Novelty Assessment Report

Paper: Fast Language Generation through Discrete Diffusion Divergence Instruct
PDF URL: https://openreview.net/pdf?id=mtdyZsa47V
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

Fast and high-quality language generation is the holy grail that people pursue in the age of AI. In this work, we introduce **Di**screte **Di**ffusion Divergence **Instruct** (**DiDi-Instruct**), a training-based method that initializes from a pre-trained (masked) discrete diffusion language model (dLLM) and distills a few-step student for fast generation. The resulting DiDi-Instruct model achieves comparable or superior performance to its dLLM teacher and the GPT-2 baseline while enabling up to 64$\times$ acceleration. The theoretical foundation of DiDi-Instruct is a novel framework based on integral KL-divergence minimization, which yields a practical training algorithm. We further introduce grouped reward normalization, intermediate-state matching, and the reward-guided ancestral sampler that significantly improve training stability, model coverage, and inference quality. On OpenWebText, DiDi-Instruct achieves perplexity from 62.2 (8 NFEs) to 18.4 (128 NFEs), which outperforms prior accelerated dLLMs and GPT-2 baseline. These gains come with a negligible entropy loss (around 1\%) and reduce additional training wall-clock time by more than $20\times$ compared to competing dLLM distillation methods. We further validate the robustness and effectiveness of DiDi-Instruct through extensive ablation studies, model scaling, downstream tasks, and the generation of discrete protein sequences. In conclusion, DiDi-Instruct is an efficient yet effective distillation method, enabling language generation in the blink of an eye. We will release our code and models along with the paper.

Core Task Landscape

This paper addresses: **Accelerating Discrete Diffusion Language Models Through Distillation**

A total of **28 papers** were analyzed and organized into a taxonomy with **25 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Discrete Diffusion Distillation Methods**
- **Image Diffusion Distillation Methods**
- **Cross-Domain and Unified Distillation Frameworks**
- **Domain-Specific Distillation Applications**
- **Survey and Overview**

Complete Taxonomy Tree

- Accelerating Discrete Diffusion Language Models Through Distillation Survey Taxonomy
- Discrete Diffusion Distillation Methods
  - KL-Divergence Based Distillation ★ (2 papers)
  - [0] Fast Language Generation through Discrete Diffusion Divergence Instruct (Anon et al., 2026) View paper
  - [4] Ultra-fast language generation via discrete diffusion divergence instruct (Zheng Haoyang, 2025) View paper
  - Dimensional Correlation and Mixture Models (1 papers)
  - [2] Distillation of discrete diffusion through dimensional correlations (Hayakawa Satoshi, 2024) View paper
  - Flow-Matching for Discrete Diffusion (1 papers)
  - [22] FS-DFM: Fast and Accurate Long Text Generation with Few-Step Diffusion Language Models (Monsefi, 2025) View paper
  - Learnable Sampler Distillation (1 papers)
  - [17] Learnable Sampler Distillation for Discrete Diffusion Models (Feiyang Fu, 2025) View paper
  - Gaussian-Guided Discrete Diffusion (1 papers)
  - [5] The diffusion duality (Sahoo, 2025) View paper
  - Autoregressive Teacher Distillation (1 papers)
  - [28] Distilled Diffusion Language Models (R ZHANG, n.d.) View paper
  - Self-Distillation Through Time (1 papers)
  - [14] Beyond Autoregression: Fast LLMs via Self-Distillation Through Time (Gulcehre, 2024) View paper
  - Discrete Diffusion Forcing (1 papers)
  - [8] Diffusion LLMs Can Do Faster-Than-AR Inference via Discrete Diffusion Forcing (Wang Xu, 2025) View paper
- Image Diffusion Distillation Methods
  - Progressive and Multi-Step Distillation (1 papers)
  - [7] Progressive distillation for fast sampling of diffusion models (Salimans, 2022) View paper
  - Score Identity and Distribution Matching (2 papers)
  - [3] Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation (Zhou, 2024) View paper
  - [25] Distribution Matching Distillation Meets Reinforcement Learning (Dengyang Jiang, 2025) View paper
  - Variational and Data-Free Distillation (2 papers)

    - [1] Swiftbrush: One-step text-to-image diffusion model with variational score distillation (Thuan Hoang Nguyen, 2024) View paper
    - [6] Dkdm: Data-free knowledge distillation for diffusion models with any architecture (Miao Zhang, 2025) View paper
    - Stochastic Consistency Distillation (1 papers)
    - [13] SCott: Accelerating Diffusion Models with Stochastic Consistency Distillation (Liu HongJian, 2025) View paper
    - Backward and Reconstruction Distillation (1 papers)
    - [16] Imagine Flash: Accelerating Emu Diffusion Models with Backward Distillation (KÃ¶hler, 2024) View paper
    - One-to-Many Knowledge Distillation (1 papers)
    - [15] Accelerating Diffusion Models with One-to-Many Knowledge Distillation (Zhang, 2024) View paper
    - Scale-Wise and Multi-Resolution Distillation (1 papers)
    - [10] Scale-wise Distillation of Diffusion Models (Starodubcev, 2025) View paper
    - Feature and Classifier Distillation (1 papers)
    - [18] Accelerating Diffusion Sampling with Classifier-based Feature Distillation (Wujie Sun, 2023) View paper
    - Guided Diffusion Distillation (2 papers)
    - [26] On Distillation of Guided Diffusion Models (Chenlin Meng, 2023) View paper
    - [27] LoRA-Enhanced Distillation on Guided Diffusion Models (Golnari, 2023) View paper
    - Idempotent and Score-Based Distillation (1 papers)
    - [24] Score-based Idempotent Distillation of Diffusion Models (Zaman, 2025) View paper
- Cross-Domain and Unified Distillation Frameworks
    - Universal Knowledge Transfer (1 papers)
    - [9] Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models (Luo Wei-jian, 2023) View paper
    - Unified Multimodal Consistency Models (1 papers)
    - [12] UniCMs: A Unified Consistency Model For Efficient Multimodal Generation and Understanding (Xu, 2025) View paper
    - Unified Discrete Diffusion Frameworks (1 papers)
    - [11] Target Concrete Score Matching: A Holistic Framework for Discrete Diffusion (Zhang Rui-xiang, 2025) View paper
- Domain-Specific Distillation Applications
    - Trajectory Prediction Distillation (1 papers)
    - [19] Collaborative-Distilled Diffusion Models (CDDM) for Accelerated and Lightweight Trajectory Prediction (Wang Bing-zhang, 2025) View paper
    - Motion Synthesis Distillation (1 papers)
    - [20] Motion synthesis via distilled absorbing discrete diffusion model (Junyi Wang, 2024) View paper
    - Text-to-Image Distillation Practical Guides (1 papers)
    - [23] Few-Step Distillation for Text-to-Image Generation: A Practical Guide (Yifan Pu, 2025) View paper
- Survey and Overview (1 papers)
    - [21] Diffusion Models Acceleration: A Quick Survey (FN Dei da Filicaia Dotti, 2025) View paper

Narrative

Core task: accelerating discrete diffusion language models through distillation. The field has evolved into several distinct branches that reflect both methodological diversity and domain specialization. Discrete Diffusion Distillation Methods focus on adapting distillation techniques specifically for language and discrete token spaces, often leveraging KL-divergence objectives or specialized sampling strategies tailored to categorical distributions. Image Diffusion Distillation Methods encompass a rich body of work on accelerating continuous diffusion models for visual generation, exploring progressive distillation schemes like Progressive Distillation[7], score-matching approaches such as Score Identity Distillation[3], and one-step or few-step generators including Swiftbrush[1] and Imagine Flash[16]. Cross-Domain and Unified Distillation Frameworks attempt to bridge discrete and continuous settings, proposing architectures or training paradigms that generalize across modalities. Domain-Specific Distillation Applications address tailored challenges in areas like video, audio, or 3D generation, while Survey and Overview papers such as Diffusion Acceleration Survey[21] synthesize emerging trends and open questions across these branches.

Within the discrete distillation landscape, a central tension revolves around balancing sample quality, inference speed, and training stability when moving from continuous to categorical spaces. Diffusion Divergence Instruct[0] sits squarely in the KL-Divergence Based Distillation cluster, emphasizing divergence minimization to compress multi-step discrete diffusion into fewer steps for language modeling. Its closest neighbor, Ultra-fast Divergence Instruct[4], shares this KL-centric philosophy but pushes toward even more aggressive step reduction. In contrast, works like Discrete Diffusion Forcing[8] and Absorbing Discrete Diffusion[20] explore alternative parameterizations or absorbing-state dynamics that sidestep some gradient estimation challenges inherent in categorical distillation. Meanwhile, methods such as DKDM[6] and Diffusion Duality[5] investigate dual formulations or knowledge transfer mechanisms that complement divergence-based objectives. The original paper thus contributes to an active subfield where researchers are refining how to faithfully distill discrete generative processes without sacrificing the expressiveness that makes diffusion models attractive for language generation.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Ultra-fast language generation via discrete diffusion divergence instruct

Authors: Zheng Haoyang, Liu Xin-yang, Haoyang Zheng, Xinyang Liu, Jiang Nan, et al. (16 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Fast and high-quality language generation is the holy grail that people pursue in the age of AI. In this work, we introduce Discrete Diffusion Divergence Instruct (DiDi-Instruct), a training-based method that initializes from a pre-trained (masked) discrete diffusion language model (dLLM) and distills a few-step student for fast generation. The resulting DiDi-Instruct model achieves comparable or superior performance to its dLLM teacher and the GPT-2 baseline while enabling up to 64$\times$ acce...

⚠ Similarity Notice

These papers appear to be the same work or very closely related variants. Both introduce 'DiDi-Instruct' (Discrete Diffusion Divergence Instruct) with identical core methodology based on integral KL-divergence minimization for distilling discrete diffusion language models. The titles, abstracts, technical approach (policy gradient formulation, grouped reward normalization, reward-guided ancestral sampler), experimental setup (OpenWebText, 169M parameter MDLM teacher), and reported results are nearly identical, strongly suggesting they are the same paper or different submission versions.

Contributions Analysis

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: DiDi-Instruct: a training-based distillation method for fast language generation

Description: The authors propose DiDi-Instruct, a novel distillation framework that trains a few-step student model from a pre-trained masked discrete diffusion language model. This method achieves comparable or superior performance to the teacher model while enabling up to 64× acceleration in generation speed.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Ultra-fast language generation via discrete diffusion divergence instruct

URL: View paper

Brief Assessment

Ultra-fast Divergence Instruct[4] presents a similar distillation framework (DiDi-Instruct) for accelerating discrete diffusion language models. Both papers share the same core methodology and appear to be the same work.

---

2. Compressed and Smooth Latent Space for Text Diffusion Modeling

URL: View paper

Brief Assessment

Smooth Latent Space[38] focuses on compressing text representations into a smooth latent space for diffusion modeling, not on distilling discrete diffusion language models for faster generation as DiDi-Instruct does.

---

3. Beyond Autoregression: Fast LLMs via Self-Distillation Through Time

URL: View paper

Brief Assessment

Self-Distillation Through Time[14] focuses on distilling discrete diffusion models for faster generation, but uses a different distillation approach (trajectory matching across time-steps) rather than the integral KL-divergence minimization framework proposed in DiDi-Instruct.

---

4. Learnable Sampler Distillation for Discrete Diffusion Models

URL: View paper

Prior Art Analysis

Learnable Sampler Distillation[17] demonstrates that prior work exists on distilling discrete diffusion models for faster generation. Both papers address the same core problem: accelerating discrete diffusion language models through distillation to enable few-step generation. Learnable Sampler Distillation[17] proposes LSD and LSD+, which train a student sampler to align its intermediate score trajectory with a teacher sampler using learnable coefficients and time schedules. This approach predates the original paper's DiDi-Instruct framework and achieves similar goals of few-step generation with comparable acceleration rates (e.g., LSD+ achieves perplexity improvements at 8-64 NFEs). The candidate paper explicitly addresses distillation of discrete diffusion models for text generation, directly overlapping with the original paper's claimed novelty.

Evidence

Evidence 1 - **Rationale**: Both papers propose training-based distillation methods for discrete diffusion models that train a few-step student from a multi-step teacher to achieve fast generation. The candidate paper's LSD approach directly addresses the same problem space as DiDi-Instruct. - **Original**: we introduce di screte di ffusion divergence instruct (didi-instruct), a training-based method that initializes from a pretrained (masked) discrete diffusion language model (dllm) and distills a few-step student for fast generation. the resulting didi-instruct model achieves comparable or superior p... - **Candidate**: we propose learnable sampler distillation (lsd), a novel approach to train fast and high-fidelity samplers for ddms. lsd employs a distillation approach where a student sampler with a few steps learns to align its intermediate score trajectory with that of a high-quality teacher sampler with numerou...

Evidence 2 - **Rationale**: Both papers report perplexity improvements at similar NFE ranges (8-128 steps), demonstrating that the candidate paper achieves comparable acceleration goals for discrete diffusion language models. - **Original**: on openwebtext, didi-instruct achieves perplexity from 62.2 (8 nfes) to 18.4 (128 nfes), which outperforms prior accelerated dllms and gpt-2 baseline. - **Candidate**: sampler nfes 8 16 32 64 euler 423.109 215.472 72.820 56.218 tweedie 404.881 177.539 64.347 50.151 jys-euler 308.123 125.283 55.842 32.943 jys-tweedie 307.382 127.232 56.382 31.192 lsd-euler 145.490 88.56431.23521.956 lsd-tweedie 168.846 86.282 35.786 21.981 lsd+-euler 128.413 51.76936.800 20.728 lsd...

---

5. Distillation of discrete diffusion through dimensional correlations

URL: View paper

Brief Assessment

Dimensional Correlations[2] focuses on distilling discrete diffusion models by capturing dimensional correlations between elements (e.g., pixel relationships in images, sequential dependencies in language) using mixture models. This is a different technical approach from DiDi-Instruct's integral KL-divergence minimization framework with adversarial discriminators for masked diffusion language models.

---

6. Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation

URL: View paper

Brief Assessment

Score Identity Distillation[3] focuses on distilling pretrained diffusion models for image generation into one-step generators, not on discrete diffusion language models for text generation. The technical domains and methodologies are fundamentally different.

---

7. Inference-Time Diffusion Model Distillation

URL: View paper

Brief Assessment

Inference-Time Distillation[39] focuses on inference-time distillation for image generation using diffusion models, not training-based distillation for discrete language models. The technical domains and approaches are fundamentally different.

---

8. The diffusion duality

URL: View paper

Brief Assessment

Diffusion Duality[5] focuses on establishing theoretical connections between continuous Gaussian diffusion and uniform-state discrete diffusion, then applying curriculum learning and consistency distillation. DiDi-Instruct uses a different distillation approach based on integral KL-divergence minimization with adversarial discriminators for masked diffusion models, representing a distinct technical framework.

---

9. Diffusion LLMs Can Do Faster-Than-AR Inference via Discrete Diffusion Forcing

URL: View paper

Brief Assessment

Discrete Diffusion Forcing[8] focuses on block-wise autoregressive generation with KV cache utilization and parallel decoding, not on distribution-matching distillation via integral KL divergence minimization as in the original paper.

Contribution 2: Theoretical framework based on integral KL-divergence minimization

Description: The authors develop a principled training method grounded in minimizing integral KL-divergence between student and teacher distributions. They reformulate the distillation objective using a policy gradient perspective, deriving a tractable update rule that uses an adversarial discriminator to estimate log-density ratios.

This contribution was assessed against **8 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Towards searching for the best student in a Knowledge Distillation framework

URL: View paper

Brief Assessment

Best Student Search[32] focuses on neural architecture search and hyperparameter optimization for knowledge distillation in general supervised learning settings, not on integral KL-divergence minimization or policy gradient methods for diffusion model distillation.

---

2. On the Design of KL-Regularized Policy Gradient Algorithms for LLM Reasoning

URL: View paper

Brief Assessment

KL-Regularized Policy Gradient[34] focuses on policy gradient methods for LLM reasoning with KL regularization, not on distilling discrete diffusion models for language generation. The candidate addresses RL optimization for reasoning tasks, while the original paper tackles fast sequence generation via diffusion model distillation.

---

3. ADPO: Anchored Direct Preference Optimization

URL: View paper

Brief Assessment

ADPO[33] focuses on preference optimization through reference anchoring in the context of RLHF alignment, not on distilling discrete diffusion models for language generation. The candidate's KL-divergence framework addresses pairwise/listwise preference learning, while the original work develops a policy gradient approach for masked diffusion distillation.

---

4. DISTILLATION AND GENERALIZATION IN DEEP REINFORCEMENT LEARNING

URL: View paper

Brief Assessment

Deep RL Distillation[36] focuses on policy distillation for deep reinforcement learning using KL-divergence between teacher and student policies, not on masked diffusion language models or discrete text generation as in the original paper.

---

5. A Distributional Approach to Controlled Text Generation

URL: View paper

Brief Assessment

Distributional Controlled Generation[35] focuses on controlled text generation from pre-trained LMs using KL divergence minimization with EBM representations and policy gradient training. The original paper addresses distillation of discrete diffusion language models using integral KL-divergence with discriminator-based reward estimation, which is a fundamentally different application domain and technical approach.

---

6. Score and Distribution Matching Policy: Advanced Accelerated Visuomotor Policies via Matched Distillation

URL: View paper

Brief Assessment

Score Distribution Matching[30] focuses on visual-motor policy learning for robotics using diffusion models, not discrete language generation. The technical domains and application contexts are fundamentally different.

---

7. KDRL: Post-Training Reasoning LLMs via Unified Knowledge Distillation and Reinforcement Learning

URL: View paper

Brief Assessment

KDRL[29] focuses on combining knowledge distillation with reinforcement learning for reasoning LLMs using reverse KL divergence and policy gradients. While both papers use policy gradient methods and KL divergence, KDRL[29] addresses a different problem domain (reasoning LLMs with teacher-student training) rather than discrete diffusion language models for fast generation.

---

8. Droid: Learning from offline heterogeneous demonstrations via reward-policy distillation

URL: View paper

Brief Assessment

Droid[31] focuses on offline learning from heterogeneous demonstrations in robotics using reward-policy distillation, not on KL-divergence minimization for model distillation via policy gradients in language generation contexts.

Contribution 3: Training and inference techniques: grouped reward normalization, intermediate-state matching, and reward-guided ancestral sampler

Description: The authors introduce three key techniques to enhance the distillation process: grouped reward normalization for training stability, intermediate-state matching to prevent mode collapse, and a reward-guided ancestral sampler (RGAS) that improves inference quality through gradient tilting and candidate re-ranking.

This contribution was assessed against **2 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Ultra-fast language generation via discrete diffusion divergence instruct

URL: View paper

Brief Assessment

Ultra-fast Divergence Instruct[4] describes identical techniques including grouped reward normalization, score decomposition (intermediate-state matching), and reward-guided ancestral sampler (RGAS). These appear to be the same contributions.

---

2. A General Framework for Inference-time Scaling and Steering of Diffusion Models

URL: View paper

Brief Assessment

Inference-time Scaling[37] focuses on inference-time steering of diffusion models using particle systems and resampling, not on training-based distillation techniques for discrete diffusion models.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Fast Language Generation through Discrete Diffusion Divergence Instruct View paper
- [1] Swiftbrush: One-step text-to-image diffusion model with variational score distillation View paper
- [2] Distillation of discrete diffusion through dimensional correlations View paper
- [3] Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation View paper
- [4] Ultra-fast language generation via discrete diffusion divergence instruct View paper
- [5] The diffusion duality View paper
- [6] Dkdm: Data-free knowledge distillation for diffusion models with any architecture View paper
- [7] Progressive distillation for fast sampling of diffusion models View paper
- [8] Diffusion LLMs Can Do Faster-Than-AR Inference via Discrete Diffusion Forcing View paper
- [9] Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models View paper
- [10] Scale-wise Distillation of Diffusion Models View paper
- [11] Target Concrete Score Matching: A Holistic Framework for Discrete Diffusion View paper
- [12] UniCMs: A Unified Consistency Model For Efficient Multimodal Generation and Understanding View paper
- [13] SCott: Accelerating Diffusion Models with Stochastic Consistency Distillation View paper
- [14] Beyond Autoregression: Fast LLMs via Self-Distillation Through Time View paper
- [15] Accelerating Diffusion Models with One-to-Many Knowledge Distillation View paper
- [16] Imagine Flash: Accelerating Emu Diffusion Models with Backward Distillation View paper
- [17] Learnable Sampler Distillation for Discrete Diffusion Models View paper
- [18] Accelerating Diffusion Sampling with Classifier-based Feature Distillation View paper
- [19] Collaborative-Distilled Diffusion Models (CDDM) for Accelerated and Lightweight Trajectory Prediction View paper
- [20] Motion synthesis via distilled absorbing discrete diffusion model View paper
- [21] Diffusion Models Acceleration: A Quick Survey View paper
- [22] FS-DFM: Fast and Accurate Long Text Generation with Few-Step Diffusion Language Models View paper
- [23] Few-Step Distillation for Text-to-Image Generation: A Practical Guide View paper
- [24] Score-based Idempotent Distillation of Diffusion Models View paper
- [25] Distribution Matching Distillation Meets Reinforcement Learning View paper
- [26] On Distillation of Guided Diffusion Models View paper
- [27] LoRA-Enhanced Distillation on Guided Diffusion Models View paper
- [28] Distilled Diffusion Language Models View paper
- [29] KDRL: Post-Training Reasoning LLMs via Unified Knowledge Distillation and Reinforcement Learning View paper
- [30] Score and Distribution Matching Policy: Advanced Accelerated Visuomotor Policies via Matched Distillation View paper
- [31] Droid: Learning from offline heterogeneous demonstrations via reward-policy distillation View paper
- [32] Towards searching for the best student in a Knowledge Distillation framework View paper
- [33] ADPO: Anchored Direct Preference Optimization View paper
- [34] On the Design of KL-Regularized Policy Gradient Algorithms for LLM Reasoning View paper
- [35] A Distributional Approach to Controlled Text Generation View paper
- [36] DISTILLATION AND GENERALIZATION IN DEEP REINFORCEMENT LEARNING View paper
- [37] A General Framework for Inference-time Scaling and Steering of Diffusion Models View paper
- [38] Compressed and Smooth Latent Space for Text Diffusion Modeling View paper
- [39] Inference-Time Diffusion Model Distillation View paper