Non-Markovian Discrete Diffusion with Causal Language Models

Yangtian Zhang[1,†] Sizhuang He[1,†]
Daniel Levine[1] Lawrence Zhao[1] David Zhang[1] Syed Asad Rizvi[1] Shiyang Zhang[1] Emanuele Zappala[2]
Rex Ying[1] David van Dijk[1]
[1] Yale University; [2] Idaho State University
[†]Equal contribution

Yale

Introduction

Discrete diffusion models offer flexible and controllable generation for structured sequences but typically rely on the Markov assumption, conditioning each step only on the current state. We propose **CaDDi**, a causal discrete diffusion model that conditions on the entire generative trajectory, unifying sequential and temporal modeling within a non-Markovian framework.

Contributions

- Introduced a **non-Markovian discrete diffusion framework** where each denoising step incorporates the full generative trajectory, improving inference robustness.
- Proposed **CaDDi, a causal discrete diffusion model that unifies sequential and temporal modeling** within a non-Markovian diffusion framework. Its further variation CaDDi-AR **generalizes traditional causal language models** as a special case and can **seamlessly adopt pretrained LLMs for discrete diffusion**, enabling more controllable and structured generation.
- Quantitative results show that CaDDi outperforms recent discrete diffusion models, achieving **lower generative perplexity** on language datasets and **stronger reasoning capabilities** when leveraging a pretrained LLM.

Non-Markovian Discrete Diffusion

Goal: Relax the Markov assumption in discrete diffusion by introducing causal dependencies across timesteps.

1. Background: Discrete Diffusion

Standard discrete diffusion models such as **D3PM** define a Markovian noising process:

$$q(x_t|x_{t-1}) = Q_t(x_t|x_{t-1}),$$

where Q_t is a pre-defined transition matrix. The reverse model $p_\theta(x_{t-1}|x_t)$ learns to denoise one step at a time.

Both the forward and reverse processes are modeled as Markov Chains.

2. Non-Markovian Forward Process

Instead of a Markovian forward process $q(x_t|x_{t-1})$, CaDDi defines

$$q(\mathbf{x}_{0:T}) := q(\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T) = q(\mathbf{x}_0)\prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{0:t-1}) = q(\mathbf{x}_0)\prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_0),$$

where independent noise is injected into the original data x_0 at each timestep t.

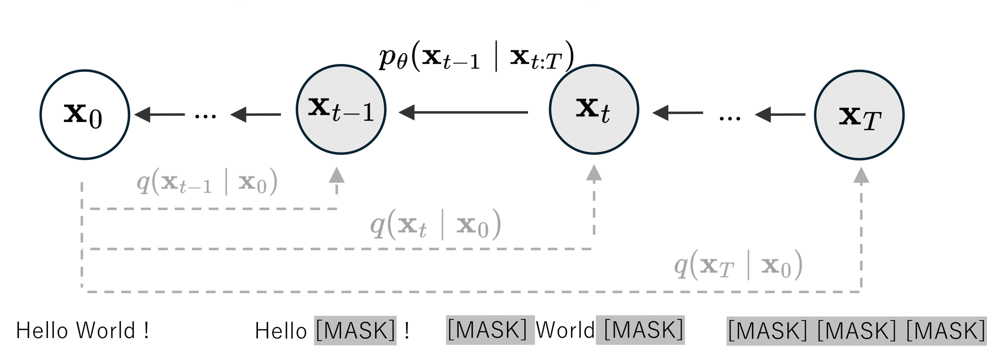


Figure 1. Illustration of Non-Markovian discrete diffusion.

3. Non-Markovian Reverse Process

The posterior of the non-Markovian discrete diffusion model is of the form:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_{t:T}) := q\big(\mathbf{x}_{t-1} \mid \mathbf{x}_{t:T}, \mathbf{x}_0 = \boldsymbol{\mu}_\theta(\mathbf{x}_{t:T}, t)\big) = q\big(\mathbf{x}_{t-1} \mid \mathbf{x}_0 = \boldsymbol{\mu}_\theta(\mathbf{x}_{t:T}, t)\big)$$

4. Autoregressive inference

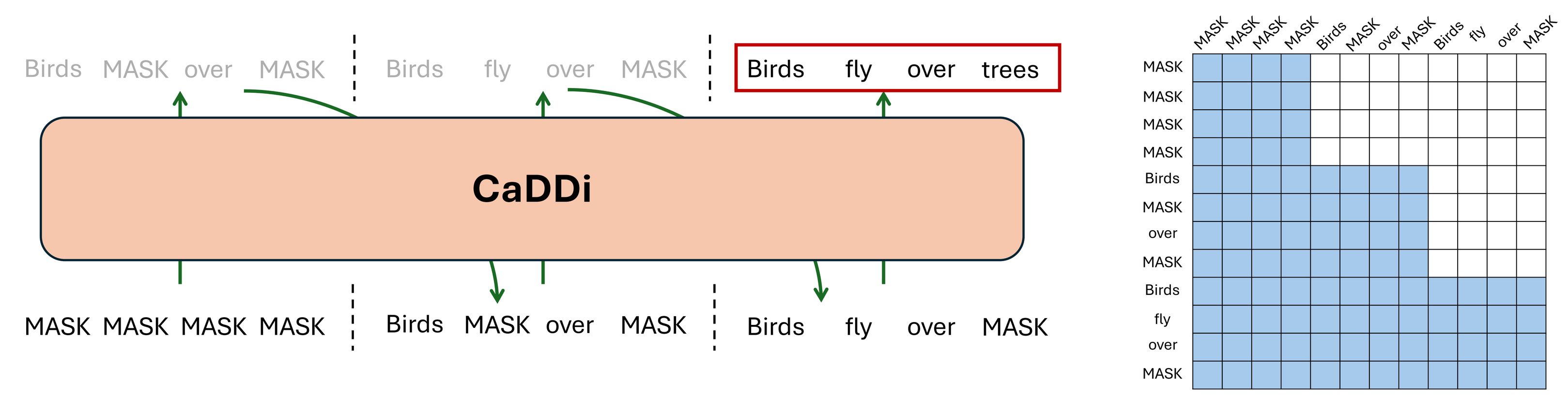


Figure 2. Autoregressive inference of non-Markovian Discrete Diffuion and the corresponding block-level attention mask

5. Evidence Lower Bound (ELBO)

We optimize:

$$\mathcal{L}_{\text{non-markov}} = \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \log p_\theta(\mathbf{x}_0 \mid \mathbf{x}_{1:T}) - \mathrm{KL}\big(q(\mathbf{x}_T \mid \mathbf{x}_0) \| p_\theta(\mathbf{x}_T)\big) - \mathcal{L}_T$$

where $\mathcal{L}_T = \sum_{t=2}^{T} \mathbb{E}_{q(\mathbf{x}_{t:T}|\mathbf{x}_0)} \mathrm{KL}\big(q(\mathbf{x}_{t-1} \mid \mathbf{x}_0) \| p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_{t:T})\big)$.

CaDDi: Causal Discrete Diffusion Model

Key idea: CaDDi unifies the **sequential** (token order) and **temporal** (diffusion timesteps) dimensions within a single causal Transformer.

1. Unified Sequential–Temporal Modeling

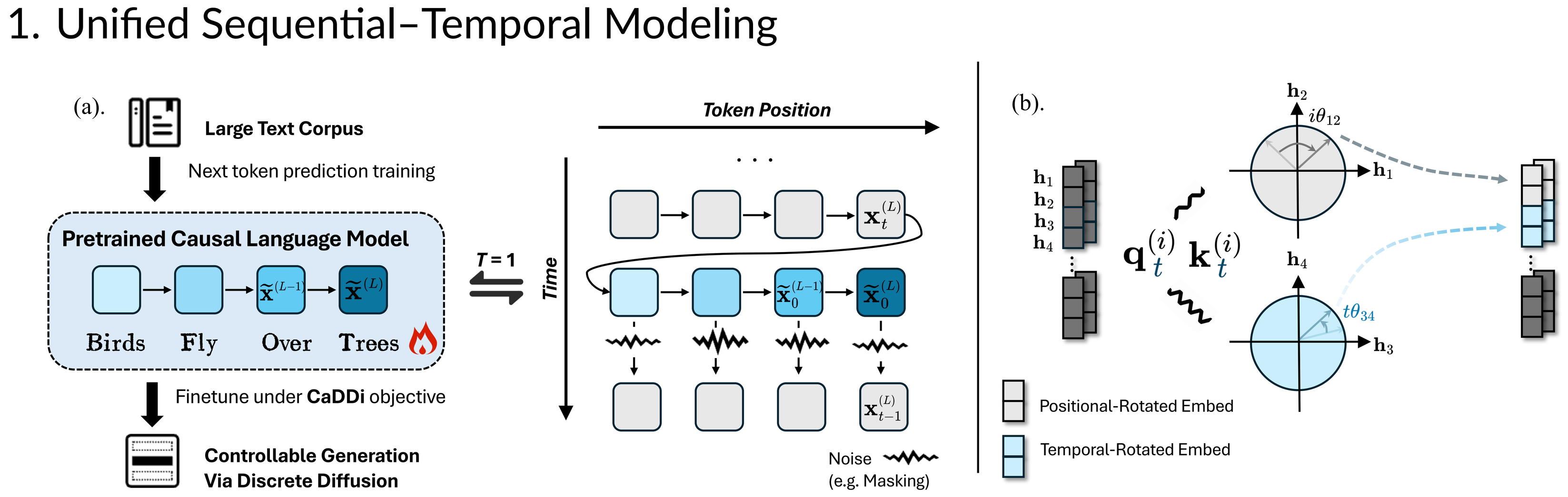


Figure 3. (a): **Unified sequential-temporal modeling** of CaDDi. A traditional autoregressive LM is a special case of CaDDi-AR with T=1. (b): **2D rotary positional encoding**

2. 2D Rotary Positional Encoding

To capture both token and timestep dependencies, we extend 1D rotary embeddings (RoPE) to a 2D variant:

$$\mathbf{R}_t^{(i)} = \begin{bmatrix} \mathbf{R}_{\text{seq}}^{(i)} & 0 \\ 0 & \mathbf{R}_{\text{time}}^{(t)} \end{bmatrix},$$

3. CaDDi-AR: Autoregression over Tokens

To better approximate the true posterior, we further factorize:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_{t:T}) = \prod_{i=0}^{L} p_\theta\left(\mathbf{x}_{t-1}^i \mid \mathbf{x}_{t-1}^{0:i-1}, \mathbf{x}_{t:T}\right)$$

enabling token-level autoregressive denoising consistent with decoder-only LMs.

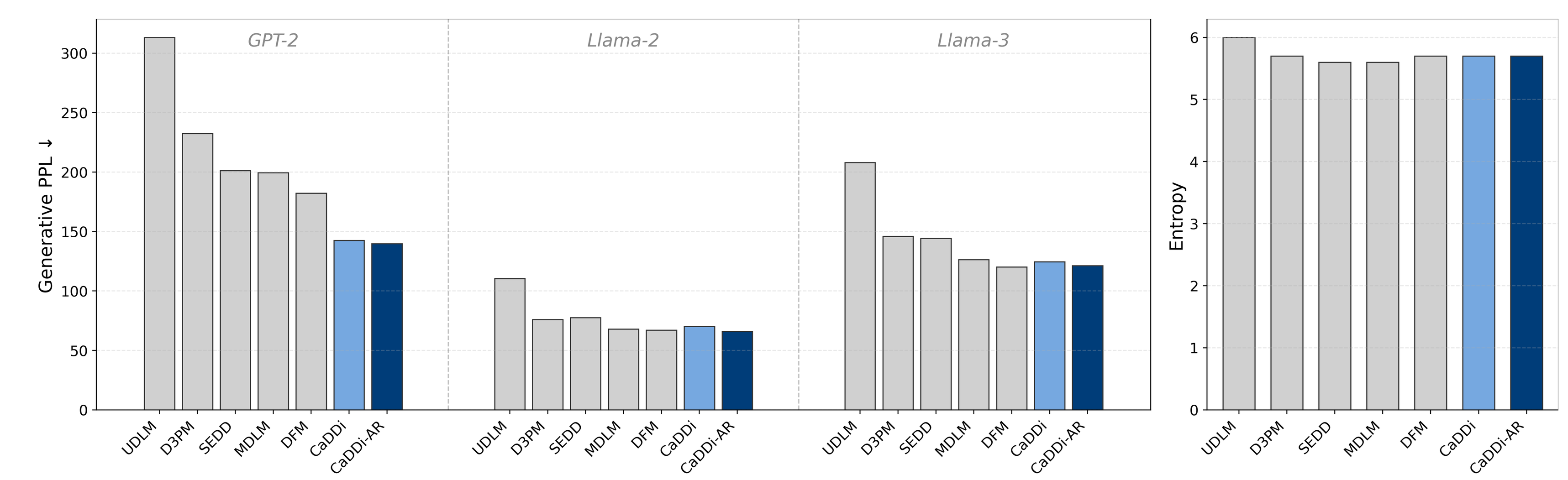
4. Semi-Speculative Decoding

CaDDi-AR reuses the previous timestep's prediction $\tilde{\mathbf{x}}_0^{\text{prev}}$ as a draft for the next step and verifies all tokens in parallel. This reduces $\mathcal{O}(L\times T)$ evaluations to nearly linear in L while preserving generation quality.

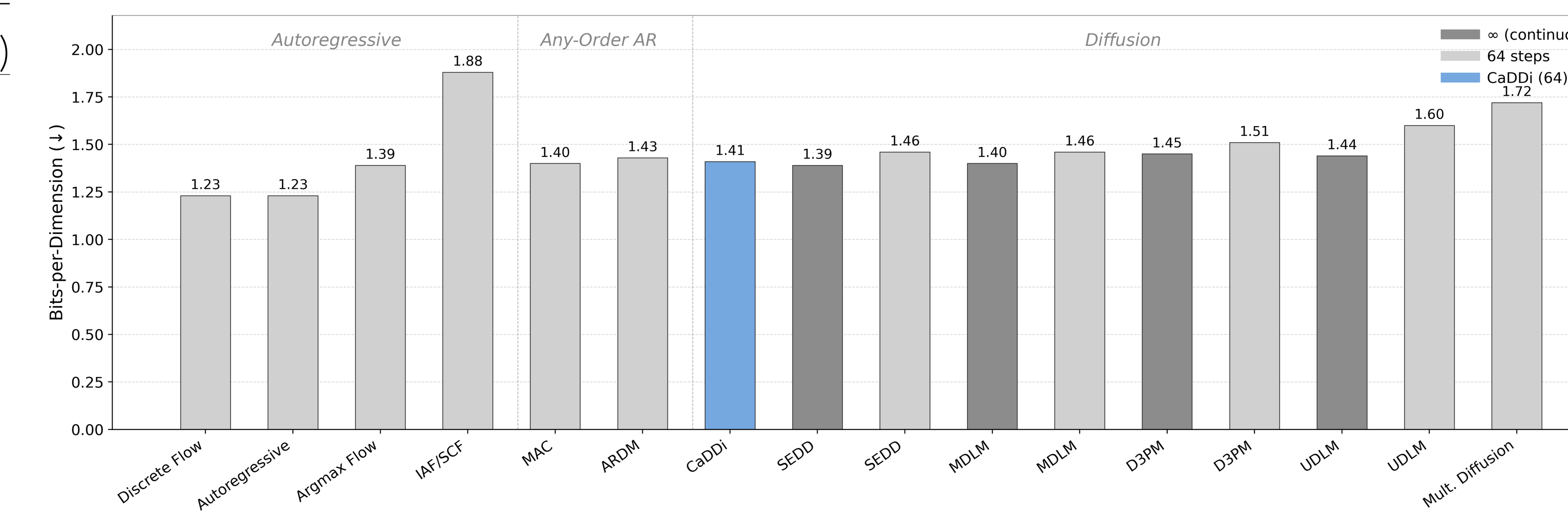
Experiments

All models use 12-layer Transformers trained with identical hyperparameters.

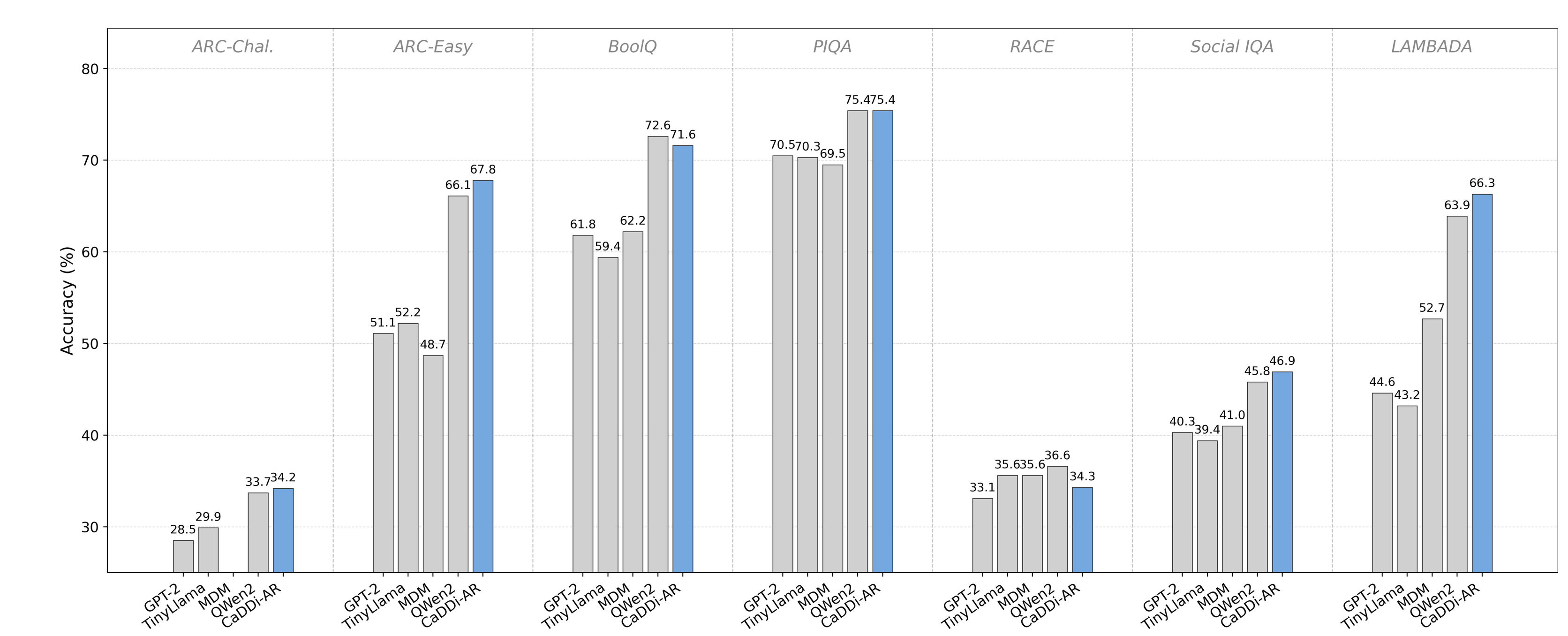
1. One Billion Words (LM1B).
CaDDi achieves the **lowest generative perplexity** while **preserving output diversity**.



2. Text8 Benchmark.
CaDDi achieves the **best bits-per-dimension (BPD)** among discrete diffusion models.

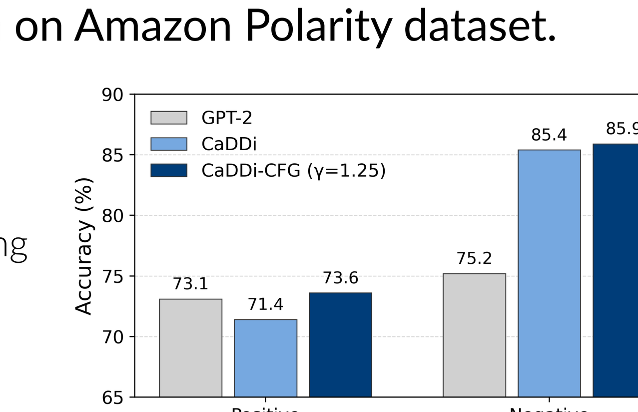


3. Reasoning with Fine-tuned LLMs.
Fine-tuning CaDDi-AR on a 1.5B QWen model yields consistent gains across reasoning datasets.



4. Conditional Text Generation on Amazon Polarity dataset.

CaDDi-CFG achieves sentiment accuracy comparable to fine-tuned GPT-2 while supporting flexible infilling from arbitrary positions.



Correspondence: rex.ying@yale.edu, david.vandijk@yale.edu