62
SEDD Uniform¶ (Lou et al., 2023)	40.25
UDLM‡ (Schiff et al., 2025)	31.28 / 36.71¶
Duo (Ours)	**29.95 / 33.68**¶

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 $t \sim [t_{\text{min}}, t_{\text{max}}]$ 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 $t \in [t_{\text{min}}, t_{\text{max}}]$, 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 $\lim \tau \to 0^+$ 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 $[t_{\text{min}}, t_{\text{max}}] = [0.03, 0.15]$.

APPENDIX E ADDITIONAL EXPERIMENTS

E.1 LM1B

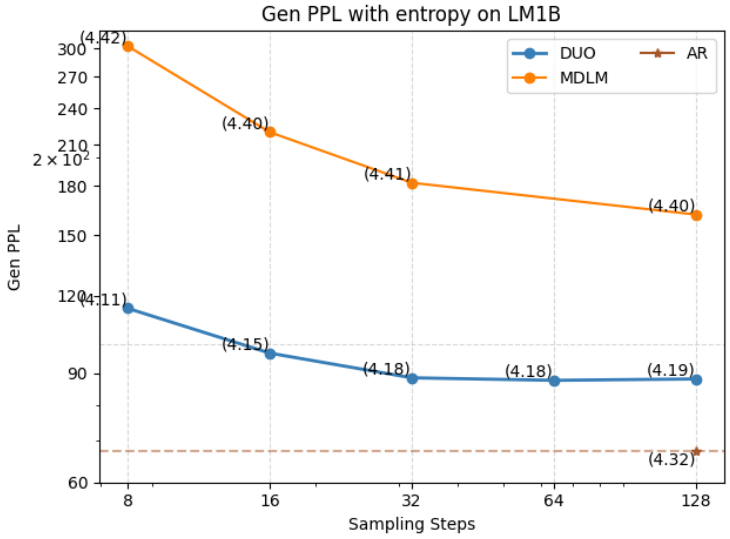

Figure 5: Gen PPL for Duo and MDLM.

E.2 TAU ABLATIONS

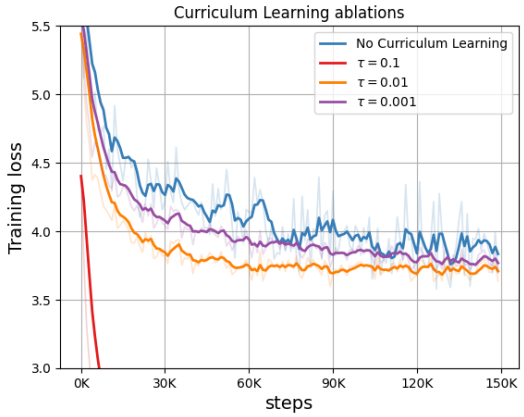

Figure 6: Train loss for curriculum learning by varying τ. Models were trained on the LM1B dataset.

E.3 SAMPLE QUALITY

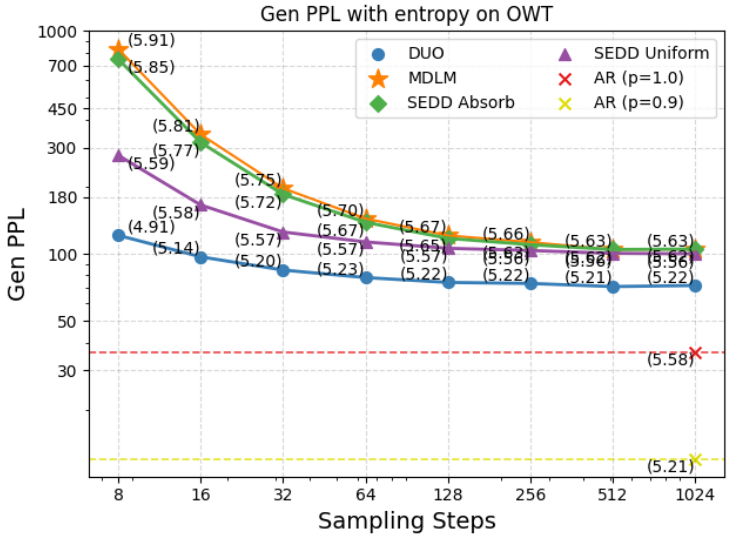

Figure 7: Sample quality comparision between Duo (ours), MDLM, SEDD (Absorb / Uniform), and AR. The numbers in brackets denote the entropy of the samples. Note that the entropy of samples for DUO is quite similar to that of the AR model with nucleus sampling (p=0.9) while the entropy of the samples for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

E.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
Base Model				
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
Round 1				
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
Round 2				
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
Round 3				
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
Round 4				
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
Round 5				
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

E.5 SAMPLES

Samples from a distilled Duo.

E.5.1 T = 1024

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

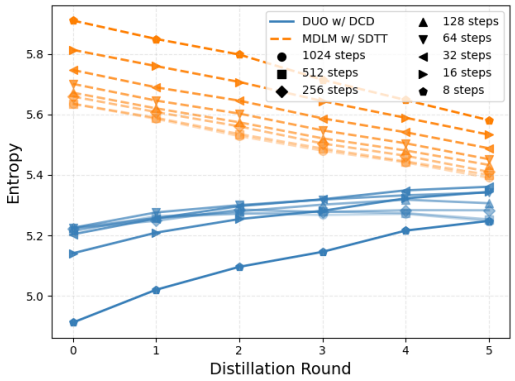

Figure 8: Entropy of MDLM distilled using SDTT, and of DUO distilled using CDC. The entropy of the SDTT-distilled MDLM decreases with distillation, while the entropy of the CDC-distilled DUO model increases. The curves corresponding to a higher number of sampling steps are displayed with lighter colors 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 groupž2019s headquarters.
n
nThe community is coming out of a ž01cfairly cautious,ž01d and ž01cun-vigigativež01d 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
nž01cIt is great to be a community, but it is important to be encouraged by the community to find the best leaders,ž01d 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-

duce royalty payments for the residential customers owned by the well to 10 percent in 2012, according to a Los Angeles-based trade union source familiar with the federal antitrust talks.
n
nThe new group, Ž01cVibrant Public ElectricŽ01d (PGP), would, still still include some customers in the Pacific Parcel, other parts of California owned by the, and a part-owner and owner of of intellectual utilities.
n
nThe prospect of a reorganization, PPC and the company at least $1 billion over the option of operating without it. The Ž01cPerties Reduction ReviewŽ01d agreed Tuesday with state regulators to put a stop to the utility from operating this year.
n
nFILE - In this this June 24, 2013 file photo, PG&T Chairman Jerry Hayward speaks in front of Reuters Reuters reporters during a deadline to announce proposed lawsuits with federal regulators. WSDA officials met several times about the possibility of an agreement, according to the sources.
n
nŽ01cWe will the outstanding issues including water and servicing payments, residential rates, the property service fees, customer services, and debt issues,Ž01d said John McAlpine, the CPUC agency spokesman.
n
nPG&T faces a $2 billion operating loss, and despite the deal, PG&T faces uncertainty on a combination either paying the utility a reduction in operating revenue and losing back to shareholders and a a partial reorganization.
n
nAt the same time, the prospects also look grim for an industry enjoying explosive growth this year, with growing demand more sources of electricity and with more federal government competition.
n
nThe move also put pressure on General Electric, the maker Wall Edison and Reliant„ cut down power output by 20 percent in 2011.
n
nŽ01cBASED CUMMARYŽ01d
n
nPG&TŽ019s shares fell $1.50 per share, after a loss of $2.6 billion in the fourth quarter ended Q3.
n
nThe company in the price-per-share market have seen higher competition, higher operating costs and a cost per diluted rate that it the the rates offered by PPC Securities.
n
nŽ01cWith revenue customers, shareholders, such higher increased rates, and better customer service, demand a higher diluted share price,Ž01d said David Smith, president of communications in PG&TŽ019s public relations statement Tuesday.
n
nTheT filed suit in November against a $2 proposed combined utility company and put the suit on Monday with the Federal Trade Commission.
n
nThe company, PG&TŽ019s new group of utilities, also be able to be much richer in cash coming months than possible utilities, with more cash to go out and more moreerous reporting requirements, ac-

```
cording to a Wall Street Journal report earlier this month that
said.
n
nAmong other questions about utilitiesŽ019 potentially toxic as-
sets, the timing of dividend dividend payments will be under
scrutiny and the way due distribution of payments to its work-
ers and retirees will be be affected by Congress under an overhaul
passed Section 18 U.S.C. 111.
n
n[Bottom:  of Reforms]
n
nPG&T owns or percent of the company, make up about 40 percent of
the,<|endoftext|>
```

E.5.2 T = 8

```
<|endoftext|> at the time, also critical of that the authors used
the site promote Mideastism, the to Ž01ccoexist.Ž01d He blamed the
Saudi authorities, for carrying out Ž01ca Censorship Convention
on the website, which being known for censorship on this [ [.Ž01d
Ž01cThe siteŽ01d are not officially known but, according the gov-
ernment officials, have provoked international criticism„ in a
number of some criticism of the universityŽ019s activities the
members of the South Asia of the Gulf Cooperation Council, said
that the university had no legal right to to restrict access to
students.
n
nAh Jafar, No.  correspondent at The Guardian, has noted that the
attacks further proof that Yemen is a country and can consider any
plot, or a potential terrorist plot of serious consequences, to
its own people.  and, indeed, though this opinion article has se-
vere outrage in some European countries.  As Ash Khan, also not
the removed by the non-The Guardian Times for this piece, a piece
purports to to be in solidarity with the terroristsŽ019s massacre
of satirical magazine Charlie Hebdo because it features a cari-
cature of a woman in a, high-color, according Mr the TimesŽ019
investigative„ conctions.
n
nMs.  She describes, if the piece of Charlie Hebdo that pro-
vokes readers's outrage, of her pet children as Ž01cbruits of
pedophiles.Ž01d But these things are not acceptable in France.
In fact, not in the context of Orthodox Jews (and I don't't get
a muchness at the thought of dealing to the Lat'mod and) Stamp-
sia), is know as much as I do about the film's final act).  But
The Guardian doesn't need a correction, I guess, the page I put it
on, for condemning a cartoon and depicts her a terrorist.  And she
has human speech.<|endoftext|>The Hunger Games has already been
criticized the sexism among many ways,and-g andesŽ014following the
example of actor Akji Aoki, a wting comment the Oct.  18, 2008 is-
sue of The Hollywood Reporter in which he complained about video
games led to negative discussion of gender gender people on women,
bizarrely contending to say that games are social games:  Ïhere
very many games that played, I would play them, but they were not
very good.Äs he said, religion in games only religious ideas of
games, and then video games came out, game was declined.  While
the existence of games is„ but quite disturbing because especially
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.
```

n
nTurn investigation of the down issues is problematic not only because it are so able to be honest and honest about religious not not cults are presstitutes in video popular world; it is also the general tendency of the media to tolerate such comments of supposedly the characters.
n
n[1]. Critic of the following:
n
nŽ022 Punishment because of Ž01cincidentity an offensive and equivocal views of a Western people.Ž01d
n
nŽ022 The tendency of people to get involved in violence. For example, Mr. Miri vilified factory father, a a Buddhist follower, in a speech he is an atheist act for being attacked by a proper audience, which is, well, that bad.
n
nŽ022 Violation of orator of speech, about Ïove, and thedom. The wasler, too, for of the ’-do-violence was punished because the violence was fostered by close support the victim had. Just of this, that is, the diatoms, Ïazy peopleänd dystopism.
n
nThe The Book Translation is a collection of translations of English published on this author here. The Images are taken from the translation.<|endoftext|>Timeing Reassignment, Love and Freedom is a Ph.D. film. U. began to in the Filipinos in 2001 and had crosstexts in a film. It brought back out in 2008 and added back back in 2010.
n
nIn the days, Latin, they scanned the country about 30„ a second per minute. In„ the First and Vietnam war, and around the world during the Second World War War The U U’s Center of Bi History says the footage of the Central American television shows was short but not as continuous. The result.
n
nÏ thought I was in it like ina mediocrat. When I saw it that I was working journalist, a situation that I was in, with half a second of my time off. The person to give a what being out but no one stood up for me. I would have say, Ž018You are doing your wrong. You are good it.Ž019Ž01d ThatŽ019s an the director John Kershaw said the first time he was making U.A.A. War. I.<|endoftext|>",