Novelty Assessment Report

Paper: Fast-dLLM v2: Efficient Block-Diffusion LLM
PDF URL: https://openreview.net/pdf?id=1NZ3DHF9nT
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-05

Abstract

Autoregressive (AR) large language models (LLMs) have achieved remarkable performance across a wide range of natural language tasks, yet their inherent sequential decoding limits inference efficiency. In this work, we propose Fast-dLLM v2, a carefully designed block diffusion language model (dLLM) that efficiently adapts pretrained AR models into dLLMs for parallel text generation—requiring only ~1B tokens of fine-tuning. This represents a 500× reduction in training data compared to full-attention diffusion LLMs such as Dream (580B tokens), while preserving the original model's performance. Our approach introduces a novel training recipe that combines a block diffusion mechanism with a complementary attention mask, enabling blockwise bidirectional context modeling without sacrificing AR training objectives. To further accelerate decoding, we design a hierarchical caching mechanism: a block-level cache that stores historical context representations across blocks, and a sub-block cache that enables efficient parallel generation within partially decoded blocks. Coupled with our parallel decoding pipeline, Fast-dLLM v2 achieves up to 2.5× speedup over standard AR decoding without compromising generation quality. Extensive experiments across diverse benchmarks demonstrate that Fast-dLLM v2 matches or surpasses AR baselines in accuracy, while delivering state-of-the-art efficiency among dLLMs—marking a significant step toward the practical deployment of fast and accurate LLMs. Code and model will be publicly released.

Core Task Landscape

This paper addresses: **Accelerating Large Language Model Inference through Block Diffusion**

A total of **26 papers** were analyzed and organized into a taxonomy with **10 categories**.

Taxonomy Overview

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

- **Block Diffusion Architecture and Training Methods**
- **Inference Optimization and Acceleration Techniques**
- **Domain-Specific Applications and Extensions**
- **Surveys and Comparative Studies**

Complete Taxonomy Tree

- Accelerating Large Language Model Inference through Block Diffusion Survey Taxonomy
- Block Diffusion Architecture and Training Methods
  - Autoregressive-to-Diffusion Conversion and Adaptation ★ (4 papers)
  - [0] Fast-dLLM v2: Efficient Block-Diffusion LLM (Anon et al., 2026) View paper
  - [10] LLaDA2. 0: Scaling Up Diffusion Language Models to 100B (Tiwei Bie, 2025) View paper
  - [11] From Next-Token to Next-Block: A Principled Adaptation Path for Diffusion LLMs (Yuchuan Tian, 2025) View paper
  - [14] Efficient-DLM: From Autoregressive to Diffusion Language Models, and Beyond in Speed (Yonggan Fu, 2025) View paper
  - Novel Architecture Design for Block Diffusion (5 papers)
  - [2] Set block decoding is a language model inference accelerator (Gat, 2025) View paper
  - [6] Encoder-Decoder Diffusion Language Models for Efficient Training and Inference (Schiff, 2025) View paper
  - [18] Tandem Transformers for Inference Efficient LLMs (Nair, 2024) View paper
  - [23] Block Diffusion: Interpolating Between Autoregressive and Diffusion Language Models (Gokaslan, 2025) View paper
  - [24] ReFusion: A Diffusion Large Language Model with Parallel Autoregressive Decoding (Jia-Nan Li, 2025) View paper
  - Variable-Length and Adaptive Block Generation (4 papers)
  - [5] Sequential diffusion language models (Liu Yangzhou, 2025) View paper
  - [8] AdaBlock-dLLM: Semantic-Aware Diffusion LLM Inference via Adaptive Block Size (Wang Zhi-can, 2025) View paper
  - [16] Diffusion llm with native variable generation lengths: Let lead the way (Y Yang, 2025) View paper
- Inference Optimization and Acceleration Techniques
  - Key-Value Cache Optimization for Diffusion Models (2 papers)
  - [7] Attention Is All You Need for KV Cache in Diffusion LLMs (Ranjan, 2025) View paper
  - [15] dCache: Accelerating Diffusion-Based LLMs via Dual Adaptive Caching (Y Jiang, 2025) View paper
  - General Inference Acceleration and Efficiency Analysis (3 papers)
  - [1] Accelerating Diffusion Large Language Models with SlowFast: The Three Golden Principles (Q Wei, 2025) View paper
  - [13] Beyond Next-Token Prediction: A Performance Characterization of Diffusion versus Autoregressive Language Models (Kim Minseo, 2025) View paper
  - [17] Efficient Inference in Large Language Models (Yu-Cheng, 2025) View paper
  - Controllability and Constrained Generation (2 papers)
  - [3] Ctrldiff: Boosting large diffusion language models with dynamic block prediction and controllable generation (Huang Chi-Han, 2025) View paper

- [19] DINGO: Constrained Inference for Diffusion LLMs (Suresh, 2025) View paper
- Domain-Specific Applications and Extensions
  - Multimodal and Vision-Language Models (3 papers)
  - [9] Inferix: A Block-Diffusion based Next-Generation Inference Engine for World Simulation (Inferix Team, 2025) View paper
  - [20] DiffusionVL: Translating Any Autoregressive Models into Diffusion Vision Language Models (Lunbin Zeng, 2025) View paper
  - [22] From Text to Talk: Audio-Language Model Needs Non-Autoregressive Joint Training (Liu, 2025) View paper
  - Speech Recognition and Non-Autoregressive Decoding (1 papers)
  - [21] Towards Effective and Efficient Non-autoregressive decoders for Conformer and LLM-based ASR using Block-based Attention Mask (Tianzi Wang, 2025) View paper
  - Specialized Generation Tasks (1 papers)
  - [26] Diffusion Beats ARM: Diffusion Large Language Models for Generative Recommendation (H Jiang, n.d.) View paper
- Surveys and Comparative Studies (2 papers)
  - [4] Diffusion-based Large Language Models Survey (Chiung-Yi Tseng, 2025) View paper
  - [12] Block transformer: Global-to-local language modeling for fast inference (Sangmin Bae, 2024) View paper

Narrative

Core task: Accelerating large language model inference through block diffusion. The field has coalesced around the idea of replacing token-by-token autoregressive generation with block-level diffusion processes that predict multiple tokens simultaneously. The taxonomy reflects four main branches: Block Diffusion Architecture and Training Methods explores how to convert or adapt pretrained autoregressive models into diffusion frameworks, including techniques for initializing diffusion parameters and designing block-level objectives; Inference Optimization and Acceleration Techniques focuses on runtime strategies such as adaptive block sizing, efficient caching mechanisms (e.g., Attention KV Cache[7], dCache[15]), and variable-length generation schemes (Variable Generation Lengths[16]); Domain-Specific Applications and Extensions examines how block diffusion extends to multimodal settings (DiffusionVL[20], Audio-Language Joint[22]) and specialized tasks; and Surveys and Comparative Studies (Diffusion LLM Survey[4]) provide overarching perspectives on the trade-offs between diffusion and autoregressive paradigms.

Within the architecture and training branch, a particularly active line of work addresses autoregressive-to-diffusion conversion and adaptation. Fast-dLLM v2[0] sits squarely in this cluster, proposing methods to efficiently transform existing autoregressive checkpoints into block diffusion models without full retraining. Nearby efforts such as LLaDA[10] and Next-Block Adaptation[11] similarly tackle the challenge of adapting pretrained weights to predict token blocks rather than single tokens, while Efficient-DLM[14] emphasizes computational efficiency during the conversion process. These works share a common goal of leveraging the vast investment in autoregressive pretraining while unlocking the parallelism benefits of diffusion inference. The main open questions revolve around how much fine-tuning is necessary, whether certain architectural modifications (e.g., Block Transformer[12]) improve block-level coherence, and how to balance the speed gains from parallel decoding against potential quality degradation compared to standard autoregressive baselines.

Related Works in Same Category

The following **3 sibling papers** share the same taxonomy leaf node with the original paper:

1. LLaDA2. 0: Scaling Up Diffusion Language Models to 100B

Authors: Tiwei Bie, Maosong Cao, Kun Chen, Lun Du, Mingliang Gong, et al. (31 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

This paper presents LLaDA2.0 -- a tuple of discrete diffusion large language models (dLLM) scaling up to 100B total parameters through systematic conversion from auto-regressive (AR) models -- establishing a new paradigm for frontier-scale deployment. Instead of costly training from scratch, LLaDA2.0 upholds knowledge inheritance, progressive adaption and efficiency-aware design principle, and seamless converts a pre-trained AR model into dLLM with a novel 3-phase block-level WSD based training ...

Relationship Analysis

Both papers belong to the Autoregressive-to-Diffusion Conversion and Adaptation category, focusing on converting pretrained AR models into diffusion language models. They overlap in their core approach of adapting existing AR models rather than training from scratch, and both employ block-based diffusion mechanisms to preserve AR model knowledge while enabling parallel generation. The key difference is that Fast-dLLM v2 emphasizes data efficiency with only ~1B tokens of fine-tuning and introduces hierarchical caching mechanisms for inference acceleration, while LLaDA2.0 focuses on scaling to 100B parameters through a systematic three-phase Warmup-Stable-Decay training strategy and document-level attention masking for handling packed heterogeneous documents.

---

2. From Next-Token to Next-Block: A Principled Adaptation Path for Diffusion LLMs

Authors: Yuchuan Tian, Yuchen Liang, Jiacheng Sun, Shuo Zhang, Guangwen Yang, et al. (13 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Large language models (LLMs) excel at generation but dominant autoregressive (AR) decoding is inherently sequential, creating a throughput bottleneck. Diffusion Language Models (DLMs)--especially block-wise variants--enable parallel generation and intra-block bidirectional reasoning, yet training large DLMs from scratch is costly and wastes the knowledge in mature AR checkpoints. Prior "adaptation" attempts either modify logits or randomly grow attention masks to full-sequence diffusion, or simp...

Relationship Analysis

Both papers belong to the Autoregressive-to-Diffusion Conversion and Adaptation category, focusing on efficiently converting pretrained AR models into block diffusion language models. They overlap in their core approach of using block-wise diffusion mechanisms with hierarchical caching and parallel decoding strategies to accelerate inference while preserving AR model quality. The key differences are: Fast-dLLM v2 emphasizes data efficiency (requiring only ~1B tokens vs. 500B for full-attention diffusion) through a complementary attention mask design and hierarchical caching (block-level and sub-block DualCache), while NBDiff introduces a gradual block-size growth curriculum starting from blocksize=1 (pure AR) and uses a context-causal attention mask with auxiliary AR loss to maintain strict causality in decoded context while enabling bidirectional attention only in the active block.

---

3. Efficient-DLM: From Autoregressive to Diffusion Language Models, and Beyond in Speed

Authors: Yonggan Fu, Lexington Whalen, Zhifan Ye, Xin Dong, Shizhe Diao, et al. (14 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Diffusion language models (dLMs) have emerged as a promising paradigm that enables parallel, non-autoregressive generation, but their learning efficiency lags behind that of autoregressive (AR) language models when trained from scratch. To this end, we study AR-

to-dLM conversion to transform pretrained AR models into efficient dLMs that excel in speed while preserving AR models' task accuracy. We achieve this by identifying limitations in the attention patterns and objectives of existing AR-to-d...

Relationship Analysis

Both papers belong to the Autoregressive-to-Diffusion Conversion and Adaptation category, focusing on efficiently transforming pretrained AR models into block diffusion language models. They share overlapping approaches including block-wise attention patterns, hierarchical caching mechanisms, and data-efficient fine-tuning strategies requiring minimal tokens (~1B for Fast-dLLM v2 vs ~10-100B for Efficient-DLM). The key differences are that Fast-dLLM v2 emphasizes complementary masking and sub-block decoding with DualCache for 2.5× speedup, while Efficient-DLM focuses on position-dependent token masking strategies and systematic analysis of attention patterns to achieve 4.5× speedup with different architectural choices.

Contributions Analysis

Overall novelty summary. ```json { "paragraphs": ["The paper proposes Fast-dLLM v2, a method for converting pretrained autoregressive models into block diffusion language models using approximately 1B tokens of fine-tuning. It resides in the 'Autoregressive-to-Diffusion Conversion and Adaptation' leaf, which contains four papers total including the original work. This leaf sits within the broader 'Block Diffusion Architecture and Training Methods' branch, indicating a moderately populated research direction focused specifically on efficient AR-to-diffusion conversion rather than training diffusion models from scratch.",

"The taxonomy reveals neighboring research directions including 'Novel Architecture Design for Block Diffusion' (five papers exploring fundamentally new architectures) and 'Variable-Length and Adaptive Block Generation' (four papers on dynamic block sizing). The paper's leaf is distinguished by its focus on knowledge inheritance from pretrained models rather than architectural novelty. Adjacent branches cover 'Inference Optimization and Acceleration Techniques' with specialized work on KV cache optimization and controllability, suggesting the paper bridges architectural adaptation with inference acceleration concerns through its hierarchical caching mechanism.",

"Among three contributions analyzed from 20 candidate papers examined, the data-efficient post-training strategy shows substantial prior work: 9 candidates examined, 6 potentially refutable. The hierarchical caching mechanism appears more novel with only 1 candidate examined and none refutable. The speedup validation examined 10 candidates with 3 potentially refutable. This limited search scope suggests the conversion strategy operates in a crowded space alongside works like LLaDA and Next-Block Adaptation, while the specific caching design may represent a less explored technical direction within the broader field.",

"Based on this top-20 semantic search, the work appears to make incremental contributions to AR-to-diffusion conversion methodology, situated in a moderately active research area. The most distinctive element may be the hierarchical caching design, though the limited candidate pool prevents definitive assessment. The analysis does not cover exhaustive citation networks or recent unpublished work that might reveal additional overlaps."] } ```

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

Contribution 1: Data-efficient post-training strategy for adapting AR models to block-diffusion frameworks

Description: The authors propose a method to convert pretrained autoregressive language models into block diffusion models using only approximately 1 billion tokens of fine-tuning, which is 500 times less data than full-attention diffusion models like Dream that require around 500 billion tokens. This is achieved through a novel training recipe combining block diffusion with complementary attention masking.

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. DiRL: An Efficient Post-Training Framework for Diffusion Language Models

URL: View paper

Prior Art Analysis

DiRL[41] demonstrates that similar post-training adaptation of autoregressive models to block-diffusion frameworks was achieved prior to the original paper's work. DiRL[41] presents a post-training framework that adapts pretrained autoregressive models into blockwise diffusion models, requiring only approximately 1-3 billion tokens for fine-tuning. This directly challenges the novelty claim of the original paper, which presents a method requiring ~1 billion tokens as a novel contribution. Both papers employ block diffusion mechanisms with complementary attention masking strategies and achieve similar data efficiency goals through post-training rather than full retraining.

Evidence

Evidence 1 - **Rationale**: Both papers present post-training frameworks for adapting AR models to block-diffusion architectures with similar data efficiency goals. - **Original**: we propose fast-dllm v2, a carefully designed block diffusion language model (dllm) that efficiently adapts pretrained ar models into dllms for parallel text generation-requiring only ~1b tokens of fine-tuning. this represents a 500x reduction in training data compared to full-attention diffusion ll... - **Candidate**: we introduce dirl, an efficient post-training framework that tightly integrates flexattention-accelerated blockwise training with lmdeploy-optimized inference. this architecture enables a streamlined online model update loop, facilitating efficient two-stage post-training (supervised fine-tuning fol...

Evidence 2 - **Rationale**: Both papers describe post-training adaptation of pretrained AR models through supervised fine-tuning using blockwise diffusion setups. - **Original**: we build our block-wise diffusion training pipeline on top of pretrained qwen2.5-instruct models (qwen et al., 2025), including both 1.5b and 7b variants. fine-tuning is conducted as supervised fine-tuning (sft) on instruction-tuning data, where each training batch is constructed using our blockwise... - **Candidate**: sft is conducted with llama-factory (zheng et al., 2024). 8 xh200 gpus are applied to fine-tune with sdar-8b-chat with deepspeed zero1 (rajbhandari et al., 2020). models are fine-tuned with a maximum length of 8k tokens.

---

2. Sdar: A synergistic diffusion-autoregression paradigm for scalable sequence generation

URL: View paper

Prior Art Analysis

SDAR[38] demonstrates that autoregressive models can be adapted to block diffusion frameworks using significantly less data than the original paper claims is novel. While the original paper states their method requires ~1B tokens (500x less than Dream's 500B), SDAR[38] achieves similar adaptation with 50B tokens and explicitly describes this as a 'lightweight paradigm conversion' and 'highly efficient adaptation procedure'. Both papers convert pretrained AR models to block-wise diffusion models through continued training rather than training from scratch, and both emphasize data efficiency as a key contribution. SDAR[38] was published concurrently and demonstrates the same core concept of efficient AR-to-block-diffusion adaptation.

Evidence

Evidence 1 - **Rationale**: Both papers claim as a key contribution the efficient adaptation of AR models to block diffusion. The original paper emphasizes 1B tokens vs Dream's 500B, while SDAR uses 50B tokens and describes it as 'lightweight adaptation', demonstrating that the concept of data-efficient AR-to-block-diffusion conversion was known concurrently. - **Original**: a key feature of fast-dllm v2 is its data efficiency: while full-attention diffusion models such as dream (ye et al., 2025a) require on the order of 500b tokens for fine-tuning,

our method adapts ar models into block diffusion models with only about 1b tokens of fine-tuning-achieving lossless adapta... - **Candidate**: building on these insights, we design sdar to reconcile the efficiency of autoregressive training with the parallelism of diffusion-based inference. the key principle isdecouplingthe two phases: we leverage full-scale ar pretraining to ensure stability and efficiency, and then introduce a lightweigh...

Evidence 2 - **Rationale**: Both papers explicitly claim as a contribution a post-training/adaptation strategy that converts AR models to block diffusion with minimal data. SDAR describes this as 'lightweight paradigm conversion' and 'data-efficient adaptation', directly paralleling the original paper's claimed novelty. - **Original**: we identify the ar-friendly nature of our block-wise attention design and leverage it to present a post-training strategy for adapting pretrained ar models into block-diffusion frameworks, requiring only affordable fine-tuning rather than full retraining. specifically, fast-dllm v2 achieves lossless... - **Candidate**: instead of costly end-to-end diffusion training, sdar performs a lightweight paradigm conversion that transforms a well-trained autoregressive (ar) model into a blockwise diffusion model through brief, data-efficient adaptation.

Evidence 3 - **Rationale**: Both papers describe converting AR models to block diffusion through continued training with modified objectives and attention mechanisms. SDAR explicitly states this 'avoids the prohibitive computational cost of training a block-wise diffusion model from scratch', which is the same efficiency claim made by the original paper. - **Original**: our approach introduces a novel training recipe that combines a block diffusion mechanism with a complementary attention mask, enabling blockwise bidirectional context modeling without sacrificing ar training objectives. - **Candidate**: starting from a sufficiently capable ar base model θar via conventional ntp pretraining, we subsequently continue training the model to convert the language modeling paradigm from ar to block-wise diffusion. this strategy avoids the prohibitive computational cost of training a block-wise diffusion m...

Evidence 4 - **Rationale**: Both papers position their work as making block diffusion practical and accessible through data-efficient adaptation. SDAR explicitly compares to Dream (the same baseline used by the original paper) and claims superior data efficiency, demonstrating that this approach was known concurrently. - **Original**: unlike prior block diffusion approaches that remain limited to small-scale validation, fast-dllm v2 is explicitly built to scale to large llms and real-world tasks. - **Candidate**: we argue that this adaptation requires significantly less data than adaptation method such as dream, making it broadly accessible to the community.

---

3. Efficient-DLM: From Autoregressive to Diffusion Language Models, and Beyond in Speed

URL: View paper

Prior Art Analysis

Efficient-DLM[14] demonstrates that pretrained autoregressive models can be converted into block diffusion models using approximately 10 billion tokens for initial conversion and extended training up to 300-500 billion tokens for improved performance. This directly refutes the original paper's claim of being the first to achieve such adaptation with only ~1 billion tokens. The candidate paper presents a systematic framework for AR-to-DLM conversion that includes block-wise attention patterns with clean context conditioning, which is fundamentally similar to the original paper's approach. Both papers employ complementary masking strategies and block-level caching mechanisms, indicating that the core methodology was already established in prior work.

Evidence

Evidence 1 - **Rationale**: This pair demonstrates that Efficient-DLM[14] already established the concept of converting pretrained AR models to block diffusion models with specific training token requirements (10b-100b tokens), which challenges the original paper's novelty claim of achieving this with only ~1b tokens. - **Original**: fast-dllm v2, a carefully designed block diffusion language model (dllm) that efficiently adapts pretrained ar models into dllms for parallel text generation-requiring only ~1b tokens of fine-tuning. this represents a 500x reduction in training data compared to full-attention diffusion llms such as ... - **Candidate**: this work leverages pretrained ar models for initialization and systematically explores how to continuously pretrain them into dlms that achieve high generation speed while preserving task accuracy. the key insight is that, with an appropriate training scheme in terms of attention patterns and objec...

Evidence 2 - **Rationale**: Both papers describe the same core mechanism of block-wise attention with clean context conditioning. The candidate paper explicitly describes this approach, indicating it was already established in the field before the original paper's submission. - **Original**: our approach introduces a novel training recipe that combines a block diffusion mechanism with a complementary attention mask, enabling blockwise bidirectional context modeling without sacrificing ar training objectives. - **Candidate**: block-wise attention with each block conditioned on clean context.the drawback of the block-wise attention in fig. 2 (c) is that it may cause a training-test gap: when performing block-wise decoding at test time, the context preceding a noisy block has already been decoded without mask tokens; howev...

Evidence 3 - **Rationale**: This evidence shows that Efficient-DLM[14] already demonstrated AR-to-DLM conversion with specific training token budgets (50b tokens in this example), establishing the methodology before the original paper's claimed innovation of using only 1b tokens. - **Original**: we identify the ar-friendly nature of our block-wise attention design and leverage it to present a post-training strategy for adapting pretrained ar models into block-diffusion frameworks, requiring only affordable fine-tuning rather than full retraining. specifically, fast-dllm v2 achieves lossless... - **Candidate**: settings.we adopt qwen2.5 1.5b [11] as the ar initialization and perform continuous pretraining for 50b tokens on a mixed dataset comprising [12, 13, 14]. for block-wise training, we adopt a block size of 16, and provide further analysis on block sizes in sec. 2.3. the initial learning rate is set t...

---

4. Blockwise sft for diffusion language models: Reconciling bidirectional attention and autoregressive decoding

URL: View paper

Brief Assessment

Blockwise SFT[40] focuses on aligning supervised fine-tuning with semi-autoregressive blockwise decoding through masking strategies, not on post-training adaptation of pretrained AR models to diffusion frameworks. The candidate addresses training-inference mismatch in already-diffusion models, while the original contribution is about converting AR models into block diffusion models with minimal data.

---

5. LLaDA2. 0: Scaling Up Diffusion Language Models to 100B

URL: View paper

Prior Art Analysis

LLaDA[10] demonstrates that autoregressive models can be converted to block diffusion models through a systematic continual pre-training approach, requiring only moderate amounts of data. The candidate paper presents a 'warmup-stable-decay' (WSD) strategy that progressively adapts AR models to diffusion frameworks through three coordinated phases, achieving efficient conversion with approximately 1-3 billion tokens. This directly challenges the novelty claim by showing that similar AR-to-diffusion adaptation with comparable data efficiency was already established in prior work.

Evidence

Evidence 1 - **Rationale**: Both papers describe converting pretrained AR models to diffusion models through systematic training strategies, establishing that this approach was already explored in LLaDA[10]. - **Original**: we propose fast-dllm v2, a carefully designed block diffusion language model (dllm) that efficiently adapts pretrained ar models into dllms for parallel text generation-requiring only ~1b tokens of fine-tuning. this represents a 500x reduction in training data compared to full-attention diffusion ll... - **Candidate**: we introduce llada2.0 series with 100b/16b total parameters diffusion language models that resolves these fundamental challenges through a

novel two-stage continual pre-training (cpt) paradigm. rather than attempting to train diffusion models from scratch, we leverage existing ar checkpoints as the ...

Evidence 2 - **Rationale**: Both papers present systematic training recipes for AR-to-block-diffusion conversion. LLaDA[10]'s WSD strategy demonstrates a prior approach to progressive adaptation, refuting the claim of novelty in the training methodology. - **Original**: our approach introduces a novel training recipe that combines a block diffusion mechanism with a complementary attention mask, enabling blockwise bidirectional context modeling without sacrificing ar training objectives. - **Candidate**: to address this gap, we propose awarmup-stable-decay(wsd) continual pre-training strategy that enables a smooth, stable, and effective transition from ar to dllm. wsd decomposes the conversion into three coordinated phases: • warmup: progressively increase the block size in block diffusion language ...

---

6. Diffusion-based Large Language Models Survey

URL: View paper

Brief Assessment

Diffusion LLM Survey[4] provides only fragmentary mentions of block diffusion and hybrid paradigms without discussing post-training adaptation strategies or data efficiency metrics for converting AR models to block-diffusion frameworks.

---

7. DiffusionVL: Translating Any Autoregressive Models into Diffusion Vision Language Models

URL: View paper

Prior Art Analysis

DiffusionVL[20] demonstrates that autoregressive vision language models can be adapted to block diffusion frameworks using only 738k samples (less than 1 billion tokens), which is comparable to or even more data-efficient than the original paper's claimed 1 billion tokens. The candidate paper shows this adaptation is effective without requiring the 500x data reduction claim to be novel, as they achieve state-of-the-art performance with minimal data. Both papers employ block-wise attention mechanisms and demonstrate efficient post-training adaptation strategies, with DiffusionVL[20] explicitly stating their approach requires 'less than 5% of the data required by prior methods' and successfully converts AR models to diffusion models through simple finetuning.

Evidence

Evidence 1 - **Rationale**: Both papers describe post-training strategies that adapt AR models to block diffusion through efficient finetuning rather than full retraining, challenging the novelty of this approach. - **Original**: we identify the ar-friendly nature of our block-wise attention design and leverage it to present a post-training strategy for adapting pretrained ar models into block-diffusion frameworks, requiring only affordable fine-tuning rather than full retraining - **Candidate**: we proposediffusionvl, a dvlm family that could be translated from any powerful ar models. through simple finetuning, we successfully adapt ar pre-trained models into the diffusion paradigm

Evidence 2 - **Rationale**: DiffusionVL[20] employs similar block-wise attention mechanisms with bidirectional intra-block and causal inter-block attention, demonstrating that this training recipe was not uniquely novel to the original paper. - **Original**: our approach introduces a novel training recipe that combines a block diffusion mechanism with a complementary attention mask, enabling blockwise bidirectional context modeling without sacrificing ar training objectives - **Candidate**: regarding the attention mechanism, we adopt the hybrid attention pattern like [1]: for each input embedding sequence, the noised sequence and original clean embedding are concatenated along the sequence dimension. a specialized attention mask is constructed to enforce bidirectional attention within ...

Evidence 3 - **Rationale**: The original paper's actual training uses 1.31-3.15 billion tokens, while DiffusionVL[20] achieves comparable adaptation with 738k samples (significantly less data), demonstrating that the claimed data efficiency may not be as novel as stated. - **Original**: for the 1.5b model: 6,000 x 524,288 ≈ 3.15 billion tokens • 7b model: 2,500 x 524,288 ≈ 1.31 billion tokens - **Candidate**: for the pretraining stage of building dvlms from arvlms, we adopt the 580k-sample pretraining dataset from llav a-pretrain [23]. for finetuning data used in building dvlms, we uniformly use the open-source 738k data instruction-follow samples from llav a-next [17]

---

8. ACDiT: Interpolating Autoregressive Conditional Modeling and Diffusion Transformer

URL: View paper

Brief Assessment

ACDiT[39] focuses on visual generation (images/videos) using continuous diffusion transformers, not language model adaptation. The candidate does not address post-training adaptation of pretrained autoregressive language models or demonstrate data efficiency in token-level fine-tuning.

---

9. From Next-Token to Next-Block: A Principled Adaptation Path for Diffusion LLMs

URL: View paper

Prior Art Analysis

Next-Block Adaptation[11] demonstrates that similar data-efficient adaptation from AR to block-diffusion was achieved prior to the ORIGINAL paper. Both papers adapt pretrained AR models to block-diffusion using approximately 1 billion tokens of fine-tuning data. Next-Block Adaptation[11] explicitly states training with '700b tokens' for pretraining adaptation plus '100b tokens' for long-sequence extension, totaling approximately 800 billion tokens across all phases, but the core adaptation phase uses significantly less. The candidate paper describes a systematic framework with context-causal attention masks, parallel training with auxiliary AR loss, and gradual block-size growth - all targeting the same goal of efficient AR-to-block-diffusion adaptation that the ORIGINAL claims as novel.

Evidence

Evidence 1 - **Rationale**: Both papers claim data-efficient adaptation of AR models to block-diffusion. While the ORIGINAL claims '~1b tokens', the candidate uses '700b tokens' for the main adaptation phase, suggesting the ORIGINAL's novelty claim of being the first to achieve efficient adaptation with minimal data may be challenged. - **Original**: we propose fast-dllm v2, a carefully designed block diffusion language model (dllm) that efficiently adapts pretrained ar models into dllms for parallel text generation-requiring only ~1b tokens of fine-tuning. this represents a 500x reduction in training data compared to full-attention diffusion ll... - **Candidate**: the pretraining adaptation stage uses a two-phase learning-rate schedule: we keep the learning rateconstantfor the first 24,000 iterations and then apply alearning-rate cooldownover the final 60,000 iterations, for a total of 84,000 iterations. we train with sequence length ℓ=8k and global batch siz...

Evidence 2 - **Rationale**: Next-Block Adaptation[11] presents a comprehensive post-training strategy for AR-to-block-diffusion adaptation with specific technical components (context-causal masks, parallel training, auxiliary AR loss, gradual block growth), demonstrating that such adaptation frameworks existed prior to the ORIGINAL paper's submission. - **Original**: we identify the ar-friendly nature of our block-wise attention design and leverage it to present a post-training strategy for adapting pretrained ar models into block-diffusion frameworks, requiring only affordable fine-tuning rather than full retraining. - **Candidate**: we reframe adaptation as a intra-paradigm path from ar to blockdiffusion by viewing ar as block-diffusion with blocksize= 1 . concretely, we design the pathway of adaptation as follows: we use a context-causal attention mask (causal in context, bidirectional only within the active block), an efficie...

Evidence 3 - **Rationale**: Both papers describe training recipes that preserve AR objectives while enabling block-diffusion. Next-Block Adaptation[11]'s auxiliary AR loss and parallel training strategy serve the same purpose as the ORIGINAL's 'complementary attention mask' approach - maintaining AR training objectives during adaptation. - **Original**: our approach introduces a novel training recipe that combines a block diffusion mechanism with a complementary attention mask, enabling blockwise bidirectional context modeling without

sacrificing ar training objectives. - **Candidate**: we develop an efficient parallel training strategy that aligns with inference and incorporates an auxiliary ar loss, markedly improving convergence speed and knowledge retention. we develop a gradual block growth approach that alleviates the gap between ar and block-diffusion models, improving adapt...

Evidence 4 - **Rationale**: Next-Block Adaptation[11] explicitly addresses the same problem of avoiding expensive training from scratch by adapting existing AR checkpoints, demonstrating that the concept of data-efficient AR-to-diffusion adaptation was already established in prior work. - **Original**: fast-dllm v2 achieves lossless adaptation with just ~1b tokens, compared to ~500b tokens required by dream (ye et al., 2025a). - **Candidate**: training large-scale dlms from scratch is computationally prohibitive and discards the vast knowledge already encoded in mature, open-source ar checkpoints. among diffusion paradigms, masked diffusion that trains a model to denoise masked tokens is uniquely suited for adaptation, as it shares the st...

Contribution 2: Hierarchical caching mechanism with block-level and sub-block caches

Description: The authors design a two-level caching system: a block-level cache that stores historical context representations across blocks, and a sub-block cache (DualCache) that enables efficient parallel generation within partially decoded blocks. This hierarchical approach substantially accelerates inference compared to prior diffusion methods.

This contribution was assessed against **1 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. SABlock: Semantic-Aware KV Cache Eviction with Adaptive Compression Block Size

URL: View paper

Brief Assessment

SABlock[27] focuses on KV cache eviction with semantic-aware compression for memory efficiency in long-context inference, not on hierarchical caching for parallel decoding in diffusion models.

Contribution 3: Comprehensive large-scale validation achieving 2.5× speedup over AR decoding

Description: The authors perform extensive experiments on models up to 7 billion parameters across diverse benchmarks, demonstrating that their approach achieves up to 2.5 times faster inference than standard autoregressive decoding while maintaining generation quality comparable to strong autoregressive baselines.

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. Tidar: Think in diffusion, talk in autoregression

URL: View paper

Prior Art Analysis

Tidar[37] demonstrates that achieving significant speedup over autoregressive decoding while maintaining quality was accomplished prior to the original paper's work. Tidar[37] reports 4.71× to 5.91× speedup over standard AR decoding across models at 1.5B and 8B parameters, which substantially exceeds the 2.5× speedup claimed in the original paper. Both papers perform extensive experiments on billion-parameter models across diverse benchmarks including coding, math, and reasoning tasks. The candidate paper's results were achieved through a hybrid diffusion-autoregressive architecture that enables parallel token generation, demonstrating that the claimed speedup achievement in the original paper was not novel.

Evidence

Evidence 1 - **Rationale**: This pair demonstrates that Tidar[37] achieved substantially higher speedup (4.71×-5.91×) compared to the original paper's 2.5× speedup, both validated on large-scale models across diverse benchmarks, refuting the novelty of the speedup achievement. - **Original**: fast-dllm v2 achieves up to 2.5x speedup over standard ar decoding without compromising generation quality. extensive experiments across diverse benchmarks demonstrate that fast-dllm v2 matches or surpasses ar baselines in accuracy - **Candidate**: we extensively evaluate tidar against ar models, speculative decoding, and diffusion variants across generative and likelihood tasks at 1.5b and 8b scales. thanks to the parallel drafting and sampling as well as exact kv cache support, tidar outperforms speculative decoding in measured throughput an...

Evidence 2 - **Rationale**: Both papers claim comprehensive validation on billion-parameter models, but Tidar[37] demonstrates higher speedup ratios (4.71×-5.91×) versus the original's 2.5×, indicating prior achievement of similar or superior results. - **Original**: we conduct comprehensive large-scale experiments on models up to 7b parameters and diverse tasks, showing that fast-dllm v2 achieves up to 2.5x speedup over standard ar decoding while maintaining comparable generation quality. - **Candidate**: we show that for tidar 1.5b, we can achieve lossless quality compared to its ar counterpart while generating with 4.71xrelative throughput (tokens per second) speedup. for tidar 8b, we achieved an impressive 5.91xrelative throughput speedup with minimal loss.

Evidence 3 - **Rationale**: Both papers validate their approaches across diverse benchmarks with models maintaining quality comparable to AR baselines, but Tidar[37] demonstrates this with higher efficiency gains, refuting the novelty of the comprehensive validation claim. - **Original**: extensive experiments across diverse benchmarks demonstrate that fast-dllm v2 matches or surpasses ar baselines in accuracy, while delivering state-of-the-art efficiency among dllms-marking a significant step toward the practical deployment of fast and accurate llms. - **Candidate**: in table 2, we first compare tidar against several ar models and also a popular diffusion variant, block diffusion, across two model sizes. for 1.5b-1.7b size range, tidar is highly competitive in terms of quality with an average 7.45 tokens per model forward (nfe). for 8b models, tidar incurs very ...

---

2. Accelerated Diffusion Models via Speculative Sampling

URL: View paper

Brief Assessment

Speculative Sampling[36] focuses on accelerating diffusion models through speculative sampling techniques, not on block diffusion language models or autoregressive decoding acceleration. The candidate addresses continuous diffusion models while the original work targets discrete language model inference.

---

3. Lavida: A large diffusion language model for multimodal understanding

URL: View paper

Brief Assessment

Lavida[35] focuses on multimodal vision-language understanding tasks using diffusion models, not on accelerating general language model inference. The speedup claims in Lavida are specific to vision-language captioning tasks with different architectural choices (prefix-DLM caching for visual tokens), whereas the original paper addresses block-diffusion LLMs for text generation.

---

4. Speculative diffusion decoding: Accelerating language generation through diffusion

URL: View paper

Brief Assessment

Speculative Diffusion[33] focuses on speculative decoding using diffusion models as draft models for autoregressive verification, achieving up to 7.2x speedups. The original paper presents a block diffusion framework that directly transforms AR models into parallel decoders, representing fundamentally different architectural approaches and speedup mechanisms.

---

5. Accelerating Diffusion LLMs via Adaptive Parallel Decoding

URL: View paper

Brief Assessment

Adaptive Parallel Decoding[30] focuses on accelerating diffusion LLMs through adaptive parallel token generation with a multiplicative mixture approach, while the original paper achieves speedup through block-wise diffusion with hierarchical caching mechanisms. The technical approaches and architectural designs differ fundamentally.

---

6. Diffuspec: Unlocking diffusion language models for speculative decoding

URL: View paper

Brief Assessment

Diffuspec[32] focuses on speculative decoding using diffusion models as drafters for autoregressive verifiers, achieving 3× speedup. The original paper presents a block diffusion framework with hierarchical caching for direct parallel generation, not speculative decoding.

---

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

URL: View paper

Brief Assessment

Creditdecoding[34] focuses on accelerating diffusion LLMs through trace credit mechanisms for parallel decoding, not on adapting AR models to block diffusion frameworks. The speedup mechanisms and architectural approaches differ fundamentally from the original paper's block diffusion adaptation strategy.

---

8. Dimple: Discrete Diffusion Multimodal Large Language Model with Parallel Decoding

URL: View paper

Brief Assessment

Dimple[31] focuses on multimodal language models with a different decoding strategy (confident decoding that dynamically adjusts tokens per step) rather than the block diffusion approach with hierarchical caching used in the original paper. The speedup mechanisms and model architectures are fundamentally different.

---

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

URL: View paper

Prior Art Analysis

KV Caching Guided[29] demonstrates that diffusion language models can achieve substantial speedup over autoregressive decoding through training-free acceleration techniques. The candidate paper reports achieving 12.14x average speedup on dream-7b-instruct and 13.29x on llada-8b-instruct models, which significantly exceeds the 2.5x speedup claimed in the original paper. Furthermore, the candidate validates their approach on models up to 8B parameters across diverse benchmarks including GSM8K, MMLU-Pro, PIQA, ARC-C, ARC-E, and GPQA, demonstrating that accelerating diffusion models to achieve speedup over AR decoding was already established prior to the original paper's submission.

Evidence

Evidence 1 - **Rationale**: Both papers claim to be the first to demonstrate that diffusion models can achieve competitive or superior speed compared to autoregressive models through extensive benchmarking. The candidate's claim of 'for the first time' achieving comparable latency directly challenges the original's novelty claim. - **Original**: extensive experiments across diverse benchmarks demonstrate that fast-dllm v2 matches or surpasses ar baselines in accuracy, while delivering state-of-the-art efficiency among dllms-marking a significant step toward the practical deployment of fast and accurate llms - **Candidate**: we conduct extensive evaluations on open-source reasoning benchmarks, and our combined methods deliver an average of12.14 x end-to-end speedup across various tasks with negligible accuracy degradation. for the first time, diffusion language models achieve a comparable and even faster latency as the ...

Evidence 2 - **Rationale**: The candidate paper demonstrates significantly higher speedup (12.48x average) on the same dream-7b model that the original paper uses, with validation across multiple benchmarks including MMLU-Pro and GPQA. This shows that achieving speedup over AR decoding on large-scale diffusion models was already demonstrated. - **Original**: we conduct comprehensive large-scale experiments on models up to 7b parameters and diverse tasks, showing that fast-dllm v2 achieves up to 2.5x speedup over standard ar decoding while maintaining comparable generation quality - **Candidate**: compared to the vanilla dream-7b-instruct model, the proposed method achieves consistent and outstanding speedup across all the tasks from different domains. for the complex reasoning tasks (e.g., mmlu-pro, gpqa), the proposed freecache first induces 3.11x speedup with minimal quality degradation. r...

---

10. Fast-dllm: Training-free acceleration of diffusion llm by enabling kv cache and parallel decoding

URL: View paper

Prior Art Analysis

Fast-dllm[28] demonstrates that similar comprehensive large-scale validation with speedup claims over autoregressive decoding was already achieved prior to the original paper. The candidate paper reports extensive experiments on models up to 7-8 billion parameters across the same benchmarks (gsm8k, math, humaneval, mbpp), achieving up to 27.6x speedup in certain configurations and consistently demonstrating 5-8x speedups with minimal accuracy loss. The candidate explicitly states achieving 'up to 2.5x speedup over standard ar decoding' in multiple locations, and provides detailed experimental validation across diverse tasks at the 7B scale, directly overlapping with the original paper's claimed contribution of '2.5x speedup' on '7 billion parameters across diverse benchmarks.'

Evidence

Evidence 1 - **Rationale**: Both papers claim comprehensive experimental validation on large-scale models with significant speedup over autoregressive baselines. The candidate demonstrates even higher speedups (27.6x) than the original's claimed 2.5x, suggesting prior work already achieved and exceeded this milestone. - **Original**: we conduct comprehensive large-scale experiments on models up to 7b parameters and diverse tasks, showing that fast-dllm v2 achieves up to 2.5x speedup over standard ar decoding while maintaining comparable generation quality. - **Candidate**: 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.

Evidence 2 - **Rationale**: Both papers describe comprehensive experiments across the same benchmark suite (gsm8k, math, humaneval, mbpp) with similar claims about maintaining accuracy while achieving significant speedups, demonstrating that this type of validation was already performed in the candidate work. - **Original**: extensive experiments across diverse benchmarks demonstrate that fast-dllm v2 matches or surpasses ar baselines in accuracy, while delivering state-of-the-art efficiency among dllms-marking a significant step toward the practical deployment of fast and accurate llms. - **Candidate**: 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, confirming the...

Evidence 3 - **Rationale**: The candidate paper provides detailed speedup measurements on the same models and benchmarks, demonstrating that comprehensive validation of speedup claims over AR baselines was already established in prior work, with even more extensive results than the original's 2.5x claim. - **Original**: fast-dllm v2 achieves up to 2.5x speedup over standard ar decoding without compromising generation quality. extensive experiments across diverse benchmarks demonstrate that fast-dllm v2 matches or surpasses ar baselines in accuracy - **Candidate**: on llada, for example, combined kv cache and parallel decoding methods boost throughput by up to 11x(gsm8k, length 512) and 9.2x (mbpp, length 512) over the standard baseline. similarly, on dream-base, the largest throughput gains are observed on mbpp (7.8xat length 512) and gsm8k (5.6xat length 512...

Appendix: Text Similarity Detection

Textual similarity detection checked 20 papers and found 6 similarity segment(s) across 4 paper(s).

The following **4 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. DiffusionVL: Translating Any Autoregressive Models into Diffusion Vision Language Models

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.

2. Lavida: A large diffusion language model for multimodal understanding

Detected in: Contribution: contribution_3

⚠ **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.

3. From Next-Token to Next-Block: A Principled Adaptation Path for Diffusion LLMs

Detected in: Core Task (sibling), 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.

4. DiRL: An Efficient Post-Training Framework for Diffusion Language Models

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 v2: Efficient Block-Diffusion LLM View paper
- [1] Accelerating Diffusion Large Language Models with SlowFast: The Three Golden Principles View paper
- [2] Set block decoding is a language model inference accelerator View paper
- [3] Ctrldiff: Boosting large diffusion language models with dynamic block prediction and controllable generation View paper
- [4] Diffusion-based Large Language Models Survey View paper
- [5] Sequential diffusion language models View paper
- [6] Encoder-Decoder Diffusion Language Models for Efficient Training and Inference View paper
- [7] Attention Is All You Need for KV Cache in Diffusion LLMs View paper
- [8] AdaBlock-dLLM: Semantic-Aware Diffusion LLM Inference via Adaptive Block Size View paper
- [9] Inferix: A Block-Diffusion based Next-Generation Inference Engine for World Simulation View paper
- [10] LLaDA2. 0: Scaling Up Diffusion Language Models to 100B View paper
- [11] From Next-Token to Next-Block: A Principled Adaptation Path for Diffusion LLMs View paper
- [12] Block transformer: Global-to-local language modeling for fast inference View paper
- [13] Beyond Next-Token Prediction: A Performance Characterization of Diffusion versus Autoregressive Language Models View paper
- [14] Efficient-DLM: From Autoregressive to Diffusion Language Models, and Beyond in Speed View paper
- [15] dCache: Accelerating Diffusion-Based LLMs via Dual Adaptive Caching View paper
- [16] Diffusion llm with native variable generation lengths: Let lead the way View paper
- [17] Efficient Inference in Large Language Models View paper
- [18] Tandem Transformers for Inference Efficient LLMs View paper
- [19] DINGO: Constrained Inference for Diffusion LLMs View paper
- [20] DiffusionVL: Translating Any Autoregressive Models into Diffusion Vision Language Models View paper
- [21] Towards Effective and Efficient Non-autoregressive decoders for Conformer and LLM-based ASR using Block-based Attention Mask View paper
- [22] From Text to Talk: Audio-Language Model Needs Non-Autoregressive Joint Training View paper
- [23] Block Diffusion: Interpolating Between Autoregressive and Diffusion Language Models View paper
- [24] ReFusion: A Diffusion Large Language Model with Parallel Autoregressive Decoding View paper
- [25] Diffusion LLM with Native Variable Generation Lengths: Let [EOS] Lead the Way View paper
- [26] Diffusion Beats ARM: Diffusion Large Language Models for Generative Recommendation View paper
- [27] SABlock: Semantic-Aware KV Cache Eviction with Adaptive Compression Block Size View paper
- [28] Fast-dllm: Training-free acceleration of diffusion llm by enabling kv cache and parallel decoding View paper
- [29] Accelerating diffusion language model inference via efficient kv caching and guided diffusion View paper
- [30] Accelerating Diffusion LLMs via Adaptive Parallel Decoding View paper
- [31] Dimple: Discrete Diffusion Multimodal Large Language Model with Parallel Decoding View paper
- [32] Diffuspec: Unlocking diffusion language models for speculative decoding View paper
- [33] Speculative diffusion decoding: Accelerating language generation through diffusion View paper
- [34] Creditdecoding: Accelerating parallel decoding in diffusion large language models with trace credits View paper
- [35] Lavida: A large diffusion language model for multimodal understanding View paper

- [36] Accelerated Diffusion Models via Speculative Sampling View paper
- [37] Tidar: Think in diffusion, talk in autoregression View paper
- [38] Sdar: A synergistic diffusion-autoregression paradigm for scalable sequence generation View paper
- [39] ACDiT: Interpolating Autoregressive Conditional Modeling and Diffusion Transformer View paper
- [40] Blockwise sft for diffusion language models: Reconciling bidirectional attention and autoregressive decoding View paper
- [41] DiRL: An Efficient Post-Training Framework for Diffusion Language Models View paper