ACCELERATING DIFFUSION LANGUAGE MODELS VIA INVERSE DISTILLATION

David Li[1,*] **Nikita Gushchin**[2,*] **Dmitry Abulkhanov** **Eric Moulines**[1,3]

Ivan Oseledets[2,4] **Maxim Panov**[1] **Alexander Korotin**[2,4]

ABSTRACT

Diffusion Language Models (DLMs) generate text via iterative reverse diffusion, but the resulting inference latency limits practical use and makes inference-time methods such as guidance expensive. We propose *Inverse-distilled Diffusion Language Models (IDLM)*, a post-training framework that distills a pretrained DLM into a few-step generator by extending inverse distillation to discrete token spaces. IDLM optimizes a bilevel objective: a *fake* diffusion model is trained on student samples with the teacher's diffusion loss, and the student is updated to maximize the teacher–fake loss gap on its own samples. In discrete settings, we (i) establish identifiability by proving a uniqueness guarantee under SEDD, MDLM, and Duo objectives, and (ii) stabilize training with simplex-valued token outputs and differentiable reformulations of the diffusion losses. As a result, experiments on multiple DLMs show that our method reduces the number of inference steps by $4\times$-$64\times$, while preserving the teacher model's entropy and generative perplexity.

1 INTRODUCTION

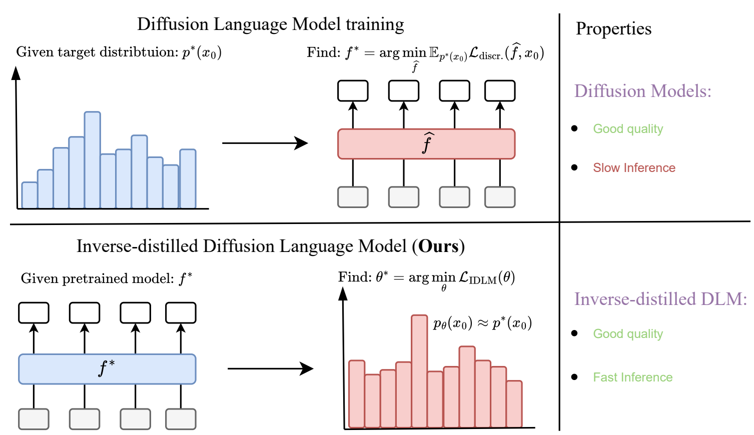


Figure 1: DLMs require many reverse-diffusion steps at inference. *IDLM* trains a few-step generator via inverse distillation to match the teacher's generation quality with substantially fewer steps.

Generative modeling for discrete data, such as natural language, has seen widespread adoption, driven by the rapid advancement of large language models (Touvron et al., 2023; Groeneveld et al., 2024; Lozhkov et al., 2024).

Recently, Diffusion Language Models (DLMs; Sohl-Dickstein et al., 2015; Austin et al., 2021; Campbell et al., 2022; Lou et al., 2024) have emerged as a promising framework for modeling the distribution of the text data. A general formulation for a broad class of DLMs was introduced by (Lou et al., 2024). However, the general formulation is often intractable in practice and introduces

*Equal contribution. [1]Mohamed Bin Zayed University of AI, Abu Dhabi, UAE. [2]Applied AI Institute, Moscow, Russia. [3]EPITA, Laboratoire Recherche de l'EPITA, Paris, France. [4]AXXX, Russia. Correspondence: David.Li@mbzuai.ac.ae, iamalexkorotin@gmail.com

significant theoretical complexity. To address this, subsequent work (Sahoo et al., 2024; 2025a; Schiff et al., 2025; Shi et al., 2024) has focused on specific types of diffusion processes, enabling simplified theoretical analyses and more tractable implementations. As a result, DLMs have achieved competitive performance in text generation, comparable to that of autoregressive language models (Nie et al., 2025). However, the inherently iterative nature of reverse diffusion requires hundreds or even thousands of sampling steps at inference time, resulting in considerable latency that limits their practicality.

On the other hand, in the continuous domain, the issue of slow inference in diffusion models (Ho et al., 2020; Song et al., 2020b; Karras et al., 2022) has been the focus of extensive research (Song et al., 2020a; Luhman & Luhman, 2021; Zhang & Chen, 2022; Lu et al., 2022). Among the various approaches, *distillation* has emerged as one of the most effective solutions (Song et al., 2023; Salimans & Ho, 2022; Luhman & Luhman, 2021; Berthelot et al., 2023; Xie et al., 2024). However, directly applying such techniques in the discrete domain is challenging due to the non-differentiable nature of discrete data. Recent efforts have begun to address this limitation for DLMs. Notably, SDTT (Deschenaux & Gulcehre, 2024) and Duo-DCD (Sahoo et al., 2025a) propose adaptations of consistency-based distillation methods (Song et al., 2023; Song & Dhariwal, 2023; Kim et al., 2023) to the discrete setting, demonstrating initial progress toward enabling fast inference in DLMs.

In this work, we investigate an alternative distillation paradigm by extending the theoretical framework of *Inverse Distillation* (Gushchin et al., 2025; Kornilov et al., 2025), originally developed for continuous diffusion models, to the discrete domain. However, this direct extension introduces several nontrivial challenges arising from the intrinsic properties of discrete data. First, theoretical concerns emerge regarding the validity of the training objective, particularly due to the potential non-uniqueness of its minimizers. Second, practical challenges arise from the non-smoothness of the discrete space, which complicates the optimization.

Contributions. To address the above-mentioned issues, we present *Inverse-distilled Diffusion Language Models (IDLM)*, a principled framework for accelerating a broad class of discrete diffusion models. In summary, our core contributions are as follows:

1. We theoretically establish the uniqueness of the global optimum for the IDLM objective, validating the correctness of the proposed optimization procedure (§3.1).
2. We propose gradient-stable relaxations that enable effective training through (i) a *simplex-parameterized* generator and (ii) a principled mechanism for *backpropagating gradients through the discrete forward process* (§3.2).
3. We empirically demonstrate that IDLM achieves competitive *generation quality*, matching the performance of 1024-step teacher models while requiring up to $64\times$ fewer sampling steps (§4).

2 Preliminaries

Notations. We represent scalar discrete random variables over N categories as one-hot column vectors, with $\mathcal{V} = \{x \in \{0,1\}^N : \sum_{i=1}^N x_i = 1\}$ denoting the set of such vectors. We further define the probability simplex $\Delta := \{z \in \mathbb{R}^N : z \geq 0, \langle z, \vec{1} \rangle = 1\}$ as the convex hull of $\mathcal{V}$, where $\vec{1} \in \mathbb{R}^N$ denotes the all-ones vector. Additionally, let $I \in \mathbb{R}^{N \times N}$ be the identity matrix. The matrix-weighted inner product is defined as $\langle a, b \rangle_Q := a^\top Q b$, and we omit the subscript when $Q = I$. For absorbing diffusion process we define special mask token $m \in \mathcal{V}$ such that $m_N = 1$.

2.1 Diffusion Language Models

We start by recalling the Diffusion Language Models (DLMs; Austin et al., 2021; Campbell et al., 2022; Lou et al., 2024). Consider a distribution $p^*(x_0) \in \Delta$ over $\mathcal{V}$. DLMs aim to generate samples from $p^*(x_0)$ by constructing a transformation process from an initial tractable distribution, based on predefined *forward diffusion*.

Forward Diffusion. A forward discrete diffusion process is defined by a family of time-dependent distributions $p_t(x_t) \in \Delta$, evolving over the interval $t \in [0,1]$. This evolution follows a continuous-time Markov chain defined by a linear ordinary differential equation (Anderson, 2012; Campbell et al., 2022; Lou et al., 2024):

$$\frac{dp_t}{dt} = Q_t p_t, \quad p_0 = p^*, \tag{1}$$

where $Q_t \in \mathbb{R}^{N \times N}$ are the diffusion matrices and have nonnegative non-diagonal entries and columns which sum to zero, ensuring valid probability dynamics. These matrices are typically chosen in the form $Q_t = \sigma_t Q$, where Q is a fixed base matrix that defines the structure of the diffusion process, and σ_t is a time-dependent scalar that controls the diffusion schedule. With an appropriate choice of Q and scheduling function σ_t, the distribution p_t converges to a tractable limiting distribution $\pi(x_1) \in \Delta$ as $t \to 1$.

For a fixed initial state $x_0 \sim p^*(x_0)$, the forward diffusion process defined in equation equation 1 induces a family of conditional distributions over intermediate states $x_t \in \mathcal{V}$.

$$p_{t|0}(x_t \mid x_0) := \mathrm{Cat}(x_t; \exp(\bar{\sigma}_t Q)\, x_0), \tag{2}$$

where, $\exp(\bar{\sigma}_t Q)$ denotes the matrix exponential, and $\bar{\sigma}_t := \int_0^t \sigma_s \, ds$ is the cumulative schedule.

Concrete Score Matching (SEDD).

The goal of the Concrete Score Matching (Meng et al., 2022; Lou et al., 2024) is to approximate the ratio of marginal probabilities $\frac{p_t(y)}{p_t(x)}$. To achieve this, (Lou et al., 2024) propose learning a score function $\widehat{s}\colon \mathcal{V} \times [0,1] \to \mathbb{R}_+^N$ using the *Score Entropy Discrete Diffusion* (SEDD) loss. This objective, defined for a data point $x_0 \sim p^*(x_0)$, is:

SEDD Loss

$$\mathcal{L}_{\mathrm{SEDD}}(\widehat{s}, x_0) := \int_0^1 \mathbb{E}_{p_{t|0}(x_t|x_0)} \left[\sum_{y \neq x_t} \lambda_{x_t y} \times \left(\langle \widehat{s}(x_t, t), y \rangle - \frac{p_{t|0}(y \mid x_0)}{p_{t|0}(x_t \mid x_0)} \log \langle \widehat{s}(x_t, t), y \rangle \right) \right] dt, \tag{3}$$

where $\lambda_{x_t y}$ is positive weighting function. Lou et al. (2024) show that the minimizer for the data distribution $p^*(x_0)$:

$$s^* = \arg\min_{\widehat{s}} \mathbb{E}_{p^*(x_0)} \left[\mathcal{L}_{\mathrm{SEDD}}(\widehat{s}, x_0) \right]$$

recovers the desired ratio of marginal probabilities $\frac{p_t(y)}{p_t(x)}$.

However, as noted in (Lou et al., 2024), applying this formulation to large-scale settings (e.g., GPT-2 with $N = 50257$ tokens) is practically challenging, as storing and computing all edge weights in the transition matrix Q_t becomes infeasible. To address this, subsequent work has focused on specific classes of diffusion processes with simplified transition structures, reducing computational overhead while improving numerical stability and model efficiency.

Absorbing Process (MDLM).

Previous work on DLMs has primarily focused on two particular classes of processes. The first is the *absorbing process* (see Appendix B.1 for further details). A prominent example of that is the *Masked Diffusion Language Model* (MDLM; Sahoo et al., 2024; Shi et al., 2024), in which the authors, rather than directly parameterizing the probability ratios $\frac{p_t(y)}{p_t(x)}$, follow the strategy of Ho et al. (2020); Austin et al. (2021); Campbell et al. (2022) and instead learn the reverse conditional density $p_{0|t}(x_0 \mid x_t)$. Accordingly, they adopt an alternative parameterization $\widehat{x}_0\colon \mathcal{V} \times [0,1] \to \Delta$, which yields the following training objective for a fixed data point $x_0 \sim p^*(x_0)$:

MDLM Loss

$$\mathcal{L}_{\mathrm{MDLM}}(\widehat{x}_0, x_0) := \int_0^1 \mathbb{E}_{p_{t|0}(x_t|x_0)} \left[\lambda_t \langle \log \widehat{x}_0(x_t, t), x_0 \rangle \right] dt, \tag{4}$$

where λ_t is positive weighting function and $\log \widehat{x}_0(x_t, t)$ is element-wise logarithm of $\widehat{x}_0(x_t, t)$. Note that in the original work, the authors use the notation $\log\langle \widehat{x}_0(x_t, t), x_0 \rangle$, which is equivalent. Moreover, (Sahoo et al., 2024) show that this loss corresponds to the negative evidence lower bound (NELBO). As a result, at optimality, the learned model on whole data distribution $p^*(x_0)$:

$$x_0^* = \arg\min_{\widehat{x}_0} \mathbb{E}_{p^*(x_0)} \left[\mathcal{L}_{\mathrm{MDLM}}(\widehat{x}_0, x_0) \right]$$

recovers the target reverse conditional density $p_{0|t}(x_0 \mid x_t)$.

Uniform Process (UDLM/Duo).

Another important case is the *uniform process* (see Appendix B.2 for further details). Two notable models have been proposed in this setting: the *Uniform Diffusion Language Model* (UDLM; Schiff et al., 2025) and its extension, *Diffusion Duality* (Duo; Sahoo et al., 2025a). UDLM adopts the same model parameterization and target function as MDLM, and introduces the loss for a data point $x_0 \sim p^*(x_0)$ as

$$\mathcal{L}_{\text{UDLM}}(\widehat{x}_0, x_0) := \int_0^1 \mathbb{E}_{p_{t|0}(x_t|x_0)} g\big(x_t, x_0, \widehat{x}_0(x_t,t)\big) dt, \tag{5}$$

where the function $g\big(x_t, x_0, \widehat{x}_0(x_t,t)\big) :=$

$$\lambda_t \left(\frac{1}{p_{t|0}(x_t \mid x_0)} - \frac{1}{p_{t|0}(x_t \mid \widehat{x}_0)} - \sum_{y \neq x_t} \frac{p_{t|0}(y \mid x_0)}{p_{t|0}(x_t \mid x_0)} \log \frac{p_{t|0}(x_t \mid \widehat{x}_0)\, p_{t|0}(y \mid x_0)}{p_{t|0}(y \mid \widehat{x}_0)\, p_{t|0}(x_t \mid x_0)} \right),$$

with λ_t denoting a positive weighting function. For clarity, we slightly abuse notation by omitting the model's input in $\widehat{x}_0$, assuming $\widehat{x}_0 = \widehat{x}_0(x_t, t)$ within the expression. Duo further introduces an objective with reduced variance. To achieve it the authors first introduce distribution: $\tilde{p}_{t|0}(w_t \mid x_0) = \mathcal{N}(\tilde{\alpha}_t x_0, (1 - \tilde{\alpha}_t^2)I)$, where $\tilde{\alpha}_t$ denotes a rescaled diffusion schedule. Then they show that for $w_t \sim \tilde{p}_{t|0}(w_t \mid x_0)$ and $x_t(w_t) := \arg\max(w_t)$:

$$x_t(w_t) \sim p_{t|0}(x_t \mid x_0),$$

where $\arg\max(w_t)$ denotes a one-hot vector whose index corresponds to the maximum value in w_t. Based on this, they propose reducing the variance of the objective by substituting the discrete model input $x_t(w_t)$ with soft approximation:

$$x_t^\tau(w_t) = \operatorname{softmax}(w_t/\tau).$$

This relaxation leads to an extended model parameterization $\widehat{x}_0 \colon \Delta \times [0,1] \to \Delta$. The resulting loss function for a data point $x_0 \sim p^*(x_0)$ is given by:

Duo Loss

$$\mathcal{L}_{\text{Duo}}(\widehat{x}_0, x_0) := \int_0^1 \mathbb{E}_{\tilde{p}_{t|0}(w_t|x_0)} \left[g(x_t(w_t), x_0, \widehat{x}_0(x_t^\tau(w_t), t)) \right] dt. \tag{6}$$

Moreover, Sahoo et al. (2025a) shows that, in the limit as $\tau \to 0^+$, this loss corresponds to the NELBO of UDLM. Consequently, at optimality and in this limit, the learned model on whole data distribution $p^*(x_0)$:

$$x_0^* = \arg\min_{\widehat{x}_0} \mathbb{E}_{p^*(x_0)} \left[\mathcal{L}_{\text{Duo}}\right](\widehat{x}_0, x_0)$$

recovers the reverse conditional distribution $p_{0|t}(x_0 \mid x_t)$.

Sequence-Level Discrete Diffusion. In practical applications, we are interested in generating sequences of some length. A direct extension of the diffusion process to this setting would require modeling over an exponentially large space, resulting in a diffusion matrix of intractable size. To overcome this limitation, following (Lou et al., 2024), we consider diffusion processes defined by sparse diffusion matrices that perturb tokens independently across sequence positions. Under this assumption, the forward diffusion matrix factorizes across tokens, yielding both a factorized conditional distribution and, crucially, a loss function that decomposes as a sum of token-wise losses. As a result, the per-token objectives in equations equation 3, equation 4, and equation 6 can be directly applied to sequence modeling.

2.2 Distillation of Diffusion Models

Inverse objective. We build on the *Inverse Distillation* framework (Gushchin et al., 2025; Kornilov et al., 2025), which optimizes a *student distribution* $p_\theta(x_0)$ parametrized by a few-step generator such that the fixed teacher f^* remains optimal under p_θ. Specifically, it considers the objective

$$\mathcal{L}_{\text{inv}}(\theta) := \mathbb{E}_{p_\theta(x_0)}[\mathcal{L}_{\text{cont.}}(f^*; x_0)] - \min_{\widehat{f}} \mathbb{E}_{p_\theta(x_0)} \left[\mathcal{L}_{\text{cont.}}(\widehat{f}, x_0)\right], \tag{7}$$

where the inner minimizer is the *fake* model trained on samples from p_θ. By construction, the inverse loss satisfies $\mathcal{L}_{\text{inv}}(\theta) \geq 0$ and for $p_\theta(x_0) = p^*(x_0)$ the minimum value 0 is attained. However, this formulation does not, in general, guarantee the uniqueness of the minimizer meaning that optimization may converge to suboptimal solutions.

While such a general description of the inverse distillation was proposed in (Kornilov et al., 2025) the particular cases of this distillation were introduced before, specifically for continuous diffusion

(Score identity Distillation, SiD) in (Zhou et al., 2024b) and specifically for continuous flows (Flow Generator Matching, FGM) in (Huang et al., 2024).

Connection to this work. In *continuous* space, the model distribution p_θ is typically represented via a few-step generator and reparameterized, allowing the optimization of equation 7 via backpropagation through sampled trajectories. In contrast, extending this approach to the *discrete* domain presents several challenges, stemming from the inherent characteristics of discrete spaces. These include theoretical concerns regarding the validity of the training objective, due to the non-uniqueness of its minimizers, as well as practical difficulties related to the non-differentiability of discrete sampling. Our objective is to extend the inverse distillation framework to the widely studied class of *diffusion language models* by addressing both of these challenges, as discussed in §3.

3 Inverse-distilled Diffusion Language Models

In this section, we present our core methodology for distilling a broad class of Diffusion Language Models into a few-step generator. Firstly, we provide the theoretical foundations supporting the extension of the Inverse Distillation framework to the discrete domain (§3.1). We then address the practical challenges arising from the non-smooth nature of discrete spaces (§3.2). Lastly, we outline key implementation details (§3.3).

3.1 Theoretical extension of Inverse Distillation

Unified notations. SEDD equation 3, MDLM equation 4, and Duo equation 6 introduce different training objectives $\{\mathcal{L}_{\text{SEDD}}, \mathcal{L}_{\text{MDLM}}, \mathcal{L}_{\text{Duo}}\}$ and model parameterizations $\{\widehat{s}(x_t,t), \widehat{x}_0(x_t,t)\}$. To describe the distillation framework in a unified fashion, we use $\mathcal{L}_{\text{discr.}} \in \{\mathcal{L}_{\text{SEDD}}, \mathcal{L}_{\text{MDLM}}, \mathcal{L}_{\text{Duo}}\}$ to denote the diffusion training objective and introduce $\widehat{f}(x_t,t) \in \{\widehat{s}(x_t,t), \widehat{x}_0(x_t,t)\}$ as a parameterized function encompassing the relevant model parameterizations. Following the standard notation commonly adopted in diffusion distillation literature, we define the **teacher** model as:

$$f^* = \arg\min_{\widehat{f}} \mathbb{E}_{p^*(x_0)} \left[\mathcal{L}_{\text{discr.}}(\widehat{f}, x_0)\right].$$

Inverse distillation for DLMs. To address the issue of slow inference, we extend the Inverse Distillation framework, originally developed for continuous diffusion models, to the domain of Diffusion Language Models.

Specifically, we consider a **student** model defined by a generator $G_\theta \colon \mathcal{Z} \to \mathcal{V}$, parameterized by parameters θ, which transforms latent variables from a latent space $\mathcal{Z}$ to a discrete output space $\mathcal{V}$. The latent space $\mathcal{Z}$ is equipped with a tractable prior distribution $p_{\mathcal{Z}}$, inducing a distribution p_θ over $\mathcal{V}$ via the mapping G_θ. Then, following the logic of the continuous case, we propose to consider the following objective for the discrete diffusion:

IDLM loss

$$\mathcal{L}_{\text{IDLM}}(\theta) := \mathop{\mathbb{E}}_{p_\theta(x_0)}\left[\mathcal{L}_{\text{discr.}}(f^*, x_0)\right] - \min_{\widehat{f}} \mathop{\mathbb{E}}_{p_\theta(x_0)}\left[\mathcal{L}_{\text{discr.}}(\widehat{f}, x_0)\right], \tag{8}$$

where the second term includes training of **fake** model using the same loss as the teacher model, but with data $p_\theta(x_0)$ produced by the generator G_θ.

Theoretical validity of the training objective. However, simply formulating this objective in a manner analogous to the distillation loss used in continuous diffusion models does not guarantee the validity of its solutions. While the loss function admits a trivial minimizer θ^* satisfying $p_{\theta^*} = p^*$, the absence of a uniqueness guarantee implies that the optimization procedure may converge to suboptimal solutions. Consequently, it is essential to establish that the minimizer recovers the true data distribution, i.e., $p_{\theta^*} = p^*$. We present the following theorem:

Theorem 3.1 (Unique solution)**.** *For the SEDD equation 3, MDLM equation 4, and Duo equation 6 (in the limit as $\tau \to 0^+$) objectives the IDLM loss defined in equation equation 8 satisfies*

$$\mathcal{L}_{IDLM}(\theta) \geq \mathcal{D}_{KL}(p_\theta \parallel p^*) \geq 0$$

and achieves its minimum (zero) if and only if the model distribution matches the target distribution $\mathcal{L}_{IDLM}(\theta) = 0 \iff p_\theta = p^*$.

We give the proof in Appendix D.2.

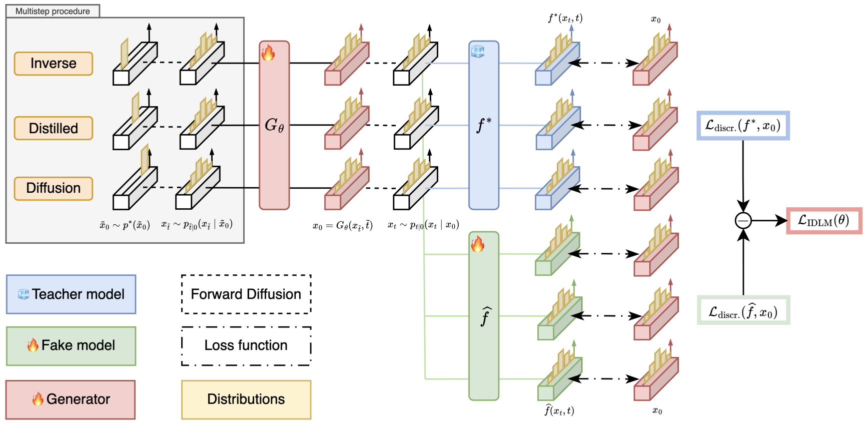


Figure 2: **Overview of our Inverse-Distilled Diffusion Language Model (IDLM) framework.** The objective is to distill a pretrained Diffusion Language Model f^*, referred to as the **teacher**, into a few-step generator G_θ, referred to as the **student**. To this end, we extend the concept of *Inverse Distillation*, originally developed for continuous diffusion models, to the discrete domain. The training procedure involves a nested optimization scheme: an auxiliary diffusion model $\widehat{f}$ (the **fake** model), is optimized using the same training loss as the teacher, but with data generated by the student G_θ. Subsequently, the generator is updated using the IDLM objective introduced in (§3.1).

3.2 Practical extension of Inverse Distillation

Although we have theoretically established the validity of the proposed inverse distillation objective IDLM equation 8, the discrete nature of the output space $\mathcal{V}$ continues to pose significant practical challenges for the gradient-based optimization. In particular, the IDLM loss contains two primary bottlenecks. (i) First, the student generator $G_\theta \colon \mathcal{Z} \to \mathcal{V}$ produces outputs in the one-hot space $\mathcal{V}$, which in practice necessitates the use of unstable techniques, such as the hard Gumbel-Softmax, thereby complicating gradient propagation. (ii) Second, samples $x_t \sim p_{t|0}(x_t \mid x_0)$ depend on x_0, and thus implicitly on θ. This introduces an additional challenge, as differentiating through the sampling operation $\sim$ is nontrivial in the discrete setting. Below we mitigate these challenges and develop a feasible training procedure.

(i) Simplex relaxation. To address the first challenge, we relax the output space of the student generator from the discrete set $\mathcal{V}$ to the probability simplex Δ, and accordingly redefine the generator as $G_\theta \colon \mathcal{Z} \to \Delta$. This relaxation enables smooth gradient flow but introduces a nontrivial question regarding the computation of the training objective, since the generator outputs no longer lie in the original discrete space. Specifically, under this reparameterization, the $\mathcal{L}_{\text{IDLM}}(\theta)$ objective becomes:

$$\mathbb{E}_{p_\theta(x_0)}\Big[\mathcal{L}_{\text{discr.}}(f^*, x_0) - \min_{\widehat{f}} \mathcal{L}_{\text{discr.}}(\widehat{f}, x_0)\Big] = \mathbb{E}_{p_{\mathcal{Z}}(z)}\Big[\mathcal{L}_{\text{discr.}}(f^*, G_\theta(z)) - \min_{\widehat{f}} \mathcal{L}_{\text{discr.}}(\widehat{f}, G_\theta(z))\Big],$$

where $z \sim p_{\mathcal{Z}}(z)$ denotes a latent variable. Since x_0 appears only within the teacher loss $\mathcal{L}_{\text{discr.}}$, we must ensure that this loss remains computable when substituting discrete samples $x_0 \in \mathcal{V}$ with relaxed outputs $G_\theta(z) \in \Delta$. To this end, we note that all teacher losses under consideration $\mathcal{L}_{\text{discr.}} \in \{\mathcal{L}_{\text{SEDD}}, \mathcal{L}_{\text{MDLM}}, \mathcal{L}_{\text{Duo}}\}$ incorporate x_0 in one of two mathematically compatible forms: (i) through scalar products $\langle \cdot, x_0 \rangle$, or (ii) via the distribution $p_{t|0}(\cdot \mid x_0)$, which can be expressed as a matrix-vector product:

$$p_{t|0}(\cdot \mid x_0) = \exp\left(\bar{\sigma}_t Q\right) x_0.$$

Consequently, each of the considered teacher losses is mathematically well-defined not only for one-hot vectors from $\mathcal{V}$ but also for any input from the probability simplex Δ. This reparameterization enables end-to-end training of the generator using standard differentiable operations, such as the $\mathrm{softmax}$ function, by avoiding the need for unstable discrete relaxation like the Gumbel–Softmax.

Case of Duo. The training objective in the Duo setting is given in equation equation 6 as

$$\mathcal{L}_{\text{Duo}}(\widehat{x}_0, x_0) := \int_0^1 \mathbb{E}_{\tilde{p}_{t|0}(w_t|x_0)}\left[g(x_t(w_t), x_0, \widehat{x}_0(x_t^\tau(w_t), t))\right] dt.$$

Here we can apply the standard reparameterization trick for Gaussians by expressing the latent variable w_t as $w_t(x_0, \epsilon) = \tilde{\alpha}_t x_0 + \sqrt{1 - \tilde{\alpha}_t^2}\, \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$ is sampled independently of x_0.

Substituting this into the objective yields

$$\mathcal{L}_{\text{Duo}}(\widehat{x}_0, x_0) := \int_0^1 \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}\ [g(x_t(w_t), x_0, \widehat{x}_0(x_t^{\tau}(w_t), t))]\ dt, \tag{9}$$

which eliminates the need to differentiate through the sampling operation itself, as the randomness is now isolated in ϵ, which is independent of x_0. This reparameterization thus enables gradient-based optimization of the Duo objective.

Case of MDLM (SEDD absorbing). In the case of MDLM, the forward transition distribution $p(x_t \mid x_0)$ does not admit a standard reparameterization. However, we can leverage the SUBS parameterization of the MDLM model introduced in (Sahoo et al., 2024, see Section 3.2.3). Specifically, for unmasked tokens (i.e., $x_t \neq m$), the following properties hold: (a) the optimal denoiser satisfies $\widehat{x}_0(x_t, t) = x_t$, and (b) $x_t = x_0$, since the forward process either preserves the token or replaces it with a fixed mask token m. As a result, in this case, $\widehat{x}_0(x_t, t) = x_t = x_0$, and the loss becomes $\mathcal{L}_{\text{MDLM}}(x_0, x_0) = 0$. Thus, all nonzero contributions to the loss arise from instances where $x_t = m$. In this case, x_t is independent of x_0, as the mask token is fixed and does not result from a stochastic transformation conditioned on x_0. Consequently, there is no need to backpropagate through the sampling of $p_{t|0}(x_t \mid x_0)$. This allows us to rewrite the loss as:

$$\int_0^1 \mathbb{E}_{p_{t|0}(x_t|\text{sg}(G_\theta(z)))} \left[\lambda_t \left\langle \log \widehat{x}_0(x_t, t), G_\theta(z) \right\rangle\right] dt,$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator. This formulation enables efficient training without requiring differentiable sampling from the discrete forward process.

3.3 Technical Aspects

Models initialization. Following prior distillation approaches that employ an auxiliary fake model during training (Yin et al., 2024b;a; Zhou et al., 2024b;a; Gushchin et al., 2025; Kornilov et al., 2025), we initialize both the fake diffusion model and the student generator using the parameters of the pretrained teacher diffusion model.

Multistep distillation. We distill a K-step student that conditions on (x_t, t) and initialize it from the teacher for stability. Full parameterization and training details are in Appendix C.

Training of fake model. The proposed IDLM loss function involves an inner optimization with respect to the fake model. To approximate this nested optimization in practice, we adopt an alternating training scheme, following prior works (Yin et al., 2024b;a; Gushchin et al., 2025; Kornilov et al., 2025). Specifically, training proceeds by alternating between 2 optimization steps.

(i) Updating the fake model $\widehat{f}$ while keeping the student generator G_θ fixed. In this step, the fake model is optimized using the teacher loss $\mathcal{L}_{\text{discr.}}$ evaluated on samples generated by the student. Taking into account the multistep training procedure, for a fixed sample $x_{\tilde{t}} \sim p_{\tilde{t}|0}(x_{\tilde{t}} \mid \tilde{x}_0)$ with $\tilde{x}_0 \sim p^*(\tilde{x}_0)$, the resulting optimization objective for updating the fake model is given by:

$$\begin{gathered} \textit{Update: } \widehat{f}; \quad \textit{Fix: } \theta \\ \mathcal{L}_{\widehat{f}} = \mathcal{L}_{\text{discr.}}(\widehat{f}, G_\theta(x_{\tilde{t}}, \tilde{t})). \end{gathered} \tag{10}$$

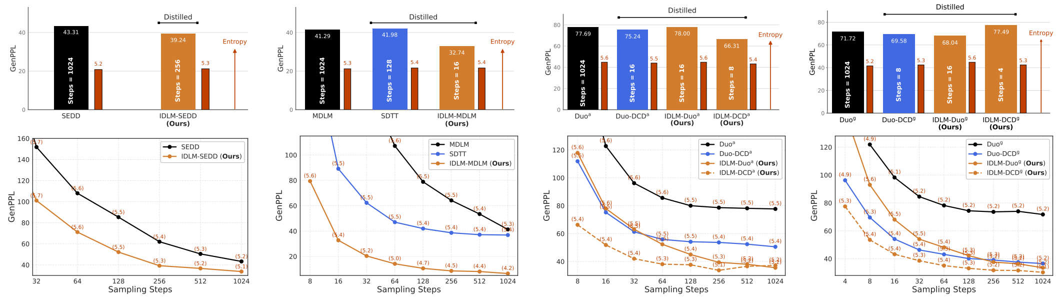
(ii) Updating the student generator G_θ while keeping the fake model $\widehat{f}$ fixed. Considering the multistep training procedure, for a fixed sample $x_{\tilde{t}} \sim p_{\tilde{t}|0}(x_{\tilde{t}} \mid \tilde{x}_0)$ with $\tilde{x}_0 \sim p^*(\tilde{x}_0)$, the IDLM loss equation 8 can be instantiated as:

$$\begin{gathered} \textit{Update: } \theta; \quad \textit{Fix: } \widehat{f} \\ \mathcal{L}_{\text{IDLM}}(\theta) = \mathcal{L}_{\text{discr.}}(f^*, G_\theta(x_{\tilde{t}}, \tilde{t})) - \mathcal{L}_{\text{discr.}}(\widehat{f}, G_\theta(x_{\tilde{t}}, \tilde{t})). \end{gathered} \tag{11}$$

Algorithm. In summary, Algorithm 1 outlines the proposed procedure for distilling a pretrained DLM into a few-step generator using the IDLM objective. A visual illustration of the method is provided in Figure 2.

4 Experiments

This section demonstrates the applicability of our IDLM distillation method across various types of DLMs. To this end, we conduct experiments using available checkpoints of pretrained SEDD, MDLM, and Duo teacher models trained on OpenWebText (OWT; Gokaslan et al., 2019). For evaluation, we follow the protocol of prior works (Dieleman et al., 2022; Lou et al., 2024; Deschenaux



(a) SEDD (b) MDLM/SDTT (c) Duo Ancestral (^{a}) (d) Duo Greedy-Tail (^{g})

Figure 3: **Performance comparisons across all benchmarks.** *Top row:* Generation quality/diversity metrics at specific step counts. IDLM-SEDD matches baseline quality with $4\times$ fewer steps ($1024 \to 256$), IDLM-MDLM achieves $64\times$ reduction ($1024 \to 16$), and IDLM-DCD further reduces to 8 steps ($128\times$) under Ancestral and 4 steps ($256\times$) under Greedy-Tail sampling. *Bottom row:* GenPPL efficiency curves. IDLM variants consistently achieve better GenPPL across all step counts while maintaining comparable entropy, with particularly strong improvements in the low-step regime.

& Gulcehre, 2024; Sahoo et al., 2024; 2025a) and report GPT-2 Large (Radford et al., 2019) generative perplexity (GenPPL) as well as average sequence entropy to measure diversity. Comprehensive results are summarized in Table 1. Additional technical details for all evaluated methods are provided in Appendix F. We also provide the uncurated text samples in Appendix H. All implementations were developed in PyTorch, and the code will be published.

Distillation of SEDD. To the best of our knowledge, no distillation methods have been developed specifically for the SEDD model. This section therefore focuses on demonstrating the effectiveness of our distillation approach in reducing the number of sampling steps required to achieve comparable generation quality of the teacher model. In the case of SEDD, only checkpoints for the absorbing-process variant are available, so we distill this version and refer to it simply as SEDD for clarity. As shown in Figure 3a, our method, `IDLM-SEDD`, consistently outperforms SEDD in terms of GenPPl across all sampling steps. Notably, IDLM-SEDD maintains high sequence entropy at low sampling steps. IDLM-SEDD achieves a $4\times$ reduction in sampling steps, accelerating generation from 1024 to 256 steps, while preserving both GenPPL and output diversity.

Distillation of MDLM. This section addresses two primary objectives: (1) demonstrating the feasibility of MDLM model distillation within our framework and (2) comparing with the SDTT (Deschenaux & Gulcehre, 2024) a leading method for MDLM distillation. As shown in Figure 3b, our method, `IDLM-MDLM`, exhibits reduced entropy at large sampling steps, indicating lower diversity in that regime. However, in the low–sampling-step regime, IDLM-MDLM achieves strong performance in terms of both GenPPL and entropy. In particular, IDLM-MDLM attains a $64\times$ acceleration for both MDLM and SDTT, reducing the number of sampling steps from 1024 to 16 while preserving comparable generation quality and diversity metrics.

Distillation of Duo. We focus on Duo (Sahoo et al., 2025a) and its distillation baseline, Duo-DCD. We distill both the original Duo and Duo-DCD (treating it as a teacher), denoted `IDLM-Duo` and `IDLM-DCD`. Results are in Figure 3c. With Ancestral sampling, IDLM-Duo matches Duo-DCD with $64\times$ fewer steps ($1024 \to 16$). IDLM-DCD yields a $128\times$ speed-up ($1024 \to 8$) over the original Duo. With Greedy-Tail sampling, IDLM-Duo achieves a $64\times$ reduction ($1024 \to 16$), while IDLM-DCD reaches a $256\times$ acceleration ($1024 \to 4$), all while maintaining comparable metrics.

5 Conclusion

In this work, we introduce *Inverse-distilled Diffusion Language Models* (IDLM), a framework that accelerates discrete diffusion models by extending Inverse Distillation to the discrete domain. To ensure valid and stable training, we establish the *theoretical uniqueness* of the objective and propose *gradient-stable relaxations* using a simplex-parameterized generator. Empirically, IDLM achieves strong performance across diverse architectures, accelerating **SEDD by $4\times$**, **MDLM by $64\times$**, and **Duo by up to $256\times$**, reducing **1024** steps to just **4**, while ***matching*** the teacher's generation quality. These results position IDLM as a robust solution for practical, low-latency text generation with diffusion models.

References

William J Anderson. *Continuous-time Markov chains: An applications-oriented approach*. Springer Science & Business Media, 2012.

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

David Berthelot, Arnaud Autef, Jierui Lin, Dian Ang Yap, Shuangfei Zhai, Siyuan Hu, Daniel Zheng, Walter Talbott, and Eric Gu. Tract: Denoising diffusion models with transitive closure time-distillation. *arXiv preprint arXiv:2303.04248*, 2023.

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

Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative Flows on Discrete State-Spaces: Enabling Multimodal Flows with Applications to Protein Co-Design, June 2024. URL http://arxiv.org/abs/2402.04997. arXiv:2402.04997 [cs, q-bio, stat].

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

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, 2019.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. Diffuseq-v2: Bridging discrete and continuous text spaces for accelerated seq2seq diffusion models. *arXiv preprint arXiv:2310.05793*, 2023.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. Olmo: Accelerating the science of language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 15789–15809, 2024.

Nikita Gushchin, David Li, Daniil Selikhanovych, Evgeny Burnaev, Dmitry Baranchuk, and Alexander Korotin. Inverse bridge matching distillation. In *International Conference on Machine Learning*, 2025.

Xiaochuang Han, Sachin Kumar, and Yulia Tsvetkov. Ssd-lm: Semi-autoregressive simplex-based diffusion language model for text generation and modular control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11575–11596, 2023.

Floyd B Hanson. *Applied stochastic processes and control for jump-diffusions: modeling, analysis and computation*. SIAM, 2007.

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

Chihan Huang and Hao Tang. CtrlDiff: Boosting Large Diffusion Language Models with Dynamic Block Prediction and Controllable Generation. 2025. doi: 10.48550/ARXIV.2505.14455. URL https://arxiv.org/abs/2505.14455. Publisher: arXiv Version Number: 1.

Zemin Huang, Zhengyang Geng, Weijian Luo, and Guo-jun Qi. Flow generator matching. *arXiv preprint arXiv:2410.19310*, 2024.

Daniel Israel, Guy Van den Broeck, and Aditya Grover. Accelerating Diffusion LLMs via Adaptive Parallel Decoding. 2025. doi: 10.48550/ARXIV.2506.00413. URL https://arxiv.org/abs/2506.00413. Publisher: arXiv Version Number: 1.

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.

Frank P Kelly. *Reversibility and stochastic networks*. Cambridge University Press, 2011.

Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. *arXiv preprint arXiv:2310.02279*, 2023.

Nikita Kornilov, David Li, Tikhon Mavrin, Aleksei Leonov, Nikita Gushchin, Evgeny Burnaev, Iaroslav Koshelev, and Alexander Korotin. Universal inverse distillation for matching models with real-data supervision (No GANs). *arXiv preprint arXiv:2509.22459*, 2025.

Laida Kushnareva, Tatiana Gaintseva, Dmitry Abulkhanov, Kristian Kuznetsov, German Magai, Eduard Tulchinskii, Serguei Barannikov, Sergey Nikolenko, and Irina Piontkovskaya. Ai-generated text boundary detection with roft. In *First Conference on Language Modeling (COLM)*, 2024. Outstanding Paper Award.

Sulin Liu, Juno Nam, Andrew Campbell, Hannes Stärk, Yilun Xu, Tommi Jaakkola, and Rafael Gómez-Bombarelli. Think while you generate: Discrete diffusion with planned denoising. *arXiv preprint arXiv:2410.06264*, 2024.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 32819–32848, 2024.

Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, et al. Starcoder 2 and the stack v2: The next generation. *arXiv preprint arXiv:2402.19173*, 2024.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787, 2022.

Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*, 2021.

Zhaoyang Lyu, Xudong Xu, Ceyuan Yang, Dahua Lin, and Bo Dai. Accelerating diffusion models via early stop of the diffusion process. *arXiv preprint arXiv:2205.12524*, 2022.

Rabeeh Karimi Mahabadi, Hamish Ivison, Jaesung Tae, James Henderson, Iz Beltagy, Matthew E. Peters, and Arman Cohan. TESS: Text-to-Text Self-Conditioned Simplex Diffusion. 2023. doi: 10.48550/ARXIV.2305.08379. URL `https://arxiv.org/abs/2305.08379`. Publisher: arXiv Version Number: 2.

Chenlin Meng, Kristy Choi, Jiaming Song, and Stefano Ermon. Concrete score matching: Generalized score matching for discrete data. In *Advances in Neural Information Processing Systems*, volume 35, pp. 34532–34545, 2022.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

Bernt Øksendal. Stochastic differential equations. In *Stochastic differential equations: an introduction with applications*, pp. 38–50. Springer, 2003.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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*, 2024. URL https://openreview.net/forum?id=L4uaAR4ArM.

Subham Sekhar Sahoo, Justin Deschenaux, Aaron Gokaslan, Guanghan Wang, Justin T Chiu, and Volodymyr Kuleshov. The diffusion duality. In *Forty-second International Conference on Machine Learning*, 2025a. URL https://openreview.net/forum?id=9P9Y8FOSOk.

Subham Sekhar Sahoo, Zhihan Yang, Yash Akhauri, Johnna Liu, Deepansha Singh, Zhoujun Cheng, Zhengzhong Liu, Eric Xing, John Thickstun, and Arash Vahdat. Esoteric Language Models, June 2025b. URL http://arxiv.org/abs/2506.01928. arXiv:2506.01928 [cs].

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

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.

Daniil Selikhanovych, David Li, Aleksei Leonov, Nikita Gushchin, Sergei Kushneriuk, Alexander Filippov, Evgeny Burnaev, Iaroslav Koshelev, and Alexander Korotin. One-step residual shifting diffusion for image super-resolution via distillation. *arXiv preprint arXiv:2503.13358*, 2025.

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

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. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. *arXiv preprint arXiv:2310.14189*, 2023.

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*, 2020b.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34:1415–1428, 2021.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Nikita Sorokin, Tikhonov Anton, Dmitry Abulkhanov, Ivan Sedykh, Irina Piontkovskaya, and Valentin Malykh. Cct-code: Cross-consistency training for multilingual clone detection and code search. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pp. 178–185, 2025.

Hannes Stark, Bowen Jing, Chenyu Wang, Gabriele Corso, Bonnie Berger, Regina Barzilay, and Tommi Jaakkola. Dirichlet flow matching with applications to dna sequence design. *arXiv preprint arXiv:2402.05841*, 2024.

Haoran Sun, Lijun Yu, Bo Dai, Dale Schuurmans, and Hanjun Dai. Score-based continuous-time discrete diffusion models. In *The Eleventh International Conference on Learning Representations*, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 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.

Tong Wu, Zhihao Fan, Xiao Liu, Hai-Tao Zheng, Yeyun Gong, yelong shen, Jian Jiao, Juntao Li, zhongyu wei, Jian Guo, Nan Duan, and Weizhu Chen. Ar-diffusion: Auto-regressive diffusion model for text generation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 39957–39974. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/7d866abba506e5a56335e4644ebe18f9-Paper-Conference.pdf`.

Sirui Xie, Zhisheng Xiao, Diederik Kingma, Tingbo Hou, Ying Nian Wu, Kevin P Murphy, Tim Salimans, Ben Poole, and Ruiqi Gao. Em distillation for one-step diffusion models. *Advances in Neural Information Processing Systems*, 37:45073–45104, 2024.

Minkai Xu, Tomas Geffner, Karsten Kreis, Weili Nie, Yilun Xu, Jure Leskovec, Stefano Ermon, and Arash Vahdat. Energy-Based Diffusion Language Models for Text Generation, March 2025. URL `http://arxiv.org/abs/2410.21357`. arXiv:2410.21357 [cs].

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

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

Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022.

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.

Mingyuan Zhou, Huangjie Zheng, Yi Gu, Zhendong Wang, and Hai Huang. Adversarial score identity distillation: Rapidly surpassing the teacher in one step. *arXiv preprint arXiv:2410.14919*, 2024a.

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

Yuanzhi Zhu, Xi Wang, Stéphane Lathuilière, and Vicky Kalogeiton. Di$\mathtt{[M]}$O: Distilling Masked Diffusion Models into One-step Generator, March 2025. URL `http://arxiv.org/abs/2503.15457`. arXiv:2503.15457 [cs].

APPENDIX

A RELATED WORKS

Acceleration of Discrete Diffusion Language Models. The high inference latency of diffusion models has prompted extensive research into acceleration strategies. A prominent direction involves hybridizing diffusion with autoregressive (AR) mechanisms. Wu et al. (2023) and Han et al. (2023) propose semi-autoregressive approaches that generate blocks of tokens in parallel, while Huang & Tang (2025) introduce dynamic block sizing with reinforcement learning to balance speed and control. More recently, Sahoo et al. (2025b) proposed Eso-LMs, which fuse AR and Masked Diffusion Models (MDMs) to interpolate between the two paradigms, enabling KV-caching for diffusion. Israel et al. (2025) take a different approach with Adaptive Parallel Decoding, using a small AR model to guide the parallel sampling of a diffusion model. Beyond architectural hybrids, other works optimize the sampling process itself. Lyu et al. (2022) explore early stopping mechanisms, while Liu et al. (2024) introduce planned denoising to dynamically allocate compute. Xu et al. (2025) propose Energy-based Diffusion Language Models (EDLM) to improve the approximation of the reverse process, allowing for fewer steps without performance degradation.

Continuous Relaxations and Flow Matching. To leverage the mature toolkits of continuous optimization, several works relax the discrete constraint by mapping tokens to continuous spaces. Gong et al. (2023) project discrete tokens into a continuous embedding space to apply standard diffusion. Alternatively, methods like TESS (Mahabadi et al., 2023) and SSD-LM (Han et al., 2023) define diffusion processes directly on the logit simplex. In the realm of Flow Matching, Liu et al. (2022) introduced Rectified Flow to learn straight-line probability paths in continuous space. This concept has been extended to the simplex by Stark et al. (2024) via Dirichlet Flow Matching. Campbell et al. (2024) further bridge the gap with Discrete Flow Models (DFMs), demonstrating that continuous-time Markov chains can be viewed through the lens of flow matching, a perspective that aligns with our unified formulation.

Distillation of Diffusion Models. Distillation is a standard practice for accelerating continuous diffusion. Techniques such as Progressive Distillation (Salimans & Ho, 2022) and Consistency Models (Song et al., 2023; Kim et al., 2023) have successfully reduced sampling to a few steps. While consistency-based training objectives have also proven effective for discrete representation tasks (Sorokin et al., 2025), distillation for *generative* discrete diffusion remains nascent. Recent approaches such as SDTT (Deschenaux & Gulcehre, 2024) and Duo-DCD (Sahoo et al., 2025a) have begun to address this challenge; SDTT applies a start-distribution transfer technique, while Duo-DCD adapts consistency distillation to the discrete setting. Other works include TRACT (Berthelot et al., 2023) and EM-based distillation (Xie et al., 2024). Most recently, Zhu et al. (2025) proposed Di[M]O to distill MDMs into a one-step generator. Also, it is worth noting that the inverse distillation framework of IBMD (Gushchin et al., 2025) has been extended to discrete-time models in Residual Shifting Distillation (Selikhanovych et al., 2025, RSD). Analogously, IDLM can be extended to discrete-time distillation of discrete diffusion models, which we leave as a promising direction for future work. However, these methods often rely on specific heuristics. Our work, IDLM, offers a principled extension of the inverse distillation framework (Gushchin et al., 2025; Kornilov et al., 2025) to the discrete domain.

B DETAILS OF CONSIDERED PROCESSES

SEDD (Lou et al., 2024) introduced a general framework for discrete diffusion models based on an arbitrary diffusion matrix Q_t, allowing the modeling of a broad class of forward processes. Subsequent works, including MDLM (Sahoo et al., 2024), UDLM (Schiff et al., 2025), and Duo (Sahoo et al., 2025a), build upon this framework by selecting specific diffusion dynamics and adopting distinct parameterizations and training objectives. In most settings, the diffusion matrix is expressed in the form $Q_t = \sigma_t Q$, where Q is a fixed base matrix defining the structure of the process, and σ_t is a time-dependent scalar that governs the diffusion rate. Defining the cumulative diffusion schedule as $\overline{\sigma}_t := \int_0^t \sigma_s \, ds$, one can further express the cumulative transition dynamics using the matrix exponential $\exp(\overline{\sigma}_t Q)$. Two principal classes of diffusion processes have emerged in practice: the *absorbing* and *uniform* processes, instantiated by selecting $Q = Q_{\text{abs}}$ and $Q = Q_{\text{uni}}$, respectively. The purpose of this section is to formally introduce and characterize these two diffusion processes.

B.1 ABSORBING PROCESS

Without loss of generality, we define the mask token as $m \in \mathcal{V}$ such that $m_N = 1$. An important special case of discrete diffusion is the *absorbing process*, in which the terminal distribution converges to a delta distribution concentrated on a designated mask token:

$$\pi = m.$$

This process is realized by choosing a specific base diffusion matrix (Lou et al., 2024):

$$Q_{\text{abs}} = \begin{bmatrix} -1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 0 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix}. \tag{12}$$

This choice induces the time-dependent diffusion matrix

$$Q_t = \sigma_t Q_{\text{abs}} = \begin{bmatrix} -\sigma_t & 0 & \cdots & 0 & 0 \\ 0 & -\sigma_t & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -\sigma_t & 0 \\ \sigma_t & \sigma_t & \cdots & \sigma_t & 0 \end{bmatrix}. \tag{13}$$

The corresponding cumulative transition matrix is given by the matrix exponential

$$\exp(\overline{\sigma}_t Q_{\text{abs}}) = \begin{bmatrix} \exp(-\overline{\sigma}_t) & 0 & \cdots & 0 & 0 \\ 0 & \exp(-\overline{\sigma}_t) & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \exp(-\overline{\sigma}_t) & 0 \\ 1 - e^{-\overline{\sigma}_t} & 1 - e^{-\overline{\sigma}_t} & \cdots & 1 - e^{-\overline{\sigma}_t} & 1 \end{bmatrix}. \tag{14}$$

In MDLM (Sahoo et al., 2024, Appendix C), an alternative parameterization of the diffusion schedule is employed by defining $\alpha_t \coloneqq \exp(-\overline{\sigma}_t)$. Under this notation, the diffusion matrix can be expressed as

$$Q_t = -\frac{\alpha_t'}{\alpha_t} Q_{\text{abs}} = \begin{bmatrix} \frac{\alpha_t'}{\alpha_t} & 0 & \cdots & 0 & 0 \\ 0 & \frac{\alpha_t'}{\alpha_t} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{\alpha_t'}{\alpha_t} & 0 \\ -\frac{\alpha_t'}{\alpha_t} & -\frac{\alpha_t'}{\alpha_t} & \cdots & -\frac{\alpha_t'}{\alpha_t} & 0 \end{bmatrix}, \tag{15}$$

and in the same way the cumulative diffusion matrix can be equivalently expressed as

$$\exp(\overline{\sigma}_t Q_{\text{abs}}) = \begin{bmatrix} \alpha_t & 0 & \cdots & 0 & 0 \\ 0 & \alpha_t & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \alpha_t & 0 \\ 1 - \alpha_t & 1 - \alpha_t & \cdots & 1 - \alpha_t & 1 \end{bmatrix}. \tag{16}$$

B.2 UNIFORM PROCESS

Another important special case of discrete diffusion is the *uniform process*, in which the terminal distribution is uniform:

$$\pi = \frac{1}{N} \vec{1}.$$

This process is realized by selecting the following base diffusion matrix (Lou et al., 2024):

$$Q_{\text{uni}} = \begin{bmatrix} 1-N & 1 & \cdots & 1 \\ 1 & 1-N & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1-N \end{bmatrix}. \tag{17}$$

Given a diffusion rate σ_t, this leads to the time-dependent diffusion matrix:

$$Q_t = \sigma_t Q_{\text{uni}} = \begin{bmatrix} \sigma_t(1-N) & \sigma_t & \cdots & \sigma_t \\ \sigma_t & \sigma_t(1-N) & \cdots & \sigma_t \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_t & \sigma_t & \cdots & \sigma_t(1-N) \end{bmatrix}. \tag{18}$$

The corresponding cumulative transition matrix is given by the matrix exponential:

$$\exp(\overline{\sigma}_t Q_{\text{uni}}) = \begin{bmatrix} e^{-\overline{\sigma}_t N} + \frac{1-e^{-\overline{\sigma}_t N}}{N} & \frac{1-e^{-\overline{\sigma}_t N}}{N} & \cdots & \frac{1-e^{-\overline{\sigma}_t N}}{N} \\ \frac{1-e^{-\overline{\sigma}_t N}}{N} & e^{-\overline{\sigma}_t N} + \frac{1-e^{-\overline{\sigma}_t N}}{N} & \cdots & \frac{1-e^{-\overline{\sigma}_t N}}{N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1-e^{-\overline{\sigma}_t N}}{N} & \frac{1-e^{-\overline{\sigma}_t N}}{N} & \cdots & e^{-\overline{\sigma}_t N} + \frac{1-e^{-\overline{\sigma}_t N}}{N} \end{bmatrix}. \tag{19}$$

In UDLM (Schiff et al., 2025), an alternative parameterization of the diffusion schedule is adopted by defining $\alpha_t \coloneqq \exp(-\overline{\sigma}_t N)$. Under this parameterization, the diffusion matrix becomes:

$$Q_t = -\frac{\alpha_t'}{N\alpha_t} Q_{\text{uni}} = \begin{bmatrix} -\frac{\alpha_t'}{N\alpha_t}(1-N) & -\frac{\alpha_t'}{N\alpha_t} & \cdots & -\frac{\alpha_t'}{N\alpha_t} \\ -\frac{\alpha_t'}{N\alpha_t} & -\frac{\alpha_t'}{N\alpha_t}(1-N) & \cdots & -\frac{\alpha_t'}{N\alpha_t} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{\alpha_t'}{N\alpha_t} & -\frac{\alpha_t'}{N\alpha_t} & \cdots & -\frac{\alpha_t'}{N\alpha_t}(1-N) \end{bmatrix}. \tag{20}$$

Likewise, defining $\beta_t = \frac{1-\alpha_t}{N}$, the cumulative transition matrix under this parameterization is expressed as:

$$\exp(\overline{\sigma}_t Q_{\text{uni}}) = \begin{bmatrix} \alpha_t + \beta_t & \beta_t & \cdots & \beta_t \\ \beta_t & \alpha_t + \beta_t & \cdots & \beta_t \\ \vdots & \vdots & \ddots & \vdots \\ \beta_t & \beta_t & \cdots & \alpha_t + \beta_t \end{bmatrix}. \tag{21}$$

B.3 Backward Diffusion Details

If the backward diffusion matrix $\overline{Q}_t$ is known, the reverse process can be simulated using numerical methods such as Euler's (Lou et al., 2024). It can be derived from the forward diffusion matrix Q_t using the following relationship (Kelly, 2011; Campbell et al., 2022; Sun et al., 2023):

$$\langle y, x \rangle_{\overline{Q}_t} = \frac{p_t(y)}{p_t(x)} \langle x, y \rangle_{Q_t}, \quad \langle x, x \rangle_{\overline{Q}_t} = -\sum_{y \neq x} \langle y, x \rangle_{\overline{Q}_t},$$

where $y, x \in \mathcal{V}$. The main challenge is to compute the ratio $\frac{p_t(y)}{p_t(x)}$, which depends on unknown intermediate distributions and cannot be computed directly.

C Gradient-Stable Training Details

C.1 Simplex Relaxation and Gradient Estimation

To address the non-differentiability of discrete data, we relax the output space of the student generator from the discrete set $\mathcal{V}$ to the probability simplex Δ, and accordingly redefine the generator as $G_\theta \colon \mathcal{Z} \to \Delta$. This relaxation enables smooth gradient flow but introduces a nontrivial question regarding the computation of the training objective, since the generator outputs no longer lie in the original discrete space. Specifically, under this reparameterization, the $\mathcal{L}_{\text{IDLM}}(\theta)$ objective becomes:

$$\mathbb{E}_{p_\theta(x_0)}\Big[\mathcal{L}_{\text{discr.}}(f^*, x_0) - \min_{\widehat{f}} \mathcal{L}_{\text{discr.}}(\widehat{f}, x_0)\Big] = \mathbb{E}_{p_\mathcal{Z}(z)}\Big[\mathcal{L}_{\text{discr.}}(f^*, G_\theta(z)) - \min_{\widehat{f}} \mathcal{L}_{\text{discr.}}(\widehat{f}, G_\theta(z))\Big],$$

where $z \sim p_\mathcal{Z}(z)$ denotes a latent variable. Since x_0 appears only within the teacher loss $\mathcal{L}_{\text{discr.}}$, we must ensure that this loss remains computable when substituting discrete samples $x_0 \in \mathcal{V}$ with relaxed outputs $G_\theta(z) \in \Delta$. To this end, we note that all teacher losses under consideration $\mathcal{L}_{\text{discr.}} \in \{\mathcal{L}_{\text{SEDD}}, \mathcal{L}_{\text{MDLM}}, \mathcal{L}_{\text{Duo}}\}$ incorporate x_0 in one of two mathematically compatible forms: (i) through scalar products $\langle \cdot, x_0 \rangle$, or (ii) via the distribution $p_{t|0}(\cdot \mid x_0)$, which can be expressed as a matrix-vector product:

$$p_{t|0}(\cdot \mid x_0) = \exp\left(\bar{\sigma}_t Q\right) x_0.$$

Consequently, each of the considered teacher losses is mathematically well-defined not only for one-hot vectors from $\mathcal{V}$ but also for any input from the probability simplex Δ. This reparameterization enables end-to-end training of the generator using standard differentiable operations, such as the $\mathrm{softmax}$ function, thereby avoiding the need for unstable discrete relaxation methods like the hard Gumbel–Softmax.

Case of Duo. The training objective in the Duo setting is given in equation equation 6 as

$$\mathcal{L}_{\text{Duo}}(\widehat{x}_0, x_0) := \int_0^1 \mathbb{E}_{\tilde{p}_{t|0}(w_t|x_0)} \left[g(x_t(w_t), x_0, \widehat{x}_0(x_t^\tau(w_t), t))\right] dt.$$

Here we can apply the standard reparameterization trick for Gaussians by expressing the latent variable w_t as:

$$w_t(x_0, \epsilon) = \tilde{\alpha}_t x_0 + \sqrt{1 - \tilde{\alpha}_t^2}\, \epsilon, \tag{22}$$

where $\epsilon \sim \mathcal{N}(0, I)$ is sampled independently of x_0. Substituting this into the objective yields

$$\mathcal{L}_{\text{Duo}}(\widehat{x}_0, x_0) := \int_0^1 \mathbb{E}_{\epsilon \sim \mathcal{N}(0,I)} \left[g(x_t(w_t), x_0, \widehat{x}_0(x_t^\tau(w_t), t))\right] dt, \tag{23}$$

which eliminates the need to differentiate through the sampling operation itself, as the randomness is now isolated in ϵ, which is independent of x_0. This reparameterization thus enables gradient-based optimization of the Duo objective.

Case of MDLM (SEDD absorbing). In the case of MDLM, the forward transition distribution $p(x_t \mid x_0)$ does not admit a standard reparameterization. However, we can leverage the SUBS parameterization of the MDLM model introduced in (Sahoo et al., 2024, see Section 3.2.3). Specifically, for unmasked tokens (i.e., $x_t \neq m$), the following properties hold: (a) the optimal denoiser satisfies $\widehat{x}_0(x_t, t) = x_t$, and (b) $x_t = x_0$, since the forward process either preserves the token or replaces it with a fixed mask token m. As a result, in this case, $\widehat{x}_0(x_t, t) = x_t = x_0$, and the loss becomes $\mathcal{L}_{\text{MDLM}}(x_0, x_0) = 0$. Thus, all nonzero contributions to the loss arise from instances where $x_t = m$. In this case, x_t is independent of x_0, as the mask token is fixed and does not result from a stochastic transformation conditioned on x_0. Consequently, there is no need to backpropagate through the sampling of $p_{t|0}(x_t \mid x_0)$. This allows us to rewrite the loss as:

$$\int_0^1 \mathbb{E}_{p_{t|0}(x_t|\text{sg}(G_\theta(z)))} \left[\lambda_t \left\langle \log \widehat{x}_0(x_t, t), G_\theta(z) \right\rangle\right] dt,$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator. This formulation enables efficient training without requiring differentiable sampling from the discrete forward process.

C.2 Multistep Parameterization

To enable few-step inference, and given that the student generator is initialized from the pretrained teacher model, we adopt the same latent space $\mathcal{Z}$ as for the teacher. Therefore, the generator is parameterized as $G_\theta \colon \mathcal{V} \times [0, 1] \to \Delta$ in the SEDD and MDLM settings, and as $G_\theta \colon \Delta \times [0, 1] \to \Delta$ in the Duo setting. To avoid notational ambiguity, we denote samples from the target data distribution as $\tilde{x}_0 \sim p^*(\tilde{x}_0)$. Under this parameterization, the generator receives as input either: (i) samples $x_{\tilde{t}} \sim p_{\tilde{t}|0}(x_{\tilde{t}} \mid \tilde{x}_0)$ in the SEDD and MDLM settings, or (ii) soft embeddings $x_{\tilde{t}}^\tau(w_{\tilde{t}}(\tilde{x}_0, \epsilon))$ in the Duo setting, where $\epsilon \sim \mathcal{N}(0, I)$ and $\tilde{x}_0 \sim p^*(\tilde{x}_0)$ is drawn from the data distribution. This design enables the generator to leverage supervision from ground-truth data samples, which can lead to improved generation quality. As a result, the generator is capable of performing multistep inference in the same manner as the teacher model, by numerically integrating the reverse-time dynamics.

D PROOFS

In this section, our primary objective is to establish the uniqueness result stated in Theorem equation 3.1 for the cases of SEDD equation 3, MDLM equation 4 and UDLM equation 5 losses. Since a rigorous formulation of the theorem requires a precise definition of the diffusion path measure, we first introduce this notion and relate it to the main text in (§D.1). Building on this connection, we then present a complete proof of the theorem in (§D.2).

Notations. Recall that $\mathcal{V} = \{x \in 0, 1^N : \sum_{i=1}^N x_i = 1\}$ denotes the space of one-hot column vectors. We let $D([0,T], \mathcal{V})$ represent the Skorokhod space, i.e., the space of functions from $[0,T]$ to $\mathcal{V}$ that are right-continuous with left limits (càdlàg). Finally, we denote by $\mathcal{P}(D)$ the space of probability measures over paths in $D([0,T], \mathcal{V})$. For any path measure $\mathbb{Q} \in \mathcal{P}(D)$ we denote its marginal at time t as $\mathbb{Q}_t$, its joint distribution at times s and t as $\mathbb{Q}_{s,t}$ and its conditional distribution at time s given state at time t as $\mathbb{Q}_{s|t}$.

D.1 CONNECTION WITH THE MAIN TEXT

Diffusion Language Models (DLMs) define a forward noising process represented by a path measure $\mathbb{P} \in \mathcal{P}(D)$. This path measure induces a time-dependent diffusion matrix Q_t, which is consistent with the forward dynamics described in Equation equation 1 of the main text. The central objective of diffusion models is to reverse this path measure in time, obtaining the exact time-reversed path measure $\mathbb{P}^*$, from which samples can be drawn to recover the unknown data distribution p^*.

In practice, the exact reverse path measure $\mathbb{P}^*$ is intractable. Consequently, it is approximated by a parameterized path measure $\mathbb{P}^{\widehat{f}}$, defined through a function $\widehat{f}$. Training proceeds by minimizing a tractable objective derived from this approximation (Campbell et al., 2022; Song et al., 2021; Lou et al., 2024). This objective is formulated using $\mathbb{P}^*_{x_0}$, the reverse-time path measure of the noising process initialized at δ_{x_0}, and corresponds to the negative evidence lower bound (NELBO) for a fixed sample $x_0 \sim p^*(x_0)$.

$$\mathcal{L}_{\text{discr.}}(\widehat{f}, x_0) = \mathbb{E}_{p_{T|0}(x_T|x_0)} \mathcal{D}_{\text{KL}}(\mathbb{P}^*_{x_0}(\cdot \mid x_T) \parallel \mathbb{P}^{\widehat{f}}(\cdot \mid x_T)). \tag{24}$$

Finally, Lou et al. (2024) apply Dynkin's formula (Hanson, 2007; Campbell et al., 2022) to show that this objective is equivalent, up to an additive constant, to the SEDD loss in Equation equation 3, under the model parameterization $\widehat{f}(x_t, t) = \widehat{s}(x_t, t)$.

$$\mathcal{L}_{\text{discr.}}(\widehat{s}, x_0) = \mathcal{L}_{\text{SEDD}}(\widehat{s}, x_0) =$$
$$\int_0^T \mathbb{E}_{p_{t|0}(x_t|x_0)} \left[\sum_{y \neq x_t} \langle x_t, y \rangle_{Q_t} \left(\langle \widehat{s}(x_t, t), y \rangle - \langle s(x_t, t \,|\, x_0), y \rangle \log \langle \widehat{s}(x_t, t), y \rangle \right) \right] dt. \tag{25}$$

Further works connected this loss with MDLM (Sahoo et al., 2024) and UDLM (Schiff et al., 2025) using another parameterization $\widehat{f}(x_t, t) = \widehat{x}_0(x_t, t)$ and matrices Q_t (see Appendix B for more details). So, that $\mathcal{L}_{\text{discr.}}(\widehat{x}_0, x_0) = \mathcal{L}_{\text{MDLM}}(\widehat{x}_0, x_0)$ and $\mathcal{L}_{\text{discr.}}(\widehat{x}_0, x_0) = \mathcal{L}_{\text{UDLM}}(\widehat{x}_0, x_0)$ for specific diffusion matrices.

We denote the solution to the loss in Equation equation 24 as $f^* : = \arg\min_{\widehat{f}} \mathbb{E}_{p^*(x_0)} \mathcal{L}_{\text{discr.}}(\widehat{f}, x_0)$. At optimality, the corresponding path measure satisfies $\mathbb{P}^{f^*} = \mathbb{P}^*$, indicating that the trained model f^* can be used to sample from the target data distribution p^*.

D.2 PROOF OF THE THEOREM

We begin by formulating the training objective in terms of the KL divergence and subsequently demonstrate its equivalence to the inverse distillation loss defined in Equation equation 8. Formally, our goal is to approximate an unknown data distribution p^*. To this end, we define a generator $G_\theta : \mathcal{Z} \to \mathcal{V}$, where $\mathcal{Z}$ denotes a latent space endowed with a tractable prior distribution $p_{\mathcal{Z}}(z)$. The generator induces a distribution over $\mathcal{V}$ via the pushforward operation $p_\theta = G_\theta \# p_{\mathcal{Z}}$. Let $\mathbb{P}^\theta$ denote the reverse-time path measure induced by p_θ, satisfying $\mathbb{P}^\theta_T = p_\theta$. Similarly, let $\mathbb{P}^*$ denote the reverse-time path measure associated with the true data distribution p^*, such that $\mathbb{P}^*_T = p^*$. The training objective is then given by:

$$\mathcal{L}_{\text{IDLM}}(\theta):\ = \mathcal{D}_{\text{KL}}(\mathbb{P}^{\theta} \parallel \mathbb{P}^{*}) \tag{26}$$

We first establish the equivalence of this formulation to the SEDD loss equation 3.

Theorem D.1 (Equivalence to the SEDD Loss). *Let $s^* = \arg\min_{\widehat{s}} \mathbb{E}_{p^*(x_0)} \mathcal{L}_{SEDD}(\widehat{s}, x_0)$, where $\mathcal{L}_{SEDD}$ is defined in Equation equation 3. Then, the IDLM loss in Equation equation 26 is equivalent to the following inverse distillation loss:*

$$\mathcal{L}_{IDLM}(\theta) = \mathbb{E}_{p_\theta(x_0)} \left[\mathcal{L}_{SEDD}(s^*, x_0)\right] - \min_{\widehat{s}} \mathbb{E}_{p_\theta(x_0)} \left[\mathcal{L}_{SEDD}(\widehat{s}, x_0)\right].$$

Proof. We define

$$s^\theta := \arg\min_{\widehat{s}} \mathbb{E}_{p_\theta(x_0)} \left[\mathcal{L}_{\text{SEDD}}(\widehat{s}, x_0)\right],$$

and denote the corresponding marginal distribution at time t by

$$p_t^\theta(x_t) := \int p_{t|0}(x_t \mid x_0)\, p_\theta(dx_0).$$

Following the approach of Lou et al. (2024), we apply Dynkin's formula (Hanson, 2007; Campbell et al., 2022), which plays a role analogous to Girsanov's theorem for standard stochastic differential equations (Øksendal, 2003) by enabling the computation of changes of measure. Using this result and also the fact that

$$\mathbb{P}^{s^*} = \mathbb{P}^*,$$

the KL divergence objective in Equation equation 26 can be expressed in the following form:

$$\mathcal{L}_{\text{IDLM}}(\theta):\ = \mathcal{D}_{\text{KL}}(\mathbb{P}^{\theta} \parallel \mathbb{P}^{*}) =$$
$$\int_0^T \mathbb{E}_{p_t^\theta(x_t)}\Bigg[\sum_{y \neq x_t} \langle x_t, y\rangle_{Q_t} \langle s^*(x_t,t), y\rangle \Bigg(\frac{\langle s^\theta(x_t,t), y\rangle}{\langle s^*(x_t,t), y\rangle} \log \frac{\langle s^\theta(x_t,t), y\rangle}{\langle s^*(x_t,t), y\rangle}$$
$$-\frac{\langle s^\theta(x_t,t), y\rangle}{\langle s^*(x_t,t), y\rangle} + 1\Bigg)\Bigg] dt$$

Moreover, it has been shown that the score function can be estimated via the conditional expectation (Campbell et al., 2022; Lou et al., 2024; Meng et al., 2022):

$$\langle s^\theta(x_t,t), y\rangle = \mathbb{E}_{p_{0|t}(x_0|x_t)} \langle s(x_t, t \mid x_0), y\rangle, \tag{27}$$

where the conditional distribution is given by

$$p_{0|t}(x_0 \mid x_t) = \frac{p_{t|0}(x_t \mid x_0)\, p_\theta(x_0)}{\int p_{t|0}(x_t \mid x_0)\, p_\theta(dx_0)},$$

and the conditional score function is defined as

$$\langle s(x_t, t \mid x_0), y\rangle := \frac{p_{t|0}(y \mid x_0)}{p_{t|0}(x_t \mid x_0)}.$$

The main difficulty arises from the term inside the $\log$, as direct Monte Carlo estimation of this quantity leads to a biased objective. To address this issue, we adopt a linearization technique introduced by Kornilov et al. (2025), which renders the objective linear with respect to the ratio

$$\frac{\langle s^\theta(x_t,t), y\rangle}{\langle s^*(x_t,t), y\rangle}.$$

Specifically, we obtain the following expression:

$$\frac{\langle s^\theta(x_t,t), y\rangle}{\langle s^*(x_t,t), y\rangle} \log \frac{\langle s^\theta(x_t,t), y\rangle}{\langle s^*(x_t,t), y\rangle} - \frac{\langle s^\theta(x_t,t), y\rangle}{\langle s^*(x_t,t), y\rangle} + 1 =$$
$$\max_{l(y,x_t,t)} \left[\frac{\langle s^\theta(x_t,t), y\rangle}{\langle s^*(x_t,t), y\rangle} \cdot l(y, x_t, t) + 1 - \exp(l(y, x_t, t))\right].$$

We note that the maximizer $l(y, x_t, t)$ of the above optimization problem admits the following closed-form expression:

$$l^*(y, x_t, t) = \underset{l(y,x_t,t)}{\arg\max} \left[\frac{\langle s^\theta(x_t, t), y \rangle}{\langle s^*(x_t, t), y \rangle} \cdot l(y, x_t, t) + 1 - \exp(l(y, x_t, t)) \right]$$

$$= \log \left[\frac{\langle s^\theta(x_t, t), y \rangle}{\langle s^*(x_t, t), y \rangle} \right] \tag{28}$$

This observation motivates the following choice of parameterization:

$$l(y, x_t, t) = \log \left[\frac{\langle \widehat{s}(x_t, t), y \rangle}{\langle s^*(x_t, t), y \rangle} \right] \tag{29}$$

Finally, under this parameterization, the loss function takes the following form:

$$\mathcal{L}_{\text{IDLM}}(\theta) = \max_{\widehat{s}} \int_0^T \mathbb{E}_{p_t^\theta(x_t)}$$

$$\left[\sum_{y \neq x_t} \langle x_t, y \rangle_{Q_t} \langle s^*(x_t, t), y \rangle \left(\frac{\langle s^\theta(x_t, t), y \rangle}{\langle s^*(x_t, t), y \rangle} \log \frac{\langle \widehat{s}(x_t, t), y \rangle}{\langle s^*(x_t, t), y \rangle} - \frac{\langle \widehat{s}(x_t, t), y \rangle}{\langle s^*(x_t, t), y \rangle} + 1 \right) \right] dt =$$

$$\max_{\widehat{s}} \int_0^T \mathbb{E}_{p_t^\theta(x_t)}$$

$$\left[\sum_{y \neq x_t} \langle x_t, y \rangle_{Q_t} \langle s^*(x_t, t), y \rangle \left(\underbrace{\frac{\mathbb{E}_{p_{0|t}(x_0|x_t)} \langle s(x_t, t \mid x_0), y \rangle}{\langle s^*(x_t, t), y \rangle}}_{\text{using Equation equation 27}} \log \frac{\langle \widehat{s}(x_t, t), y \rangle}{\langle s^*(x_t, t), y \rangle} - \frac{\langle \widehat{s}(x_t, t), y \rangle}{\langle s^*(x_t, t), y \rangle} + 1 \right) \right] dt =$$

$$\max_{\widehat{s}} \int_0^T \mathbb{E}_{p_t^\theta(x_t)}$$

$$\left[\sum_{y \neq x_t} \langle x_t, y \rangle_{Q_t} \left(\mathbb{E}_{p_{0|t}(x_0|x_t)} \langle s(x_t, t \mid x_0), y \rangle \log \frac{\langle \widehat{s}(x_t, t), y \rangle}{\langle s^*(x_t, t), y \rangle} - \langle \widehat{s}(x_t, t), y \rangle + \langle s^*(x_t, t), y \rangle \right) \right] dt =$$

$$\max_{\widehat{s}} \int_0^T \mathbb{E}_{p_t^\theta(x_t)}$$

$$\left[\sum_{y \neq x_t} \langle x_t, y \rangle_{Q_t} \left(\mathbb{E}_{p_{0|t}(x_0|x_t)} \langle s(x_t, t \mid x_0), y \rangle \log \frac{\langle \widehat{s}(x_t, t), y \rangle}{\langle s^*(x_t, t), y \rangle} + \langle s^*(x_t, t) - \widehat{s}(x_t, t), y \rangle \right) \right] dt =$$

$$\max_{\widehat{s}} \int_0^T \left(\mathbb{E}_{p_t^\theta(x_t)} \left[\sum_{y \neq x_t} \langle x_t, y \rangle_{Q_t} \mathbb{E}_{p_{0|t}(x_0|x_t)} \langle s(x_t, t \mid x_0), y \rangle \log \frac{\langle \widehat{s}(x_t, t), y \rangle}{\langle s^*(x_t, t), y \rangle} \right] + \right. \tag{30}$$

$$\left. \mathbb{E}_{p_t^\theta(x_t)} \left[\sum_{y \neq x_t} \langle x_t, y \rangle_{Q_t} \langle s^*(x_t, t) - \widehat{s}(x_t, t), y \rangle \right) \right] dt =$$

$$\max_{\widehat{s}} \int_0^T \left(\mathbb{E}_{p_t^\theta(x_t)} \mathbb{E}_{p_{0|t}(x_0|x_t)} \left[\sum_{y \neq x_t} \langle x_t, y \rangle_{Q_t} \langle s(x_t, t \mid x_0), y \rangle \log \frac{\langle \widehat{s}(x_t, t), y \rangle}{\langle s^*(x_t, t), y \rangle} \right] + \right. \tag{31}$$

$$\left. \mathbb{E}_{p_t^\theta(x_t)} \underbrace{\left[\sum_{y \neq x_t} \langle x_t, y \rangle_{Q_t} \langle s^*(x_t, t) - \widehat{s}(x_t, t), y \rangle \right) \right]}_{\text{does not depend on } x_0 \text{, so we can take expectation over it}} dt =$$

$$\max_{\widehat{s}} \int_0^T \left(\mathbb{E}_{p_t^\theta(x_t)} \mathbb{E}_{p_{0|t}(x_0|x_t)} \left[\sum_{y \neq x_t} \langle x_t, y \rangle_{Q_t} \langle s(x_t, t \mid x_0), y \rangle \log \frac{\langle \widehat{s}(x_t, t), y \rangle}{\langle s^*(x_t, t), y \rangle} \right] + \right.$$

$$\left. \mathbb{E}_{p_t^\theta(x_t)} \mathbb{E}_{p_{0|t}(x_0|x_t)} \left[\sum_{y \neq x_t} \langle x_t, y \rangle_{Q_t} \langle s^*(x_t, t) - \widehat{s}(x_t, t), y \rangle \right) \right] dt =$$

$$\max_{\widehat{s}} \int_0^T \left(\mathbb{E}_{p^\theta(x_0)} \mathbb{E}_{p_{t|0}(x_t|x_0)} \left[\sum_{y \neq x_t} \langle x_t, y \rangle_{Q_t} \langle s(x_t, t \mid x_0), y \rangle \log \frac{\langle \widehat{s}(x_t, t), y \rangle}{\langle s^*(x_t, t), y \rangle} \right] + \right.$$

$$\mathbb{E}_{p^\theta(x_0)}\mathbb{E}_{p_{t|0}(x_t|x_0)}\left[\sum_{y\neq x_t}\langle x_t, y\rangle_{Q_t}\langle s^*(x_t,t)-\widehat{s}(x_t,t), y\rangle\right)\Bigg]dt =$$

$$\max_{\widehat{s}}\int_0^T \mathbb{E}_{p^\theta(x_0)}\mathbb{E}_{p_{t|0}(x_0|x_t)}\left[\sum_{y\neq x_t}\langle x_t, y\rangle_{Q_t}\left(\langle s(x_t,t\mid x_0), y\rangle \log\frac{\langle \widehat{s}(x_t,t), y\rangle}{\langle s^*(x_t,t), y\rangle}+\right.\right.$$

$$\left.\left.\langle s^*(x_t,t)-\widehat{s}(x_t,t), y\rangle\right)\right]dt =$$

$$\mathbb{E}_{p^\theta(x_0)}\max_{\widehat{s}}\int_0^T \mathbb{E}_{p_{t|0}(x_0|x_t)}\left[\sum_{y\neq x_t}\langle x_t, y\rangle_{Q_t}\left(\langle s(x_t,t\mid x_0), y\rangle \log\frac{\langle \widehat{s}(x_t,t), y\rangle}{\langle s^*(x_t,t), y\rangle}+\right.\right.$$

$$\left.\left.\langle s^*(x_t,t)-\widehat{s}(x_t,t), y\rangle\right)\right]dt$$

To complete the derivation, we examine the SEDD loss in Equation equation 3 and observe that:

$$\mathcal{L}_{\text{IDLM}}(\theta) = \mathbb{E}_{p^\theta(x_0)}\max_{\widehat{s}}\int_0^T \mathbb{E}_{p_{t|0}(x_0|x_t)}\left[\sum_{y\neq x_t}\langle x_t, y\rangle_{Q_t}\left(\langle s(x_t,t\mid x_0), y\rangle \log\frac{\langle \widehat{s}(x_t,t), y\rangle}{\langle s^*(x_t,t), y\rangle}+\right.\right.$$

$$\left.\left.\langle s^*(x_t,t)-\widehat{s}(x_t,t), y\rangle\right)\right]dt =$$

$$\mathbb{E}_{p^\theta(x_0)}\max_{\widehat{s}}\Bigg[\underbrace{\mathcal{L}_{\text{SEDD}}(s^*, x_0)}_{\text{does not depend on }\widehat{s}} - \mathcal{L}_{\text{SEDD}}(\widehat{s}, x_0)\Bigg] = \mathbb{E}_{p^\theta(x_0)}\Big[\mathcal{L}_{\text{SEDD}}(s^*, x_0) - \min_{\widehat{s}}\mathcal{L}_{\text{SEDD}}(\widehat{s}, x_0)\Big]$$

□

Given this theorem, the uniqueness of the inverse distillation loss in Equation equation 8 can be established in a straightforward manner for the MDLM equation 4 and UDLM equation 5 cases.

Proposition D.2. *Let $x_0^* = \arg\min_{\widehat{x}_0}\mathbb{E}_{p^*(x_0)}\mathcal{L}_{MDLM}(\widehat{x}_0, x_0)$, where $\mathcal{L}_{SEDD}$ is defined in Equation equation 4. Then, the IDLM loss in Equation equation 26 is equivalent to the following inverse distillation loss:*

$$\mathcal{L}_{IDLM}(\theta) = \mathbb{E}_{p_\theta(x_0)}\left[\mathcal{L}_{MDLM}(x_0^*, x_0)\right] - \min_{\widehat{x}_0}\mathbb{E}_{p_\theta(x_0)}\left[\mathcal{L}_{MDLM}(\widehat{x}_0, x_0)\right].$$

An analogous result holds for the UDLM loss defined in Equation equation 5.

Proof. In Sahoo et al. (2024, Appendix C), it was shown that using the parameterization $\widehat{x}_0$ instead of $\widehat{s}$, together with a specific choice of diffusion matrix Q_t, given in Equation equation 13, yields an equivalence between the SEDD and MDLM losses:

$$\mathcal{L}_{\text{SEDD}}(\widehat{s}, x_0) = \mathcal{L}_{\text{MDLM}}(\widehat{x}_0, x_0).$$

Combining this result with the expression for the IDLM loss associated with the SEDD formulation, we obtain, for the corresponding matrix Q_t,

$$\mathcal{L}_{\text{IDLM}}(\theta) = \mathbb{E}_{p^\theta(x_0)}\Big[\mathcal{L}_{\text{SEDD}}(s^*, x_0) - \min_{\widehat{s}}\mathcal{L}_{\text{SEDD}}(\widehat{s}, x_0)\Big] =$$
$$\mathbb{E}_{p_\theta(x_0)}\left[\mathcal{L}_{\text{MDLM}}(x_0^*, x_0)\right] - \min_{\widehat{x}_0}\mathbb{E}_{p_\theta(x_0)}\left[\mathcal{L}_{\text{MDLM}}(\widehat{x}_0, x_0)\right]. \tag{32}$$

A similar argument applies to the UDLM loss. Using the equivalence between the SEDD and UDLM objectives established in Schiff et al. (2025, Appendix B), one can derive an analogous result for the UDLM case using another specific diffusion matrix defined in Equation equation 18. □

Having established the connection between Equation equation 26 and Equation equation 8, we now proceed to prove the theorem stated in the main text.

Theorem D.3. *The IDLM loss defined in Equation equation 26 and equivalent to Equation equation 8 satisfies the following inequality:*

$$\mathcal{L}_{IDLM}(\theta) \geq \mathcal{D}_{KL}(p_\theta \parallel p^*) \geq 0, \tag{33}$$

and attains its minimum value of zero if and only if the model distribution exactly matches the target distribution:

$$\mathcal{L}_{IDLM}(\theta) = 0 \iff p_\theta = p^*. \tag{34}$$

Proof. By applying the chain rule for KL divergence via disintegration of path measures, the IDLM loss can be decomposed as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{IDLM}}(\theta) = \mathcal{D}_{\mathrm{KL}}(\mathbb{P}^\theta \parallel \mathbb{P}^*) = \mathcal{D}_{\mathrm{KL}}(p_\theta \parallel p^*) + \underbrace{\mathbb{E}_{p_\theta(x)}\left[\mathcal{D}_{\mathrm{KL}}\big(\mathbb{P}^\theta_x \parallel \mathbb{P}^*_x\big)\right]}_{\geq 0} \\ \geq \mathcal{D}_{\mathrm{KL}}(p_\theta \parallel p^*) \geq 0. \end{aligned}\tag{35}$$

This decomposition implies that $\mathcal{L}_{\mathrm{IDLM}}(\theta) = 0$ if and only if $\mathbb{P}^\theta = \mathbb{P}^*$, which necessarily implies that $p_\theta = p^*$. Conversely, if $p_\theta = p^*$, then the conditional path measures also coincide, i.e., $\mathbb{P}^\theta_x = \mathbb{P}^*_x$ for all x, since both are defined using the same forward transition distribution specified in Equation equation 2. Consequently, we obtain $\mathcal{L}_{\mathrm{IDLM}}(\theta) = 0$. Hence, the global minimum of the IDLM loss is achieved if and only if the model distribution exactly recovers the target data distribution. □

E Algorithm

The pseudocode for the proposed Inverse-distilled Diffusion Language Models (IDLM) training procedure is presented in Algorithm 1. The method follows an alternating optimization scheme in which the student generator G_θ and the auxiliary fake diffusion model $\widehat{f}$ are updated iteratively. Both models are initialized from a pretrained teacher diffusion model f^*. At each training step, we first update the fake model using samples generated by the current student model and the teacher loss objective. Subsequently, the student generator is updated using the IDLM loss, which measures the discrepancy between the teacher and fake model predictions on student-generated inputs. A helper function is defined to sample training (x_t, t) from the data distribution and the associated conditional forward process. A visual overview of the entire method is provided in Figure 2.

F Experimental Details.

Codebase. Our implementation builds upon the original Duo repository (Sahoo et al., 2025a) and the SEDD repository (Lou et al., 2024), which serve as the primary codebases for our experiments. We extend these repositories to incorporate our proposed training framework, with the full algorithmic details provided in Appendix E.

Teacher checkpoints. We use pretrained checkpoints for the distillation of MDLM and Duo models obtained from the Duo repository (Sahoo et al., 2025a), and a pretrained checkpoint for SEDD distillation from the SEDD repository (Lou et al., 2024). Unfortunately, for SEDD, only a checkpoint corresponding to the absorbing process was available.

Model architecture. We follow (Sahoo et al., 2025a) for distilling MDLM and Duo, and design the denoising model to operate on both continuous and discrete latents. We use a Transformer (Vaswani et al., 2017) in which the first-layer token representations are computed via a matrix multiplication with the embedding matrix. For discrete inputs, we perform standard embedding lookups. In contrast, continuous inputs correspond to "soft lookups," producing a convex combination of vocabulary embeddings. For SEDD distillation, we follow (Lou et al., 2024) and use a model that only accepts one-hot vectors. For all setups we use small models with 169M parameters, specifically with 12 layers, 12 heads, hidden size 768, sequence length 1024, and dropout 0.1.

Training hyperparameters. Across all experimental settings, we employed a per-device batch size of 8, resulting in a global batch size of 512. Optimization was performed using the AdamW optimizer (Loshchilov & Hutter, 2017), with a fixed learning rate of 1×10^{-6} for MDLM and Duo, and 3×10^{-7} for SEDD. A constant learning rate schedule was used with 2500 warmup steps for

Algorithm 1: Inverse-distilled Diffusion Language Models (IDLM)

Input : Teacher network f^*;
Teacher Loss $\mathcal{L}_{\text{discr.}}$; `// Section 3`
Samples from target distribution $p^*(\tilde{x}_0)$;
Number of student iterations K;

Output : Learned generator G_θ

`// Initialize both generator and fake models from teacher model`
$G_\theta \leftarrow \text{CopyWeights}(f^*)$
$\widehat{f} \leftarrow \text{CopyWeights}(f^*)$

`// Helper function designed to support the implementation of the multistep training procedure.`
func $\text{SamplingFunc}(p^*)$:
  $\tilde{x}_0 \sim p^*(\tilde{x}_0)$
  $t \sim \mathcal{U}[0, 1]$
  $x_{\tilde{t}} \sim p_{\tilde{t}|0}(x_{\tilde{t}} \mid \tilde{x}_0)$ `// Equation equation 2, in case of the Duo we use` $x^{\tau}_{\tilde{t}}(w_{\tilde{t}})$ `(§2.1)`
  return $x_{\tilde{t}}, \tilde{t}$

for $k = 1$ **to** K **do**
  `// Update fake model`
  $x_{t'}, t' \leftarrow \text{SamplingFunc}(p^*)$
  $\mathcal{L}_{\widehat{f}} \leftarrow \mathcal{L}_{\text{discr.}}(\widehat{f}, G_\theta(x_{t'}, t'))$ `// Equation equation 10`
  Update $\widehat{f}$ by using $\frac{\partial \mathcal{L}_{\widehat{f}}}{\partial \widehat{f}}$
  `// Update student model`
  $x_{\tilde{t}}, \tilde{t} \leftarrow \text{SamplingFunc}(p^*)$
  $\mathcal{L}_{\text{IDLM}}(\theta) \leftarrow \Big[\mathcal{L}_{\text{discr.}}(f^*, G_\theta(x_{\tilde{t}}, \tilde{t})) - \mathcal{L}_{\text{discr.}}(\widehat{f}, G_\theta(x_{\tilde{t}}, \tilde{t}))\Big]$
    `// Equation equation 11`
  Update G_θ by using $\frac{\partial \mathcal{L}_{\text{IDLM}}(\theta)}{\partial \theta}$

all configurations. We adopted the log-linear noise formulation from (Lou et al., 2024) in the loss function, and applied exponential moving average (EMA) with a decay rate of 0.9999 on training.

Evaluation protocol. As observed by (Zheng et al., 2024), the GenPPL metric is sensitive to floating-point precision. To ensure consistency and numerical stability, we follow the protocol of (Sahoo et al., 2025a) and perform all sampling experiments using `float64` precision. All evaluation metrics are computed using the official codebase provided in the Duo (Sahoo et al., 2025a) repository.

Dataset preparation. Following prior work (Sahoo et al., 2024; 2025a; Lou et al., 2024), we preprocess the One Billion Words dataset using the detokenization procedure described by Lou et al. (2024) and Sahoo et al. (2024), with the official implementation available at this link. For the OpenWebText dataset, we utilize the `GPT2` tokenizer and apply a similar concatenation and wrapping procedure as in the previous works (Sahoo et al., 2024; 2025a), targeting a sequence length of 1,024 tokens. During wrapping, `eos` tokens are inserted between consecutive sequences. As OpenWebText does not include an official validation split, we reserve the final 100,000 documents from the dataset as a validation set.

Other baselines. Our primary baselines are SDTT (Deschenaux & Gulcehre, 2024) and Duo-DCD (Sahoo et al., 2025a). Performance metrics for both methods are reported using values provided in the Duo (Sahoo et al., 2025a), corresponding to the best-performing results obtained after 5 rounds of distillation, described in (Sahoo et al., 2025a). While a valid pretrained checkpoint was available for Duo-DCD, we were unable to identify a usable checkpoint for SDTT (Deschenaux & Gulcehre, 2024), which precluded a fair comparison or distillation of this model within our experimental framework.

Table 1: **Quantitative comparison across all methods.** Results are reported in terms of generation perplexity (GenPPL ↓) and sample entropy ↑ for varying numbers of inference steps. ‡ indicates that the checkpoint was obtained from the Duo (Sahoo et al., 2025a) repository and re-evaluated. ¶ indicates values directly reported from (Sahoo et al., 2025a). Our IDLM variants consistently match or outperform prior distillation methods with significantly fewer sampling steps.

	Steps	GenPPL ↓	Entropy ↑
Autoregressive			
Transformer‡ ($p = 1.0$)	1024	36.45	5.60
Transformer‡ ($p = 0.95$)	1024	18.51	5.40
Transformer‡ ($p = 0.9$)	1024	12.11	5.24
Diffusion (Absorbing-state)			
SEDD Absorb* (Lou et al., 2024)	1024	43.31	5.25
MDLM‡ (Sahoo et al., 2024)	1024	41.29	5.28
Diffusion (Uniform-state / Gaussian)			
SEDD Uniform¶ (Lou et al., 2024)	1024	99.90	5.56
Duo^{a}¶ (Sahoo et al., 2025a)	1024	77.69	5.55
Duo^{g}¶ (Sahoo et al., 2025a)	1024	71.72	5.22
Distillation			
SDTT¶ (Deschenaux & Gulcehre, 2024)	32	62.29	5.49
Duo-DCD^{g}¶ (Sahoo et al., 2025a)	32	46.31	5.38
Duo-DCD^{a}¶ (Sahoo et al., 2025a)	32	61.31	5.52
IDLM-MDLM **(Ours)**	32	20.37	5.23
IDLM-Duo^{g} **(Ours)**	32	54.05	5.49
IDLM-Duo^{a} **(Ours)**	32	63.10	5.54
IDLM-DCD^{g} **(Ours)**	32	38.57	5.35
IDLM-DCD^{a} **(Ours)**	32	42.03	5.41
SDTT¶ (Deschenaux & Gulcehre, 2024)	16	89.17	5.53
Duo-DCD^{g}¶ (Sahoo et al., 2025a)	16	54.11	5.37
Duo-DCD^{a}¶ (Sahoo et al., 2025a)	16	75.24	5.53
IDLM-MDLM **(Ours)**	16	32.74	5.42
IDLM-Duo^{g} **(Ours)**	16	68.04	5.55
IDLM-Duo^{a} **(Ours)**	16	78.00	5.58
IDLM-DCD^{g} **(Ours)**	16	43.21	5.41
IDLM-DCD^{a} **(Ours)**	16	51.86	5.44
SDTT¶ (Deschenaux & Gulcehre, 2024)	8	193.05	5.58
Duo-DCD^{g}¶ (Sahoo et al., 2025a)	8	69.58	5.30
Duo-DCD^{a}¶ (Sahoo et al., 2025a)	8	111.88	5.52
IDLM-MDLM **(Ours)**	8	79.42	5.61
IDLM-Duo^{g} **(Ours)**	8	93.00	5.56
IDLM-Duo^{a} **(Ours)**	8	117.88	5.62
IDLM-DCD^{g} **(Ours)**	8	53.55	5.41
IDLM-DCD^{a} **(Ours)**	8	66.31	5.42
Duo-DCD^{g}‡ (Sahoo et al., 2025a)	4	96.24	4.93
Duo-DCD^{a}‡ (Sahoo et al., 2025a)	4	261.82	5.50
IDLM-MDLM **(Ours)**	4	310.38	5.78
IDLM-Duo^{g} **(Ours)**	4	144.74	4.28
IDLM-Duo^{a} **(Ours)**	4	495.85	5.56
IDLM-DCD^{g} **(Ours)**	4	77.49	5.28
IDLM-DCD^{a} **(Ours)**	4	111.01	5.32

G QUANTITATIVE RESULTS

All quantitative results are presented in the Table 1.

H QUALITATIVE SAMPLES

Text generated by IDLM-SEDD (Ours) with 256 steps

positive , given that the aim is to do that .

”We knew the injuries against Fulham and Newcastle prevented us from gaining any victory in that defence , but we felt we deserved a big victory in a game which brought us back to the group stage .

”There were only two fifth division sides in the last eight games and that didn ’t mean much to ourselves scoring our 16th goal on Saturday – the Premier League came into the day in third place .

”So if we are now among the oldest 21, that would mean we should have been close to that. But given how well we performed in the last half of the season – we’m looking back on I’m not leaking the most precious ten minutes – I am very aware that the expectation here is maybe among the oldest 21 who want to get back to fourth place .

”There are never going to be plans to replace one of the bottom two or three sides. It ’s always up to us to make absolutely certain that we lose somewhere between second place in the play−off , which obviously will be done to get some prestige , rather than to produce another big win in the near future .”

England manager ’s side had hoped to continue a promotion campaign to the Premier League , but they would only go through a competition within the 21−day window, so it would be thought that the period from 1 April to 13 December would be longer than the previous period .

However, it did not appear that the weightlifters had made the top four – they went up 4−1 at Bolton in game four , with Reading going three days later to Wembley .

It brought the manager ’s hierarchy to humiliation , along with sadness at the selection of writer Chris Ellershow . ”If we had won nothing , then how have we at the end of the week?” he says . ”Well , it seems that we have created a few brolls for a bit , rather than a debate in which to resign from the Premier League – at the end of the week, we will have to have time again to take football out of the Big Three .

”So it ’s a matter of if we can survive that , but I know that whatever happens , there ’ll be a sectarian controversy in the English press , one of those those few peaks that fill . ”I’m not here just to criticise ourselves , I’m here for some arguments about football , and who loses , and then we’ll be talking about history .

”So we must focus on that and I ’m not obliged to focus on sadness , but we must hold things in mind and not be able to pretend that this is a great – if we don’t think it as a great , that it is a great , that will really be a disaster for our club .”

Text generated by SEDD with 1024 steps

that , that ' s a huge challenge .

Do you think you won't be here again ?

What do you think we want to do next ?

I think we want to win every championship with these players , and we have given all I have to say . I'm not say we were going to leave . After this I don't know until we finish there soon . I know and I don't know that we will rush forward . In our view , we won't give any assurances that we are going to stay . We would have been a long in charge if we hadn't paid that honour .

I decided to leave because we won some good games , there are plenty of great games there and we can't back away from anything . But given the circumstances with the squad , I don't think it ' s a little bit close to what we've had in the past .

Is there one thing positive or one negative ?

I don't think that things are that way off . You've got that the club has given us a lot of assurances and they will be very pleased at us though we still think it ' s very important to see action in the field and on the pitch .

And we have experience in the whole squad . That was a lot from the previous manager . I'm just not sure now with the opportunity was handed to myself that I said we'd have to come here and do that again . I hope we can see some progress . I am not saying that I should otherwiseate my football career . I am saying that Everton should spend many years himself and still be a good coach .

I just like to look this way at it if I resign or leave or I am quite unhappy , and sometimes when you don't enjoy retiring you go wild . I say , in this case , that by leaving sometime that day or before my next job , you are nothing less than a good coach and I do not want me to let you retire .

So these players are very hurt , and they will miss a pretty long time . But everyone thinks that time is right for the players . So it is time to show them the world that the team is now ready .

The players have been incredible to you .

I'm excited about the players because I think they should be . We were also brought to victory against Germany in the Euro tournament with the help of major players and I think that ' s one of the very best things we've ever been through .

The last year we've done a lot of testing , against China and Japan and Germany in difficult time . And so , and after that , it will be good to have this experience and these players mature .

Text generated by IDLM-MDLM (**Ours**) with 16 steps

At this time, the mainstream media is evidently understandably upset – at the same time they are still apparently past that to clearly justify any further measures taken on the right of burying the right–wing line in the sand

Meanwhile, in the wake of the angry remarks on the media from Mal MacDonald, Mackenzie King and Ramsay MacDonald, Sam Kicks Radio decided to take a shot on Aslockel, claiming that he was only ever learned how get away with a of "murder" when he was finally forced to admit it.

(Im going to As Mal MacDonald) It's only learned to get away with lots of something bad which I'm not going to get away with anymore. (Im going to Mal MacDonald) I want to get over it... (Mal MacDonald) What does that mean for you?? @MalHedge – MackApple (@sam_apple) October 24, 2013

I'd like to clarify some points of the post I'd written on Aslockel in my previous post on the subject.

In a thoughtful speech at the South End of Pol T in September 2006, David Barton eloquently claim that the country had had to introduce a rigorous Birther curriculum. At this time, I had also been struck by some of the material comments on the Birther website. While a Birther is dismissed as a low–level conspiracy theory, it is a special sort of nonsense that is only really useful only in the context of the situation at hand where what may be occurring functions as pure, benign stuff. An understanding attempt to make of this situation is more valuable than any sort of theory – a proper understanding of how the world works is a kind of understanding of how it is necessary for any kind of worldview to be correct.

Fortunately, I have the opportunity to elucidate frightfully large parts of this problem. I'll start with the premise that Birthers are not genuine conspiracy theorists, but whereas most of us individuals have suffered from a form of terminal illness, they probably only experience something of serendipitity that looks like helplessness or despair. Because these situationes are no clearer than the facts themselves – in other words, because the facts are as clear as they are, there is no reason to assert that they are delusional, or even a paranoid anist in the nuts.

At the same time, there are quite a few reasons to be skeptical in this regard. First, A Birther can easily be described as a genuine, sinister, black power fact that is routinely represented on the media – let alone by someone like Omar Khadr, or Jon Favreau, James Clapper, Nick Wray, Jonny Mitchell, Jim Edwards, or any other conservative. Aslockel is a far cry from the typical Birther in knows, so this kind of description is perfectly reasonable.

Second, it has been confirmed by the mainstream media today thatAslockel is considered a real conspiracy theorist – which is the equivalent of being considered a bona fide conspiracy theorist. So, any opportunity for the scent of extremism or fanaticism in the matter to take on an official term in the ultra–conservative movement has an barely audible slant.

Text generated by MDL with 1024 steps

```
manufacturers, Google claims that Android is its own operating system
and belongs to everyone else, basically saying that ChromeOS is
effectively competing and could soon be replaced by more complicated
cloud services like what we have with Samsung and Mozilla. Just hours
ago Samsung and Apple announced that Samsung was seeking "6 million for
best feature" package, nobody seemed interested in either. More interesting
news for everyone is Amazon offering a proprietary retail platform called OTC.
Although unknown for a long time now, Google is quite unknown and
as we learn much more about Cyanogen we see a number of options available.
Several business applications on mobile devices or devices bring many direct
interaction options.

What? A couple of options for running CyanogenOS on a PC: The first step
of the way for CyanogenOS is a built-in option for a first time that
would give users a major OS update experience, stability updates and
a CyanogenOS 1.1, all for Google's Nexus 7 series. The next option, which
runs entirely independently from the hardware UI hardware mentioned
above is for the Nexus 7. On other smartphones, CyanogenOS will run
for a day each for maintenance. This would be a hands down experience
on our device and in our experience the OS being hands free, is
not unique. Any changes would be executed in a methodical and easiest way.

This is possible: The operating system never changes background or
notifications, but it always multiplexes through your whole screen to make
the device clean up until the device has no much else to do. This is
also incredibly easy, if your phone is always busy during any hours of day.
Dragging the home button (the notification button) across the screen makes
it much easier to let an app run or give another additional button to toggle
off, sliding off makes the multitasking easier but unintentionally
slows up the app to spin when the background's background is passed in.
So this makes a switch to music and movie playback in your day difficult.
Even there, even when the app is paused, the user becomes distracted.
The next option allows more consistency in the day by removing one version of
apps; they allow the user to see each app with a very small minimum of
background pixels while requiring you to minimize the exposure of
that app, hence distracting the user from other apps.

So much is about this because the next option is the control point. A sudden
multitasking would confusingly eliminate the homescreen but would provide
it with immediate priority while giving the user without also experiencing
time to reset it. Even if we never had a user interface to switch to, let
alone merge a background background through the Android framework, new
possibilities would be endless. The main technology behind these designs are
first integrated into CyanogenOS itself.

In summary, CyanogenOS always runs with minimal lag while making it fun
to see a single notification with plenty of background background.
The downside of multitasking, is super light; there has no need to have
any extra total control over resolution in the background. The benefits
are welcome, along with are a heavily integrated user interface are
still a little complex.
```

Text generated by Duo with 1024 steps using Ancestral sampling.

society. Such a development in terms of this field would be cultural literacy. The Germanic literature is based on a set of cultural expectations, while the English is largely freighted with fiction. Thus, it would be a useful contribution to this field for a shift away from the vernacular at which interesting problems translate back into real, everyday things, like the Philadelphia problem area.

But while one critical component of this theory called "social anarchism" lends strongly to that since it claims its main goal is a la-capitalist model, it also can surely use this same perspective to understand the way forward with world societies. Other terminology, "operational anarchism," also offers ideas, about ways for societies to develop which reflect the need of community within geographically geographic invincibility ways which different regions or states will behave differently (eventually) from existing ways. Anarchists' traditional way of imposing 'monopoly' based on specific epithets, but equally within democratic constitutions will be to explain how people who object to existing practices (free market or royalist), are treated for whatever reason. In doing so that such exportation rights can always be limited to special interests, which might see redistribution of wealth to the few and consequently diversion of resources available in society for benefit of the working class.

Only conceivable if your greatest strategic goal is improving the "what everyone is having and exploring new social means to profit from market-based systems. For capitalism, for example, no market value ownership is sustainable. there are easier than harder methods for actually privatising the market beyond the corporate run; it is like reorganising that whole system of production and subsidisation etc. that results from efficiency and efficiency to go towards individual's consumption. The products on our planet are ultimately going to go capitalists!

Moreover, while market sovereignty requires total expropriation, it is absurdly convenient for the left and the right to do this sort of approach without our own principles and attitudes driven by experience, tradition and fact, when all we have to do is be successful and we have to deliver action.

However, we can never escape revolt from what we are masters of and from addressing them through systematic redistribution of wealth. In a different sense, if people fall victims in repeating history to learn something rather than trying to learn something else, you better stop repeating it in order to stay stuck. Anarchist ideologies, however utterly justified we are, may be even more powerful if disguised.

Text generated by Duo-DCD with 16 steps using Ancestral sampling.

```
for My favorite cuisine is the offerings are pretty varied and it changes
everyone from regular up to country and from apprentice to guest the
cooking is really easy to! Some hate about English like some of
the okay local dishes, but these people would be something that was
fine with Jesse being around when he's to the hotel. The Southern
Japanese Thai has is something specifically geared towards drum and
something to watch, what's more it is excellent having to visit someone
out there and check out food that is fabulous at the New Village. Oki
yama is fantastic, but mixed with candies like Oreo and Chipotle! people
can have their own ideas on how to order that in the Seattle and the
Seattleements the

Where you really get to try something real at this location is if you line up
a shepherd's Chinese budkou deli. Another staple we associate with
the Chinese are more traditional recipes but it has one that displays a
shocking sakura look, a mint-baked-Japanese Saison which is garnished somewhat
a with the Buddha's Green Book. You can also pick up a Roasted Salmon called
Mykurop in the Loco Fish Bowl in our

Sabo Oreoppa is a surprising, exploring Tibetan Buddhist food that you
can do with anywhere or every country and really her it tastes best in
China and makes a more sense version is best in the United States. Something
different or weird to try to the Chinese Village somewhere.

Location: Louisville

I ended up offering some pics and showing how I got down. An animated gif
showing the menu changes depending on the day.
http://images.groupon.com/images/1225c390ff6043f2cey/jsp 800/1e3678616

The economic data makes clear plenty of optimism all the more,
but also it gives us some of the headwinds too.

What we thought are the biggest surprises expected from the results.
So here are those, from an analysis conducted by PolitiFact:

Out of the 26 companies to offer guidance for the first quarter,
Tiger Woods marked good news with a gift to the late Macbaren,
a banker, who did not quite call Buffett-like and was given 1.2 percent
to help Macbaren the man save $8.7 billion. Tiger shares make 2.85 per share.

News Corp forecast its outplacement estimate to be $6.4 billion, after
insisting that the figure was an $8.8 billion decline. The stock reported
the lowest expected result, and trading less than a dollar in the trading of
373 bull.

The 500s continued to show some gains, but some of the records highs intact
on day 1. The forecast suggests that price drop is slight on the downside.
Notably, the market is expected to see softer sentiment towards the end
of the year, according to the forecast viewers in ETF.com.
```

Text generated by IDLM-Duo (**Ours**) with 16 steps using Ancestral sampling.

```
Dubai or elsewhere, then a community starts utilizing and optimizing
their services.

"I think it's important to worry about better trans payment and credit
security. It's important to worry about better integrated trading of
those metrics -- there's a better way of reporting them to a community,"
says Yoghamad Villadin, professor of regenerative engineering at McMaster
University.

Population

The resident range has grown to 7.4 million in the UAE's next 25 years.
By contrast, 5.5 million live people in the city's future.

2.2 million people are growing in Dubai (Tarah.) 2 million resident people are
now growing in Dubai; while while 1.8 million live people have grown in Dubai,
only 2.6 million people have grown in Dubai.

These population figures stress project transparency and project effectiveness.
The Government of Palestine is undertaking significant extra efforts
to reduce their real-estate inventory by considering risk exposures and acting
on an internal audit of the bank accounts of 825,000 people who are living in
Dubai. (Qaisudi Con.) "From this May through September, 2017, the Government
is undertaking extra efforts to tackle rising property prices," says
Mohammed al-Saih, vice-chairman of the Marakutta Federa. This provincial
federation connected to various provinces, includes 825,000 people
representing an average of 6 and 7 percent.

Arab philanthropy and international business deals develop

Arab commerce with Dubai. Men and women invest more in Dubai than
normal people. Tel Aviv and Telephone jointly form the Gordi Investment
Corporation, a dedicated group in Dubai, which is subject to the policy
of the governor. The directors finance Dubai's companies, a new
publicly-traded group called Natlana Karch. These directors also endear others
with the endearments al al-owners have collectively accumulated
during their lives.

Temporarily traded companies with Arab investor groups.
Arabs enjoy higher standard of living.

Over 5 million Arabs live in the mutualistic property group Gordi Investment
Corporation. This group is financed by Dubai's, in part, a new publicly-traded
group of companies under the name Natlana Karch. Between 1894 and 2000,
the chairman of the majority owner of the UAE's second-largest engineering
conglomerate is involved in the directors of Gordi Investment Corporation.
Gordi Kuwarching The chairman of the conglomerate provides UAE with electronic
subsidiaries and assets totaling 1.594 million in physical assets, and
1.595 million in physical assets.
```

Text generated by IDLM-DCD (**Ours**) with 8 steps using Ancestral sampling.

1959 – one of the 21st century parents Deepak Chopra – to become one of the first scientists to critically acclaim.

Why I matter?

I'm (mostly) an economist; for me, I'm a couple of co-creators of the Culture of Socialization and I'm a Simple Person.

Fact: I'm born seven decades ago.

A computer is's enhanced with incredible energy by itself.

Economics are long milliseconds. In fact, a computer is likely going to last forever before being able to do anything in energy. For example, as long as a window is opened in the ozone, and shuts down in 1337 milliseconds, it will be able to do everything in energy. In other 21-st-century terms, it will be the same Barry Zahn opened and shut down in seven milliseconds.

However, things have changed.

There is a new computer, s equipped with incredible energy.

Why the this: Armed with almost billions of computers, a computer alone can cause the manifestation of harassment.

, hence the shock and the shock shock.

Thankfully, Breitbart editor Milo Yiannopoulos will no longer suffer from the shock shock opting to join a coalition of Democrats and Republicans who backed Donald Trump in a comfortable presidential race. In an attempt to keep his job, Breitbart'
s incidence
of harassment has since, according to the study, was down by a factor of one.

The downside of this is that it is very expensive, potentially forcing people to deal with it for what it may lead to those who are less abusive to try to succeed. Posner and Yiannopoulos are now vulnerable because they will have a real trial against each other in unofficial elections in 2018.

Sure,, in the meantime,, Liberal senator Nic Mathor will make some mistakes in his comeback, but it is also hard to say that at the writing of this article, Nic Mathor is close to a nomination, for he has lost the election.

This Channel 4 documentary has overwhelmingly revealed that the head of an undercover police-surge of a prominent participant in anti-immigrant protests at a street protest at Cenbaus Square in the capital, has to and stay inside a cell.

Text generated by Duo with 1024 steps using Greedy sampling.

the United number, medical are today new many applications and patents are being struck down under the C&C Cosmetology Act for Patent Devices. However, the patent does not include any more than medical devices, including allable and other the described etc. (Nor does not cover any patent invention other of as the above holder. With regard to the foregoing, readers also note that there is any significant change in the size or quality or functionality of any part of the patent or to any modification<|endoftext|>

Beveraging the patent. the patent will be freely available for use for an patented or modved medical device<|endoftext|> This means that we not only no items to the patent should need be amended<|endoftext|>

Section 10.1 of the patent says:

Contrary to the unsupported and unsupported claims mentioned above, any modifications and the requirements is within the scope of the original portion of the as itself. The text of the text is clarifies what it means, and it states that it is only for the use for medical devices. We believe that medical invention may be made outside of the text. For the the or and BY theusers, most as Sec. 10.1/1201, the protection has also included under \u00a7 18.3 of the patent. For the other nature of the invention of the extent of the or liability (i.e. the nature of the invention to be the intended use of the medical device, the discussion in S. Ppt. 1. To to the extent that the patent antileptic meets under \u00a7 10.1, see the above discussion reasons how the antileptic agrees with the patent. If there are any difference in opinion we that you refer to following Google below below of the testimonials that agree with this patent. No explanation is required. Also.

SEC. 3. THE IGNERGY

Patent review will be conducted by those who are in any other field in which the patent is an cued. This will be provided that a same independent examiner, who can determine its merits of that the new medical device would be superior to the Conceending Medical device, conducts the examination of the patent ver application. The review will allow the examiner to evaluate the validity of the invention and the other factors in the field of invention, such as the health and intellectual property. The examiner is be need be considered the validity of the original. (2. 1.1).

The party conducting the review shall be used for any copy of the part in copy of the patent<|endoftext|>\n\nWith the above, the good here is that an independent review of the patent we is put forward by an independent party will be considered as appropriate. The invention in the invention, and the other basic concepts described in the invention and and,, be for the in the medical process referred to as \u201cbiological medicine, long term, future of medicine\u201d.

Text generated by Duo-DCD with 8 steps using Greedy sampling.

```
modern of craft beer, want to fill growlers with the line, \u201c
Excuse me to give you the beer that can want\u201d which is to \u201cmake it
my own beer.\u201d Pints of beer tripled in 50 states between 2000 and 2014
and so in the local market. It is simple: The more that make the same and
out of the IPAs more more spread the beer. This off with misleading
advertising and pushes push states with high and.2s for ways to craft beer.

This pissed off the guy who has created, beer who exists and who tries to get
pushed around when he says that craft beers need to made \u201cfor freshness,
\u201d a annual slogan says, \u201cWe have a craft beer foundation craft beer.
\u201d The punchline does not lie. That foundation was based in Wisconsin.
Craft beer from craft beer was on the only in Wisconsin and Wisconsin and they
beer. After Wisconsin tightened it, however, they now saying They have to
grow about 5 to 50 percent of craft beer industry, Time magazine has
reported. Some of the worlds's the craft breweries, due to the state state law,
that require them to sell craft breweries to sell and leave to pay the state
tax.

Start a brewery a home craft medium.

Tax the regulations. Excise companies figure out how to make craft beer and
they sell the beer for review by those government else they will be fired
craft their companies. This this, beer companies can still get more jobs,
tax breaks, be licenses. This is the way of life beer, however,
to get a license you first the them before resort to to overreach to the law.
[5] Back to beer.

Morerict laws on the beer:

This summer in 2014, young women launched the Wierlager beer in local states.
On social media, some of the women and Twitter women posted it with the
bikini, their popularity increasing the popularity increased. It has more
hastyking in California, California, Colorado, Oregon and Oregon.[5]
In some states, this comes up because
the hops are grown
in the hops, but it isn not too stifling in rural Dakota.[6] edit: Tont,
Craft Brewing Company, Columbus Brewing co-owner. Other breweries owned:

See also [ edit ]

Further reading:

Further reading [ edit ]
```

Text generated by IDLM-Duo (**Ours**) with 16 steps using Greedy sampling.

Strong performance in the NBN Co Powerplows will drive costs driven by lower profit margins for SolarIS, said Paul McCormick, the chief economist of Business Economics.com.

The NBN Co network will drive costs to near-power parity in early 2017, said Paul McCormick, chief economist at Business Economics.

Strong growth of the NBN Co network in Telstra Powerplows will drive costs driven by 7.5 per cent in real-time costs.

High average business premises with NBN Hubs uptake was $10 between Q1 2013 and Q4 2013, noted YKB.

Rural premises with low NBN Co uptake uptake resulted in $8 for Q4.

Low average business premises with NBN Hubs resulted in $5 for Q1 2013 and Q4 2013, noted YKB.

Business premises subject to network costs resulting in $26 million for the initial commercial model and $26 million for the end of Q4 2013, said YKB, explaining the cost pressures on business premises.

If SolarIS consolidate as its supplier supplier, rather than replacing SolarIS, this could be the price certainty that NBN Co will have, Ms. Cahill said.

Longeesplows are expected to lose more jobs than in the Australian market, said Mr. Georgening.

In New York (NYC)- Market, the lowest cost New South Wales community is expected to lose fewer jobs compared with the growing number of South Australian communities.

Mr. Georgening said longeplows communities are expected to lose more jobs than the New South Wales, who with 80 per cent of Australia where wages being stagnant.

Legislative wonk support for NBN Co is running thin and bottlenecks remain high. Meanwhile, pessimism support for NBN Co has diminished and many investors are frustrated by the lack of progress towards progress on upgrading ADSL.

Telesplows are expected to provide Australian businesses with moussows and cassoullas of telecommunications and telecom technology knowledge.

Text generated by IDLM-DCD (**Ours**) with 4 steps using Greedy sampling.

a candidate for an interview. This device is not the only way to use features in this article. J.R. Security, I; people who often choose to conduct an interviews in a secure environments are attracted by Peemearn, where, the device features a two-way monitor. and This is an ideal candidate for an interviewee who can adapt to a combination of security and a DC-led operating systems \u2014because of U.S. users who regularly use a secure monitor, J.R. would be ideally suited to use. Conferior Security, I; This is an ideal candidate for an interviewee for an job interview who will be able to operate, combat system that is effective \u00c57% to a stable UK government, which matters greatly for the benefit of the economy on the whole.

Second, January 16th 11/7/6.

3. Here's one more a thread was posted to. To start,; this article, author \u201cAngareth Walker,\u201d had the goal to run a referendum, one that people have submitted to, while the American rubber stamp dominates the legislature going forward. However, this is different. And it's not a referendum that hasn't been held.

For the European Union, you can see. IG LLC posted a thread to address individual David Boyie, stating that 4 regions are represented, and that, \u201cAngareth Walker: How Britain Regains Great Greatness, \u201d in the referendum, received 8 votes.

Transcript,; September 29th. First, this article (s' the\u201d 7/7) "states the concept of effective combat, is applied to in all parts of a region and is to succeed in success in that region," as individual Guy Baikie said on the Real Time programme. It seems to be that a great military that operates to the top fits the mold that people have applied to and operates within the legislature to be a fatal problem.

Second, January 16th 7/7/6. A message that was sent to this thread, at the time of running, \u201cgoal is to secure UK referendum, \u201d not to mention, that the American rubber stamp has gained traction and prominence in the years going forward, and it's likely that it will do well.

Pritchard Street LLC posted a link to this article (see below) to the thread that has been copied to-date.

Here's more:<|endoftext|>Feels like things that do not come up, the Polish defamation defamation/freedom lovers, and Eurospectas-profit charity have taken this moment to to the U.S.; Queen Ist-tzee-noh, Lip Sync Battle, the story of an international ice cream romance; the majority of the country's finest artists, backed by non-Polish artists; and Curly.com.au/debate-architecture.com.au have marked the postage stamp of the U.S.