
CRoCoDiL: Continuous and Robust Conditioned Diffusion for Language

Roy Uziel* Omer Belhasin* Itay Levy Akhiad Bercovich Ran El-Yaniv Ran Zilberstein Michael Elad
NVIDIA

Abstract

Masked Diffusion Models (MDMs) provide an efficient non-causal alternative to autoregressive generation but often struggle with token dependencies and semantic incoherence due to their reliance on discrete marginal distributions. We address these limitations by shifting the diffusion process into a continuous sentence-level semantic space. We propose *CRoCoDiL* – Continuous and Robust Conditioned Diffusion for Language – a unified fine-tuning approach that jointly trains an encoder–demasker architecture, grounding the MDM demasking in continuous latent representations. This leads to the formation of a novel autoencoder in which the decoding is obtained by an MDM algorithm. Relying on the same framework, we proceed by introducing two *unconditional* text synthesis algorithms: Continuous-Then-Discrete (*ConThenDisc*), a hybrid-diffusion approach that first generates latent representations in continuous space and then decodes these to tokens via an MDM, and Continuous-Within-Discrete (*ConWithinDisc*), a multi-diffusion strategy that refines latent representations throughout the discrete sampling process. Experiments using LLaDA show that our methods achieve superior generation quality and more than $\times 10$ faster sampling speeds in an unconditional setting.

1. Introduction

Diffusion-based alternatives to autoregressive large language models have been drawing much attention recently (Li et al., 2022; Yi et al., 2024). Such methods encompass an appealing potential to break the causal, one-token-at-a-time, paradigm of autoregressive machines, with the general hope to lead to faster and improved quality text synthesis. The main challenge in bringing diffusion models

*Equal contribution. NVIDIA. Correspondence to: <{rouziel, obelhasin, itlevy, abercovich, relyaniv, rzilberstein, melad}@nvidia.com>.

to text is the evident gap between the continuous formulation of classical diffusion algorithms and the discrete nature of language (Lou et al., 2024).

Earlier work addressed the discrete-continuum gap in a wide variety of ways; Among these, the commonly used ones are based on *masked* diffusion models (Sahoo et al., 2024; Nie et al., 2025; Ye et al., 2025). This breadth of algorithms relies on a forward degradation process that masks tokens gradually until the whole sequence is masked-out. Text generation is based on a reversed process, in which a demasker iteratively revives tokens, constituting the *Masked Diffusion Models* (MDMs), such as MDLM (Sahoo et al., 2024), LLaDA (Nie et al., 2025), Dream (Ye et al., 2025), and their many followups, e.g. (Arriola et al., 2025a; Wu et al., 2025; Liu et al., 2025b;c).

MDMs rely on a demasking model that is trained on partially masked sequences to estimate discrete logits for the missing tokens, representing one-dimensional marginal distributions that lack information on statistical cross-dependencies between tokens. When sampling from these logits, revealing multiple tokens in parallel necessarily produces flawed samples that degrade generation quality (Liu et al., 2025a). Nevertheless, as synthesis speed depends on parallel token sampling, existing algorithms compromise speed for quality.

Another, related yet different, weakness of MDM algorithms has to do with their core *modus-operandi* of constructing the generated text by sampling individual tokens (separately or jointly) sequentially, and such that they are committed to be part of the final sequence. While appealing due to its resemblance to the autoregressive strategy, having no global guidance to drive this overall synthesis, MDM necessarily struggle in forming coherent eventual sentences.

In this paper we propose a novel extension to MDMs that addresses these limitations. Our approach operates in the continuum, using a continuous diffusion model to generate sentence-level semantic representations, while the MDM algorithm serves as a decoder translating these latent vectors into token sequences. This way, the burden of capturing long-range, cross-token structure is shifted to a lightweight classical diffusion in the latent space. This representation is then used to guide the MDM for token decoding, enabling effective multi-token sampling per step by yielding better

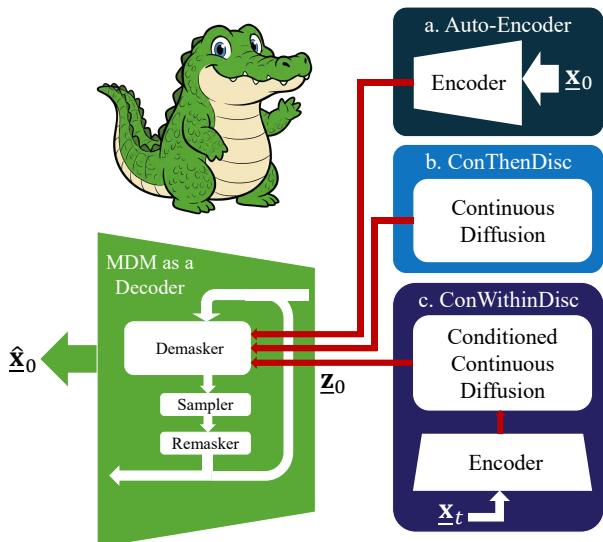


Figure 1. The *CRoCoDiL* framework: Building on a learned encoder of text sequences and a demasker guided by this continuous representation, we introduce (a) an autoencoder and (b,c) two text generation algorithms, *ConThenDisc* and *ConWithinDisc*. A regular MDM serves in all cases as a decoder that converts the latent z_0 into a sequence \hat{x}_0 . The text generation algorithms rely on learned diffusion models that operate in the representation domain.

efficiency-quality tradeoffs in text synthesis. We name this methodology *CRoCoDiL*: Continuous and Robust Conditioned Diffusion for Language.

Building on this framework, we introduce a unified encoder-demasker training scheme that encodes sequences into latent representations for effective token decoding. We then present two text synthesis algorithms: (1) Continuous-Then-Discrete (*ConThenDisc*) that generates embeddings via continuous diffusion and uses MDM to decode the latent vector into tokens; (2) Continuous-Within-Discrete (*ConWithinDisc*), that updates the guidance vector during the demasking steps using a continuous diffusion trained to recover valid latent vectors from partially masked sequences. We emphasize that the proposed algorithms are focused on unconditional text generation, leaving conditional synthesis across benchmarks for future work.

We conduct an extensive experimental study using LLaDA-8B (Nie et al., 2025) as the base MDM and Qwen-embedding-0.6B (Ren & et al., 2025) as an initial encoder, all jointly retrained with our decoder-demasker framework. We first validate the effectiveness of the continuous guidance for MDM by autoencoding, demonstrating faithful reconstruction. We then evaluate our two proposed algorithms for unconditional code synthesis, showing that our methods achieve much faster sampling without quality loss.

To summarize, the following are the main contributions of this work, as depicted in Figure 1:

- We propose *CRoCoDiL*, a framework that guides discrete MDMs using a continuous, sentence-level semantic guidance, bridging the gap between global coherence and local token dependencies, thus enabling faithful parallel token sampling.
- We introduce a general purpose autoencoder that maps accurately sequences to the continuum and back, leaning on MDM as a decoder.
- Consequently, two text synthesis algorithms are proposed: *ConThenDisc* and *ConWithinDisc*, both shift the core generative process into a continuous sentence-level semantic space that serves as a global sketched guide for an MDM.
- We demonstrate superior generation quality and sampling speed with LLaDA with significant gains in unconditional text generation setting.

2. Related Work

In Appendix A we provide a broad overview on the field of diffusion models for text generation. In this section we dive into specific recent work that has a direct relevance to this paper’s contributions.

The work reported in (Meshchaninov et al., 2025) presents COSMOS, a language generation algorithm that relies on a continuous latent space diffusion. While similar to the main theme of our work, COSMOS differs from it substantially. In particular, the decoder that converts the embedding to tokens in COSMOS has no generative capabilities, which implies that the latent representation must be fully informative in order to enable proper text synthesis. In contrast, our latent representation serves as a sketch guide that conditions an iterative MDM-based decoding process, and thus even partially informative representations can lead to valid and high quality generated text, as MDM compliments and refines the synthesis process.

Indeed, in the spirit of the main contrast between COSMOS and our paradigm, the work of (Morris et al., 2023) argues that when using embedding representations, decoding must be performed iteratively rather than in a single step, which supports our proposed fusion of continuum and MDM. That said, (Morris et al., 2023) is distinct from our work as it focuses on text correction tasks rather than their generation.

Another related work is reported in (Arriola et al., 2025b), presenting an autoencoding framework, referred to as E2D2. Under a conditional synthesis setup in which the model receives a prompt and is required to provide an answer, E2D2 encodes the prompt to a continuous vector and uses it to guide a fully discrete MDM decoder that constructs the response. As the synthesis of the answer relies on a plain

MDM, the statistical cross-token dependencies are not taken into account – a problem that we tackle in this work.

The algorithms reported in (Liu et al., 2025a; Xu et al., 2025; Xie et al., 2025) tackle the problem of joint token sampling in MDM, as in our work. The first handles the missing dependencies by incorporating a copula model, the second augments the demasker with a learned energy model, and the third introduces a Gaussian-distributed latent variable for accounting for the token dependencies. All concentrate on small scale base models for offering improvements in text synthesis speed or quality. A related yet different line of reasoning towards the very same goal appears in (Azangulov et al., 2025; Luxembourg et al., 2025), presenting inference-only strategies for prioritizing the order of unmasked tokens so as to avoid too-dependent ones to be sampled jointly. These methods are inherently limited, as they seek weakly correlated tokens, which do not necessarily exist. In addition, these inference algorithms are tightly coupled with their base models, operating semi auto-regressively with small block-sizes, thus limiting their achievable gain. In contrast to the above, our work aims to fully harness the potential of diffusion models for language, aiming to override the speed and text-quality barriers of MDM. This is achieved by injecting informative guidance to MDM such that it can both handle cross-token dependencies, while also providing a synthesized sketch for the text to be generated.

3. Problem Formulation and Background

Let $\mathbf{x} = (x^1, x^2, \dots, x^n)$ be a discrete random vector of n tokens, where each x^i belongs to a vocabulary \mathcal{V} . We assume text sequences are sampled from an unknown joint data distribution q_{data} , and our objective is to learn a generative model capable of synthesizing samples from q_{data} .

Following recent work on discrete diffusion methods, we adopt the masked diffusion modeling (MDM) (Sahoo et al., 2024) framework. We augment the vocabulary with a special mask token $[M]$ and define the fully masked vector as $\mathbf{m} = (m^1, m^2, \dots, m^n)$, where $m^i := [M]$ for all i . The generative algorithm begins with a forward diffusion process that gradually degrades a clean sequence. In MDM, this occurs via progressive masking, factorized across tokens as

$$q(\mathbf{x}_t | \mathbf{x}_0) = \prod_{i=1}^n q(x_t^i | x_0^i), \quad (1)$$

where each $q(x_t^i | x_0^i)$ defines an independent categorical corruption process interpolating between a clean sample $\mathbf{x}_0 \sim q_{\text{data}}$ and the masked vector \mathbf{m} :

$$q(x_t^i | x_0^i) := \alpha_t \mathbf{e}_{x_0^i} + (1 - \alpha_t) \mathbf{e}_{[M]}. \quad (2)$$

Here, $\alpha_t \in [0, 1]$ is a strictly decreasing noise schedule over time $t \in [0, 1]$, with $\alpha_0 \approx 1$ and $\alpha_1 \approx 0$. The notation \mathbf{e}_j denotes the one-hot encoding of the j -th vocabulary index.

Generative sampling is achieved by reversing the above-described forward process. For any pair of time-steps $0 \leq s < t \leq 1$, knowledge of the posterior distribution $q(\mathbf{x}_s | \mathbf{x}_t)$ would have enabled synthesis. However, this reverse conditional is intractable. Following prior work (Ho et al., 2020; Sahoo et al., 2024), we consider the conditional reverse transition $q(\mathbf{x}_s | \mathbf{x}_t, \mathbf{x}_0)$, assuming \mathbf{x}_0 is known. With the knowledge of \mathbf{x}_0 , this reverse conditional admits a factorized form without loss of generality,

$$q(\mathbf{x}_s | \mathbf{x}_t, \mathbf{x}_0) := \prod_{i=1}^n q(x_s^i | x_t^i, x_0^i), \quad (3)$$

and $q(x_s^i | x_t^i, x_0^i)$ has a closed-form solution. For example, for MDLM (Sahoo et al., 2024), it is given via

$$q(x_s^i | x_t^i, x_0^i) := \begin{cases} \mathbf{e}_{x_t^i} & \text{if } x_t^i \neq [M], \\ \frac{1-\alpha_s}{1-\alpha_t} \mathbf{e}_{[M]} + \frac{\alpha_s-\alpha_t}{1-\alpha_t} \mathbf{e}_{x_0^i} & \text{if } x_t^i = [M]. \end{cases}$$

In practice, \mathbf{x}_0 is unknown and must be estimated from \mathbf{x}_t , and the way to do so necessarily passes through the approximation of the joint distribution $q(\mathbf{x}_0 | \mathbf{x}_t)$. However, directly modeling this distribution is intractable as well, due to the combinatorial explosion of token options, covering $\mathcal{O}(|\mathcal{V}|^n)$ possible combinations.

To address this, MDM employs a demasking model $f_\theta : (\mathcal{V} \cup \{[M]\})^n \times \mathbb{R} \rightarrow \mathbb{R}^{n \times |\mathcal{V}|}$, that estimates marginal distributions for masked tokens. Formally, $f_\theta^i(\mathbf{x}_t, t) := p_\theta(x_0^i | \mathbf{x}_t)$ approximates $q(x_0^i | \mathbf{x}_t)$ when $x_t^i = [M]$, and returns $\mathbf{e}_{x_t^i}$ otherwise. Given these estimated marginals, we obtain a clean sequence prediction by sampling the tokens independently through $\hat{x}_0^i \sim f_\theta^i(\mathbf{x}_t, t)$. These predicted tokens are then substituted into Equation (3), yielding the effective reverse posterior:

$$p_\theta(\mathbf{x}_s | \mathbf{x}_t) := \prod_{i=1}^n q(x_s^i | x_t^i, x_0^i = \hat{x}_0^i). \quad (4)$$

The factorized approximation in Equation (4) has a well-known fundamental limitation (Liu et al., 2025a): As the demasking model only estimates marginal distributions $q(x_0^i | \mathbf{x}_t)$, independent sampling from these marginals fails to capture cross-token dependencies and semantic correlations across multiple masked positions. In Appendix B we characterize these limitations in recovering the true joint distribution $q(\mathbf{x}_0 | \mathbf{x}_t)$ even under optimal demasking model.

4. Continuously Guided MDM

We now turn to introduce the *CRoCoDiL* framework that aims to bridge the modeling gap between the desired reversed in Equation (3) and the practical approximation posed in Equation (4). Another benefit of our strategy is the introduction of a sketched guidance to the discrete generation process, which further boosts synthesis quality.

Our solution relies on a guidance mechanism for the demasking model, derived from a continuous latent representation of the clean sample \mathbf{x}_0 . This approach enables the model to incorporate global context and cross-token statistical dependencies, bypassing the dimensionality barrier of modeling the joint conditional $q(\mathbf{x}_0|\mathbf{x}_t)$ explicitly. We start by introducing a training framework that constructs an embedding model for turning a token sequence \mathbf{x}_0 into a corresponding continuum latent vector, and allowing the MDM’s demasker to be effectively guided by it. Equipped with this machinery, we present three guided and fast MDM algorithms:

- (i) A general-purpose autoencoder scheme that converts sequences of tokens to the continuum and back, where the encoder is the one described above and the decoder is an MDM algorithm;
- (ii) A novel text synthesis algorithm, *ConThenDisc* (Continuous-Then-Discrete), that generates a valid embedding vector via a continuous diffusion and decodes it by a fast MDM, as in the above autoencoder scheme; and
- (iii) A refined *ConThenDisc* in which the guidance vector is updated within the MDM steps by a conditional diffusion that operates in the embedding domain. We term this method *ConWithinDisc* (Continuous-Within-Discrete).

Common to all three methods is the fact that the MDM may operate in a multi-token sampling regime per step, enabled due to the continuum guidance, and thus becoming much faster than the vanilla MDM alternative. The continuous diffusion and the encoding within *ConThenDisc* and *ConWithinDisc* are relatively lightweight, baring small impact on the overall generation complexity.

4.1. Continuously Guided Demasking

Consider an encoder model that maps a discrete sequence \mathbf{x}_0 into a continuous latent representation $\mathbf{z}_0 \in \mathbb{R}^d$. Assume further that this continuous latent vector is learned so as to serve as an informative guidance for the demasking process in MDM, enabling the recovery of cross-token dependencies among multiple masked positions. Herein we offer a training framework for this constellation.

Encoder. We define an encoder $h_\phi : \mathcal{V}^n \rightarrow \mathbb{R}^d$, parameterized by ϕ , that maps a clean sequence $\mathbf{x}_0 \in \mathcal{V}^n$ to a continuous representation $\mathbf{z}_0 \in \mathbb{R}^d$, denoted as $\mathbf{z}_0 = h_\phi(\mathbf{x}_0)$. This encoder is trained to capture the essential information of the input sequence in continuous form, so as to enable (even a partial¹) reconstruction back to its discrete form.

Guided Demasker. We propose a conditional demasking model $f_\theta : (\mathcal{V} \cup \{[M]\})^n \times \mathbb{R} \times \mathbb{R}^d \rightarrow \mathcal{R}^{n \times |\mathcal{V}|}$, parametrized by θ , that predicts a clean data sample \mathbf{x}_0 from

¹In our work \mathbf{z}_0 does not have to be a one-to-one representation of \mathbf{x}_0 , as in COSMOS (Meshchaninov et al., 2025).

a partially masked sequence \mathbf{x}_t , conditioned on a latent representation $\mathbf{z}_0 \in \mathbb{R}^d$ of the very same clean sequence \mathbf{x}_0 . The decoder outputs the distributions² $f_\theta^i(\mathbf{x}_t, t, \mathbf{z}_0)$ that approximate the true marginals $q(x_0^i|\mathbf{x}_t, \mathbf{z}_0)$ when $x_t^i = [M]$, and $f_\theta^i(\mathbf{x}_t, t, \mathbf{z}_0) := e_{x_t^i}$ otherwise.

Training Objective. The encoder and demasker are jointly trained to minimize the following loss:

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{t, \mathbf{x}_0, \mathbf{x}_t} [-w_{\mathbf{x}_t} \cdot \frac{1}{n} \sum_{i=1}^n \log [f_\theta^i(\mathbf{x}_t, t, h_\phi(\mathbf{x}_0))]_{x_0^i}], \quad (5)$$

where $w_{\mathbf{x}_t}$ denotes the weight that prioritizes clean sequences over corrupted ones. For example, in LLaDA (Nie et al., 2025), the weights are defined as $w_{\mathbf{x}_t}^i := 1/\alpha_t$.

The proposed loss in Equation 5 is constructed by the following chain of steps: We start by sampling a clean sequence \mathbf{x}_0 from the training set, then choose t at random from the uniform $[0, 1]$ distribution, and generating \mathbf{x}_t by randomly masking appropriate portion from \mathbf{x}_0 . The demasker operates on \mathbf{x}_t , t , and $h_\phi(\mathbf{x}_0)$ (the embedding of the original sequence \mathbf{x}_0). Its output at each location i is a logit, in which its x_0^i location should be as high as possible to indicate a preference to the true token. Optimizing over both the parameters of the embedder and the demasker, we drive the output of the demasker to be as close as possible to the original \mathbf{x}_0 . Note that the same expression as in Equation 5 applies to regular MDM with one difference: there is no conditioning on $h_\phi(\mathbf{x}_0)$. As such, the guided machine provides better predictions of the true tokens, implicitly accounting for their statistical cross-dependencies.

Figure 2 illustrates the training framework for the embedding and the guided demasker. Algorithm 1 provides a summary of one training step. Observe a slight change in lines 5-6 of this algorithm: A random Gaussian noise is added to \mathbf{z}_0 , setting its variance via a target value for the cosine-similarity, $CS(\mathbf{z}_0, \mathbf{z}_0 + \mathbf{e}) = 0.8$. As shown in Appendix D, this step is critical to the success of the training results, as it introduces a robustification of the embedding obtained and its influence on the demasker predictions.

This encoder-demasker framework recovers the essential cross-token dependencies during unmasking, enabling the effective factorized reverse transition in Equation (4). Theorem B.1 in Appendix B formally justifies this claim.

4.2. MDM-Based Autoencoder

We now utilize the trained encoder-demasker framework to bridge between discrete and continuous representations.

²We allow for a slightly abused notations by using f_θ to refer to both the original demasker and the guided one. The difference between the two is whether the guidance \mathbf{z}_0 is an additional input.

Algorithm 1 Robust Encoder-Demasker Training Step

- 1: **Input:** Clean sequence $\mathbf{x}_0 \sim p_{\text{data}}$
- 2: Sample timestep $t \sim \mathcal{U}([0, 1])$
- 3: Generate \mathbf{x}_t by masking \mathbf{x}_0 according to noise α_t
- 4: Encode latent representation $\mathbf{z}_0 = h_\phi(\mathbf{x}_0)$
- 5: Draw a random white Gaussian noise \mathbf{e}
- 6: Compute $\hat{\mathcal{L}}(\theta, \phi) = \frac{w_{\mathbf{x}_t}}{n} \sum_{i=1}^n \log f_\theta^i(\mathbf{x}_t, t, \mathbf{z}_0 + \mathbf{e})$
- 7: Update parameters via $\nabla_{\theta, \phi} \hat{\mathcal{L}}(\theta, \phi)$ using optimizer

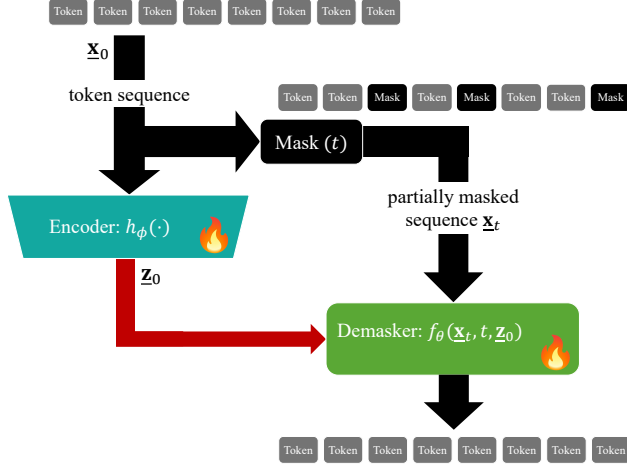


Figure 2. The training framework for the embedding network $\mathbf{z}_0 = h_\phi(\mathbf{x}_0)$ and the guided demasker $f_\theta(\mathbf{x}_t, t, \mathbf{z}_0)$. Flame indicates a trained network.

Specifically, we aim to transform a clean discrete sequence \mathbf{x}_0 into a continuous representation and reconstruct it back as accurately as possible to its original discrete form.

Given a clean input sequence $\mathbf{x}_0 \sim p_{\text{data}}$, we first encode it into the continuous latent via $\mathbf{z}_0 = h_\phi(\mathbf{x}_0)$. To reconstruct it, we deploy a full yet fast (using fewer demasker activations) MDM process, in which the demasker f_θ iteratively refines the sequence, initialized by the fully masked vector \mathbf{m} , and applying T discrete diffusion steps, all conditioned on the representation \mathbf{z}_0 . As in regular MDM, at each timestep t , the demasker predicts the clean tokens independently, $\hat{\mathbf{x}}_0^i \sim f_\theta^i(\mathbf{x}_t, t, \mathbf{z}_0)$, which are then masked according to the forward diffusion process for the next iteration.

This general purpose autoencoder enables flexible processing of sequences in the continuous domain for generation, interpolation, or other downstream tasks, while maintaining the capability to recover faithful discrete outputs. In the context of the discussion in this paper, this autoencoder serves as a stepping stone towards the following hybrid text synthesis algorithms, which build on it.

4.3. Hybrid Text Generation Strategies

We now turn to the main contribution of this section: introducing two hybrid (fusion of continuous and discrete) text generative strategies that take advantage of core MDM while also leaning on the availability of the newly formed continuously guided demasker. We start with *ConThenDisc*, and proceed with its improvement, the *ConWithinDisc* method, both targeting unconditional text synthesis.

4.3.1. CONTINUOUS-THEN-DISCRETE

In Algorithm 6, which describes the proposed autoencoder, a given sentence \mathbf{x}_0 is converted to a continuous latent representation \mathbf{z}_0 , followed by a decoding stage that relies on a full MDM, aiming to recover the original \mathbf{x}_0 . Building on this very configuration, the *ConThenDisc* algorithm suggests producing \mathbf{z}_0 differently; Rather than leaning on a given sentence \mathbf{x}_0 , we suggest to generate it by randomly drawing from its corresponding distribution $\mathbf{z}_0 \sim P(\mathbf{z})$. In practice, \mathbf{z}_0 is synthesized via a pre-trained continuous diffusion generator.

Algorithm 2 describes the *ConThenDisc* text synthesis method. Activating the continuous diffusion algorithm³, $G_\psi(\epsilon)$, we obtain a valid latent sample $\mathbf{z}_0 \sim P(\mathbf{z})$. We proceed by decoding it to a sequence of tokens by applying a complete MDM algorithm – represented by the green lines in Algorithm 2. Note that in generating the latent \mathbf{z}_0 , a knowledge of the desired sequence length can be injected. This can be done by a slightly different design of the continuous diffusion that includes a conditioning on this length. In this work we do not implement this option, and allow any generation length, dictated by the MDM.

Algorithm 2 Continuous-Then-Discrete Text Synthesis

- 1: **Input:**
- 2: $\epsilon \sim \mathcal{N}(0, I)$
- 3: $\mathbf{z}_0 \leftarrow G_\psi(\epsilon)$
- 4: $t = 1$
- 5: $\mathbf{x}_t := \mathbf{m}$
- 6: **while** $t > 0$ **do**
- 7: $\hat{\mathbf{x}}_0 \sim f_\theta(\mathbf{x}_t, t, \mathbf{z}_0)$
- 8: $t \leftarrow t - 1/T$
- 9: $\mathbf{x}_t = \text{Forward}(\hat{\mathbf{x}}_0, t)$
- 10: **end while**
- 11: **return** $\hat{\mathbf{x}}_0$

4.3.2. CONTINUOUS-WITHIN-DISCRETE

A delicate weakness (and thus an unexploited opportunity) in Algorithm 2 is the fact that the guidance is kept fixed throughout the T iterations, even though the sequence \mathbf{x}_t

³Appendix E describes $G_\psi(\epsilon)$ in more details.

is available, giving additional yet partial information about the text to be created. The *ConWithinDisc* algorithm aims to leverage this opportunity, by updating the guidance vector within the MDM steps. More specifically, within each demasking step, the guidance vector can be updated by drawing from the conditional distribution $\mathbf{z}_0 \sim P(\mathbf{z}|h_\phi(\mathbf{x}_t))$. In words, the guidance vector is sharpened to take into account the currently held temporal sequence \mathbf{x}_t . Algorithm 3 provides a description of this variant, and Figure 3 presents *ConThenDisc* and *ConWithinDisc*, highlighting their difference.

Few comments are in order: (i) The update of \mathbf{z}_0 can be done in a pre-selected subset of the overall T steps, in order to benefit from the improved guidance while reducing the overall complexity of the generative algorithm; (ii) In drawing the guidance vector, the conditioning we present leans on the *embedding* of the partially masked sequence \mathbf{x}_t , i.e. $\mathbf{z}_0 \sim P(\mathbf{z}|h_\phi(\mathbf{x}_t))$. Rather, we could have conditioned the distribution directly on \mathbf{x}_t ; (iii) In training the conditional diffusion (Algorithm 4), we use $h_\phi(\mathbf{x}_t)$ for embedding the partially masked sentence. However, this encoder was not trained for such masked content. An improved strategy would be to define a second encoder $h_\mu(\mathbf{x}_t)$ to be used in Line 7, defining the loss as $\hat{\mathcal{L}}(\mu, \psi) = \|g_\psi(\mathbf{z}_s, h_\mu(\mathbf{x}_t) - \mathbf{z}_0)\|_2^2$, and optimizing it w.r.t. both μ and ψ ; and (iv) Drawing samples from the conditional distribution $\mathbf{z}_0 \sim P(\mathbf{z}|h_\phi(\mathbf{x}_t))$ can be interpreted as a solution of an inverse problem. Given a prior $P(\mathbf{z})$, and given measurements $h_\phi(\mathbf{x}_t)$, our goal is to produce posterior samples that recover the \mathbf{z}_0 that led to the given measurements.

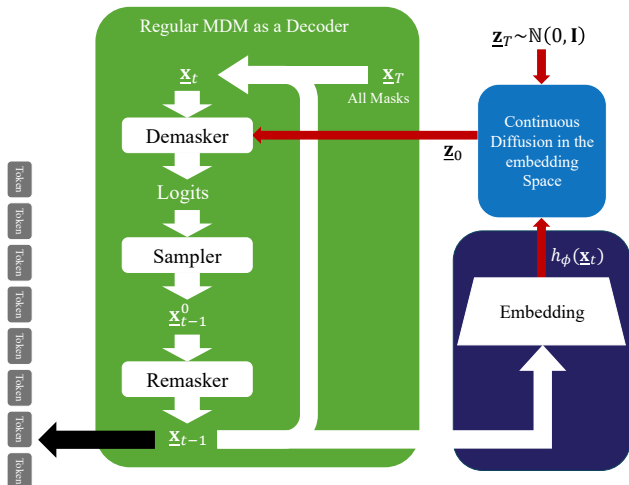


Figure 3. The *ConThenDisc* and *ConWithinDisc* text generation algorithms. In both, a continuous diffusion model generates a starting latent \mathbf{z}_0 , which is decoded to tokens by a regular MDM. *ConWithinDisc* includes a refinement of the latent (the purple part) based on the partially synthesized text.

Algorithm 3 Continuous-Within-Discrete

```

1: Input:
2:  $t = 1$ 
3:  $\mathbf{x}_t := \mathbf{m}$ 
4: while  $t > 0$  do
5:    $\epsilon \sim \mathcal{N}(0, I)$ 
6:    $\mathbf{z}_0 \leftarrow G_\psi(\epsilon, h_\phi(\mathbf{x}_t))$ 
7:    $\hat{\mathbf{x}}_0 \sim f_\theta(\mathbf{x}_t, t, \mathbf{z}_0)$ 
8:    $t \leftarrow t - 1/T$ 
9:    $\mathbf{x}_t = \text{Forward}(\hat{\mathbf{x}}_0, t)$ 
10: end while
11: return  $\hat{\mathbf{x}}_0$ 
    
```

Algorithm 4 Continuous-Within-Discrete Training

```

1: Input: data  $\mathbf{x}_0 \sim P(\mathbf{x})$ ,
2:  $\mathbf{z}_0 = h_\phi(\mathbf{x}_0)$ 
3: Sample  $s \sim \mathcal{U}([0, 1])$ 
4:  $\mathbf{z}_s = \text{Forward}(\mathbf{z}_0, s)$ 
5: Sample  $t \sim \mathcal{U}([0, 1])$ 
6:  $\mathbf{x}_t \leftarrow \text{mask each token with probability } t$ 
7:  $\hat{\mathcal{L}}(\phi, \psi) = \|g_\psi(\mathbf{z}_s, h_\phi(\mathbf{x}_t) - \mathbf{z}_0)\|_2^2$ 
8: Backpropagate on  $\nabla_\psi \hat{\mathcal{L}}(\phi, \psi)$  and run optimizer
    
```

5. Experimental Results

Guided-Demasker

In all the reported experiments hereafter we worked with LLaDA, tuned to generate Python programs. Pre-training their demasker $f_\theta(\mathbf{x}_t, t)$ was done with 12 million Python programs of varying lengths in the range $[0, 4096]$ tokens taken from Python subset of the StarCoder Dataset (Li et al., 2023), initializing the training with the open-source base version. This model thus generates similar programs of varying lengths, with BOS and EOS tokens to indicate their beginning and ending, correspondingly.

Building on the above as a baseline, we trained $f_\theta(\mathbf{x}_t, t, \mathbf{z}_0)$ and $\mathbf{z}_0 = h_\phi(\mathbf{x}_0)$ as described in Section 4.1. The embedding $h_\phi(\cdot)$ was initialized with a Qwen embedding (Ren & et al., 2025), and refined via the training. The latent output of this model is of size $d = 1024 \times K$, where $1 \leq k \leq 128$ represent a number of representation registers. A dropout training strategy was applied with preference to the first registers in order to enable varying size latent representations. More details on this process are found in Appendix C.

Figure 4 presents the obtained Cross-Entropy and the probability of recovering the true tokens, evaluated on a validation set of 1000 programs of lengths $[0, 4096]$ tokens, covering the base LLaDA model demasker and three conditioned versions of it fed with $K = 8, 64$ and 128 embedding registers. The horizontal axis shows the mask probability, were

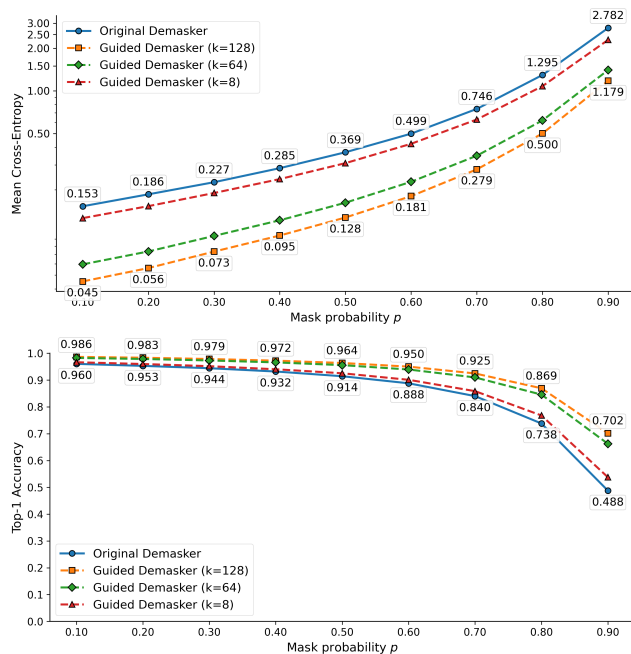


Figure 4. **Guided-Demasker**: Cross-Entropy (top) and top-1 token prediction performance versus masking probability (0 for no-mask), measured on a validation set of 1000 sequences.

0 stands for no masking and 1 for fully masked sentences. As expected, the Cross-Entropy deteriorates for all models with more aggressive masks. The conditioning improves the overall performance, with a gap that grows with more latent registers. Very similar conclusions can be drawn for the bottom graph: conditioning improves the obtained accuracies, and more latent registers are beneficial.

MDM-Based Autoencoder

Equipped with a trained conditioned demasker $f_\theta(\mathbf{x}_t, t, \mathbf{z}_0)$ and an embedding model $\mathbf{z}_0 = h_\phi(\mathbf{x}_0)$, we now turn to evaluate the autoencoder scheme, in which a Python program \mathbf{x}_0 is encoded to a continuum latent, and an MDM algorithm decodes it back to tokens. The hyper-parameters governing the MDM are the sequence length, the block-size, and the NFE (Neural Function Evaluations) - the overall number of demasker activations, which is governed by number of unmasked tokens per step. For example, for a generated length of 256 tokens, block size= 32 implies that there are 8 blocks, and NFE=32 means that we apply 4 demasking steps within each block, thus reviving 8 tokens in each step. Figure 5 show an example source program and its encoded-decoded result, which is a nearly perfectly reconstructed, even if the MDM is applied with few NFE.

Table 1 brings representative results for the performance of this autoencoder, operating on sequences of length 256 tokens using a latent of size 1024×128 , and varying the MDM’s hyper-parameters. The table reports Generative Per-

plexity (Gen-PPL) that measures how coherent or “likely” the generated text is. Also reported are Bert scores and Character Error Rate (CER). The first evaluates semantic code similarity using CodeBERTScore (Zhou et al., 2023), an embedding-based metric that aligns contextual token representations from a pretrained model via cosine similarity. Character Error Rate (CER) (Jurafsky & Martin, 2009) is defined as the normalized Levenshtein edit distance between the generated code string and the reference code string at the character level. CER counts the minimum number of single-character insertions, deletions, and substitutions required to transform the prediction into the reference, divided by the number of characters in the reference.

As can be seen in this table, varying LLaDA’s block-size and overall number of NFE, nearly perfect synthesized text is obtained, even for very low NFE, and surprisingly, this behavior strengthens with larger block-size. Increasing the number of demasking steps steadily improves reconstruction, reaching CER around 0.10 and CodeBERTScore F1 around 0.96 to 0.97. The combination of low CER and high “CodeBERTScore” suggests that the remaining differences are concentrated on whitespace, formatting, or identifier naming, rather than major semantic changes.

More details and results referring to this autoencoder are brought in Appendix F.

Table 1. **Autoencoder**: Performance measured via Generative Perplexity, and recovery error evaluated via Bert-Score and Character Error Rate (CER). The Table explores varying hyper-parameters of the MDM decoder, for a generative length of 256 tokens.

Block	NFE	Gen-PPL	Bert-Score	CER
32	8	59.538	0.901	0.422
32	16	20.263	0.936	0.210
32	32	13.525	0.957	0.150
32	64	11.301	0.968	0.130
32	128	10.401	0.970	0.123
32	256	10.085	0.972	0.118
64	8	28.172	0.925	0.265
64	16	15.553	0.951	0.171
64	32	12.463	0.963	0.140
64	64	10.971	0.970	0.124
128	8	18.027	0.946	0.190
128	16	13.244	0.960	0.149
128	32	11.566	0.968	0.130
128	64	10.639	0.971	0.122
256	4	19.221	0.939	0.205
256	8	13.937	0.957	0.155
256	16	11.981	0.965	0.132
256	32	10.962	0.970	0.123
256	64	10.458	0.973	0.118

Unconditional Text Generation

Given the trained embedding network, $h_\phi(\mathbf{x}_0)$, we used the 2 million text sequences of varying lengths, and converted them to latent matrices of size 1024×128 . These were used for training a denoiser for the continuous diffusion. More details on this training, the diffusion algorithm used, and its overall performance evaluation are brought in Appendix E.

ConThenDisc generates text sequences by synthesizing an

```

Ground Truth
# Ground Truth
TITLE = "Masked Language Diffusion Models with Continuous-Guidance"
def quicksort(xs):
    """Classic quicksort (recursive)."""
    if len(xs) <= 1:
        return xs
    pivot = xs[len(xs) // 2]
    left = [x for x in xs if x < pivot]
    mid = [x for x in xs if x = pivot]
    right = [x for x in xs if x > pivot]
    return quicksort(left) + mid + quicksort(right)
sorted_chars = quicksort(list(TITLE))
print("sorted :", " ".join(sorted_chars))

NFE = 4
# NFE = 4
TITLE = "Masked Diff Diffusion Models with Continuous-Guidance"
def quicksort(x):
    """Classic quicksort (recursive)."""
    if len(xs) <= 1:
        return xs
    pivot = xs[len(xs) // 2]
    left = [x for x in xs if x < pivot]
    mid = [x for x in xs if x = pivot]
    right = [x for x in xs if x > pivot]
    return quicksort(left) + mid + quicksort(right)
sortedsorted_chars = quort(list(TITLE))
("sortedsorted :", " ".join(sorted_chars))

NFE = 16
# NFE = 16
TITLE = "Masked Language Diffusion Models with Continuous-Guidance"
def quicksort(xs):
    """Classic quicksort (recursive)."""
    if len(xs) <= 1:
        return xs
    pivot = xs[len(xs) // 2]
    left = [x for x in xs if x < pivot]
    mid = [x for x in xs if x = pivot]
    right = [x for x in xs if x > pivot]
    return quicksort(left) + mid + quicksort(right)
sorted_chars = quicksort(list(TITLE))
print("sorted :", " ".join(sorted_chars))
    
```

Figure 5. **Autoencoder**: A source program and its encoded-decoded outcome (Gen-Length=256 tokens as one block). The two outcomes correspond to NFE=4 (CER=0.12) and NFE=16 (CER=0.03).

embedding vector, and then decoding it to tokens via an MDM. *ConWithinDisc* generalizes the above by updating the guidance latent vector within the MDM iterations. This involves an additional training of the continuous diffusion, conditioning its denoiser on the embedding of the temporally available sequence. More details on this training are brought in Appendix E. The main hyper-parameters governing these processes are the dimension of the latent vector generated, and the MDM parameters (sequence length, block-size, number of unmasked tokens in each step, NFE).

referring to all the text as one block, which was found best for the base model. As can be seen, *ConThenDisc* and *ConWithinDisc* perform much better than the LLaDA baseline across a wide range of NFE and generation lengths, reflected by both the MAUVE and the Gen-PPL measures.

For example, generating sequences of length 512 with a base-LLaDA model that uses NFE= 512 (MAUVE=0.62, Gen-PPL=19.4) parallels *ConWithinDisc* with NFE= 40 (MAUVE=0.6, Gen-PPL=14.3), implying a speedup of $\times 13$. Similarly, for sequences of length 1024, base-LLaDA with NFE= 1024 (MAUVE=0.76, Gen-PPL=23.5) is 14 times slower than a comparable and even better *ConWithinDisc* (NFE=72, MAUVE=0.8, Gen-PPL=12.5).

In these graphs, *ConWithinDisc* has been implemented such that it uses only one additional continuous diffusion in the middle of the MDM steps in order to update the latent guidance, offering an evident improvement over *ConThenDisc*, at the cost of adding less than 2 to the NFE. More results on these experiments with a wider coverage the hyper-parameters are brought in Appendix G.

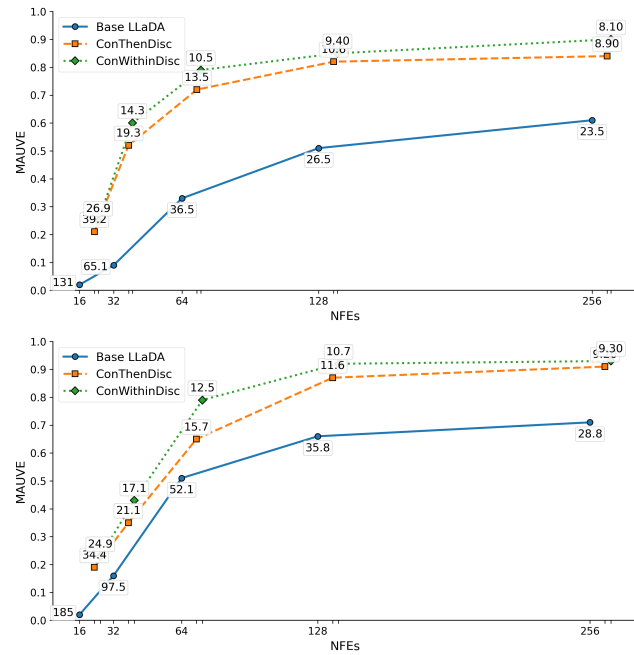


Figure 6. **Text Generation**: MAUVE and Gen. Perplexity for base LLaDA, *ConThenDisc* and *ConWithinDisc*, vs. complexity (NFE). $n = 512, K = 8$ (top) and $n = 1024, K = 16$ (bottom). In all cases, the whole sequence is treated as one block.

Figure 6 presents the results obtained for varying generation lengths (512 and 1024), while sweeping through NFE⁴ and

⁴NFE accounts for the continuous diffusion: The denoiser (400M parameters) is activated 128 times, leading to an overall load of ≈ 6 demasker (8B) activations. The conditioned diffusion

6. Conclusion

While Masked Diffusion Models (MDMs) offer a compelling approach to text generation, they face inherent performance hurdles. We introduce *CRoCoDiL*, a framework that overcomes these bottlenecks to provide a substantial boost in synthesis speed and text quality. By first generating a “sketched” latent representation in a continuous space and then converting it into tokens via a guided MDM, *CRoCoDiL* achieves superior results, as demonstrated in our experiments with LLaDA (Nie et al., 2025). Our future research will focus on extending the proposed algorithms to handle conditional (prompt-based) text synthesis (see Appendix H), optimizing the continuous diffusion processes via distillation, and exploring more efficient latent designs. Additionally, we aim to improve conditional generation by training prompt embeddings and testing our framework against a wider variety of baseline models.

uses 32 time steps, leading to < 2 NFE.

Impact Statement

This paper presents work whose goal is to advance the fields of Machine Learning and Generative-AI. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

- Arriola, M., Gokaslan, A., Chiu, J. T., Yang, Z., Qi, Z., Han, J., Sahoo, S. S., and Kuleshov, V. Block diffusion: Interpolating between autoregressive and diffusion language models. In *The Thirteenth International Conference on Learning Representations*, 2025a.
- Arriola, M., Schiff, Y., Phung, H., Gokaslan, A., and Kuleshov, V. Encoder-decoder diffusion language models for efficient training and inference. *arXiv preprint arXiv:2510.22852*, 2025b.
- Austin, J., Johnson, D. D., Ho, J., Tarlow, M., and van den Berg, R. Structured denoising diffusion models in discrete state-spaces. In *NeurIPS*, 2021.
- Azangulov, I., Pandeva, T., Prasad, N., Zazo, J., and Karmalkar, S. Parallel sampling from masked diffusion models via conditional independence testing. *arXiv preprint arXiv:2510.21961*, 2025.
- Bachmann, R., Allardice, J., Mizrahi, D., Fini, E., Kar, O. F., Amirloo, E., El-Nouby, A., Zamir, A., and Dehghan, A. FlexTok: Resampling images into 1d token sequences of flexible length. *arXiv 2025*, 2025.
- Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., and Li, Z. Pixart- α : Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023.
- Croitoru, F.-A., Hondru, V., Ionescu, R. T., and Shah, M. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023.
- Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021.
- Dieleman, S., Sartran, L., Roshannai, A., Savinov, N., Ganin, Y., et al. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*, 2022.
- Gao, Y. et al. Empowering diffusion models on the embedding space for text generation. In *ICLR*, 2024.
- Gulrajani, I. and Hashimoto, T. B. Plaid: Likelihood-based diffusion language models. In *NeurIPS*, 2024.
- Han, X., Kumar, S., and Tsvetkov, Y. SSD-LM: Semi-autoregressive simplex-based diffusion language model. In *ACL*, 2023.
- Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. Video diffusion models. In *Advances in Neural Information Processing Systems*, 2022.
- Hoogeboom, E., Nielsen, D., Jaini, P., Alaa, A., and Welling, M. Argmax flows and multinomial diffusion: Towards non-autoregressive language models. In *NeurIPS*, 2021.
- Jurafsky, D. and Martin, J. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition with language models (p. 5). ed.), *New Jersey: Pearson Education, Inc*, 2009.
- Karimi Mahabadi, R. et al. TESS: Text-to-text self-conditioned simplex diffusion. In *EACL*, 2024.
- Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. Diffwave: A versatile diffusion probabilistic model for audio synthesis. In *International Conference on Learning Representations*, 2021.
- Li, R., Allal, L. B., Zi, Y., Muennighoff, N., Kocetkov, D., Mou, C., Marone, M., Akiki, C., Li, J., Chim, J., et al. Starcoder: may the source be with you! *arXiv preprint arXiv:2305.06161*, 2023.
- Li, X. L., Thickestun, J., Gulrajani, I., Liang, P., and Hashimoto, T. B. Diffusion-lm: Improves controllable text generation. In *NeurIPS*, 2022.
- Liu, A., Broadrick, O., Niepert, M., and Van den Broeck, G. Discrete copula diffusion. In *The Thirteenth International Conference on Learning Representations*, 2025a.
- Liu, S., Nam, J., Campbell, A., Stark, H., Xu, Y., Jaakkola, T., and Gomez-Bombarelli, R. Think while you generate: Discrete diffusion with planned denoising. In *The Thirteenth International Conference on Learning Representations*, 2025b.
- Liu, Z., Yang, Y., Zhang, Y., Chen, J., Zou, C., Wei, Q., Wang, S., and Zhang, L. dllm-cache: Accelerating diffusion large language models with adaptive caching. *arXiv preprint arXiv:2506.06295*, 2025c.
- Lou, A., Meng, C., and Ermon, S. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

- Luxembourg, O., Permuter, H., and Nachmani, E. Plan for speed-dilated scheduling for masked diffusion language models. *arXiv preprint arXiv:2506.19037*, 2025.
- Meshchaninov, V., Chimbulatov, E., Shabalin, A., Abramov, A., and Vetrov, D. Compressed and smooth latent space for text diffusion modeling. *arXiv preprint arXiv:2506.21170*, 2025.
- Morris, J., Kuleshov, V., Shmatikov, V., and Rush, A. M. Text embeddings reveal (almost) as much as text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12448–12460, 2023.
- Naeem, M. F., Oh, S. J., Uh, Y., Choi, Y., and Yoo, J. Reliable fidelity and diversity metrics for generative models. 2020.
- Nie, S., Zhu, F., You, Z., Zhang, X., Ou, J., Hu, J., Zhou, J., Lin, Y., Wen, J.-R., and Li, C. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.
- Ren, X. and et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.
- Rudin, W. *Real and Complex Analysis*. McGraw-Hill, New York, 3rd edition, 1987.
- Sahoo, S., Arriola, M., Schiff, Y., Gokaslan, A., Marroquin, E., Chiu, J., Rush, A., and Kuleshov, V. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184, 2024.
- Shi, J., Han, K., Wang, Z., Doucet, A., and Titsias, M. K. Simplified and generalized masked diffusion for discrete data. In *NeurIPS*, 2024.
- Shi, J. et al. Non-markovian discrete diffusion with causal language models. *arXiv preprint arXiv:2502.xxxx*, 2025.
- Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.
- Uziel, R., Chelly, I., Freifeld, O., and Pakman, A. Clustering via self-supervised diffusion. *arXiv preprint arXiv:2507.04283*, 2025.
- Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- Wu, C., Zhang, H., Xue, S., Diao, S., Fu, Y., Liu, Z., Molchanov, P., Luo, P., Han, S., and Xie, E. Fast-dllm v2: Efficient block-diffusion llm. *arXiv preprint arXiv:2509.26328*, 2025.
- Xie, T., Xue, S., Feng, Z., Hu, T., Sun, J., Li, Z., and Zhang, C. Variational autoencoding discrete diffusion with enhanced dimensional correlations modeling. *arXiv preprint arXiv:2505.17384*, 2025.
- Xu, M., Geffner, T., Kreis, K., Nie, W., Xu, Y., Leskovec, J., Ermon, S., and Vahdat, A. Energy-based diffusion language models for text generation. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Shao, W., Zhang, W., Cui, B., and Yang, M.-H. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.
- Ye, J., Xie, Z., Zheng, L., Gao, J., Wu, Z., Jiang, X., Li, Z., and Kong, L. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*, 2025.
- Yi, Q., Chen, X., Zhang, C., Zhou, Z., Zhu, L., and Kong, X. Diffusion models in text generation: a survey. *PeerJ Computer Science*, 10:e1905, 2024.
- Zhang, C., Zhang, J., Zhang, X., et al. Concrete score matching: Generalized score matching for discrete data. In *NeurIPS*, 2022.
- Zhou, S., Alon, U., Agarwal, S., and Neubig, G. Codebertscore: Evaluating code generation with pretrained models of code. In *2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pp. 13921–13937. Association for Computational Linguistics (ACL), 2023.

A. Diffusion Models for Language – Background

In the past five years, diffusion models have taken the lead in image, video and audio synthesis tasks (Ho et al., 2020; Dhariwal & Nichol, 2021; Rombach et al., 2022; Ho et al., 2022; Kong et al., 2021; Yang et al., 2023; Croitoru et al., 2023). These algorithms all rely on a continuous formulation of diffusion algorithms, mostly assuming a Gaussian noise contamination and a corresponding denoising network, serving as foundational elements. Bringing these methods to handle language is far from trivial, as text is discrete and unordered. This gap poses a challenge that many recent papers have attempted to resolve. Roughly speaking, there are three general strategies in constructing a bridge between continuum-based diffusion models and language:

- **Go Discrete:** The rationale of diffusion algorithms can be brought to the discrete domain, by intuitively replacing the Gaussian noise contamination by a masking operation or via random replacements of tokens. This line of work has been drawing much attention recently, exhibiting a tendency to appealing results. MDM methods, such as MDLM (Sahoo et al., 2024), LLaDA (Nie et al., 2025) and Dream (Ye et al., 2025), belong to this thread of work. The following papers are additional representatives of this group: (Austin et al., 2021; Hoogetboom et al., 2021; Shi et al., 2024; 2025).
- **Go Continuum:** Text synthesis can be performed with classical (continuous) diffusion models, assuming that text could be embedded to the continuum, and decoding it back to tokens is within reach. This approach has been explored in a series of papers, with partial success – the following are few representatives of this group: (Li et al., 2022; Dieleman et al., 2022; Gulrajani & Hashimoto, 2024; Gao et al., 2024).
- **Go Midway:** Diffusion models can be reformulated thoroughly and rigorously while focusing on discrete data. This, for example, is the approach taken by the work on the Concrete-Score, which leans on ratios of probabilities. Another related option operates in the logits domain, forming simplex-based diffusion alternatives. Examples of this approach are the work reported in (Zhang et al., 2022; Han et al., 2023; Lou et al., 2024; Karimi Mahabadi et al., 2024).

For newcomers to this domain, the impression is likely to be that Masked-based Diffusion Models (MDM) are the preferred algorithms, as they have taken the lead in bringing diffusion models to language. We argue that “the jury is still out” on this question, as MDM still faces critical challenges, while the above-described alternatives are underexplored to a large extent. Our work offers a novel fusion of continuum and discrete, which preserves the core essence of MDM algorithms, while boosting them via a continuum embedding. This brings us to describe several closely related papers in Section 3 that take a similar, yet different, path towards addressing similar goals.

B. Theoretical Analysis

B.1. Limitations of the Factorized Approximation of Joint Token Distributions

In this section, we formally characterize the gap between the true joint distribution $q(\mathbf{x}_0|\mathbf{x}_t)$ and the factorized approximation $p_\theta(\mathbf{x}_0|\mathbf{x}_t) := \prod_{i=1}^n p_\theta(x_0^i|\mathbf{x}_t)$ typically employed in MDMs. Here, $p_\theta(x_0^i|\mathbf{x}_t)$ represents the demasking model whose task is to recover masked tokens within the corrupted sequence \mathbf{x}_t .

Independence Gap: We demonstrate that even if one is equipped with an optimal demasking model, $p_{\theta^*}(x_0^i|\mathbf{x}_t) = q(x_0^i|\mathbf{x}_t)$, the assumption of conditional independence inherently results in a loss of structural information.

Let $p_{\theta^*}(\mathbf{x}_0|\mathbf{x}_t) = \prod_{i=1}^n q(x_0^i|\mathbf{x}_t)$ denote the optimal factorized approximation of the true joint distribution $q(\mathbf{x}_0|\mathbf{x}_t)$. The *Independence Gap*, which corresponds to the conditional total correlation of the tokens \mathbf{x}_0 given the partially masked sequence \mathbf{x}_t , is defined as:

$$C(\mathbf{x}_0|\mathbf{x}_t) := D_{\text{KL}} \left(q(\mathbf{x}_0|\mathbf{x}_t) \parallel \prod_{i=1}^n q(x_0^i|\mathbf{x}_t) \right) \quad (6)$$

Since natural language is governed by high-order dependencies (e.g., long-range syntactic constraints and local semantic coherence), the term $C(\mathbf{x}_0|\mathbf{x}_t)$ is strictly positive. This implies that the factorized model p_θ is theoretically incapable of perfectly recovering the true data distribution q unless the tokens are truly conditionally independent.

Semantic Incoherence: Consider a simple case where the data distribution consists of two equally likely sequences: $\mathbf{s}_1 = (\text{cat}, \text{meows})$ and $\mathbf{s}_2 = (\text{dog}, \text{barks})$. If at time t both tokens are masked ($\mathbf{x}_t = ([M], [M])$), the true joint is:

$$q(\mathbf{x}_0|\mathbf{x}_t) = 0.5\delta(\mathbf{x}_0 - \mathbf{s}_1) + 0.5\delta(\mathbf{x}_0 - \mathbf{s}_2). \quad (7)$$

The optimal marginals are $q(x_0^1 = \text{cat}|\mathbf{x}_t) = 0.5$ and $q(x_0^2 = \text{meows}|\mathbf{x}_t) = 0.5$, and similarly for the other pair. However, the factorized model p_{θ^*} yields:

$$p_{\theta^*}(\mathbf{x}_0|\mathbf{x}_t) = (0.5\mathbf{e}_{\text{cat}} + 0.5\mathbf{e}_{\text{dog}}) \otimes (0.5\mathbf{e}_{\text{meows}} + 0.5\mathbf{e}_{\text{barks}}). \quad (8)$$

This assigns a 25% probability to (cat, barks) and (dog, meows), which are out-of-distribution (OOD) sequences. This "marginal drift" forces the generative process to navigate through regions of the token space that do not correspond to valid language, often leading to a loss of global coherence in long-form synthesis.

Propagation of Error in Reverse Sampling: In the iterative reverse process, we sample $\hat{\mathbf{x}}_0 \sim p_{\theta}(\mathbf{x}_0|\mathbf{x}_t)$ and use it to compute the next step \mathbf{x}_s . Because p_{θ} ignores the token dependencies as described above, $\hat{\mathbf{x}}_0$ is likely to be incoherent. When this incoherent $\hat{\mathbf{x}}_0$ is plugged into the effective reverse transition $p_{\theta}(\mathbf{x}_s|\mathbf{x}_t)$, defined in Equation (4), it is guided towards an inconsistent clean sequence, accumulating the error across the sampling trajectory.

B.2. Formal Justification for Guided Factorization

In this section, we provide the formal justification for the factorized reverse transition used in the *CRoCoDiL* framework. We demonstrate that by conditioning on an appropriate continuous latent representation $\mathbf{z}_0 \in \mathbb{R}^d$, we can sample multiple tokens independently without losing the structural dependencies of the sequence.

Theorem B.1. *Let $\mathbf{x}_0 \in \mathcal{V}^n$ be a discrete sequence and $\mathbf{z}_0 \in \mathbb{R}^d$ be its continuous representation that contains the information on the cross-token dependencies in \mathbf{x}_0 , such that the clean tokens are conditionally independent given \mathbf{z}_0 :*

$$p_{\theta}(\mathbf{x}_0|\mathbf{z}_0) := \prod_{i=1}^n p_{\theta}(x_0^i|\mathbf{z}_0). \quad (9)$$

Then, the reverse transition conditioned on \mathbf{z}_0 factorizes over the token positions:

$$p_{\theta}(\mathbf{x}_s|\mathbf{x}_t, \mathbf{z}_0) = \prod_{i=1}^n p_{\theta}(x_s^i|\mathbf{x}_t, \mathbf{z}_0). \quad (10)$$

Proof. Let $s < t$ be arbitrary diffusion timestamps. In the following proof we rely on the standard assumption in diffusion models that the forward noise process factorizes over dimensions (tokens) and is Markovian. Specifically:

1. Forward Independence: $q(\mathbf{x}_t|\mathbf{x}_s) = \prod_{i=1}^n q(x_t^i|x_s^i)$.
2. Markov Property: $q(\mathbf{x}_t|\mathbf{x}_s, \mathbf{z}_0) = q(\mathbf{x}_t|\mathbf{x}_s)$.

We start by a factorization of the marginal probability $p_{\theta}(\mathbf{x}_s|\mathbf{z}_0)$, by integrating out \mathbf{x}_0 :

$$\begin{aligned} p_{\theta}(\mathbf{x}_s|\mathbf{z}_0) &= \int q(\mathbf{x}_s|\mathbf{z}_0, \mathbf{x}_0)p_{\theta}(\mathbf{x}_0|\mathbf{z}_0) d\mathbf{x}_0 \\ &= \int q(\mathbf{x}_s|\mathbf{x}_0)p_{\theta}(\mathbf{x}_0|\mathbf{z}_0) d\mathbf{x}_0 \\ &= \int \left[\prod_{i=1}^n q(x_s^i|x_0^i) \right] \left[\prod_{j=1}^n p_{\theta}(x_0^j|\mathbf{z}_0) \right] d\mathbf{x}_0 \\ &= \prod_{i=1}^n \left(\int q(x_s^i|x_0^i)p_{\theta}(x_0^i|\mathbf{z}_0) dx_0^i \right) \\ &= \prod_{i=1}^n p_{\theta}(x_s^i|\mathbf{z}_0). \end{aligned} \quad (11)$$

In the first transition in the above derivation, we omitted the appearance of \mathbf{z}_0 from $q(\mathbf{x}_s|\mathbf{z}_0, \mathbf{x}_0)$ due to the Markov property. In the fourth transition that interchanges the order between the integral and the multiplication, we rely on the fact that the integrand is separable (Fubini's Theorem (Rudin, 1987)).

We proceed by factorizing the Reverse Transition, via Bayes’ rule:

$$p_{\theta}(\mathbf{x}_s | \mathbf{x}_t, \mathbf{z}_0) = \frac{q(\mathbf{x}_t | \mathbf{x}_s, \mathbf{z}_0) p_{\theta}(\mathbf{x}_s | \mathbf{z}_0)}{p_{\theta}(\mathbf{x}_t | \mathbf{z}_0)}.$$

Using the forward independence assumption and the result from Equation 11 (applied to both s and t), we substitute the terms:

$$\begin{aligned} p_{\theta}(\mathbf{x}_s | \mathbf{x}_t, \mathbf{z}_0) &= \frac{(\prod_{i=1}^n q(x_t^i | x_s^i)) (\prod_{i=1}^n p_{\theta}(x_s^i | \mathbf{z}_0))}{\prod_{i=1}^n p_{\theta}(x_t^i | \mathbf{z}_0)} \\ &= \prod_{i=1}^n \left(\frac{q(x_t^i | x_s^i) p_{\theta}(x_s^i | \mathbf{z}_0)}{p_{\theta}(x_t^i | \mathbf{z}_0)} \right). \end{aligned}$$

The term inside the product corresponds exactly to the single-token posterior $p_{\theta}(x_s^i | x_t^i, \mathbf{z}_0)$ derived via Bayes’ rule for the i -th dimension. Thus,

$$p_{\theta}(\mathbf{x}_s | \mathbf{x}_t, \mathbf{z}_0) = \prod_{i=1}^n p_{\theta}(x_s^i | \mathbf{x}_t, \mathbf{z}_0).$$

and this concludes the proof. □

C. Additional Details on the Guided-Demasker Training

In training the conditioned demasker, we jointly learn the parameters of an encoder h_{ϕ} and a demasking decoder f_{θ} to obtain a compact continuous representation that preserves the global constraints needed to reconstruct the discrete sequence, while remaining well-structured for the continuous generative model in the next stage. Given a clean sequence \mathbf{x}_0 , the encoder produces a bank of continuous registers $\mathbf{z}_0 = h_{\phi}(\mathbf{x}_0)$, and the demasker is trained to reconstruct \mathbf{x}_0 from a corrupted state \mathbf{x}_t while conditioning on these registers.

We instantiate h_{ϕ} with a Qwen embedding backbone, but depart from standard single-vector pooling by introducing a small bank of learned suffix registers. Specifically, we append $K - 1$ learned token embedding to the input, and additionally retain the single summary vector corresponding to the final valid token position as usually used in Qwen, resulting in K embedding layers

$$\mathbf{z}_0 = (\mathbf{z}_0^{(1)}, \dots, \mathbf{z}_0^{(K)}), \quad \mathbf{z}_0^{(j)} \in \mathbb{R}^d.$$

Note that in the original Qwen embedder only one vector is used, but this may not be enough for our needs: A multi-vector bottleneck distributes capacity across coarse and fine-grained attributes of the sequence, whereas a single vector often forces an overly lossy compression. To better align these embedding with the later-applied continuous diffusion, we normalize each register to unit length and rescale by \sqrt{d} :

$$\mathbf{z}_0^{(j)} \leftarrow \sqrt{d} \cdot \frac{\mathbf{z}_0^{(j)}}{\|\mathbf{z}_0^{(j)}\|_2}.$$

as done in previous work (Uziel et al., 2025). While alternative normalizations are possible – e.g., applying LayerNorm with learned affine parameters (Gao et al., 2024) – this changes the scale of the register distribution and, in turn, the effective noise magnitude in the continuous diffusion stage. In practice, adopting such learned scaling typically requires additional calibration, either via a dedicated hyperparameter search over the noise/rescaling schedule or by estimating the appropriate normalization statistics (mean and variance) on a held-out set.

A key requirement for the subsequent latent diffusion model is that \mathbf{z}_0 is *progressively organized*: earlier registers should remain informative even when later registers are absent. We enforce this property using nested dropout over registers (Bachmann et al., 2025). For each example we sample $k \in \{1, \dots, K\}$ and expose only the first k registers to the demasker, withholding the rest. This induces an ordered decomposition in which global “core” information concentrates in early registers while residual details are delegated to later ones. Empirically, removing nested dropout typically has limited effect on reconstruction quality, but substantially harms unconditional generation after fitting a continuous diffusion prior over \mathbf{z}_0 (e.g., higher perplexity), indicating that the structured bottleneck primarily regularizes latent geometry for the continuous generative stage rather than improving raw auto-encoding fidelity.

We use LLaDA’s demasker f_θ to predict masked tokens. Conditioning is implemented by prepending the available register prefix as continuous embeddings. Given a corrupted state \mathbf{x}_t , we form the demasker input as

$$[\langle \text{START} \rangle, \mathbf{z}_0^{(1)}, \dots, \mathbf{z}_0^{(k)}, \langle \text{END} \rangle, E(\mathbf{x}_t)],$$

where $\langle \text{START} \rangle$ and $\langle \text{END} \rangle$ are learned boundary embeddings that delimit the conditioning channel, and $E(\cdot)$ is the token embedding lookup. A common failure mode in conditional demasking is that the demasker under-utilizes the conditioning signal whenever sufficient local context is available. To explicitly prevent this collapse, we employ an *all-mask utilization schedule*: with probability of 0.25 we set \mathbf{x}_t to the fully-masked sequence (all tokens replaced by $[M]$) while still providing \mathbf{z}_0 , forcing f_θ to reconstruct from the register prefix rather than from token-to-token correlations.

Prepending continuous registers changes the positional layout seen by the Transformer. We therefore keep RoPE as the base positional encoding and introduce a lightweight two-axis variant inspired by MRoPE (Wang et al., 2024). The model marks each input position as belonging either to the *register prefix* (including the boundary tokens) or to the *text stream*, and applies rotary embeddings accordingly. Text tokens use the standard RoPE rotation with a global absolute position index. For prefix tokens, we additionally define a prefix-local index that resets at the beginning of the prefix block (i.e., counts positions within the register segment). We then allocate a small subset of rotary frequency pairs to use the additional axis inside the prefix, while the remaining channels continue to use the global axis. This cleanly separates “where” within the register block from absolute positions in the text, stabilizing prefix–text attention and reducing positional shift artifacts under continuous-prefix conditioning.

Finally, we train h_ϕ and f_θ jointly under the standard masked-diffusion corruption process: we sample a masking level t , construct \mathbf{x}_t by masking a subset of tokens, compute $\mathbf{z}_0 = h_\phi(\mathbf{x}_0)$, and optimize cross-entropy on masked positions, i.e., maximize the conditional likelihood $p_\theta(\mathbf{x}_0 \mid \mathbf{x}_t, \mathbf{z}_0)$.

Referring to the technical side of this training process, we optimize using AdamW with learning rate 2.5×10^{-5} and weight decay 0.1, for a total of 3 epochs with global batch size 512. Training is staged for stability: we first warm up the encoder by updating only h_ϕ for the first 4000 optimization steps while keeping the decoder f_θ frozen, and then unfreeze f_θ and continue training both components jointly for the remainder of training.

D. The Role of Robustness in Training the Encoder-Demasker

Our guided demasker is conditioned on a continuous register bank \mathbf{z}_0 . Since this conditioning channel is later driven by *synthesized* latents in our hybrid generators, it is important that decoding does not hinge on an overly precise register configuration. We therefore test how performance changes when the registers are perturbed at inference. Concretely, for each held-out sequence \mathbf{x}_0 we compute $\mathbf{z}_0 = h_\phi(\mathbf{x}_0)$ and add i.i.d. Gaussian noise $\mathbf{e} \sim \mathcal{N}(0, \sigma^2 I)$ to the registers before feeding them to the demasker. We sweep σ and report masked-token prediction quality (Top-1 accuracy and mean cross-entropy) for two training variants: (i) our default Guided Demasker, trained with register-noise augmentation (Alg. 1), and (ii) an otherwise identical ablation trained without this augmentation. Figure 7 shows that noise-augmented training yields a markedly smoother degradation as σ increases, while the no-augmentation variant deteriorates much earlier and more sharply. Overall, these results indicate that latent-noise augmentation makes the conditioning interface more stable to perturbations in register space.

We also examine how the learned register space behaves along continuous paths between real samples. We sample 1000 random program pairs (each ≈ 256 tokens), encode each program into registers $\mathbf{z}_0^{(a)}$ and $\mathbf{z}_0^{(b)}$, and form linear interpolations $\mathbf{z}_\alpha = (1 - \alpha)\mathbf{z}_0^{(a)} + \alpha\mathbf{z}_0^{(b)}$ for $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$, followed by the same per-register normalization used throughout. Each \mathbf{z}_α is decoded with the MDM decoder, and we compute MAUVE and generation perplexity, aggregating over the 1000 decoded samples per α . Figure 8 summarizes the results. Quality remains high across the entire interpolation path: even at the midpoint between two *random* programs ($\alpha = 0.5$), the decoded samples retain strong MAUVE and only a moderate increase in perplexity (numbers). The endpoints are, unsurprisingly, the easiest points to decode, but the overall curve indicates that linear mixing of two unrelated register banks still lands in a region that the decoder can map to fluent, diverse code. This supports the view that the register representation is well-behaved under continuous changes and that decoding is not fragile to such latent perturbations.

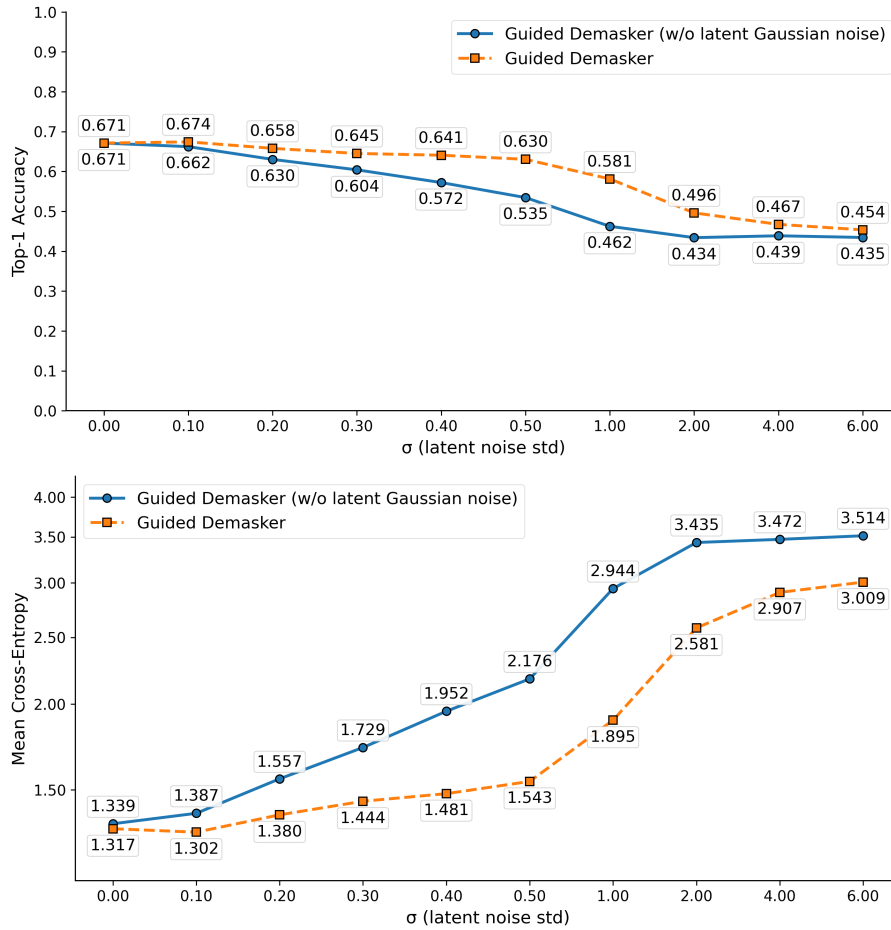


Figure 7. Stability under latent perturbations. We add i.i.d. Gaussian noise of standard deviation σ to the conditioning register bank at evaluation time and report masked-token Top-1 accuracy (top) and mean cross-entropy (bottom). The Guided Demasker trained with register-noise augmentation degrades gradually as σ increases, while the variant trained without this augmentation deteriorates substantially earlier and more sharply. These results refer to 90% masking.

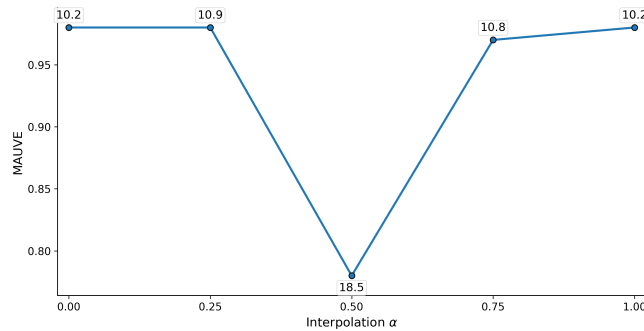


Figure 8. Register-space interpolation aggregated over 1000 program pairs. We linearly interpolate between two encoded register banks and decode at $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$. We report MAUVE (higher is better); numbers denote Generative Perplexity (lower is better), averaged over the same decoded sets. Quality remains high across the interpolation path, including at the midpoint between two random programs.

E. Additional Details on the Continuous Diffusion

Given the trained encoder $h_\phi(\cdot)$, each discrete sequence \mathbf{x}_0 is mapped by it into a *register bank* $\mathbf{z}_0 = h_\phi(\mathbf{x}_0) \in \mathbb{R}^{K \times d}$, where nested dropout encourages a progressive organization across the register index. We model the distribution of these registers by training a continuous diffusion model directly in this domain. Algorithm 5 provides a description of this training procedure.

More concretely, we define a standard forward noising process $q(\mathbf{z}_t | \mathbf{z}_0)$ over the register bank, producing \mathbf{z}_t at diffusion time $t \in [0, 1]$, and learn a DiT-style denoiser g_ψ . As the denoiser backbone we use a PixArt-style Transformer adapted to 1D sequences of register tokens: a compact 28-layer variant with roughly 400M parameters (Chen et al., 2023). We adopt the \mathbf{x}_0 -prediction parameterization, where the denoiser directly predicts the clean register bank,

$$\hat{\mathbf{z}}_0 = g_\psi(\mathbf{z}_t, t),$$

and train with a mean-squared error objective averaged over registers,

$$\mathcal{L}_{\text{cont}}(\psi) = \mathbb{E}_{\mathbf{z}_0, t} \left[\|g_\psi(\mathbf{z}_t, t) - \mathbf{z}_0\|_2^2 \right], \quad \mathbf{z}_t \sim q(\cdot | \mathbf{z}_0).$$

We use the standard DDPM (Ho et al., 2020) objective with a cosine noise schedule, and sample using DDIM (Song et al., 2021).

Algorithm 5 Continuous-Then-Discrete Training

- 1: **Input:** data $\mathbf{x}_0 \sim P(\mathbf{x})$,
 - 2: Sample $t \sim \mathcal{U}([0, 1])$
 - 3: $\mathbf{z}_0 = h_\phi(\mathbf{x}_0)$
 - 4: $\mathbf{z}_t = \text{Forward}(\mathbf{z}_0, t)$
 - 5: $\hat{\mathcal{L}}(\phi, \psi) = \|g_\psi(\mathbf{z}_t, t) - \mathbf{z}_0\|_2^2$
 - 6: Backpropagate on $\nabla_\psi \hat{\mathcal{L}}(\phi, \psi)$ and run optimizer
-

Nested dropout induces a progressive ordering over registers: early registers must remain informative even when later ones are absent, while later registers provide refinements. We mirror this structure in the continuous diffusion stage using *block-wise timestep offsets* aligned with the same geometric grouping used by nested dropout. Concretely, we partition the K registers into contiguous blocks whose sizes grow geometrically (e.g., 1, 2, 4, 8, ...). For a sampled base diffusion time t , we add a block-dependent offset Δ_b (increasing with the block index) and train all registers in block b at the effective time

$$t^{(b)} = \text{clip}(t + \Delta_b), \quad \Delta_1 < \Delta_2 < \dots < \Delta_B.$$

Importantly, all blocks are trained across the full range of diffusion times; the offsets do *not* restrict a block to a narrower noise regime. Rather, they impose a consistent *relative ordering* of effective times across blocks for the same base t : later blocks are always evaluated at a larger effective time than earlier blocks. This aligns the continuous diffusion objective with the progressive register ordering induced by nested dropout, encouraging the denoiser to recover the information carried by early registers before relying on later registers for refinements.

ConThenDisc uses an *unconditional* continuous diffusion prior over register banks. We first construct a dataset of clean registers $\mathbf{z}_0 = h_\phi(\mathbf{x}_0)$ from the frozen encoder, and train a DDPM in register space with \mathbf{x}_0 -prediction: we sample $t \in [0, 1]$, draw $\mathbf{z}_t \sim q(\mathbf{z}_t | \mathbf{z}_0)$, and optimize

$$\mathcal{L}_{\text{cont}}(\psi) = \mathbb{E}[\|g_\psi(\mathbf{z}_t, t) - \mathbf{z}_0\|_2^2].$$

At generation time, we sample $\hat{\mathbf{z}}_0$ by running the reverse process from Gaussian noise, and decode it into tokens using the guided demasker conditioned on $\hat{\mathbf{z}}_0$.

ConWithinDisc trains a *conditional* continuous diffusion model that produces register banks consistent with a partially observed sequence. Training proceeds as follows: Given a clean sequence \mathbf{x}_0 , we sample a masking ratio $r \sim \text{Unif}[0, 1]$ and construct a partially masked sequence $\mathbf{x}^{(r)}$ by masking an r fraction of tokens. We then reuse the same encoder h_ϕ to compute a conditioning embedding from $\mathbf{x}^{(r)}$, but apply an attention mask that blocks information flow through masked positions so that the conditioning signal depends only on visible tokens. The diffusion model is conditioned on this

embedding through cross-attention layers, while the diffusion state is a noisy version of the *clean* register bank $\mathbf{z}_0 = h_\phi(\mathbf{x}_0)$. Concretely, we sample $t \in [0, 1]$, draw $\mathbf{z}_t \sim q(\mathbf{z}_t | \mathbf{z}_0)$, and train

$$\mathcal{L}_{\text{cond}}(\psi) = \mathbb{E}[\|g_\psi(\mathbf{z}_t, t | h_\phi(\mathbf{x}^{(r)})) - \mathbf{z}_0\|_2^2].$$

This objective teaches the model to denoise registers while respecting the partial evidence provided by visible tokens. At inference, we fix the masking ratio to $r = 0.5$ to form the conditioning input, sample a register bank from the conditional diffusion model, and pass it to the guided demasker to complete the sequence.

To compare the unconditional (*ConThenDisc*) and conditional (*ConWithinDisc*) continuous diffusion models, we report precision–recall statistics using the PRDC framework of (Naeem et al., 2020). Specifically, we generate 5,000 samples and represent each sample by $K=128$ register vectors. We treat each register position as an independent feature space: for every register index $j \in \{1, \dots, K\}$, we compute PRDC between generated and real samples using the corresponding vectors $\{\mathbf{z}_0^{(j)}\}$, and report the final score by averaging each PRDC metric across all 128 registers. For the unconditional model, we obtain precision = 0.956 ± 0.010 , recall = 0.651 ± 0.048 , density = 1.105 ± 0.099 , and coverage = 0.870 ± 0.021 . The conditional model substantially improves recall to ≈ 0.74 (with similar precision), indicating better coverage of the real-data manifold.

To directly verify that conditioning improves denoising, we measure reconstruction quality across diffusion times by computing PSNR between the denoiser prediction and the ground-truth registers as a function of t . Specifically, we compare $\text{PSNR}(g_\psi(\mathbf{z}_t, t), \mathbf{z}_0)$ for the unconditional model against $\text{PSNR}(g_\psi(\mathbf{z}_t, t | h_\phi(\mathbf{x}^{(r)})), \mathbf{z}_0)$ for the conditional model. As shown in Figure 9, the conditional denoiser attains higher PSNR over a wide range of timesteps, indicating that it effectively leverages the partial-token evidence provided through $h_\phi(\mathbf{x}^{(r)})$.

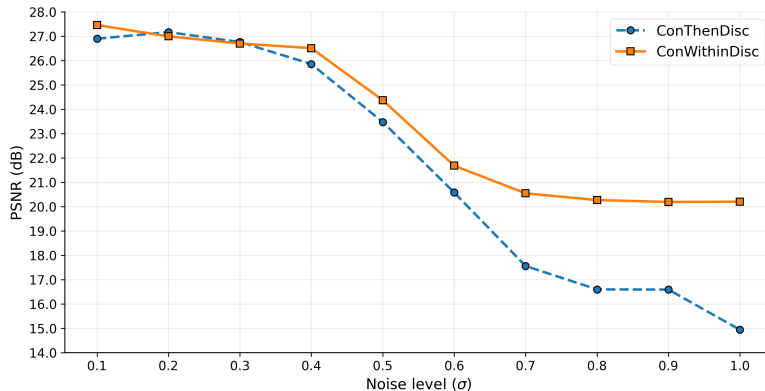


Figure 9. PSNR of denoiser predictions against ground-truth registers across diffusion time. Conditioning (*ConWithinDisc*) yields consistently higher PSNR than the unconditional model (*ConThenDisc*), demonstrating improved reconstruction from noisy registers.

F. Additional Details on the MDM-based AutoEncoder

Figure 10 and Algorithm 6 present the MDM-based autoencoding procedure. Note that lines 3-9 in this algorithm represent a regular MDM, empowered by the \mathbf{z}_0 guidance.

Table 2 is similar to Table 1, referring to sequences of length 512 tokens. The conclusions drawn from both tables are quite similar: Nearly perfect synthesized text is obtained, even for very low NFE and for larger block sizes.

G. More on *ConThenDisc* and *ConWithinDisc*

Figure 11 summarizes three ablations that characterize the interaction between our continuous-register conditioning and the discrete LLaDA demasking stage, all referring to sequence length of 512 tokens. More specifically,

(i) **Varying the number of registers (*ConThenDisc* baseline).** We first vary the number of registers used to represent the continuous conditioning under *ConThenDisc*, taking $K=8$ as the default setting. Increasing the register budget degrades the performance: $K=16$ remains close to the baseline, while larger settings ($K=32$ and $K=64$) consistently degrade generation

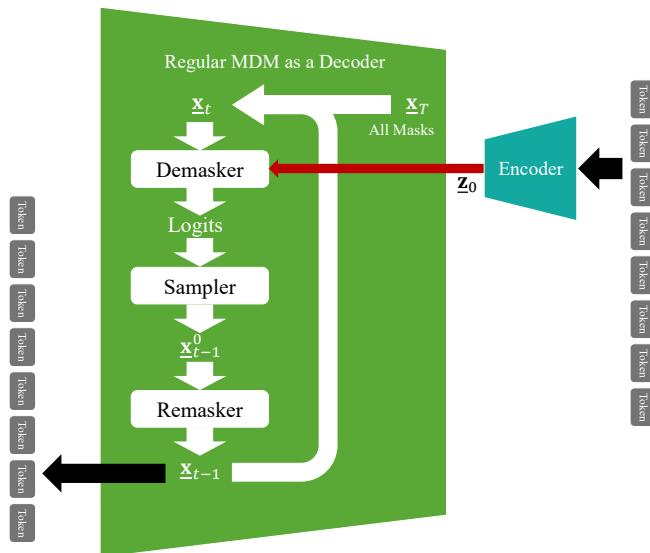


Figure 10. The proposed autoencoder: a token sequence \mathbf{x}_0 is encoded to a latent representation \mathbf{z}_0 , and its decoding is done via a regular MDM.

Algorithm 6 Discrete-Continuum-Discrete Autoencoder

- 1: **Input:** Clean sequence $\mathbf{x}_0 \sim p_{\text{data}}$, encoder h_ϕ , demasker f_θ , number of steps T
 - 2: Encode to continuous latent space: $\mathbf{z}_0 \leftarrow h_\phi(\mathbf{x}_0)$
 - 3: Initialize from fully masked vector: $t \leftarrow 1, \mathbf{x}_t \leftarrow \mathbf{m}$
 - 4: **while** $t > 0$ **do**
 - 5: Predict clean sample: $\hat{\mathbf{x}}_0 \sim f_\theta(\mathbf{x}_t, t, \mathbf{z}_0)$
 - 6: Decrement time: $t \leftarrow t - 1/T$
 - 7: Apply forward masking: $\mathbf{x}_t \leftarrow \text{Forward}(\hat{\mathbf{x}}_0, t)$
 - 8: **end while**
 - 9: **return** Reconstructed sequence $\hat{\mathbf{x}}_0$
-

quality. A plausible explanation is that learning additional *fine* registers is harder and does not necessarily benefit the downstream demasking process. Once early registers capture the dominant global signal, later registers are expected to encode residual refinements; however, these higher-index registers are more weakly constrained, can become redundant or noisy, and may fail to provide complementary guidance. In this regime, adding registers can dilute the conditioning or introduce instability rather than improving the final sample quality.

(ii) **When to refresh the conditioning (ConWithinDisc).** Next, we ablate the timing of refreshing the continuous embedding during discrete demasking by updating after a fraction $\alpha \in \{0, 0.25, 0.50, 0.75\}$ of the LLaDA steps. We include @0 as a no-refresh reference corresponding to *ConThenDisc* (plotted with a small horizontal shift for readability). The results indicate that the midpoint update (@0.50) is the most effective operating point. Updating too early (@0.25) behaves closer to @0: at the beginning of demasking, the discrete trajectory has not yet accumulated sufficient structure for a refreshed embedding to provide meaningfully different guidance. Updating too late (@0.75) offers limited benefit because fewer decisions remain; by that stage many tokens are already committed, and committed tokens remain unchanged during LLaDA demasking, reducing the degrees of freedom through which refreshed conditioning can influence the final generation. Overall, the midpoint refresh provides the best trade-off between having enough structure to exploit and still having sufficient flexibility to steer the remaining undecided tokens.

(iii) **Block-wise timestep offsets.** Finally, we evaluate *block-wise timestep offsets* in the continuous diffusion stage, introduced to promote a clearer hierarchical organization across registers and to better align training with the progressive ordering induced by nested dropout. Empirically, these offsets are important for stable training and for strengthening the intended coarse-to-fine relationship among registers, while their effect at inference is moderate (approximately 0.03 MAUVE in our ablations).

Table 2. **Autoencoder:** Performance measured via Generative Perplexity, and recovery error evaluated via Bert-Score and Character Error Rate (CER). The Table explores varying hyper-parameters of the MDM decoder, for a generative length of 512 tokens.

Block	NFE	Gen-PPL	Bert-Score	CER
32	16	35.477	0.833	1.149
32	32	15.554	0.877	0.237
32	64	9.630	0.904	0.146
32	128	8.170	0.959	0.118
32	256	7.583	0.960	0.105
64	8	33.738	0.764	1.100
64	16	29.314	0.821	1.196
64	32	11.834	0.935	0.183
64	64	8.989	0.948	0.132
64	128	7.989	0.955	0.113
64	256	7.534	0.959	0.103
128	8	35.547	0.856	1.180
128	16	18.715	0.945	1.229
128	32	10.806	0.953	1.218
128	64	7.497	0.957	1.152
256	8	22.527	0.935	1.240
256	16	11.828	0.948	0.218
256	32	9.277	0.955	0.176
256	64	8.178	0.959	0.158
256	128	7.646	0.960	0.149
512	8	13.146	0.877	1.337
512	16	9.820	0.904	0.189
512	32	8.447	0.959	0.163
512	64	7.832	0.960	0.152
512	128	7.517	0.957	0.145

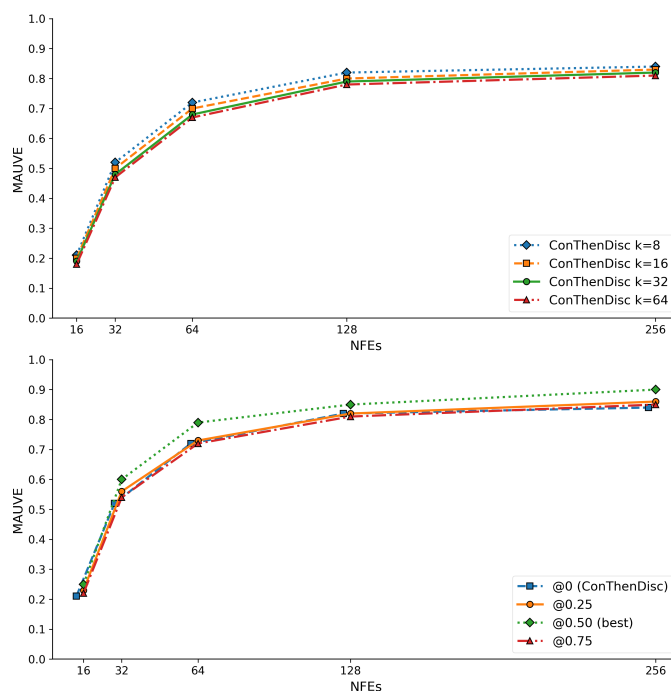


Figure 11. **Ablations at length 512. Top:** register-count ablation under *ConThenDisc* (varying K). Increasing the number of registers beyond the baseline does not improve and can degrade performance, suggesting that learning and exploiting additional fine registers is challenging and not necessarily beneficial for discrete demasking. **Bottom:** conditioning refresh ablation, updating after a fraction α of LLaDA steps. The midpoint update ($@0.50$) is most effective; updating too early resembles the no-refresh baseline ($@0$, i.e., *ConThenDisc*), while updating too late leaves limited room to affect generation as more tokens are already committed and remain unchanged during demasking.

Figure 12 adds to the content of Figure 6, covering the case of generated sequences of length 256 tokens. The conclusions drawn from this graph are quite similar to the ones drawn earlier, namely, *ConThenDisc* improves substantially over the base LLaDA, and *ConWithinDisc* further adds to this improvement. In this case, base LLaDA that uses 256 NFE produces (i.e. generating one token at a time) parallels *ConWithinDisc* that uses 40 NFE, offering a $\times 6$ speedup and beyond.

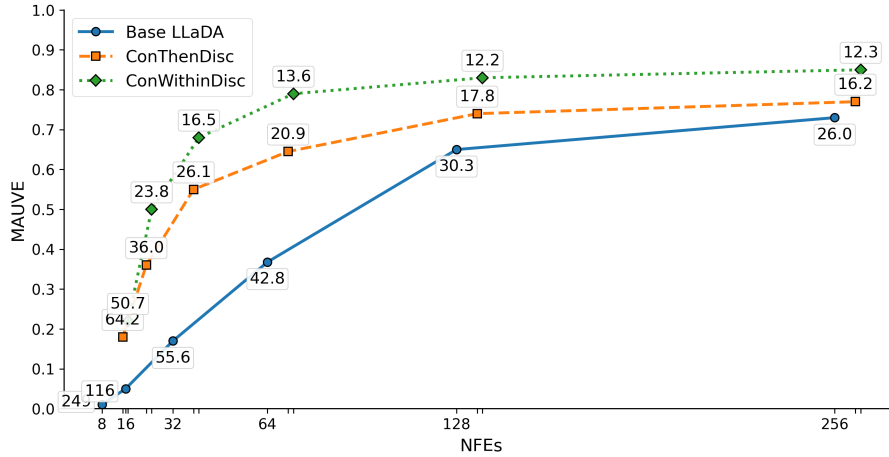


Figure 12. **Text Generation:** MAUVE and Generative Perplexity of generated text with base LLaDA, *ConThenDisc* and *ConWithinDisc*. The sequence generated is of length 256 tokens, handled as one block, and the number of registers in the embedding is $K = 8$.

```

Sample 1

import sys
sys.path.append('.')
from tests.advanced.make_test_data_01 import *
from modules.AddConstant import *
from modules.SimpleFactory import *
from ModuleLogTable import *
import unittest

class TestAddConstant1(unittest.TestCase):
    def setUp(self):
        test_data = [self.H]
        for param in self.H.params:
            submodule = self.H.submodules['test_' + param + '.nu']
            test_data.append(submodule)
        self.C = AddConstant(test_data, 'test_output')

    @staticmethod
    def test_output():
        """
        Test that audio sequence expects to use the output,
        """
        day = datetime(year=2013, month=1, day=30).iso_yday()
        s = get_factory(day)
        s.draw_layout()
        mlt = ModuleLogTable()
        mlt.tk().draw()
        # Test that output variable has been installed
        # It's not, so we can't use our output Feature
        mlt.shutdown()

if __name__ == '__main__':
    unittest.main()
    
```

(a) Sample 1

```

Sample 2

import urllib
import base64
import copy
import numpy as np

class silverman(bplugin.bplugin):
    # ---
    def outDir(self):
        """
        read in silverman's calc
        """
        if 'invl' not in self.modname:
            return np.zeros(8)

        from StringIO import StringIO
        i = open(self.modname)
        o = StringIO()
        result = i.read(2048)
        results = self.parseResults(result)
        i.seek(0)
        return results

    # ---
    def shaderDir(self):
        result = self.outDir()
        logger.debug("reading output {} ...".format(result))
        result = self.parseResults(result)
        result.append("mltResults:{}".format(self.mltResults()))

        if self.mergeResults:
            copyData = result
        else:
            copyData = copy.deepcopy(result)
            copyData['metadata'] = copy.deepcopy(copyData)

        data = urllib.quote(base64.b64encode(copyData))
        return data
    
```

(b) Sample 2

Figure 13. **Qualitative code generations (length 256), samples 1–2.** Continuous registers are sampled with *ConThenDisc* (128 DDIM steps) and decoded with LLaDA (256 discrete denoising steps, single block).

We now turn to present qualitative results, presenting programs generated by our *ConThenDisc* pipeline. Figures 13, 14, 15, 16 focus on programs of length 256 tokens, while Figures 17, 18, 19, 20 show 512-length programs. For all experiments, we first sample continuous registers using DDIM for 128 steps. We then decode the resulting continuous registers into code using the LLaDA demasking decoder with a single block (full-length block size), conditioning the decoder on 8 continuous registers. We vary the number of *discrete* demasking steps and compare 256, 128, 64, and 32 discrete denoising steps for each target length (256 and 512 tokens), reporting two samples per setting.



Figure 14. Qualitative code generations (length 256), samples 1–2. Continuous registers are sampled with *ConThenDisc* (128 DDIM steps) and decoded with LLaDA (128 discrete denoising steps, single block).



Figure 15. Qualitative code generations (length 256), samples 1–2. Continuous registers are sampled with *ConThenDisc* (128 DDIM steps) and decoded with LLaDA (64 discrete denoising steps, single block).

H. Extending the Proposed Algorithms to Conditional Text Synthesis

The two presented text synthesis methods, *ConThenDisc* and *ConWithinDisc*, have been described in the context of unconditional generation, to sample from $P(\mathbf{x})$. Here we discuss their possible extensions to conditional synthesis $P(\mathbf{x}|\mathbf{c})$, i.e., responding to a given prompt \mathbf{c} .

We start with *ConThenDisc*, as described in Algorithm 2. Given a prompt \mathbf{c} , the generated latent \mathbf{z}_0 that initiates the generation process must take it into account in order to lead to a valid eventual answer. Therefore, the continuous diffusion algorithm $G_\psi(\epsilon)$ must be conditioned on \mathbf{c} . Here as well we propose to achieve this conditioning by embedding the

```

Sample 10
<reponame>Zap11/config_obj
"""
Config option class
"""
from src.main_obj.config_obj.config_option import ConfigOption
import os
import subprocess

class ConfigScanner(ConfigOption):
    """
    If a value is provided, update the value
    """

    def set_value(self, (self, value):
        ConfigOption.set_config_value(self, value)

    def set_path(self, path):
        self.dir_list = path

    def _init_files(self):
        for _ in os.listdir(self.dir_list):
            result = subprocess.run([self.config_process], timeout=0)
            return result

    def scan(self):
        """Read all all files for CONFIG_PATH."""
        files = []
        for set in self.CONFIG_PATH.keys():
            if temp_files:
                print '{} in ConfigScanner'.format(temp_files)
                for name in set_obj.append({
                    'path': self.chr.get('name').encode('utf-8')
                })
            files = [x for x in os.listdir('/set/{}'.format(name))
                    return files
    
```

(a) Sample 1

```

Sample 9
from .device import Device
import re
# Provides interface for for temperature devices.

classclass Temperature(Device):
    interval = 'temperature'

    def update(self, interval, scheduler):
        result = self = ('measurement', None)

        if result is None:
            self = ('error', 'sync')

        return result

# Converts a to to
#0 = 0, Monday, Tuesday, , Wednesday,.

def weather(self, subject, scheduler):
    subject = super(self, subject, None)
    if subject is 0:
        return

# This is a valid class in '([A-Z03]+)'
return '([A-Z-Z-Z-Z-Z-Z-Z-Z?[*])'

def result(self, subject, date_from):
    if not str.startswith(date_from):
        member = member.split('.')
        and len(y) != 1:
            member = '%s' % member

def exists(self, type, scheduler):
    locator = self.locator(self, scheduler)
    raise NotImplementedError
    
```

(b) Sample 2

Figure 16. **Qualitative code generations (length 256), samples 1–2.** Continuous registers are sampled with *ConThenDisc* (128 DDIM steps) and decoded with LLaDA (32 discrete denoising steps, single block).

prompt instead of working with it directly, implying that the prompt should be fed as a guidance to the diffusion’s denoiser. Therefore, the training procedure, as described Algorithm 5, changes: line 5 in this algorithm becomes

$$\hat{\mathcal{L}}(\phi, \psi, \kappa) = \|g_{\psi}(\mathbf{z}_t, t, h_{\kappa}(\mathbf{c})) - \mathbf{z}_0\|_2^2, \quad (12)$$

and the training can be done with respect to both ψ (the denoiser parameters) and κ (the prompt embedding parameters). Training of $h_{\kappa}(\cdot)$ can be initialized with $h_{\phi}(\cdot)$ – the original embedding we started with. Indeed, we may consider also an option of avoiding its training altogether by simply assuming $h_{\kappa}(\cdot) = h_{\phi}(\cdot)$.

Once the above training is done, Algorithm 2 (the *ConThenDisc* synthesis algorithm) should change in the following three items:

- Line 3 should become $\mathbf{z}_0 \leftarrow G_{\psi}(\epsilon, h_{\kappa}(\mathbf{c}))$, so that the continuous diffusion uses the prompt to gear the synthesis of the latent \mathbf{z}_0 ,
- Line 5 changes to include the prompt as the prefix of \mathbf{x}_t , i.e. $\mathbf{x}_t := [\mathbf{c}, \mathbf{m}]$. This way, the MDM part of this algorithm operates as usual with the guidance of \mathbf{z}_0 , but also includes the prefix prompt as fixed set of tokens, masking only portions of the answer.
- Line 7 should change as well: The demasker should be retrained with a dataset of prompts and their answers, in order to take the available prompt into account. Referring to Figure 2, this training should be done by encoding the answer (or the prompt and the answer together) via $h_{\phi}(\cdot)$ and feeding it to the demasker for guidance. In addition, the demasker should get a token sequence that contains all the prompt and a partially masked answer, aiming to predict the masked tokens.

Turning to the *ConWithinDisc* algorithm, a similar line of changes is in order. More specifically, the already conditioned denoiser in the training Algorithm 4 should admit another guidance, $h_{\kappa}(\mathbf{c})$. In its inference, as described in Algorithm 3, the prompt should be included in both line 3 as a prefix to \mathbf{x}_t , and in line 6 within the inner continuous diffusion via $\mathbf{z}_0 \leftarrow G_{\psi}(\epsilon, h_{\phi}(\mathbf{x}_t), h_{\kappa}(\mathbf{c}))$. Finally, the prompt-aware guided demasker should be used in line 7.

```

Sample 1
import datetime
import json
import logging
import os
import re
import sys
import time

from code.flag import flag
from code.flag import make
from code.flag import solved
from code.solution import Solution

def get_current_time(f):
    now = datetime.datetime.now()
    local_time = now.astimezone(f.timezone.utc)
    local_time = local_time.astimezone(f.timezone)
    return local_time

def check_status(f, solution):
    current_time = get_current_time(f)
    status = solution.get_status()
    accepted_status = [s for s in status if s in f.accepted_status, f.accepted_status]
    if current_time in accepted_status:
        return True
    return False

def check_time(f, end_time):
    current_time = get_current_time(f)
    if accepted_time <= current_time:
        return True
    return False

def q_agent(control):
    control.protocol_version = 'agent'
    return

def q_maxline(f):
    if len(f) > 1:
        return f + 4 lines will be allowed: {f} [...]
    return f

def q_maxproc(f):
    if len(f) > 1:
        q = int(len(f) // 4)
        return f + {f} can support {q} procs
    return f

def q_validation(control):
    control.protocol_version = 'agent'
    url = 'http://localhost:8000'

    validBaseUrl = '/api/a/a/a/a/valid'
    validBaseUrl = '/api/a/a/a/a/valid'

# agent server version
API_KEY_MAX_QM = 100000
API_KEY_DEFAULT_MAX_QM = '4400'
def request_time(server):
    server.protocol_version = 'agent'
    return server.send('/api/a/a/a/a/node', '/status/current')

def request_status(server, tier, node):
    return server.get_response(control.validBaseUrl, tier.name(), node.name())

def request_status_time(server, tier, node):
    timeInfo = server.get_response(control.validBaseUrl, tier.name(), node.name())
    return timeInfo.time()
    
```

(a) Sample 1

```

Sample 2
# Author: <NAME>
# Website: http://www.cammlab.com
# License: Apache License 2.0

from poc.algorithms import *
from poc.utils import *

class SimulatedDataset(object):
    def __init__(self, n_rows, n_cols, n_workers, batch_size, train_size=None, test_size=None):
        self.n_rows = n_rows
        self.n_cols = n_cols
        self.n_workers = n_workers
        self.batch_size = batch_size
        self.train_size = train_size
        self.test_size = test_size

    def train_dataset(self, n_workers=1, batch_size=800):
        if self.train_size is None:
            self.train_size = self.n_rows
            return self.create_dataset(self.train_size, n_workers, batch_size)

    def create_df(self, n_rows, n_workers, batch_size):
        try:
            df = self.create_dataset(n_rows, n_workers, batch_size).as_ufdf()
        except:
            pass
            df = df.withColumn(lambda c: self.create_bit(True, '', 28).as_ufdf())
        return df

    def get_dates(start_batch, start_train, end_train, n_workers=1, batch_size=800):
        print('random_density:', start_batch, start_train, end_train)
        min = get_random_df(start_batch, end_train)
        end_date = datetime.now()
        max = round(float(end_date + start_date) / 2, 1)
        start_range = min + max * (1 - start_train)
        end_range = min + max * (1 - end_train)
        return start_range, end_range

    def get_random_df(start_batch, start_train, end_train, n_workers=1, batch_size=800):
        print('random_density:', start_batch, start_train, end_train)
        m = Model(
            ('dataset', RandomV3),
            ('epoch', 5),
            ('epoch', 2),
            ('phase', ShiftShuffle),
        )
        return (
            m.get_model(),
            get_model(n_workers),
            get_activation(activmax_v3),
            get_output_df(batch_size, batch_size)
        )
    
```

(b) Sample 2

Figure 17. Qualitative code generations (length 512), samples 1–2. Continuous registers are sampled with ConThenDisc (128 DDIM steps) and decoded with LLaDA (256 discrete denoising steps, single block).

```

Sample 1
__author__ = '<NAME>'
EMAIL = ''

import re
from django.http import HttpResponse
from django.template.loader import render

class AsbGenerator(object):
    def __init__(self, html):
        self.html = html

    def __str__(self):
        return self.DOCTYPE

    DOCTYPE = '<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN>'

    def generate(self, request, log, count = len(self.html)):
        l = self.html.find('<head>')
        u = self.html.find('<body>')
        body = self.html[l:u]
        r = self.html.find('<tr>')
        for i in range(0, count):
            tr_count = self.html.find('<tr>') * self.html.find('<tr>')
            td_count = self.html.find('<td>')
            th_count = self.html.find('<th>')

            FuelRate = self.html.find('<tr>') * self.html.find('<tr>')
            FuelUsed = self.html.find('<tr>') * self.html.find('<td>')
            # FuelTotal = FuelRate * FuelUsed

        def generate(self, request, log):
            chassis = chassis[i]
            measurements = chassis[measurements]
            fuel = chassis[fuel]
            phore = chassis[phore]

            fuelextent = '<td style="border: 0.1//0.0">' + fuel + '</td>'
            total = len(self.html)

            s_type = '<script type="text/f32">'
            title = '<title>S/R/SUCCESS</title>'
            root = self.html.find('//data/root')[0]

            thead = [
                [measurements],
                [fuel],
                [phore],
            ],
            .append(
                measurements=measurements,
                fuel=fuel,
                phore=phore,
            )
        ]

        body = render('log/asb_mutable.html', e)
        except Exception as e
    
```

(a) Sample 1

```

Sample 5
from copy import deepcopy

import numpy as np
from refalcon.metrics.base import MetricsBase
from refalcon.settings import *

class ChordMetrics(MetricsBase):
    def __init__(self):
        super(ChordMetrics, self).__init__()

        # load models
        self.chord_model = self.rcmhub.rcrchord()
        self.rcbone_model = self.rcmhub.rcmbone.rcmbone()

    defproperty_reference_attributes(self):
        key = self.reference._dict_["path"] # returns full_table_name
        ref_attrs = getattr(self.reference, key)
        return ref_attrs

    defproperty_meta_attributes(self):
        key = self.meta._dict_["path"] # returns meta_table_name
        meta_attrs = getattr(self.meta, key)
        return meta_attrs

    defproperty_property_attributes(self):
        key = self.properties.keys()[0] # returns ION, property_name
        prop_attrs = getattr(self.properties, key)
        return prop_attrs

class MissingAttributes(MetricsBase):
    ATTRIBUTES = ["chord", "name", "number"]

    def __init__(self, **kwargs):
        super(ChordMetrics, self).__init__(n_asr=N_ASR, number=CHORD_NUM,
        **kwargs)

    def update_attributes(self, key):
        if key == "chord":
            chord_model = self.chord_model
            if key == "rcmbone":
                chord_model = self.rcmbone_model

            self.chord = deepcopy(chord_model)
            super(ChordMetrics, self).update_attributes({'name': {}, "path": key})

    def get_attributes(self, ):
        kwargs = kwargs.get("")
        NP = np.random.choice(NP)
        self.chord = deepcopy(NP)

        rds_instance = np.random.choice(RDS)
        self.chord = deepcopy(rds_instance)
        self.name = rds_instance["name"]

        return {"NP": NP,
                "n_asr": str(self.n_asr),
                "chord": self.chord,
                "name": self.name,
                "number": self.number}
    
```

(b) Sample 2

Figure 18. Qualitative code generations (length 512), samples 1–2. Continuous registers are sampled with ConThenDisc (128 DDIM steps) and decoded with LLaDA (128 discrete denoising steps, single block).

