Adaptive Guidance for Retrieval-Augmented Masked Diffusion Models

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) improves factual grounding by incorporating external knowledge into language model generation. However, when retrieved context is noisy, unreliable, or inconsistent with the model's parametric knowledge, it introduces retrieval-prior conflicts that can degrade generation quality. While this problem has been studied in autoregressive language models, it remains largely unexplored in diffusion-based language models, where the iterative denoising process introduces unique challenges for integrating retrieved context. In this work, we propose **Adaptive Retrieval-Augmented Masked Diffusion (ARAM)**, a training-free adaptive guidance framework for Masked Diffusion Models (MDMs) in RAG settings. ARAM dynamically calibrates the guidance scale during denoising according to the Signal-to-Noise Ratio (SNR) of the distributional shift induced by retrieved context. Intuitively, the model strengthens guidance when the retrieved context provides reliable corrective evidence and suppresses it when the contextual signal is noisy or non-supportive. Extensive experiments on multiple knowledge-intensive QA benchmarks show that ARAM improves overall QA performance over decoding-based RAG baselines.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities across various natural language processing tasks (Grattafiori et al., 2024; Hui et al., 2024). However, their reliance on static parametric knowledge often leads to hallucinations and outdated responses on knowledge-intensive queries. Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Zhao et al., 2026) addresses this limitation by grounding generation in external knowledge retrieved at inference time.

Recently, diffusion-based language generation has emerged as a promising alternative to autoregressive models. In particular, Masked Diffusion Models (MDMs) generate text through an iterative denoising process that progressively fills masked tokens while leveraging bidirectional attention (Nie et al., 2025; Ye et al., 2025). This paradigm enables parallel token updates and iterative refinement of intermediate predictions.

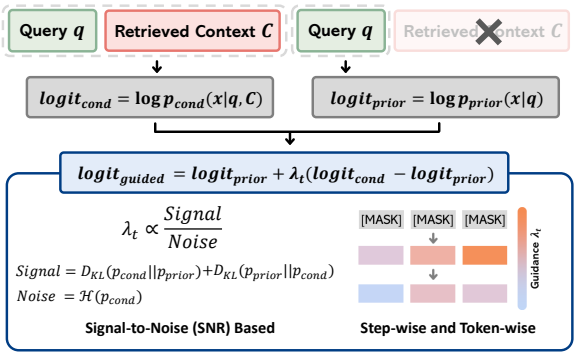


Figure 1: **Overview of ARAM.** ARAM mitigates knowledge conflict via step-wise and token-wise adaptive guidance driven by the Signal-to-Noise Ratio.

While recent research has focused on improving the generative capabilities of MDMs to narrow the gap with autoregressive models (Lou et al., 2023), applying RAG to MDMs remains underexplored. Preliminary studies indicate that MDM-based RAG demonstrates remarkable in-context learning capabilities that often surpass those of autoregressive models (Dai et al., 2026; Yu et al., 2026). Despite this potential, existing approaches assume that the retrieved documents are reliable. In realistic RAG settings, however, the retrieved context is often noisy, partially irrelevant, or contradictory to the model's internal knowledge.

A fundamental challenge is how to ground the iterative denoising process in retrieved knowledge, especially when this external information conflicts with the model's parametric prior. When such heterogeneous context is injected into generation, the model must dynamically decide how much to trust the retrieved context versus its parametric prior. We refer to this tension as *knowledge conflict* (Shi

et al., 2024; Wang et al., 2025b; Yuan et al., 2024). This raises a key question:

> How should a diffusion language model determine when to trust retrieved context and when to rely on its parametric knowledge?

To address this question, we propose **Adaptive Retrieval-Augmented Masked Diffusion (ARAM)**, an adaptive guidance framework that self-calibrates the guidance scale at each denoising step and token position. To inject conditional information and control the generation quality in MDMs, we adopt Classifier-Free Guidance (CFG) (Ho and Salimans, 2022) for RAG. However, applying a static guidance scale under retrieved contexts leads to a critical dilemma: (1) over-guidance on noisy context, which causes hallucinations, and (2) under-guidance on valid context, which fails to correct the flawed prior.

Our key insight is that the influence of retrieved context can be interpreted as a distributional shift from a retrieval-free prior to a retrieval-conditioned posterior. We analyze this interaction through an information-theoretic lens and derive a Signal-to-Noise Ratio (SNR) of the distributional shift to adapt the guidance strength. Consequently, when retrieved context consistently supports the conditional distribution, guidance can be strengthened; when the retrieved context is uncertain or non-supportive, guidance can be attenuated to prevent the over-amplification of noise. We illustrate the overview of our proposed method in Fig. 1.

We evaluate our adaptive guidance mechanism on knowledge-intensive QA benchmarks. Empirical results demonstrate that our method outperforms both static CFG and baseline MDMs, achieving improvements in exact matching (EM) and F1 scores for MDM-based RAG. Our main contributions are as follows:

- We propose ARAM, an adaptive CFG framework for MDMs in RAG, which dynamically modulates guidance scale based on the SNR.
- We analyze the interaction between retrieval context and parametric prior in diffusion-based generation through an information-theoretic lens and derive a guidance rule motivated by signal and uncertainty.
- Extensive experiments across multiple QA benchmarks validate that our method improves EM and F1 over other baselines and mitigates retrieval-prior conflicts.

2 Related Work

2.1 Knowledge Conflict in RAG

While Retrieval-Augmented Generation (RAG) improves factual grounding by incorporating external context, it can also introduce knowledge conflict (Longpre et al., 2021) between the model's internal knowledge and the retrieved external context, leading to degraded performance and hallucinations (Cheng et al., 2024). To mitigate these conflicts, current approaches rely on external context assessors (Choi et al., 2025), multi-agent debates (Wang et al., 2025b), or contrastive decoding (Shi et al., 2024; Wang et al., 2025a). A complementary line of work trains the model to self-reflect or self-reason over retrieved evidence (Asai et al., 2024; Xia et al., 2025; Luo et al., 2025), requiring additional supervision over reflection or reasoning trajectories. CAD (Shi et al., 2024) is most closely related to our work, as it contrasts output distributions with and without context to penalize reliance on parametric priors. However, applying a static contrastive penalty often overcorrects when the retrieved context already aligns with the model's parametric knowledge, inadvertently degrading performance on non-conflicting data (Yuan et al., 2024). To address this overcorrection, dynamic decoding strategies like AdaCAD (Wang et al., 2025a) and COIECD (Yuan et al., 2024) have emerged. While effective, these approaches are designed for autoregressive language models and cannot be trivially extended to the non-causal, iterative denoising dynamics of diffusion-based language models. Motivated by the need for dynamic adjustment (Wang et al., 2025a; Yuan et al., 2024), we introduce a training-free, discrete-diffusion-specific framework that modulates the guidance scale at each denoising step using a Signal-to-Noise Ratio (SNR).

2.2 Classifier-Free Guidance in MDMs

Classifier-Free Guidance (CFG) (Ho and Salimans, 2022) is widely used to strengthen conditional generation by combining conditional and prior score functions. Since CFG was originally formulated for continuous spaces, defining an analogous procedure for discrete diffusion is nontrivial, as the discrete state space is non-differentiable. Recent theoretical works (Schiff et al., 2024; Rojas et al., 2025) have formalized the validity of CFG at the

logit level in discrete diffusion models, but standard CFG applies a static guidance scale throughout the entire denoising process. Recognizing the dynamic uncertainty during generation, recent methods like A-CFG (Li et al., 2025) dynamically alter the unconditional input by re-masking low-confidence tokens. Distinct from modifying the input formulation, our approach directly calibrates the guidance itself, preventing over-guidance on noisy context and under-guidance on valid corrections.

2.3 RAG in Masked Diffusion Models

Recent studies demonstrate the potential of integrating RAG with MDMs to enhance knowledge utilization beyond autoregressive (AR) models. For example, recent analysis on Attention Floating (Dai et al., 2026) reveals that MDMs exhibit distinctive attention dynamics, where dominant attention tokens are dispersed across positions and denoising steps. This behavior allows MDMs to effectively track relevant context, resulting in stronger utilization of external knowledge than AR models. SPREAD (Yu et al., 2026) further shows that MDMs exhibit a strong reliance on contextual information and improve answer precision by enforcing query-level semantic coherence. Despite this potential, integrating RAG into MDMs remains unexplored. To the best of our knowledge, our work is the first to address knowledge conflict in MDM-based RAG through an intrinsic step-wise adaptive guidance mechanism.

3 Preliminaries

Masked Diffusion Language Models. Masked Diffusion Models (MDMs) generate text through an iterative denoising process over discrete token sequences. Let $\mathbf{x}_0 = (x_1, \ldots, x_L)$ denote the target token sequence of length L, where each token x_i belongs to a vocabulary $\mathcal{V}$. The forward process $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ gradually corrupts the sequence over discrete timesteps $t \in \{1, \ldots, T\}$ by replacing a subset of tokens with a special mask token `[MASK]`. The reverse process is parameterized by a neural network p_θ, which predicts token distributions $p_\theta(\mathbf{x}_0|\mathbf{x}_t)$ for the masked positions. The state transition to the previous step $\mathbf{x}_{t-1}$ is then sampled using the posterior distribution:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \sum_{\mathbf{x}_0} q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) p_\theta(\mathbf{x}_0|\mathbf{x}_t). \quad (1)$$

In MDMs, this unmasking process is typically executed in parallel across multiple masked positions based on token-level confidence. For notational simplicity, the following definitions are written for a single masked position, where x_t denotes the noised token at timestep t and x denotes the token at the target position.

Retrieval-Augmented Generation. In Retrieval-Augmented Generation (RAG), the model conditions generation on a user query q and a set of K retrieved documents $\mathcal{C} = \{d_1, \ldots, d_K\}$. The retrieved context $\mathcal{C}$ is concatenated with the query and provided as input to the language model. We denote this conditional token distribution as:

$$p_{\text{cond}}(x) := p_\theta(x|x_t, q, \mathcal{C}). \quad (2)$$

Conversely, when the model relies on its parametric prior without retrieval context, the distribution is:

$$p_{\text{prior}}(x) := p_\theta(x|x_t, q). \quad (3)$$

Classifier-Free Guidance in Discrete Diffusion. While Classifier-Free Guidance (CFG) was originally designed for continuous diffusion by extrapolating score differences, CFG in discrete state spaces operates directly on the prediction logits. Let ℓ_{cond} and ℓ_{prior} denote the logits corresponding to p_{cond} and p_{prior}, respectively. The guided logits are computed as follows, where $\lambda \geq 0$ is the guidance scale:

$$\ell_\lambda = \ell_{\text{prior}} + \lambda(\ell_{\text{cond}} - \ell_{\text{prior}}). \quad (4)$$

4 Method

In this section, we present the Adaptive Retrieval-Augmented Masked Diffusion (ARAM) framework. We first define the knowledge conflict and the limitations of static guidance (Sec. 4.1), and then interpret CFG as an exponential tilting family (Sec. 4.2). Finally, we derive an adaptive guidance rule based on an information-theoretic Signal-to-Noise Ratio (SNR) (Sec. 4.3 and Sec. 4.4).

4.1 Dilemma of Static Guidance

In standard discrete CFG (Eq. (4)), the guidance scale λ uniformly amplifies the logit difference across all tokens and diffusion steps. However, in realistic RAG settings, the retrieved context $\mathcal{C}$ introduces varying degrees of reliability, creating a knowledge conflict between the parametric prior and external evidence (Longpre et al., 2021). We can categorize these generation dynamics into three distinct scenarios:

1. **Reliable Context:** When $\mathcal{C}$ contains highly relevant and accurate supporting facts, p_{cond} confidently shifts toward the correct tokens. This produces a large divergence between p_{cond} and p_{prior} with low uncertainty. In this scenario, a larger λ is required to fully incorporate the retrieved facts and update missing or outdated parametric knowledge.

2. **Irrelevant Context:** When $\mathcal{C}$ is topically unrelated to the query, the model largely ignores the external context, resulting in minimal distributional divergence ($p_{\text{cond}} \approx p_{\text{prior}}$). The guidance scale λ should remain close to zero to prevent unnecessary perturbation.

3. **Conflicting Context:** When $\mathcal{C}$ contains misleading or contradictory information relative to the query or the model prior, p_{cond} can deviate substantially from p_{prior}. However, this deviation is typically accompanied by high uncertainty or internal inconsistency. If a large λ is applied, it overwrites the parametric prior p_{prior} and exacerbates hallucinations. Therefore, λ must be adaptively suppressed considering the uncertainty of p_{cond}.

To resolve these dynamics, the model requires an intrinsic, token-level mechanism to evaluate the trustworthiness of $\mathcal{C}$ and adaptively modulate λ.

4.2 Information-Theoretic View of CFG

We consider an MDM generating a discrete token sequence. Throughout this section and the next, we develop the analysis at a single masked position i and denoising step t; the same construction is applied independently to every masked position at every step in our final algorithm. To quantify the influence of the retrieved context $\mathcal{C}$ at diffusion step t, we define the *context score* as a log-likelihood ratio:

$$s(x_t^i; \mathcal{C}) := \log \frac{p_{\text{cond}}(x_t^i)}{p_{\text{prior}}(x_t^i)}. \tag{5}$$

For brevity, we drop the (t, i) subscript when no confusion arises and write $x = x_t^i$. Using this term with Eq. (4), the guided distribution $p_\lambda(x) \propto p_{\text{cond}}(x)^\lambda p_{\text{prior}}(x)^{1-\lambda}$ can then be written as:

$$p_\lambda(x) = \frac{p_{\text{prior}}(x) \exp(\lambda s(x; \mathcal{C}))}{Z_\lambda}, \tag{6}$$

where $Z_\lambda = \sum_{x \in \mathcal{V}} p_{\text{prior}}(x) \exp(\lambda s(x; \mathcal{C}))$.

To quantify the influence of retrieved context, we refer to the retrieval information gain as:

$$\text{IG}_t := D_{\text{KL}}(p_{\text{cond}} \| p_{\text{prior}}) = \mathbb{E}_{p_{\text{cond}}}[s(x; \mathcal{C})]. \tag{7}$$

This quantity measures how strongly the retrieved context alters the model's posterior belief. If the retrieved context $\mathcal{C}$ is treated as a random variable produced by the retriever, the Conditional Mutual Information (CMI) satisfies

$$I(x_0; \mathcal{C} \mid x_t) = \mathbb{E}_{\mathcal{C}}[D_{\text{KL}}(p_{\text{cond}} \| p_{\text{prior}})] = \mathbb{E}_{\mathcal{C}}[\text{IG}_t]. \tag{8}$$

Thus, IG_t serves as an information-theoretic signal measuring the usefulness of retrieved context. In the following, we first analyze the ideal choice of λ that maximizes a local approximation to this information gain (Sec. 4.3), and then introduce a practical surrogate that preserves the theoretical structure while remaining stable in MDM decoding (Sec. 4.4).

4.3 Lower Bound and the Ideal SNR

To understand the optimal choice of λ, we relate the information gain to the Donsker-Varadhan (DV) variational representation (Donsker and Varadhan, 1983). All proofs are provided in Appendix A.

Theorem 1 (DV Variational Lower Bound)**.** *For any λ, the retrieval information gain satisfies:*

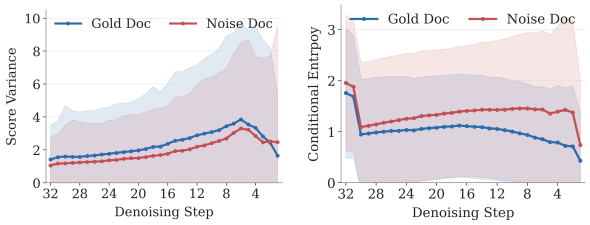
$$\text{IG}_t \geq \mathcal{L}(\lambda) := \lambda \mathbb{E}_{p_{cond}}[s(x; \mathcal{C})] - \log Z_\lambda. \tag{9}$$

Corollary 1.1 (Global Maximizer of the DV Bound)**.** *Let $\mathcal{L}(\lambda)$ be defined in Eq. (9). Then $\mathcal{L}(1) = D_{\text{KL}}(p_{cond} \| p_{prior})$. Hence, $\lambda = 1$ maximizes the variational lower bound $\mathcal{L}(\lambda)$.*

If the retrieved context $\mathcal{C}$ were a perfect oracle, Corollary 1.1 suggests that setting $\lambda = 1$ would be globally optimal. However, in practical RAG settings (Sec. 4.1), the retriever may return noisy or conflicting documents in $\mathcal{C}$. Therefore, setting $\lambda = 1$ assumes full trust in p_{cond} and leaves the generation highly vulnerable to hallucinations.

To mitigate this risk, we treat the prior distribution p_{prior} (where $\lambda = 0$) as a safe parametric anchor. Instead of jumping to the global maximum at $\lambda = 1$, we analyze the local geometry around our anchor ($\lambda = 0$) by taking a second-order Taylor expansion as:

$$\mathcal{L}(\lambda) \approx \lambda \underbrace{\left(\mathbb{E}_{p_{\text{cond}}}[s] - \mathbb{E}_{p_{\text{prior}}}[s]\right)}_{\text{Signal}} - \frac{\lambda^2}{2} \underbrace{\text{Var}_{p_{\text{prior}}}(s)}_{\text{Noise}}. \tag{10}$$



(a) $\mathrm{Var}_{p_{\text{prior}}}(s)$ (b) $\mathcal{H}(p_{\text{cond}})$

Figure 2: **Comparison of noise proxies during the denoising process. (a)** The variance of the log-likelihood ratio s attenuates the guidance scale even when correction is required. **(b)** In contrast, conditional entropy provides a stable measure of uncertainty, more clearly distinguishing between informative and noisy contexts.

using the derivatives $\mathcal{L}'(0) = \mathbb{E}_{p_{\text{cond}}}[s] - \mathbb{E}_{p_{\text{prior}}}[s]$ and $\mathcal{L}''(0) = -\mathrm{Var}_{p_{\text{prior}}}(s)$. Maximizing this concave approximation yields

$$\lambda^* = \frac{\text{Signal}}{\text{Noise}} = \frac{\mathbb{E}_{p_{\text{cond}}}[s] - \mathbb{E}_{p_{\text{prior}}}[s]}{\mathrm{Var}_{p_{\text{prior}}}(s)}. \tag{11}$$

This equation forms the Signal-to-Noise Ratio (SNR): an expected information gain (Signal) penalized by the dispersion introduced by the context (Noise). Detailed proofs are provided in Appendix A.

4.4 Practical Adaptive Guidance

Although Eq. (11) provides a theoretically optimal λ^*, directly using $\mathrm{Var}_{p_{\text{prior}}}(s)$ as the noise term is empirically unstable in masked diffusion decoding. As shown in Fig. 2a, the variance often becomes larger for *Gold Doc* (documents containing the correct answer) than for *Noise Doc* (documents not containing the answer), which undesirably suppresses λ_t^i even when the retrieved context is informative. The variance also exhibits substantial stepwise fluctuations, making the resulting guidance scale unstable across denoising steps. Experimental details are provided in Sec. 5.4.

To obtain a more stable estimate of the dispersion introduced by the context, we instead use the conditional entropy $\mathcal{H}(p_{\text{cond}})$ as the noise surrogate. When the retrieved context clearly points to the correct answer, p_{cond} becomes peaked on a few tokens and $\mathcal{H}(p_{\text{cond}})$ is small, giving a large SNR that amplifies guidance. When the retrieved context is unhelpful or contradicts itself (Sec. 4.1), p_{cond} spreads across many tokens and $\mathcal{H}(p_{\text{cond}})$ becomes large, which weakens guidance. Fig. 2b shows this directly: $\mathcal{H}(p_{\text{cond}})$ stays low for *Gold Doc* and rises for *Noise Doc*, unlike the variance in Fig. 2a. This is also in line with prior findings that

Algorithm 1 Adaptive Retrieval-Augmented Masked Diffusion (ARAM)

Require: Query q, Retrieved context $\mathcal{C}$, Masked Diffusion Model p_θ, Timesteps T, Maximum Guidance Scale $\lambda_{\max}$, Sensitivity β, A Small Positive ϵ

Initialize $\mathbf{x}_T \leftarrow$ ([MASK], ..., [MASK])
for $t = T$ **to** 1 **do**
  $p_{\text{prior}}, p_{\text{cond}} \leftarrow p_\theta(\cdot|\mathbf{x}_t, q), p_\theta(\cdot|\mathbf{x}_t, q, \mathcal{C})$
  for Each masked position i **do**
    Signal $\leftarrow D_{\mathrm{KL}}(p_{\text{cond}}\|p_{\text{prior}}) + D_{\mathrm{KL}}(p_{\text{prior}}\|p_{\text{cond}})$
    Noise $\leftarrow \mathcal{H}(p_{\text{cond}})$
    $\lambda_t^i \leftarrow \lambda_{\max} \cdot \tanh\left(\beta \cdot \frac{\text{Signal}}{\text{Noise}+\epsilon}\right)$
    Compute ℓ_{guided} using Eq. (13)
  end for
  Sample $\mathbf{x}_{t-1}$ using the guided logits ℓ_{guided} according to the unmasking policy
end for
return Generated sequence $\mathbf{x}_0$

language model predictions usually stay in a narrow entropy range, and that moving outside this range signals uncertain or conflicting context (Arora et al., 2023; Yuan et al., 2024). We use this substitution as a theory-motivated design, and compare it to $\mathrm{Var}_{p_{\text{prior}}}(s)$ and other candidates in Table 9.

In practice, for each denoising step t and each position i, we compute the following signal and noise proxies:

$$\begin{aligned}\text{Signal} &:= \mathbb{E}_{p_{\text{cond}}}[s] - \mathbb{E}_{p_{\text{prior}}}[s] \\ &= D_{\mathrm{KL}}(p_{\text{cond}}\|p_{\text{prior}}) + D_{\mathrm{KL}}(p_{\text{prior}}\|p_{\text{cond}}), \\ \text{Noise} &:= \mathcal{H}(p_{\text{cond}}) = -\sum_{x\in\mathcal{V}} p_{\text{cond}}(x)\log p_{\text{cond}}(x).\end{aligned}$$

To bound the guidance scale and prevent over-amplification when the noise term is small, we apply a tanh-based scaling:

$$\lambda_t^i = \lambda_{\max} \cdot \tanh\left(\beta \cdot \frac{\text{Signal}}{\text{Noise}+\epsilon}\right), \tag{12}$$

where $\lambda_{\max}$ defines the maximum guidance scale, β controls the sensitivity to the SNR, and $\epsilon > 0$ ensures numerical stability. An ablation against unbounded SNR scaling is reported in Appendix D (Table 10). Finally, we apply this token-wise adaptive λ_t^i to the logit-space CFG interpolation:

$$\ell_{\text{guided}}(x) = \ell_{\text{prior}}(x) + \lambda_t^i\big(\ell_{\text{cond}}(x) - \ell_{\text{prior}}(x)\big). \tag{13}$$

The complete procedure for our proposed ARAM method is detailed in Algorithm 1.

5 Experiments and Results

5.1 Experimental Protocol

Datasets and evaluation metrics. Following Dai et al. (2026), we evaluate our method on

Table 1: Performance comparison across QA benchmarks. Bold indicates the best score within each backbone.

Method	Open-Domain QA						Multi-hop QA		Slot Filling	
	NQ		TriviaQA		MARCO QA		HotpotQA		T-REx	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Autoregressive Models										
LLaMA	20.5	30.9	64.4	69.3	1.4	14.7	15.0	22.8	34.2	39.6
+ RAG (Lewis et al., 2020)	31.9	48.2	71.7	78.3	2.4	19.0	21.2	31.4	30.3	35.3
Qwen	14.9	25.2	48.1	52.7	1.1	11.8	21.2	29.0	32.6	36.5
+ RAG (Lewis et al., 2020)	31.3	46.6	70.8	76.6	2.3	19.0	24.2	33.7	30.1	34.7
Masked Diffusion Models - LLaDA										
LLaDA	9.1	15.9	27.7	32.0	1.1	9.7	15.7	21.5	23.9	26.4
+ RAG (Lewis et al., 2020)	17.5	31.9	44.8	55.1	1.4	**25.6**	6.5	15.6	17.8	23.8
+ CAD (Shi et al., 2024)	1.4	16.6	7.1	23.9	0.4	24.7	0.3	9.0	1.3	7.4
+ AdaCAD (Wang et al., 2025a)	9.3	18.2	27.1	33.6	1.2	14.3	15.0	21.4	21.7	24.7
+ COIECD (Yuan et al., 2024)	13.0	28.4	37.4	49.2	0.9	25.4	7.1	15.7	11.1	17.1
+ SPREAD (Yu et al., 2026)	19.3	34.4	51.7	61.0	1.7	25.2	12.6	21.2	18.0	24.0
+ A-CFG (Li et al., 2025)	19.1	33.9	51.8	61.2	1.4	24.5	12.6	21.5	18.2	**35.9**
+ **Ours**	**30.3**	**43.2**	**67.7**	**73.1**	**2.5**	20.0	**20.7**	**28.9**	**31.4**	34.8
Masked Diffusion Models - Dream										
Dream	14.9	24.9	36.7	43.5	1.9	11.0	16.9	26.2	30.4	35.1
+ RAG (Lewis et al., 2020)	35.9	50.5	72.7	78.4	2.8	17.5	25.0	35.2	29.7	35.0
+ CAD (Shi et al., 2024)	23.9	40.9	55.5	67.0	1.7	20.9	14.1	24.8	19.3	26.8
+ AdaCAD (Wang et al., 2025a)	37.4	50.7	72.5	77.4	2.9	**21.7**	24.3	34.5	31.4	36.7
+ COIECD (Yuan et al., 2024)	30.2	45.9	68.5	75.4	2.7	**21.7**	19.1	29.4	24.4	30.6
+ SPREAD (Yu et al., 2026)	35.9	50.1	73.7	78.5	2.9	17.8	25.0	34.7	**34.5**	**38.0**
+ A-CFG (Li et al., 2025)	34.0	49.2	70.6	77.8	2.6	17.4	24.2	34.6	28.1	34.1
+ **Ours**	**37.7**	**51.4**	**74.4**	**79.5**	**3.3**	18.0	**25.8**	**36.1**	30.0	35.5

five diverse knowledge-intensive QA datasets: NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), MARCO QA (Bajaj et al., 2016a), HotpotQA (Yang et al., 2015), and T-REx (Elsahar et al., 2018). These datasets cover a wide range of scenarios, including open-domain QA (NQ, TriviaQA, MARCO QA), multi-hop reasoning (HotpotQA), and slot filling (T-REx), making them suitable for evaluating knowledge utilization in language models. For all datasets, we adopt Exact Match (EM) and F1 score as our primary evaluation metrics. We use bge-large (Xiao et al., 2024) to retrieve documents from the MS MARCO 2.1 (Bajaj et al., 2016b) following Li et al., 2024.

Models and baselines. We compare our method across Autoregressive (AR) models and Masked Diffusion Models (MDMs). For AR models, we utilize LLaMA-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen2.5-7B-Instruct (Hui et al., 2024). For MDMs, we evaluate LLaDA-8B-Instruct (Nie et al., 2025) and Dream-v0-Instruct-7B (Ye et al., 2025). We benchmark ARAM against the following decoding strategies, which can be broadly categorized into AR-adapted contrastive methods and MDM-specific generation mechanisms:

- **Standard RAG:** The conventional generation method for MDMs utilizing the logit conditioned on the retrieved context $\ell_{\text{cond}}(x)$.
- **AR-Based Baselines:** We adapt AR-based methods to the diffusion generation process. These include CAD (Shi et al., 2024), COIECD (Yuan et al., 2024), and AdaCAD (Wang et al., 2025a).
- **MDM-Specific Baselines:** We compare against guidance mechanisms tailored for discrete diffusion. These include SPREAD (Yu et al., 2026), which guides token selection to maintain semantic coherence, and A-CFG (Li et al., 2025), which dynamically re-masks low-confidence tokens to construct a localized unconditional input for more effective guidance.

5.2 Benchmark Results

Table 1 presents the overall EM and F1 scores across the five QA benchmarks. The results demonstrate the strong capability of our proposed method in enhancing MDM-based RAG. Compared to the standard MDM-based RAG (which is mathematically equivalent to fixed CFG with $\lambda = 1$ since $\ell_{\text{guided}} = \ell_{\text{cond}}$), our adaptive guidance improves overall EM and F1 trends.

Applying AR-specific dynamic decoding methods (CAD, COIECD, and AdaCAD) to MDMs results in suboptimal performance. CAD exhibits

Table 2: Categorization rates of retrieval-prior interaction outcomes.

Method	Positive (↑)	Negative (↓)	Cons. Correct	Cons. Wrong
		LLaDA		
RAG	0.135	0.051	0.040	0.774
Ours	**0.234**	**0.022**	**0.069**	**0.675**
		Dream		
RAG	0.259	0.049	0.100	0.592
Ours	**0.266**	**0.038**	**0.111**	**0.585**

Table 3: Performance comparison across subsets stratified by retrieval context quality.

Method	Gold		Non-Answering		Irrelevant	
	LLaDA	Dream	LLaDA	Dream	LLaDA	Dream
RAG	24.2	49.7	0.77	1.54	0.00	0.00
Ours	**41.7**	**52.0**	**1.92**	**1.92**	**0.00**	**0.00**

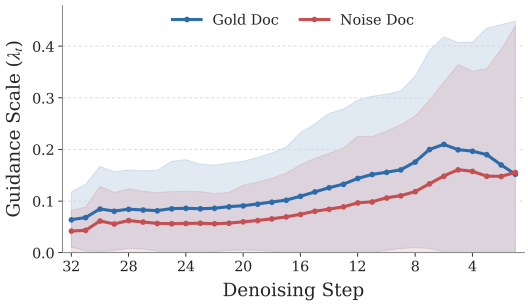
consistent degradation, suggesting that uniform contrastive guidance can over-correct in MDM-RAG settings. While ADACAD dynamically adjusts the guidance scale across generation steps, it applies a uniform scale across tokens. This limits performance in MDMs, where token-wise uncertainty varies across positions and denoising steps.

We also compare our method against SPREAD and A-CFG tailored for MDMs. While SPREAD and A-CFG improve upon static RAG by intervening in the token unmasking process, our method achieves higher overall performance across most benchmarks. Although SPREAD shows higher scores in the T-REx slot-filling task on Dream, our method outperforms the baselines in open-domain and multi-hop reasoning tasks. These results show that calibrating logit-space guidance via SNR provides a strong alternative to re-masking strategies.

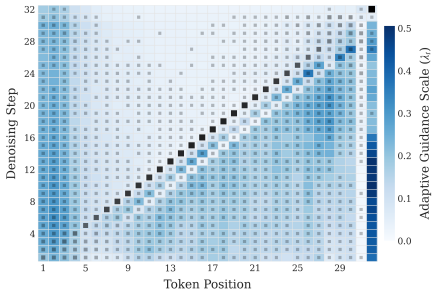
5.3 Analysis of Retrieval-Prior Conflict

To validate how ARAM resolves the knowledge conflict between the parametric prior and retrieved context, we analyze both model behavior and retrieval context quality. Detailed settings are provided in Appendix B.

First, we analyze the interaction between retrieval context and prior knowledge by comparing retrieval-based generation with the retrieval-free baseline. We categorize the outcomes into four cases: *Positive* (baseline incorrect → retrieval-based correct), *Negative* (baseline correct → retrieval-based incorrect), *Consistently Correct*, and *Consistently Wrong*. As shown in Table 2, ARAM increases the *Positive* rate while reducing the *Negative* rate for both models compared to static RAG. This indicates that ARAM strengthens



(a) Average guidance trajectory.



(b) Token-level guidance heatmap.

Figure 3: **Dynamics of adaptive guidance**. **(a)** Average guidance scale (λ_t^i) per step. ARAM maintains high guidance for gold documents but suppresses λ_t^i for noisy ones. **(b)** Token-wise guidance heatmap. Black squares mark unmasking positions, demonstrating a right-to-left generation tendency. ARAM assigns larger guidance to positions that align with the unmasking pattern.

guidance when retrieval provides useful corrective evidence while suppressing it when the retrieved context conflicts with the parametric prior.

Second, we stratify the evaluation samples by retrieval context quality. We define *Gold* as cases where the retrieved context contains the answer. For the remaining samples (equivalent to *Noise Doc* in Sec. 4.4), we further distinguish *Non-Answering* as cases where the retrieved context is related to the query but does not provide sufficient evidence for the correct answer, and *Irrelevant* as cases where the retrieved context is off-topic. Table 3 shows that ARAM improves over static RAG in both the *Gold* and *Non-Answering* settings. On *Gold* samples, ARAM more effectively incorporates valid external evidence, mitigating the under-guidance problem of static RAG. On *Non-Answering*, ARAM avoids over-amplifying unreliable contextual signals. On the *Irrelevant* subset, ARAM does not improve over static RAG, indicating that adaptive guidance cannot recover performance when the retrieved context is entirely uninformative.

Taken together, adaptive SNR-based guidance better resolves the dilemma of static guidance under knowledge conflict. ARAM reduces harmful

Table 4: Ablation study on $\lambda_{\max}$ and β. Bold denotes the best score and underlined represents second-best.

$\lambda_{\max}$	β	LLaDA EM	LLaDA F1	Dream EM	Dream F1
1.0	1.0	**24.9**	**36.1**	**33.9**	**43.6**
2.0		13.9	27.3	31.4	42.4
3.0		9.1	22.8	26.3	38.9
1.0	0.1	**30.4**	**39.3**	33.5	42.0
	0.3	28.4	38.7	**34.3**	43.4
	0.5	26.9	37.7	34.2	43.5
	0.7	26.1	37.0	34.0	43.5
	1.0	24.9	36.1	33.9	**43.6**
	2.0	22.3	34.1	33.6	43.3

retrieval-induced changes while preserving beneficial corrections and remains effective when the retrieved context is related but non-supportive.

5.4 Dynamics of Adaptive Guidance

To analyze the behavior of the proposed adaptive guidance, we monitor the step-wise and token-wise guidance scales during denoising (Fig. 3). We sampled 200 instances from the NQ dataset for both *Gold Doc* (documents containing the answer) and *Noise Doc* (documents not containing the answer) settings using the LLaDA model. At each timestep t, we recorded the token-wise average λ_t^i values.

As illustrated in Fig. 3a, the step-wise average guidance trajectory aligns with the SNR formulation. When provided with a *Gold Doc*, the model maintains a larger guidance scale. When confronted with a *Noise Doc*, the guidance scale is decreased, mitigating the risk of hallucinating based on irrelevant context.

We also visualize the token-level guidance heatmap to examine the spatial distribution of λ_t. In Fig. 3b, LLaDA exhibits a right-to-left unmasking pattern, and the highest guidance values cluster adjacent to recently unmasked tokens. Since newly generated tokens provide strong local context for predicting nearby masked tokens, ARAM allocates stronger guidance to positions that remain uncertain near recently resolved tokens.

5.5 Ablation Study

Effect of guidance hyperparameters. We investigate the impact of the maximum guidance scale $\lambda_{\max}$ and the sensitivity parameter β. As shown in Table 4, varying β while fixing $\lambda_{\max} = 1.0$ reveals that the two models exhibit different sensitivities to the context signal; Dream achieves balanced performance at $\beta = 0.5$, while LLaDA achieves optimal results at $\beta = 0.1$.

Table 5: Comparison of noise proxies. Average EM and F1 scores are reported.

Noise Proxy	LLaDA EM	LLaDA F1	Dream EM	Dream F1
$\mathrm{Var}_{p_{\text{prior}}}(s)$	20.3	32.5	33.5	43.2
$\mathcal{H}(p_{\text{cond}})$	**24.9**	**36.1**	**33.9**	**43.6**

Table 6: Average inference time per query (seconds).

Method	NFE / Step	LLaDA	Dream
No-RAG	1	0.93	0.94
RAG	1	1.55	1.56
CFG	2	2.02	3.06
ARAM (Ours)	2	2.07	3.29

Choice of noise proxy. To validate the use of conditional entropy $\mathcal{H}(p_{\text{cond}})$ as a noise proxy, we compare it against the score variance $\mathrm{Var}_{p_{\text{prior}}}(s)$ in Eq. (11), setting $\lambda_{\max} = 1.0$ and $\beta = 1.0$. As shown in Table 5, the entropy proxy yields higher generation quality for both models. This gap stems from the behavior of noise proxies during unmasking (Fig. 2a), empirically supporting the design choice in Sec. 4.4: the conditional entropy provides a more stable measure of uncertainty, resulting in robust guidance. Additional ablations are provided in Appendix D.

Computational Cost Analysis. We measure inference latency to analyze the overhead introduced by retrieval and adaptive guidance. Table 6 reports the average execution time over three runs. The gap between No-RAG and static RAG mainly comes from the longer input sequence from concatenating retrieved context. Methods based on classifier-free guidance, including standard CFG and ARAM, require both conditional and unconditional forward passes at each step (2 NFE). Compared with standard CFG, ARAM introduces additional overhead from computing the adaptive guidance scale, including the entropy and the SNR.

6 Conclusion

We introduce Adaptive Retrieval-Augmented Masked Diffusion (ARAM), a decoding framework that addresses knowledge conflict in MDM-based RAG. We propose an adaptive guidance rule based on a Signal-to-Noise Ratio, suppressing guidance when retrieved context is conflicting or uninformative. Experiments demonstrate that ARAM improves robustness under retrieval-prior conflict and yields strong gains over MDM-based baselines.

Limitations

While our findings demonstrate that ARAM effectively mitigates knowledge conflict in MDM-based RAG, we acknowledge several limitations. Our empirical claims are scoped to English-language, knowledge-intensive QA tasks using open-source MDMs, and the backbone coverage is restricted to LLaDA, LLaDA 1.5, and Dream as these are the primary publicly available instruction-tuned MDMs with sufficient capability for retrieval-augmented QA. Extension to broader diffusion families remains future work. It also remains an open question whether our method translates effectively to open-ended generation tasks or other decoding strategies. From a practical perspective, ARAM uses the same number of forward passes per denoising step as standard CFG, but introduces a small additional latency from computing the entropy and SNR terms as quantified in Section 5.5. Due to this overhead, it remains an important consideration for deploying ARAM in highly latency-sensitive production environments.

References

Kushal Arora, Timothy J O'Donnell, Doina Precup, Jason Weston, and Jackie CK Cheung. 2023. The stable entropy hypothesis and entropy-aware decoding: An analysis and algorithm for robust natural language generation. *arXiv preprint arXiv:2302.06784*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pages 9112–9141.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, and 1 others. 2016a. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, and 1 others. 2016b. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Sitao Cheng, Liangming Pan, Xunjian Yin, Xinyi Wang, and William Yang Wang. 2024. Understanding the interplay between parametric and contextual knowledge for large language models. *arXiv preprint arXiv:2410.08414*.

Eunseong Choi, June Park, Hyeri Lee, and Jongwuk Lee. 2025. Conflict-aware soft prompting for retrieval-augmented generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26969–26983.

Xin Dai, Pengcheng Huang, Zhenghao Liu, Shuo Wang, Yukun Yan, Chaojun Xiao, Yu Gu, Ge Yu, and Maosong Sun. 2026. Revealing the attention floating mechanism in masked diffusion models. *arXiv preprint arXiv:2601.07894*.

Monroe D Donsker and SR Srinivasa Varadhan. 1983. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on pure and applied mathematics*, 36(2):183–212.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-rex: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Pengxiang Li, Shilin Yan, Joey Tsai, Renrui Zhang, Ruichuan An, Ziyu Guo, and Xiaowei Gao.

2025. Adaptive classifier-free guidance via dynamic low-confidence masking. *arXiv preprint arXiv:2505.20199*.

Xinze Li, Sen Mei, Zhenghao Liu, Yukun Yan, Shuo Wang, Shi Yu, Zheni Zeng, Hao Chen, Ge Yu, Zhiyuan Liu, and 1 others. 2024. Rag-ddr: Optimizing retrieval-augmented generation using differentiable data rewards. *arXiv preprint arXiv:2410.13509*.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 7052–7063.

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

Pengfei Luo, Jingbo Zhou, Tong Xu, Yuan Xia, Linli Xu, and Enhong Chen. 2025. Imagescope: Unifying language-guided image retrieval via large multimodal model collective reasoning. In *Proceedings of the ACM on Web Conference 2025*, pages 1666–1682.

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

Kevin Rojas, Ye He, Chieh-Hsin Lai, Yuta Takida, Yuki Mitsufuji, and Molei Tao. 2025. Theory-informed improvements to classifier-free guidance for discrete diffusion models. *arXiv preprint arXiv:2507.08965*.

Yair Schiff, Subham Sekhar Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dalla-torre, Bernardo P de Almeida, Alexander Rush, Thomas Pierrot, and Volodymyr Kuleshov. 2024. Simple guidance mechanisms for discrete diffusion models. *arXiv preprint arXiv:2412.10193*.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025a. Adacad: Adaptively decoding to balance conflicts between contextual and parametric knowledge. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11636–11652.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025b. Retrieval-augmented generation with conflicting evidence. *arXiv preprint arXiv:2504.13079*.

Yuan Xia, Jingbo Zhou, Zhenhui Shi, Jun Chen, and Haifeng Huang. 2025. Improving retrieval augmented language model with self-reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pages 25534–25542.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2013–2018.

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

Chuanyue Yu, Jiahui Wang, Yuhan Li, Heng Chang, Ge Lan, Qingyun Sun, Jia Li, Jianxin Li, and Ziwei Zhang. 2026. Unlocking the potentials of retrieval-augmented generation for diffusion language models. *arXiv preprint arXiv:2601.11342*.

Xiaowei Yuan, Zhao Yang, Yequan Wang, Shengping Liu, Jun Zhao, and Kang Liu. 2024. Discerning and resolving knowledge conflicts through adaptive decoding with contextual information-entropy constraint. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3903–3922.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. 2026. Retrieval-augmented generation for ai-generated content: A survey. *Data Science and Engineering*, pages 1–29.

Fengqi Zhu, Rongzhen Wang, Shen Nie, Xiaolu Zhang, Chunwei Wu, Jun Hu, Jun Zhou, Jianfei Chen, Yankai Lin, Ji-Rong Wen, and 1 others. 2025. Llada 1.5: Variance-reduced preference optimization for large language diffusion models. *arXiv preprint arXiv:2505.19223*.

A Proofs and Derivations

A.1 Proof of Theorem 1

The Donsker-Varadhan (DV) variational representation (Donsker and Varadhan, 1983) of the Kullback-Leibler divergence for two distributions P and Q states for the function class $f : \mathcal{X} \to \mathbb{R}$ that:

$$D_{\mathrm{KL}}(P\|Q) = \sup_f \left(\mathbb{E}_P[f(x)] - \log \mathbb{E}_Q[\exp\left(f(x)\right)]\right).$$

We set $P = p_{\text{cond}}$, $Q = p_{\text{prior}}$, and $f(x) = \lambda s(x;\mathcal{C}) = \lambda \log \frac{p_{\text{cond}}(x)}{p_{\text{prior}}(x)}$ for $\lambda \in \mathbb{R}$. Then, substituting these into the DV representation yields:

$$D_{\mathrm{KL}} \geq \mathbb{E}_{p_{\text{cond}}}[\lambda s(x;\mathcal{C})] - \log \mathbb{E}_{p_{\text{prior}}}[\exp(\lambda s(x;\mathcal{C}))].$$

By definition, $\mathrm{IG}_t = D_{\mathrm{KL}}(p_{\text{cond}}\|p_{\text{prior}})$ and $Z_\lambda = \mathbb{E}_{p_{\text{prior}}}[\exp(\lambda s(x;\mathcal{C}))]$. Therefore, we obtain the lower bound:

$$\mathrm{IG}_t \geq \lambda \mathbb{E}_{p_{\text{cond}}}[s(x;\mathcal{C})] - \log Z_\lambda =: \mathcal{L}(\lambda).$$

□

A.2 Proof of Corollary 1.1

To evaluate $\mathcal{L}(\lambda)$ at $\lambda = 1$, we first expand the partition term Z_1:

$$\begin{aligned} Z_1 &= \mathbb{E}_{p_{\text{prior}}}[\exp(s(x;\mathcal{C}))] \\ &= \sum_{x\in\mathcal{V}} p_{\text{prior}}(x) \exp\left(\log \frac{p_{\text{cond}}(x)}{p_{\text{prior}}(x)}\right) \\ &= \sum_{x\in\mathcal{V}} p_{\text{cond}}(x) \\ &= 1. \end{aligned}$$

Consequently, the log-partition function is $\log Z_1 = 0$. Substituting this into $\mathcal{L}(1)$ gives:

$$\begin{aligned} \mathcal{L}(1) &= 1 \cdot \mathbb{E}_{p_{\text{cond}}}[s(x;\mathcal{C})] - \log(1) \\ &= \mathbb{E}_{p_{\text{cond}}}[s(x;\mathcal{C})] \\ &= \sum_{x\in\mathcal{V}} p_{\text{cond}} \log \frac{p_{\text{cond}}}{p_{\text{prior}}} \\ &= D_{\mathrm{KL}}(p_{\text{cond}}\|p_{\text{prior}}). \end{aligned}$$

Since $\mathcal{L}(\lambda)$ is bounded above by $D_{\mathrm{KL}}(p_{\text{cond}}\|p_{\text{prior}})$ according to Theorem 1, $\lambda = 1$ is a global maximizer. □

A.3 Derivation of Eq. (11)

We analyze the local geometry of $\mathcal{L}(\lambda)$ around the parametric anchor $\lambda = 0$. The objective function is:

$$\mathcal{L}(\lambda) = \lambda \mathbb{E}_{p_{\text{cond}}}[s(x;\mathcal{C})] - \log Z_\lambda,$$

where the partition function Z_λ is:

$$\begin{aligned} Z_\lambda &= \mathbb{E}_{p_{\text{prior}}}[\exp(\lambda s(x;\mathcal{C}))] \\ &= \sum_{x\in\mathcal{V}} p_{\text{prior}} \exp(\lambda s(x;\mathcal{C})). \end{aligned}$$

and $s(x;\mathcal{C}) = \log \frac{p_{\text{cond}}(x)}{p_{\text{prior}}(x)}$.

The first derivative with respect to λ is:

$$\mathcal{L}'(\lambda) = \mathbb{E}_{p_{\text{cond}}}[s(x;\mathcal{C})] - \frac{\partial}{\partial\lambda} \log Z_\lambda.$$

Using the property of the log-partition function, its derivative is the expectation under the tilted distribution $p_\lambda(x) := \frac{p_{\text{prior}}(x)\exp(\lambda s(x;\mathcal{C}))}{Z_\lambda}$:

$$\begin{aligned} \frac{\partial}{\partial\lambda} \log Z_\lambda &= \frac{1}{Z_\lambda} \sum_{x\in\mathcal{V}} p_{\text{prior}}(x) s(x;\mathcal{C}) \exp(\lambda s(x;\mathcal{C})) \\ &= \sum_{x\in\mathcal{V}} p_\lambda(x) s(x;\mathcal{C}) \\ &= \mathbb{E}_{p_\lambda}[s(x;\mathcal{C})]. \end{aligned}$$

Thus, $\mathcal{L}'(\lambda) = \mathbb{E}_{p_{\text{cond}}}[s(x;\mathcal{C})] - \mathbb{E}_{p_\lambda}[s(x;\mathcal{C})]$. At $\lambda = 0$, $p_0(x) = p_{\text{prior}}(x)$, which yields:

$$\mathcal{L}'(0) = \mathbb{E}_{p_{\text{cond}}}[s(x;\mathcal{C})] - \mathbb{E}_{p_{\text{prior}}}[s(x;\mathcal{C})].$$

Using the first derivative, the second derivative with respect to λ is:

$$\begin{aligned} \mathcal{L}''(\lambda) &= \frac{\partial}{\partial\lambda}\mathcal{L}' \\ &= \frac{\partial}{\partial\lambda}\left(\mathbb{E}_{p_{\text{cond}}}[s(x;\mathcal{C})] - \mathbb{E}_{p_\lambda}[s(x;\mathcal{C})]\right) \\ &= -\frac{\partial}{\partial\lambda}\mathbb{E}_{p_\lambda}[s(x;\mathcal{C})] \\ &= -\sum_{x\in\mathcal{V}} s(x;\mathcal{C}) \frac{\partial p_\lambda(x)}{\partial\lambda}. \end{aligned}$$

To get the derivative of $p_\lambda(x)$, we first calculate derivative of log-probability as:

$$\begin{aligned} \frac{\partial}{\partial\lambda} \log p_\lambda(x) &= \frac{\partial}{\partial\lambda}\Big(\log p_{\text{prior}}(x) + \lambda s(x;\mathcal{C}) \\ &\qquad - \log Z_\lambda\Big) \\ &= s(x;\mathcal{C}) - \frac{\partial}{\partial\lambda} \log Z_\lambda \\ &= s(x;\mathcal{C}) - \mathbb{E}_{p_\lambda}[s(x;\mathcal{C})]. \end{aligned}$$

Then, we obtain

$$\begin{aligned} \frac{\partial}{\partial\lambda} p_\lambda(x) &= p_\lambda(x) \frac{\partial}{\partial\lambda} \log p_\lambda(x) \\ &= p_\lambda(x)\Big(s(x;\mathcal{C}) - \mathbb{E}_{p_\lambda}[s(x;\mathcal{C})]\Big). \end{aligned}$$

Therefore, we calculate

$$\begin{aligned}\mathcal{L}''(\lambda) &= -\sum_{x\in\mathcal{V}} s(x;\mathcal{C})p_\lambda(x)\Big(s(x;\mathcal{C}) - \mathbb{E}_{p_\lambda}[s(x;\mathcal{C})]\Big) \\ &= -\Big(\mathbb{E}_{p_\lambda}[s(x;\mathcal{C})^2] - (\mathbb{E}_{p_\lambda}[s(x;\mathcal{C})])^2\Big) \\ &= -\mathrm{Var}_{p_\lambda}(s(x;\mathcal{C})).\end{aligned}$$

Evaluating at $\lambda = 0$ gives:

$$\mathcal{L}''(0) = -\mathrm{Var}_{p_{\text{prior}}}(s(x;\mathcal{C}))$$

Since $Z_0 = 1$, we have $\mathcal{L}(0) = 0$. The second-order Taylor expansion of $\mathcal{L}(\lambda)$ around $\lambda = 0$ is:

$$\begin{aligned}\mathcal{L}(\lambda) &\approx \mathcal{L}(0) + \lambda\mathcal{L}'(0) + \frac{\lambda^2}{2}\mathcal{L}''(0) \\ &= \lambda\left(\mathbb{E}_{p_{\text{cond}}}[s] - \mathbb{E}_{p_{\text{prior}}}[s]\right) - \frac{\lambda^2}{2}\mathrm{Var}_{p_{\text{prior}}}(s).\end{aligned}$$

To find the local maximum of this concave quadratic approximation, we set the derivative with respect to λ to zero:

$$\mathbb{E}_{p_{\text{cond}}}[s] - \mathbb{E}_{p_{\text{prior}}}[s] - \lambda\mathrm{Var}_{p_{\text{prior}}}(s) = 0.$$

Solving for λ yields the optimal scale λ^* formulated as the Signal-to-Noise Ratio:

$$\lambda^* = \frac{\mathbb{E}_{p_{\text{cond}}}[s] - \mathbb{E}_{p_{\text{prior}}}[s]}{\mathrm{Var}_{p_{\text{prior}}}(s)}.$$

B Implementation Details

Hyperparameters. In the retrieval component, we fetch the top-3 relevant documents per query. All benchmark evaluations are conducted on a randomly sampled subset of 1,000 instances per dataset. For all baseline methods, we adopt the default hyperparameters specified in their original papers. For the autoregressive baselines adapted to MDMs, we set the guidance scale to $\lambda = 2$ for CAD, and use $\alpha = 1$ with a threshold of 0.25 for COIECD. In AdaCAD, we substitute the static λ with the token-level Jensen-Shannon Divergence (JSD). For MDM-specific baselines, SPREAD modifies the re-masking strategy without altering the pre-softmax logits. For A-CFG, we use a re-masking proportion of $\rho = 0.7$ and a guidance scale of $\lambda = 1.5$ (corresponding to $w = 0.5$ in the original formulation). For the proposed ARAM, we calculate λ_t^i at each denoising step, setting the sensitivity parameter to $\beta = 0.1$ for LLaDA and $\beta = 0.5$ for Dream, and the maximum guidance scale to $\lambda_{\max} = 1.0$ for both LLaDA and Dream.

Prompt and Decoding Settings. We follow the decoding configuration and prompt design of Attention Floating (Dai et al., 2026). We utilize a temperature of 0, nucleus sampling with $p = 0.9$, a maximum generation length of 32 tokens, and 32 denoising steps. For the unmasking strategy, we adopt the default policies of the respective backbone models: low-confidence for LLaDA and entropy for Dream. We employ a zero-shot instruction prompt for LLaDA (Figure 5) and a few-shot prompt with in-context examples for Dream and other AR baselines (Figure 6). When retrieved documents are available, they are concatenated sequentially with index labels (e.g., "Passage 1: [Text]"). In the no-retrieval setting, the context field is populated with the string "No relevant context available."

Experiment Details for Robustness Analysis. For the analysis in Sec. 5.3, we utilize the outputs from the main experiments: the retrieval-free baseline, static RAG, and ARAM. For each backbone model, we align predictions across methods on the NQ benchmark and determine correctness using the exact-match criterion.

For the retrieval-prior interaction analysis (Table 2), we compare each retrieval-based method against the retrieval-free baseline on an instance-by-instance basis. Each sample is assigned to one of four categories: *Positive* if the no-retrieval prediction is incorrect but the retrieval-based method is correct, *Negative* if the no-retrieval prediction is correct but the retrieval-based method is incorrect, *Consistently Correct* if both are correct, and *Consistently Wrong* if both are incorrect. We report the proportion of instances in each category.

For the analysis of the retrieval context quality (Table 3), we stratify samples according to the quality of the retrieved context. We first define a sample as *Gold* if at least one retrieved document contains the answer. This is determined by normalized string matching, including lowercasing, article removal, and punctuation stripping, between the reference answer set and the retrieved documents. For the remaining non-gold (equivalent to *Noise Doc* in Sec. 5.4) samples, we further distinguish between *Non-Answering* and *Irrelevant* cases using an LLM-based annotation procedure that utilizes the prompt template detailed in Figure 4. Given the question, reference answer, and retrieved documents, the LLM is prompted to output one of two labels: *irrelevant*, if the documents are off-topic and provide no useful evidence for answering the

Table 7: Performance on LLaDA 1.5.

Method	EM	F1
LLaDA 1.5	15.8	22.0
LLaDA 1.5 + RAG	17.6	30.6
LLaDA 1.5 + Ours	**24.2**	**35.9**

Table 8: Comparison with Self-RAG under the same retrieval setup. Llama-2-7B-chat with standard RAG is included as a same-backbone reference.

Method	NQ		TriviaQA		MARCO QA		HotpotQA		T-REx	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Baseline	16.8	28.0	47.8	58.0	2.1	22.3	15.6	24.7	2.4	9.1
Self-RAG-7B	27.1	38.1	24.5	40.6	0.6	14.7	5.9	16.2	7.8	14.6

question, or *Non-Answering*, if the documents are topically related but fail to provide sufficient support for the correct answer. For stable annotation, we use GPT-4.1-mini with deterministic decoding.

We further verify its stability by re-annotating each non-gold example three times with the order of the top-3 passages randomly permuted, and using majority voting as the final label. The trial-level agreement with the majority label is 97.2%. This suggests that the stratification used in Table 3 is robust to annotation noise.

C Additional Experiments

C.1 Scalability across model configurations

To verify the generalizability of ARAM, we evaluate the framework on a more recent architecture, LLaDA 1.5 (Zhu et al., 2025). As shown in Table 7, ARAM improves the EM and F1 scores over the standard RAG baseline using $\lambda_{\max} = 1.0, \beta = 1.0$. This outcome indicates that our SNR-based adaptive guidance mechanism functions consistently across different model variants.

C.2 Training-based Conflict-Aware Methods

Beyond decoding-time interventions, recent work mitigates retrieval-prior conflict by training the model over retrieved evidence (Asai et al., 2024; Xia et al., 2025). As a representative example, Self-RAG (Asai et al., 2024) trains a language model to generate reflection tokens that decide whether to retrieve, critique retrieved passages, and assess the supportiveness of its own outputs. Unlike ARAM, which keeps the backbone fixed and resolves conflicts during inference, these methods require additional training on reflection-token trajectories. We separate this comparison from the main results in Table 1 because Self-RAG (i) is built on a different backbone (Llama-2-7B-chat) and (ii) belongs to a fundamentally different category of training-based methods. For a fair comparison, we evaluate the publicly released Self-RAG-7B checkpoint using the same retrieval corpus and top-3 document setting as ARAM. We additionally evaluate Llama-2-7B-chat with standard RAG as a baseline.

In Table 8, Self-RAG improves over baseline (Llama-2-7B-chat + RAG) on NQ and T-REx, where reflection-token reranking benefits single-hop factoid retrieval. However, it underperforms on TriviaQA, HotpotQA, and MARCO QA, where the task requires aggregating evidence across multiple passages rather than reranking a single source. This pattern indicates that Self-RAG's learned reflection is best matched to focused single-hop retrieval, while the broader robustness gains seen on multi-document tasks are not automatically transferred. When compared against ARAM in Table 1, ARAM matches or exceeds Self-RAG on the majority of these benchmarks in our setting, without requiring reflection-token supervision.

Table 9: Comparison of noise proxies. Average EM and F1 scores are reported.

Noise Proxy	LLaDA		Dream	
	EM	F1	EM	F1
$\mathrm{Var}_{p_{\text{prior}}}(s)$	20.3	32.5	33.5	43.2
$\mathcal{H}(p_{\text{prior}})$	20.5	32.4	33.6	42.8
$\lvert\mathcal{H}(p_{\text{prior}}) - \mathcal{H}(p_{\text{cond}})\rvert$	23.7	35.3	33.4	42.8
$\mathcal{H}(p_{\text{cond}})$	**24.9**	**36.1**	**33.9**	**43.6**

Table 10: Comparison of stability function. Average EM and F1 scores are reported.

Signal Proxy	LLaDA		Dream	
	EM	F1	EM	F1
- (Raw SNR)	21.5	33.5	33.5	43.4
Tanh	**24.9**	**36.1**	**33.9**	**43.6**

D Additional Ablation Studies

Extended Comparison of Noise Proxies. To further investigate the formulation of the noise proxy, we compare the conditional entropy $\mathcal{H}(p_{\text{cond}})$ against three alternative metrics: score variance $\mathrm{Var}_{p_{\text{prior}}}(s)$, prior entropy $\mathcal{H}(p_{\text{prior}})$, and the absolute entropy difference $|\mathcal{H}(p_{\text{prior}}) - \mathcal{H}(p_{\text{cond}})|$. As shown in Table 9, $\mathcal{H}(p_{\text{cond}})$ yields the highest EM and F1 scores across both Dream and LLaDA models. The prior entropy $\mathcal{H}(p_{\text{prior}})$ omits the uncertainty introduced by the external context, leading to lower guidance accuracy. The absolute difference formulation also underperforms compared to direct conditional entropy. This indicates that $\mathcal{H}(p_{\text{cond}})$

alone provides a sufficient and stable measure for the noise introduced by the retrieved context.

Effect of the Stability Function. Table 10 presents the ablation on the stability function used to bound the adaptive guidance scale. We compare the raw SNR scaling against the Tanh-scaled formulation. The results demonstrate that applying the Tanh function improves generation performance for both models. This improvement is attributed to the bounding property of Tanh, which prevents numerical explosion of λ_t when the denominator (noise) approaches zero during the unmasking steps. Bounding the guidance scale ensures stable generation without over-amplifying the contextual logits.

LLM Annotation Prompt

You are labeling retrieval context quality for a QA example.

Question:
{query}

Reference answers:
{answers}

Retrieved passages:
{passages}

Task:
Return exactly one label:
- irrelevant: the passages are largely unrelated to the question and do not provide useful evidence.
- related_non_answering: the passages are related to the question but do not clearly contain the correct answer.

Rules:
- If the passages are about the same entity/topic/question but fail to answer it clearly, return non answering.
- Use irrelevant only when the passages are genuinely off-topic or retrieval junk.
- Return only the label, with no explanation.

Figure 4: Prompt template for LLM-based retrieval context quality annotation.

Prompt for LLaDA

Use the following passages as context if they are helpful, but provide a SHORT and DIRECT answer.

RULES:
- First, check the context passages for the answer
- If the context contains the answer, use it
- If not, use your general knowledge
- Answer must be 1-10 words maximum
- Do NOT include any explanations or extra text
- Do NOT repeat or summarize the passages

Context:
{context}

Question: {query}

Short Answer:

Figure 5: Prompt template used for LLaDA.

Prompt for Dream

You are a helpful assistant that answers questions.

INSTRUCTIONS:
1. First, check if the answer is in the provided context passages
2. If the answer is in the context, use it
3. If the context doesn't contain the answer, use your general knowledge
4. Always provide a direct, concise answer (typically 1-10 words)
5. Do NOT include explanations, reasoning, or phrases like "based on" or "according to"
6. Never say "no answer found" - always attempt to answer using available information

Context:
{context}

Here are examples showing the expected answer format:

Example 1:
Context:
Passage 1: The Eiffel Tower is located in Paris, France. It was completed in 1889.
Question: Where is the Eiffel Tower located?
Answer: Paris, France

Example 2:
Context:
Passage 1: Albert Einstein was born on March 14, 1879 in Ulm, Germany.
Question: When was Albert Einstein born?
Answer: March 14, 1879

Example 3:
Context:
Passage 1: The Great Wall of China stretches over 13,000 miles.
Question: What is the capital of Japan?
Answer: Tokyo
(Note: This answer uses general knowledge since the context doesn't contain it)

—
Now answer the following question in the same concise format:
Question: {query}
Answer:

Figure 6: Few-shot prompt template used for the Dream and other AR backbones.