

DAWN: Dependency-Aware Fast Inference for Diffusion LLMs

Lizhuo Luo^{1,2} Zhuoran Shi³ Jiajun Luo⁴ Zhi Wang⁴ Shen Ren⁵ Wenya Wang¹ Tianwei Zhang¹

Abstract

Diffusion large language models (dLLMs) have shown advantages in text generation, particularly due to their inherent ability for parallel decoding. However, constrained by the quality-speed trade-off, existing inference solutions adopt conservative parallel strategies, leaving substantial efficiency potential underexplored. A core challenge is that parallel decoding assumes each position can be filled independently, but tokens are often semantically coupled. Thus, the correct choice at one position constrains valid choices at others. Without modeling these inter-token dependencies, parallel strategies produce deteriorated outputs. Motivated by this insight, we propose **DAWN**, a training-free, dependency-aware decoding method for fast dLLM inference. DAWN extracts token dependencies and leverages two key motivations: (1) positions dependent on unmasked certain positions become more reliable, (2) simultaneously unmasking strongly coupled uncertain positions induces errors. Given those findings, DAWN leverages a dependency graph to select more reliable unmasking positions at each iteration, achieving high parallelism with negligible loss in generation quality. Extensive experiments across multiple models and datasets demonstrate that DAWN speedups the inference by **1.80 - 8.06** \times over baselines while preserving the generation quality. Code is released at <https://github.com/lizhuo-luo/DAWN>.

1. Introduction

Diffusion models have achieved remarkable success in image (Podell et al., 2023; Xie et al., 2024) and video (Blattmann et al., 2023; Kong et al., 2024) generation, and

¹Nanyang Technological University ²AUMOVIO-NTU Corporate Lab, Nanyang Technological University ³Southern University of Science and Technology ⁴SIGS, Tsinghua University

⁵AUMOVIO Singapore Pte. Ltd.. Correspondence to: Tianwei Zhang <tianwei.zhang@ntu.edu.sg>.

Preprint. February 9, 2026.

have recently been extended to text generation. Unlike autoregressive (AR) models (OpenAI et al., 2024; Qwen et al., 2025) that generate tokens sequentially, diffusion large language models (dLLMs) (Nie et al., 2025; Ye et al., 2025) adopt full attention over all positions and refine an entire sequence through multiple denoising iterations, achieving surprisingly strong performance on text generation tasks. dLLMs offer potential solutions to longstanding limitations of AR models, including the inability to decode in parallel and the reversal curse (Berglund et al., 2023). These characteristics have drawn growing research interest in further advancing dLLMs (Google DeepMind, 2025; Song et al., 2025b; Labs et al., 2025; Bie et al., 2025).

Despite these efforts, dLLMs still exhibit performance gaps in practical deployments compared to state-of-the-art AR models (Wu et al., 2025b; Li et al., 2025b). These gaps are largely attributed to two main factors: KV-Cache management (Wu et al., 2025b; Ma et al., 2025; Liu et al., 2025b) and nonindependent position predictions (Song & Zhou, 2025; Wu et al., 2025b). First, dLLMs employ bidirectional attention, which fundamentally contradicts the causal assumption underlying standard KV-Cache mechanisms. Second, the marginal distributions at each position produced by dLLMs often violate the independence assumption underlying parallel decoding.

This work aims to improve the efficiency of parallel decoding in dLLMs, with a particular focus on *nonindependent position predictions*. Most existing parallel decoding strategies (Wu et al., 2025b; Ben-Hamu et al., 2025) select masked positions using heuristics and relatively coarse-grained criteria, such as confidence and entropy, to ensure that the selected positions behave approximately independently. However, overly conservative selection criteria can substantially limit the achievable parallelism, leaving much of the potential efficiency of parallel decoding underexploited. To address this problem, a natural alternative is to account for positional relationships more directly. Since the main difficulty stems from position coupling, improving parallel decoding requires approximating positional dependencies during inference rather than evaluating each position in isolation. Attention maps (Zhang et al., 2025a; 2024) provide an approximate yet cheap signal of token interactions that is readily available from each forward pass, making them a practical proxy for dependency estimation. From this per-

spective, two observations are particularly relevant: (i) we verify that dLLMs can exhibit abnormal attention concentration patterns (akin to attention sinks (Xiao et al., 2024; Rulli et al., 2025)) that are largely semantically irrelevant, which mislead the attention-based dependency estimates; and (ii) positions that are strongly dependent to previously unmasked high-confidence tokens can remain highly consistent with the final output even when their confidence is relatively low. Moreover, avoiding simultaneous unmasking of strongly coupled low-confidence positions can substantially reduce errors induced by parallel decoding under marginal probabilities. These observations provide a new insight: *dependency-aware inference rules can expand safe parallelism beyond those conservative threshold methods.*

Motivated by this, we propose DAWN, a training-free Dependency-Aware fast inference method for diffusion LLMs. It improves parallel decoding by explicitly accounting for position dependency. DAWN consists of three cooperating components: *Dependency Graph Construction*, *Anchor-Guided Decoding*, and *Conflict-Based Scheduling*. At each denoising iteration, a dependency graph is constructed from processed attention maps via thresholding, which captures salient token coupling relations. Based on this graph, the subsequent two components select two disjoint sets of positions for parallel updates, $\mathcal{U}_{\text{anchor}}$ and $\mathcal{U}_{\text{conflict}}$. *Anchor-Guided Decoding* first selects approximately independent high-confidence positions, and then treats previously unmasked high-confidence tokens as anchors and relaxes the confidence threshold for strongly anchor-coupled masked positions. *Conflict-Based Scheduling* identifies conflicts in the dependency graph and greedily constructs a large non-conflicting update set from the remaining candidates that satisfy a lower confidence threshold. Positions in $\mathcal{U}_{\text{anchor}}^{(t)} \cup \mathcal{U}_{\text{conflict}}^{(t)}$ are then unmasked in parallel.

Compared with prior approaches that rely on strict positional criteria, DAWN relaxes selection constraints by incorporating dependency information, thereby enabling additional unmasking that would otherwise be filtered out while preserving generation quality. Meanwhile, it substantially mitigates failures caused by non-independent position predictions during parallel decoding. Extensive experiments validate that DAWN improves the quality-speed trade-off of dLLM inference across multiple models and settings.

The main contributions are summarized as follows:

- We show that token dependencies can be estimated from attention maps during inference, enabling more aggressive parallel decoding. We identify two key findings: (1) attention sinks—positions that absorb disproportionate attention regardless of semantics—distort dependency estimates and must be filtered; (2) once a high-confidence token is committed, positions that depend on it become reliably predictable, even at lower confidence.

- We propose **DAWN**, a training-free, dependency-aware method for fast inference of diffusion LLMs. It uses the estimated dependencies in two ways: relaxing confidence thresholds for positions anchored by committed high-confidence tokens, and preventing strongly coupled low-confidence positions from being unmasked together, thereby enabling more efficient inference.
- Extensive experiments across multiple models, datasets, and representative baselines demonstrate the effectiveness of DAWN, achieving **1.80 - 8.06**× speedup over the baseline. Ablation studies further validate the contributions of each component and the impact of key hyperparameters.

2. Preliminaries

2.1. Inference Process of dLLMs

Most recent diffusion LLMs adopt the discrete masked diffusion model paradigm (Shi et al., 2025; Sahoo et al., 2024), where generation is cast as an iterative unmasking process. Unlike autoregressive models that commit tokens sequentially, dLLMs start from a heavily masked sequence and progressively recover masked positions over several denoising steps until no [MASK] remains. At each step, the model predicts token distributions for all masked positions, conditioned on the current state.

Given a prompt X , we initialize the response as a fully masked sequence of predefined length L :

$$y^{(0)} = ([\text{MASK}], \dots, [\text{MASK}]) \in (\mathcal{V} \cup \{[\text{MASK}]\})^L.$$

Here, \mathcal{V} denotes the vocabulary and [MASK] is the special mask token. In the naive setting, the sampler unmasks exactly one token with the highest confidence per step. At each denoising step $t = 0, 1, \dots, L-1$, it concatenates the prompt X and the current response state $y^{(t)}$ as the model input, and commits the corresponding token at the [MASK] response position with the highest confidence:

$$\begin{aligned} c_i^{(t)} &\triangleq \max_{v \in \mathcal{V}} p_\theta(y_i = v \mid X, y^{(t)}), \quad i \in M^{(t)}, \\ i_t &= \arg \max_{i \in M^{(t)}} c_i^{(t)}, \\ y_i^{(t+1)} &= \begin{cases} \arg \max_{v \in \mathcal{V}} p_\theta(y_i = v \mid X, y^{(t)}), & \text{if } i = i_t, \\ y_i^{(t)}, & \text{otherwise.} \end{cases} \end{aligned}$$

where $M^{(t)} \triangleq \{i \mid y_i^{(t)} = [\text{MASK}]\}$ denotes the set of masked response positions at step t . Repeating this procedure for L steps yields a fully unmasked response $y^{(L)}$.

Despite the inherent parallelism within each step, practical dLLM inference exhibits a pronounced quality-speed trade-off (Qian et al., 2026): directly unmasking multiple tokens in one step leads to substantial quality degradation.

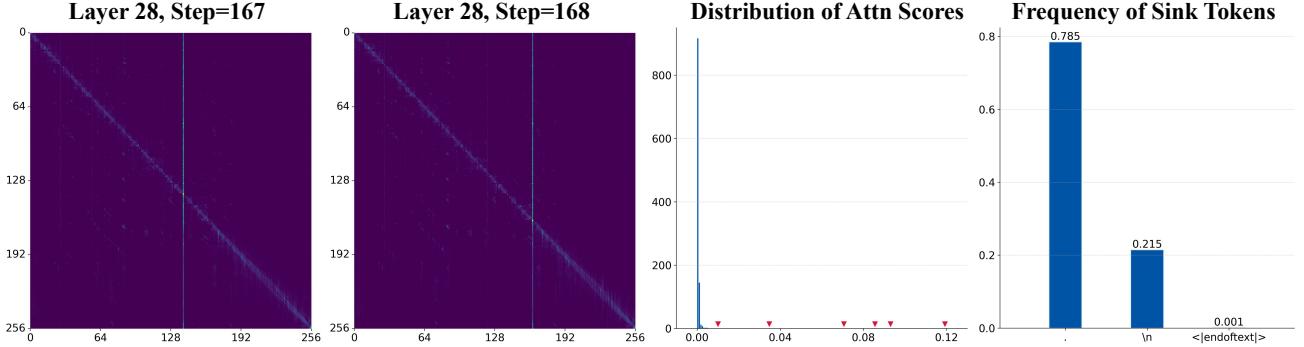


Figure 1. Attention Sinks in dLLMs. We conduct experiments on multiple samples using LLaDA-8B-Instruct. Left: The two heatmaps show partial attention maps from the same layer at different denoising steps, illustrating that the attention sink shifts across denoising iterations. Right: The third plot reports the distribution of attention scores corresponding to the first plot, and the rightmost plot reports the frequency of sink tokens from multiple samples.

2.2. Nonindependent Position Predictions

Unmasking a single token per step is inherently slow, and runs counter to the full position computation paradigm of dLLMs, where each denoising step produces predictions for all positions simultaneously. A common approach is to commit tokens at different positions independently according to the model’s predictive distributions at each position:

$$p_{\theta}\left(\{y_i\}_{i \in U^{(t)}} \mid X, y^{(t)}\right) \approx \prod_{i \in U^{(t)}} p_{\theta}\left(y_i \mid X, y^{(t)}\right),$$

where $U^{(t)} \subseteq M^{(t)}$ is a set of parallel unmasking positions. However, position predictions in masked refinement are often statistically coupled. Committing multiple tokens at strongly coupled positions in the same step can introduce inconsistencies and degrade generation quality. As the classic example in prior work (Song & Zhou, 2025; Wu et al., 2025b), two-word poker hands (e.g., “high card,” “two pair,” “full house,” “straight flush”) must match as a pair, while parallel sampling can produce invalid combinations such as “high house”. However, if we first decode one of them like “[MASK] house”, then the probability of recovering a valid pair in the next step becomes much higher, which aligns with the conditional probability factorization: $p(y_i, y_j \mid X, y^{(t)}) = p(y_j \mid X, y^{(t)}) p(y_i \mid y_j, X, y^{(t)})$. This nature of dLLMs suggests that parallel updates should account for position dependencies, avoiding strongly coupled positions while increasing the number of safely updated positions per step.

3. Observations

3.1. Attention Sinks Bias Dependency Proxies

Attention sinks (Xiao et al., 2024; Rulli et al., 2025) are common in dLLMs. Across multiple methods and diverse samples, we repeatedly observe an abnormal concentration of attention on a small subset of keys. Moreover, the dis-

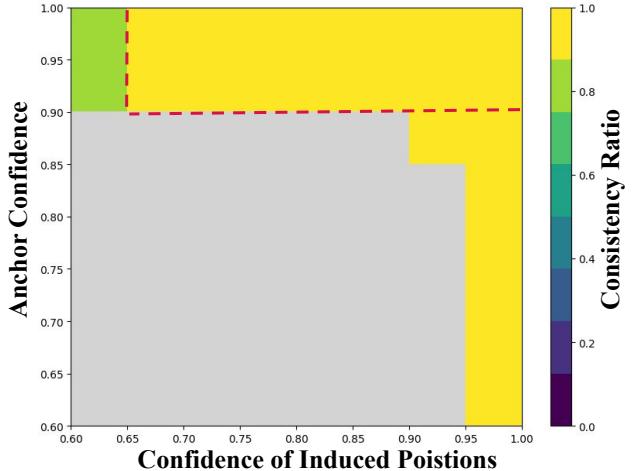


Figure 2. Heatmap of Induced Consistency We conduct experiments with LLaDA-8B-Instruct on sampled instances from GSM8K and HumanEval. For each request, we identify coupled pairs where anchors (prompts or previously unmasked tokens) influence induced positions (currently masked positions), and measure whether each induced token’s prediction matches the final decoded output (consistency ratio). Gray cells indicate bins with a negligible fraction of samples and are excluded from analysis.

tribution of concentration is not static. The specific sink tokens and the strength of aggregation can shift as the denoising step progresses. However, this phenomenon appears largely unrelated to the semantics of sink tokens. As shown in Fig. 1, each heatmap exhibits an abnormal concentration of attention mass. Comparing attention maps across denoising steps further indicates that the sink location shifts over iterations. In the rightmost plot, most identified sink tokens are punctuation marks or special tokens with little lexical meaning, suggesting that attention-sink formation is largely independent of token semantics.

Attention sinks can be problematic when attention maps are used as a proxy for token dependencies. Sink tokens attract

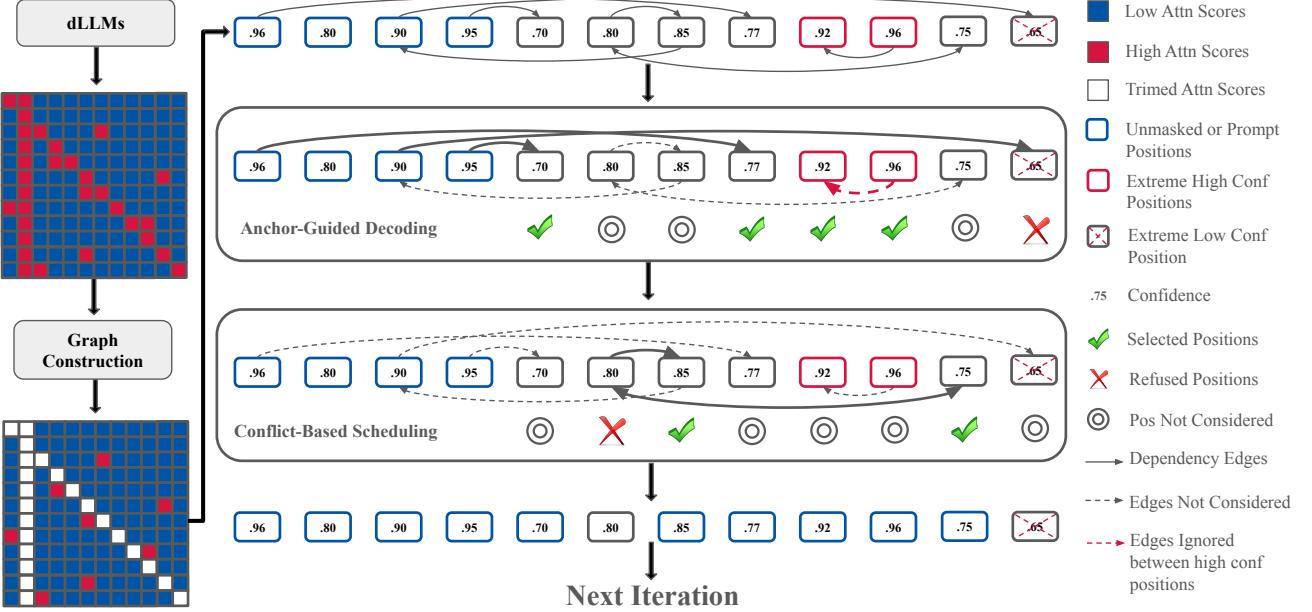


Figure 3. Overview of DAWN. Left: *Dependency Graph Construction* preprocesses the attention map and extracts a sparse directed dependency graph by retaining only salient (high-score) attention links. Middle: guided by this graph, *Anchor-Guided Decoding* and *Conflict-Based Scheduling* select two sets of positions, and the union of selected positions is unmasked simultaneously.

a large amount of attention and can be misinterpreted as exerting strong influence over many other tokens.

3.2. High-Confidence Positions as Anchors

During dLLM inference, masked positions can exhibit high consistency even when their confidence is not particularly high. More importantly, we find that the induced positions that are strongly dependent on prompts or previously unmasked tokens with high confidence (anchors) often exhibit high consistency despite relatively low instantaneous confidence. This is validated by the highlighted region in the upper area of Fig. 2: the induced positions can remain consistent with the final output when their confidence is low. In particular, when the corresponding anchors have confidence above 0.9, the induced tokens are consistently correct at a relatively low confidence. Therefore, the safety of parallel updates depends not only on confidence, but also on whether the prediction is sufficiently conditioned on reliable context.

4. Methodology

Driven by the observations, we propose DAWN, a training-free, dependency-aware solution to accelerate dLLM inference while maintaining the generation quality. The key idea is to *extract positional dependencies and leverage them to select a larger set of reliable positions for update at each iteration*. DAWN realizes efficient inference by three cooperating modules, as shown in Fig 3. *Dependency Graph Construction* builds a sparse directed graph from a lightweight proxy

of token dependencies based on attention maps. *Anchor-Guided Decoding* first selects high-confidence masked positions that are approximately independent, and then leverages high-confidence prompts or committed positions as anchors, enabling strongly coupled positions (induced positions) to be unmasked with a relatively low confidence. These together produce a parallel set $\mathcal{U}_{\text{anchor}}^{(t)}$. For the remaining tokens, *Conflict-Based Scheduling* uses conflict relations induced by the dependency graph to select a maximum independent set $\mathcal{U}_{\text{conflict}}^{(t)}$ from positions that meet a lower confidence threshold. Finally, the positions in $\mathcal{U}_{\text{anchor}}^{(t)} \cup \mathcal{U}_{\text{conflict}}^{(t)}$ are unmasked in parallel in this iteration. These three modules are pipelined within each denoising iteration to enable accurate and efficient inference. The following subsections describe their mechanisms in detail.

4.1. Dependency Graph Construction

Many prior works (Chen et al., 2024; Zhang et al., 2025a; 2024; Xi et al., 2025) use attention maps to characterize interactions among tokens, thereby enabling more efficient inference. Inspired by these applications, we treat attention maps as a lightweight, approximate proxy for token dependencies during decoding. This coupling signal is readily available from each forward pass and can be converted into a sparse directed dependency graph, which serves as the basis for subsequent efficient scheduling.

Since attention patterns evolve across denoising iterations, the dependency graph is constructed at each iteration from

the model’s attention maps. To obtain a signal that is closer to the final prediction and less noisy, attention weights are averaged across the last few layers and all heads. As discussed in Sec. 3.1, dLLMs often exhibit attention sinks, where a small set of positions absorbs most of the attention mass due to systematic bias rather than semantic relevance. This can lead to misleading dependent relations when attention is used as a proxy for dependency. To mitigate this effect, sink positions are identified via outlier detection and filtering of their incoming attention mass: given an aggregated attention matrix $A^{(t)} \in \mathbb{R}^{L \times L}$ at iteration t , the incoming attention mass of position j is defined as

$$\bar{A}_j^{(t)} = \frac{1}{L} \sum_{i=1}^L A_{i,j}^{(t)},$$

and position j is marked as a sink if $\bar{A}_j^{(t)}$ is larger than a predefined threshold τ_{sink} . Self-attention on the diagonal is ignored as it does not capture cross-position dependencies.

Given the processed attention proxy at iteration t , a directed sparse dependency graph is constructed to capture salient token couplings during decoding. Specifically, to keep the graph sparse and focus on the most informative relations, edges are retained based on thresholded (τ_{edge}) attention scores. A directed edge $j \rightarrow i$ is added if query position i places a sufficiently large attention mass on key position j , indicating that the prediction at token i can be significantly conditioned by token j . The resulting graph provides an approximate representation of positional dependencies and will be used by subsequent modules.

4.2. Anchor-Guided Decoding

Prior work (Wu et al., 2025b) suggests that sufficiently high-confidence positions are approximately independent as well as non-conflicting, motivating a high conservative threshold τ_{high} for selecting them for parallel updates, where τ_{high} is usually set to 0.9. Meanwhile, some low-confidence tokens at masked positions can still be consistent with the final response, indicating that only confidence thresholding remains a limit for dLLM inference. Motivated by this, *Anchor-Guided Decoding* is introduced to accept approximately independent positions and reliable positions under a relaxed confidence criterion.

As discussed in Sec. 3.2, low-confidence positions that are strongly coupled to previously unmasked positions meet the confidence threshold τ_{high} and often match the final result. Following this observation, we define anchors as positions that have been unmasked and their confidence meets τ_{high} . Using the dependency graph, we identify the induced positions \mathcal{I} as masked positions that are reachable via directed edges from anchor tokens. Intuitively, higher-confidence anchors provide more reliable context, allowing the corre-

sponding induced positions to be unmasked at lower confidence. Accordingly, the set of positions eligible to be unmasked at the denoising step t is defined as:

$$\mathcal{U}_{\text{anchor}}^{(t)} = \left\{ i \mid \begin{array}{l} i \in M^{(t)}, c_i \geq \tau_{high} \text{ or} \\ i \in \mathcal{I}^{(t)}, c_i \geq \tau_{induced} \end{array} \right\}.$$

where $\tau_{induced}$ is the confidence threshold for unmasking induced positions. Overall, *Anchor-Guided Decoding* selects approximately independent positions under a high confidence threshold and relaxes the threshold for induced positions, enabling more efficient inference.

4.3. Conflict-Based Scheduling

For the remaining positions that are strongly coupled and satisfy a lower confidence threshold τ_{low} , simultaneously unmasking them may introduce inconsistent commitments, as discussed in Sec. 2.2. To mitigate such errors while retaining high parallelism, *Conflict-Based Scheduling* is introduced to prevent lower-confidence but highly coupled positions from being decoded in the same step.

Since the dependency graph captures salient positional dependencies, it can be used to identify strongly coupled position pairs. A conflict relation is defined between two positions connected by an edge in the dependency graph, regardless of direction: if either $i \rightarrow j$ or $j \rightarrow i$ is present, then positions i and j are considered to be in conflict and should not be unmasked simultaneously.

Algorithm 1 shows the detailed procedure. Specifically, at iteration t , based on these conflicts and the graph topology, a greedy independent set is constructed from the remaining positions that (i) are not yet included in and not conflicting neighbors of $\mathcal{U}_{\text{anchor}}^{(t)}$, and (ii) satisfy a lower confidence threshold, yielding additional positions that can be decoded in parallel. Concretely, positions are greedily selected in descending order of confidence: the highest-confidence position is added to the update set $\mathcal{U}_{\text{conflict}}^{(t)}$, and all of its conflicting neighbors are removed from further consideration. This procedure repeats until no candidate positions remain.

In practice, quality degradation from naively lowering the confidence threshold is largely driven by positional coupling. By explicitly avoiding simultaneous unmasking of strongly coupled positions, *Conflict-Based Scheduling* helps maintain decoding quality while allowing a lower confidence threshold τ_{low} , thus speeding up the inference.

5. Experiments

5.1. Setups

Models and Benchmarks. We evaluate our approach on several variants of two models: LLaDA-8B-Instruct (Nie et al., 2025), LLaDA-1.5 (Zhu et al., 2025), Dream-v0-Base-

Algorithm 1 Conflict-Based Scheduling

```

1: Input: remaining candidate positions  $\mathcal{C}^{(t)}$ , anchor update set  $\mathcal{U}_{\text{anchor}}^{(t)}$ , confidence scores  $\{c_i\}_{i \in \mathcal{C}^{(t)}}$ , conflict neighbors  $\{\mathcal{N}(i)\}_{i \in \mathcal{C}^{(t)}}$ , lower threshold  $\tau_{\text{low}}$ 
2: Output: parallel update set  $\mathcal{U}_{\text{conflict}}^{(t)}$ 
3: Initialize  $\mathcal{U}_{\text{conflict}}^{(t)} \leftarrow \emptyset$ 
4:  $\mathcal{X} \leftarrow \mathcal{U}_{\text{anchor}}^{(t)} \cup \bigcup_{i \in \mathcal{U}_{\text{anchor}}^{(t)}} \mathcal{N}(i)$  {selected positions and their conflicts}
5:  $\mathcal{R} \leftarrow \{i \in \mathcal{C}^{(t)} \mid c_i \geq \tau_{\text{low}}\} \setminus \mathcal{X}$  {remaining candidates}
6: while  $\mathcal{R} \neq \emptyset$  do
7:   Select  $i^* \leftarrow \arg \max_{i \in \mathcal{R}} c_i$ 
8:    $\mathcal{U}_{\text{conflict}}^{(t)} \leftarrow \mathcal{U}_{\text{conflict}}^{(t)} \cup \{i^*\}$ 
9:    $\mathcal{R} \leftarrow \mathcal{R} \setminus (\{i^*\} \cup \mathcal{N}(i^*))$ 
10: end while
11: return  $\mathcal{U}_{\text{conflict}}^{(t)}$ 

```

7B (Ye et al., 2025), Dream-v0-Instruct-7B. Benchmarks include diverse datasets: GSM8K (5-shot) (Cobbe et al., 2021), MATH (4-shot) (Hendrycks et al., 2021), HumanEval (0-shot) (Chen, 2021), and MBPP (3-shot) (Austin et al., 2021), covering a range of reasoning and code generation tasks. We report tokens per second (TPS), relative speedup ratio (Speedup) and number of function evaluations (NFE) to reflect efficiency, along with task accuracy (Acc.).

Baselines. We compare DAWN against four baselines: the Original sampling method (Top-1 Sampling), which unmasks the top-1 confidence position at each iteration; the Confidence-Aware Parallel proposed by Fast-dLLM (Wu et al., 2025b), selecting positions whose confidence exceeds a predefined threshold; KLASS (Kim et al., 2025), using both confidence and KL divergence to select positions; and LocalLeap (Kong et al., 2025), identifying anchors and performing localized relaxed parallel decoding.

Hardware and Implementation Details. Our experiments are conducted on a NVIDIA H100 80G GPU. We set the generation length to 256 and the block length to 32 for all methods except KLASS, which uses its best-performing block length. All baselines are evaluated under their default hyperparameter settings. For DAWN, we average the attention maps from the last 4 layers. τ_{high} is set to 0.9 according to Fast-dLLM. Confidence thresholds are set from 0.7 to 0.85. Full configurations and the justifications can be found in Appendix A. All evaluations are conducted using the standardized lm-eval (Gao et al., 2024) library.

5.2. Main Results

Table 1 reports the accuracy and efficiency of each method across different models on four benchmarks.

Overall, DAWN achieves a substantial improvement in inference speed while maintaining accuracy comparable to or even slightly higher than the original method. DAWN achieves substantially higher throughput than prior decoding baselines, with up to **8.06**× speedup on MBPP using the LLaDA-1.5 model. Moreover, these gains do not come at the cost of quality. On LLaDA-8B-Instruct, DAWN achieves 77.94 accuracy on GSM8K, matching the original method, and attains 30.80 on MBPP, slightly higher than the original 29.60. In other settings, DAWN incurs only negligible quality loss, indicating a more favorable quality-speed trade-off.

Compared with confidence-aware parallel and KLASS, DAWN significantly improves throughput while achieving nearly identical accuracy. Compared with LocalLeap, DAWN shows clear advantages in both quality and speed: across most benchmarks, it improves throughput by approximately 0.05 – 5.17 tokens per second while enhancing accuracy by up to 3.04%. It unlocks additional safe parallelism by explicitly accounting for positional dependencies during refinement, enabling more efficient decoding.

5.3. Ablation Study

We conduct extensive ablation studies to assess the contributions of key components in DAWN. We further examine the sensitivity of DAWN to the generation length and the lower confidence threshold τ_{low} to understand how these factors affect its effectiveness. Discussions of other hyperparameter choices are provided in Appendix A. All other experiment settings follow the main experiment.

Effectiveness of key components. Since *Dependency Graph Construction* serves as the foundation for the other two modules, we focus on validating the effectiveness of components for selecting parallel update positions. Table 2 summarizes ablations of DAWN on two representative dLLMs and benchmarks. Overall, the full solution consistently improves efficiency while maintaining comparable accuracy. In particular, removing *Anchor-Guided Decoding* (AGD) causes a substantial efficiency drop across both models and tasks. On LLaDA-8B-Instruct, TPS decreases from 44.72 to 22.31 on GSM8K, indicating that AGD is a primary contributor to the speedup by expanding the set of positions that can be safely unmasked under reliable anchor context. On Dream-v0-Instruct-7B, removing *Conflict-Based Scheduling* (CBS) yields higher accuracy but lower efficiency, suggesting that CBS mainly unlocks additional parallelism by avoiding inconsistent joint updates among strongly coupled positions, trading a small amount of accuracy for further acceleration.

Impact of generation length. Figure 4 reports the accuracy and throughput of DAWN and the original baseline with different generation lengths L . Across both LLaDA and Dream models, DAWN consistently improves throughput over the

Table 1. Performance comparison between DAWN and baselines across 4 datasets and 4 models. We report Accuracy, TPS, and Speedup to assess their efficiency and generation quality.

Method	GSM8K			MATH			HumanEval			MBPP		
	Acc. \uparrow	TPS \uparrow	Speedup \uparrow	Acc. \uparrow	TPS \uparrow	Speedup \uparrow	Acc. \uparrow	TPS \uparrow	Speedup \uparrow	Acc. \uparrow	TPS \uparrow	Speedup \uparrow
LLaDA-8B-Instruct												
Original	77.94	10.32	1.00 \times	33.20	14.10	1.00 \times	40.24	26.46	1.00 \times	29.60	9.46	1.00 \times
Confidence	78.39	33.44	3.24 \times	32.98	36.88	2.62 \times	40.85	86.73	3.28 \times	30.00	34.81	3.68 \times
KLASS	76.12	23.23	2.25 \times	32.32	25.27	1.79 \times	36.59	56.91	2.15 \times	27.00	22.62	2.39 \times
LocalLeap	77.33	44.19	4.28 \times	32.42	47.12	3.34 \times	39.63	109.80	4.15 \times	30.80	44.13	4.66 \times
DAWN	77.94	44.72	4.33 \times	32.36	48.16	3.42 \times	40.24	108.99	4.12 \times	30.80	45.17	4.77 \times
LLaDA-1.5												
Original	81.12	9.65	1.00 \times	33.36	12.75	1.00 \times	43.90	9.26	1.00 \times	38.80	3.45	1.00 \times
Confidence	80.74	32.49	3.37 \times	33.38	32.87	2.58 \times	42.68	23.39	2.53 \times	38.80	20.40	5.91 \times
KLASS	77.63	22.63	2.35 \times	31.80	23.29	1.83 \times	39.63	16.49	1.78 \times	30.00	13.25	3.84 \times
LocalLeap	80.14	42.60	4.41 \times	32.44	42.36	3.32 \times	42.07	30.21	3.26 \times	39.20	27.00	7.83 \times
DAWN	80.82	43.24	4.48 \times	31.80	42.84	3.36 \times	42.07	29.53	3.19 \times	37.60	27.80	8.06 \times
Dream-v0-Base-7B												
Original	76.42	14.37	1.00 \times	34.70	18.16	1.00 \times	38.41	20.90	1.00 \times	53.40	11.85	1.00 \times
Confidence	73.84	23.41	1.63 \times	34.38	44.66	2.46 \times	39.63	61.54	2.94 \times	53.80	50.37	4.25 \times
KLASS	72.00	12.58	0.88 \times	32.80	23.16	1.27 \times	42.68	31.32	1.50 \times	55.80	19.46	1.64 \times
LocalLeap	72.55	25.40	1.77 \times	34.48	51.32	2.83 \times	36.59	68.28	3.27 \times	53.60	59.38	5.01 \times
DAWN	73.54	25.88	1.80 \times	34.00	51.45	2.83 \times	39.63	68.96	3.29 \times	53.20	64.55	5.45 \times
Dream-v0-Instruct-7B												
Original	76.35	7.30	1.00 \times	37.80	17.73	1.00 \times	53.66	22.66	1.00 \times	54.20	8.37	1.00 \times
Confidence	75.20	28.81	3.95 \times	38.24	37.92	2.14 \times	57.32	52.11	2.30 \times	55.60	39.09	4.67 \times
KLASS	71.72	12.93	1.77 \times	34.66	18.74	1.06 \times	54.88	26.71	1.18 \times	56.80	18.53	2.21 \times
LocalLeap	73.24	32.94	4.51 \times	38.10	43.66	2.46 \times	54.27	59.28	2.62 \times	55.00	42.71	5.10 \times
DAWN	73.16	32.99	4.52 \times	38.22	44.18	2.49 \times	54.88	60.23	2.66 \times	55.80	44.03	5.26 \times

Table 2. Ablation study on the effectiveness of key modules in DAWN across 2 datasets and 2 models. We report Accuracy, TPS, and NFE to assess their quality and speed. The settings are the same as in the main experiments.

Method	GSM8K			HumanEval		
	Acc. \uparrow	TPS \uparrow	NFE \downarrow	Acc. \uparrow	TPS \uparrow	NFE \downarrow
LLaDA-8B-Instruct						
DAWN	77.94	44.72	55.76	40.24	109.0	61.63
- AGD	76.80	22.31	112.9	41.46	51.51	131.0
- CBS	78.47	33.83	74.69	41.46	92.52	72.93
Original	77.94	10.32	256	40.24	26.46	256
Dream-v0-Instruct-7B						
DAWN	73.16	32.99	51.92	54.88	60.23	77.12
- AGD	73.31	29.33	58.13	54.88	54.79	83.34
- CBS	75.28	27.63	62.19	57.31	51.97	93.01
Original	76.35	7.30	256	53.66	22.66	256

original sampler while maintaining comparable accuracy. As L increases, throughput decreases for both methods due to higher denoising costs, yet DAWN continues to boost the efficiency. Overall, it preserves a more favorable quality-speed trade-off across a wide range of lengths.

Impact of block length. Figure 5 reports the accuracy and throughput of DAWN with different block lengths. As

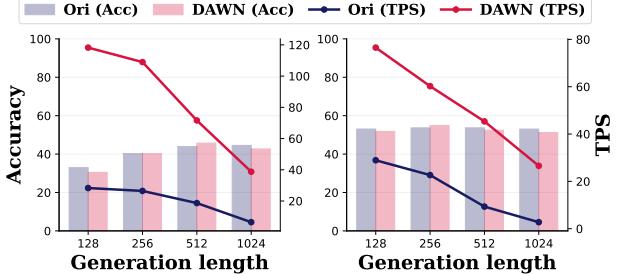


Figure 4. Effectiveness of DAWN and the original sampler on HumanEval under different generation lengths ($L \in \{128, 256, 512, 1024\}$). Bars report accuracy (left y-axis) and solid lines report TPS (right y-axis). Left and right figures correspond to LLaDA-8B-Instruct and Dream-v0-Instruct-7B.

the block length increases, throughput increases for both methods as higher parallelism, but accuracy first increases and then decreases. Overall, it preserves a robust quality-speed trade-off across a wide range of lengths.

Impact of lower threshold. Figure 6 reports the evaluation under different values of lower threshold τ_{low} . We observe a clear quality-speed trade-off on both models. Reducing τ_{low} increases TPS by admitting more parallel updates, but can hurt accuracy due to less reliable low-confidence com-

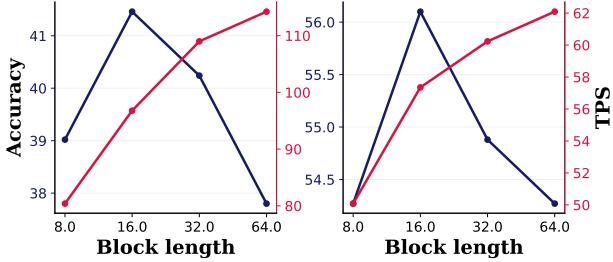


Figure 5. Effectiveness of DAWN on HumanEval under different block lengths ($L \in \{8, 16, 32, 64\}$). We report accuracy (blue, left y-axis) and TPS (red, right y-axis). Left and right figures correspond to LLaDA-8B-Instruct and Dream-v0-Instruct-7B.

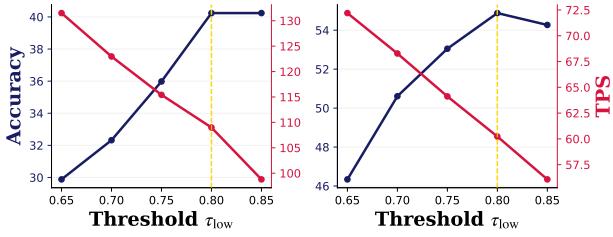


Figure 6. We vary the lower threshold τ_{low} and report accuracy (blue, left y-axis) and throughput TPS (red, right y-axis) on HumanEval. The left and right figures correspond to LLaDA-8B-Instruct and Dream-v0-Instruct-7B. The dashed line (yellow) marks the default setting $\tau_{low} = 0.80$.

mitments. The default $\tau_{low} = 0.80$ (dashed line) maintains a high generation quality and improved efficiency.

6. Related Work

Diffusion Large Language Models. Autoregressive (AR) models have long been the dominant paradigm for natural language generation, largely due to the discrete and sequential nature of text. In contrast, diffusion models have achieved remarkable success in continuous domains such as image and video generation. Recently, diffusion-based language models have gained renewed interest and emerged as a competitive alternative for text generation, demonstrating promising progress across a wide range of tasks.

Representative approaches include pretraining dLLMs from scratch (Nie et al., 2025) and building dLLMs on top of existing AR models (Ye et al., 2025). In parallel, several commercial systems (Song et al., 2025b; Labs et al., 2025; Google DeepMind, 2025) highlight the feasibility and practical potential of diffusion-based generation. Beyond these early successes, recent works (Wu et al., 2025a; Bie et al., 2025; Liu et al., 2025a) continue to advance dLLMs along multiple dimensions, including scaling to larger model size and exploring alternative training-inference paradigms for faster refinement. Moreover, diffusion-based modeling has been extended to multimodal settings, where multimodal

dLLMs (You et al., 2025; Yang et al., 2025a) demonstrate strong performance across a variety of tasks and point toward more unified diffusion-based generative models.

Efficient Inference of dLLMs. Despite dLLMs’s ability to update multiple positions in parallel, they still face practical challenges during inference, motivating more efforts on faster and more reliable decoding.

Existing works focus on the KV Cache optimization (Wu et al., 2025b; Ma et al., 2025), early stopping (Yang et al., 2025b; Li et al., 2025a), distillation-based (Chen et al., 2025) acceleration, and others (Zhang et al., 2025b; Song et al., 2025a; Luo et al., 2026). These directions have shown promising gains in improving the efficiency of dLLM inference. Beyond these optimizations, numerous studies have explored optimized sampling strategies for dLLM inference. Fast-dLLM (Wu et al., 2025b) adopts a confidence-aware strategy that unmasks multiple positions when their scores are sufficiently high, thereby making the independence approximation more reliable. EB-Sampler (Ben-Hamu et al., 2025) uses the entropy of predictive distributions to decide which positions are safe to update in parallel. KLASS (Kim et al., 2025) further incorporates temporal stability by comparing distributions across iterations via the KL divergence. WINO (Hong et al., 2025) follows a draft-and-verify style: it drafts many tokens in parallel and selectively regenerates those that fail verification. Spiffy and related works (Agrawal et al., 2026; Wei et al., 2025) apply speculative decoding style strategies to dLLMs to accelerate diffusion decoding. LocalLeap (Kong et al., 2025) leverages a local determinism hypothesis, observing that positions adjacent to high-confidence commits tend to stabilize earlier and can therefore be updated more aggressively. Unlike the above methods, this work explicitly approximates token coupling during inference and leverages this approximation to guide efficient parallel sampling.

7. Conclusion

This work leads to a better quality-speed trade-off for dLLM inference, narrowing the gap to state-of-the-art language models in practical generation settings. Specifically, we propose **DAWN**, a training-free, dependency-aware fast inference method for dLLMs. It is primarily motivated by the inherent challenge of nonindependent position predictions in dLLMs and mitigates this issue by adopting a positional-dependency perspective, offering a complementary approach to alleviate failures in parallel unmasking. Guided by a sparse directed dependency graph, DAWN selects unmasking positions at each iteration and enables highly parallel updates while preserving generation quality. Extensive experiments across multiple models and datasets validate the effectiveness of DAWN, demonstrating consistent speedups while maintaining comparable quality.

8. Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

- Agrawal, S., Garrepalli, R., Goel, R., Lee, M., Lott, C., and Porikli, F. Spiffy: Multiplying diffusion lilm acceleration via lossless speculative decoding, 2026. URL <https://arxiv.org/abs/2509.18085>.
- Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C., Terry, M., Le, Q., et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.
- Ben-Hamu, H., Gat, I., Severo, D., Nolte, N., and Karrer, B. Accelerated sampling from masked diffusion models via entropy bounded unmasking, 2025. URL <https://arxiv.org/abs/2505.24857>.
- Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., and Evans, O. The reversal curse: Llms trained on “a is b” fail to learn “b is a”. *arXiv preprint arXiv:2309.12288*, 2023.
- Bie, T., Cao, M., Chen, K., Du, L., Gong, M., Gong, Z., Gu, Y., Hu, J., Huang, Z., Lan, Z., Li, C., Li, C., Li, J., Li, Z., Liu, H., Liu, L., Lu, G., Lu, X., Ma, Y., Tan, J., Wei, L., Wen, J.-R., Xing, Y., Zhang, X., Zhao, J., Zheng, D., Zhou, J., Zhou, J., Zhou, Z., Zhu, L., and Zhuang, Y. Llada2.0: Scaling up diffusion language models to 100b, 2025. URL <https://arxiv.org/abs/2512.15745>.
- Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.
- Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., and Chang, B. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models, 2024. URL <https://arxiv.org/abs/2403.06764>.
- Chen, M. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.
- Chen, Z., Fang, G., Ma, X., Yu, R., and Wang, X. dparallel: Learnable parallel decoding for dllms. *arXiv preprint arXiv:2509.26488*, 2025.
- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac'h, A., Li, H., McDonell, K., Muennighoff, N., Ociepa, C., Phang, J., Reynolds, L., Schoelkopf, H., Skowron, A., Sutawika, L., Tang, E., Thite, A., Wang, B., Wang, K., and Zou, A. The language model evaluation harness, 07 2024. URL <https://zenodo.org/records/12608602>.
- Google DeepMind. Gemini diffusion: Our state-of-the-art, experimental text diffusion model. <https://deepmind.google/models/gemini-diffusion/>, 2025. Accessed: 2026-01-10.
- Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.
- Hong, F., Yu, G., Ye, Y., Huang, H., Zheng, H., Zhang, Y., Wang, Y., and Yao, J. Wide-in, narrow-out: Revokable decoding for efficient and effective dllms. *arXiv preprint arXiv:2507.18578*, 2025.
- Kim, S. H., Hong, S., Jung, H., Park, Y., and Yun, S.-Y. Klass: Kl-guided fast inference in masked diffusion models, 2025. URL <https://arxiv.org/abs/2511.05664>.
- Kong, F., Zhang, J., Liu, Y., Wu, Z., Tian, Y., W., V., and Zhou, G. Accelerating diffusion lilm inference via local determinism propagation, 2025. URL <https://arxiv.org/abs/2510.07081>.
- Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al. Hunyuan-video: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.
- Labs, I., Khanna, S., Kharbanda, S., Li, S., Varma, H., Wang, E., Birnbaum, S., Luo, Z., Miraoui, Y., Palrecha, A., et al. Mercury: Ultra-fast language models based on diffusion. *arXiv preprint arXiv:2506.17298*, 2025.
- Li, J., Dong, X., Zang, Y., Cao, Y., Wang, J., and Lin, D. Beyond fixed: Training-free variable-length denoising for diffusion large language models, 2025a. URL <https://arxiv.org/abs/2508.00819>.
- Li, Y., Wei, F., Zhang, C., and Zhang, H. Eagle-3: Scaling up inference acceleration of large language models via training-time test, 2025b. URL <https://arxiv.org/abs/2503.01840>.

- Liu, A., He, M., Zeng, S., Zhang, S., Zhang, L., Wu, C., Jia, W., Liu, Y., Zhou, X., and Zhou, J. Wedlm: Reconciling diffusion language models with standard causal attention for fast inference, 2025a. URL <https://arxiv.org/abs/2512.22737>.
- Liu, Z., Yang, Y., Zhang, Y., Chen, J., Zou, C., Wei, Q., Wang, S., and Zhang, L. dllm-cache: Accelerating diffusion large language models with adaptive caching. *arXiv preprint arXiv:2506.06295*, 2025b.
- Luo, L., Li, S., Wen, Y., and Zhang, T. Dsb: Dynamic sliding block scheduling for diffusion llms, 2026. URL <https://arxiv.org/abs/2602.05992>.
- Ma, X., Yu, R., Fang, G., and Wang, X. dkv-cache: The cache for diffusion language models. *arXiv preprint arXiv:2505.15781*, 2025.
- Nie, S., Zhu, F., You, Z., Zhang, X., Ou, J., Hu, J., Zhou, J., Lin, Y., Wen, J.-R., and Li, C. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.
- OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O’Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotstetd, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2024. URL <https://arxiv.org/abs/2303.08774>.
- Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- Qian, Y.-Y., Su, J., Hu, L., Zhang, P., Deng, Z., Zhao, P., and Zhang, H. d3llm: Ultra-fast diffusion lilm using pseudo-trajectory distillation. *arXiv preprint arXiv:2601.07568*, 2026.
- Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. Qwen2.5 technical report, 2025. URL <https://arxiv.org/abs/2412.15115>.
- Rulli, M. E., Petruzzi, S., Michielon, E., Silvestri, F., Scardapane, S., and Devoto, A. Attention sinks in diffusion language models, 2025. URL <https://arxiv.org/abs/2510.15731>.
- Sahoo, S. S., Arriola, M., Schiff, Y., Gokaslan, A., Marroquin, E., Chiu, J. T., Rush, A., and Kuleshov, V. Simple

- and effective masked diffusion language models, 2024.
URL <https://arxiv.org/abs/2406.07524>.
- Shi, J., Han, K., Wang, Z., Doucet, A., and Titsias, M. K. Simplified and generalized masked diffusion for discrete data, 2025. URL <https://arxiv.org/abs/2406.04329>.
- Song, J. and Zhou, L. Ideas in inference-time scaling can benefit generative pre-training algorithms, 2025. URL <https://arxiv.org/abs/2503.07154>.
- Song, Y., Liu, X., Li, R., Liu, Z., Huang, Z., Guo, Q., He, Z., and Qiu, X. Sparse-dllm: Accelerating diffusion llms with dynamic cache eviction, 2025a. URL <https://arxiv.org/abs/2508.02558>.
- Song, Y., Zhang, Z., Luo, C., Gao, P., Xia, F., Luo, H., Li, Z., Yang, Y., Yu, H., Qu, X., Fu, Y., Su, J., Zhang, G., Huang, W., Wang, M., Yan, L., Jia, X., Liu, J., Ma, W.-Y., Zhang, Y.-Q., Wu, Y., and Zhou, H. Seed diffusion: A large-scale diffusion language model with high-speed inference, 2025b. URL <https://arxiv.org/abs/2508.02193>.
- Wei, L., Chen, W., Tang, P., Guo, X., Ye, L., Wang, R., and Li, M. Orchestrating dual-boundaries: An arithmetic intensity inspired acceleration framework for diffusion language models. *arXiv preprint arXiv:2511.21759*, 2025.
- Wu, C., Zhang, H., Xue, S., Diao, S., Fu, Y., Liu, Z., Molchanov, P., Luo, P., Han, S., and Xie, E. Fast-dllm v2: Efficient block-diffusion llm, 2025a. URL <https://arxiv.org/abs/2509.26328>.
- Wu, C., Zhang, H., Xue, S., Liu, Z., Diao, S., Zhu, L., Luo, P., Han, S., and Xie, E. Fast-dllm: Training-free acceleration of diffusion llm by enabling kv cache and parallel decoding, 2025b. URL <https://arxiv.org/abs/2505.22618>.
- Xi, H., Yang, S., Zhao, Y., Xu, C., Li, M., Li, X., Lin, Y., Cai, H., Zhang, J., Li, D., et al. Sparse videogen: Accelerating video diffusion transformers with spatial-temporal sparsity. *arXiv preprint arXiv:2502.01776*, 2025.
- Xiao, G., Tian, Y., Chen, B., Han, S., and Lewis, M. Efficient streaming language models with attention sinks, 2024. URL <https://arxiv.org/abs/2309.17453>.
- Xie, E., Chen, J., Chen, J., Cai, H., Tang, H., Lin, Y., Zhang, Z., Li, M., Zhu, L., Lu, Y., and Han, S. Sana: Efficient high-resolution image synthesis with linear diffusion transformers, 2024. URL <https://arxiv.org/abs/2410.10629>.
- Yang, L., Tian, Y., Li, B., Zhang, X., Shen, K., Tong, Y., and Wang, M. Mmada: Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*, 2025a.
- Yang, Y., Wang, C., Wang, S., Wen, Z., Qi, B., Xu, H., and Zhang, L. Diffusion llm with native variable generation lengths: Let [eos] lead the way, 2025b. URL <https://arxiv.org/abs/2510.24605>.
- Ye, J., Xie, Z., Zheng, L., Gao, J., Wu, Z., Jiang, X., Li, Z., and Kong, L. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*, 2025.
- You, Z., Nie, S., Zhang, X., Hu, J., Zhou, J., Lu, Z., Wen, J.-R., and Li, C. Llada-v: Large language diffusion models with visual instruction tuning. *arXiv preprint arXiv:2505.16933*, 2025.
- Zhang, J., Wei, J., Huang, H., Zhang, P., Zhu, J., and Chen, J. Sageattention: Accurate 8-bit attention for plug-and-play inference acceleration. *arXiv preprint arXiv:2410.02367*, 2024.
- Zhang, J., Xiang, C., Huang, H., Wei, J., Xi, H., Zhu, J., and Chen, J. Spargeattn: Accurate sparse attention accelerating any model inference. *arXiv preprint arXiv:2502.18137*, 2025a.
- Zhang, T., Li, Z., Yan, X., Qin, H., Guo, Y., and Zhang, Y. Quant-dllm: Post-training extreme low-bit quantization for diffusion large language models, 2025b. URL <https://arxiv.org/abs/2510.03274>.
- Zhu, F., Wang, R., Nie, S., Zhang, X., Wu, C., Hu, J., Zhou, J., Chen, J., Lin, Y., Wen, J.-R., and Li, C. Llada 1.5: Variance-reduced preference optimization for large language diffusion models, 2025. URL <https://arxiv.org/abs/2505.19223>.

A. Experiment Details

We conduct our experiments on several variants of two models: LLaDA-8B-Instruct (Nie et al., 2025), LLaDA-1.5 (Zhu et al., 2025), Dream-v0-Base-7B (Ye et al., 2025), Dream-v0-Instruct-7B. Benchmarks include diverse datasets: GSM8K (5-shot) (Cobbe et al., 2021), MATH (4-shot) (Hendrycks et al., 2021), HumanEval (0-shot) (Chen, 2021), and MBPP (3-shot) (Austin et al., 2021), covering a range of reasoning and code generation tasks. Across all settings, we fix the block size to 32 and generation length to 256 tokens except KLASS, which uses its best-performing block length.

All baselines are evaluated under their default hyperparameter settings. To identify the optimal hyperparameter configuration for our method DAWN on different models, including τ_{edge} , τ_{induced} and τ_{sink} , we conduct a grid search on the HumanEval dataset. Specifically, we evaluate a range of candidate values for each parameter and plot the corresponding Accuracy–TPS curves, visualized in Fig 7, 8, 9, 10.

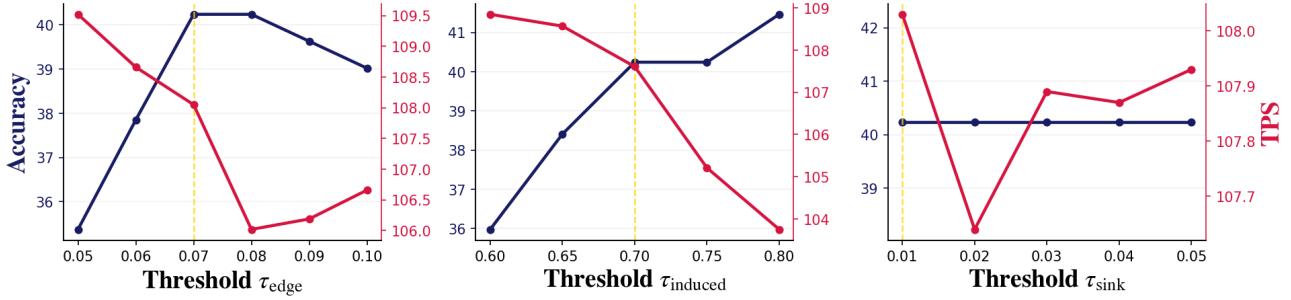


Figure 7. We vary the threshold τ_{edge} , τ_{induced} , τ_{sink} and report accuracy (blue, left y-axis) and throughput TPS (red, right y-axis) on LLaDA-8B-Instruct. The dashed lines (yellow) mark the final settings $\tau_{\text{edge}} = 0.07$, $\tau_{\text{induced}} = 0.70$, $\tau_{\text{sink}} = 0.01$.

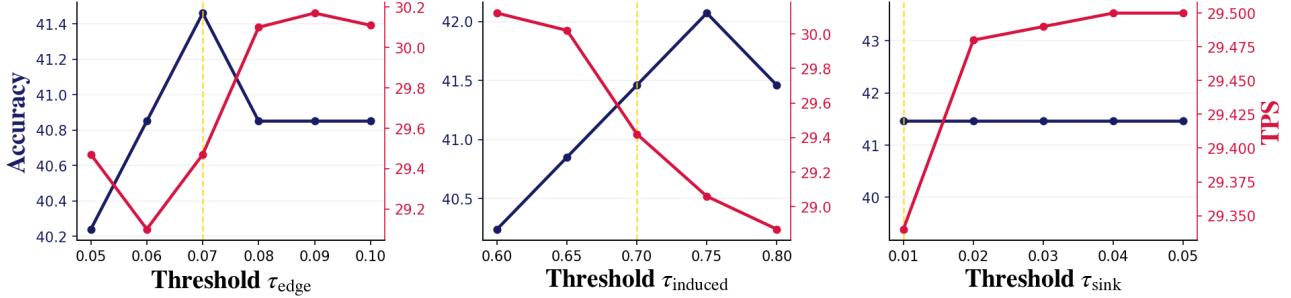


Figure 8. We vary the threshold τ_{edge} , τ_{induced} , τ_{sink} and report accuracy (blue, left y-axis) and throughput TPS (red, right y-axis) on LLaDA-1.5. The dashed lines (yellow) mark the final settings $\tau_{\text{edge}} = 0.07$, $\tau_{\text{induced}} = 0.70$, $\tau_{\text{sink}} = 0.01$.

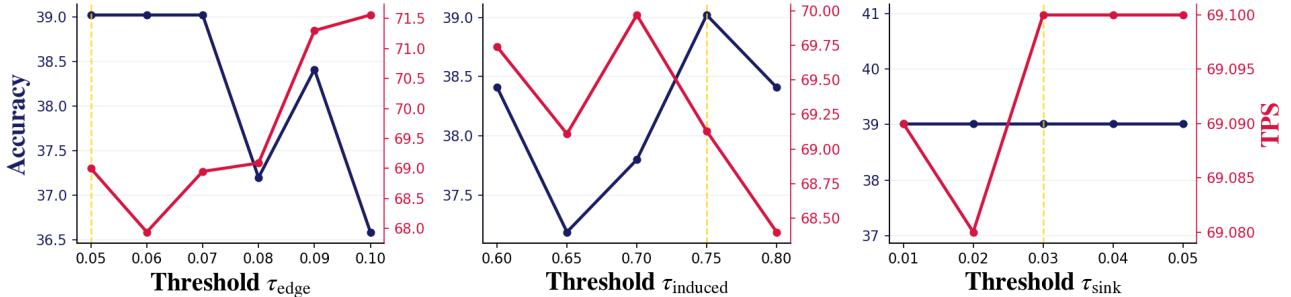


Figure 9. We vary the threshold τ_{edge} , τ_{induced} , τ_{sink} and report accuracy (blue, left y-axis) and throughput TPS (red, right y-axis) on Dream-v0-Base-7B. The dashed lines (yellow) mark the final settings $\tau_{\text{edge}} = 0.05$, $\tau_{\text{induced}} = 0.75$, $\tau_{\text{sink}} = 0.03$.

In most cases, we observe a clear trade-off between throughput and accuracy: higher throughput is generally achieved at the cost of reduced accuracy. To balance this trade-off, we select hyperparameter values that lie near the Pareto frontier,

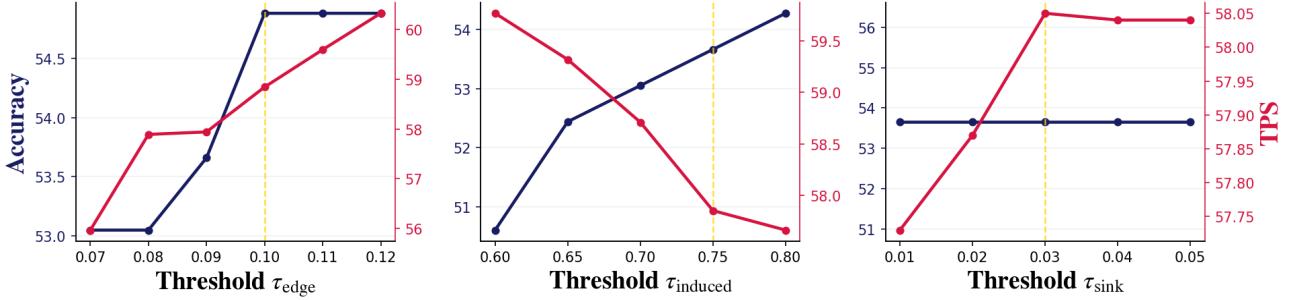


Figure 10. We vary the threshold τ_{edge} , τ_{induced} , τ_{sink} and report accuracy (blue, left y-axis) and throughput TPS (red, right y-axis) on Dream-v0-Instruct-7B. The dashed lines (yellow) mark the final settings $\tau_{\text{edge}} = 0.10$, $\tau_{\text{induced}} = 0.75$, $\tau_{\text{sink}} = 0.03$.

favoring configurations that preserve high accuracy while providing meaningful efficiency gains. For example, on the LLaDA-8B-Instruct model, we select $\tau_{\text{edge}} = 0.07$, which achieves the highest accuracy among the tested values while offering higher throughput compared to $\tau_{\text{edge}} = 0.08$. This criterion ensures that the chosen configuration does not sacrifice accuracy for marginal speed improvements.

For the the lower confidence threshold τ_{low} hyperparameter, we perform a grid search over both models and benchmarks, resulting in 16 experimental settings in total. Within each setting, we further tune the method-specific hyperparameters to obtain the best-performing configuration. The final selected results are reported in Table 3.

Table 3. Final hyperparameter settings for DAWN, where rows denote models and columns denote hyperparameter types. The τ_{induced} , τ_{sink} , τ_{edge} are shared across benchmarks for each model, while τ_{low} adopt benchmark-specific configurations.

Model	τ_{induced}	τ_{sink}	τ_{edge}	τ_{low}			
				GSM8K	MATH	HumanEval	MBPP
LLaDA-8B-Instruct	0.70	0.01	0.07	0.75	0.75	0.8	0.7
LLaDA-1.5	0.70	0.01	0.07	0.75	0.75	0.8	0.75
Dream-v0-Base-7B	0.75	0.03	0.05	0.75	0.8	0.8	0.8
Dream-v0-Instruct-7B	0.75	0.03	0.10	0.8	0.8	0.8	0.8

B. Discussion

This work is primarily motivated by the challenge of nonindependent position predictions in dLLMs, which suggests that decoding order should account for dependencies among positions. In many existing analyses and methods, high confidence is often treated as a sufficient indicator of per-position consistency or approximate independence, and the resulting update rules largely rely on such conservative criteria, which can limit achievable parallelism and leave a substantial portion of dLLM parallel potential underexploited. In contrast, this work attempts to capture positional coupling more directly and to derive corresponding decoding strategies from these dependency relations.

Viewing dLLM decoding through the lens of positional dependencies provides a complementary perspective that can more directly address the persistent difficulty of parallel decoding under non-independence. It is hoped that this study will encourage further investigation into dependency structures in dLLMs and inspire more efficient inference methods that leverage them.