IMPLICIT EOS DENSITY FOR ONLINE BIDIRECTIONAL LENGTH ADAPTATION IN MASKED DIFFUSION LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Masked diffusion LLMs (e.g., LLaDA) typically rely on a fixed generation length (i.e., a fixed `MASK` canvas), struggling to accommodate responses whose required lengths vary substantially across inputs: a short length may constrain content generation, while a long length wastes computation. We study the denoising dynamics and find that the evolving implicit density (ρ) of end-of-sequence (`EOS`) tokens during denoising reveals whether the current length is excessive or insufficient. Building on this observation, we propose **ρ-EOS**, a training-free, online bidirectional length adaptation strategy for masked diffusion LLMs: low density triggers expansion, while high density triggers contraction. This enables online length adaptation within the original denoising loop, without retraining or a separate length-adjustment stage. Extensive experiments on mathematical reasoning and code generation tasks show that **ρ-EOS** yields a more favorable Pareto frontier in the quality-efficiency trade-off. Additionally, because **ρ-EOS** targets online length adaptation, it is orthogonal to pruning, caching, and other acceleration-oriented decoding strategies, making it potentially compatible with and complementary to them.

1 INTRODUCTION

Large language models Brown et al. (2020); Bai et al. (2023); Touvron et al. (2023); Achiam et al. (2023) have exhibited scaling laws and versatile capabilities in language understanding and generation. Recently, diffusion large language models (dLLMs) Nie et al. (2025); Khanna et al. (2025); DeepMind (2025); Song et al. (2025) have emerged as a promising alternative paradigm, attracting increasing attention. In contrast to next-tokens prediction paradigm, masked dLLMs formulate text generation as a discrete probabilistic denoising process. Starting from a fully noisy sequence (i.e., `MASK` canvas), dLLMs iteratively unmask `MASK` tokens in parallel, which enables global context modeling and flexible generation orders.

Despite these advantages, masked dLLMs remain constrained by a fixed-length inference paradigm. Specifically, decoding starts from a fixed number of `MASK` tokens, which determines the available generation length (i.e., the `MASK` canvas) for subsequent denoising. In practice, this fixed generation length can mismatch the semantically required response length, particularly for long-chain reasoning Lightman et al. (2023); Muennighoff et al. (2025) or tasks that require adaptive switching between short and long thinking modes (hybrid reasoning) Anthropic (2025a;b). If the initial `MASK` length is insufficient, the model may lack sufficient budget to complete complex tasks. Conversely, an overly long generation length wastes computation and often yields low token utilization on easy tasks. This rigid fixed-length setup therefore creates an inherent trade-off between quality and cost, in contrast to autoregressive models, which can naturally continue generation token by token. To address this issue, DAEDAL Li et al. (2025a) proposes a training-free two-stage solution that supports only unidirectional expansion through separate initial length adjustment and iterative mask insertion phases. Its first iteratively expands an initial short `MASK` canvas to a coarse target length by checking the confidence of trailing `EOS` tokens. Then the second stage performs iterative `MASK` insertion to further expand low-confidence regions. This two-stage (offline adaptation) pipeline introduces additional inference overhead and does not support length contraction, which becomes limiting when the initial canvas is overly long or when earlier expansion decisions need to be corrected.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

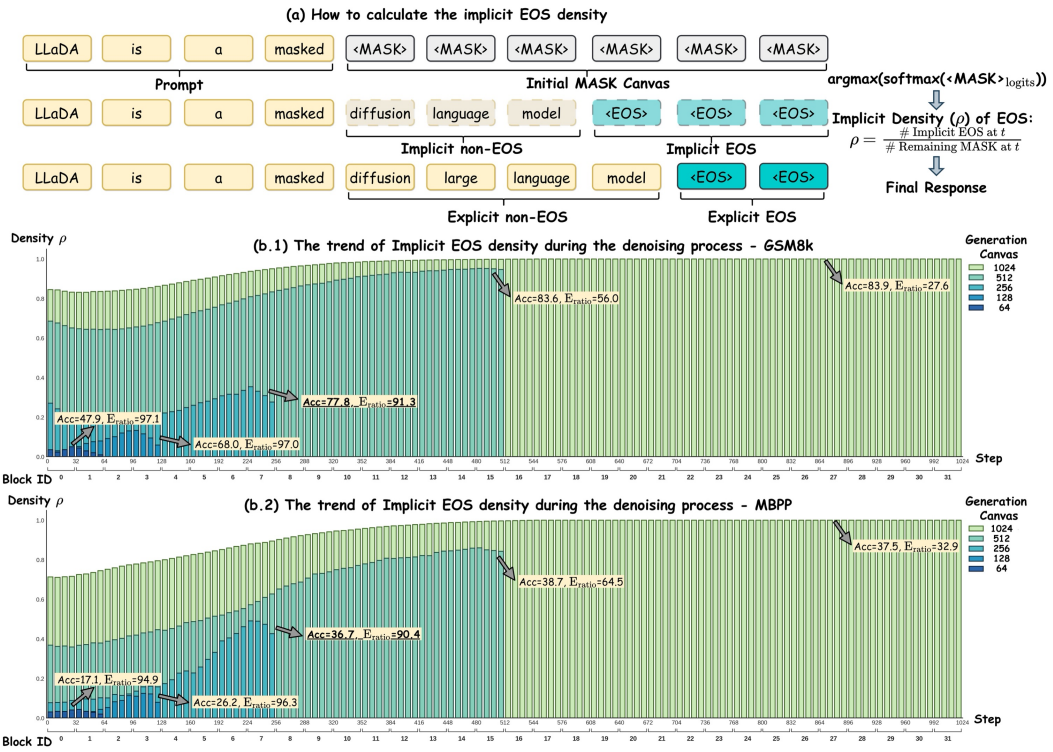


Figure 1: **The Evolution Trend of Implicit `EOS` Density ($\rho_{\texttt{EOS}}$) During the Denoising Process.** (a) Illustration of the calculation for $\rho_{\texttt{EOS}}$. (b.1) Trends of $\rho_{\texttt{EOS}}$ on GSM8k. (b.2) Trends of $\rho_{\texttt{EOS}}$ on MBPP.

To mitigate these limitations, we revisit the denoising dynamics of masked dLLMs and identify an implicit signal of length sufficiency. As illustrated in Figure 1, under different fixed generation canvases, the implicit `EOS` density exhibits distinguishable trends throughout the denoising process. When the implicit `EOS` density in the early denoising phase is very low, it indicates that the remaining generation length is insufficient. In contrast, when the density in the early denoising phase becomes high, it indicates that the current length is excessive and contains surplus generation capacity. When the length is appropriate, the density reaches an intermediate equilibrium value. Crucially, this signal emerges naturally during denoising and reflects the model's internal assessment of length sufficiency. Motivated by this observation, we introduce **ρ-EOS**, a training-free, online adaptation strategy for bidirectional length adaptation, which jointly performs denoising and length adaptation by monitoring the implicit `EOS` density.

Extensive experiments show that **ρ-EOS** achieves a superior Pareto frontier in the quality–efficiency trade-off compared to fixed-length baselines and DAEDAL. Furthermore, compared to DAEDAL, **ρ-EOS** supports bidirectional length adaptation within a single denoising loop and remains compatible with acceleration-oriented decoding methods (e.g., pruning Chen et al. (2025), caching Wu et al. (2025a), or early-exit Li et al. (2025b)), indicating its greater generality.

2 Related Work

Diffusion Large Language Models. Recent work has extended diffusion model to discrete domain for text generation Austin et al. (2021a); He et al. (2023); Gong et al. (2022); Lou et al. (2023); Zheng et al. (2024); Sahoo et al. (2024). A prominent paradigm is masked diffusion language modeling (i.e., absorbing-state), which iteratively denoise a `MASK` canvas Nie et al. (2025); Zhu et al. (2025a;b); Gong et al. (2024); Ye et al. (2025). However, it typically requires a predefined fixed generation length, which can lead to either redundant computation or degraded quality Li et al. (2025a). Another line of dLLMs, known as block diffusion language models Arriola et al. (2025); Cheng et al. (2025); Liu et al. (2025), supports variable-length generation by adopting a hybrid

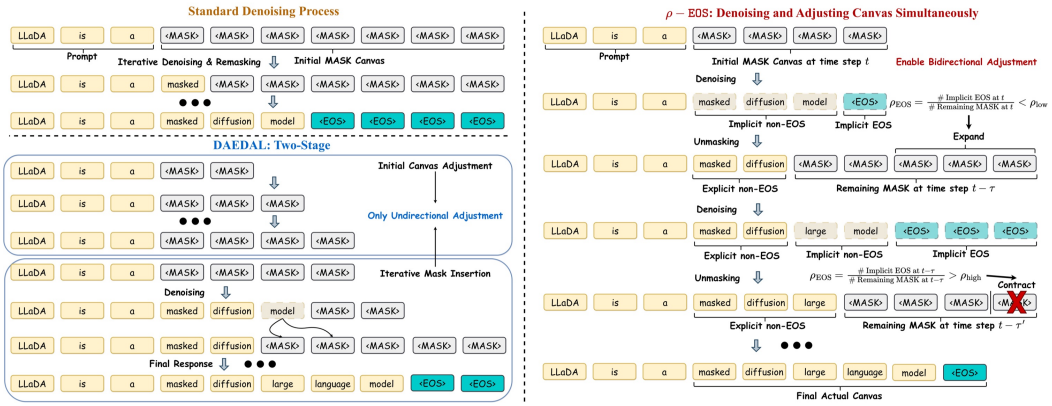


Figure 2: **The Formulation of Implicit and Explicit Concept for Tokens and Density. (Upper Left)** The standard denoising process of masked dLLMs (e.g., LLaDA). **(Lower Left)** The two-stage process of DAEDAL. **(Right)** The bidirectional variable-length denoising process of ρ-**EOS**.

paradigm that divides a sequence into blocks, performs discrete diffusion within each block, and conditions autoregressively on all preceding blocks. In this work, we focus on masked dLLMs and aim to address fixed-length limitation.

Variable-Length Strategies for Masked Diffusion LLMs. To alleviate the fixed-length constraint of masked diffusion LLMs, recent works have explored several approaches. For instance, FlexMDM Kim et al. (2025) decomposes the denoising process into insertion and unmasking steps, learning to predict the number of new `MASK` tokens to insert before each position. DreamOn Wu et al. (2025b) introduces special control tokens, where `EXPAND` splits into two `MASK` tokens to extend the sequence, and `DELETE` is removed to shorten it. dLLM-Var Yang et al. (2025b) enables masked dLLMs to perform inference in a block diffusion manner by training the models to accurately predict the `EOS` token. In contrast, DAEDAL Li et al. (2025a) proposes a training-free alternative that decouples length adjustment from denoising. It first performs length expansion then follow with adaptive denoising. While it improves inference efficiency to some extent, its two-stage design potentially introduces additional latency. Moreover, it supports only unidirectional length expansion, which limits flexibility. Beyond that, length flexibility also plays a critical role in reinforcement learning (RL) fine-tuning, where diverse reasoning depths and response structures can facilitate exploration Zhao et al. (2025); Gong et al. (2025); Yang et al. (2025a); Wang et al. (2025). These considerations further motivate the need for a single-stage, training-free, and bidirectional control strategy.

3 Methods

3.1 Overview

To address the fixed-length limitation in standard masked dLLMs inference, we introduce ρ-**EOS**, a training-free, online bidirectional length adaptation strategy. Our method enables dLLMs to adapt their generation length (i.e., the `MASK` canvas) *on the fly* during denoising, rather than relying on a fixed predefined length or separate adjustment.

3.2 Implicit Token

We first clarify the concept of **Implicit Token**. At each denoising step, the input $\mathbf{x}_t$ at time step t is fed into the model $p_\theta(\cdot)$ to obtain the logits, denoted as $\mathbf{x}_t^{\text{logits}}$. These logits are then processed by

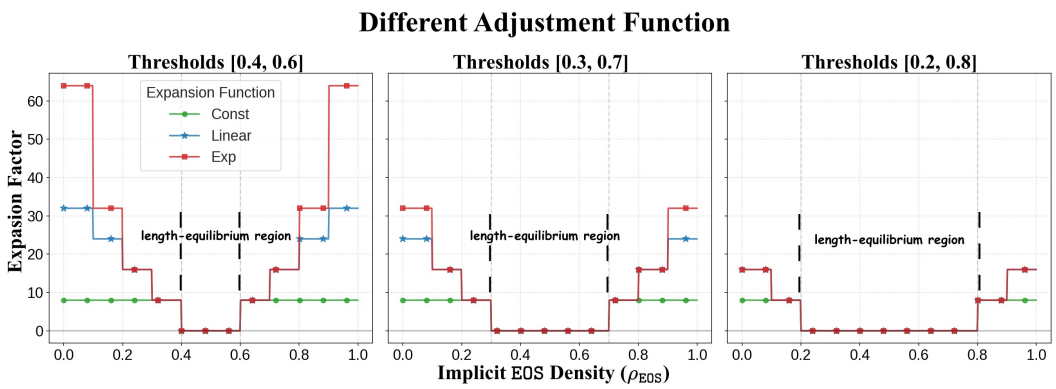


Figure 3: Examples of piecewise adjustment factor function with constant, linear, and exponential forms.

$\text{argmax}(\text{softmax}(\cdot))$, yielding the tokens:

$$\mathbf{x}_t^{\text{logits}} = p_\theta(\mathbf{x}_t), \tag{1}$$

$$\hat{\mathbf{x}}_0 = \text{argmax}(\text{softmax}(\mathbf{x}_t^{\text{logits}})), \tag{2}$$

$$\mathbf{x}_{t-\tau} = \text{Remask}(\hat{\mathbf{x}}_0), \tag{3}$$

where $p_\theta(\cdot)$ denotes the model, and $\hat{\mathbf{x}}_0$ represents the model's prediction of the clean sequence. We refer to $\hat{\mathbf{x}}_0$ as the implicit $\mathbf{x}_0$ because it is not immediately committed as the final response. Moreover, the tokens in $\hat{\mathbf{x}}_0$ that correspond to the remaining `MASK` positions in $\mathbf{x}_t$ are defined as the **Implicit Tokens** at time step t (i.e., $\{i|\hat{\mathbf{x}}_0^i \wedge (\mathbf{x}_t^i = \texttt{MASK})\}$). After applying Remask$(\cdot)$, only a subset of $\hat{\mathbf{x}}_0$ is decoded into **Explicit Tokens**, while the remaining tokens in $\hat{\mathbf{x}}_0$ are remasked to form $\mathbf{x}_{t-\tau}$. This iterative process continues until a fully denoised sequence is obtained, which constitutes the standard inference procedure of masked dLLMs. A schematic illustration is shown in Figure 2.

3.3 Implicit EOS Density

Based on the notion of implicit tokens, we define **Implicit Density** (ρ) as the ratio between implicit tokens with specific properties and remaining masked tokens at time step t. As depicted in Figure 2, the implicit token density is defined as:

$$\rho = \frac{\text{\# Implicit Token at } t}{\text{\# Remaining } \texttt{MASK} \text{ Token at } t}, \tag{4}$$

where # Implicit Token at t denotes the number of implicit tokens with certain properties at time step t, and # Remaining MASK Token at t represents the number of remaining `MASK` tokens in $\mathbf{x}_t$ that still need to be denoised. Accordingly, we define the implicit `EOS` density as:

$$\rho_{\texttt{EOS}} = \frac{\text{\# Implicit } \texttt{EOS} \text{ Token at } t}{\text{\# Remaining } \texttt{MASK} \text{ Token at } t}. \tag{5}$$

3.4 Bidirectional Length Adjustment

Figure 1 shows the trend of $\rho_{\texttt{EOS}}$ on GSM8K and MBPP under fixed generation lengths from 64 to 1024, and masked dLLMs implicitly encode length-sufficiency information in $\rho_{\texttt{EOS}}$. When the generation length is sufficient to complete the response, the model assigns higher probability mass to `EOS`, resulting in a larger $\rho_{\texttt{EOS}}$ (e.g., 512 or 1024). Conversely, when the available length is insufficient (e.g., 64 or 128), the model tends to fully utilize the masked space for semantic content, leading to a persistently low $\rho_{\texttt{EOS}}$. When the available length is appropriate, the implicit `EOS` density reaches an intermediate **equilibrium state**, achieving a better trade-off between quality and token utilization. This observation suggests that $\rho_{\texttt{EOS}}$ serves as a reliable internal signal for assessing whether the current generation length is appropriate. To apply this criterion robustly, we empirically define an **equilibrium region** using two thresholds ρ_{low} and ρ_{high}. When $\rho_{\texttt{EOS}} \in [\rho_{\text{low}}, \rho_{\text{high}}]$, the

Algorithm 1 The Inference Process of ρ-**EOS**

Input: Prompt $\mathbf{p}$, model p_θ, initial/max canvas $L_{\text{init}}/L_{\text{max}}$, $[\rho_{\text{low}}, \rho_{\text{high}}]$, confidence threshold τ_{high}, $\mathrm{E}_{\text{factor}}(\cdot)$, max adjustment step N
Output: clean sequence $\mathbf{x}$
$L_{\text{cur}} \leftarrow L_{\text{init}}$, $n \leftarrow 0$
$\mathbf{x}^t \leftarrow [\mathbf{p}, \underbrace{\texttt{MASK}, \ldots, \texttt{MASK}}_{L_{\text{init}}}]$ ▷ Initialization
while ContainsMask($\mathbf{x}_t$) **do**
  $\mathbf{x}_t^{\text{logits}} \leftarrow p_\theta(\mathbf{x}_t)$
  $\rho_{\texttt{EOS}} \leftarrow \text{ImplicitEOSDensity}(\mathbf{x}_t^{\text{logits}})$
  $\mathbf{x}_t^{\text{conf}}, \hat{\mathbf{x}}_0 \leftarrow \text{Decode}(\mathbf{x}_t^{\text{logits}})$
  for $i \in \{\mathbf{x}_t^i = \texttt{MASK} \wedge \mathbf{x}_t^{\text{conf},i} > \tau_{\text{high}}\}$ **do**
    $\mathbf{x}_{t-\tau}^i \leftarrow \hat{\mathbf{x}}_0^i$
  end for
  if $0 < L_{\text{cur}} < L_{\text{max}}$ and $n < N$ **then**
    if $\rho_{\texttt{EOS}} \in [\rho_{\text{low}}, \rho_{\text{high}}]$ **then**
      $\mathbf{x}_t, L_{\text{cur}} \leftarrow \mathbf{x}_{t-\tau}, L_{\text{cur}}$ ▷ Hold L_{cur}
    else if $\rho_{\texttt{EOS}} < \rho_{\text{low}}$ **then** ▷ Expand L_{cur}
      $\mathbf{x}_{t-\tau} \leftarrow \mathbf{x}_{t-\tau} \oplus \underbrace{[\texttt{MASK}, \ldots, \texttt{MASK}]}_{\mathrm{E}_{\text{factor}}(\rho_{\texttt{EOS}})}$
      $L_{\text{cur}} \leftarrow L_{\text{cur}} + \mathrm{E}_{\text{factor}}(\rho_{\texttt{EOS}})$
    else ▷ Contract L_{cur}
      $\mathbf{x}_{t-\tau} \leftarrow \mathbf{x}_{t-\tau}^{0:(|\mathbf{x}| - \mathrm{E}_{\text{factor}}(\rho_{\texttt{EOS}}))}$
      $L_{\text{cur}} \leftarrow L_{\text{cur}} - \mathrm{E}_{\text{factor}}(\rho_{\texttt{EOS}})$
    end if
  end if
  $n \leftarrow n + 1$
end while
return $\mathbf{x}$

model suspends length modification and focuses solely on denoising, enabling stable bidirectional control without oscillatory behavior. When $\rho_{\texttt{EOS}} < \rho_{\text{low}}$, the remaining masked space is insufficient and length expansion is triggered via MASK insertion. Conversely, when $\rho_{\texttt{EOS}} > \rho_{\text{high}}$, the model indicates surplus generation capacity, and the sequence is contracted by removing partial trailing MASK.

Formally, the sequence update under bidirectional length adjustment is defined as:

$$\mathbf{x} = \begin{cases} \mathbf{x} \leftarrow \mathbf{x} \oplus \underbrace{[\texttt{MASK}, \ldots, \texttt{MASK}]}_{\mathrm{E}_{\text{factor}}(\rho_{\texttt{EOS}})} & \text{if } \rho_{\texttt{EOS}} < \rho_{\text{low}} \\ \mathbf{x} \leftarrow \mathbf{x}^{0:(|\mathbf{x}| - \mathrm{E}_{\text{factor}}(\rho_{\texttt{EOS}}))} & \text{else if } \rho_{\texttt{EOS}} > \rho_{\text{high}} \\ \mathbf{x} & \text{otherwise} \end{cases} \tag{6}$$

with the update to current generation length L_{cur}:

$$L_{\text{cur}} = \begin{cases} L_{\text{cur}} + \mathrm{E}_{\text{factor}}(\rho_{\texttt{EOS}}) & \text{if } \rho_{\texttt{EOS}} < \rho_{\text{low}} \\ L_{\text{cur}} - \mathrm{E}_{\text{factor}}(\rho_{\texttt{EOS}}) & \text{else if } \rho_{\texttt{EOS}} > \rho_{\text{high}} \\ L_{\text{cur}} & \text{otherwise} \end{cases} \tag{7}$$

Beyond the adjustment rule itself, we further investigate how aggressively the generation length should be adjusted through different **adjustment-factor functions** $\mathrm{E}_{\text{factor}}(\cdot)$. DAEDAL Li et al. (2025a) adopts a constant expansion factor, a design that is simple and stable but leads to multiple adjustment iterations when the initial length is severely mismatched. Motivated by Figure 1, we

Table 1: **Results on LLaDA-Instruct-8B across Four Benchmarks.** We compare the ρ-**EOS** against DAEDAL and baselines. The best setting for the baseline is **orange**. DAEDAL is **blue**, and ρ-**EOS** is **red**. Under the ρ-**EOS** setting, Sym and Asym denote the symmetric and asymmetric lower-upper threshold centered around 0.5. The **best** results are **bold**, and the second-best results are underlined.

Benchmark	Metric	Fixed-Length Denoising (Baseline)						DAEDAL	ρ-EOS	ρ-EOS
		64	128	256	512	1024	2048	64	64 Sym	64 Asym
GSM8K	$Acc\uparrow$	47.9	68.0	77.8	83.6	83.9	82.7	**84.6**	81.8	84.2
	E_{token}	62.1	124.1	233.7	286.8	282.9	297.3	246.0	198.4	252.5
	N_{token}	64	128	256	512	1024	2048	331.0	231.8	361.0
	$E_{\text{ratio}}\uparrow$	97.1%	97.0%	91.3%	56.0%	27.6%	14.5%	74.5%	**85.6%**	70.0%
	$T_{\text{runtime}}\downarrow$	174	360	860	2458	8238	30305	1090	**645**	823
MATH500	$Acc\uparrow$	24.0	29.0	35.4	38.4	40.4	40.0	40.0	37.0	**40.6**
	E_{token}	61.7	123.3	244.7	423.9	585.8	706.5	464.6	428.8	558.7
	N_{token}	64	128	256	512	1024	2048	618.3	486.3	682.2
	$E_{\text{ratio}}\uparrow$	96.4%	96.3%	95.6%	82.8%	57.2%	34.5%	75.2%	**88.2%**	81.9%
	$T_{\text{runtime}}\downarrow$	98	208	462	1202	3663	12764	2207	**1398**	1968
MBPP	$Acc\uparrow$	21.0	28.8	36.7	38.7	37.5	38.8	39.4	39.6	**40.6**
	E_{token}	61	122	232	331	335	336	308.4	324.0	335.9
	N_{token}	64	128	256	512	1024	2048	558.3	507.4	594.4
	$E_{\text{ratio}}\uparrow$	94.9%	95.5%	90.4%	64.5%	32.9%	16.3%	55.2%	**63.9%**	55.5%
	$T_{\text{runtime}}\downarrow$	176	332	713	1728	4760	15172	3551	2031	**2030**
HUMANEVAL	$Acc\uparrow$	17.1	26.2	36.6	45.1	45.1	**48.2**	44.5	43.3	43.9
	$E_{\text{token}}\uparrow$	59.6	125	247.3	462	631.1	691.1	494.7	570.4	643.1
	N_{token}	64	128	256	512	1024	2048	813.0	654.0	769.9
	$E_{\text{ratio}}\uparrow$	93.2%	97.7%	96.6%	90.2%	61.6%	33.8%	64.8%	**88.4%**	83.5%
	$T_{\text{runtime}}\downarrow$	111	230	543	1474	4569	16046	1283	**580**	593

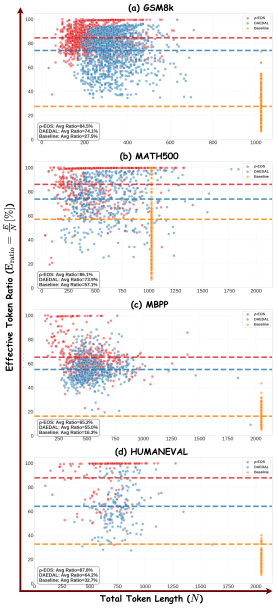


Figure 4: E_{ratio} **per sample.**

additionally consider linear and exponential designs, as depicted in Figure 3. The piecewise linear function scales the adjustment magnitude proportionally to the distance between ρ_{EOS} and the boundaries of the equilibrium region. The piecewise exponential function further amplifies this effect when ρ_{EOS} is extremely low or high, enabling more aggressive expansion or contraction. The key motivation is to rapidly drive the generation into the stable ρ_{EOS} region, so that the model can enter the pure denoising phase early without repeatedly interleaving length adjustment steps. Compared to unidirectional expansion, bidirectional control improves flexibility and token utilization. As illustrated in Figure 2, if the generation length becomes overly large after expansion, the proposed mechanism can contract it, leading to more efficient and stable generation.

3.5 Online Bidirectional Variable-Length Adaptation Denoising

Overall, by integrating implicit EOS density estimation with bidirectional length adjustment, ρ-**EOS** realizes **training-free** and **bidirectional variable-length** generation within a unified single-stage denoising process, i.e., it jointly performs unmasking and length adjustment. At each denoising step, the model $p_\theta(\cdot)$ performs a forward pass on $\mathbf{x}_t$ to obtain $\mathbf{x}_t^{\text{logits}}$ and the implicit prediction $\hat{\mathbf{x}}_0$ (Equation 3). After applying the $\text{Remask}(\cdot)$ operation to produce $\mathbf{x}_{t-\tau} = \text{Remask}(\hat{\mathbf{x}}_0)$, we compute ρ_{EOS} according to Equation 5 and adjust the current generation length using the bidirectional strategy in Equations 6 and 7. This step-level denoising process is repeated until denoising is complete. By coupling denoising and length adjustment within a single inference loop, ρ-**EOS** enables dynamic expansion when additional reasoning space is needed and contraction when redundancy is detected, without introducing extra stages. Algorithm 1 provides a detailed description of the full procedure.

4 Experiments

4.1 Experimental Setups

Implementation Details. We follow the settings of DAEDAL Li et al. (2025a) to ensure fairness and reproducibility. To reduce randomness from a single evaluation run, each experiment is repeated three times, and we report the averaged results.

Benchmarks and Metrics. To comprehensively evaluate the effectiveness of ρ-**EOS**, we conduct experiments on four benchmarks covering mathematics and code tasks. For mathematical reasoning,

Table 2: **Results on LLaDA-1.5-8B across Four Benchmarks.** We compare the ρ-**EOS** against DAEDAL and baselines. The best setting for the baseline is **orange**. DAEDAL is **blue**, and ρ-**EOS** is **red**. Under the ρ-**EOS** setting, Sym and Asym denote the symmetric and asymmetric lower-upper threshold centered around 0.5. The **best** results are **bold**, and the second-best results are underlined.

Benchmark	Metric	Fixed-Length Denoising (Baseline)						DAEDAL	ρ-EOS	ρ-EOS
		64	128	256	512	1024	2048	64	64 Sym	64 Asym
GSM8K	Acc ↑	49.9	71.2	80.5	83.6	83.9	84.2	**85.1**	83.3	84.5
	E_{token}	62.4	124.8	236.7	292.6	287	293.9	254.7	214.5	231.6
	N_{token}	64	128	256	512	1024	2048	348.3	254.3	362.4
	E_{ratio} ↑	97.1%	97.0%	91.2%	56.0%	27.7%	14.4%	72.9%	**84.4%**	63.9%
	T_{runtime} ↓	149	335	844	2488	8450	30875	1056	**721**	990
MATH500	Acc ↑	22.8	30.2	35.2	39.2	**42.8**	39.4	42.6	38.2	41.0
	E_{token}	62.1	124.8	246.5	428.6	583.5	713.7	489.1	339.5	583.4
	N_{token}	64	128	256	512	1024	2048	662.6	373.1	711.3
	E_{ratio} ↑	97.0%	97.5%	96.3%	83.7%	57.0%	34.9%	73.8%	**91.0%**	82.0%
	T_{runtime} ↓	97	192	444	1177	3616	12610	2460	850	**637**
MBPP	Acc ↑	21.2	30.4	39.2	38.8	39.4	39.4	38.2	**40.8**	**40.8**
	E_{token}	60.9	123.7	238.2	347.3	343.5	356.2	321.8	336.1	334.1
	N_{token}	64	128	256	512	1024	2048	587.1	535.6	540.8
	E_{ratio} ↑	95.1%	96.6%	93.0%	67.8%	33.6%	17.4%	54.8%	**62.8%**	61.8%
	T_{runtime} ↓	156	313	690	1689	4688	14895	4124	3001	**2565**
HUMANEVAL	Acc ↑	17.6	20.7	38.4	45.1	**49.4**	48.8	43.9	44.5	44.5
	E_{token}	60.3	124.9	251.7	473.8	672.1	761.6	523	460.8	686.6
	N_{token}	64	128	256	512	1024	2048	806.4	509.7	809.5
	E_{ratio} ↑	94.2%	97.5%	98.3%	92.5%	65.6%	37.2%	64.4%	**90.4%**	84.8%
	T_{runtime} ↓	47	78	162	406	1212	4187	743	**433**	647

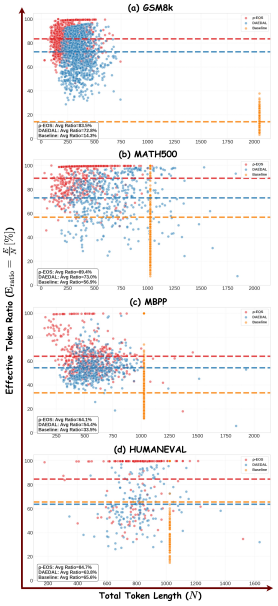


Figure 5: E_{ratio} **per sample.**

we evaluate on GSM8K (Cobbe et al., 2021) and MATH500 (Lightman et al., 2023). For code generation, we employ MBPP(Austin et al., 2021b) and HumanEval(Chen et al., 2021). Accuracy (Acc) is adopted as the evaluation metric. In addition, we introduce four extra metrics: the total number of generated tokens (N_{token}), the number of effective tokens after removing EOS padding (E_{token}), the effective token ratio ($E_{\text{ratio}} = \frac{E_{\text{token}}}{N_{\text{token}}}$), and the evaluation runtime (T_{runtime}, in seconds).

4.2 Main Results

Table 1 and Table 2 compare the fixed-length baselines, DAEDAL, and ρ-**EOS**. The baselines are evaluated under six predefined generation lengths ranging from 64 to 2048. Both DAEDAL and ρ-**EOS** employ $L_{\text{init}} = 64$, ensuring a fair comparison.

Performance of ρ-EOS. In Table 1 and Table 2, ρ-**EOS** achieves performance comparable to DAEDAL and the best-performing baselines across most benchmarks. On MATH500 and MBPP in Table 1, ρ-**EOS** with $L_{\text{init}} = 64$ Asym outperforms both DAEDAL and the best baseline, while consistently demonstrating higher efficiency. On MBPP and HumanEval in Table 2, ρ-**EOS** with $L_{\text{init}} = 64$ Asym also outperforms DAEDAL on accuracy and efficiency. More importantly, both the E_{ratio} and the T_{runtime} consistently compare favorably DAEDAL and the strongest baselines on nearly all tasks.

Token utilization and efficiency of ρ-EOS. For the best-performing baselines, higher accuracy often comes at the cost of sharply increased T_{runtime} and substantially reduced E_{ratio}. In particular, results of MBPP-2048, HumanEval-2048 in Table 1, and GSM8k-2048 in Table 2 exhibit extremely low E_{ratio}, accompanied by a dramatic increase in T_{runtime}. While DAEDAL alleviates this inefficiency to some extent, ρ-**EOS** goes a step further. By enabling bidirectional length adjustment, ρ-**EOS** consistently achieves higher E_{ratio} than both DAEDAL and the best-performing baselines at comparable accuracy levels. Figure 4 and Figure 5 visualize the per-sample distribution of N_{token} and the corresponding E_{ratio}. Across all tasks, a substantial portion of the samples generated by ρ-**EOS** achieve an E_{ratio} close to 100%, approaching autoregressive counterparts. Such behavior is rarely observed in DAEDAL or fixed-length baselines.

4.3 Robustness and Ablation Studies

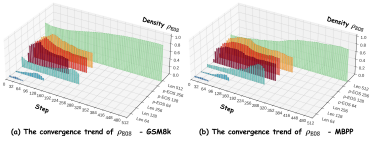
Robustness to the initial length and favorable Pareto frontier. Table 3 demonstrates that ρ-**EOS** is robust to the choice of the initial generation length. When varying L_{init} from 128 to 1024, ρ-**EOS**

Table 3: **Different-Initial Length Results on LLaDA-Instruct-8B and LLaDA-1.5-8B.** We compare ρ-**EOS** against DAEDAL under initial lengths from 128 to 1024. The **best** results are **bold**, and the second results are underlined.

Benchmark	Metric	LLaDA Various-Init-Length										LLaDA1.5 Various-Init-Length									
		128		256		512		1024		Average		128		256		512		1024		Average	
		DAEDAL	ρ-EOS	DAEDAL	ρ-EOS	DAEDAL	ρ-EOS	DAEDAL	ρ-EOS	DAEDAL	ρ-EOS	DAEDAL	ρ-EOS	DAEDAL	ρ-EOS	DAEDAL	ρ-EOS	DAEDAL	ρ-EOS	DAEDAL	ρ-EOS
GSM8K	$\boldsymbol{Acc}\uparrow$	84.6	84.6	84.0	83.7	85.3	84.4	84.8	84.8	**84.7**	84.4	85.1	85.2	85.0	83.8	84.2	84.3	85.9	84.9	**85.1**	84.5
	$\boldsymbol{E}_{\text{token}}$	246.3	231.8	250.2	284.7	286.8	288.6	281.3	282.6	266.2	271.9	254.7	257.8	256.9	261.0	275.7	293.7	284.0	292.5	267.8	276.3
	$\boldsymbol{N}_{\text{token}}$	330.8	313.0	341.1	336.4	530.0	438.6	1040.0	666.8	560.5	438.7	349.3	370.5	356.4	321.0	531.4	474.2	1040.0	398.0	569.3	390.9
	$\boldsymbol{E}_{\text{ratio}}\uparrow$	74.5%	74.1%	73.4%	84.6%	50.9%	65.8%	27.0%	42.4%	56.5%	**66.7%**	72.9%	69.6%	72.1%	81.3%	51.9%	61.9%	27.3%	73.5%	56.1%	**71.6%**
	$\boldsymbol{T}_{\text{runtime}}\downarrow$	1035	835	1014	834	1090	1044	5656	1809	2199.0	**1131.0**	1093	990	1008	638	1223	1122	5476	1281	2200.0	**1007.8**
MATH500	$\boldsymbol{Acc}\uparrow$	42.6	40.0	43.2	40.4	41.8	42.8	41.6	39.8	**42.3**	40.8	42.6	40.6	43.0	41.2	41.4	42.0	40.0	40.6	**41.8**	41.1
	$\boldsymbol{E}_{\text{token}}$	464.6	410.8	465.7	492.2	479.0	556.6	541.4	588.4	487.7	512.0	489.1	528.1	490.0	559.2	499.4	586.7	534.3	602.3	503.2	569.1
	$\boldsymbol{N}_{\text{token}}$	618.4	480.8	621.0	549.2	686.8	642.0	1058.8	874.1	746.3	636.5	662.6	624.1	664.2	664.3	717.7	692.9	1064.9	792.1	777.4	693.4
	$\boldsymbol{E}_{\text{ratio}}\uparrow$	75.1%	89.2%	75.0%	89.5%	69.7%	86.7%	51.1%	67.3%	67.7%	**83.2%**	73.8%	84.6%	73.8%	84.2%	69.6%	84.7%	50.2%	76.0%	66.9%	**82.4%**
	$\boldsymbol{T}_{\text{runtime}}\downarrow$	2199	1062	2175	1543	2096	1980	6543	1938	3253.3	**1630.8**	2447	1805	2423	1912	2345	2074	2315	2199	2382.5	**1997.5**
MBPP	$\boldsymbol{Acc}\uparrow$	39.4	39.0	39.4	40.0	38.8	39.4	37.8	38.6	38.8	**39.3**	38.2	41.8	38.2	41.4	37.8	39.0	40.0	40.4	38.6	**40.7**
	$\boldsymbol{E}_{\text{token}}$	308.4	341.6	308.4	344.0	313.0	342.5	324.7	345.0	313.6	343.3	321.8	345.1	321.8	350.2	325.4	362.1	331.5	350.7	325.1	352.0
	$\boldsymbol{N}_{\text{token}}$	558.3	605.1	558.3	611.6	604.6	638.3	1046.7	773.5	692.0	657.1	587.1	556.5	587.1	497.7	621.4	662.8	1046.8	943.6	710.6	665.2
	$\boldsymbol{E}_{\text{ratio}}\uparrow$	55.2%	56.5%	55.2%	56.3%	51.8%	53.6%	31.0%	44.6%	48.3%	**52.8%**	54.8%	62.0%	54.8%	70.4%	52.4%	54.6%	31.7%	37.2%	48.4%	**56.1%**
	$\boldsymbol{T}_{\text{runtime}}\downarrow$	3518	2836	3462	2169	3303	2389	5674	2558	3989.3	**2396.0**	4102	2824	4060	1853	3907	2585	4471	2501	4135.0	**2440.8**
HUMANEVAL	$\boldsymbol{Acc}\uparrow$	48.2	48.8	48.2	47.0	48.2	45.1	47.0	47.6	**47.9**	47.1	43.9	47.6	43.9	47.0	43.9	48.2	43.9	49.4	43.9	**48.1**
	$\boldsymbol{E}_{\text{token}}$	497.5	618.0	497.5	613.0	497.3	613.7	540.0	638.8	508.1	620.9	519.2	634.7	519.2	654.5	518.8	623.6	528.4	677.6	521.4	647.6
	$\boldsymbol{N}_{\text{token}}$	769.5	693.5	769.5	698.7	772.4	717.8	1046.9	927.9	839.6	759.5	806.4	734.9	806.4	743.0	808.4	692.2	1049.4	951.5	867.7	780.4
	$\boldsymbol{E}_{\text{ratio}}\uparrow$	64.7%	89.1%	64.7%	87.7%	64.4%	85.5%	51.6%	68.8%	61.4%	**82.8%**	64.4%	86.4%	64.4%	88.1%	64.2%	90.1%	50.4%	71.2%	60.9%	**84.0%**
	$\boldsymbol{T}_{\text{runtime}}\downarrow$	746	528	789	542	725	548	1644	618	976.0	**559.0**	786	590	730	593	705	623	798	606	754.8	**603.0**

Table 4: **Comparison on EOS Density and Confidence and Single Stage v.s. Two Stage.** Ablation on GSM8k and MBPP with LLaDA-Instruct-8B, $L_{\text{init}} = 64$.

Metric	GSM8k				MBPP			
	Single Stage		Two Stage		Single Stage		Two Stage	
	Confidence	Density (ρ)	Confidence	Density (ρ)	Confidence	Density (ρ)	Confidence	Density (ρ)
$\boldsymbol{Acc}\uparrow$	73.7	**84.2**	**84.6**	84.4	40.0	**40.6**	38.2	**40.4**
$\boldsymbol{E}_{\text{token}}$	224.3	252.2	246.0	219.9	340.7	335.9	321.8	335.3
$\boldsymbol{N}_{\text{token}}$	256.6	361.0	331.0	414.9	546.5	594.4	587.1	602.3
$\boldsymbol{E}_{\text{ratio}}\uparrow$	87.4%	70.0%	74.5%	53.0%	62.3%	55.5%	54.8%	55.7%
$\boldsymbol{T}_{\text{runtime}}\downarrow$	888	823	1090	1049	2560	2030	4124	2030



(a) The convergence trend of ρ_{EOS} - GSM8k (b) The convergence trend of ρ_{EOS} - MBPP

Figure 6: **The convergence trend of ρ_{EOS}:** baseline with different fixed-length and ρ-EOS with different initial lengths.

maintains remarkably stable accuracy, while consistently achieving a higher $\boldsymbol{E}_{\text{ratio}}$ and significantly lower $\boldsymbol{T}_{\text{runtime}}$ than DAEDAL across all tasks. In contrast, DAEDAL shows clear sensitivity to the initial length in terms of both $\boldsymbol{E}_{\text{ratio}}$ and $\boldsymbol{T}_{\text{runtime}}$. This robustness stems from its bidirectional length-adjustment mechanism driven by ρ_{EOS}. ρ-**EOS** dynamically contracts or expands the generation length based on whether the current length is insufficient or excessive. As a result, even when initialized with an aggressive or conservative length, the model can rapidly converge toward a task-adaptive length regime. We present the accuracy-latency curves of each method with different (initial) lengths across various benchmarks in Figure 7. ρ-**EOS** Pareto dominates DAEDAL and baselines across most tasks, achieving a more favorable Pareto frontier in the quality-cost trade-off.

Comparison of EOS Confidence and Density as Control Signals, and of Single-Stage vs. Two-Stage Strategies. Table 4 compares EOS confidence and density as length-control signals under both single-stage and two-stage adjustment strategies. Across both GSM8K and MBPP, ρ-EOS consistently outperforms confidence-based control in most settings. As discussed in Yang et al. Yang et al. (2025a), masked dLLMs suffer from the EOS Trap during the early denoising phase, where the confidence of EOS is significantly higher than that of non-EOS tokens. Therefore, using confidence may lead to inaccurate length control during early decoding. In contrast, ρ-**EOS** provides a more reliable estimate of termination likelihood across masked positions. Moreover, the results further demonstrate that implicit EOS density is also an effective signal for two-stage strategy. Overall, these results show that implicit EOS density is a superior signal to EOS confidence.

The convergence trend of ρ-EOS. Figure 6 visualizes the evolution of ρ_{EOS} throughout the denoising. Across different tasks and initial lengths, we observe that, a short fixed-length baselines causes ρ_{EOS} to converge toward 0, while an overly long length causes ρ_{EOS} to converge toward 1. Only when the generation length is appropriate does the ρ_{EOS} converge to an equilibrium region. In contrast, ρ-**EOS** (red, orange, and yellow bars) exhibits a distinct pattern: after a small number of adjustment steps, the estimated ρ_{EOS} stabilizes within the equilibrium region before final decoding.

Ablation on [ρ_{low}, ρ_{high}]. We investigate the sensitivity of ρ-**EOS** to the thresholds $[\rho_{\text{low}}, \rho_{\text{high}}]$. In Table 5, under Sym settings (i.e., symmetrically setting ρ_{low} and ρ_{high} at intervals of 0.1), varying the thresholds over a wide range leads to minor fluctuations in $\boldsymbol{Acc}$ and $\boldsymbol{T}_{\text{runtime}}$, indicating low sensitivity to threshold tuning. Across different thresholds, ρ-**EOS** maintains stable $\boldsymbol{Acc}$. In particular, wider equilibrium regions (e.g., $[0.2, 0.8]$) favor conservative length adjustment, resulting in higher

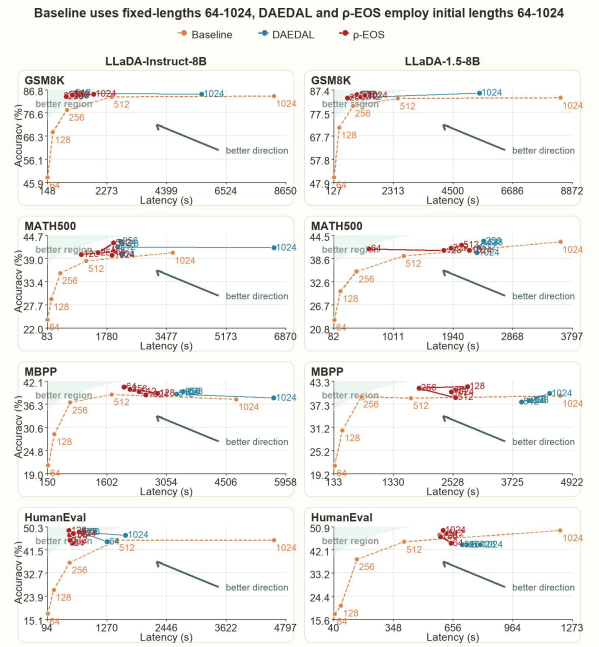


Figure 7: **The Pareto frontier in accuracy and latency trade-off.** Horizontal axis is latency, vertical axis is accuracy.

Table 5: **Ablation of density thresholds** on GSM8K with LLaDA-8B and $L_{\text{init}} = 64$. The best setting ($[\rho_{\text{low}}, \rho_{\text{high}}] = [0.4, 0.8]$) is highlighted in red.

Metric	Density Threshold $[\rho_{\text{low}}, \rho_{\text{high}}]$			
	[0.2,0.8]	[0.3,0.7]	[0.4,0.6]	[0.4,0.8]
$\boldsymbol{Acc}$ ↑	81.8	81.7	81.7	84.2
$\boldsymbol{E}_{\text{token}}$	198.4	210.5	218.4	252.5
$\boldsymbol{N}_{\text{token}}$	231.8	251.2	267.5	361.0
$\boldsymbol{E}_{\text{ratio}}$ ↑	85.6%	83.8%	81.6%	70.0%
$\boldsymbol{T}_{\text{runtime}}$ ↓	645	722	845	823

Table 6: **Ablation of adjustment-factor functions** on GSM8K with LLaDA-Instruct-8B, $L_{\text{init}} = 64$, and $[\rho_{\text{low}}, \rho_{\text{high}}] = [0.4, 0.6]$. $\boldsymbol{Step}$ is the average adjustment steps.

Metric	Adjustment Factor $\mathrm{E}_{\text{factor}}(\cdot)$			
	Base_{256}	Const.	Linear	Exp.
$\boldsymbol{Acc}$ ↑	77.8	79.8	80.9	81.7
$\boldsymbol{E}_{\text{token}}$	233.7	183.4	208.0	218.4
$\boldsymbol{N}_{\text{token}}$	256	213.3	251.0	267.5
$\boldsymbol{E}_{\text{ratio}}$ ↑	91.3%	86.0%	82.9%	81.6%
$\boldsymbol{T}_{\text{runtime}}$ ↓	860	589	757	845
$\boldsymbol{Step}$	–	61.0	54.8	52.4

$\boldsymbol{E}_{\text{ratio}}$ but slightly reduced $\boldsymbol{E}_{\text{token}}$ and $\boldsymbol{T}_{\text{runtime}}$. Conversely, narrower or asymmetric ones allow more aggressive adjustment, increasing $\boldsymbol{E}_{\text{token}}$ at the cost of additional denoising steps. This robustness suggests that ρ_{EOS} is a reliable and well-calibrated signal of length sufficiency, making ρ-**EOS** resilient to moderate variations in hyperparameter choices. Moreover, the asymmetric threshold pair $[0.4, 0.8]$ reaches the best performance ($L_{\text{init}} = 64$). We attribute this to the fact that the model starts from a short initial length, so it tends to favor expansion and therefore benefits from a looser lower bound and a tighter upper bound. In experiments with longer initial lengths, this prior is weaker.

Ablation to the adjustment factor function. Table 6 compares different $\mathrm{E}_{\text{factor}}(\cdot)$. The $\mathrm{E}_{\text{const}}(\cdot)$ improves over the baseline but remains limited. In contrast, the $\mathrm{E}_{\text{linear}}(\cdot)$ scales the adjustment magnitude proportionally to the deviation of ρ_{EOS} from the equilibrium region, enabling faster correction when the current length is moderately misaligned. The $\mathrm{E}_{\text{exp}}(\cdot)$ further amplifies this effect, allowing the model to reach a stable denoising regime more rapidly. As evidenced by the higher accuracy and fewer average adjustment steps in Table 6, these results confirm that adapting the adjustment magnitude based on implicit EOS density is crucial, and that aggressive early correction can significantly accelerate convergence without sacrificing accuracy.

5 Conclusion

In this work, we propose ρ-**EOS**, a training-free, online bidirectional length adaptation strategy for masked diffusion large language models. By leveraging the implicit EOS density, ρ-**EOS** continuously assesses length sufficiency and enables both expansion and contraction within a unified inference loop. Extensive experiments on mathematics and coding benchmarks show that ρ-**EOS** achieves a superior Pareto frontier to fixed-length baselines and DAEDAL in the performance-efficiency trade-off. Bidirectional control also improves token utilization and reduces sensitivity to the initial canvas length, helping sustain favorable trade-offs across different decoding configurations. More broadly, our analysis reveals that masked diffusion LLMs inherently possess a strong internal estimate of the appropriate response length for a given task. Using this internal signal for inference-time control provides a practical way to adapt fixed-length denoising to the varying generation requirements of individual inputs. Implicit EOS density provides a reliable signal for exposing and exploiting this latent capability, allowing fixed-length denoising to be seamlessly transformed into a flexible, adaptive-length generation process.

AI USE STATEMENT

We used generative AI tools, including large language models (LLMs) and AI agents, to assist with language polishing, drafting and revising text, improving the manuscript's organization, identifying relevant literature, and preparing and refining figures and tables. We did not use these tools to originate the research ideas, design experiments, or conduct the primary analyses of experimental data. AI assistance with the discussion and conclusion was limited to summarizing and articulating interpretations of results already obtained by the authors. The authors take full responsibility for the final manuscript, including the accuracy and integrity of its text, scientific claims, references, figures, and tables.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Claude 4 Anthropic. System card: Claude opus 4 & claude sonnet 4, 2025a. URL https://www.anthropic.com/claude-4-system-card.

Claude 4.5 Anthropic. System card: Claude sonnet 4.5, 2025b. URL https://www.anthropic.com/claude-sonnet-4-5-system-card.

Marianne Arriola, Aaron Gokaslan, Justin T Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Block diffusion: Interpolating between autoregressive and diffusion language models. *arXiv preprint arXiv:2503.09573*, 2025.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021a.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021b.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Xinhua Chen, Sitao Huang, Cong Guo, Chiyue Wei, Yintao He, Jianyi Zhang, Hai Li, Yiran Chen, et al. Dpad: Efficient diffusion language models with suffix dropout. *arXiv preprint arXiv:2508.14148*, 2025.

Shuang Cheng, Yihan Bian, Dawei Liu, Linfeng Zhang, Qian Yao, Zhongbo Tian, Wenhai Wang, Qipeng Guo, Kai Chen, Biqing Qi, et al. Sdar: A synergistic diffusion-autoregression paradigm for scalable sequence generation. *arXiv preprint arXiv:2510.06303*, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems, 2021. *URL https://arxiv. org/abs/2110.14168*, 9, 2021.

DeepMind. Gemini diffusion, 2025. URL https://deepmind.google/models/gemini-diffusion/.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pp. 70–79. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-5409.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*, 2022.

Shansan Gong, Shivam Agarwal, Yizhe Zhang, Jiacheng Ye, Lin Zheng, Mukai Li, Chenxin An, Peilin Zhao, Wei Bi, Jiawei Han, et al. Scaling diffusion language models via adaptation from autoregressive models. *arXiv preprint arXiv:2410.17891*, 2024.

Shansan Gong, Ruixiang Zhang, Huangjie Zheng, Jiatao Gu, Navdeep Jaitly, Lingpeng Kong, and Yizhe Zhang. Diffucoder: Understanding and improving masked diffusion models for code generation. *arXiv preprint arXiv:2506.20639*, 2025.

Zhengfu He, Tianxiang Sun, Qiong Tang, Kuanning Wang, Xuan-Jing Huang, and Xipeng Qiu. Diffusionbert: Improving generative masked language models with diffusion models. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 4521–4534, 2023.

Samar Khanna, Siddhant Kharbanda, Shufan Li, Harshit Varma, Eric Wang, Sawyer Birnbaum, Ziyang Luo, Yanis Miraoui, Akash Palrecha, Stefano Ermon, et al. Mercury: Ultra-fast language models based on diffusion. *arXiv preprint arXiv:2506.17298*, 2025.

Jaeyeon Kim, Lee Cheuk-Kit, Carles Domingo-Enrich, Yilun Du, Sham Kakade, Timothy Ngotiaoco, Sitan Chen, and Michael Albergo. Any-order flexible length masked diffusion. *arXiv preprint arXiv:2509.01025*, 2025.

Jinsong Li, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jiaqi Wang, and Dahua Lin. Beyond fixed: Training-free variable-length denoising for diffusion large language models. *arXiv preprint arXiv:2508.00819*, 2025a.

Pengxiang Li, Yefan Zhou, Dilxat Muhtar, Lu Yin, Shilin Yan, Li Shen, Soroush Vosoughi, and Shiwei Liu. Diffusion language models know the answer before decoding. *arXiv preprint arXiv:2508.19982*, 2025b.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.

Aiwei Liu, Minghua He, Shaoxun Zeng, Sijun Zhang, Linhao Zhang, Chuhan Wu, Wei Jia, Yuan Liu, Xiao Zhou, and Jie Zhou. Wedlm: Reconciling diffusion language models with standard causal attention for fast inference. *arXiv preprint arXiv:2512.22737*, 2025.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*, 2023.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1797–1807. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1206.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

Subham Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184, 2024.

Yuxuan Song, Zheng Zhang, Cheng Luo, Pengyang Gao, Fan Xia, Hao Luo, Zheng Li, Yuehang Yang, Hongli Yu, Xingwei Qu, et al. Seed diffusion: A large-scale diffusion language model with high-speed inference. *arXiv preprint arXiv:2508.02193*, 2025.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971, 2023. URL https://arxiv.org/abs/2302.13971.

Chenyu Wang, Paria Rashidinejad, DiJia Su, Song Jiang, Sid Wang, Siyan Zhao, Cai Zhou, Shannon Zejiang Shen, Feiyu Chen, Tommi Jaakkola, et al. Spg: Sandwiched policy gradient for masked diffusion language models. *arXiv preprint arXiv:2510.09541*, 2025.

Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, Ligeng Zhu, Ping Luo, Song Han, and Enze Xie. Fast-dllm: Training-free acceleration of diffusion llm by enabling kv cache and parallel decoding. *arXiv preprint arXiv:2505.22618*, 2025a.

Zirui Wu, Lin Zheng, Zhihui Xie, Jiacheng Ye, Jiahui Gao, Yansong Feng, Zhenguo Li, Victoria W., Guorui Zhou, and Lingpeng Kong. Dreamon: Diffusion language models for code infilling beyond fixed-size canvas, 2025b. URL https://hkunlp.github.io/blog/2025/dreamon.

Jingyi Yang, Guanxu Chen, Xuhao Hu, and Jing Shao. Taming masked diffusion language models via consistency trajectory reinforcement learning with fewer decoding step. *arXiv preprint arXiv:2509.23924*, 2025a.

Yicun Yang, Cong Wang, Shaobo Wang, Zichen Wen, Biqing Qi, Hanlin Xu, and Linfeng Zhang. Diffusion llm with native variable generation lengths: Let [eos] lead the way. *arXiv preprint arXiv:2510.24605*, 2025b.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b, 2025. URL https://hkunlp.github.io/blog/2025/dream.

Wei Zhang, Zhenhong Zhou, Junfeng Fang, Rongwu Xu, Kun Wang, Yuanhe Zhang, Rui Wang, Ge Zhang, Xinfeng Li, Li Sun, Lingjuan Lyu, Yang Liu, and Sen Su. LIFEBench: Evaluating length instruction following in large language models. *arXiv preprint arXiv:2505.16234*, 2025.

Siyan Zhao, Devaansh Gupta, Qinqing Zheng, and Aditya Grover. d1: Scaling reasoning in diffusion large language models via reinforcement learning. *arXiv preprint arXiv:2504.12216*, 2025.

Kaiwen Zheng, Yongxin Chen, Hanzi Mao, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling. *arXiv preprint arXiv:2409.02908*, 2024.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Fengqi Zhu, Rongzhen Wang, Shen Nie, Xiaolu Zhang, Chunwei Wu, Jun Hu, Jun Zhou, Jianfei Chen, Yankai Lin, Ji-Rong Wen, et al. Llada 1.5: Variance-reduced preference optimization for large language diffusion models. *arXiv preprint arXiv:2505.19223*, 2025a.

Fengqi Zhu, Zebin You, Yipeng Xing, Zenan Huang, Lin Liu, Yihong Zhuang, Guoshan Lu, Kangyu Wang, Xudong Wang, Lanning Wei, et al. Llada-moe: A sparse moe diffusion language model. *arXiv preprint arXiv:2509.24389*, 2025b.

A LIMITATIONS

ρ-**EOS** introduces density thresholds and an adjustment-factor function, which together determine the aggressiveness of expansion and contraction. Our ablations suggest that ρ-**EOS** is reasonably robust to moderate variations of these hyperparameters, but the best quality-efficiency trade-off may still depend on the initial canvas and the target task. In addition, the practical efficiency gains may vary with batching strategy, hardware, and decoding configuration.

B SUPPLEMENTARY EXPERIMENTS

Ablation on ρ thresholds [ρ_{low}, ρ_{high}] on HumanEval. We also conduct an ablation of density thresholds on HumanEval. As shown in Table 7, under symmetric settings, varying the thresholds over a wide range leads to minor fluctuations in accuracy (within 0.8%) and moderate increases in runtime, indicating low sensitivity to threshold tuning. Across different threshold settings, ρ-**EOS** consistently maintains stable accuracy. For wider equilibrium regions (e.g., $[0.2, 0.8]$) favor conservative canvas adjustment, yielding a higher $\boldsymbol{E}_{\text{ratio}}$ (90.4%) but slightly reduced $\boldsymbol{E}_{\text{token}}$ and $\boldsymbol{T}_{\text{runtime}}$. Regarding narrower or asymmetric settings allow more aggressive adjustment, substantially increasing $\boldsymbol{E}_{\text{token}}$ (up to 686.6) at the cost of additional denoising steps (up to 647s). Moreover, the asymmetric threshold pair $[0.4, 0.8]$ achieves the best accuracy (44.5%, tied with $[0.2, 0.8]$) when $L_{\text{init}} = 64$. We attribute this to the fact that the model starts from a short initial length, so it tends to favor expansion and therefore benefits from a looser lower bound and a tighter upper bound. In experiments with longer initial lengths, this prior is weaker, and such an asymmetric setting is less necessary.

Compatibility of ρ-EOS with acceleration-oriented decoding methods. As mentioned in the main text, ρ-**EOS** is orthogonal to acceleration-oriented decoding methods and can be stacked seamlessly with them. Table 8 validates this compatibility by combining ρ-**EOS** ($L_{\text{init}} = 64$) with three representative speedup methods: Fast-dLLM Wu et al. (2025a), Prophet Li et al. (2025b), and DPad Chen et al. (2025). Under the same adaptive-length setting, stacking DPad achieves runtime reductions ranging from roughly $2\times$ to $10\times$, while Fast-dLLM yields $1.3\times$ to $5.3\times$ speedups across the four benchmarks. These gains naturally incur moderate performance trade-offs: on GSM8K, for instance, DPad and Fast-dLLM retain 77.4% and 75.2% accuracy, respectively, versus the original 81.8%; on HumanEval, they achieve 38.4% and 41.5% compared to 43.3%. By contrast, Prophet exhibits larger accuracy drops (e.g., falling to 17.1% on HumanEval), indicating that its aggressive early-exit strategy interacts less favorably with length-adaptive generation. Overall, the results demonstrate that ρ-**EOS** serves as a compatible foundation—it can be layered mainstream acceleration techniques on top of it to realize $2\times$–$10\times$ (and beyond) speedups with only minor sacrifice in task performance.

Table 7: **Ablation on ρ_{high} and ρ_{low} on HumanEval.** Ablations are conducted on HumanEval with LLaDA-Instruct-8B and $L_{\text{init}} = 64$. The threshold of density (ρ) ranges from 0 to 1. We highlight the best-performance Asym setting ($[\rho_{\text{low}} = 0.4,\ \rho_{\text{high}} = 0.8]$) in **red**.

Metric	Density Threshold [ρ_{low}, ρ_{high}]			
	[0.2, 0.8]	[0.3, 0.7]	[0.4, 0.6]	[0.4, 0.8]
$\boldsymbol{Acc}$ ↑	44.5%	43.9%	43.7%	44.5%
$\boldsymbol{E}_{\text{token}}$	460.8	543.1	640.1	686.6
$\boldsymbol{N}_{\text{token}}$	509.7	619.9	758.5	809.5
$\boldsymbol{E}_{\text{ratio}}$ ↑	90.4%	87.6%	84.4%	84.8%
$\boldsymbol{T}_{\text{runtime}}$ ↓	433	522	595	647

B.1 WHY DOES ρ-EOS WORK?

Implicit length awareness in masked dLLMs by treating padding EOS as a part of response. At the core of ρ-**EOS** lies a simple but critical hypothesis: *masked diffusion language models implicitly estimate how much generation space they need for a given input, even before the sequence is fully denoised.* Importantly, this behavior is not an emergent artifact of our inference strategy, but is deeply rooted in the training engineering implementation of masked dLLMs. In masked diffusion

Table 8: **Compatibility of ρ-EOS with acceleration-oriented decoding methods.** We compare the original ρ-EOS with its combinations with several strong decoding acceleration methods under the same variable-length setting (L_{init}=64). The results show that ρ-EOS is largely orthogonal to these acceleration strategies: when paired with strong accelerators such as Fast-dLLM or DPad, it achieves substantial runtime reduction while maintaining comparable task performance. $\boldsymbol{Acc}$ denotes accuracy, $\boldsymbol{E}_{\text{token}}$ is the average effective tokens, $\boldsymbol{N}_{\text{token}}$ is the average total tokens, $\boldsymbol{E}_{\text{ratio}}$ is the effective token ratio, and $\boldsymbol{T}_{\text{runtime}}$ is the evaluation runtime in seconds. The **best** results within each benchmark block are **bold**, and the second-best results are underlined.

Benchmark	Method	$\boldsymbol{Acc}\uparrow$	$\boldsymbol{E}_{\text{token}}$	$\boldsymbol{N}_{\text{token}}$	$\boldsymbol{E}_{\text{ratio}}\uparrow$	$\boldsymbol{T}_{\text{runtime}}\downarrow$
GSM8K	ρ-EOS	**81.8**	198.4	231.8	85.6%	645
	ρ-EOS + Fast-dLLM	75.2	242.7	264.9	91.6%	498
	ρ-EOS + Prophet	62.5	205.0	264.0	77.7%	506
	ρ-EOS + DPad	77.4	156.9	**171.1**	**91.7%**	**324**
MATH500	ρ-EOS	**37.0**	428.8	486.3	88.2%	1398
	ρ-EOS + Fast-dLLM	36.2	376.3	411.2	91.5%	275
	ρ-EOS + Prophet	25.0	222.0	407.1	54.5%	239
	ρ-EOS + DPad	34.0	247.2	260.2	**95.0%**	**163**
MBPP	ρ-EOS	**39.6**	324.0	507.4	63.9%	2031
	ρ-EOS + Fast-dLLM	37.0	344.9	564.0	61.2%	382
	ρ-EOS + Prophet	27.2	300.4	547.0	54.9%	221
	ρ-EOS + DPad	37.2	264.1	283.3	**93.2%**	**208**
HumanEval	ρ-EOS	**43.3**	570.4	654.0	88.4%	580
	ρ-EOS + Fast-dLLM	41.5	636.7	753.6	84.5%	143
	ρ-EOS + Prophet	17.1	418.3	722.8	57.9%	101
	ρ-EOS + DPad	38.4	431.2	446.4	**96.6%**	**87**

Table 9: **Ablation of the contraction operation.** Expansion and all other decoding settings are held fixed; only contraction is enabled or disabled. All experiments use LLaDA-Instruct-8B, L_{init}=64, and L_{max}=2048. Acc. denotes accuracy for GSM8K and MATH500 and pass@1 for MBPP and HumanEval. $\boldsymbol{E}_{\text{token}}$ and $\boldsymbol{N}_{\text{token}}$ are the average effective and total tokens, respectively.

Benchmark	Contraction	Acc. ↑	$\boldsymbol{E}_{\text{token}}$	$\boldsymbol{N}_{\text{token}}\downarrow$	$\boldsymbol{E}_{\text{ratio}}\uparrow$
GSM8K	Enabled	81.8	198.4	**231.8**	**85.6%**
	Disabled	**83.1**	239.9	326.6	73.4%
MATH500	Enabled	37.0	428.8	**486.3**	**88.2%**
	Disabled	**38.0**	442.4	520.8	84.9%
MBPP	Enabled	39.6	324.0	**507.4**	**63.9%**
	Disabled	**39.8**	335.2	597.0	56.2%
HumanEval	Enabled	43.3	570.4	**654.0**	**88.4%**
	Disabled	**45.1**	645.2	769.2	83.9%

large language models such as LLaDA Nie et al. (2025), padding EOS tokens are explicitly treated as part of the response sequence and are randomly masked during training. As a result, the model is repeatedly trained to jointly predict semantic tokens and EOS tokens under partial observation. This training paradigm forces the model to reason not only about ***what*** content should be generated, but also ***where*** meaningful content should end. Consequently, the prediction of EOS becomes intrinsically coupled with the model's internal estimate of length sufficiency.

Crucially, this latent estimate is not exposed as an explicit scalar length prediction. Instead, it manifests implicitly through property of special functional token–the density of EOS over the remaining masked positions. From a denoising perspective, when the model assigns high probability mass to

Table 10: **Detailed contraction statistics.** "Generated" counts removed positions that had already been decoded, while "EOS" is the subset of those generated positions containing EOS. "Gen. samples" counts samples in which at least one generated position is removed. With contraction disabled, all reported contraction statistics are zero by construction.

Benchmark	Events	Contracted samples	Removed	MASK	Generated	EOS	Gen. samples	Removed/event
GSM8K	7,893	663 (50.23%)	109,192	48,483	60,709	56,949	447	13.83
MATH500	825	118 (23.60%)	9,032	5,820	3,212	3,045	31	10.95
MBPP	398	53 (10.60%)	5,648	4,117	1,531	1,454	22	14.19
HumanEval	89	11 (6.71%)	1,208	816	392	392	3	13.57

EOS at masked locations, it signals that the semantic content is largely complete and that the remaining generation space is redundant. Conversely, a consistently low implicit EOS density indicates that the model still requires additional capacity to express the intended response, and is therefore reluctant to terminate. Viewed from this angle, ρ-**EOS** does not introduce an external heuristic for length control. Instead, it ***reads out*** and ***amplifies*** a signal that the model has already learned to encode during training. By monitoring the implicit EOS density and feeding it back into the denoising process, our method aligns inference-time length adaptation with the model's internal notion of response completeness. This close alignment between training and inference is a key reason why ρ-**EOS** enables stable and effective bidirectional length control without additional supervision or multi-stage procedures. ***In essence, variable-length inference emerges naturally when the inference process respects the semantics of EOS prediction learned by masked diffusion large language models.***

B.2 Why does $\rho_{\texttt{EOS}}$ outperform EOS confidence for length control?

A side effect of treating padding EOS as part of the response. While treating padding EOS tokens as part of the response endows masked diffusion language models with implicit length awareness, it also introduces a non-trivial side effect that overconfidence on EOS may lead to instability in using it as a single-stage length control signal.

In particular, as observed in Yang et al. (2025a), masked dLLMs often exhibit an EOS Trap phenomenon: during early denoising steps, the predicted confidence of EOS is significantly higher than that of non-EOS ones. This behavior is a direct consequence of the training paradigm. Since padding EOS tokens may occupy a large fraction of the sequence during training and are randomly masked, the model is frequently optimized to recover EOS under high uncertainty. As a result, in early decoding steps—when most positions remain masked and semantic structure has not yet emerged—the model tends to assign disproportionately high probability mass to EOS. Importantly, this high EOS confidence does not indicate semantic completeness, but rather reflects the model's uncertainty and its bias toward the statistically termination token.

When EOS confidence is directly used as a length control signal, this early-stage overconfidence can prematurely trigger contraction or suppress necessary expansion. Such misaligned length decisions are particularly harmful in a single-stage setting, where length adjustment and semantic denoising are tightly interleaved. This explains the performance degradation observed when EOS confidence is employed as the control signal for single-stage strategy. In contrast, the implicit EOS density ρ aggregates EOS mass over all remaining masked positions, providing a global and temporally smoother estimate of length sufficiency. Therefore, implicit density ρ serves as a more robust length control signal than confidence.

B.3 Additional Experimental Results.

Effect of contraction on quality and efficiency. We isolate the contribution of contraction by keeping expansion and all other decoding settings unchanged and disabling only the contraction operation. As shown in Table 9, disabling contraction slightly improves accuracy or pass@1 by 0.2–1.8 points across the four main benchmarks. However, enabling contraction reduces the average total canvas length by 6.6–29.0% and improves the effective-token ratio by 3.3–12.2 percentage points. Contraction should therefore be understood as a quality-efficiency balancing mechanism rather than a component intended to maximize accuracy in isolation: disabling it may be preferable

Table 11: **Results on additional generation domains with LLaDA-Instruct-8B.** Fixed-length decoding uses generation lengths of 256 or 512, while ρ-EOS uses L_{init}=64 and L_{max}=512. R-1/2/L denote ROUGE-1/2/L. IFEval reports prompt-level (P) and instruction-level (I) strict and loose accuracy. The **best** result in each row is **bold**, and the second-best is underlined.

Domain	Benchmark	Metric	Fixed 256	Fixed 512	ρ-EOS
Summarization	XSum	R-1 ↑	4.15	0.76	**11.77**
		R-2 ↑	1.05	0.19	**2.99**
		R-L ↑	3.12	0.60	**8.74**
	SAMSum	R-1 ↑	3.30	1.26	**15.54**
		R-2 ↑	1.24	0.46	**6.16**
		R-L ↑	2.60	0.97	**12.37**
Long-form	LIFEBench	Length score ↑	**20.84**	17.79	18.46
		Acc@10 ↑	**7.50**	4.50	5.50
		Relative error ↓	**1.15**	4.45	1.16
Instruction following	IFEval	P-Strict ↑	34.01	**34.20**	32.57
		P-Loose ↑	37.15	**39.56**	38.27
		I-Strict ↑	46.64	**47.72**	45.53
		I-Loose ↑	49.76	**51.68**	50.53

when quality is the sole objective, whereas enabling it removes redundant capacity and improves token utilization.

Table 10 further reports how contraction behaves. The fraction of samples experiencing contraction is task dependent, ranging from 6.71% on HumanEval to 50.23% on GSM8K, while each event removes 10.95–14.19 tokens on average. Contraction can truncate both trailing MASK positions and already decoded positions; among the latter, 93.8–100% are EOS tokens across the four benchmarks. These statistics show that contraction primarily trims EOS-heavy tail capacity, while also clarifying that it is not restricted to removing only unresolved MASK positions.

Generalization to additional generation domains. To examine whether adaptive canvas control generalizes beyond mathematical reasoning and code generation, we evaluate summarization (XSum (Narayan et al., 2018) and SAMSum (Gliwa et al., 2019)), long-form generation (LIFEBench (Zhang et al., 2025)), and instruction following (IFEval (Zhou et al., 2023)). We compare fixed-length decoding with $L \in \{256, 512\}$ against ρ-**EOS** initialized from a short canvas (L_{init}=64) and allowed to expand up to L_{max}=512. The complete results are reported in Table 11.

On summarization, ρ-**EOS** consistently improves all three ROUGE metrics. It obtains ROUGE-1/2/L scores of 11.77/2.99/8.74 on XSum and 15.54/6.16/12.37 on SAMSum, compared with 4.15/1.05/3.12 and 3.30/1.24/2.60, respectively, for the stronger fixed-length baseline. On LIFEBench, ρ-**EOS** improves over the 512-token baseline in length score (18.46 vs. 17.79), Acc@10 (5.50 vs. 4.50), and relative length error (1.16 vs. 4.45); the 256-token baseline remains strongest on the first two metrics, while its relative error is nearly identical (1.15). On IFEval, ρ-**EOS** remains within 1.15–2.19 points of the strongest fixed-length result across the four prompt- and instruction-level metrics. Thus, the added results show strong gains for summarization and broadly competitive behavior for long-form generation and instruction following, without implying uniform improvements on every metric.