STABILITY-WEIGHTED DECODING FOR DIFFUSION LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion large language models (dLLMs) enable parallel text generation by iteratively denoising a fully masked sequence. However, existing decoding strategies rely on single-step static confidence metrics, ignoring temporal history and often prematurely unmasking unstable tokens. Under idealized ground-truth context revelation, we establish that the expected KL divergence between true conditional distributions equals the information gained about a token and lower-bounds its dependence on the previously masked context. Inspired by this result, we propose Stability-Weighted Decoding (SWD), a training-free, plug-and-play strategy that incorporates temporal stability into token scoring, using model-based temporal KL as an empirical sensitivity signal rather than a dependency guarantee. Experiments on code generation and mathematical reasoning benchmarks demonstrate that SWD improves generation accuracy across representative scoring metrics and exhibits exceptional robustness under aggressive acceleration ratios.

1 INTRODUCTION

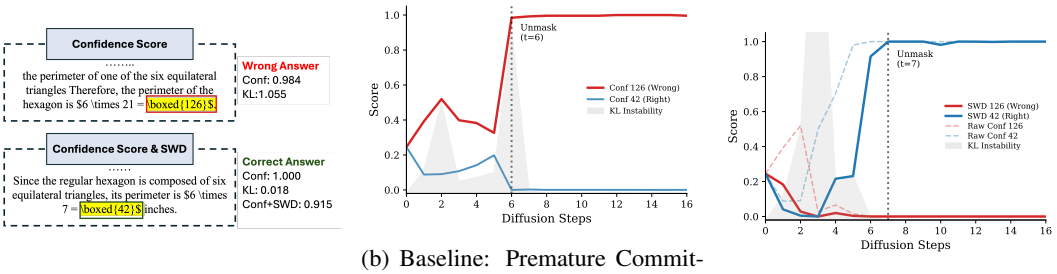
Discrete diffusion models (Austin et al., 2021a) have emerged as a compelling paradigm for generative tasks, with Masked Diffusion Models (MDMs) (Sahoo et al., 2024) demonstrating exceptional performance in text generation. By scaling this specific architecture, recent works like LLaDA (Nie et al., 2025) and Dream (Ye et al., 2025) have established the class of Diffusion Large Language Models (dLLMs) (Bie et al., 2025; Zhu et al., 2025; Gong et al., 2025), challenging the dominance of autoregressive (AR) methods. Unlike AR models constrained to left-to-right processing, dLLMs leverage bidirectional attention and an iterative denoising process, and conceptually characterized as an any-order autoregressive process (Ou et al., 2024; Kim et al., 2025a), possess higher potential than their fixed-order counterparts. This unique mechanism enables parallel decoding capability to predict and unmask multiple tokens simultaneously (Ben-Hamu et al., 2025; Wu et al., 2025). However, realizing this potential requires a critical decision at each step: accurately identifying which subset of tokens is sufficiently reliable to unmask.

Current decoding strategies predominantly rely on static confidence metrics (e.g., confidence (Nie et al., 2025), margin (Kim et al., 2025a), entropy (Koh et al., 2025)) derived from a single timestep. These scores measure certainty at the current step but do not explicitly measure how predictions change as the context evolves. As illustrated in Figure 1, tokens often exhibit temporal instability — transient high confidence with significant fluctuations. Relying solely on static metrics may lead to premature commitment, where unstable tokens are unmasked too early, locking in errors that propagate globally.

To motivate the use of temporal dynamics, we establish an information-theoretic result under idealized context revelation: the expected KL divergence between true conditional distributions equals the information gained from newly revealed tokens and lower-bounds dependence on the previously masked context. Inspired by this result, we introduce Stability-Weighted Decoding (SWD), which modulates base scores using an exponential penalty on temporal KL between model predictions. The exponential penalty is a design choice intended to reduce premature commitment, whose effectiveness we evaluate empirically.

We comprehensively validate SWD across state-of-the-art dLLMs (LLaDA-8B and Dream-7B) on diverse code generation and mathematical reasoning benchmarks. Empirical results demonstrate that

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.



(a) Case Study: Hexagon Problem

(b) Baseline: Premature Commitment

(c) SWD: Stability-Aware Selection

Figure 1: We visualize the decoding trajectory for a hexagon geometry problem in Math500. **(a)** The Baseline (Confidence-only) generates an incorrect answer "126", while SWD correctly outputs "42". **(b)** Although the incorrect token '126' exhibits high instability (large KL divergence, grey area), the model prematurely unmasks it at $t = 6$ due to transient high confidence. **(c)** SWD applies a stability penalty. The high instability of '126' suppresses its score (Red solid line), preventing the error. The correct token '42' is selected only after it stabilizes later in the diffusion process.

SWD improves generation accuracy and exhibits exceptional robustness under aggressive acceleration regimes. Our main contributions are:

- We establish an information-gain identity and dependency lower bound under idealized ground-truth context revelation, motivating temporal sensitivity as a decoding signal.
- We propose SWD, a training-free strategy that seamlessly integrates with any score-based policy by explicitly penalizing temporally unstable tokens.
- Comprehensive evaluations demonstrate that SWD enhances generation accuracy across diverse models, datasets, and selection strategies.

2 Related Work

The application of diffusion models to discrete data (Austin et al., 2021a; Campbell et al., 2022; Lou et al., 2024; Arriola et al., 2025) has culminated in masked diffusion models (MDMs) (Sahoo et al., 2024), which utilize absorbing states and demonstrate strong empirical performance in language generation. This paradigm has recently been scaled to large instruction-tuned models, such as the LLaDA and Dream families (Nie et al., 2025; Ye et al., 2025), challenging the long-standing dominance of autoregressive language models. While ongoing efforts continue to scale capabilities (Gong et al., 2025; Zhu et al., 2025; Bie et al., 2025; Song et al., 2025), extend to multimodal settings (You et al., 2025; Yang et al., 2025; Yu et al., 2025), and explore theoretical limitations (Feng et al., 2025; Zheng et al., 2025; Kim et al., 2025a), realizing the full parallel generation potential of dLLMs remains fundamentally bottlenecked by their decoding strategies.

Decoding quality and efficiency in dLLMs are governed by how masked tokens are selected for unmasking at each iteration. Early approaches using uniform unmasking schedules (Nie et al., 2025) have been superseded by score-based criteria—such as confidence, entropy (Koh et al., 2025), or margin (Kim et al., 2025a; Huang et al., 2026). To accelerate generation, methods like Fast-dLLM (Wu et al., 2025) and EB-Sampler (Ben-Hamu et al., 2025) unmask multiple tokens simultaneously using threshold-based or entropy-bound rules. Instead of using static scores, recent studies exploit temporal dynamics (Kim et al., 2025b; Wang et al., 2025b; Chen et al., 2025). For instance, Credit Decoding (Wang et al., 2025a) aggregates historical logits across multiple steps to smooth predictions and KLASS (Kim et al., 2025b) use step-wise information and confidence score to select tokens. Our method (SWD) is inspired by an information-theoretic bound under idealized context revelation. It uses instantaneous step-wise divergence as an empirical sensitivity penalty without requiring a multi-step history buffer.

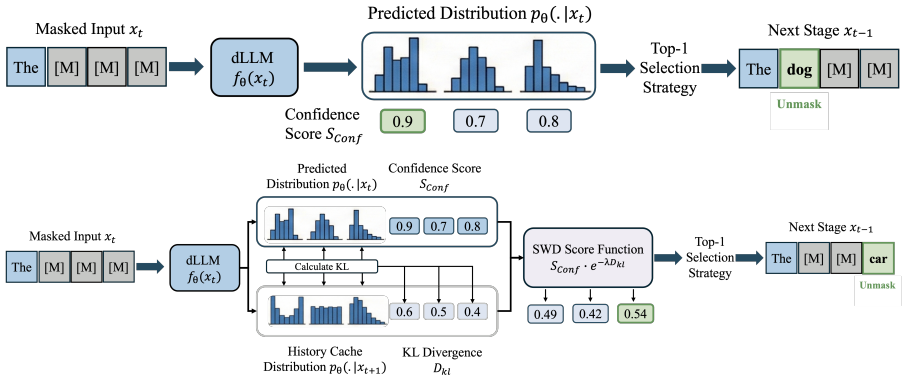


Figure 2: **Overview of decoding pipelines using Confidence Score and Top-1 Selection. (Top) Standard dLLM:** The model relies solely on the static confidence score at the current step. In this case, it prematurely commits to the high-confidence token "dog" (0.9) despite its instability. **(Bottom) Stability-Weighted Decoding (SWD):** By incorporating the historical distribution, SWD identifies the instability of "dog" (high KL divergence) and penalizes its score ($0.9 \rightarrow 0.49$). Consequently, the selection shifts to the more stable candidate "car" ($0.8 \rightarrow 0.54$), effectively correcting the error.

3 METHODOLOGY

We introduce the notation (Section 3.1), establish the information-theoretic result (Section 3.2), and present the stability-weighted scoring rule inspired by this result (Section 3.3).

3.1 NOTATION

Let $\mathcal{V}$ denote a vocabulary set of size $K = |\mathcal{V}|$. We denote $\mathbf{x}_t = (x_t^1, \ldots, x_t^L)$ as a sequence of length L at time step t, where each token $x_t^i \in \mathcal{V} \cup \{[\texttt{MASK}]\}$. The diffusion process operates over discrete time steps $t \in \{0, \ldots, T\}$, where $t = 0$ represents the clean data $\mathbf{x}_0$ and $t = T$ represents the fully masked state $\mathbf{x}_T = [\texttt{MASK}]^L$. We let $M_t \subseteq \{1, \ldots, L\}$ denote the set of masked indices at step t. The denoising network f_θ maps a masked sequence $\mathbf{x}_t$ to predict the original clean token, leading to the categorical distribution:

$$p_\theta^i(\cdot \mid \mathbf{x}_t) = \mathrm{Cat}\left(\mathrm{Softmax}\left([f_\theta(\mathbf{x}_t)]^i\right)\right). \tag{1}$$

3.2 THEORETICAL FRAMEWORK: INSTABILITY AS DEPENDENCY

We first establish a relation between distributional changes and contextual dependence under idealized ground-truth revelation. Throughout the theorem, p denotes conditionals of the same true joint distribution, distinct from the model predictions p_θ^i in Eq. (1).

Theorem 3.1 (**Sensitivity-Dependency Bound**). *Fix nested masked-index sets $M_t \subseteq M_{t+1}$ and a target $i \in M_t$. Let $\mathbf{U}_{t+1} = (x_0^j)_{j \in M_{t+1} \setminus \{i\}}$ contain the other masked tokens. The update reveals their ground-truth values $\Delta\mathbf{I}_t = (x_0^j)_{j \in M_{t+1} \setminus M_t}$, so $\mathbf{x}_t$ retains $\mathbf{x}_{t+1}$ and includes these reveals. Under this true joint distribution,*

$$\begin{aligned} &\mathbb{E}\left[D_{\mathrm{KL}}\left(p(x_0^i \mid \mathbf{x}_t) \,\|\, p(x_0^i \mid \mathbf{x}_{t+1})\right)\right] \\ &\quad = I(x_0^i; \Delta\mathbf{I}_t \mid \mathbf{x}_{t+1}) \leq I(x_0^i; \mathbf{U}_{t+1} \mid \mathbf{x}_{t+1}). \end{aligned} \tag{2}$$

The expectation averages over both contexts induced by these fixed masks and the true data distribution.

Proof. See Appendix A for the information-gain identity and the bound from the chain rule of mutual information. □

Table 1: **Main Results on Parallel Decoding.** Performance comparison of Stability-Weighted Decoding (SWD) against standard baselines across three representative scoring metrics: Confidence, Margin, and Entropy, as well as history-aware baselines (KLASS and CreditDecoding). Base denotes the vanilla EB Sampler strategy, and +SWD denotes our method. Values in parentheses indicate the absolute accuracy improvement. SWD improves Accuracy (**Acc**) while maintaining high efficiency (lower **NFE**). The best Accuracy for each model across all configurations is marked in **bold**.

Model	Metric	HumanEval		MBPP		GSM8K		MATH500		Average	
		Acc	NFE	Acc	NFE	Acc	NFE	Acc	NFE	Acc	NFE
LLaDA	Confidence	28.0	102.9	23.4	110.3	31.5	224.1	19.6	226.6	25.6	166.0
	+ SWD	**36.0** (↑8.0)	77.2	**34.8** (↑11.4)	66.8	**67.3** (↑35.8)	163.2	**26.2** (↑6.6)	205.7	**41.1** (↑15.5)	128.2
	Margin	23.2	100.8	26.0	108.3	38.7	215.8	21.8	221.0	27.4	161.5
	+ SWD	29.9 (↑6.7)	66.4	34.0 (↑8.0)	65.9	64.4 (↑25.7)	166.9	25.8 (↑4.0)	198.4	38.5 (↑11.1)	124.4
	Entropy	28.7	108.9	18.4	114.1	26.3	226.6	19.6	230.7	23.3	170.1
	+ SWD	32.3 (↑3.6)	103.5	26.6 (↑8.2)	108.0	48.7 (↑22.4)	191.8	20.4 (↑0.8)	224.3	32.0 (↑8.8)	156.9
	KLASS	28.1	115.4	24.0	117.9	58.8	136.8	18.4	304.7	32.3	168.7
	Credit	29.3	66.1	24.2	52.5	59.4	81.4	18.2	181.8	32.8	95.5
Dream	Confidence	12.2	179.7	29.4	212.4	29.3	286.6	10.8	364.1	20.4	260.7
	+ SWD	24.4 (↑12.2)	67.4	26.2 (↓3.2)	86.2	61.6 (↑32.3)	119.3	35.0 (↑24.2)	151.6	36.8 (↑16.4)	106.1
	Margin	10.4	173.4	23.0	202.6	29.5	275.6	11.4	354.9	18.6	251.6
	+ SWD	14.6 (↑4.2)	63.9	17.6 (↓5.4)	69.9	61.0 (↑31.5)	111.5	33.8 (↑22.4)	135.4	31.8 (↑13.2)	95.2
	Entropy	21.3	188.4	33.2	174.7	28.4	292.3	8.4	382.4	22.8	259.5
	+ SWD	**40.9** (↑19.6)	84.9	**42.6** (↑9.4)	109.4	**65.0** (↑36.6)	132.1	**38.6** (↑30.2)	181.6	**46.8** (↑24.0)	127.0
	KLASS	22.6	165.7	42.8	156.8	31.9	211.1	13.0	330.8	27.6	216.1
	Credit	26.2	58.5	41.4	79.0	31.4	107.6	13.6	113.8	28.2	89.7

Expected distributional change measures the information acquired from the revealed context and lower-bounds dependence on the other tokens masked before the update.

Inspired by this result, we use changes in model predictions as an empirical signal of contextual sensitivity.

3.3 Stability-Weighted Scoring

Inspired by Theorem 3.1, we define the practical temporal sensitivity of a still-masked token using the current and previous model predictions:

$$\mathcal{D}_{\text{temp}}^{(i)} = D_{\text{KL}}\big(p_\theta^i(\cdot \mid \mathbf{x}_t) \,\|\, p_\theta^i(\cdot \mid \mathbf{x}_{t+1})\big)\,. \tag{3}$$

Let $\mathcal{S}_{\text{base}}(i) \geq 0$ be a utility score whose larger values indicate higher selection priority. The multiplicative formulation requires scores on a compatible nonnegative scale; a score such as raw negative entropy requires a corresponding transformation and consistent threshold interpretation.

We choose an exponential penalty to combine base utility with temporal sensitivity:

$$\mathcal{S}_{\text{SWD}}(i) = \underbrace{\mathcal{S}_{\text{base}}(i)}_{\text{Base Utility}} \cdot \underbrace{\exp\left(-\lambda \cdot \mathcal{D}_{\text{temp}}^{(i)}\right)}_{\text{Stability Modulator}}. \tag{4}$$

Here $\lambda \geq 0$ controls the penalty strength. The factor equals one at zero divergence and decreases as divergence grows; $\lambda = 0$ recovers the base score. Tokens with large temporal divergence receive a stronger exponential penalty, while those with small divergence retain scores close to their original values. This encourages the decoder to defer temporally unstable predictions, using temporal KL as an empirical signal of contextual sensitivity. SWD can be incorporated into existing token-selection policies without modifying or retraining the underlying denoiser.

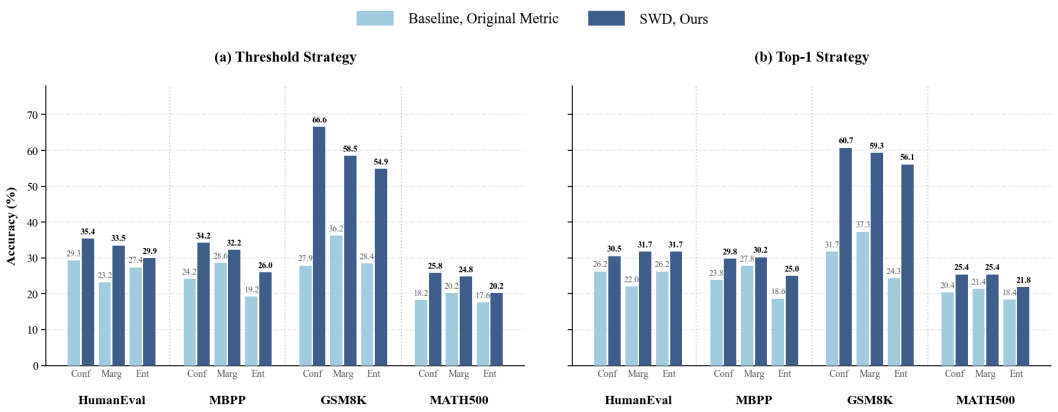


Figure 3: **Detailed selection strategies analysis.** Comparisons across selection policies: **(a) Threshold Strategy** and **(b) Top-1 Strategy**. SWD Blue outperforms the Baseline Gray.

Algorithm 1 Stability-Weighted Decoding (SWD)

Input: Masked sequence $\mathbf{x}_T$ ($T = L$ for Top-1), denoiser f_θ, penalty λ
Output: Generated sequence $\mathbf{x}_0$
Initialize: $\mathbf{p}_{T+1} \leftarrow \text{Uniform}(\mathcal{V})$, $M_T \leftarrow \{1, \dots, L\}$
for $t = T$ **down to** 1 **do**
  // 1. Predict & Score
  Compute probabilities: $\mathbf{p}_t \leftarrow \text{softmax}(f_\theta(\mathbf{x}_t))$
  Compute base scores: $\mathcal{S} \leftarrow \Phi(\mathbf{p}_t)$ *(e.g., Confidence)*
  // 2. Stability Modulation (SWD)
  Compute instability: $\mathcal{D}_{\text{KL}} \leftarrow \sum_{v \in \mathcal{V}} \mathbf{p}_t(v) \log\big(\mathbf{p}_t(v)/\mathbf{p}_{t+1}(v)\big)$
  Update scores: $\mathcal{S} \leftarrow \mathcal{S} \cdot \exp(-\lambda \cdot \mathcal{D}_{\text{KL}})$
  // 3. Select & Update (Example: Top-1)
  Identify best candidate: $i^* \leftarrow \arg\max_{i \in M_t} \mathcal{S}^{(i)}$
  Unmask: $x_{t-1}^{(i^*)} \leftarrow \arg\max \mathbf{p}_t^{(i^*)}$
  Update Mask: $M_{t-1} \leftarrow M_t \setminus \{i^*\}$
  Cache history: $\mathbf{p}_{t+1} \leftarrow \mathbf{p}_t$
end for
return $\mathbf{x}_0$

4 Experiments

We empirically validate Stability-Weighted Decoding (SWD) to demonstrate its effectiveness and universality. Following our experimental setup (Section 4.1), we first show SWD's seamless integration with arbitrary scoring metrics and state-of-the-art dLLMs (Section 4.2). We then evaluate its robustness across diverse decoding configurations and aggressive acceleration regimes (Sections 4.3 and 4.4). Finally, we analyze its computational efficiency and sensitivity to the stability penalty λ (Sections 4.5 and 4.6).

4.1 Experimental Setup

We evaluate SWD on LLaDA-8B-Instruct (Nie et al., 2025) and Dream-v0-Instruct-7B (Ye et al., 2025) across code (HumanEval (Chen et al., 2021), MBPP (Austin et al., 2021b)) and reasoning (GSM8K (Cobbe et al., 2021), MATH500 (Lightman et al., 2023)) benchmarks. Following Li et al. (2025), unless otherwise stated, generation lengths are 256 for HumanEval and MBPP, and 512 for GSM8K and MATH500. To demonstrate universality, we compare SWD against three static metrics (Confidence, Margin, Entropy) and two history-aware methods (KLASS (Kim et al., 2025b), CreditDecoding (Wang et al., 2025a)). Unless specified, we use EB-Sampler (Ben-Hamu et al.,

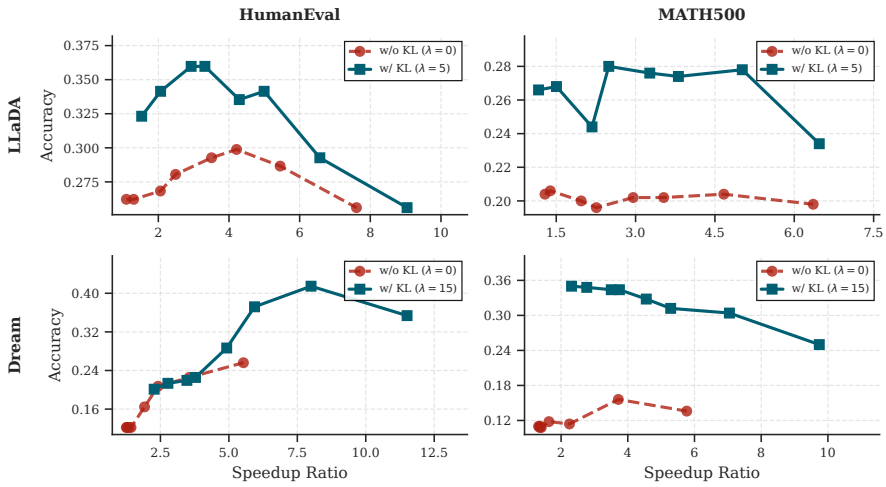


Figure 4: **Efficiency-Accuracy Trade-off.** Accuracy vs. Speedup Ratio by varying the uncertainty budget γ on HumanEval and MATH500. Solid Blue: SWD; Dashed Red: Confidence baseline. SWD achieves a superior Pareto front and effectively prevents performance collapse under high-acceleration regimes.

2025) with $\gamma = 0.1$; penalty settings are listed in the detailed result tables. All experiments run on 4 NVIDIA A100 GPUs.

We study a validation protocol that selects λ by accuracy on 250 MATH training examples for each model and decoding regime, then freezes it across test tasks. Appendix B gives the full sweep, selected values, and their relation to the metric-specific operating points.

4.2 Universal Enhancement across Scoring Metrics

As shown in Table 1, Stability-Weighted Decoding (+SWD) yields substantial Pareto improvements, simultaneously boosting accuracy and reducing computational overhead (NFE). For instance, under the Dream model with Entropy, SWD doubles the average accuracy (22.8% to 46.8%) while halving NFE (259.5 to 127.0). This demonstrates that our explicit stability penalty effectively prunes unproductive token explorations. Although a minor accuracy drop occurs in the Dream-MBPP setting, it is offset by drastically reduced computational cost (e.g., Margin NFE drops from 202.6 to 69.9), and recalibrating the budget ($\gamma = 0.005$) easily recovers and exceeds baseline performance.

The most striking gains are observed in complex mathematical reasoning tasks (GSM8K and MATH500), where SWD yields absolute accuracy surges of up to +36.6% and +30.2%, respectively. By penalizing step-wise instability, SWD acts as a crucial regularizer against error accumulation in lengthy logical deductions. Furthermore, SWD achieves higher accuracy than the history-aware baselines KLASS and CreditDecoding across the evaluated tasks, while exhibiting a better accuracy–efficiency trade-off.

Interestingly, while the optimal base metric varies by model architecture—LLaDA peaks with Confidence (41.1% Avg Acc) whereas Dream synergizes best with Entropy (46.8%)—SWD demonstrates robust adaptability. Regardless of the underlying confidence estimation strategy, SWD serves as an orthogonal enhancement that amplifies baseline performance.

We also calibrate the baseline γ on a fixed 50-example subset per model and benchmark to approach the fixed SWD NFE, using NFE rather than accuracy for selection. SWD improves accuracy in all eight reported pairs; individual NFEs remain approximate matches (Appendix C).

4.3 Generalization to Diverse Decoding Configurations

To assess the generality of SWD beyond the EB-Sampler selection method, we evaluate it under diverse token selection paradigms. We consider two selection policies: a Threshold Strategy (Wu

Table 2: **Semi-autoregressive decoding.** LLaDA-8B, Block=32. Accuracy (%). Avg parentheses show changes from Base.

Method	HEval	MBPP	GSM8K	MATH	Avg
Confidence	39.0	37.8	83.5	38.2	49.6
+ SWD	42.7	38.8	82.9	39.0	50.9 (↑1.3)
Margin	40.9	38.4	82.3	38.6	50.1
+ SWD	41.5	37.4	82.3	38.8	50.0 (↓0.1)
Entropy	42.7	38.2	83.5	37.0	50.4
+ SWD	43.9	38.0	82.5	39.6	51.0 (↑0.6)
History-Aware Baselines					
KLASS	39.6	36.8	79.2	39.6	48.8
Credit	38.4	38.2	79.1	40.4	49.0

Table 3: **Efficiency and resource profiling.** LLaDA-8B on HumanEval. Times are in ms; VRAM is in GB.

Length	256		512	
	Base	SWD	Base	SWD
Acc (%)	28.0	36.0	15.9	28.0
NFE	102.9	75.7	149.2	82.8
T_m	7,745.9	5,976.7	16,374.0	9,148.2
T_{kl}	—	193.4	—	419.6
T_{total}	7,745.9	6,170.1	16,374.0	9,567.8
VRAM avg.	17.7	18.4	20.0	21.1
VRAM peak	23.1	24.2	27.0	28.4

et al., 2025), and a Top-1 Strategy (Nie et al., 2025), using confidence, margin, and entropy-based scores. As shown in Figure 3, SWD outperforms the corresponding baselines across all benchmarks and scoring metrics(see Tables 9 and 10 for detailed results). The improvements are especially pronounced on reasoning-intensive tasks; for example, on GSM8K under the Threshold strategy, SWD increases accuracy from approximately 27.9% to over 66.6%, and nearly doubles performance under the Top-1 setting. These results indicate that temporal stability provides a robust and beneficial signal, independent of the specific selection policy or scoring metric.

We further evaluate SWD in a semi-autoregressive setting (LLaDA, Block=32). Table 2 shows SWD still improves average accuracy, outperforming history-aware baselines like KLASS and CreditDecoding. Notably, SWD achieves substantial gains on reasoning-heavy tasks, such as HumanEval (+3.7% with Confidence) and MATH500 (+2.6% with Entropy). Despite minor fluctuations on saturated benchmarks (e.g., GSM8K) or with Margin-based scoring, penalizing temporal instability remains a highly effective mechanism under strict block-wise settings.

We further evaluate SWD on LLaDA-8B, full-sequence IFEval (Zhou et al., 2023) prompt-level accuracy rises from 34.0% to 48.4%, and BBH (Suzgun et al., 2022) accuracy from 11.7% to 50.8%, with lower NFE. Appendix E reports the complete results.

4.4 Robustness across Acceleration Regimes

We rigorously evaluate SWD's speed-accuracy trade-off by sweeping the EB-Sampler's uncertainty budget $\gamma \in \{0.0005, 0.005, 0.05, 0.1, 0.5, 1, 2\}$ to simulate diverse acceleration ratios. As illustrated in Figure 4, SWD achieves a superior Pareto front over the standard Confidence baseline, with its curve lying strictly above the baseline across all configurations.

Notably, we observe distinct behaviors between models. While LLaDA initially exhibits higher accuracy, it suffers from rapid degradation under aggressive acceleration, with both the baseline and SWD dropping to $\sim$25% on HumanEval and $\sim$20% on MATH500. In contrast, Dream enhanced by SWD demonstrates exceptional robustness, overtaking LLaDA and maintaining $\sim$30% accuracy on HumanEval even at extreme speedups where LLaDA falters. These results confirm that explicitly penalizing temporal instability effectively mitigates the performance collapse typically associated with aggressive parallel decoding.

4.5 Efficiency and Resource Consumption Analysis

We evaluate the computational efficiency and memory overhead of SWD using LLaDA-8B on HumanEval, with generation lengths set to 256 and 512 ($B = L$). Table 3 reports model forward time T_m, KL computation time T_{kl}, and total latency T_{total}. While SWD introduces a local computational layer for KL divergence, its temporal cost T_{kl} is remarkably marginal, accounting for only 3.1% (193.4 ms) and 4.4% (419.6 ms) of the total latency under $L = 256$ and $L = 512$ configurations, respectively. Crucially, by preventing premature token commitments, SWD drastically reduces the

total NFE. The time saved by avoiding these expensive model forward passes (T_m) far outweighs the local overhead, leading to an overall reduction of 20.3% ($L = 256$) and 41.6% ($L = 512$) in end-to-end latency (T_{total}). This confirms that SWD scales efficiently with sequence length, trading minimal local arithmetic overhead for substantial global acceleration.

Furthermore, the VRAM consumption of SWD remains highly practical for 8B-parameter models. For $L = 256$, caching the previous step's probability distribution incurs a negligible increase of 4.8% in peak VRAM (from 23.1 GB to 24.2 GB). This footprint is highly competitive with alternative history-aware methods, such as CreditDecoding, which maintains a logit trace matrix $C \in \mathbb{R}^{L\times|V|}$, or KLASS, which similarly tracks historical divergence. Given the massive reduction in both NFE and total wall-clock time, this minor memory trade-off is fully justified for practical deployment.

4.6 Ablation Studies

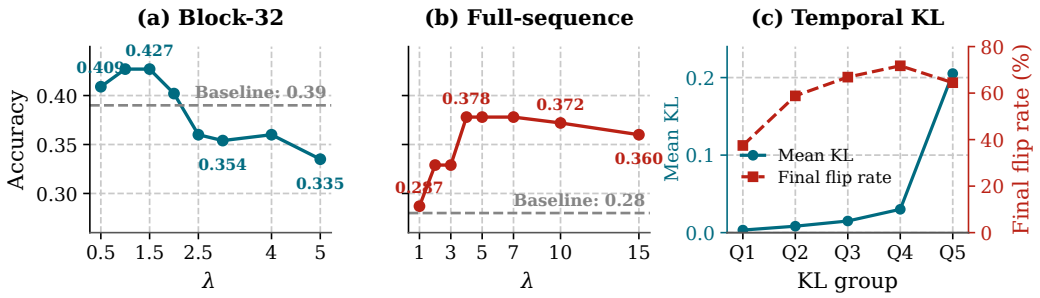


Figure 5: **Penalty sensitivity and temporal KL analysis.** (a,b) LLaDA-8B with EB-Sampler: dashed grey lines denote the Confidence baseline ($\lambda = 0$). (c) Vanilla full-sequence HumanEval decoding: Q1–Q5 order observations by increasing KL. Mean KL (left) and final-flip rate (right) use independent axis scales.

Sensitivity to λ. Figure 5(a,b) evaluates LLaDA-8B with EB-Sampler. Block-32 accuracy peaks at $\lambda = 1, 1.5$ and falls below the baseline at the tested values $\lambda \geq 2.5$. Full-sequence accuracy peaks at $\lambda = 4, 5, 7$ and exceeds the baseline throughout the tested range. The results exhibit an inverted U shaped trend, demonstrating that while SWD is robust across a wide range, λ requires appropriate tuning to reach the optimal balance. Appendix B separately gives the validation-based selection protocol and search grids.

Temporal KL and prediction changes. We analyze vanilla LLaDA-8B full-sequence EB decoding on HumanEval, without SWD. A final flip means that a masked position's current argmax differs from its final generated token, not a reference-answer error. Figure 5(c) shows a generally increasing, final-flip rate across KL quintiles (37.48% in the lowest versus 64.47% in the highest). Higher-KL predictions are thus more likely to change during decoding, supporting temporal KL as an empirical indicator of prediction instability. Within the confidence > 0.8 subset, the corresponding rates are 0.53% and 1.12%. Appendix F gives the full analysis.

KL direction. Using the same model predictions as Eq. (3), we compare

$$\text{Forward} = D_{\text{KL}}\big(p_\theta^i(\cdot \mid \mathbf{x}_t) \,\|\, p_\theta^i(\cdot \mid \mathbf{x}_{t+1})\big)\,,$$
$$\text{Reverse} = D_{\text{KL}}\big(p_\theta^i(\cdot \mid \mathbf{x}_{t+1}) \,\|\, p_\theta^i(\cdot \mid \mathbf{x}_t)\big)\,,$$
$$\text{Symmetric} = \tfrac{1}{2}\,(\text{Forward} + \text{Reverse})\,.$$

Figure 6 shows that Forward has the highest average accuracy in both regimes. No direction is uniformly best in all measures. We retain Forward KL as the main formulation, consistent with the orientation of the motivating information-gain identity. Appendix G gives the full results.

Block size. Table 4 evaluates block sizes $\{8, 16, 64, 128\}$ on LLaDA-8B with EB-Sampler. Using a fixed $\lambda = 1$, SWD improves average accuracy at every tested size. This supports the effectiveness of temporal stability weighting across different local unmasking constraints, with gains varying by block size.

Table 4: **Performance across varying Block Sizes.** LLaDA-8B with EB-Sampler. Base uses $\lambda = 0$; SWD uses $\lambda = 1$. Bold marks the best accuracy within each block.

Strategy	HEval	MBPP	GSM8K	MATH	Avg
Block 8					
Base	37.8	**35.8**	77.7	38.2	47.4
+ SWD	**40.2**	34.8	**79.1**	**39.0**	**48.3**
Block 16					
Base	37.2	**36.0**	77.4	**41.4**	48.0
+ SWD	**39.0**	35.4	**78.2**	39.6	**48.1**
Block 64					
Base	37.2	**35.2**	**79.4**	37.4	47.3
+ SWD	**41.5**	34.2	77.8	**40.6**	**48.5**
Block 128					
Base	**33.5**	**30.0**	76.1	37.2	44.2
+ SWD	**33.5**	**30.0**	**77.1**	**39.0**	**44.9**

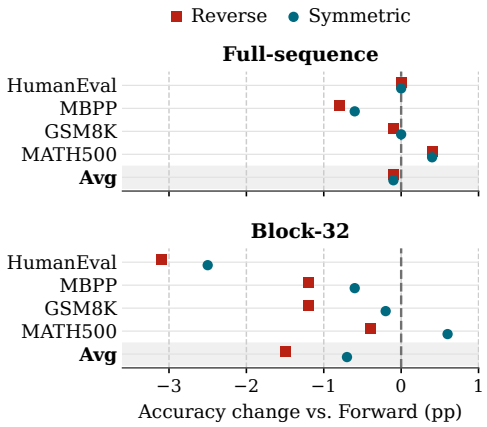


Figure 6: **KL-direction ablation.** LLaDA-8B, Confidence scoring, EB-Sampler. Accuracy changes are relative to Forward KL (zero line), in percentage points. Absolute accuracy/NFE appears in Appendix G.

Longer history with EMA-SWD. While standard SWD evaluates temporal stability from adjacent predictions (t versus $t+1$), we also explore longer historical trajectories through an Exponential Moving Average variant (EMA-SWD). The reported comparisons favor the single-step formulation. One possible explanation is that averaging carries early-stage fluctuations into later steps, diluting the current stability signal. Appendix H provides the formulation, momentum sweep, and empirical comparisons.

5 Conclusion

In this work, we introduced Stability-Weighted Decoding (SWD), a training-free and plug-and-play approach for reducing premature commitment in diffusion language model decoding. We established an information-theoretic connection between distributional change and contextual dependence under idealized ground-truth revelation, which motivates using temporal changes in model predictions as a practical sensitivity signal. Building on this insight, SWD penalizes temporally unstable token predictions and can be integrated into existing score-based decoding strategies without retraining the underlying model. Across code generation, mathematical reasoning, and instruction-following benchmarks, SWD improves the accuracy–efficiency trade-off across a range of decoding settings, with particularly strong gains under aggressive full-sequence parallel decoding. These results highlight temporal stability as a useful signal for token selection and suggest a simple direction for more reliable parallel decoding in diffusion language models.

Limitations

While SWD mitigates premature commitment, two limitations remain. First, it applies a constant stability penalty (λ) throughout reverse diffusion, although uncertainty is typically higher early in generation and lower near convergence. A fixed penalty may therefore be suboptimal. Future work could explore a timestep-dependent schedule (λ_t) that adjusts the penalty as generation proceeds.

Second, SWD is sensitive to λ. Although it improves performance across many tested settings, the best value varies by model and decoding regime. Applying SWD to new architectures or task domains may require validation-based tuning rather than relying on a universal default.

Reproducibility Statement

The experimental models, benchmarks, generation lengths, decoding configuration, and hardware are described in Section 4.1. Algorithm 1 and Figure 7 provide pseudocode and a minimal implementation example. Appendix A states the idealized assumptions and proves the information-gain identity and dependency bound; Appendix H and Appendix D report additional ablations and detailed experimental results.

AI Use Statement

Generative AI tools assisted with refining the paper narrative and and drafting code for analysis and figures. The authors reviewed the AI-assisted outputs, independently checked all reported experimental results, and take responsibility for the paper's content, claims, and artifacts.

References

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. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2503.09573.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021a. Curran Associates Inc. ISBN 9781713845393.

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, 2021b. URL https://arxiv.org/abs/2108.07732.

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

Tiwei Bie, Maosong Cao, Kun Chen, Lun Du, Mingliang Gong, Zhuochen Gong, Yanmei Gu, Jiaqi Hu, Zenan Huang, Zhenzhong Lan, Chengxi Li, Chongxuan Li, Jianguo Li, Zehuan Li, Huabin Liu, Lin Liu, Guoshan Lu, Xiaocheng Lu, Yuxin Ma, Jianfeng Tan, Lanning Wei, Ji-Rong Wen, Yipeng Xing, Xiaolu Zhang, Junbo Zhao, Da Zheng, Jun Zhou, Junlin Zhou, Zhanchao Zhou, Liwang Zhu, and Yihong Zhuang. Llada2.0: Scaling up diffusion language models to 100b, 2025. URL https://arxiv.org/abs/2512.15745.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Tom Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Kecheng Chen, Ziru Liu, Xijia Tao, Hui Liu, Xinyu Fu, Suiyun Zhang, Dandan Tu, Lingpeng Kong, Rui Liu, and Haoliang Li. Beyond confidence: Adaptive and coherent decoding for diffusion language models, 2025. URL https://arxiv.org/abs/2512.02044.

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.

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.

Guhao Feng, Yihan Geng, Jian Guan, Wei Wu, Liwei Wang, and Di He. Theoretical benefit and limitation of diffusion language model, 2025. URL https://arxiv.org/abs/2502.09622.

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, 2025. URL https://arxiv.org/abs/2506.20639.

Pengcheng Huang, Tianming Liu, Zhenghao Liu, Yukun Yan, Shuo Wang, Tong Xiao, Zulong Chen, and Maosong Sun. Empirical analysis of decoding biases in masked diffusion models, 2026. URL https://arxiv.org/abs/2508.13021.

Jaeyeon Kim, Kulin Shah, Vasilis Kontonis, Sham M. Kakade, and Sitan Chen. Train for the worst, plan for the best: Understanding token ordering in masked diffusions. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net, 2025a. URL https://openreview.net/forum?id=DjJmre5IkP.

Seo Hyun Kim, Sunwoo Hong, Hojung Jung, Youngrok Park, and Se-Young Yun. Klass: Kl-guided fast inference in masked diffusion models, 2025b. URL https://arxiv.org/abs/2511.05664.

Hyukhun Koh, Minha Jhang, Dohyung Kim, Sangmook Lee, and Kyomin Jung. Conditional [mask] discrete diffusion language model, 2025. URL https://arxiv.org/abs/2411.06438.

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*, 2025.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023. URL https://arxiv.org/abs/2305.20050.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

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.

Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. Your absorbing discrete diffusion secretly models the conditional distributions of clean data, 2024.

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. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Yuxuan Song, Zheng Zhang, Cheng Luo, Pengyang Gao, Fan Xia, Hao Luo, Zheng Li, Yuehang Yang, Hongli Yu, Xingwei Qu, Yuwei Fu, Jing Su, Ge Zhang, Wenhao Huang, Mingxuan Wang, Lin Yan, Xiaoying Jia, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Yonghui Wu, and Hao Zhou. Seed diffusion: A large-scale diffusion language model with high-speed inference, 2025. URL https://arxiv.org/abs/2508.02193.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them, 2022. URL https://arxiv.org/abs/2210.09261.

Kangyu Wang, Zhiyun Jiang, Haibo Feng, Weijia Zhao, Lin Liu, Jianguo Li, Zhenzhong Lan, and Weiyao Lin. Creditdecoding: Accelerating parallel decoding in diffusion large language models with trace credits, 2025a. URL https://arxiv.org/abs/2510.06133.

Wen Wang, Bozhen Fang, Chenchen Jing, Yongliang Shen, Yangyi Shen, Qiuyu Wang, Hao Ouyang, Hao Chen, and Chunhua Shen. Time is a feature: Exploiting temporal dynamics in diffusion language models, 2025b. URL https://arxiv.org/abs/2508.09138.

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.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models, 2025. URL https://arxiv.org/abs/2505.15809.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*, 2025.

Zebin You, Shen Nie, Xiaolu Zhang, Jun Hu, Jun Zhou, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. Llada-v: Large language diffusion models with visual instruction tuning, 2025. URL https://arxiv.org/abs/2505.16933.

Runpeng Yu, Xinyin Ma, and Xinchao Wang. Dimple: Discrete diffusion multimodal large language model with parallel decoding, 2025. URL https://arxiv.org/abs/2505.16990.

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. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=CTC7CmirNr.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

Fengqi Zhu, Rongzhen Wang, Shen Nie, Xiaolu Zhang, Chunwei Wu, Jun Hu, Jun Zhou, Jianfei Chen, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Llada 1.5: Variance-reduced preference optimization for large language diffusion models, 2025. URL https://arxiv.org/abs/2505.19223.

A Information-Theoretic Foundation of Stability-Weighted Decoding

We prove Theorem 3.1 using the existing decoding notation, under idealized ground-truth context revelation.

A.1 Preliminaries and Notation

Fix $M_t \subseteq M_{t+1}$ and $i \in M_t$. These index sets are held fixed independently of token values for this argument. Draw $\mathbf{x}_0$ from its true joint distribution and form $\mathbf{x}_{t+1}$ and $\mathbf{x}_t$ by masking the corresponding positions. All conditionals denoted by p below come from this same joint distribution, rather than the learned denoiser p_θ^i.

- x_0^i is the ground-truth token at the target position, which remains masked in both contexts.
- $\mathbf{x}_{t+1}$ is the previous context; $\mathbf{x}_t$ is the updated context, retaining all previously revealed values.
- $\mathbf{U}_{t+1} = (x_0^j)_{j \in M_{t+1} \setminus \{i\}}$ contains the other previously masked tokens. Similarly, $\mathbf{U}_t = (x_0^j)_{j \in M_t \setminus \{i\}}$ contains the other tokens still masked after the update. The target is excluded from both tuples.
- $\Delta\mathbf{I}_t = (x_0^j)_{j \in M_{t+1} \setminus M_t}$ contains the ground-truth values revealed by this update. Given the fixed masks, $\mathbf{x}_t$ and $(\mathbf{x}_{t+1}, \Delta\mathbf{I}_t)$ carry the same information.

Every expectation below averages over the contexts induced by the true joint distribution, including $\mathbf{x}_{t+1}$. Thus the conditional mutual information uses the same averaging convention. Conditional probabilities are evaluated on their support.

A.2 Equivalence of Instability and Information Gain

Lemma A.1 (Information gain under ground-truth revelation)**.** *Under the assumptions above, the expected KL divergence between true conditional distributions equals the conditional mutual information acquired from the revealed tokens:*

$$\begin{aligned} &\mathbb{E}\big[D_{\mathrm{KL}}\big(p(x_0^i \mid \mathbf{x}_t) \,\|\, p(x_0^i \mid \mathbf{x}_{t+1})\big)\big] \\ &\qquad = I(x_0^i; \Delta\mathbf{I}_t \mid \mathbf{x}_{t+1}) = I(x_0^i; \mathbf{x}_t \mid \mathbf{x}_{t+1}). \end{aligned} \tag{5}$$

Proof. Because the updated context retains the previous one, $p(x_0^i \mid \mathbf{x}_t) = p(x_0^i \mid \mathbf{x}_{t+1}, \Delta\mathbf{I}_t)$. Expanding the expectation gives

$$\begin{aligned} &\mathbb{E}\big[D_{\mathrm{KL}}\big(p(x_0^i \mid \mathbf{x}_t) \,\|\, p(x_0^i \mid \mathbf{x}_{t+1})\big)\big] \\ &= \sum_{\mathbf{x}_{t+1}, \Delta\mathbf{I}_t} p(\mathbf{x}_{t+1}, \Delta\mathbf{I}_t) \sum_{v \in \mathcal{V}} p(v \mid \mathbf{x}_{t+1}, \Delta\mathbf{I}_t) \log \frac{p(v \mid \mathbf{x}_{t+1}, \Delta\mathbf{I}_t)}{p(v \mid \mathbf{x}_{t+1})} \\ &= \sum_{\mathbf{x}_{t+1}, \Delta\mathbf{I}_t, v} p(\mathbf{x}_{t+1}, \Delta\mathbf{I}_t, v) \log \frac{p(v, \Delta\mathbf{I}_t \mid \mathbf{x}_{t+1})}{p(v \mid \mathbf{x}_{t+1}) p(\Delta\mathbf{I}_t \mid \mathbf{x}_{t+1})} \\ &= I(x_0^i; \Delta\mathbf{I}_t \mid \mathbf{x}_{t+1}). \end{aligned} \tag{6}$$

The second equality uses Bayes' rule. Given $\mathbf{x}_{t+1}$, the revealed tuple and updated context determine one another, yielding the final equality in the lemma. □

A.3 The Sensitivity-Dependency Bound

Theorem A.2 (Restatement of Theorem 3.1)**.** *Under the same assumptions, the expected KL divergence is a lower bound on dependence on the other previously masked tokens:*

$$I(x_0^i; \mathbf{U}_{t+1} \mid \mathbf{x}_{t+1}) \geq \mathbb{E}\big[D_{\mathrm{KL}}\big(p(x_0^i \mid \mathbf{x}_t) \,\|\, p(x_0^i \mid \mathbf{x}_{t+1})\big)\big]\,. \tag{7}$$

Proof. For the fixed index sets, $\mathbf{U}_{t+1}$ is partitioned into $\Delta\mathbf{I}_t$ and $\mathbf{U}_t$. The chain rule gives

$$\begin{aligned} I(x_0^i; \mathbf{U}_{t+1} \mid \mathbf{x}_{t+1}) &= I(x_0^i; \Delta\mathbf{I}_t, \mathbf{U}_t \mid \mathbf{x}_{t+1}) \\ &= I(x_0^i; \Delta\mathbf{I}_t \mid \mathbf{x}_{t+1}) + I(x_0^i; \mathbf{U}_t \mid \mathbf{x}_{t+1}, \Delta\mathbf{I}_t) \\ &= \mathbb{E}\big[D_{\mathrm{KL}}\big(p(x_0^i \mid \mathbf{x}_t) \,\|\, p(x_0^i \mid \mathbf{x}_{t+1})\big)\big] + I(x_0^i; \mathbf{U}_t \mid \mathbf{x}_t). \end{aligned} \tag{8}$$

The last equality follows from Lemma A.1 and the equivalence of the updated context and the previous context plus reveals. Nonnegativity of the residual conditional mutual information proves the bound. Equality can hold when this residual term is zero; the lower bound need not be strict. □

B HYPERPARAMETER SELECTION

We study a validation protocol using 250 examples sampled from the MATH training set. For each model and decoding regime, we select λ by validation accuracy and freeze the selected value across test tasks. Table 5 reports the full search grids and validation accuracies. Table 6 summarizes the selected values. This protocol does not imply that SWD has a universal default penalty.

Table 5: **Penalty selection on 250 MATH training examples.** Validation accuracy (%); bold marks the maximum for each model and decoding regime.

Full-sequence decoding						
λ	1	3	5	10	15	30
LLaDA-8B	22.8	23.2	**23.6**	22.4	22.0	21.6
Dream-7B	17.2	20.8	27.2	28.0	26.4	**29.2**
Block-32 decoding						
λ	0	0.5	1	1.5	2	3
LLaDA-8B	27.2	29.6	**31.2**	29.2	27.6	30.4
Dream-7B	30.8	**32.8**	30.8	31.2	29.2	25.2

Table 6: **Validation-selected penalties.** Maxima of the grids in Table 5.

Model	Regime	Selected λ	Validation accuracy (%)
LLaDA-8B	Full	5	23.6
LLaDA-8B	Block-32	1	31.2
Dream-7B	Full	30	29.2
Dream-7B	Block-32	0.5	32.8

The approximately NFE-matched comparison calibrates a different parameter, the baseline budget γ, by NFE on a fixed 50-example subset. Its selected budgets appear in Table 7; the SWD penalty stays fixed during that comparison.

C Comparisons at Approximately Matched NFE

For each model and benchmark, we sweep the baseline EB-Sampler uncertainty budget γ on a fixed 50-example subset and choose the value whose NFE is closest to the fixed SWD configuration. This selection uses NFE alone, not accuracy. Table 7 lists the selected budgets and all eight comparisons. The SWD operating points correspond to the full-sequence Confidence rows of Table 1.

Table 7: **Approximately NFE-matched full-sequence comparison.** Each result is accuracy (%) / NFE. Base budgets are calibrated separately for each model and benchmark.

Model	Dataset	Base γ	Matched Base	SWD
LLaDA-8B	HumanEval	0.30	29.3 / 73.0	36.0 / 77.2
	MBPP	0.30	24.8 / 70.7	34.8 / 66.8
	GSM8K	0.30	31.5 / 161.7	67.3 / 163.2
	MATH500	0.15	19.6 / 206.1	26.2 / 205.7
	Average	—	26.3 / 127.9	41.1 / 128.2
Dream-7B	HumanEval	1.00	22.6 / 72.1	24.4 / 67.4
	MBPP	0.75	23.4 / 78.0	26.2 / 86.2
	GSM8K	1.00	29.0 / 110.3	61.6 / 119.3
	MATH500	0.75	12.8 / 159.2	35.0 / 151.6
	Average	—	22.0 / 104.9	36.8 / 106.1

SWD improves accuracy in every reported pair, including pairs where it uses fewer evaluations.

D Detailed Experimental Results

In this section, we provide the complete experimental data supporting the main claims of the paper. We present granular performance metrics (Accuracy and NFE) across all models (LLaDA-8B, Dream-7B), decoding settings (Full-Sequence, Block-32), and selection strategies (EB-Sampler, Static, Threshold).

D.1 LLaDA-8B Comprehensive Results

Table 8 presents the extended results for LLaDA-8B under the Full-Sequence setting. We include a wider range of penalty strengths (λ) and compare against the Static decoding baseline. Table 9 details the study on the Threshold-based selection strategy, demonstrating SWD's robustness across different settings. We changed the threshold τ to make sure NFE are comparable, as SWD changes the value of scores. Table 10 provides specific results for the Top-1 selection strategy. Table 11 provides the granular breakdown of results for LLaDA-8B under the Semi-Autoregressive (Block-32) setting. We compare the Static strategy (fixed block generation), the standard EB-Sampler (Base), and SWD-enhanced versions with varying λ. The data confirms that SWD ($\lambda = 1.0$ or 1.5) boosts performance over the adaptive baseline, particularly in reasoning tasks like HumanEval and MATH500, while maintaining comparable inference costs.

Table 8: **LLaDA-8B Full-Sequence Comprehensive Results.** Performance comparison of different scoring targets (Confidence, Margin, NegEntropy) and strategies (Static, EB-Sampler) with expanded penalty strengths (λ). **Acc** is Accuracy (%) and **NFE** is Average Number of Function Evaluations. Best accuracy within each target block is bolded.

Target	Strategy	HumanEval		MBPP		GSM8K		MATH500	
		Acc ↑	NFE ↓	Acc ↑	NFE ↓	Acc ↑	NFE ↓	Acc ↑	NFE ↓
Confidence	EB (Baseline)	28.0	102.9	23.4	110.3	31.5	224.1	19.6	226.6
	+ SWD ($\lambda = 1.5$)	32.9	91.2	27.6	100.5	45.3	195.4	21.2	213.4
	+ SWD ($\lambda = 5.0$)	**36.0**	77.2	**34.8**	66.8	**67.3**	163.2	**26.2**	205.7
Margin	EB (Baseline)	23.2	100.8	26.0	108.3	38.7	215.8	21.8	221.0
	+ SWD ($\lambda = 1.0$)	23.8	92.8	27.8	103.5	39.5	203.1	20.4	214.0
	+ SWD ($\lambda = 1.5$)	23.8	88.2	28.0	99.5	42.5	195.9	21.8	209.7
	+ SWD ($\lambda = 5.0$)	**29.9**	66.4	**34.0**	65.9	**64.4**	166.9	**25.8**	198.4
Entropy	EB (Baseline)	28.7	108.9	18.4	114.1	26.3	226.6	19.6	230.7
	+ SWD ($\lambda = 1.0$)	28.0	106.7	21.4	112.6	27.9	221.1	19.2	230.2
	+ SWD ($\lambda = 1.5$)	29.3	106.1	22.6	111.7	29.5	216.8	19.8	229.3
	+ SWD ($\lambda = 5.0$)	**32.3**	103.5	**26.6**	108.0	**48.7**	191.8	**20.4**	224.3

Table 9: **LLaDA on Threshold Strategy.** Performance with varying confidence thresholds (τ) and penalty strengths (λ). Bold indicates the best result (Highest Acc, Lowest NFE) within each metric group.

Metric	Params	HumanEval		MBPP		GSM8K		MATH500	
		Acc	NFE	Acc	NFE	Acc	NFE	Acc	NFE
Confidence	$\tau = 0.9, \lambda = 0$	29.27	66.16	24.20	52.45	27.90	150.40	18.20	181.79
	$\tau = 0.8, \lambda = 1$	31.10	61.96	26.60	42.17	38.06	131.25	18.40	172.48
	$\tau = 0.8, \lambda = 5$	**35.37**	56.12	**34.20**	34.44	**66.64**	101.55	**25.80**	149.25
Margin	$\tau = 0.9, \lambda = 0$	23.17	65.77	28.60	60.62	36.24	167.82	20.20	195.64
	$\tau = 0.8, \lambda = 1$	24.39	80.59	28.60	76.05	39.35	205.15	20.60	240.16
	$\tau = 0.8, \lambda = 5$	**33.54**	53.99	**32.20**	42.26	**58.45**	139.56	**24.80**	176.85
Neg_Entropy	$\tau = -0.4, \lambda = 0$	27.44	85.01	19.20	71.70	28.40	90.49	17.60	199.78
	$\tau = -0.6, \lambda = 1$	**30.49**	79.84	22.00	58.12	**57.99**	88.88	17.60	194.32
	$\tau = -0.7, \lambda = 5$	29.88	77.39	**26.00**	59.83	54.89	81.81	**20.20**	183.36

Table 10: **LLaDA Top-1 Strategy Analysis.** Performance comparison using the **Top-1 selection strategy** (unmasking only the single highest-scoring token per step). We compare the baseline ($\lambda = 0$) against SWD ($\lambda = 5$). Note that Top-1 decoding results in fixed NFEs equal to the generation length: 256 for HumanEval and MBPP, and 512 for GSM8K and MATH500.

Metric	Method	HumanEval		MBPP		GSM8K		MATH500	
		Acc	NFE	Acc	NFE	Acc	NFE	Acc	NFE
Confidence	Top-1 (Base)	26.2	256.0	23.8	256.0	31.7	512.0	20.4	512.0
	Top-1 + SWD ($\lambda = 5$)	**30.5**	256.0	**29.8**	256.0	**60.7**	512.0	**25.4**	512.0
Margin	Top-1 (Base)	22.0	256.0	27.8	256.0	37.3	512.0	21.4	512.0
	Top-1 + SWD ($\lambda = 5$)	**31.7**	256.0	**30.2**	256.0	**59.3**	512.0	**25.4**	512.0
Entropy	Top-1 (Base)	26.2	256.0	18.6	256.0	24.3	512.0	18.4	512.0
	Top-1 + SWD ($\lambda = 5$)	**31.7**	256.0	**25.0**	256.0	**56.1**	512.0	**21.8**	512.0

Table 11: **LLaDA-8B Block-32 Detailed Results.** Comprehensive performance comparison under the semi-autoregressive setting (Block Size = 32). We compare Static decoding, the standard EB-Sampler baseline, and SWD with various penalty strengths (λ). Bold indicates the best accuracy within each metric group.

Metric	Method	HumanEval		MBPP		GSM8K		MATH500	
		Acc ↑	NFE ↓	Acc ↑	NFE ↓	Acc ↑	NFE ↓	Acc ↑	NFE ↓
Confidence	Static (Block)	39.6	256.0	38.6	256.0	82.9	512.0	**40.4**	512.0
	Base (EB)	39.0	76.6	37.8	55.9	**83.5**	116.0	38.2	173.2
	+ SWD ($\lambda = 1.0$)	**42.7**	76.3	**38.8**	55.1	82.9	110.6	39.0	163.3
	+ SWD ($\lambda = 1.5$)	**42.7**	73.0	38.6	54.3	82.4	109.6	38.2	162.4
Margin	Static (Block)	40.9	256.0	37.8	256.0	82.2	512.0	38.6	512.0
	Base (EB)	40.9	74.7	**38.4**	54.2	82.3	111.2	38.6	164.4
	+ SWD ($\lambda = 1.0$)	**41.5**	74.6	37.4	52.3	82.3	108.3	**38.8**	159.4
	+ SWD ($\lambda = 1.5$)	40.2	72.3	38.2	51.6	**83.2**	107.2	36.8	157.6
NegEntropy	Static (Block)	41.5	256.0	37.2	256.0	83.3	512.0	**40.4**	512.0
	Base (EB)	42.7	80.4	38.2	58.5	**83.5**	123.5	37.0	180.6
	+ SWD ($\lambda = 0.5$)	**43.9**	83.5	38.0	57.9	82.5	121.7	39.6	179.7
	+ SWD ($\lambda = 1.0$)	39.6	86.8	**38.4**	59.5	82.9	119.8	37.0	176.9
	+ SWD ($\lambda = 1.5$)	38.4	87.5	35.8	60.9	82.7	118.7	39.4	174.8

D.2 DREAM-7B COMPREHENSIVE RESULTS

Table 12 details the results for Dream-7B under the Full-Sequence setting. We observe significant improvements with higher λ values (e.g., $\lambda = 30$), particularly for the Entropy metric. Table 13 presents the results for Dream-7B under the Block=32 setting.

Table 12: **Dream-7B Full-Sequence Comprehensive Results.** Comparison of baselines and SWD with various penalty strengths. Note the effectiveness of stronger penalties (e.g., $\lambda = 15, 30$) for this model. Bold indicates the best result in each group.

Metric	Strategy	HumanEval		MBPP		GSM8K		MATH500	
		Acc	NFE	Acc	NFE	Acc	NFE	Acc	NFE
Confidence	EB ($\lambda = 0$)	12.20	179.65	29.40	212.37	29.34	286.60	10.80	364.08
	+ SWD ($\lambda = 1.0$)	14.02	160.84	39.40	154.71	29.04	249.50	12.20	305.93
	+ SWD ($\lambda = 1.5$)	14.63	154.23	41.40	152.61	31.24	224.87	13.80	277.19
	+ SWD ($\lambda = 5.0$)	17.07	73.56	**44.40**	87.57	52.69	123.84	25.80	115.06
	+ SWD ($\lambda = 15.0$)	22.56	67.73	26.60	86.81	59.29	116.94	34.40	136.58
	+ SWD ($\lambda = 30.0$)	**24.39**	67.41	26.20	86.24	**61.64**	119.26	**35.00**	151.64
Margin	EB ($\lambda = 0$)	10.37	173.40	**23.00**	202.60	29.49	275.62	11.40	354.94
	+ SWD ($\lambda = 15$)	13.41	60.87	14.00	74.24	58.07	105.94	**34.40**	109.84
	+ SWD ($\lambda = 30$)	**14.63**	63.93	17.60	69.93	**61.03**	111.53	33.80	135.41
Neg_Entropy	EB ($\lambda = 0$)	21.34	188.44	33.20	174.73	28.36	292.32	8.40	382.38
	+ SWD ($\lambda = 15$)	28.05	154.53	**43.20**	204.61	56.03	155.55	29.80	214.48
	+ SWD ($\lambda = 30$)	**40.85**	84.93	42.60	109.39	**64.97**	132.09	**38.60**	181.55

Table 13: **Dream-7B Block-32 Results.** Performance comparison with a local block constraint ($K = 32$). Bold indicates the best result in each group.

Metric	Strategy	HumanEval		MBPP		GSM8K		MATH500	
		Acc	NFE	Acc	NFE	Acc	NFE	Acc	NFE
Confidence	EB (baseline)	**50.00**	55.86	46.20	53.16	**76.12**	138.58	**38.00**	170.98
	+ SWD ($\lambda = 0.5$)	48.78	53.59	**47.00**	51.98	72.18	106.28	35.40	167.58
	+ SWD ($\lambda = 1.0$)	44.51	54.10	45.00	49.36	75.89	132.33	35.40	164.39
	+ SWD ($\lambda = 1.5$)	42.68	52.49	44.60	48.34	75.89	130.41	33.80	166.15
	+ SWD ($\lambda = 5.0$)	36.59	51.27	41.60	42.00	73.62	131.55	30.40	164.28
Margin	EB ($\lambda = 0$)	**49.39**	57.49	42.20	50.30	**75.21**	135.67	37.00	169.82
	+ SWD ($\lambda = 0.5$)	48.78	56.29	**45.00**	51.59	72.71	105.02	**37.80**	166.15
	+ SWD ($\lambda = 1.0$)	48.17	55.30	43.80	49.94	71.27	103.91	36.00	162.65
Neg_Entropy	EB ($\lambda = 0$)	50.61	57.22	46.40	53.36	73.62	112.11	40.00	176.26
	+ SWD ($\lambda = 0.5$)	50.00	56.75	**47.20**	**52.87**	73.39	110.80	**41.20**	172.89
	+ SWD ($\lambda = 1.0$)	**52.44**	55.98	**47.20**	54.75	**74.15**	109.66	40.80	170.47

E Instruction Following and Broader Reasoning

We additionally evaluate LLaDA-8B on IFEval (Zhou et al., 2023) and BBH (Suzgun et al., 2022), extending the task coverage to instruction following and broader reasoning. We use the default evaluation settings for both benchmarks. IFEval reports prompt-level and instruction-level accuracy separately; BBH reports accuracy.

Table 14: Additional LLaDA-8B results. IFEval accuracy is prompt-level / instruction-level (%); BBH uses a single accuracy (%).

Dataset	Setting	Method	Accuracy (%)	NFE
IFEval	Full	Base	34.0 / 42.8	290.6
	Full	SWD	48.4 / 59.7	224.9
	Block-32	Base	56.4 / 66.3	285.2
	Block-32	SWD	56.8 / 66.6	272.1
BBH	Full	Base	11.7	219.7
	Full	SWD	50.8	186.7
	Block-32	Base	57.2	137.1
	Block-32	SWD	57.4	132.3

The full-sequence gains are substantial: IFEval prompt-level accuracy rises by 14.4 percentage points, and BBH accuracy rises by 39.1 points, both with lower NFE. The Block-32 gains are much smaller: 0.4 points on IFEval prompt-level accuracy and 0.2 points on BBH, again with lower NFE. These results extend the empirical scope while reinforcing the distinction between aggressive full-sequence decoding and stronger block-wise baselines.

F TEMPORAL KL AND PREDICTION CHANGES

We analyze vanilla LLaDA-8B full-sequence EB-Sampler trajectories on HumanEval without applying SWD. Each observation corresponds to a masked position at a decoding step. We record temporal KL between consecutive predictions, current confidence, the current argmax, and the final generated token. A final flip is disagreement between the current argmax and that final token. This measures change relative to the decoder's own output, not error relative to a reference answer.

Table 15 reports the available aggregate measurements. Figure 5 plots the first five groups, which partition all observations by temporal KL; the two series use independent vertical scales, so their heights and slopes should not be compared directly. For the subset with confidence > 0.8, the reported groups are the lowest and highest KL quintiles within that subset; intermediate quintiles are not available here.

Table 15: **Temporal KL and final-flip aggregates.** Vanilla LLaDA-8B on HumanEval, full-sequence EB decoding. A dash denotes an unreported mean confidence.

Observations	KL quintile	Mean KL	Mean confidence	Flip rate (%)
All masked observations	Lowest 20%	0.0033	—	37.48
	20%–40%	0.0083	—	58.81
	40%–60%	0.0150	—	66.87
	60%–80%	0.0301	—	71.74
	Highest 20%	0.2052	—	64.47
Confidence > 0.8	Lowest 20%	0.0010	0.946	0.53
	Highest 20%	0.2333	0.907	1.12

All four higher-KL groups have larger final-flip rates than the lowest quintile. Temporal KL is therefore an effective empirical indicator of prediction change. In the high-confidence subset, the highest-KL group has a larger flip rate, but also a lower mean confidence. Showing these candidates are unstable and this approximation is effective.

G Ablation of KL Direction

We compare three KL variants using LLaDA-8B with EB-Sampler. Forward KL uses $D_{\mathrm{KL}}(p_\theta^i(\cdot \mid \mathbf{x}_t) \| p_\theta^i(\cdot \mid \mathbf{x}_{t+1}))$, the current $\|$ previous direction of the main method. Reverse KL swaps the two arguments, and Symmetric KL takes the arithmetic mean of the forward and reverse divergences.

Table 16: KL-direction ablation on LLaDA-8B. Each entry reports accuracy (%) / NFE. Averages cover the four benchmarks.

Setting	Dataset	Forward	Reverse	Symmetric
Full	HumanEval	36.0 / 77.2	36.0 / 73.5	36.0 / 70.5
	MBPP	34.8 / 66.8	34.0 / 46.0	34.2 / 51.0
	GSM8K	67.3 / 163.2	67.2 / 135.1	67.3 / 141.8
	MATH500	26.2 / 205.7	26.6 / 199.0	26.6 / 187.3
	Average	41.1 / 128.2	41.0 / 113.4	41.0 / 112.7
Block-32	HumanEval	42.7 / 76.3	39.6 / 66.3	40.2 / 70.8
	MBPP	38.8 / 55.1	37.6 / 51.2	38.2 / 51.2
	GSM8K	82.9 / 110.6	81.7 / 102.4	82.7 / 105.4
	MATH500	39.0 / 163.3	38.6 / 154.6	39.6 / 154.1
	Average	50.9 / 101.3	49.4 / 93.6	50.2 / 95.4

The variants exhibit different accuracy–efficiency trade-offs. Forward KL has the highest average accuracy in both decoding regimes. In full-sequence decoding, Symmetric KL has a slightly lower average accuracy (41.0% versus 41.1%). We retain Forward KL as the main formulation, consistent with the orientation of the motivating information-gain identity.

H Extended Formulation of Multi-Step History (EMA-SWD)

A natural theoretical extension to our proposed Stability-Weighted Decoding (SWD) is the incorporation of long-range historical information. While the standard SWD evaluates instability based on the immediate temporal shift between $p_\theta(\cdot|x_t)$ and $p_\theta(\cdot|x_{t+1})$, one might intuitively assume that tracking the distribution trajectory across multiple steps could smooth out transient noise and provide a more robust unmasking signal.

To rigorously explore this hypothesis, we formulated an Exponential Moving Average (EMA) variant of our method, termed **EMA-SWD**. Instead of using a static single-step divergence, EMA-SWD maintains a dynamic historical instability weight $C_{kl}^{(t)}$ for each token, iteratively updated during the reverse diffusion process:

$$C_{kl}^{(t)} = (1-\beta) \cdot D_{KL}\left(p_\theta(\cdot|x_t) \parallel p_\theta(\cdot|x_{t+1})\right) + \beta \cdot C_{kl}^{(t+1)} \tag{9}$$

where $\beta \in [0, 1)$ acts as a momentum parameter controlling the influence of past history. The stability-weighted score is then modulated using this accumulated trajectory:

$$\mathcal{S}_{EMA-SWD}^{(i)} = \mathcal{S}_{base}^{(i)} \cdot \exp\left(-\lambda \cdot C_{kl}^{(t)}\right) \tag{10}$$

H.1 Empirical Analysis and Failure Mode

To determine the optimal momentum, we first conducted a grid search over $\beta \in \{0.0, 0.1, \ldots, 0.9\}$ on the HumanEval benchmark using LLaDA-8B (Confidence score, Block=Full, $\lambda = 5.0$). As shown in Table 17, incorporating longer history ($\beta > 0$) generally degrades performance, with $\beta = 0.0$ (our original single-step SWD) and $\beta = 0.1$ yielding the best results.

We further extended this evaluation using the best-performing EMA variant ($\beta = 0.1$) across all four datasets to verify if it provides consistent gains. As detailed in Table 18, the original single-step formulation ($\beta = 0.0$) consistently outperforms the EMA-SWD across both semi-autoregressive (Block 32) and full-sequence settings.

Table 17: **Momentum search.** LLaDA-8B on HumanEval; Full, $\lambda = 5$.

β	Acc (%)	β	Acc (%)
0.0	**35.98**	0.5	31.71
0.1	**35.98**	0.6	28.05
0.2	34.76	0.7	27.44
0.3	32.93	0.8	32.32
0.4	28.66	0.9	31.10

Table 18: **Original SWD versus EMA-SWD.** Accuracy (%) on LLaDA-8B across four benchmarks. Original SWD uses $\beta = 0$; EMA-SWD uses $\beta = 0.1$.

Block	Method	β	HEval	MBPP	GSM8K	MATH
32	EMA-SWD	0.1	42.07	34.60	79.00	38.20
	SWD	0.0	**42.70**	**38.80**	**82.90**	**39.00**
Full	EMA-SWD	0.1	35.98	33.20	60.42	26.00
	SWD	0.0	**36.00**	**34.80**	**67.30**	**26.20**

Discussion. We attribute this phenomenon to the inherent information dynamics of masked diffusion models. In the early stages of the reverse process ($t \approx T$), the model is highly uncertain, yielding essentially uniform or chaotic categorical distributions. The KL divergence between these early steps is dominated by random noise rather than meaningful structural convergence. By employing EMA, this chaotic information from the early phases is propagated into the critical mid-to-late stages of decoding. This accumulated noise dilutes the high-fidelity stability signal that emerges when the token is actually converging.

Similarly, baseline methods like CreditDecoding also attempt to aggregate multi-step historical logits, which often limits their performance for the same reason. Consequently, the instantaneous temporal shift (t vs $t-1$) provides the highest signal-to-noise ratio. This confirms that our current single-step SWD formulation successfully maintains a superior balance of simplicity, computational efficiency, and generation accuracy without the need for multi-step historical caching.

I CODE IMPLEMENTATION

Key code is illustrated in Figure 7. SWD only needs a slight change to the original code.

```python
def step(logits_t, k):
    # 1. Compute Probabilities
    probs = F.softmax(logits_t, dim=-1)

    # 2. Base Score: Confidence
    scores, _ = probs.max(dim=-1)

    # 3. Select Top-k
    val, idx = torch.topk(scores, k)

    return idx
```

```python
def swd_step(logits_t, logits_prev, k, lam):
    # 1. Compute Probabilities
    probs = F.softmax(logits_t, dim=-1)

    # 2. Base Score: Confidence
    scores, _ = probs.max(dim=-1)

    # 3. Apply Stability Modulator (New)
    p_prev = F.softmax(logits_prev, dim=-1)
    kl = (probs * (probs.log() - p_prev.log())).sum(-1)
    scores = scores * torch.exp(-lam * kl)

    # 4. Select Top-k
    val, idx = torch.topk(scores, k)
    return idx
```

Figure 7: **Implementation Comparison.** Python code for a single decoding step. **Left:** Standard approach relying solely on confidence. **Right:** Our SWD method. By adding the stability modulator (Lines 9-11), we suppress unstable tokens with minimal code changes.

J MODEL AND DATASET LICENSES

All pre-trained models (LLaDA, Dream) and datasets (HumanEval, MBPP, GSM8K, MATH500) used in this work are publicly available under open-source licenses (e.g., MIT, Apache 2.0) and are utilized strictly for academic research purposes. All pre-trained models and datasets utilized in this study were strictly used for academic research and methodological evaluation, which is completely consistent with their intended use cases.