Novelty Assessment Report

Paper: Latent Refinement Decoding: Enhancing Diffusion-Based Language Models by Refining Belief States
PDF URL: https://openreview.net/pdf?id=55oAbWpTcO
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-30

Abstract

Autoregressive (AR) models remain the standard for natural language generation but still suffer from high latency due to strictly sequential decoding. Recent diffusion-inspired approaches, such as LLaDA and Dream, mitigate this by generating in parallel, yet they suffer from two core limitations: information loss, as predictive distributions for non-finalised tokens are discarded at each step, and a lack of well-behaved commitment dynamics, where local decisions are not properly coordinated at the global level. We introduce Latent Refinement Decoding (LRD), a two-stage framework with Latent Refinement and a Predictive Feedback Loop. The first stage maintains masked positions as distributional mixtures of predicted tokens and the mask embedding, allowing the model to establish more globally consistent beliefs. The second stage progressively finalises confident tokens while retaining uncertain ones for iterative feedback. KL-divergence dynamics provide a principled and reliable criterion for convergence and early stopping. Experiments across coding (HumanEval +6.3, MBPP +2.6) and reasoning (GSM8K +2.9, MATH500 +3.8) benchmarks show that LRD improves accuracy while delivering speedups of up to 10.6×. Moreover, LRD is orthogonal to system-level optimisation: when combined with KV-cache and parallel-based accelerators (e.g., Fast-dLLM), it improves accuracy and yields up to 2.4× additional speedup, making it a strong and versatile alternative for parallel sequence generation.

Core Task Landscape

This paper addresses: **Parallel Text Generation Using Diffusion Language Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **18 categories**.

Taxonomy Overview

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

- **Foundational Diffusion Language Model Architectures**
- **Inference Acceleration and Parallel Decoding**
- **Domain-Specific Applications and Adaptations**
- **Controllability and Fine-Grained Generation Control**
- **Theoretical Foundations and Comparative Analysis**

Complete Taxonomy Tree

- Parallel Text Generation Using Diffusion Language Models Survey Taxonomy
- Foundational Diffusion Language Model Architectures
  - Discrete Diffusion Frameworks for Text (4 papers)
  - [11] Discrete diffusion models for language generation (Weligalle, 2025) View paper
  - [17] Diffusion-lm improves controllable text generation (Li, 2022) View paper
  - [21] Tess: Text-to-text self-conditioned simplex diffusion (Mahabadi, 2024) View paper
  - [50] Diffusion-NAT: Self-Prompting Discrete Diffusion for Non-Autoregressive Text Generation (Zhou, 2023) View paper
  - Latent Space Diffusion Methods (3 papers)
  - [8] Non-Autoregressive Text Generation Using Diffusion Models (Krishnamurthy, 2025) View paper
  - [32] Symbolic Music Generation with Diffusion Models (Mittal, 2021) View paper
  - [43] Compressed and Smooth Latent Space for Text Diffusion Modeling (Meshchaninov, 2025) View paper
  - Hybrid Autoregressive-Diffusion Architectures (4 papers)
  - [2] Block diffusion: Interpolating between autoregressive and diffusion language models (Gokaslan, 2025) View paper
  - [18] Ctrldiff: Boosting large diffusion language models with dynamic block prediction and controllable generation (Huang Chi-Han, 2025) View paper
  - [33] Fast-dLLM v2: Efficient Block-Diffusion LLM (WU Chengyue, 2025) View paper
  - [44] ReFusion: A Diffusion Large Language Model with Parallel Autoregressive Decoding (Jia-Nan Li, 2025) View paper
  - Large-Scale Pretrained Diffusion LLMs (3 papers)
  - [4] Dream 7b: Diffusion large language models (Ye, 2025) View paper
  - [6] Seed diffusion: A large-scale diffusion language model with high-speed inference (Song Yuxuan, 2025) View paper
  - [34] Muddit: Liberating Generation Beyond Text-to-Image with a Unified Discrete Diffusion Model (Shi Qingyu, 2025) View paper
- Inference Acceleration and Parallel Decoding
  - Adaptive Parallel Decoding Strategies (5 papers)
  - [12] dparallel: Learnable parallel decoding for dllms (Chen Zigeng, 2025) View paper
  - [19] AdaBlock-dLLM: Semantic-Aware Diffusion LLM Inference via Adaptive Block Size (Guanxi Lu, 2025) View paper
  - [23] Accelerating Diffusion LLMs via Adaptive Parallel Decoding (Israel, 2025) View paper
  - [27] Learning to Parallel: Accelerating Diffusion Large Language Models via Adaptive Parallel Decoding (Bao Wenrui, 2025) View paper

    - Block-Based and Semi-Autoregressive Decoding (2 papers)
    - [15] Diffusion llm with native variable generation lengths: Let lead the way (Y Yang, 2025) View paper
    - [22] Diffusion llms can do faster-than-ar inference via discrete diffusion forcing (Wang Xu, 2025) View paper
    - Conditional Independence and Sampling Optimization (2 papers)
    - [9] Parallel Sampling from Masked Diffusion Models via Conditional Independence Testing (Azangulov, 2025) View paper
    - [14] Parallelbench: Understanding the trade-offs of parallel decoding in diffusion llms (Kang, 2025) View paper
    - Computational Efficiency and KV-Cache Utilization (2 papers)
    - [10] Accelerating diffusion language model inference via efficient kv caching and guided diffusion (Meng Jian, 2025) View paper
    - [40] DPad: Efficient Diffusion Language Models with Suffix Dropout (Chen Xin-hua, 2025) View paper
    - Refinement and Feedback-Based Decoding ★ (3 papers)
    - [0] Latent Refinement Decoding: Enhancing Diffusion-Based Language Models by Refining Belief States (Anon et al., 2026) View paper
    - [7] Free Draft-and-Verification: Toward Lossless Parallel Decoding for Diffusion Large Language Models (Wu, 2025) View paper
    - [31] From Denoising to Refining: A Corrective Framework for Vision-Language Diffusion Model (Ji, 2025) View paper
- Domain-Specific Applications and Adaptations
    - Code Generation with Diffusion Models (2 papers)
    - [5] Beyond autoregression: An empirical study of diffusion large language models for code generation (Zhang Yitong, 2025) View paper
    - [29] Codefusion: A pre-trained diffusion model for code generation (Singh, 2023) View paper
    - Multimodal and Cross-Modal Diffusion (3 papers)
    - [39] DiffCollage: Parallel Generation of Large Content with Diffusion Models (Qin-sheng Zhang, 2023) View paper
    - [42] Emage: Non-Autoregressive Text-to-Image Generation (Feng, 2023) View paper
    - [46] UniD3: unified discrete diffusion for simultaneous vision-language generation (Hu Minghui, 2023) View paper
    - Speech and Audio Applications (2 papers)
    - [25] E1 tts: Simple and fast non-autoregressive tts (Zhijun Liu, 2025) View paper
    - [48] Cross-Modality Diffusion Modeling and Sampling for Speech Recognition (Chia-Kai Yeh, 2024) View paper
    - Specialized Generation Tasks (2 papers)
    - [13] IPAD: Iterative, parallel, and diffusion-based network for scene text recognition (Xiaomeng Yang, 2025) View paper
    - [36] Topology Sculptor, Shape Refiner: Discrete Diffusion Model for High-Fidelity 3D Meshes Generation (Song Kai-yu, 2025) View paper
- Controllability and Fine-Grained Generation Control (2 papers)
    - [16] LAD: LoRA-Adapted Diffusion (Ruurd Jan Anthonius Kuiper, 2025) View paper
    - [30] Contrastive Parallel Denoising for Improving Attribute Alignment of Diffusion models (Xie, 2025) View paper
- Theoretical Foundations and Comparative Analysis
    - Computational Power and Reasoning Analysis (2 papers)
    - [24] On the Reasoning Abilities of Masked Diffusion Language Models (Svete, 2025) View paper
    - [41] On Powerful Ways to Generate: Autoregression, Diffusion, and Beyond (Yang Chenxiao, 2025) View paper
    - Unified Frameworks and Model Relationships (3 papers)
    - [26] Efficient Parallel Samplers for Recurrent-Depth Models and Their Connection to Diffusion Language Models (Geiping, 2025) View paper
    - [37] Unifying Continuous and Discrete Text Diffusion with Non-simultaneous Diffusion Processes (Li Bocheng, 2025) View paper
    - [47] Autoregressive Diffusion Models (Emiel Hoogeboom, 2021) View paper
    - Benchmarking and Evaluation Studies (2 papers)
    - [35] Jailbreaking Large Language Diffusion Models: Revealing Hidden Safety Flaws in Diffusion-Based Text Generation (Zhang Yuan-he, 2025) View paper
    - [38] EmbDiffounder: Semantic-Transfer Enhanced Sequence Diffusion Model for Text Generation (J Huang, 2025) View paper
    - Survey Literature (5 papers)
    - [1] A survey on parallel text generation: From parallel decoding to diffusion language models (Zhang Ling-zhe, 2025) View paper
    - [3] A survey on diffusion language models (LI, 2025) View paper
    - [20] Diffusion-based Large Language Models Survey (Chiung-Yi Tseng, 2025) View paper
    - [45] Diffusion Models in a Nutshell: A Tutorial (Anas M. Ali, 2025) View paper
    - [49] A Survey of Diffusion Models in Natural Language Processing (Zou Hao, 2023) View paper

Narrative

Core task: parallel text generation using diffusion language models. The field has organized itself around several complementary directions. Foundational Diffusion Language Model Architectures establish the basic modeling frameworks—ranging from continuous-space formulations like Diffusion-LM Controllable[17] to discrete variants such as Discrete Diffusion Models[11]—that enable non-autoregressive generation. Inference Acceleration and Parallel Decoding focuses on making these models practical by reducing the number of denoising steps or refining outputs more efficiently, with works like Block Diffusion[2] and Parallel Sampling Masked[9] exploring different strategies for faster sampling. Domain-Specific Applications and Adaptations tailor diffusion approaches to specialized tasks such as code generation (Diffusion Code Generation[5], CodeFusion[29]) or symbolic music (Symbolic Music Diffusion[32]), while Controllability and Fine-Grained Generation Control investigates how to steer outputs toward desired attributes (CtrlDiff[18]). Theoretical Foundations and Comparative Analysis provides surveys and unifying perspectives (Parallel Text Generation Survey[1], Diffusion Language Models Survey[3]) that clarify trade-offs between autoregressive and parallel paradigms.

Within the acceleration branch, a particularly active line of work explores refinement and feedback-based decoding, where models iteratively improve draft outputs rather than generating from scratch. Latent Refinement Decoding[0] exemplifies this approach by operating in a compressed latent space to refine text efficiently, positioning itself alongside methods like Denoising to Refining[31] that reframe the diffusion process as progressive refinement and Free Draft Verification[7] that leverages verification signals to guide iterative improvement. These refinement-oriented techniques contrast with one-shot parallel samplers (Parallel Sampling Masked[9]) and adaptive strategies (Adaptive Parallel Decoding[23]) that dynamically adjust decoding depth. The central tension across these directions is balancing generation quality, controllability, and computational cost: while some works prioritize speed through aggressive parallelism, others like Latent Refinement Decoding[0] emphasize maintaining fidelity by carefully refining intermediate representations, reflecting broader questions about how best to exploit the flexibility of diffusion models for practical text generation.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Free Draft-and-Verification: Toward Lossless Parallel Decoding for Diffusion Large Language Models

Authors: Wu, Shutong, Zhang, Jiawei, Shutong Wu, et al. (6 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Diffusion Large Language Models (DLLMs) have emerged as a new paradigm of language modeling beyond autoregressive next-token prediction. Thanks to their bidirectional attention mechanism, DLLMs are more capable of capturing the connection of context, and thus show unique advantages in challenges like the famous"reversal curse"or learning under data-constrained scenarios. In addition, taking advantage of their inherent modeling foundations, DLLMs have the great potential of efficient inference wi...

Relationship Analysis

Both papers belong to the Refinement and Feedback-Based Decoding category, addressing iterative improvement in parallel text generation using diffusion language models. They overlap in their focus on improving generation quality through feedback mechanisms during parallel decoding, with both maintaining belief states across iterations. The key difference is that the original paper (Latent Refinement Decoding) introduces a two-phase framework with soft embedding mixtures and entropy-based early stopping for adaptive refinement, while the candidate paper (Free Draft-and-Verification) proposes a draft-and-verification approach that generates multiple candidate sequences in parallel and verifies them through the DLLM itself, focusing on lossless parallel decoding without model modification.

---

2. From Denoising to Refining: A Corrective Framework for Vision-Language Diffusion Model

Authors: Ji, Yatai, Wang Teng, Yatai Ji, Ge, et al. (18 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Discrete diffusion models have emerged as a promising direction for vision-language tasks, offering bidirectional context modeling and theoretical parallelization. However, their practical application is severely hindered by a train-inference discrepancy, which leads to catastrophic error cascades: initial token errors during parallel decoding pollute the generation context, triggering a chain reaction of compounding errors and leading to syntactic errors and semantic hallucinations. To address ...

Relationship Analysis

Both papers belong to the Refinement and Feedback-Based Decoding category, focusing on iterative refinement processes to improve generation quality in diffusion-based language models. While the original paper (Latent Refinement Decoding) introduces a two-phase framework with soft embedding mixtures and KL-divergence-based monitoring for parallel text generation, the candidate paper (ReDiff) addresses vision-language diffusion models by training the model to actively refine and correct its own errors through a two-stage process involving synthetic error injection and online self-correction learning. The key difference is that the original paper operates on pure text generation with continuous embedding refinement, whereas the candidate focuses on vision-language tasks with an emphasis on correcting hallucinations and syntactic errors through expert-guided revision.

Contributions Analysis

Overall novelty summary. The paper proposes Latent Refinement Decoding (LRD), a two-stage framework combining distributional mixture representations with iterative feedback loops for parallel text generation. It resides in the 'Refinement and Feedback-Based Decoding' leaf, which contains only three papers total, indicating a relatively sparse research direction within the broader inference acceleration landscape. This leaf sits alongside more populated areas like 'Adaptive Parallel Decoding Strategies' (five papers) and 'Block-Based and Semi-Autoregressive Decoding' (two papers), suggesting refinement-based approaches represent an emerging rather than saturated research thread.

The taxonomy reveals that LRD's neighbors include adaptive strategies that dynamically select tokens for parallel decoding and block-based methods that partition generation into sequential chunks. The refinement leaf explicitly excludes single-pass parallel methods and training-based improvements, positioning LRD within iterative quality-enhancement approaches rather than one-shot generation or architectural innovations. Nearby leaves like 'Conditional Independence and Sampling Optimization' (two papers) and 'Computational Efficiency and KV-Cache Utilization' (two papers) address complementary concerns—identifying independent token sets and reducing memory overhead—that LRD does not directly target, clarifying its distinct focus on belief-state maintenance and progressive commitment.

Among twenty candidates examined across three contributions, none were flagged as clearly refuting LRD's novelty. The 'Latent Refinement Decoding framework' contribution examined ten candidates with zero refutations, as did the 'Adaptive two-phase sampling with KL-based monitoring' contribution. The 'Soft diffusion mechanism' examined zero candidates, likely due to limited semantic overlap in the search. This suggests that within the examined scope—drawn from top-K semantic matches and citation expansion—LRD's specific combination of distributional mixtures, predictive feedback, and KL-divergence-based convergence criteria does not have direct precedents, though the limited search scale (twenty papers from a fifty-paper taxonomy) means unexplored prior work may exist.

The analysis covers a focused subset of the field, emphasizing refinement-oriented methods and their immediate neighbors in the taxonomy. The sparse population of the refinement leaf and absence of refutations among examined candidates suggest LRD introduces mechanisms not prominently represented in the surveyed literature. However, the twenty-candidate scope leaves open the possibility of relevant work in adjacent areas—such as latent-space diffusion methods or hybrid architectures—that were not surfaced by semantic search, warranting caution in generalizing these findings beyond the examined sample.

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

Contribution 1: Latent Refinement Decoding (LRD) framework

Description: A two-stage decoding framework for diffusion language models that first refines global beliefs in continuous embedding space through distributional mixtures of predicted tokens and mask embeddings, then progressively finalizes confident tokens while retaining uncertain ones for iterative feedback, using KL-divergence dynamics for convergence monitoring and early stopping.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. From Skeleton to Flesh: Aggregated Relational Transformer Towards Controllable Video Captioning with Two-Step Decoding

URL: View paper

Brief Assessment

Skeleton to Flesh[58] focuses on video captioning with relation-based two-step decoding (skeleton relations → full captions), not diffusion language models with continuous embedding refinement and KL-based convergence monitoring.

---

2. Beyond fixed: Training-free variable-length denoising for diffusion large language models

URL: View paper

Brief Assessment

Training-free Variable-length[54] addresses the static generation length constraint in diffusion LMs through dynamic length expansion, whereas the original paper's LRD focuses on a two-stage refinement process with soft embeddings and KL-based convergence monitoring for improved decoding quality.

---

3. From Denoising to Refining: A Corrective Framework for Vision-Language Diffusion Model

URL: View paper

Brief Assessment

Denoising to Refining[31] focuses on vision-language diffusion models with a two-stage training framework (foundational revision training and online self-correction learning) to address error cascades in parallel generation. The ORIGINAL paper's LRD operates on pure language models with continuous embedding mixtures and KL-divergence monitoring for convergence, representing a fundamentally different technical approach and application domain.

---

4. Text-guided molecule generation with diffusion language model

URL: View paper

Brief Assessment

Text-guided Molecule Generation[51] focuses on molecule generation from text descriptions using diffusion models for SMILES strings, not general language model decoding. The two-phase approach in Text-guided Molecule Generation[51] (text-guided generation followed by correction of invalid SMILES) serves a domain-specific purpose distinct from LRD's general framework for refining beliefs in continuous embedding space for language generation.

---

5. Riv: Recursive introspection mask diffusion vision language model

URL: View paper

Brief Assessment

Riv[53] focuses on vision-language models with introspection-based self-correction for multimodal tasks, not on two-stage decoding frameworks for diffusion language models with iterative refinement in continuous embedding space.

---

6. Think while you generate: Discrete diffusion with planned denoising

URL: View paper

Brief Assessment

Think While Generate[55] focuses on decomposing discrete diffusion into planning (selecting which positions to denoise) and denoising (predicting token values), using a planner model to identify corrupted positions. In contrast, the original paper's LRD operates in continuous embedding space through distributional mixtures without explicit planning models, representing a fundamentally different architectural approach to diffusion-based generation.

---

7. DiffuGR: Generative Document Retrieval with Diffusion Language Models

URL: View paper

Brief Assessment

DiffuGR[56] focuses on document retrieval using diffusion models to generate document identifiers (docids), not on general-purpose language generation with iterative refinement in continuous embedding space. The candidate addresses a different application domain (information retrieval) rather than the general text generation framework proposed in the original paper.

---

8. Diffusion-based Large Language Models Survey

URL: View paper

Brief Assessment

Diffusion LLMs Survey[20] provides only a brief mention of latent diffusion for language generation without describing any two-stage decoding framework or the specific mechanisms of LRD (distributional mixtures, progressive token finalization, KL-divergence monitoring).

---

9. Encoder-Decoder Diffusion Language Models for Efficient Training and Inference

URL: View paper

Brief Assessment

Encoder-Decoder Diffusion[57] focuses on architectural separation of clean token representation (encoder) and noisy token denoising (decoder) for efficiency, not on two-stage iterative refinement with distributional mixtures and KL-divergence monitoring as in the original paper's LRD framework.

---

10. Beyond tokens: A survey on decoding methods for large language models and large vision-language models

URL: View paper

Brief Assessment

Decoding Methods Survey[52] discusses various decoding approaches including iterative refinement methods, but focuses on surveying existing techniques rather than proposing novel frameworks. The survey mentions optimization-based and two-stage methods in passing, but does not present a specific framework combining latent refinement with distributional mixtures and KL-divergence monitoring as described in the original paper's LRD contribution.

Contribution 2: Soft diffusion mechanism for continuous denoising

Description: A mechanism that maintains masked positions as distributional mixtures rather than hard assignments, preserving distributional information across denoising steps and enabling cross-position refinement through self-attention in the embedding space.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Contribution 3: Adaptive two-phase sampling with KL-based monitoring

Description: A sampling strategy that automatically transitions from soft embedding refinement to hard token commitment based on KL-divergence convergence criteria, enabling adaptive early stopping that adjusts generation length based on problem complexity rather than using fixed iteration counts.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Adaptive and Efficient Continual Learning in Dynamic Environments

URL: View paper

Brief Assessment

Adaptive Continual Learning[64] focuses on continual training of diffusion models in dynamic environments, not adaptive sampling strategies for text generation. The candidate's KL divergence usage appears in a different context (continual learning convergence) rather than monitoring denoising convergence for early stopping in language generation tasks.

2. KLASS: KL-Guided Fast Inference in Masked Diffusion Models

URL: View paper

Brief Assessment

KLASS[63] uses KL-divergence to identify stable predictions for multi-token unmasking within a single-phase sampling approach, whereas the original paper proposes a distinct two-phase framework that transitions from soft embedding refinement to hard token commitment based on KL convergence criteria.

3. Entropy-Adaptive Diffusion Policy Optimization with Dynamic Step Alignment

URL: View paper

Brief Assessment

Entropy-Adaptive Diffusion Policy[65] focuses on RL fine-tuning of diffusion models for image generation with entropy-based exploration, not on adaptive sampling strategies for language model decoding with KL-divergence convergence monitoring as in the original paper.

4. A network traffic data generation model based on AOT-DDPM for abnormal traffic detection

URL: View paper

Brief Assessment

AOT-DDPM Traffic[61] focuses on network traffic data generation for anomaly detection, not language model decoding. The candidate's adaptive sampling and KL-divergence monitoring are applied in a completely different domain (network security) with different technical objectives than the original paper's text generation framework.

5. How can Diffusion Models Evolve into Continual Generators?

URL: View paper

Brief Assessment

Continual Generators[66] focuses on continual learning for diffusion models across sequential tasks, not adaptive sampling strategies within a single generation process. The KL-divergence monitoring in the candidate serves a different purpose (detecting task convergence in continual learning) rather than controlling phase transitions during individual sample generation.

6. A Comprehensive Survey on Continual Learning in Generative Models

URL: View paper

Brief Assessment

Continual Learning Generative[59] is a survey paper on continual learning methods for generative models (LLMs, MLLMs, VLAs, diffusion models), not a research paper proposing adaptive sampling strategies with KL-divergence monitoring for diffusion model convergence.

7. Diffusion-based Large Language Models Survey

URL: View paper

Brief Assessment

The candidate paper does not discuss adaptive sampling strategies, KL-divergence convergence criteria, or early stopping mechanisms for diffusion models. The survey's scope does not cover these technical contributions.

8. KL-Divergence Guided Temperature Sampling

URL: View paper

Brief Assessment

KL-Divergence Temperature Sampling[67] uses KL-divergence to guide temperature adjustments for attribution vs. diversity trade-offs in grounded generation, not for adaptive early stopping in diffusion-based iterative refinement as in the original paper.

9. Tuning Sequential Monte Carlo Samplers via Greedy Incremental Divergence Minimization

URL: View paper

Brief Assessment

Tuning Sequential Monte[60] focuses on tuning SMC samplers for static Bayesian inference by minimizing incremental KL divergence between proposal and target paths. The original paper addresses diffusion language model decoding with soft embedding refinement and hard token commitment phases. These are fundamentally different application domains and technical approaches.

10. Adding Conditional Control to Diffusion Models with Reinforcement Learning

URL: View paper

Brief Assessment

Conditional Control Reinforcement[62] uses KL divergence as a reward function component in RL-based fine-tuning of pre-trained diffusion models, not as a convergence monitoring criterion for adaptive phase transitions in sampling strategies.

Appendix: Text Similarity Detection

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

References

- [0] Latent Refinement Decoding: Enhancing Diffusion-Based Language Models by Refining Belief States View paper
- [1] A survey on parallel text generation: From parallel decoding to diffusion language models View paper
- [2] Block diffusion: Interpolating between autoregressive and diffusion language models View paper

- [3] A survey on diffusion language models View paper
- [4] Dream 7b: Diffusion large language models View paper
- [5] Beyond autoregression: An empirical study of diffusion large language models for code generation View paper
- [6] Seed diffusion: A large-scale diffusion language model with high-speed inference View paper
- [7] Free Draft-and-Verification: Toward Lossless Parallel Decoding for Diffusion Large Language Models View paper
- [8] Non-Autoregressive Text Generation Using Diffusion Models View paper
- [9] Parallel Sampling from Masked Diffusion Models via Conditional Independence Testing View paper
- [10] Accelerating diffusion language model inference via efficient kv caching and guided diffusion View paper
- [11] Discrete diffusion models for language generation View paper
- [12] dparallel: Learnable parallel decoding for dllms View paper
- [13] IPAD: Iterative, parallel, and diffusion-based network for scene text recognition View paper
- [14] Parallelbench: Understanding the trade-offs of parallel decoding in diffusion llms View paper
- [15] Diffusion llm with native variable generation lengths: Let lead the way View paper
- [16] LAD: LoRA-Adapted Diffusion View paper
- [17] Diffusion-lm improves controllable text generation View paper
- [18] Ctrldiff: Boosting large diffusion language models with dynamic block prediction and controllable generation View paper
- [19] AdaBlock-dLLM: Semantic-Aware Diffusion LLM Inference via Adaptive Block Size View paper
- [20] Diffusion-based Large Language Models Survey View paper
- [21] Tess: Text-to-text self-conditioned simplex diffusion View paper
- [22] Diffusion llms can do faster-than-ar inference via discrete diffusion forcing View paper
- [23] Accelerating Diffusion LLMs via Adaptive Parallel Decoding View paper
- [24] On the Reasoning Abilities of Masked Diffusion Language Models View paper
- [25] E1 tts: Simple and fast non-autoregressive tts View paper
- [26] Efficient Parallel Samplers for Recurrent-Depth Models and Their Connection to Diffusion Language Models View paper
- [27] Learning to Parallel: Accelerating Diffusion Large Language Models via Adaptive Parallel Decoding View paper
- [28] Learning to Parallel: Accelerating Diffusion Large Language Models via Learnable Parallel Decoding View paper
- [29] Codefusion: A pre-trained diffusion model for code generation View paper
- [30] Contrastive Parallel Denoising for Improving Attribute Alignment of Diffusion models View paper
- [31] From Denoising to Refining: A Corrective Framework for Vision-Language Diffusion Model View paper
- [32] Symbolic Music Generation with Diffusion Models View paper
- [33] Fast-dLLM v2: Efficient Block-Diffusion LLM View paper
- [34] Muddit: Liberating Generation Beyond Text-to-Image with a Unified Discrete Diffusion Model View paper
- [35] Jailbreaking Large Language Diffusion Models: Revealing Hidden Safety Flaws in Diffusion-Based Text Generation View paper
- [36] Topology Sculptor, Shape Refiner: Discrete Diffusion Model for High-Fidelity 3D Meshes Generation View paper
- [37] Unifying Continuous and Discrete Text Diffusion with Non-simultaneous Diffusion Processes View paper
- [38] EmbDiffounder: Semantic-Transfer Enhanced Sequence Diffusion Model for Text Generation View paper
- [39] DiffCollage: Parallel Generation of Large Content with Diffusion Models View paper
- [40] DPad: Efficient Diffusion Language Models with Suffix Dropout View paper
- [41] On Powerful Ways to Generate: Autoregression, Diffusion, and Beyond View paper
- [42] Emage: Non-Autoregressive Text-to-Image Generation View paper
- [43] Compressed and Smooth Latent Space for Text Diffusion Modeling View paper
- [44] ReFusion: A Diffusion Large Language Model with Parallel Autoregressive Decoding View paper
- [45] Diffusion Models in a Nutshell: A Tutorial View paper
- [46] UniD3: unified discrete diffusion for simultaneous vision-language generation View paper
- [47] Autoregressive Diffusion Models View paper
- [48] Cross-Modality Diffusion Modeling and Sampling for Speech Recognition View paper
- [49] A Survey of Diffusion Models in Natural Language Processing View paper
- [50] Diffusion-NAT: Self-Prompting Discrete Diffusion for Non-Autoregressive Text Generation View paper
- [51] Text-guided molecule generation with diffusion language model View paper
- [52] Beyond tokens: A survey on decoding methods for large language models and large vision-language models View paper
- [53] Riv: Recursive introspection mask diffusion vision language model View paper
- [54] Beyond fixed: Training-free variable-length denoising for diffusion large language models View paper
- [55] Think while you generate: Discrete diffusion with planned denoising View paper
- [56] DiffuGR: Generative Document Retrieval with Diffusion Language Models View paper
- [57] Encoder-Decoder Diffusion Language Models for Efficient Training and Inference View paper
- [58] From Skeleton to Flesh: Aggregated Relational Transformer Towards Controllable Video Captioning with Two-Step Decoding View paper
- [59] A Comprehensive Survey on Continual Learning in Generative Models View paper
- [60] Tuning Sequential Monte Carlo Samplers via Greedy Incremental Divergence Minimization View paper
- [61] A network traffic data generation model based on AOT-DDPM for abnormal traffic detection View paper
- [62] Adding Conditional Control to Diffusion Models with Reinforcement Learning View paper
- [63] KLASS: KL-Guided Fast Inference in Masked Diffusion Models View paper
- [64] Adaptive and Efficient Continual Learning in Dynamic Environments View paper
- [65] Entropy-Adaptive Diffusion Policy Optimization with Dynamic Step Alignment View paper
- [66] How can Diffusion Models Evolve into Continual Generators? View paper
- [67] KL-Divergence Guided Temperature Sampling View paper