Attention-Discounted Adaptive Sampler for Masked Diffusion Language Models

Yusuf Sahin[1] **Ahmed Rockey Saikia**[2]
Volkan Cevher[2] **Paolo Favaro**[1]
[1]University of Bern, Bern, Switzerland
[2]EPFL, Lausanne, Switzerland

Abstract

Masked diffusion language models can reduce inference steps by revealing multiple tokens per denoising iteration, but this parallelism is fragile: positions that are individually confident may be unsafe to commit together when their predictions are coupled. Existing training-free samplers such as Top-k, Fast-dLLM, and EB-Sampler mainly control how many tokens to reveal, while often ranking candidates by token-wise scores that ignore interactions within the selected set. We propose ADAS, a training-free reranking rule that leaves the base sampler's stopping rule unchanged and greedily discounts each token-wise confidence score according to its attention to already selected positions, weighted by their prediction uncertainty. Across LLaDA-8B-Base and Dream-7B-Base on the reasoning benchmarks GSM8K and MATH500 and the code benchmarks HumanEval and MBPP, plugging ADAS into all three samplers improves low-NFE performance at matched denoiser evaluations by 9.11 and 10.46 percentage points on average, respectively, with 3.1% per-forward runtime overhead. Code is available at `https://github.com/yusufsahin99/ADAS`.

1 Introduction

Masked diffusion language models (MDLMs) [Sahoo et al., 2024, Shi et al., 2024] promise a different speed–quality tradeoff from causal autoregressive models: each denoising step can reveal a subset of positions rather than a single next token. This parallelism is especially appealing for reasoning and code-generation tasks, where long outputs make inference latency costly. The central decoding question is therefore not only *what* token to place at each masked position, but also *which positions* can be safely committed in the same step.

This second question is where aggressive parallel decoding becomes fragile. Many practical samplers rank candidate positions by token-wise quantities such as confidence, margin, or entropy, and then use a stopping rule or budget to decide how many positions to reveal. Such rules implicitly treat high-scoring candidates as compatible. Yet successful joint unmasking depends not only on individual token scores, but also on whether the selected predictions are mutually compatible: revealing one position may substantially change the predictive distribution at another.

Recent work has made this failure mode increasingly visible. Some methods treat diffusion decoding as planning or search [Peng et al., 2025, Lee et al., 2025]; others control the amount of parallelism through schedules [Luxembourg et al., 2025, Israel et al., 2025] or learned unmasking policies [Jazbec et al., 2025, Hong et al., 2026, Bao et al., 2026]. The theoretical lens of Ben-Hamu et al. [2025] also separates *model error* from *joint dependence error*, clarifying that good marginal predictions are not sufficient when several positions are committed simultaneously.

We take this perspective one step further. Widely used decoders often differ in the *constraint* imposed on the selected subset, but still rely on a token-wise objective for constructing that subset. Top-k

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

fixes the cardinality, Fast-dLLM [Wu et al., 2025] uses a confidence-threshold condition, and EB-Sampler [Ben-Hamu et al., 2025] enforces an entropy budget. These mechanisms decide *when to stop adding tokens*; they do not, by themselves, make the marginal value of a candidate depend on which other positions have already been selected.

Our proposal is to make subset construction dependency-aware while leaving the stopping rule intact. We introduce ADAS, a training-free Attention-Discounted Adaptive Sampler that greedily builds the set of positions to unmask. Starting from standard token confidence scores, ADAS discounts a remaining candidate when it attends strongly to already selected, uncertain positions. Confidence supplies an individual-value proxy, while attention supplies a directional compatibility proxy for each append decision.

This design deliberately separates ranking from budget control. ADAS does not train a planner, add a verifier, perform lookahead search, or redesign the sampler's stopping criterion. Instead, it upgrades the greedy ranking step inside existing decoders, so the same attention-discounted update can be used with fixed-cardinality, confidence-threshold, or entropy-budget rules. This places ADAS in a complementary position to learned unmasking policies [Jazbec et al., 2025, Hong et al., 2026, Asano et al., 2026], lookahead methods [Lee et al., 2025], and schedule- or verifier-based approaches [Luxembourg et al., 2025, Israel et al., 2025].

The experiments bear this out: ADAS helps most when few denoising steps (low-NFE) are used and many positions are revealed at once, which is precisely the setting in which ranking positions by individual token scores breaks down. Across mathematical reasoning and code-generation benchmarks, adding the ADAS score update consistently improves Top-k, EB-Sampler, and Fast-dLLM at matched denoiser-evaluation budgets.

Our contributions are as follows.

- We analyze real decoding trajectories and measure how much the predicted distribution at a masked position changes when another masked position is revealed. We find that this change is well predicted by the attention weight from the first position to the second. The prediction becomes more consistent across decoding states when the attention weight is multiplied by the uncertainty of the revealed token. Further supporting results on the relationship between attention and stochastic dependence in synthetic data are reported in the appendix.
- We introduce ADAS, a training-free attention-discounted greedy selector that decouples *ranking* from *stopping*, plugs into Top-k, Fast-dLLM, and EB-Sampler, and adds only 3.1% runtime overhead per model forward.
- Across two MDLMs and four reasoning/code benchmarks, ADAS improves matched-NFE performance by 9.11 and 10.46 points on average.

2 Related work

Recent work on masked diffusion decoding has shown that generation quality depends strongly on the order and grouping of positions unmasked at each step. Several approaches therefore treat decoding as an explicit planning or policy-learning problem. Peng et al. [2025] formulate sampling as planner-guided decoding, Asano et al. [2026] separate where-to-unmask from what-to-unmask and learn a supervised planner, Lee et al. [2025] use lookahead search over candidate decoding trajectories, and Jazbec et al. [2025], Hong et al. [2026], Bao et al. [2026] learn unmasking policies or filters that improve over hand-designed heuristics. These methods show that token-wise confidence order is often not enough, especially outside weakly parallel regimes.

A complementary line of work improves the speed–quality tradeoff of diffusion decoding through better inference schedules or auxiliary signals. Luxembourg et al. [2025] propose a structural scheduler that reduces harmful interactions through dilated grouping, while Israel et al. [2025] use an auxiliary autoregressive model to adapt the amount of parallelism during decoding. Ben-Hamu et al. [2025] provide a useful theoretical framing by decomposing decoding error into model error and joint dependence error, motivating adaptive multi-token unmasking rules such as EB-Sampler. Kim et al. [2025] propose KLASS, a training-free sampler that combines token confidence with temporal KL divergence between consecutive denoising distributions to select stable tokens for unmasking.

Recent dependency-aware samplers provide especially close comparisons. PUNT [Azangulov et al., 2026] performs approximate conditional-independence testing and removes lower-confidence tokens from conflicting groups. DAWN [Luo et al., 2026] extracts a dependency graph, favoring positions supported by revealed context while avoiding simultaneous updates of coupled uncertain positions. DOS [Zhou et al., 2026] instead scores attention to already revealed tokens as positive support for masked-token updates. DAPD [Kim et al., 2026], the closest method to ADAS, symmetrizes and thresholds masked-position attention into a binary dependency graph, then selects an independent set for parallel decoding; an attention edge is therefore a hard conflict.

ADAS uses the same broad source of information—self-attention—but instantiates a different decoding principle. It does not construct a binary graph, threshold attention, or impose independent-set constraints. Instead, ADAS keeps attention continuous and uses it as a marginal penalty inside greedy subset construction. Thus, ADAS does not declare two positions incompatible. A strongly coupled candidate can still be selected when its confidence gain outweighs the estimated commitment risk. This soft formulation is important because attention is an imperfect proxy for harmful dependence: not every high-attention pair should be forbidden, and not every low-attention pair is guaranteed independent.

The two methods also differ in their integration point. DAPD is a standalone dependency-aware decoder that defines its own parallel batches through graph coloring. ADAS is a sampler-agnostic reranking module: it leaves the stopping rule, budget, and admissibility condition of Top-k, Fast-dLLM, or EB-Sampler unchanged, and only changes the order in which candidates are proposed. In short, DAPD uses attention to construct a graph of forbidden simultaneous updates; ADAS uses attention to compute a soft marginal discount inside existing samplers.

3 Preliminaries

3.1 Notation

Let $\mathcal{V} = \{1, \dots, K\} \uplus \{m\}$ denote a finite vocabulary augmented with a special mask token m. A sequence is denoted by $x \in \mathcal{V}^d$, where d is the sequence length. A position i is *masked* if $x_i = m$ and *unmasked* otherwise. For a partially masked sequence, let $\mathcal{M} \subseteq \{1, \dots, d\}$ be the set of masked positions, and let $\bar{\mathcal{M}}$ denote its complement. We write $x^{\bar{\mathcal{M}}}$ for the observed tokens. We use $q(\cdot \mid x^{\bar{\mathcal{M}}})$ to denote the true conditional distribution over masked positions given the currently revealed context and $p_\theta(\cdot \mid x^{\bar{\mathcal{M}}})$ to denote the model's conditional distribution.

3.2 Masked Diffusion Language Models

Masked diffusion language models [Sahoo et al., 2024, Shi et al., 2024] generate text through iterative denoising. Starting from a highly masked sequence, the model repeatedly predicts token distributions at masked positions and progressively unmasks a subset of them until no masks remain. A decoding step therefore involves two coupled decisions: *what* token to place at a selected position, and *where* to unmask next.

At a decoding step, an MDLM with parameters θ predicts, for each masked position $i \in \mathcal{M}$, a conditional distribution $p_\theta(x^i \mid x^{\bar{\mathcal{M}}})$. Many existing decoding strategies then assign each masked position a token-wise score derived from this distribution, and construct a subset $S \subseteq \mathcal{M}$ of positions to unmask according to a sampler-specific stopping rule. A common choice is the confidence score

$$c_i = \max_{x^i} p_\theta(x^i \mid x^{\bar{\mathcal{M}}}), \tag{1}$$

namely the probability assigned to the most likely token value at position i. Different samplers mainly differ in how the subset S is constrained, e.g., by fixing its size, thresholding confidence, or enforcing an entropy budget.

3.3 Model error and joint dependence error

A central difficulty in parallel decoding is that the model provides token-wise conditional marginals for masked positions, while decoding multiple positions in the same step requires reasoning about

Table 1: Real-data dependency diagnostic. $\bar{\rho}_s$ is the mean statewise Spearman correlation and ρ correlates all pairs. Positive Gain means the ADAS score is more negatively associated with intervention KL than confidence alone.

Task	Model	$\bar{\rho}_s(A_{ji}, K_j)$	$\rho(A_{ji}, K_j)$	$\rho(u_{ji}, K_j)$	$\bar{\rho}_s(c_j^0, K_j)$	$\bar{\rho}_s(s_j, K_j)$	Gain
GSM8K	Dream	.358	.311	.823	.045	-.281	**+.325**
	LLaDA	.360	.463	.835	.076	-.330	**+.405**
HumanEval	Dream	.305	.423	.782	-.030	-.209	**+.178**
	LLaDA	.389	.544	.841	.100	-.363	**+.463**
MBPP	Dream	.204	.284	.791	-.114	-.211	**+.097**
	LLaDA	.251	.327	.794	-.023	-.215	**+.193**
MATH500	Dream	.316	.298	.806	-.079	-.271	**+.193**
	LLaDA	.436	.545	.815	.116	-.279	**+.395**

their *joint* behavior. As emphasized by Ben-Hamu et al. [2025], the error incurred by unmasking a subset $S \subseteq \mathcal{M}$ can be decomposed into a *model error* term and a *joint dependence error* term:

$$\underbrace{\sum_{i \in S} D_{\mathrm{KL}}\big(q(x^i \mid x^{\bar{\mathcal{M}}}),\, p_\theta(x^i \mid x^{\bar{\mathcal{M}}})\big)}_{\text{model error}} \;+\; \underbrace{D_{\mathrm{KL}}\big(q(x^S \mid x^{\bar{\mathcal{M}}}),\, \textstyle\prod_{i \in S} q(x^i \mid x^{\bar{\mathcal{M}}})\big)}_{\text{joint dependence error}}$$

The model error captures inaccuracies in the model's token-wise predictions, while the joint dependence error measures the discrepancy introduced by treating the selected positions as conditionally independent. The latter is the multi-information of the true conditional distribution q over S: it is model-independent and characterizes the intrinsic coupling among positions selected for simultaneous commitment.

Since q is unobserved at decoding time, ADAS uses the denoiser self-attention A_{ij} as a model-side proxy for coupling strength. When this proxy is informative, discounting candidates that attend strongly to already selected, uncertain positions should reduce the expected dependence cost of the next parallel commitment. Section 4 tests this premise on the target benchmarks; additional synthetic experiments validating the relationship between attention and stochastic dependence are provided in Appendix F.

4 Real-data dependency diagnostic

We test whether attention reflects *conditional dependence* along genuine decoding trajectories: if masked position j depends on masked position i, revealing i should change the model distribution at j. We generate answers for 100 examples each from GSM8K, MATH500, HumanEval, and MBPP with LLaDA-8B-Base and Dream-7B-Base using confidence Top-k decoding ($k = 8$). A state is the partially masked output at a particular unmasking step. We probe zero-indexed steps $0, 7, 15, 23, 31$ of the 32-step GSM8K trajectories and steps $0, 15, 31, 47, 63$ of the 64-step trajectories for the other tasks. At each state, we choose the most-confident masked position i. Before revealing it, we record final-layer attention averaged across heads, A_{ji}, from each of the other seven co-selected masked positions j to i, then reveal only i in a counterfactual copy. Let p_j^0 and p_j^i be the distributions before and after this intervention, and $c_j^a = \max_x p_j^a(x)$. We measure

$$K_j = D_{\mathrm{KL}}(p_j^0 \| p_j^i), \qquad \Delta c_j = c_j^i - c_j^0, \qquad u_{ji} = A_{ji}(1 - c_i^0), \qquad s_j = c_j^0 - 40u_{ji}.$$

The baseline trajectory is unchanged, yielding 500 states and 3,500 pairs per benchmark/model; Appendix B gives full details.

Let ρ denote Spearman correlation over all 3,500 pairs and $\bar{\rho}_s$ the mean of the 500 correlations computed separately among the seven pairs in each state. The latter holds the chosen token and partially masked output fixed. Because c_i^0 is constant within a state, uncertainty gating cannot alter the statewise rank of A_{ji}, but it can calibrate effects across states. A stronger dependence effect should produce a lower s_j; Gain is $\bar{\rho}_s(c_j^0, K_j) - \bar{\rho}_s(s_j, K_j)$.

Attention is positively correlated with intervention KL in every setting ($\bar{\rho}_s = .204$–$.436$). Uncertainty-gated attention u_{ji} is more correlated than raw attention over all pairs (.782–.841 versus .284–.545),

showing improved cross-state calibration. Finally, the attention-discounted confidence score s_j is more negatively correlated with KL than the token-wise confidence c_j^0 in all eight statewise comparisons (Gain = .097–.463); in all eight benchmark–model settings, the 95% paired bootstrap interval for this improvement is strictly positive. Thus, attention is predictive of conditional dependence among masked positions, and uncertainty-weighted attention is even more predictive across decoding states.

To test whether attention reranking defers tokens that benefit from additional context, we examine its first divergence from confidence ranking in each state. We denote the confidence-ranked token it defers by d and the replacement it selects by r, then reveal the common first choice i. Thus, K_d and K_r measure the resulting distributional changes at the deferred and replacement tokens, while Δc_d and Δc_r measure their confidence changes. Across 1,098 states with such a divergence, deferred tokens versus replacements have mean KL .793 versus .230, mean confidence gain $+.092$ versus $+.024$, and gain confidence in 86.8% versus 62.9% of cases. Attention reranking therefore preferentially delays tokens whose distributions change more and whose confidence increases more after receiving the selected token as context, while choosing less affected replacements for immediate parallel commitment; model details appear in Appendix B.

5 Method

5.1 Problem statement

Consider a masked diffusion decoding step with current set of masked positions $\mathcal{M}$. The goal is to select a subset $S \subseteq \mathcal{M}$ of positions to unmask in parallel. In most existing samplers, this selection problem is implicitly decomposed into two parts: a *token-wise utility* used to rank candidate positions, and a *stopping criterion* that determines when enough positions have been selected.

A common choice is to use the confidence scores c_i introduced in Eq. (1) to construct S so as to maximize an additive objective of the form

$$U(S) = \sum_{i \in S} c_i, \tag{2}$$

subject to a sampler-specific constraint. Different decoding rules mainly differ in this constraint. For example, Top-k enforces a fixed cardinality constraint $|S| = k$, Fast-dLLM uses a confidence-based stopping condition, and EB-Sampler uses an entropy-based bound.

Our starting point is that this token-wise objective ignores dependencies among positions selected in the same step. A candidate's reliability should depend not only on its own token-wise score, but also on the uncertainty of the positions with which it is jointly committed and the strength of their interactions. We therefore retain the general subset-selection view above, but replace the purely additive token-wise objective with a dependency-aware utility that accounts for pairwise interactions.

5.2 Attention-Discounted Adaptive Sampler (ADAS)

ADAS replaces token-wise candidate ranking with dependency-aware greedy subset construction, while leaving the sampler's stopping rule unchanged. At each decoding step, let $\mathcal{M}$ be the set of masked positions and let c_i denote the usual confidence score for position $i \in \mathcal{M}$. Let $A \in [0,1]^{|\mathcal{M}| \times |\mathcal{M}|}$ be an attention-based interaction matrix over masked positions, where A_{ij} measures attention from position i to position j. In our implementation, A is obtained from the denoiser's final self-attention layer by averaging attention scores across heads, following the ablation in Appendix D.

Given an already selected set S, ADAS scores a remaining candidate $i \in \mathcal{M} \setminus S$ by the marginal utility

$$u(i \mid S, A) = c_i - \alpha \sum_{s \in S} A_{is}(1 - c_s), \tag{3}$$

where $\alpha \geq 0$ controls the strength of the attention discount. The first term favors positions with high token-wise confidence. The second term discounts candidates that attend strongly to selected positions whose predictions are still uncertain. Thus, a candidate is penalized not simply for being coupled to another selected position, but for being coupled to one whose commitment is less reliable.

Algorithm 1 Attention-Discounted Adaptive Sampler (ADAS)

Require: $\mathcal{M}$, scores $\{c_i\}$, attention A, weight α, admissibility rule Adm
  $S \leftarrow \emptyset$
  while $\mathcal{M} \setminus S \neq \emptyset$ **do**
    $i^\star \leftarrow \arg\max_{i \in \mathcal{M} \setminus S} \left[c_i - \alpha \sum_{s \in S} A_{is}(1 - c_s) \right]$
    if not $\mathrm{Adm}(S \cup \{i^\star\})$ **then**
      break
    end if
    $S \leftarrow S \cup \{i^\star\}$
  end while
  return S

For an ordered sequence of selections $\pi = (\pi_1, \ldots, \pi_m)$, define the append-only utility

$$U(\pi, A) = \sum_{t=1}^{m} \left[c_{\pi_t} - \alpha \sum_{r<t} A_{\pi_t \pi_r}(1 - c_{\pi_r}) \right]. \tag{4}$$

Appending candidate i to π has marginal value exactly Eq. (3). Equation (4) is therefore an ordered motivation for the operational greedy rule, not an equivalent unordered set utility: an unordered pairwise objective would introduce reverse-direction terms absent from ADAS. Algorithm 1 defines the method; we claim no global optimization or approximation guarantee.

At each iteration, ADAS selects the remaining candidate with largest current marginal utility and checks whether adding it satisfies the base sampler's admissibility rule. If the enlarged set is admissible, the candidate is added; otherwise, construction stops and the selected positions are unmasked in parallel.

In our experiments, ADAS is combined with three admissibility rules:

$$\text{Top-}k: \quad |S \cup \{i^\star\}| \leq k, \tag{5}$$

$$\text{Fast-dLLM}: \quad |S \cup \{i^\star\}| \Big(1 - \min_{s \in S \cup \{i^\star\}} c_s\Big) < f, \tag{6}$$

$$\text{EB-Sampler}: \quad \sum_{s \in S \cup \{i^\star\}} H\big(p_\theta(x^s \mid x^{\widetilde{\mathcal{M}}})\big) - \max_{s \in S \cup \{i^\star\}} H\big(p_\theta(x^s \mid x^{\widetilde{\mathcal{M}}})\big) < \gamma. \tag{7}$$

Here k, f, and γ are the fixed-cardinality, confidence, and entropy-budget thresholds, respectively. These rules determine when subset construction stops; ADAS only changes how candidates are ranked before that stopping decision.

We use $\alpha = 40$ as a single global default in every main experiment. It was selected once by averaging an ablation over six LLaDA-8B HumanEval settings and then reused without task-, model-, or stopping-rule-specific retuning (Appendix D). The attention discount can be updated incrementally. After adding a selected position s, each remaining candidate's score is decreased by $\alpha A_{is}(1 - c_s)$, giving $O(|S|\,|\mathcal{M}|)$ selection cost for a final subset S. This overhead is small relative to denoiser forward passes; we measure it in Section 6.3 and ablate α and the uncertainty term in Appendix D.

6 Experiments

6.1 Experimental setup

We evaluate ADAS on four benchmarks spanning mathematical reasoning and code generation: GSM8K [Cobbe et al., 2021], MATH500 [Hendrycks et al., 2021], MBPP [Austin et al., 2021], and HumanEval [Chen et al., 2021]. We use two strong open-weight masked diffusion language models, Dream-7B-Base [Ye et al., 2025] and LLaDA-8B-Base [Nie et al., 2025].

Our goal is to test whether attention-discounted subset construction helps precisely where parallel unmasking is most useful but most error-prone. We therefore instantiate ADAS with three representative stopping criteria: Top-k, Fast-dLLM, and EB-Sampler, yielding three variants we denote Top-k+AD, Fast-dLLM+AD, and EB+AD. Each variant is compared against its corresponding baseline over a range of parallel decoding settings.

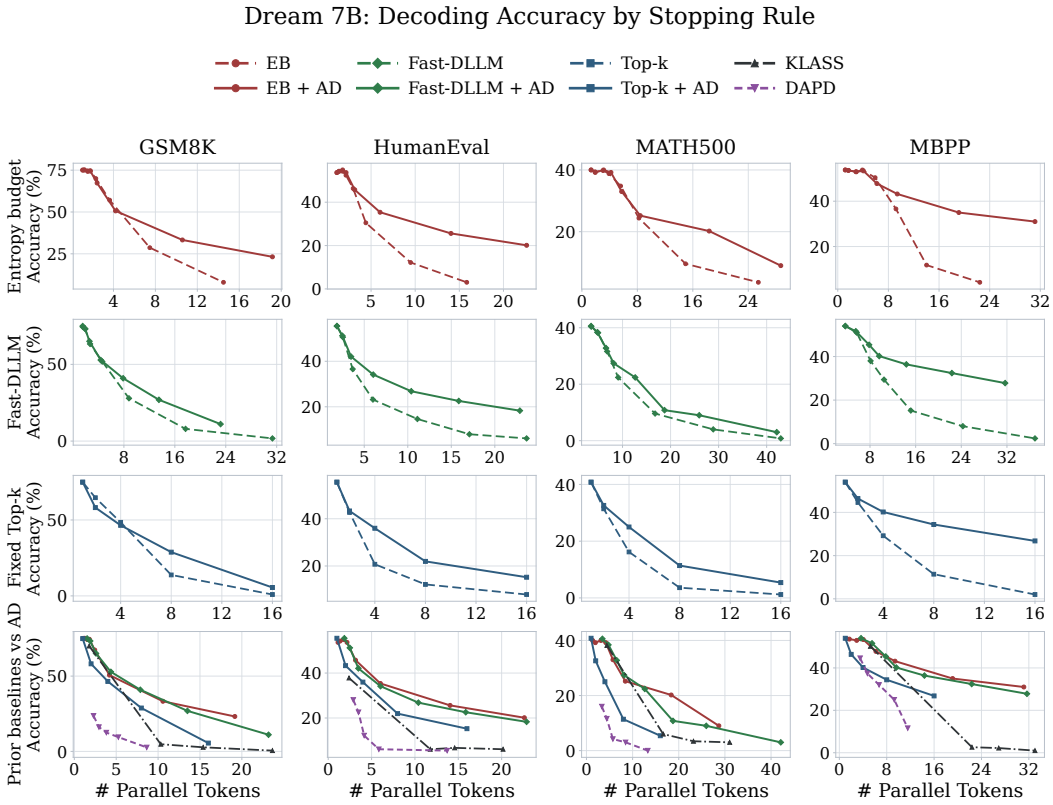

Figure 1: Effect of attention-discounted selection on Dream-7B-Base. Rows correspond to entropy-bounded decoding, Fast-dLLM, fixed Top-k, and comparison against KLASS and DAPD; columns correspond to GSM8K, HumanEval, MATH500, and MBPP. The same pattern observed for LLaDA-8B-Base persists: attention-discounted selection is most beneficial when many tokens are revealed in parallel, where the original samplers degrade sharply.

We further evaluate KLASS [Kim et al., 2025] and DAPD [Kim et al., 2026] as additional training-free baselines.

6.2 Main results

Figures 1 and 2 summarize the main speed–quality tradeoff. Moving to the right corresponds to revealing more tokens per denoising step, and therefore to lower NFE. The central question is whether the sampler can preserve quality in this highly parallel regime, where jointly selected dependent positions are most likely to induce errors.

In each figure, the first three rows compare EB-Sampler, Fast-dLLM, and Top-k with their ADAS variants; the final row compares ADAS-augmented samplers with KLASS and DAPD. Together, these plots test two claims: ADAS should improve its own base sampler, and the resulting decoders should remain competitive with recent training-free samplers.

Across both models, ADAS consistently improves robustness in the high-parallelism regime. For the adaptive stopping rules EB and fast-dLLM, the baseline curves often deteriorate rapidly as the entropy budget increases, whereas their attention-discounted variants degrade more gradually.

Fixed Top-k decoding shows that the benefit is not tied to adaptive stopping criteria. Since k directly controls the number of tokens unmasked at each step, this setting isolates the effect of subset construction. Across both base models, Top-k+AD improves or preserves performance at larger k, where vanilla Top-k suffers the largest degradation. Thus, reranking candidates by attention-discounted utility selects safer groups of tokens even when the subset size is fixed in advance.

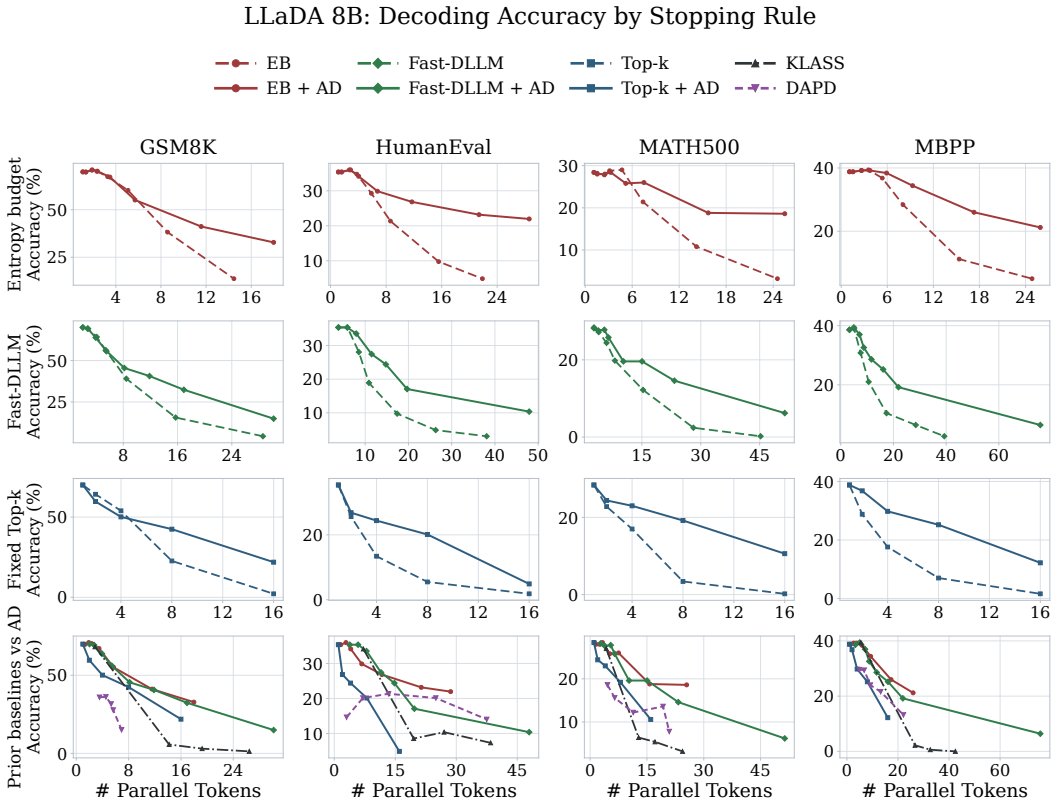

Figure 2: Effect of attention-discounted selection on LLaDA-8B-Base. Rows correspond to entropy-bounded decoding, Fast-dLLM, fixed Top-k, and comparison against KLASS and DAPD; columns correspond to GSM8K, HumanEval, MATH500, and MBPP. Across stopping rules and datasets, attention-discounted selection preserves substantially higher accuracy in the highly parallel regime, while matching the original samplers more closely when the number of parallel tokens is small.

Table 2: Matched-NFE absolute gains in the low-NFE regime, reported as LLaDA-8B-Base / Dream-7B-Base percentage points.

Dataset	EB	Fast-dLLM	Top-k	Avg. over methods
GSM8K	+9.22 / +14.31	+6.92 / +6.36	+11.90 / +5.88	+9.35 / +8.85
MATH500	+6.49 / +4.84	+5.70 / +2.15	+10.73 / +6.93	+7.64 / +4.64
HumanEval	+10.45 / +15.63	+6.90 / +12.03	+9.55 / +10.77	+8.97 / +12.81
MBPP	+10.37 / +15.01	+7.39 / +12.04	+13.67 / +19.60	+10.48 / +15.55
Avg. over datasets	+9.13 / +12.45	+6.73 / +8.14	+11.46 / +10.80	+9.11 / +10.46

The comparison with prior training-free baselines further supports this interpretation. In the bottom row of each figure, the ADAS-augmented decoders generally remain competitive with or above KLASS and DAPD at moderate and large parallel-token counts. Both baselines often degrade rapidly as parallelism increases, whereas ADAS tends to degrade more smoothly. One possible explanation for the KLASS trend is that temporal-stability signals become less reliable when many tokens are committed at once, since consecutive denoising distributions can shift substantially. The DAPD trend suggests that hard attention-derived constraints may be brittle in some high-parallel regimes. Overall, these results indicate that soft attention-discounted reranking is a robust way to reduce harmful joint commitments among dependent positions.

To quantify the low-NFE regime, for baseline points averaging at least four tokens per step, we interpolate the ADAS curve at matched NFE. Table 2 reports mean gains across the three stopping rules as LLaDA / Dream.

The gains hold across entropy-, confidence-, and fixed-budget stopping rules and are strongest on code generation, where simultaneous coupled commitments can create inconsistent outputs. The smaller but positive MATH500 gains suggest that longer sequential chains, rather than pairwise incompatibilities within one update, may govern some errors.

Across 90 matched operating points, ADAS has mean gain $+9.27$ with a 95% bootstrap interval of $[+7.74, +10.84]$, improving in 80 cases and regressing in 10. This measures cross-operating-point stability, not fixed-setting significance; Appendix E gives the full breakdown.

6.3 Analysis of runtime overhead

ADAS reads last-layer attention and greedily updates token scores; accessing these weights requires disabling FlashAttention only in the final denoiser layer. Comparing each variant with its vanilla decoder under the same stopping rule and similar NFE, full-dataset wall-clock measurements show that attention extraction and reranking add 3.1% per model forward on average.

7 Discussion and limitations

ADAS is a training-free, dependency-aware sampler that improves highly parallel decoding across models, tasks, and stopping rules. Its central limitation is the proxy itself: attention may reflect syntax, position, or formatting, omits higher-order interactions, and weak attention does not guarantee compatibility. Although our intervention validates conditional influence on all four target benchmarks, evidence remains limited to two base MDLMs and reasoning/code; head- or layer-specific signals, larger models, and open-ended generation remain untested. The global $\alpha = 40$ was chosen on one model/benchmark, and confidence miscalibration can directly affect its uncertainty gate.

ADAS also commits irreversibly. Revise-or-remask could repair early errors after neighboring context arrives, but would change commitment dynamics and possibly NFE, making it complementary future work. Deterministic temperature-zero runs have no seed variance: the operating-point bootstrap measures cross-configuration stability, not fixed-setting significance, and sensitivity to resampled few-shot orderings remains untested. Higher-order signals, revision, and auxiliary verification are promising extensions.

The broader impacts of ADAS are mostly inherited from the underlying language models. ADAS may improve the inference efficiency and output quality of existing masked diffusion language models, which can reduce decoding cost and make these models more practical to use. At the same time, faster or higher-quality generation could indirectly increase misuse risks already associated with language models, such as automated spam, low-quality content generation, or other harmful downstream applications. ADAS does not introduce a new model, dataset, or deployed system, so these risks are inherited from the underlying models rather than specific to a newly released artifact.

Acknowledgments and Disclosure of Funding

This work was supported as part of the Swiss AI Initiative by a grant from the Swiss National Supercomputing Centre (CSCS) under project ID a137 on Alps.

References

Hikaru Asano, Tadashi Kozuno, Kuniaki Saito, and Yukino Baba. Where-to-unmask: Ground-truth-guided unmasking order learning for masked diffusion language models, 2026. URL `https://arxiv.org/abs/2602.09501`.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models, 2021. URL `https://arxiv.org/abs/2108.07732`.

Iskander Azangulov, Teodora Pandeva, Niranjani Prasad, Javier Zazo, and Sushrut Karmalkar. Parallel sampling from masked diffusion models via conditional independence testing. In *International Conference on Learning Representations*, 2026. URL `https://arxiv.org/abs/2510.21961`.

Wenrui Bao, Zhiben Chen, Dan Xu, and Yuzhang Shang. Learning to parallel: Accelerating diffusion large language models via learnable parallel decoding. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.25188.

Heli Ben-Hamu, Itai Gat, Daniel Severo, Niklas Nolte, and Brian Karrer. Accelerated sampling from masked diffusion models via entropy bounded unmasking. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2505.24857.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

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

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset, 2021. URL https://arxiv.org/abs/2103.03874.

Chunsan Hong, Seonho An, Min-Soo Kim, and Jong Chul Ye. Improving discrete diffusion unmasking policies beyond explicit reference policies. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.05725.

Daniel Israel, Guy Van den Broeck, and Aditya Grover. Accelerating diffusion LLMs via adaptive parallel decoding. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2506.00413.

Metod Jazbec, Theo X. Olausson, Louis Béthune, Pierre Ablin, Michael Kirchhof, João Monteiro, Victor Turrisi, Jason Ramapuram, and Marco Cuturi. Learning unmasking policies for diffusion language models, 2025. URL https://arxiv.org/abs/2512.09106.

Bumjun Kim, Dongjae Jeon, Moongyu Jeon, and Albert No. Dependency-aware parallel decoding via attention for diffusion LLMs, 2026. URL https://arxiv.org/abs/2603.12996.

Seo Hyun Kim, Sunwoo Hong, Hojung Jung, Youngrok Park, and Se-Young Yun. KLASS: KL-guided fast inference in masked diffusion models. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2511.05664.

Sanghyun Lee, Seungryong Kim, Jongho Park, and Dongmin Park. Lookahead unmasking elicits accurate decoding in diffusion language models, 2025. URL https://arxiv.org/abs/2511.05563.

Lizhuo Luo, Zhuoran Shi, Jiajun Luo, Zhi Wang, Shen Ren, Wenya Wang, and Tianwei Zhang. DAWN: Dependency-aware fast inference for diffusion LLMs, 2026. URL https://arxiv.org/abs/2602.06953.

Omer Luxembourg, Haim Permuter, and Eliya Nachmani. Plan for speed: Dilated scheduling for masked diffusion language models, 2025. URL https://arxiv.org/abs/2506.19037.

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, 2025. URL https://arxiv.org/abs/2502.09992.

Fred Zhangzhi Peng, Zachary Bezemek, Sawan Patel, Jarrid Rector-Brooks, Sherwood Yao, Avishek Joey Bose, Alexander Tong, and Pranam Chatterjee. Path planning for masked diffusion model sampling, 2025. URL https://arxiv.org/abs/2502.03540.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models, 2024. URL https://arxiv.org/abs/2406.07524.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. Simplified and generalized masked diffusion for discrete data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=xcqSOfHt4g.

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, 2025. URL https://arxiv.org/abs/2505.22618.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7B: Diffusion large language models, 2025. URL https://arxiv.org/abs/2508.15487.

Xueyu Zhou, Yangrong Hu, and Jian Huang. DOS: Dependency-oriented sampler for masked diffusion language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 17404–17419, 2026. URL https://aclanthology.org/2026.findings-acl.861/.

A Evaluation details

We evaluate on GSM8K, HumanEval, MBPP, and MATH500 with Dream 7B and LLaDA 8B. We use lm-eval-harness 0.4.8 with its default few-shot seed 1234: 8-shot prompting for GSM8K, 4-shot for MATH500, 3-shot for MBPP, and 0-shot for HumanEval. For batched inference, we pad each prompt to the task sequence length. Maximum generation length is 256 for GSM8K and 512 for HumanEval, MBPP, and MATH500. All evaluations use temperature 0. Dream uses Transformers 4.57.1 with flash SDPA disabled when attention extraction is required; its prompts and one-example generation were checked against the literal harness request.

Decoding controls. For the entropy-budget (EB) and Fast-dLLM families, we sweep the budget parameters γ and f. For the fixed-subset Top-k family, we sweep the subset size k directly.

Table 3: Hyperparameter ranges used for the three decoder families.

Method family	Sweep values
EB / EB+AD	$\gamma \in \{10^{-3}, 10^{-2}, 10^{-1}, 0.3, 1, 3, 10, 20\}$
Fast-dLLM	$f \in \{0.3, 1, 2, 3, 5, 8, 12\}$
Fast-dLLM+AD	$f \in \{0.3, 1, 2, 3, 5, 8, 12, 20\}$
Top-k / Top-k+AD	$k \in \{1, 2, 4, 8, 16\}$

For EB and Fast-dLLM, larger threshold permits more aggressive parallel unmasking. For Top-k, the level of parallelism is controlled directly by k, i.e., the number of tokens unmasked per denoising step. We include the extra $f = 20$ point for Fast-dLLM+AD only to probe the very high-parallel regime reached by the vanilla decoder at smaller thresholds; matched-NFE summaries use interpolation at baseline operating points rather than threshold-wise comparisons.

Compute resources. All evaluations were run as batched inference jobs on one 4 $\times$ GH200 GPU node per operating point. We used padding to the task maximum sequence length for batched inference. Dream 7B used batch size 4 per GPU, and LLaDA 8B used batch size 2 per GPU.

Hyperparameters for KLASS and DAPD. For KLASS and DAPD, we did not perform per-task accuracy tuning. Instead, we first ran small pilot evaluations to identify threshold values that place each method in the same broad low-NFE regime as our main curves, and then reused the same

Table 4: Actual wall-clock duration by model and dataset, aggregated over the completed full-dataset operating points used in the paper figures.

Model	Dataset	Min–max hrs/run
Dream 7B	GSM8K	0.13–2.21
Dream 7B	HumanEval	0.06–0.57
Dream 7B	MATH500	0.08–2.00
Dream 7B	MBPP	0.10–2.15
LLaDA 8B	GSM8K	0.26–4.86
LLaDA 8B	HumanEval	0.07–1.26
LLaDA 8B	MATH500	0.14–3.67
LLaDA 8B	MBPP	0.15–4.27

hyperparameter grids for all models and datasets. For KLASS, which accepts tokens satisfying both a confidence and KL criterion, we evaluated

$$(\theta_{\text{conf}}, \theta_{\text{KL}}) \in \{(0.70, 1.00), (0.198, 0.51), (0.18, 0.60), (0.16, 0.70)\}.$$

For DAPD, using the masked max-normalized attention dependency score, we evaluated

$$(\tau_{\min}, \tau_{\max}) \in \{(0.002, 0.02), (0.005, 0.05), (0.01, 0.10), (0.02, 0.20), (0.05, 0.50)\}.$$

These settings were chosen to span a range of parallelism levels rather than to maximize accuracy on any individual benchmark.

B Real-data dependency diagnostic details

Trajectory construction. For every example and model, we run the ordinary confidence Top-k trajectory with $k = 8$ and collect the partially masked outputs at the five unmasking steps listed below. MATH500 uses a 100-example subset balanced across its seven subjects (15 examples each from algebra and counting/probability and 14 from each remaining subject).

Table 5: Settings for the real-data intervention diagnostic.

Task	Examples	Shots	Length	Steps	Probed steps
GSM8K	100	8	256	32	0, 7, 15, 23, 31
MATH500	100	4	512	64	0, 15, 31, 47, 63
HumanEval	100	0	512	64	0, 15, 31, 47, 63
MBPP	100	3	512	64	0, 15, 31, 47, 63

At each state, let M be the masked positions and p_m^{before} the first-pass predictive distribution. We choose $i = \arg\max_{m \in M} c_m$, where $c_m = \max_v p_m^{\text{before}}(v)$, and set its deterministic prediction to $\hat{x}_i = \arg\max_v p_i^{\text{before}}(v)$. From this same pre-reveal pass, we extract head-mean final-layer A_{ji}, with query j and key i. In a non-mutating counterfactual copy, we reveal only i, rerun the model, and record

$$D_{\text{KL}}\left(p_j^{\text{before}} \middle\| p_j^{\text{after}}\right), \qquad \Delta c_j = \max_v p_j^{\text{after}}(v) - \max_v p_j^{\text{before}}(v).$$

The original trajectory then commits its ordinary Top-k set. The primary scope contains the other seven co-selected positions, giving 3,500 pairs per task/model. This scope directly measures dependencies among positions the baseline would commit together.

Estimands and uncertainty. The correlation ρ combines all 3,500 pairs. To avoid conflating token ranking with differences across examples and denoising stages, we also compute Spearman correlation independently within each state and report their mean $\bar{\rho}_s$. Intervals for these means are normal intervals across states. For paired comparisons, we first average the five states per example and bootstrap the 100 example-level means. Since i and c_i are fixed within a state, A_{ji} and $A_{ji}(1-c_i)$ have identical statewise ranks; gating is evaluated as cross-state calibration through ρ.

Table 6: Statewise intervention results with uncertainty intervals. Gain is $\bar{\rho}_s(c_j^0, K_j) - \bar{\rho}_s(s_j, K_j)$. Gain intervals use a paired bootstrap over 100 examples.

Task	Model	$\bar{\rho}_s(A_{ji}, K_j)$ [95% interval]	Gain [95% interval]
GSM8K	Dream	.358 [.312, .403]	.325 [.295, .358]
	LLaDA	.360 [.312, .409]	.405 [.368, .444]
HumanEval	Dream	.305 [.265, .346]	.178 [.143, .216]
	LLaDA	.389 [.340, .437]	.463 [.402, .525]
MBPP	Dream	.204 [.161, .247]	.097 [.074, .123]
	LLaDA	.251 [.202, .301]	.193 [.165, .221]
MATH500	Dream	.316 [.274, .359]	.193 [.164, .223]
	LLaDA	.436 [.393, .480]	.395 [.347, .445]

Delayed-token stability. After the common first token i, we compare the next candidate under confidence Top-8 with the exact Top-8+AD rule. In 1,098 states they differ, defining exactly one deferred token and one ADAS replacement per state. We analyze this first divergence, not all eight selections, and measure both tokens after revealing i.

Table 7: Change after receiving newly revealed context, reported as deferred token / ADAS replacement.

Model	KL change	Mean Δc_j	Confidence increased
LLaDA	.671 / .197	+.084 / +.035	86.2% / 64.5%
Dream	.914 / .263	+.101 / +.013	87.4% / 61.2%
Overall	**.793 / .230**	**+.092 / +.024**	**86.8% / 62.9%**

The confidence-gain direction holds in all eight task/model settings, while KL is higher for deferred tokens in seven of eight (Dream/MBPP is effectively tied). Because free-form solutions do not provide a unique token-aligned target, this diagnostic measures stability rather than token correctness; correctness is measured by end-to-end task scores.

C Theoretical motivation for the ADAS utility

We provide a local derivation motivating the ADAS utility. The argument is not meant to establish a global guarantee for an arbitrary transformer; rather, it shows that under a standard first-order approximation of the denoiser around the current decoding state, the natural dependence penalty has the same form as the ADAS discount.

Fix the current partially unmasked sequence $x^{\bar{\mathcal{M}}}$. For each masked position s, let

$$p_s(a) = p_\theta(x^s = a \mid x^{\bar{\mathcal{M}}}), \qquad \hat{x}_s = \arg\max_{a \in \mathcal{V}} p_s(a), \qquad c_s = p_s(\hat{x}_s).$$

Let $Z_s = \phi_s(X^s)$ be the representation inserted at position s, and let $\hat{z}_s = \phi_s(\hat{x}_s)$. Since the vocabulary is finite, define

$$D_\phi^2 = \max_{s \in \mathcal{M}} \max_{a,b \in \mathcal{V}} \|\phi_s(a) - \phi_s(b)\|_2^2.$$

Then

$$\begin{aligned}
\mathbb{E}\big[\|Z_s - \hat{z}_s\|_2^2\big] &= \sum_{a \in \mathcal{V}} p_s(a) \|\phi_s(a) - \phi_s(\hat{x}_s)\|_2^2 \\
&\le D_\phi^2 \sum_{a \neq \hat{x}_s} p_s(a) \\
&= D_\phi^2 (1 - c_s).
\end{aligned}$$

Thus $1 - c_s$ upper-bounds the representation uncertainty of the selected source token s.

Now consider appending a candidate position i after an already selected set S. Let $\ell_i(z_S)$ denote the logits at position i when the selected positions are filled with representations $z_S = \{z_s : s \in S\}$. Around the current decoding state $\hat{z}_S$, a first-order expansion gives

$$\ell_i(z_S) \approx \ell_i(\hat{z}_S) + \sum_{s \in S} J_{is}(z_s - \hat{z}_s),$$

where J_{is} is the Jacobian of the logits at position i with respect to the representation at position s.

For a self-attention layer, the direct message from source s to target i is weighted by the attention coefficient A_{is}. We therefore use the empirical attention matrix as a tractable proxy for the relative sensitivity of i to perturbations at s, and write the local approximation

$$\|J_{is}\|_2^2 \approx L_i^2 A_{is},$$

where L_i absorbs the remaining local linear maps, residual transformations, and normalization factors. This approximation says that source positions receiving larger attention from i have larger local influence on the logits at i.

Using this approximation,

$$\begin{aligned}\mathbb{E}\big[\|\ell_i(Z_S) - \ell_i(\hat{z}_S)\|_2^2\big] &\approx \mathbb{E}\left[\left\|\sum_{s\in S} J_{is}(Z_s - \hat{z}_s)\right\|_2^2\right] \\ &\lesssim L_i^2 \sum_{s\in S} A_{is}\mathbb{E}\big[\|Z_s - \hat{z}_s\|_2^2\big] \\ &\le L_i^2 D_\phi^2 \sum_{s\in S} A_{is}(1-c_s).\end{aligned}$$

Finally, for softmax probabilities, local logit perturbations induce KL changes that are second order in the perturbation size:

$$D_{\mathrm{KL}}\left(\mathrm{softmax}(\ell_i + \delta) \,\|\, \mathrm{softmax}(\ell_i)\right) \le \frac{1}{2}\|\delta\|_2^2.$$

Therefore, the incremental dependence risk of adding candidate i after S is locally controlled by a quantity proportional to

$$\sum_{s\in S} A_{is}(1-c_s).$$

This motivates the risk-adjusted marginal utility

$$u(i \mid S, A) = c_i - \alpha \sum_{s\in S} A_{is}(1-c_s),$$

where $\alpha \ge 0$ absorbs the constants in the local approximation and controls the strength of the dependence penalty.

D Hyperparameter and ablation studies

Attention-discount strength We ablate the attention-discount strength α on LLaDA-8B HumanEval under three decoding families: fixed Top-k, AD+EB, and Fast-dLLM+AD. For fixed Top-k, we test $k \in \{8, 16\}$. For AD+EB, we use two entropy-budget settings chosen to match these operating regimes approximately: $\gamma = 10$ (low NFE) and $\gamma = 3$ (medium NFE). For AD Fast-dLLM, we similarly use $f = 5$ and $f = 2$. In all cases, we sweep

$$\alpha \in \{1, 5, 10, 20, 40, 80\},$$

and compare against the corresponding no-AD baseline, which we denote as $\alpha = 0$.

The best α depends on the decoding regime when viewed per setting. For example, fixed Top-k prefers smaller values in some cases, whereas the adaptive stopping-rule methods are more favorable to larger values. However, when averaging across all six matched settings in Table 8, $\alpha = 40$ achieves the highest overall accuracy. Relative to an oracle that selects the best α separately for each row, the global default loses only 0.72 points on average and at most 2.44 points in the worst case. We therefore use $\alpha = 40$ as a single global default in the main experiments, avoiding task- or regime-specific retuning while remaining competitive across the tested settings.

Table 8: Pass@1 accuracy (%) for the α sweep on LLaDA HumanEval across fixed Top-k, AD+EB, and AD Fast-dLLM. The no-AD baseline is shown as $\alpha = 0$.

Setting	$\alpha = 0$	$\alpha = 1$	$\alpha = 5$	$\alpha = 10$	$\alpha = 20$	$\alpha = 40$	$\alpha = 80$
Top-k, $k = 16$	1.83	4.88	7.32	5.49	5.49	4.88	4.88
Top-k, $k = 8$	5.49	9.15	18.90	18.29	21.34	20.12	18.90
AD+EB, $\gamma = 10$	9.76	9.15	20.12	23.78	23.78	23.17	23.78
AD+EB, $\gamma = 3$	21.34	20.73	23.17	22.56	26.22	26.83	22.56
Fast-dLLM+AD, $f = 5$	9.76	9.15	21.95	23.78	24.39	24.39	21.95
Fast-dLLM+AD, $f = 2$	28.05	29.27	27.44	28.05	30.49	33.54	33.54
Average over 6 settings	–	13.72	19.82	20.33	21.95	**22.15**	20.93

Table 9: Ablation of the uncertainty weighting term in the attention discount on LLaDA-8B-Base HumanEval with Top-k+AD decoding at $k = 8$. The default selected-token uncertainty $1 - c_s$ outperforms using the average uncertainty over masked positions $1 - c_{\text{mask}}$.

Uncertainty weighting term	Pass@1 (%)
Selected-token uncertainty $1 - c_s$	**20.12**
Average masked-position uncertainty $1 - c_{\text{mask}}$	17.07

Ablation on the uncertainty weighting term. We additionally ablate the uncertainty factor used inside the attention discount. The default ADAS utility discounts a candidate position according to the uncertainty of the already selected position, using the factor $1 - c_s$. We compare this choice against a variant that replaces $1 - c_s$ with a global average uncertainty over the currently masked positions, denoted $1 - c_{\text{mask}}$. This ablation is conducted on LLaDA-8B-Base with HumanEval under Top-k+AD decoding at $k = 8$.

The selected-token uncertainty weighting improves Pass@1 by 3.05 absolute points over the global masked-position uncertainty variant. This suggests that the discount benefits from conditioning on the uncertainty of the specific selected position rather than using a single global uncertainty scalar.

Effect of attention source We ablate which layer is used to construct the attention matrix in ADAS. We use LLaDA-8B-Base on HumanEval with fixed Top-k decoding at $k = 8$, so all variants use the same number of denoising steps. The no-ADAS baseline ranks tokens only by confidence, while the ADAS variants use mean attention over heads from the specified transformer layer.

Table 10: Attention-source ablation for ADAS on LLaDA-8B HumanEval with fixed Top-k decoding at $k = 8$. Final-layer attention performs best, suggesting that later attention patterns provide the most useful dependency signal for selecting tokens to unmask in parallel.

Selection rule	Pass@1 (%)
Top-k confidence, no ADAS	5.49
Top-k+AD with first-layer attention	9.76
Top-k+AD with middle-layer attention	15.24
Top-k+AD with final-layer attention	**20.12**

E Robustness across operating points

We perform an additional robustness analysis over matched operating points. Each operating point corresponds to a model–dataset–stopping-rule–hyperparameter configuration. This analysis is intended to assess whether the observed gains are stable across the evaluated settings, rather than driven by a small number of favorable thresholds or tasks.

For each vanilla baseline point in the low-NFE regime, defined as having at least four parallel tokens per denoising step on average, we linearly interpolate the corresponding ADAS curve at the same NFE and compute the matched gain

$$\Delta = \text{score}_{\text{ADAS}}(\text{NFE}) - \text{score}_{\text{base}}(\text{NFE}).$$

Baseline points outside the interpolation range of the corresponding ADAS curve are excluded. We then bootstrap the matched gains over operating points with 10,000 resamples and report the mean gain and central 95% bootstrap interval. Because operating points from the same model, dataset, or decoding family are correlated, this is a descriptive stability summary across evaluated configurations, not a confidence interval for a fixed setting or a paired per-example significance test.

Table 11: Robustness of ADAS gains across matched operating points. Gains are absolute accuracy/pass@1 differences in percentage points after matching by NFE. Bootstrap intervals use 10,000 operating-point resamples and summarize cross-configuration stability.

Method family	Matched points	Mean gain	95% bootstrap interval	Positive / negative
EB-Sampler	28	+10.42	[+7.30, +13.67]	24 / 4
Fast-dLLM	38	+7.24	[+5.42, +9.14]	34 / 4
Top-k	24	+11.13	[+8.31, +13.97]	22 / 2
All	90	+9.27	[+7.74, +10.84]	80 / 10

Table 12: Robustness of ADAS gains across datasets under the same matched-operating-point analysis as Table 11.

Dataset	Matched points	Mean gain	95% bootstrap interval	Positive / negative
GSM8K	19	+8.60	[+4.71, +12.48]	14 / 5
HumanEval	22	+10.33	[+8.10, +12.50]	21 / 1
MATH500	24	+5.73	[+3.84, +7.64]	21 / 3
MBPP	25	+12.24	[+8.99, +15.59]	24 / 1

F Controlled synthetic dependency diagnostics

We complement the real-data intervention with arithmetic predicates whose dependency structure is known. For the first diagnostic, each example contains three sum or product predicates and exactly two masked numeric literals. In the dependent condition, both masks occur in the same predicate; in the independent condition, they occur in different predicates. Dream-7B-Base predicts both literals in one forward pass without confidence ordering, adaptive stopping, or attention discounting. Under constraint-satisfaction scoring, dependent pairs are substantially harder: accuracy falls from 71% for independent pairs to 30% for dependent pairs.

The second diagnostic includes sum, product, minimum, and maximum predicates. We compare final-layer, head-mean attention between masked tokens from the same predicate with pairs from different predicates. Figure 3 and Table 13 show higher attention for known-dependent pairs on both models. This controlled result isolates dependency structure, while Section 4 establishes the corresponding intervention relationship on the target tasks.

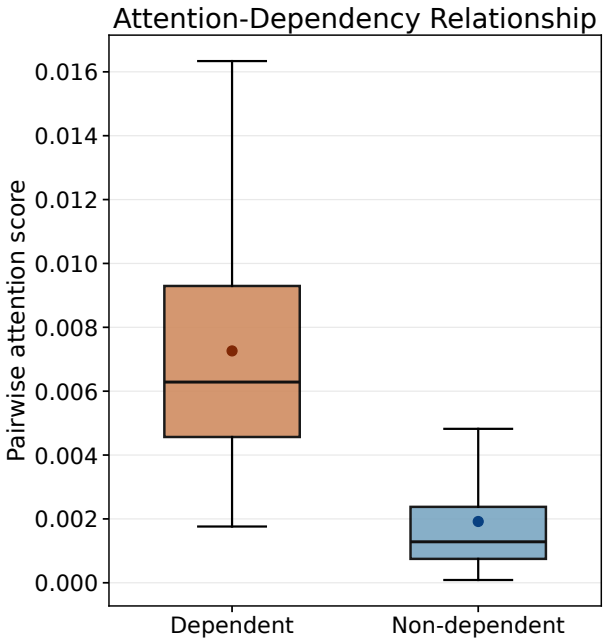

Figure 3: Pairwise attention for dependent and non-dependent masked-token pairs in the synthetic diagnostic on LLaDA-8B-Base.

Dataset Construction

Assume the set of equation types

$$E = \{A + B + C = D,\ A \cdot B \cdot C = D,\ \min\{A, B, C\} = D,\ \max\{A, B, C\} = D\}$$

is given, together with the number of predicates n and masking ratio r.

1. Sample predicate indices. We sample a sequence

$$\pi = (\pi_1, \ldots, \pi_n), \qquad \pi_i \in \{0, \ldots, |E| - 1\},$$

and construct the i-th predicate from the equation type $E[\pi_i]$.

2. Instantiate the predicates. Each predicate contains four integer slots (A, B, C, D), so the full sequence contains $4n$ integer variables. We sample positive integers for A, B, C and compute D according to the selected equation type.

3. Mask integer variables. We randomly mask

$$r \cdot 4n$$

of the integer positions in the sequence. Only the integers are masked; the operators and relation symbols $(+, \cdot, \min, \max, =)$ are never masked.

Figure 4: Illustration of the toy data-generation process. The construction yields a masked sequence containing masked-token pairs with known dependency structure.

Table 13: Attention-dependency relationship on the shared toy dataset. Reported values are mean pairwise attention for dependent and non-dependent masked pairs aggregated as mean of attention heads from the last layer.

Model	Dep. Attn.	Non-Dep. Attn.
LLaDA-8B-Base	0.007261	0.001919
Dream-7B-Base	0.009517	0.004842

Table 14: Results on **Dream-7B-Base** with EB stopping rule. Each cell reports **score** on the first line and average NFE over samples on the second line. For GSM8K and MATH500, score is accuracy; for HumanEval and MBPP, score is Pass@1.

	GSM8K		HumanEval		MATH500		MBPP	
γ	EB	EB + AD	EB	EB + AD	EB	EB + AD	EB	EB + AD
0.001	**75.06** 240.50	**75.06** 240.37	**53.66** 453.44	**53.66** 453.06	**40.00** 386.65	**40.00** 384.99	**53.80** 432.49	**53.80** 431.56
0.01	**75.13** 209.23	**75.13** 208.93	**54.27** 372.90	**54.27** 372.83	**39.40** 269.48	**39.20** 266.06	**53.60** 305.49	**53.60** 302.70
0.1	**74.60** 166.87	**74.30** 165.76	**54.27** 289.84	**54.88** 287.95	**39.80** 168.08	**40.00** 163.52	**53.00** 179.38	**53.00** 174.84
0.3	**74.68** 143.42	**74.45** 142.59	**52.44** 237.53	**53.66** 236.08	**38.80** 131.00	**39.20** 122.17	**53.80** 131.58	**53.40** 126.62
1	**69.98** 110.42	**67.32** 104.99	**46.34** 172.41	**45.73** 162.49	**34.80** 91.65	**33.00** 89.23	**50.40** 86.70	**47.80** 82.90
3	**57.09** 70.41	**50.64** 61.00	**30.49** 116.12	**35.37** 85.06	**24.40** 62.27	**25.20** 60.81	**36.60** 55.71	**43.20** 54.47
10	**28.66** 34.29	**33.28** 24.21	**12.20** 54.01	**25.61** 36.43	**9.60** 34.24	**20.20** 27.88	**11.80** 36.50	**35.00** 26.77
20	**7.96** 17.66	**23.20** 13.36	**3.05** 32.32	**20.12** 22.63	**3.60** 20.12	**9.00** 17.85	**4.20** 22.82	**31.00** 16.43

G Detailed results

Table 15: Results on **LLaDA-8B-Base** with EB stopping rule. Each cell reports **score** on the first line and average NFE over samples on the second line. For GSM8K and MATH500, score is accuracy; for HumanEval and MBPP, score is Pass@1.

	GSM8K		HumanEval		MATH500		MBPP	
γ	EB	EB + AD	EB	EB + AD	EB	EB + AD	EB	EB + AD
0.001	**69.90** 237.17	**69.90** 236.92	**35.37** 457.12	**35.37** 456.04	**28.40** 460.24	**28.40** 459.09	**38.80** 475.93	**38.80** 475.47
0.01	**69.83** 190.83	**69.83** 189.83	**35.37** 326.08	**35.37** 321.84	**28.20** 336.50	**28.00** 332.56	**38.80** 338.33	**38.80** 334.52
0.1	**70.81** 136.13	**70.96** 134.03	**35.98** 190.35	**35.98** 178.60	**27.80** 209.92	**28.00** 203.39	**39.20** 197.32	**39.20** 191.63
0.3	**70.20** 110.54	**70.28** 107.93	**34.76** 135.54	**34.15** 126.72	**28.80** 164.38	**28.40** 157.04	**39.40** 144.04	**39.20** 136.83
1	**67.40** 78.10	**67.17** 73.56	**29.27** 87.19	**29.88** 75.84	**29.00** 109.11	**25.80** 98.88	**36.80** 95.70	**38.40** 86.68
3	**60.20** 50.36	**55.12** 44.89	**21.34** 59.38	**26.83** 43.71	**21.40** 69.33	**26.00** 68.29	**28.40** 63.66	**34.40** 55.08
10	**38.21** 29.82	**41.17** 22.14	**9.76** 32.90	**23.17** 23.95	**10.80** 36.02	**18.80** 32.59	**11.20** 33.34	**26.00** 29.61
20	**13.72** 17.70	**32.83** 14.24	**4.88** 23.41	**21.95** 17.90	**3.20** 20.86	**18.60** 20.09	**5.00** 20.60	**21.20** 19.75

Table 16: Results on **Dream-7B-Base** with Fast-dLLM stopping rule. Each cell reports **score** on the first line and average NFE over samples on the second line. For GSM8K and MATH500, score is accuracy; for HumanEval and MBPP, score is Pass@1.

	GSM8K		HumanEval		MATH500		MBPP	
f	FD	FD + AD	FD	FD + AD	FD	FD + AD	FD	FD + AD
0.3	**74.45** 167.07	**74.98** 167.59	**55.49** 273.27	**55.49** 275.57	**40.40** 146.80	**40.60** 148.19	**54.00** 138.66	**54.00** 139.72
1	**73.01** 132.59	**73.31** 134.97	**50.61** 201.44	**51.22** 205.78	**38.20** 104.11	**38.40** 107.50	**51.00** 90.36	**51.60** 92.76
2	**63.31** 95.36	**65.05** 98.78	**36.59** 139.48	**42.07** 147.07	**31.60** 75.19	**32.80** 78.06	**38.00** 63.20	**45.40** 65.30
3	**51.86** 55.60	**52.84** 58.71	**23.17** 85.32	**34.15** 84.87	**22.40** 56.29	**27.40** 62.73	**29.40** 48.96	**40.20** 53.23
5	**27.90** 29.12	**40.94** 32.49	**14.63** 46.10	**26.83** 49.23	**9.60** 30.52	**22.40** 40.51	**15.20** 33.78	**36.40** 35.64
8	**7.96** 14.47	**26.91** 18.96	**7.93** 29.98	**22.56** 32.28	**4.00** 17.73	**10.80** 27.31	**8.00** 21.09	**32.40** 22.91
12	**1.74** 8.17	**11.07** 11.04	**6.10** 21.65	**18.29** 22.38	**0.80** 11.92	**9.00** 19.72	**2.40** 13.87	**27.80** 16.16

Table 17: Results on **LLaDA-8B-Base** with Fast-dLLM stopping rule. Each cell reports **score** on the first line and average NFE over samples on the second line. For GSM8K and MATH500, score is accuracy; for HumanEval and MBPP, score is Pass@1.

	GSM8K		HumanEval		MATH500		MBPP	
f	FD	FD + AD	FD	FD + AD	FD	FD + AD	FD	FD + AD
0.3	**69.98** 125.16	**69.98** 126.02	**35.37** 132.63	**35.37** 134.80	**28.40** 182.54	**28.20** 184.13	**38.60** 153.71	**38.60** 154.44
1	**69.22** 91.28	**69.29** 93.16	**35.37** 86.65	**35.37** 87.59	**27.20** 128.43	**27.60** 131.62	**39.00** 101.49	**39.40** 102.71
2	**64.22** 62.60	**63.84** 65.10	**28.05** 60.01	**33.54** 64.59	**24.40** 85.34	**27.80** 94.18	**30.80** 67.83	**37.00** 71.91
3	**55.57** 46.39	**56.18** 47.26	**18.90** 47.09	**27.44** 44.57	**19.80** 63.04	**25.80** 79.03	**21.00** 48.29	**32.60** 58.54
5	**39.04** 30.29	**45.49** 31.37	**9.76** 29.43	**24.39** 34.63	**12.20** 33.48	**19.60** 49.76	**10.40** 29.45	**28.60** 43.96
8	**15.62** 16.29	**40.64** 21.56	**4.88** 19.46	**17.07** 26.01	**2.40** 18.20	**19.60** 34.05	**6.40** 17.98	**25.20** 31.80
12	**4.32** 8.96	**32.37** 15.12	**3.05** 13.44	**10.37** 10.68	**0.20** 11.32	**14.60** 21.99	**2.60** 13.03	**19.20** 23.37
20						**6.20** 9.96		**6.40** 6.76

Table 18: Results on **Dream-7B-Base** with Top-k stopping rule. Each cell reports **score** on the first line and average NFE over samples on the second line. For GSM8K and MATH500, score is accuracy; for HumanEval and MBPP, score is Pass@1.

	GSM8K		HumanEval		MATH500		MBPP	
k	Top-k	Top-k + AD	Top-k	Top-k + AD	Top-k	Top-k + AD	Top-k	Top-k + AD
1	**74.91** 256.00	**74.91** 256.00	**55.49** 512.00	**55.49** 512.00	**40.80** 512.00	**40.80** 512.00	**54.00** 512.00	**54.00** 512.00
2	**64.67** 128.00	**58.15** 128.00	**42.68** 256.00	**43.29** 256.00	**31.40** 256.00	**32.60** 256.00	**44.60** 256.00	**46.40** 256.00
4	**48.45** 64.00	**46.47** 64.00	**20.73** 128.00	**35.98** 128.00	**16.20** 128.00	**25.00** 128.00	**29.20** 128.00	**40.20** 128.00
8	**13.80** 32.00	**28.81** 32.00	**12.20** 64.00	**21.95** 64.00	**3.60** 64.00	**11.40** 64.00	**11.40** 64.00	**34.40** 64.00
16	**0.91** 16.00	**5.53** 16.00	**7.93** 32.00	**15.24** 32.00	**1.20** 32.00	**5.40** 32.00	**2.00** 32.00	**26.80** 32.00

Table 19: Results on **LLaDA-8B-Base** with Top-k stopping rule. Each cell reports **score** on the first line and average NFE over samples on the second line. For GSM8K and MATH500, score is accuracy; for HumanEval and MBPP, score is Pass@1.

	GSM8K		HumanEval		MATH500		MBPP	
k	Top-k	Top-k + AD	Top-k	Top-k + AD	Top-k	Top-k + AD	Top-k	Top-k + AD
1	**69.98** 256.00	**69.98** 256.00	**35.37** 512.00	**35.37** 512.00	**28.40** 512.00	**28.40** 512.00	**38.80** 512.00	**38.80** 512.00
2	**64.22** 128.00	**59.67** 128.00	**25.61** 256.00	**26.83** 256.00	**22.80** 256.00	**24.40** 256.00	**28.80** 256.00	**36.80** 256.00
4	**53.90** 64.00	**50.11** 64.00	**13.41** 128.00	**24.39** 128.00	**17.00** 128.00	**23.00** 128.00	**17.60** 128.00	**29.80** 128.00
8	**22.67** 32.00	**42.46** 32.00	**5.49** 64.00	**20.12** 64.00	**3.40** 64.00	**19.20** 64.00	**7.00** 64.00	**25.20** 64.00
16	**2.20** 16.00	**21.91** 16.00	**1.83** 32.00	**4.88** 32.00	**0.20** 32.00	**10.60** 32.00	**1.60** 32.00	**12.20** 32.00

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims are stated in the abstract and introduction and are supported by the method description in Section 5.2, the main experimental results in Section 6.2, and the runtime analysis in Section 6.3. The scope is limited to masked diffusion language models, the evaluated benchmarks, and the stopping criteria considered in the experiments.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in Section 7, including the use of self-attention only as a proxy for stochastic dependence, the limited model and task scope, and the fact that robustness intervals summarize variation across matched operating points rather than repeated seeds or per-example uncertainty.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present formal theoretical results, theorems, or proofs. The error decomposition in Section 3.3 is used as motivation from prior work, while the proposed ADAS method is evaluated empirically in Section 6.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide a detailed description of the evaluation settings and proposed method in the paper, and include in the supplemental material a link to a public repository containing the code, commands, and instructions needed to reproduce the main experiments. The implementation builds on the standard language model evaluation framework `lm-eval`.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide reproduction code and instructions in the supplemental material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The proposed ADAS decoding method is described in Section 5.2, including the attention-based interaction matrix, marginal utility, greedy subset construction, stopping-rule-agnostic formulation, and computational complexity. Algorithm 1 gives pseudocode, and Appendix A specifies the experimental methodology, including benchmarks, models, prompting setup, decoding sweep values, batch sizes, and compute resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: The paper reports a robustness analysis over matched low-NFE operating points in Appendix E. Specifically, for each baseline operating point, we interpolate the corresponding ADAS curve at the same NFE and compute the matched accuracy/pass@1 gain. We then bootstrap these matched gains over operating points using 10,000 resamples and report the mean gain, 95% bootstrap interval, and the number of positive and negative matched comparisons for each stopping-rule family. These intervals are intended as a robustness summary across evaluated model–dataset–sampler configurations rather than as repeated-seed uncertainty estimates.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide extensive information regarding the compute resources in Section A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research complies with the NeurIPS Code of Ethics: it does not involve human subjects, crowdsourcing, private or newly collected personal data, surveillance data, or high-risk deployment. Experiments use standard public benchmarks and existing

open-weight models, and the main ethical considerations are indirect risks from improving language-model generation speed and quality, which are discussed as inherited from downstream use of the underlying models.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 7 discusses broader impacts. ADAS may improve the inference efficiency and output quality of existing masked diffusion language models, reducing decoding cost and making these models more practical to use. The paper also discusses potential negative impacts: faster or higher-quality generation could indirectly increase misuse risks already associated with language models, such as automated spam, low-quality content generation, or other harmful downstream applications. The paper clarifies that ADAS does not introduce a new model, dataset, or deployed system, so these risks are inherited from the underlying models rather than specific to a newly released artifact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: ADAS does not release a new model, dataset, or deployed system requiring additional release safeguards. The indirect misuse risks discussed in Section 7 are inherited from the underlying language models rather than from a newly released high-risk artifact.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use only existing public benchmarks, models, and evaluation tooling. GSM8K is released under the MIT License; MATH/MATH500 is derived from the MATH dataset repository, released under the MIT License; MBPP is released under CC-BY 4.0; HumanEval is released under the MIT License. We evaluate open-weight Dream-7B-Base, released under Apache-2.0, and LLaDA-8B-Base, released under the MIT License. We use the EleutherAI LM Evaluation Harness, released under the MIT License. We use these assets only for evaluation and cite the corresponding papers or repositories.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do not introduce a new dataset, model, or benchmark. The reproduction code is documented through the supplemental material, while all evaluated datasets and models are existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our work doesn't involve research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our work doesn't involve research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Our usage of LLMs in this work is limited to proofreading of the manuscript.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.