Novelty Assessment Report

Paper: Fast-dLLM: Training-free Acceleration of Diffusion LLM by Enabling KV Cache and Parallel Decoding
PDF URL: https://openreview.net/pdf?id=3Z3Is6hnOT
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-30

Abstract

Diffusion-based large language models (Diffusion LLMs) have shown promise for non-autoregressive text generation. However, the practical inference speed of open-sourced Diffusion LLMs often lags behind autoregressive models due to the lack of Key-Value (KV) Cache and quality degradation when decoding multiple tokens simultaneously. To bridge this gap, we introduce Fast-dLLM, a method that incorporates a novel block-wise approximate KV Cache mechanism tailored for bidirectional diffusion models, enabling cache reuse with negligible performance drop. Additionally, we identify the root cause of generation quality degradation in parallel decoding as the disruption of token dependencies under the conditional independence assumption. To address this, Fast-dLLM also proposes a confidence-aware parallel decoding strategy that selectively decodes tokens exceeding a confidence threshold, mitigating dependency violations and maintaining generation quality. Experimental results on LLaDA and Dream models across multiple LLM benchmarks demonstrate up to 27.6× throughput improvement with minimal accuracy loss, closing the performance gap with autoregressive models and paving the way for practical deployment of Diffusion LLMs.

Core Task Landscape

This paper addresses: **Accelerating Inference of Diffusion-Based Large Language Models**
A total of **50 papers** were analyzed and organized into a taxonomy with **21 categories**.

Taxonomy Overview

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

- **Decoding Strategy Optimization**
- **Cache-Based Acceleration**
- **Architectural Acceleration**
- **Speculative and Hybrid Decoding**
- **Training and Objective Optimization**
- **Model Compression and Quantization**
- **Diffusion Model Foundations and Theory**
- **Multimodal Diffusion Models**
- **Application-Specific Diffusion Models**
- **Benchmarking and Comparative Analysis**
- ... and 1 more categories

Complete Taxonomy Tree

- Accelerating Inference of Diffusion-Based Large Language Models Survey Taxonomy
- Decoding Strategy Optimization
  - Adaptive Parallel Decoding (3 papers)
  - [8] Creditdecoding: Accelerating parallel decoding in diffusion large language models with trace credits (Wang Kangyu, 2025) View paper
  - [16] Accelerating Diffusion LLMs via Adaptive Parallel Decoding (Israel, 2025) View paper
  - [25] Accelerating Diffusion LLM Inference via Local Determinism Propagation (Zhang Jingyuan, 2025) View paper
  - Confidence-Based Token Selection ★ (3 papers)
  - [0] Fast-dLLM: Training-free Acceleration of Diffusion LLM by Enabling KV Cache and Parallel Decoding (Anon et al., 2026) View paper
  - [12] Self Speculative Decoding for Diffusion Large Language Models (Gao Yifeng, 2025) View paper
  - [19] Saber: An Efficient Sampling with Adaptive Acceleration and Backtracking Enhanced Remasking for Diffusion Language Model (Dong Yihong, 2025) View paper
  - Planning and Trajectory Optimization (3 papers)
  - [7] Revolutionizing reinforcement learning framework for diffusion large language models (Wang Yinjie, 2025) View paper
  - [11] Plan for Speed--Dilated Scheduling for Masked Diffusion Language Models (Permuter, 2025) View paper
  - [39] Planner Aware Path Learning in Diffusion Language Models Training (Bezemek, 2025) View paper
  - Early Stopping and Convergence Detection (2 papers)
  - [27] Accelerating Diffusion Large Language Models with SlowFast: The Three Golden Principles (Q Wei, 2025) View paper
  - [42] Diffusion Language Models Know the Answer Before Decoding (Li Pengxiang, 2025) View paper
- Cache-Based Acceleration
  - Adaptive KV Cache for Diffusion Models (2 papers)
  - [10] dLLM-Cache: Accelerating Diffusion Large Language Models with Adaptive Caching (Liu Zhi-Yuan, 2025) View paper
  - [26] dCache: Accelerating Diffusion-Based LLMs via Dual Adaptive Caching (Y Jiang, 2025) View paper

- Block-Wise KV Cache (2 papers)
  - [15] Accelerating diffusion language model inference via efficient kv caching and guided diffusion (Meng Jian, 2025) View paper
  - [36] AdaBlock-dLLM: Semantic-Aware Diffusion LLM Inference via Adaptive Block Size (Guanxi Lu, 2025) View paper
- Architectural Acceleration
  - Sparse Attention for Diffusion Models (2 papers)
  - [4] SparseD: Sparse Attention for Diffusion Language Models (Wang ZeQing, 2025) View paper
  - [43] Sparse-LaViDa: Sparse Multimodal Discrete Diffusion Language Models (Shufan Li, 2025) View paper
  - Encoder-Decoder Architectures (1 papers)
  - [33] Encoder-Decoder Diffusion Language Models for Efficient Training and Inference (Schiff, 2025) View paper
  - Token-Level Pruning (1 papers)
  - [21] Few-shot temporal pruning accelerates diffusion models for text generation (B Li, 2024) View paper
- Speculative and Hybrid Decoding
  - Diffusion-Based Speculative Decoding (2 papers)
  - [14] Diffuspec: Unlocking diffusion language models for speculative decoding (Li, 2025) View paper
  - [17] Speculative diffusion decoding: Accelerating language generation through diffusion (Bartoldson, 2025) View paper
  - Block-Autoregressive Hybrid (4 papers)
  - [2] Sequential diffusion language models (Liu Yangzhou, 2025) View paper
  - [31] Ctrldiff: Boosting large diffusion language models with dynamic block prediction and controllable generation (Huang Chi-Han, 2025) View paper
  - [32] Fast-dllm v2: Efficient block-diffusion llm (WU Chengyue, 2025) View paper
  - [46] Diffusion llms can do faster-than-ar inference via discrete diffusion forcing (Wang Xu, 2025) View paper
- Training and Objective Optimization
  - Simplified Training Objectives (3 papers)
  - [1] Simple and effective masked diffusion language models (Sahoo, 2024) View paper
  - [13] Efficient Perplexity Bound and Ratio Matching in Discrete Diffusion Language Models (Haxholli, 2025) View paper
  - [24] Your Absorbing Discrete Diffusion Secretly Models the Conditional Distributions of Clean Data (Nie Shen, 2024) View paper
  - Reinforcement Learning for Diffusion (1 papers)
  - [40] Taming Masked Diffusion Language Models via Consistency Trajectory Reinforcement Learning with Fewer Decoding Step (Yang, 2025) View paper
  - Distillation-Based Acceleration (2 papers)
  - [38] FS-DFM: Fast and Accurate Long Text Generation with Few-Step Diffusion Language Models (Monsefi, 2025) View paper
  - [49] DLM-One: Diffusion Language Models for One-Step Sequence Generation (Chen, 2025) View paper
  - Context-Aware Initialization (1 papers)
  - [45] Context-Aware Initialization for Reducing Generative Path Length in Diffusion Language Models (Tongyuan Miao, 2025) View paper
- Model Compression and Quantization (2 papers)
  - [23] DLLMQuant: Quantizing Diffusion-based Large Language Models (Xu Chen, 2025) View paper
  - [47] Quantization meets dllms: A systematic study of post-training quantization for diffusion llms (Lin Haokun, 2025) View paper
- Diffusion Model Foundations and Theory (2 papers)
  - [22] A Convergence Theory for Diffusion Language Models: An Information-Theoretic Perspective (Li Gen, 2025) View paper
  - [30] Energy-based diffusion language models for text generation (Xu, 2024) View paper
- Multimodal Diffusion Models (2 papers)
  - [18] Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding (Xin Yi, 2025) View paper
  - [41] Lavida: A large diffusion language model for multimodal understanding (Li Shu-Fan, 2025) View paper
- Application-Specific Diffusion Models (2 papers)
  - [44] Discrete diffusion language model for efficient text summarization (Do Huu Dat, 2025) View paper
  - [50] A2D: Any-Order, Any-Step Safety Alignment for Diffusion Language Models (Wonje Jeung, 2025) View paper
- Benchmarking and Comparative Analysis (4 papers)
  - [3] Dream 7b: Diffusion large language models (Ye, 2025) View paper
  - [34] Beyond Next-Token Prediction: A Performance Characterization of Diffusion versus Autoregressive Language Models (Kim Minseo, 2025) View paper
  - [35] How Efficient Are Diffusion Language Models? A Critical Examination of Efficiency Evaluation Practices (Peng Han, 2025) View paper
  - [48] Set Block Decoding is a Language Model Inference Accelerator (Gat, 2025) View paper
- Survey and Framework Papers (7 papers)
  - [5] A survey on diffusion language models (LI, 2025) View paper
  - [6] A survey on parallel text generation: From parallel decoding to diffusion language models (Zhang Ling-zhe, 2025) View paper
  - [9] dinfer: An efficient inference framework for diffusion language models (Ma Yuxin, 2025) View paper
  - [20] Diffusion models in text generation: a survey (Qiuhua Yi, 2024) View paper
  - [28] Diffusion-based Large Language Models Survey (Chiung-Yi Tseng, 2025) View paper
  - [29] Discrete diffusion models for language generation (Weligalle, 2025) View paper
  - [37] Discrete Diffusion in Large Language and Multimodal Models: A Survey (Yu, 2025) View paper

Narrative

Core task: Accelerating inference of diffusion-based large language models. The field has organized itself around several complementary acceleration strategies. Decoding Strategy Optimization explores how to intelligently select or schedule tokens during the iterative diffusion process, including confidence-based selection and adaptive scheduling approaches. Cache-Based Acceleration (e.g., dLLM-Cache[10], dCache[26]) focuses on reusing intermediate computations across diffusion steps to reduce redundant calculations. Architectural Acceleration and Model Compression branches address efficiency through structural modifications and quantization techniques (DLLMQuant[23]), while Speculative and Hybrid Decoding methods (Diffuspec[14], Self Speculative Decoding[12]) attempt to predict multiple tokens or steps ahead to reduce sequential dependencies. Training and Objective Optimization investigates how learning procedures can be redesigned for faster convergence, and foundational branches cover theoretical underpinnings (Convergence

Theory[22], Discrete Diffusion Models[29]) alongside multimodal extensions and application-specific adaptations. Survey works (Diffusion Language Survey[5], Diffusion LLM Survey[28]) provide broader perspectives on these evolving directions.

Particularly active lines of work center on reducing the number of diffusion steps required and exploiting token-level confidence to skip unnecessary computations. Fast-dLLM[0] sits within the Confidence-Based Token Selection cluster, emphasizing early stopping or selective refinement of high-confidence tokens to avoid redundant denoising iterations. This approach contrasts with neighboring methods like Saber[19], which may prioritize different scheduling heuristics, and Self Speculative Decoding[12], which leverages draft-and-verify mechanisms rather than confidence thresholds. The trade-off across these branches often involves balancing generation quality against wall-clock speedup: confidence-based strategies can yield substantial gains when token predictions stabilize quickly, but may require careful tuning to avoid premature convergence. Fast-dLLM[0] exemplifies this balance by targeting scenarios where diffusion models exhibit predictable confidence dynamics, positioning it as a practical complement to cache-based and speculative techniques that address orthogonal bottlenecks in the inference pipeline.

Related Works in Same Category

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

1. Self Speculative Decoding for Diffusion Large Language Models

Authors: Gao Yifeng, Yifeng Gao, Wang Yu-xuan, Ziang Ji, Qi, et al. (13 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

Diffusion-based Large Language Models (dLLMs) have emerged as a competitive alternative to autoregressive models, offering unique advantages through bidirectional attention and parallel generation paradigms. However, the generation results of current parallel decoding methods deviate from stepwise decoding, introducing potential performance degradation, which limits their practical deployment. To address this problem, we propose \textbf{S}elf \textbf{S}peculative \textbf{D}ecoding (SSD), a lossl...

Relationship Analysis

Both papers belong to the Confidence-Based Token Selection category, using model confidence to guide token decoding in diffusion LLMs. They overlap in addressing the quality degradation problem of parallel decoding through confidence-aware strategies—Fast-dLLM uses a confidence threshold to selectively decode tokens exceeding a global threshold, while SSD employs self-speculative decoding with hierarchical verification trees. The key difference is that Fast-dLLM combines confidence-based parallel decoding with KV caching mechanisms for acceleration, whereas SSD focuses on a self-drafting and verification approach without auxiliary modules, achieving lossless acceleration by using the model itself as both drafter and verifier.

---

2. Saber: An Efficient Sampling with Adaptive Acceleration and Backtracking Enhanced Remasking for Diffusion Language Model

Authors: Dong Yihong, Ma Zhaoyu, Yihong Dong, Jiang Xue, Zhaoyu Ma, et al. (26 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

Diffusion language models (DLMs) are emerging as a powerful and promising alternative to the dominant autoregressive paradigm, offering inherent advantages in parallel generation and bidirectional context modeling. However, the performance of DLMs on code generation tasks, which have stronger structural constraints, is significantly hampered by the critical trade-off between inference speed and output quality. We observed that accelerating the code generation process by reducing the number of sa...

Relationship Analysis

Both papers belong to the Confidence-Based Token Selection category, using model confidence scores to guide token decoding in diffusion LLMs. While Fast-dLLM uses a confidence threshold to select which tokens to unmask in parallel across general text generation tasks, Saber focuses specifically on code generation and introduces an adaptive acceleration mechanism combined with backtracking-enhanced remasking. The key difference is that Saber addresses code generation's structural constraints through adaptive sampling and error correction via backtracking, whereas Fast-dLLM emphasizes general parallel decoding with KV caching for broader text generation scenarios.

Contributions Analysis

Overall novelty summary. The paper proposes Fast-dLLM, which combines a block-wise approximate KV cache mechanism with a confidence-aware parallel decoding strategy to accelerate diffusion-based large language models. According to the taxonomy, it resides in the 'Confidence-Based Token Selection' leaf under 'Decoding Strategy Optimization', alongside two sibling papers. This leaf represents a moderately populated research direction within a broader taxonomy of 50 papers across approximately 36 topics, suggesting that confidence-based approaches are an established but not overcrowded area of investigation in diffusion LLM acceleration.

The taxonomy reveals that Fast-dLLM sits at the intersection of two major acceleration paradigms: 'Decoding Strategy Optimization' (which includes adaptive parallel decoding and planning-based methods) and 'Cache-Based Acceleration' (covering adaptive and block-wise KV cache techniques). Neighboring leaves include 'Adaptive Parallel Decoding' and 'Block-Wise KV Cache', indicating that the paper bridges token selection strategies with caching mechanisms. The taxonomy's scope notes clarify that confidence-based methods focus on model confidence scores for token unmasking, distinguishing them from planning-based trajectory optimization or purely architectural modifications.

Among 28 candidates examined through limited semantic search, the analysis identified 11 refutable pairs across three contributions. The block-wise KV cache mechanism examined 10 candidates with 7 appearing to provide overlapping prior work, suggesting substantial existing research on caching for diffusion models. The confidence-aware parallel decoding strategy examined 9 candidates with only 2 refutable matches, indicating potentially greater novelty in this specific combination. The overall Fast-dLLM framework also examined 9 candidates with 2 refutable pairs, though the limited search scope means these statistics reflect top-K semantic matches rather than exhaustive coverage.

Based on the limited literature search of 28 candidates, the work appears to synthesize existing acceleration paradigms—caching and confidence-based decoding—in a novel combination tailored for diffusion LLMs. The higher refutation rate for the caching component suggests this aspect builds more directly on established techniques, while the confidence-aware strategy may represent a less explored integration. The analysis does not cover the full breadth of diffusion LLM research, and a more comprehensive search might reveal additional overlapping work in either component.

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

Contribution 1: Block-wise approximate KV Cache mechanism for bidirectional diffusion models

Description: The authors propose a novel KV caching strategy tailored for masked diffusion language models that use full bidirectional attention. By adopting block-wise generation and caching prefix (and optionally suffix) tokens, the method enables substantial computational reuse across decoding steps with negligible performance degradation.

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. dCache: Accelerating Diffusion-Based LLMs via Dual Adaptive Caching

URL: View paper

Prior Art Analysis

dCache[26] demonstrates that similar block-wise KV caching strategies for bidirectional diffusion models were proposed prior to the original paper. Both papers address the fundamental challenge that bidirectional attention in diffusion LLMs prevents standard KV cache reuse, and both propose block-wise generation with approximate KV caching as the solution. dCache[26] explicitly describes partitioning sequences into blocks, caching KV states of non-current blocks, and reusing them across decoding steps—the same core mechanism claimed as novel in the original paper.

Evidence

Evidence 1 - **Rationale**: Both papers introduce approximate KV cache frameworks specifically designed for diffusion-based LLMs with bidirectional attention, addressing the same fundamental problem. - **Original**: we introduce fast-dllm, a method that incorporates a novel block-wise approximate kv cache mechanism tailored for bidirectional diffusion models, enabling cache reuse with negligible performance drop. - **Candidate**: we introducedual adaptive cache(d 2cache), which is a training-free approximate kv cache framework for accelerating dllm inference.

Evidence 2 - **Rationale**: Both papers identify the identical core challenge: bidirectional attention in diffusion LLMs prevents standard KV cache reuse because updating any token requires recomputing all KV states. - **Original**: first, diffusion llms do not support key-value (kv) caching, a critical component in ar models for speeding up inference. second, the generation quality tends to degrade when decoding multiple tokens in parallel. - **Candidate**: due to bidirectional attention, dllms cannot benefit from the standard key-value (kv) cache as arms do. as shown in figure 1 (a), arms leverage causal attention to sequentially generate new tokens and append each new token to the end of the sequence. this autoregressive process naturally enables the...

Evidence 3 - **Rationale**: dCache[26] explicitly references prior work (including wu et al., 2025, which appears to be the original paper based on timing) that explored approximate KV cache for DLLMs based on KV state similarity across steps—the same foundational observation. - **Original**: as shown in figure 2, we adopt a block-wise decoding strategy to support the use of a key-value (kv) cache. initially, we compute and store the kv cache for the prompt, which is reused throughout block 0. within each block, the same cache is reused for multiple decoding steps. after completing the d... - **Candidate**: to address the above efficiency challenges, recent studies (ma et al., 2025; wu et al., 2025; liu et al., 2025; hu et al., 2025) have explored approximate kv cache to accelerate dllm inference. these studies build on the following key observation:for a subset of tokens, their kv states often exhibit...

Evidence 4 - **Rationale**: Both papers rely on the identical observation that KV states show high similarity across adjacent steps in diffusion models, and both cite this as the foundation for approximate caching strategies. - **Original**: the effectiveness of our approximate kv cache approach stems from the observation that kv activations exhibit high similarity across adjacent inference steps, as illustrated in figure 3. the red boxed region in figure 3a highlights the similarity scores within a block, which are consistently close t... - **Candidate**: recent studies on approximate kv cache in dllms have shown that the kv states of certain tokens exhibit high similarity across adjacent decoding steps (wu et al., 2025; liu et al., 2025). leveraging this redundancy, they first partition the entire sequence into a static segment and a dynamic segment...

Evidence 5 - **Rationale**: Both papers propose caching strategies that go beyond simple prefix caching to include multiple segments of the sequence, demonstrating similar architectural approaches to the problem. - **Original**: we further propose a dualcache version that caches keys and values for both prefix and suffix tokens. - **Candidate**: d2cache features a two-stage fine-grained selection strategy to identify tokens and adaptively update their kv states at each decoding step, while caching the kv states of the remaining tokens for reuse.

---

2. dLLM-Cache: Accelerating Diffusion Large Language Models with Adaptive Caching

URL: View paper

Prior Art Analysis

dLLM-Cache[10] demonstrates that prior work exists on KV caching strategies for diffusion language models with bidirectional attention. Both papers propose training-free caching mechanisms that exploit the observation that KV activations remain similar across adjacent denoising steps in masked diffusion models. dLLM-Cache[10] explicitly describes 'long-interval prompt caching with partial response updates guided by feature similarity' and achieves 'up to 9.1 x speedup' on the same base models (llada 8b and dream 7b). The original paper's claim to be first in proposing block-wise approximate KV cache for bidirectional diffusion models is refuted by dLLM-Cache[10]'s earlier demonstration of adaptive caching that addresses the identical technical challenge of bidirectional attention incompatibility with traditional KV caching.

Evidence

Evidence 1 - **Rationale**: Both papers propose training-free KV caching mechanisms for diffusion LLMs that enable cache reuse without performance degradation. dLLM-Cache[10] explicitly addresses the same technical challenge of bidirectional attention incompatibility. - **Original**: we introduce fast-dllm, a method that incorporates a novel block-wise approximate kv cache mechanism tailored for bidirectional diffusion models, enabling cache reuse with negligible performance drop. - **Candidate**: we propose dllm-cache, a training-free adaptive caching framework that combines long-interval prompt caching with partial response updates guided by feature similarity. this design enables efficient reuse of intermediate computations without compromising model performance.

Evidence 2 - **Rationale**: Both papers identify the identical technical challenge: KV caching incompatibility with bidirectional attention in diffusion LLMs. dLLM-Cache[10] explicitly states this problem and proposes a solution based on token stability across steps. - **Original**: first, diffusion llms do not support key-value (kv) caching, a critical component in ar models for speeding up inference. second, the generation quality tends to degrade when decoding multiple tokens in parallel. - **Candidate**: traditional arm acceleration techniques, such as key-value caching, are incompatible with dllms due to their bidirectional attention mechanism. to address this specific challenge, our work begins with a key observation that dllm inference involves a static prompt and a partially dynamic response, wh...

Evidence 3 - **Rationale**: Both papers base their caching mechanisms on the same fundamental observation about stability/similarity across adjacent steps in diffusion model inference. - **Original**: the effectiveness of our approximate kv cache approach stems from the observation that kv activations exhibit high similarity across adjacent inference steps, as illustrated in figure 3. - **Candidate**: our work begins with a key observation that dllm inference involves a static prompt and a partially dynamic response, where most tokens remain stable across adjacent denoising steps.

Evidence 4 - **Rationale**: Both papers evaluate on the same base models (llada and dream) and report significant speedups with negligible performance loss, demonstrating that dLLM-Cache[10] already validated this approach on identical architectures. - **Original**: experimental results on llada and dream models across multiple llm benchmarks demonstrate up to 27.6x throughput improvement with minimal accuracy loss - **Candidate**: extensive experiments on representative dllms, including llada 8b and dream 7b, show that dllm-cache achieves up to 9.1 x speedup over standard inference without compromising output quality.

---

3. Diffusion llm with native variable generation lengths: Let lead the way

URL: View paper

Brief Assessment

Native Variable Lengths[62] focuses on enabling variable-length generation through fixed EOS masking and multi-sample packing, not on KV cache mechanisms for bidirectional attention. Their KV cache reuse is a consequence of block-wise generation for variable lengths, not the core contribution.

---

4. ReFusion: A Diffusion Large Language Model with Parallel Autoregressive Decoding

URL: View paper

Brief Assessment

ReFusion[64] operates at the slot level with causal attention and full KV cache reuse, fundamentally different from the original paper's block-wise approximate KV cache for bidirectional attention in masked diffusion models. The architectural approaches are distinct.

---

5. dKV-Cache: The Cache for Diffusion Language Models

URL: View paper

Prior Art Analysis

dKV-Cache[63] demonstrates that prior work exists on KV caching mechanisms for bidirectional attention in diffusion language models. Both papers propose KV cache strategies tailored for masked diffusion models with full bidirectional attention, adopting block-wise or delayed generation strategies to enable cache reuse across decoding steps. The candidate paper explicitly addresses the same core challenge: enabling KV cache in diffusion models despite bidirectional attention, and proposes delayed caching strategies with negligible performance degradation—directly overlapping with the original paper's claimed novelty.

Evidence

Evidence 1 - **Rationale**: Both papers claim to introduce novel KV cache mechanisms specifically designed for bidirectional diffusion models. The candidate's 'delayed kv-cache' and the original's 'block-wise approximate kv cache' both address the same fundamental challenge of enabling cache reuse in non-autoregressive diffusion models. - **Original**: we introduce fast-dllm, a method that incorporates a novel block-wise approximate kv cache mechanism tailored for bidirectional diffusion models, enabling cache reuse with negligible performance drop. - **Candidate**: we address this bottleneck by proposing a kvcache-like mechanism, delayed kv-cache, for the denoising process of dlms. our approach is motivated by the observation that different tokens have distinct representation dynamics throughout the diffusion process. accordingly, we propose a delayed and cond...

Evidence 2 - **Rationale**: Both papers explicitly identify bidirectional attention as the core obstacle to KV caching in diffusion models and propose approximation strategies to overcome this limitation. This demonstrates prior recognition of the same technical challenge. - **Original**: first, fast-dllm introduces an approximate kv cache tailored to diffusion llms. while the bidirectional nature of attention in diffusion llms precludes a fully equivalent kv cache, our approximation closely resembles an ideal cache in practice. to support kv cache, we adopt a block-wise generation m... - **Candidate**: we identify two core reasons that prevent the direct usage of kv-cache in dlms. (1) kv-cache hinges on the assumption that the key and value states of previously generated tokens remain fixed during subsequent decoding steps. this property is preserved in autoregressive models through the use of a c...

Evidence 3 - **Rationale**: Both papers provide empirical evidence that KV activations show high similarity across adjacent steps in diffusion models, which is the foundational observation justifying their respective caching strategies. This demonstrates prior establishment of the same empirical insight. - **Original**: the effectiveness of our approximate kv cache approach stems from the observation that kv activations exhibit high similarity across adjacent inference steps, as illustrated in figure 3. the red boxed region in figure 3a highlights the similarity scores within a block, which are consistently close t... - **Candidate**: we investigate in dlms whether k and v can be reused. we focus on the dynamics of k and v for each token, and the results are shown in figure 2. interestingly, we observe several noteworthy patterns in the dynamics of qkv states: (1) despite step-to-step differences, the key and value embeddings, k ...

---

6. From slow bidirectional to fast autoregressive video diffusion models

URL: View paper

Brief Assessment

Bidirectional to Autoregressive[60] focuses on distilling bidirectional video diffusion models into autoregressive generators for streaming video generation. While it mentions KV caching for efficient inference, the paper does not propose a novel KV cache mechanism for bidirectional attention in masked diffusion language models, which is the core contribution of the original paper.

---

7. d2Cache: Accelerating Diffusion-Based LLMs via Dual Adaptive Caching

URL: View paper

Prior Art Analysis

d2Cache[65] demonstrates that prior work exists on approximate KV caching mechanisms for bidirectional diffusion language models. Both papers address the fundamental challenge that bidirectional attention in diffusion LLMs prevents direct application of standard KV cache (as used in autoregressive models). d2Cache[65] presents a fine-grained token-level selection strategy that adaptively updates KV states at each decoding step, while the original paper proposes block-wise generation with prefix/suffix caching. The core innovation claimed by the original paper—enabling KV cache reuse in bidirectional diffusion models through approximate caching—is demonstrated to have been explored in d2Cache[65], which was developed concurrently or prior to the original submission.

Evidence

Evidence 1 - **Rationale**: Both papers rely on the same foundational observation about KV state similarity across adjacent steps. d2Cache[65] explicitly acknowledges this as established knowledge in the field, suggesting the observation itself was not novel to the original paper. - **Original**: The effectiveness of our approximate kv cache approach stems from the observation that kv activations exhibit high similarity across adjacent inference steps, as illustrated in figure 3. the red boxed region in figure 3a highlights the similarity scores within a block, which are consistently close t... - **Candidate**: recent studies on approximate kv cache in dllms have shown that the kv states of certain tokens exhibit high similarity across adjacent decoding steps (wu et al., 2025; liu et al., 2025). leveraging this redundancy, they first partition the entire sequence into a static segment and a dynamic segment...

---

8. Fast-dllm v2: Efficient block-diffusion llm

URL: View paper

Prior Art Analysis

Fast-dllm v2[32] demonstrates that similar block-wise KV caching mechanisms for diffusion language models were developed prior to the ORIGINAL paper. Both papers propose caching strategies that exploit temporal similarity across adjacent decoding steps in masked diffusion models. Fast-dllm v2[32] explicitly describes a hierarchical caching mechanism with block-level cache and sub-block cache (dualcache) that stores prefix and suffix tokens, enabling computational reuse across blocks—matching the core innovation claimed in the ORIGINAL paper. The evidence shows Fast-dllm v2[32] was published earlier and addresses the same technical challenge of enabling KV cache in bidirectional attention diffusion models.

Evidence

Evidence 1 - **Rationale**: Both papers propose block-wise caching mechanisms that exploit similarity across steps. Fast-dllm v2[32] explicitly describes a hierarchical cache including block-level and sub-block components, directly corresponding to the ORIGINAL's

claimed innovation. - **Original**: we introduce a block-wise approximate kv cache mechanism specifically designed for bidirectional attention. our approach reuses cached activations from previously decoded blocks by exploiting the high similarity of kv activations between adjacent steps. - **Candidate**: we design a hierarchical caching mechanism: a block-level cache that stores historical context representations across blocks, and a sub-block cache that supports efficient parallel decoding within partially generated blocks which adopts the dualcache in fast-dllm

Evidence 2 - **Rationale**: The ORIGINAL paper's dualcache innovation—caching both prefix and suffix tokens—is explicitly described in Fast-dllm v2[32], indicating prior work on this exact mechanism. - **Original**: we implement a bidirectional version of our kv caching mechanism, named dualcache, that caches not only the prefix tokens but also the suffix tokens, which consist entirely of masked tokens under our block-wise decoding scheme. - **Candidate**: dualcache. this method caches the kv activations for both the preceding text (prefix) and the subsequent masked tokens (suffix).

Evidence 3 - **Rationale**: The block-wise generation and caching strategy described in the ORIGINAL paper is present in Fast-dllm v2[32], which caches decoded blocks for reuse in subsequent blocks. - **Original**: to support kv cache, we adopt a block-wise generation manner. before generating a block, we compute and store kv cache of the other blocks to reuse. after generating the block, we recompute the kv cache of all the blocks. - **Candidate**: since each block in fast-dllm v2 is decoded in a causal order, we naturally preserve left-to-right semantics across blocks. after decoding each block, its unmasked tokens are cached as read-only context for future blocks.

---

9. Accelerating diffusion language model inference via efficient kv caching and guided diffusion

URL: View paper

Prior Art Analysis

KV Caching Guided[15] demonstrates that similar KV caching strategies for diffusion language models were proposed prior to the ORIGINAL paper. Both papers propose training-free KV caching mechanisms that exploit the temporal stability of key-value projections across denoising steps in masked diffusion models. The candidate paper's 'freecache' method uses a reducing window caching strategy where KV projections are frozen for completed blocks, which is conceptually similar to the ORIGINAL paper's block-wise generation with prefix/suffix caching. Both methods observe high similarity of KV activations across adjacent inference steps and leverage this for computational reuse.

Evidence

Evidence 1 - **Rationale**: Both papers propose KV caching mechanisms specifically designed for diffusion language models that reuse projections across denoising steps, demonstrating similar core contributions. - **Original**: we introduce fast-dllm, a method that incorporates a novel block-wise approximate kv cache mechanism tailored for bidirectional diffusion models, enabling cache reuse with negligible performance drop. - **Candidate**: we propose freecache, a key-value (kv) approximation caching technique that reuses stable kv projections across denoising steps, effectively reducing the computational cost of dlm inference.

Evidence 2 - **Rationale**: Both papers describe block-wise caching strategies where KV projections are computed for blocks, reused during generation, and updated after block completion. The mechanisms are functionally equivalent. - **Original**: as shown in figure 2, we adopt a block-wise decoding strategy to support the use of a key-value (kv) cache. initially, we compute and store the kv cache for the prompt, which is reused throughout block 0. within each block, the same cache is reused for multiple decoding steps. after completing the d... - **Candidate**: we propose freecache, a reducing window caching strategy where the set of actively computed tokens dynamically shrinks as blocks of the sequence are finalized. the algorithm proceeds as follows:initialization and partitioning:the post-prompt generation sequence is partitioned into fixed-size blocks ...

Evidence 3 - **Rationale**: Both papers provide the same empirical observation and justification: KV projections show high temporal stability/similarity across adjacent denoising steps, enabling safe reuse with minimal approximation error. - **Original**: the effectiveness of our approximate kv cache approach stems from the observation that kv activations exhibit high similarity across adjacent inference steps, as illustrated in figure 3. the red boxed region in figure 3a highlights the similarity scores within a block, which are consistently close t... - **Candidate**: a key insight for optimizing the generation process in diffusion language models is that the key and value (kv) projections for the clean portions of a sequence exhibit high temporal stability. although these projections are not strictly unchanging, they quickly converge and undergo only negligible ...

Evidence 4 - **Rationale**: The ORIGINAL paper's dualcache (prefix + suffix) and the candidate's reducing window approach (caching completed blocks while computing active window) both exploit bidirectional caching opportunities in diffusion models. - **Original**: we further propose a dualcache version that caches keys and values for both prefix and suffix tokens. - **Candidate**: windowed re-computation:to generate a block bi, theactive computation windowis defined as bi and all subsequent blocks (bi+1, ..., bn). kv projections are recomputed only for tokens within this window until bi is fully unmasked, using all prior frozen blocks and the prompt as context.

---

10. Attention is all you need for kv cache in diffusion llms

URL: View paper

Prior Art Analysis

Attention KV Cache[61] demonstrates that prior work exists on KV caching mechanisms for diffusion LLMs with bidirectional attention. Both papers propose caching strategies that reuse key-value states across decoding steps in masked diffusion models. The candidate paper introduces sliding window-based caching with adaptive updates based on attention patterns, while the original proposes block-wise caching with prefix/suffix reuse. Both address the same fundamental challenge: enabling KV cache in bidirectional attention contexts where traditional autoregressive caching fails.

Evidence

Evidence 1 - **Rationale**: Both papers explicitly address the challenge of KV caching in bidirectional diffusion models, acknowledging that traditional autoregressive caching assumptions break down. - **Original**: we introduce a block-wise approximate kv cache mechanism specifically designed for bidirectional attention. our approach reuses cached activations from previously decoded blocks by exploiting the high similarity of kv activations between adjacent steps. - **Candidate**: diffusion llms differ from autoregressive decoders in that their key-value (kv) states evolve across denoising steps due to bidirectional dependencies. our objective is to adaptively decidewhen andwhereto recompute the kv cache to preserve accuracy while minimizing latency.

Contribution 2: Confidence-aware parallel decoding strategy

Description: The authors introduce a dynamic decoding approach that selectively decodes tokens based on confidence thresholds rather than a fixed count per step. This strategy mitigates token-dependency violations under the conditional independence assumption and maintains generation quality while accelerating inference by up to 13.3×.

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. Dynamic early exit in reasoning models

URL: View paper

Brief Assessment

Dynamic Early Exit[53] focuses on early termination of reasoning chains in large reasoning models (LRMs) during test-time scaling, not on parallel token decoding for diffusion-based language models. The candidate addresses when to stop generating reasoning steps based on confidence thresholds, while the original paper addresses how many tokens to decode simultaneously in masked diffusion models.

---

2. Diffgrm: Diffusion-based generative recommendation model

URL: View paper

Brief Assessment

DiffGRM[54] applies confidence-based decoding to recommendation systems with semantic IDs, not general language model inference acceleration. The technical domain and application context differ fundamentally from the original paper's focus on diffusion LLMs.

---

3. Deep think with confidence

URL: View paper

Brief Assessment

Deep Think Confidence[52] focuses on filtering reasoning traces based on confidence scores in ensemble voting scenarios for test-time scaling, not on accelerating diffusion LLM inference through parallel token decoding. The candidate addresses a different problem domain (reasoning trace selection) rather than the original's focus on mitigating token-dependency violations during parallel decoding in diffusion models.

---

4. Accelerating Diffusion LLM Inference via Local Determinism Propagation

URL: View paper

Prior Art Analysis

Local Determinism Propagation[25] demonstrates that prior work exists on confidence-based dynamic parallel decoding for diffusion LLMs. Both papers address the same core problem: using confidence thresholds to selectively decode tokens in parallel while maintaining generation quality. The candidate paper explicitly describes a 'confidence-aware parallel decoding' approach that dynamically selects tokens based on confidence criteria, directly overlapping with the original paper's claimed novelty. The candidate's methodology of identifying high-confidence 'anchors' and performing localized parallel decoding within bounded neighborhoods represents a specific implementation of confidence-based dynamic parallel decoding that predates or is contemporaneous with the original work.

Evidence

Evidence 1 - **Rationale**: Both papers describe confidence-based parallel decoding strategies for diffusion LLMs. The candidate explicitly discusses 'confidence-aware parallel decoding' as a foundational concept, demonstrating prior work on this approach. - **Original**: we propose a novel confidence-aware parallel decoding method. Unlike prior approaches that select a fixed number of tokens per step, our method dynamically selects tokens whose confidence exceeds a global threshold, enabling safe and effective parallel decoding. - **Candidate**: confidence-aware parallel decoding.the bidirectional attention and arbitrary-position update capability of dllms make them inherently well-suited for parallel decoding, but conservative quality safeguards often limit decoding speed. Suppose a set of n to be predicted positions given a conditioning c...

Evidence 2 - **Rationale**: Both papers propose confidence-threshold-based strategies for parallel decoding. The candidate's LocalLeap method uses confidence thresholds to determine which tokens can be decoded in parallel, demonstrating that this approach existed prior to or contemporaneously with the original work. - **Original**: To address this issue and fully exploit the parallelism potential of diffusion llms, we propose a novel confidence-thresholding strategy to select which tokens can be safely decoded simultaneously. instead of selecting the tokens with top k confidence to decode as in llada, we select tokens with con... - **Candidate**: we proposelocalleap, a training-free adaptive parallel decoding strategy. our method identifies anchors in the sequence and performs parallel decoding within a bounded radius around each anchor, using a relaxed confidence threshold to define the decoding boundary. this adaptive sampling strategy ens...

Evidence 3 - **Rationale**: Both papers describe methods that dynamically select tokens for parallel decoding based on confidence thresholds. The candidate's approach of identifying high-confidence tokens (anchors) and using confidence criteria (κ and τ) for parallel decoding demonstrates prior work on confidence-aware dynamic parallel decoding strategies. - **Original**: confidence-aware parallel decoding.we propose a novel confidence-aware parallel decoding method. Unlike prior approaches that select a fixed number of tokens per step, our method dynamically selects tokens whose confidence exceeds a global threshold, enabling safe and effective parallel decoding. Th... - **Candidate**: At decoding step s, let bs be the active decoding block and ms ⊆ bs the set of masked indices awaiting decoding. For each masked position i∈ ms, mask predictor pθ produces a confidence score ci s as defined in eq. 2. Our objective is to identify a subset ds ⊆ ms that can be safely decoded in paralle...

---

5. Dimple: Discrete diffusion multimodal large language model with parallel decoding

URL: View paper

Prior Art Analysis

Dimple[55] demonstrates prior work on confidence-based dynamic parallel decoding for diffusion language models. Both papers propose selecting tokens based on confidence thresholds rather than fixed counts per step. The candidate paper's 'confident decoding' dynamically adjusts token counts using a threshold γ, where tokens with confidence ≥ γ are decoded simultaneously—directly paralleling the original paper's confidence-thresholding strategy. Both approaches aim to mitigate token-dependency violations under conditional independence assumptions while accelerating inference through adaptive parallel decoding.

Evidence

Evidence 1 - **Rationale**: Both papers propose threshold-based token selection where tokens exceeding a confidence threshold are decoded in parallel, rather than selecting a fixed top-k count. This demonstrates the same core innovation. - **Original**: we propose a novel confidence-thresholding strategy to select which tokens can be safely decoded simultaneously. instead of selecting the tokens with top k confidence to decode as in llada, we select tokens with confidence larger than a threshold. - **Candidate**: we therefore propose confident decoding, which dynamically adjusts the number of tokens updated per step based on a fixed confidence threshold $\gamma \in (0, 1)$. formally, at each step t, we define it = {i | c(i) t ≥ γ}, where c(i) t is the confidence score for position i. if it is non-empty, all tokens at...

Evidence 2 - **Rationale**: Both approaches justify the strategy by enabling safe parallel decoding when confidence is high while preserving quality by avoiding uncertain predictions. - **Original**: To address this issue and fully exploit the parallelism potential of diffusion llms, we propose a novel confidence-thresholding strategy to select which tokens can be safely decoded simultaneously. - **Candidate**: This method enables: • decoding multiple tokens simultaneously when model is highly confident, improving efficiency; • avoiding low-confidence updates, preserving generation quality.

Evidence 3 - **Rationale**: Both papers explicitly contrast their dynamic threshold approach against prior fixed-count methods, highlighting the same technical distinction and motivation. - **Original**: second,confidence-aware parallel decoding.we propose a novel confidence-aware parallel decoding method. unlike prior approaches that select a fixed number of tokens per step, our method dynamically selects tokens whose confidence exceeds a global threshold, enabling safe and effective parallel decod... - **Candidate**: in previous confidence-based decoding [5, 39, 31], the number of tokens decoded per step is fixed. however, we argue that decoding should adapt to the

semantic structure of the text: some steps may allow many tokens to be confidently predicted, while others may necessitate more caution.

---

6. Spec-LLaVA: Accelerating Vision-Language Models with Dynamic Tree-Based Speculative Decoding

URL: View paper

Brief Assessment

Spec-LLaVA[59] focuses on speculative decoding for vision-language models using a draft-target model pair with tree-based verification, whereas the original paper addresses diffusion-based language models with confidence thresholding for direct parallel token generation. These are fundamentally different acceleration paradigms for different model architectures.

---

7. Collaborative Speculative Inference for Efficient LLM Inference Serving

URL: View paper

Brief Assessment

Collaborative Speculative Inference[57] focuses on confidence-based token fusion for combining outputs from multiple specialized drafters in a distributed speculative inference system, not on confidence-thresholding for selective parallel decoding within a single diffusion model's generation process.

---

8. Confidence-Modulated Speculative Decoding for Large Language Models

URL: View paper

Brief Assessment

Confidence-Modulated Speculative[51] focuses on speculative decoding with draft-then-verify for autoregressive models, using entropy and margin-based uncertainty measures. The original paper addresses parallel decoding in diffusion LLMs using confidence thresholds to select tokens, which is a fundamentally different architecture and decoding paradigm.

---

9. Introducing dynamic token embedding sampling of large language models for improved inference accuracy

URL: View paper

Brief Assessment

Dynamic Token Embedding[56] focuses on dynamic token embedding sampling during inference, not confidence-based parallel decoding strategies for diffusion language models. The candidate addresses embedding layer operations rather than the selective token decoding mechanism described in the original paper.

Contribution 3: Fast-dLLM framework achieving state-of-the-art acceleration for Diffusion LLMs

Description: The authors present Fast-dLLM, an integrated framework combining block-wise KV caching and confidence-aware parallel decoding. Experiments show up to 27.6× end-to-end speedup on multiple benchmarks with minimal accuracy loss, closing the performance gap with autoregressive models and enabling practical deployment of Diffusion LLMs.

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. dinfer: An efficient inference framework for diffusion language models

URL: View paper

Prior Art Analysis

dinfer[9] demonstrates that similar acceleration techniques for Diffusion LLMs were developed independently and published prior to or contemporaneously with the original paper. Both papers address the same core challenges: implementing KV caching for bidirectional attention in diffusion models and enabling parallel decoding strategies. dinfer[9] presents a complete framework with block-wise KV caching (prefix and dual cache variants), confidence-aware parallel decoding strategies (threshold, hierarchical, credit), and achieves comparable or superior speedups (10x over baseline, 2-3x over AR models). The substantial overlap in technical approach, problem formulation, and experimental methodology indicates that the original paper's novelty claim of being first to propose this integrated framework is refuted by dinfer[9]'s prior or concurrent work.

Evidence

Evidence 1 - **Rationale**: Both papers present frameworks that incorporate block-wise KV caching mechanisms for diffusion models. dinfer[9] explicitly describes a modular KV-cache management component as part of its core architecture. - **Original**: we introduce fast-dllm, a method that incorporates a novel block-wise approximate kv cache mechanism tailored for bidirectional diffusion models, enabling cache reuse with negligible performance drop. - **Candidate**: dinfer modularizes inference into four components-model, diffusion iteration manager, decoding strategy, and kv-cache management-and provides well-designed apis for flexible combinations of algorithms in each component.

Evidence 2 - **Rationale**: Both papers identify the same fundamental challenge of KV cache incompatibility in bidirectional diffusion models and propose block-wise caching solutions to address it. - **Original**: first,key-value cache for block-wise decoding.we introduce a block-wise approximate kv cache mechanism specifically designed for bidirectional attention. our approach reuses cached activations from previously decoded blocks by exploiting the high similarity of kv activations between adjacent steps. - **Candidate**: A central challenge in dllm inference is kv-cache incompatibility. in ar models, causal attention allows kv states to be computed once and reused; in dllms, however, token representations evolve across denoising steps, making static reuse infeasible. without caching, inference must perform transform...

Evidence 3 - **Rationale**: dinfer[9] implements multiple confidence-aware parallel decoding strategies including threshold-based decoding, demonstrating prior work on this approach. The citation to 'fast-dllm (wu et al., 2025)' suggests this technique was known in the field. - **Original**: second,confidence-aware parallel decoding.we propose a novel confidence-aware parallel decoding method. unlike prior approaches that select a fixed number of tokens per step, our method dynamically selects tokens whose confidence exceeds a global threshold, enabling safe and effective parallel decod... - **Candidate**: dinfer supports three strategies for parallel decoding: • threshold decoding (from fast-dllm (wu et al., 2025)): commits tokens whose confidence exceeds a preset threshold. • hierarchical decoding (ours): recursively partitions masked spans, ensuring at least one token is decoded per region, thereby...

Evidence 4 - **Rationale**: Both papers report similar magnitude speedups (10x+ over baselines, 2-3x over AR models) using comparable techniques, indicating parallel development of similar acceleration frameworks. - **Original**: experimental results on llada and dream models across multiple llm benchmarks demonstrate up to 27.6x throughput improvement with minimal accuracy loss, closing the performance gap with autoregressive models and paving the way for practical deployment of diffusion llms. - **Candidate**: On humaneval, dinfer achieves over 1,100 tps at batch size 1, and averages more than 800 tps across six benchmarks on a single node with 8x h800 gpus. compared to fast-dllm (wu et al., 2025), dinfer delivers more than a 10x speedup while maintaining accuracy; on llada-moe it provides a 2 - 3x speedu...

Evidence 5 - **Rationale**: dinfer[9] explicitly mentions dual cache as an existing approach, indicating this technique was already known in the field before the original paper's submission. - **Original**: By caching both prefix and suffix blocks, the dualcache strategy enables

substantial computational reuse. - **Candidate**: Earlier approaches introduced training-free strategies such as blockwise caching and dual cache (wu et al., 2025), which reuse kv states for decoded tokens or suffixes of masked tokens.

---

2. A survey on diffusion language models

URL: View paper

Brief Assessment

Diffusion Language Survey[5] is a survey paper that reviews existing techniques rather than proposing novel methods. It discusses acceleration techniques like caching and parallel decoding as existing work in the field, not as original contributions.

---

3. Learning-to-cache: Accelerating diffusion transformer via layer caching

URL: View paper

Brief Assessment

Learning-to-cache[69] focuses on accelerating diffusion transformers for image generation tasks (DiT, U-ViT) through layer caching mechanisms, not on diffusion-based language models. The candidate addresses visual generation while the original targets text generation with Diffusion LLMs.

---

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

URL: View paper

Brief Assessment

Free Draft-and-Verification[67] focuses on lossless parallel decoding through draft-and-verification mechanisms, while Fast-dLLM combines block-wise KV caching with confidence-aware parallel decoding. These are complementary approaches addressing different bottlenecks in diffusion LLM inference.

---

5. Accelerating Diffusion LLMs via Adaptive Parallel Decoding

URL: View paper

Brief Assessment

Adaptive Parallel Decoding[16] focuses on a fundamentally different acceleration approach using multiplicative mixtures with auxiliary autoregressive models for adaptive token selection, rather than the block-wise KV caching and confidence-thresholding strategies proposed in Fast-dLLM.

---

6. Smoothcache: A universal inference acceleration technique for diffusion transformers

URL: View paper

Brief Assessment

SmoothCache[68] focuses on accelerating Diffusion Transformers (DiT) for image, video, and audio generation tasks, not on Diffusion-based Large Language Models (LLMs) for text generation. The candidate addresses different model architectures and application domains than the original paper's focus on language modeling.

---

7. Creditdecoding: Accelerating parallel decoding in diffusion large language models with trace credits

URL: View paper

Brief Assessment

CreditDecoding[8] addresses parallel decoding acceleration through trace credit accumulation and historical logit fusion, whereas Fast-dLLM focuses on block-wise KV caching and confidence-aware parallel decoding. These are complementary approaches to different bottlenecks in Diffusion LLM inference.

---

8. Deepcache: Accelerating diffusion models for free

URL: View paper

Brief Assessment

DeepCache[66] focuses on accelerating image generation in diffusion models (Stable Diffusion, LDM) by caching high-level features in U-Net architectures across denoising steps. Fast-dLLM targets text generation in Diffusion Language Models using block-wise KV caching and confidence-aware parallel decoding for token sequences. These are fundamentally different application domains and technical approaches.

---

9. Attention is all you need for kv cache in diffusion llms

URL: View paper

Prior Art Analysis

Attention KV Cache[61] presents an acceleration framework (Elastic-Cache) that combines KV caching with confidence-aware parallel decoding for diffusion LLMs, achieving comparable or superior speedups to the original Fast-dLLM. Both frameworks integrate caching mechanisms with parallel decoding strategies to accelerate diffusion LLM inference. The candidate reports speedups up to 45.1× on similar benchmarks (GSM8K, MATH, HumanEval, MBPP) using the same base models (LLaDA), demonstrating that prior work achieved state-of-the-art acceleration before the original submission.

Evidence

Evidence 1 - **Rationale**: The candidate achieves higher speedups (45.1× vs 27.6×) on the same benchmarks and models, demonstrating that comparable or superior acceleration was achieved in prior work. - **Original**: experimental results on llada and dream models across multiple llm benchmarks demonstrate up to 27.6x throughput improvement with minimal accuracy loss, closing the performance gap with autoregressive models - **Candidate**: across tables 1, 3, and 5, our proposedelastic-cachedelivers substantial throughput gains for diffusion llms with minimal accuracy loss. by adaptively updating the cache only when necessary, it achieves a speedup of up to 45.1xover the standard baseline.

Evidence 2 - **Rationale**: Both papers evaluate on identical benchmarks (GSM8K, MATH, HumanEval, MBPP) and models (LLaDA, Dream), with the candidate achieving comparable or superior results, indicating prior state-of-the-art acceleration existed. - **Original**: we conduct comprehensive experiments on multiple open-source diffusion llms (llada, dream) and four mainstream benchmarks (gsm8k, math, humaneval, mbpp). Results demonstrate that our fast-dllm consistently deliver order-of-magnitude speedups with minimal or no degradation in accuracy - **Candidate**: we evaluateelasticcacheon llada-instruct (nie et al., 2025a), llada-1.5 (zhu et al., 2025), and multimodal llada-v (you et al., 2025) across mbpp (austin et al., 2021b), humaneval (chen et al., 2021), math (hendrycks et al., 2021), gsm8k (cobbe et al., 2021)

Appendix: Text Similarity Detection

Textual similarity detection checked 31 papers and found 4 similarity segment(s) across 2 paper(s).

The following **2 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Self Speculative Decoding for Diffusion Large Language Models

Detected in: Core Task (sibling)

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

---

2. Fast-dllm v2: Efficient block-diffusion llm

Detected in: Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Fast-dLLM: Training-free Acceleration of Diffusion LLM by Enabling KV Cache and Parallel Decoding View paper
- [1] Simple and effective masked diffusion language models View paper
- [2] Sequential diffusion language models View paper
- [3] Dream 7b: Diffusion large language models View paper
- [4] SparseD: Sparse Attention for Diffusion Language Models View paper
- [5] A survey on diffusion language models View paper
- [6] A survey on parallel text generation: From parallel decoding to diffusion language models View paper
- [7] Revolutionizing reinforcement learning framework for diffusion large language models View paper
- [8] Creditdecoding: Accelerating parallel decoding in diffusion large language models with trace credits View paper
- [9] dinfer: An efficient inference framework for diffusion language models View paper
- [10] dLLM-Cache: Accelerating Diffusion Large Language Models with Adaptive Caching View paper
- [11] Plan for Speed--Dilated Scheduling for Masked Diffusion Language Models View paper
- [12] Self Speculative Decoding for Diffusion Large Language Models View paper
- [13] Efficient Perplexity Bound and Ratio Matching in Discrete Diffusion Language Models View paper
- [14] Diffuspec: Unlocking diffusion language models for speculative decoding View paper
- [15] Accelerating diffusion language model inference via efficient kv caching and guided diffusion View paper
- [16] Accelerating Diffusion LLMs via Adaptive Parallel Decoding View paper
- [17] Speculative diffusion decoding: Accelerating language generation through diffusion View paper
- [18] Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding View paper
- [19] Saber: An Efficient Sampling with Adaptive Acceleration and Backtracking Enhanced Remasking for Diffusion Language Model View paper
- [20] Diffusion models in text generation: a survey View paper
- [21] Few-shot temporal pruning accelerates diffusion models for text generation View paper
- [22] A Convergence Theory for Diffusion Language Models: An Information-Theoretic Perspective View paper
- [23] DLLMQuant: Quantizing Diffusion-based Large Language Models View paper
- [24] Your Absorbing Discrete Diffusion Secretly Models the Conditional Distributions of Clean Data View paper
- [25] Accelerating Diffusion LLM Inference via Local Determinism Propagation View paper
- [26] dCache: Accelerating Diffusion-Based LLMs via Dual Adaptive Caching View paper
- [27] Accelerating Diffusion Large Language Models with SlowFast: The Three Golden Principles View paper
- [28] Diffusion-based Large Language Models Survey View paper
- [29] Discrete diffusion models for language generation View paper
- [30] Energy-based diffusion language models for text generation View paper
- [31] Ctrldiff: Boosting large diffusion language models with dynamic block prediction and controllable generation View paper
- [32] Fast-dllm v2: Efficient block-diffusion llm View paper
- [33] Encoder-Decoder Diffusion Language Models for Efficient Training and Inference View paper
- [34] Beyond Next-Token Prediction: A Performance Characterization of Diffusion versus Autoregressive Language Models View paper
- [35] How Efficient Are Diffusion Language Models? A Critical Examination of Efficiency Evaluation Practices View paper
- [36] AdaBlock-dLLM: Semantic-Aware Diffusion LLM Inference via Adaptive Block Size View paper
- [37] Discrete Diffusion in Large Language and Multimodal Models: A Survey View paper
- [38] FS-DFM: Fast and Accurate Long Text Generation with Few-Step Diffusion Language Models View paper
- [39] Planner Aware Path Learning in Diffusion Language Models Training View paper
- [40] Taming Masked Diffusion Language Models via Consistency Trajectory Reinforcement Learning with Fewer Decoding Step View paper
- [41] Lavida: A large diffusion language model for multimodal understanding View paper
- [42] Diffusion Language Models Know the Answer Before Decoding View paper
- [43] Sparse-LaViDa: Sparse Multimodal Discrete Diffusion Language Models View paper
- [44] Discrete diffusion language model for efficient text summarization View paper
- [45] Context-Aware Initialization for Reducing Generative Path Length in Diffusion Language Models View paper
- [46] Diffusion llms can do faster-than-ar inference via discrete diffusion forcing View paper
- [47] Quantization meets dllms: A systematic study of post-training quantization for diffusion llms View paper
- [48] Set Block Decoding is a Language Model Inference Accelerator View paper
- [49] DLM-One: Diffusion Language Models for One-Step Sequence Generation View paper
- [50] A2D: Any-Order, Any-Step Safety Alignment for Diffusion Language Models View paper
- [51] Confidence-Modulated Speculative Decoding for Large Language Models View paper
- [52] Deep think with confidence View paper
- [53] Dynamic early exit in reasoning models View paper
- [54] Diffgrm: Diffusion-based generative recommendation model View paper
- [55] Dimple: Discrete diffusion multimodal large language model with parallel decoding View paper

- [56] Introducing dynamic token embedding sampling of large language models for improved inference accuracy View paper
- [57] Collaborative Speculative Inference for Efficient LLM Inference Serving View paper
- [58] Speculative Decoding via Hybrid Drafting and Rollback-Aware Branch Parallelism View paper
- [59] Spec-LLaVA: Accelerating Vision-Language Models with Dynamic Tree-Based Speculative Decoding View paper
- [60] From slow bidirectional to fast autoregressive video diffusion models View paper
- [61] Attention is all you need for kv cache in diffusion llms View paper
- [62] Diffusion llm with native variable generation lengths: Let lead the way View paper
- [63] dKV-Cache: The Cache for Diffusion Language Models View paper
- [64] ReFusion: A Diffusion Large Language Model with Parallel Autoregressive Decoding View paper
- [65] d2Cache: Accelerating Diffusion-Based LLMs via Dual Adaptive Caching View paper
- [66] Deepcache: Accelerating diffusion models for free View paper
- [67] Free Draft-and-Verification: Toward Lossless Parallel Decoding for Diffusion Large Language Models View paper
- [68] Smoothcache: A universal inference acceleration technique for diffusion transformers View paper
- [69] Learning-to-cache: Accelerating diffusion transformer via layer caching View paper