Unlocking the Potentials of Retrieval-Augmented Generation for Diffusion Language Models: A Semantic Drift Perspective

Anonymous ACL submission

Abstract

Diffusion Language Models (DLMs) have recently demonstrated remarkable capabilities in natural language processing tasks. However, the potential of Retrieval-Augmented Generation (RAG), which shows great successes for enhancing large language models (LLMs), has not been well explored, due to the fundamental difference between LLM and DLM decoding. To fill this critical gap, we systematically test the performance of DLMs within the RAG framework. Our findings reveal that DLMs coupled with RAG show promising potentials with stronger dependency on contextual information, but suffer from limited generation precision. We identify a key underlying issue: Response Semantic Drift (RSD), where the generated answer progressively deviates from the query's original semantics, leading to low precision content. We trace this problem to the denoising strategies in DLMs, which fail to maintain semantic alignment with the query throughout the iterative denoising process. To address this, we propose Semantic-Preserving REtrieval-Augmented Diffusion (SPREAD), a novel framework that introduces a query-relevance-guided denoising strategy. By actively guiding the denoising trajectory, SPREAD ensures the generation remains anchored to the query's semantics and effectively suppresses drift. Experimental results demonstrate that SPREAD significantly enhances the precision and effectively mitigates RSD of generated answers within the RAG framework.

1 Introduction

While auto-regressive Large Language Models (LLMs) have dominated natural language processing, recent advancements in Diffusion Language Models (DLMs) (Nie et al., 2025; Ye et al., 2025; Austin et al., 2021) present an attractive alternative, demonstrating superior performance in tasks requiring fine-grained control due to their parallel decoding and iterative refinement capabilities (Li et al., 2025b). However, similar to LLMs, DLMs are also confronted with challenges such as knowledge gaps and hallucination. On the other hand, Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) has become a foundational technology for enabling LLMs to generate reliable outputs grounded in external knowledge sources (Fan et al., 2024; Zhao et al., 2024). Consequently, integrating DLMs with RAG to mitigate these issues is a pivotal direction, yet remains largely unexplored in the literature. This problem is also highly challenging as DLMs and LLMs have fundamentally different generation processes. Whether the unique architecture of DLMs necessitates different RAG design warrants further investigation.

To bridge this gap, we conduct preliminary experiments to combine DLMs with RAG for the question answering task (the details are introduced in Section 3.1). As shown in Table 1, DLMs can achieve scores comparable to or even surpassing those of LLMs on metrics such as F1. A particularly notable finding is their stronger tendency for high copy rate (CR), which we define as the proportion of words in the generated answer that appear identically within the provided source context. However, as further shown in Figure 1, the absolute precision of DLM-generated answers remains a bottleneck. The results suggest that while the copied retrieved content is factually grounded in the context, it may often include redundant details. Consequently, the answer's relevance to the query's semantics is diminished.

We identify a key issue underlying this phenomenon: Response Semantic Drift (RSD) within the generated answer. In contrast to the semantic drift in (Spataru, 2024), which measures the separation between correct and incorrect facts in generated text, or other task-specific semantic drifts (Spataru, 2024; Khan and Syed, 2025; Mollah et al., 2025), our RSD focuses on the gradual drift in semantic between consecutive sentences

Method	NQ		Trivia		HotpotQA		MuSiQue		Multihop		UltraDomain	
	F1	CR	F1	CR	F1	CR	F1	CR	F1	CR	F1	CR
Qwen2.5-7B	28.33	77.59	40.95	82.68	47.44	77.79	22.25	64.87	26.98	58.67	22.21	45.42
Qwen2.5-32B	23.17	71.69	33.19	78.58	38.42	72.80	20.56	59.76	24.32	58.07	22.33	44.28
Llama3-8B	21.59	70.14	36.73	62.36	39.77	61.91	17.24	59.18	15.11	57.78	28.85	59.59
LLaDA-8B	33.09	81.01	39.80	82.24	44.36	79.82	27.03	72.36	**46.34**	**63.39**	30.91	61.78
Dream-7B	**38.42**	**84.81**	**46.47**	**84.41**	**52.40**	**81.88**	**30.56**	**77.65**	38.46	58.63	**36.92**	**71.35**

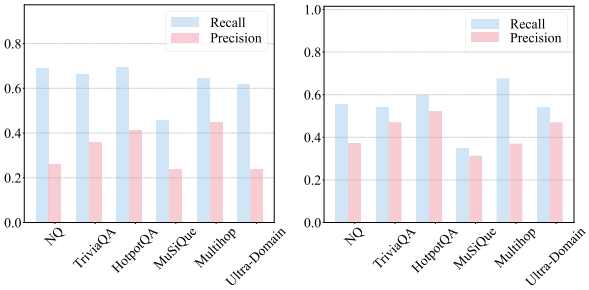
Table 1: Performance comparison of LLMs and DLMs equipped with RAG. CR indicates copy rate. DLMs in RAG achieve high F1 scores and exhibit high copy rates from the provided context.

within the generated content, which can lead to outputs that are coherent at a local level but lack query semantics relevance and precision. Our analysis further traces this issue to the inherent denoising strategies in DLMs, particularly confidence-based methods (Nie et al., 2025; Ye et al., 2025; Wei et al., 2025; Li et al., 2025a; Wang et al., 2023). These strategies determine denoising priorities based on global and token-level certainty estimates. However, they fundamentally overlook the alignment of the denoising process with the query semantics. Consequently, the generation process may gradually deviate from the original semantics, resulting in the observed semantic drift and precision loss.

To address this challenge and fully unlock the potentials of RAGs for DLMs, we propose Semantic-Preserving REtrieval-Augmented Diffusion (SPREAD). Our core idea lies in assessing the semantic relevance between generated token and query, and then utilizes these scores to determine token selection. This strategy actively guides the iterative denoising process of DLMs at each denoising step, ensuring that the decoded text remains closely aligned with query semantics throughout the generation process. Compared to previous denoising strategies that rely on global contexts and token-level certainty, SPREAD dynamically adjusts the denoising process, using query semantics as a constraint to suppress semantic drift. Experimental results show that our method consistently and substantially enhances the precision of the generated answers and effectively mitigates the issue of RSD across all datasets.

Our contributions are summarized as follows:

- We investigate the integration of DLMs and RAG and identify Response Semantic Drift as a key bottleneck, leading to redundant and incoherent outputs. To the best of our knowledge, we are the first to study this critical issue.



(a) Results of LLaDA. (b) Results of Dream.

Figure 1: Precision and recall of DLMs equipped with RAG. While both LLaDA and Dream achieve high recall, their precision is consistently relatively low.

- We propose SPREAD, a novel framework that introduces a query-relevance-guided denoising strategy to dynamically align the generation process with the query semantics.
- We conduct extensive experiments to validate the effectiveness of SPREAD. Our method significantly mitigates semantic drift and enhances output quality for both LLaDA and Dream across 6 benchmark datasets.

2 Related Work

2.1 Retrieval-Augmented Generation

RAG enhances language models by integrating external knowledge through joint modeling of retrieval and generation (Lewis et al., 2020). Subsequent work has pursued better knowledge integration via richer representations (Huang et al., 2025) and end-to-end retriever optimization to align retrieval with generation goals (Zhou and Chen, 2025; Li et al., 2025c). Other efforts coordinate the RAG pipeline through joint training (Gao et al., 2025), reinforcement learning (Wang et al., 2024a), or unified ranking-generation schemes (Yu et al., 2024). Furthermore, research has expanded to specialized

data forms, such as developing RAG frameworks for table-based knowledge (Chen et al., 2024), and explored novel paradigms like organizing knowledge into multiple partitions for enhanced agentic reasoning (Wang et al., 2024b).

While RAG is a vital solution for knowledge supplementation and hallucination reduction in generative models, its application to the DLM paradigm is unexplored. A key research question is whether the DLM's unique architecture and decoding process require new methodologies for such integration.

2.2 Diffusion Language Models

DLMs offer a non-autoregressive alternative for LLMs, which generate text through iterative denoising for parallel decoding and enhanced controllability (Li et al., 2025b). Recent research on DLMs has branched along several paths. Initial explorations with continuous DLMs operating in latent spaces demonstrated advantages in efficiency and controllability (Xu et al., 2025; Chen et al., 2025). However, for textual data, discrete diffusion methods have been developed as a more direct approach to better model text tokens (Lou et al., 2024). Building on this, masked and likelihood-based methods have demonstrated competitive performance, narrowing the gap with autoregressive models (Sahoo et al., 2024; Gulrajani and Hashimoto, 2023). Hybrid approaches like block diffusion further balance efficiency and flexibility (Arriola et al., 2025).

However, the integration of DLMs with RAG remains a notably underexplored area. Even with a preliminary attempt for continuous DLMs (Anonymous, 2025), the fundamental challenge of how to effectively ground the iterative denoising process of DLMs on retrieved knowledge has not been systematically explored. A core challenge is how to ensure that the iterative denoising process over tokens remains semantically anchored to the dynamically retrieved knowledge throughout the generation.

3 Exploratory Experiments and Analyses

In this section, we conduct a detailed analysis to explore the potential and identify challenges of combining DLMs with RAG. We begin by presenting preliminary experiments, which reveal a key performance bottleneck, i.e., low answer precision, and motivate a deeper investigation into their generation process.

3.1 Preliminary Results

To investigate the performance of DLMs equipped with RAG, we conducted some preliminary experiments on. For the retrieval component of RAG, we employ a dense vector retriever powered by the NV-Embed-v2 encoder (Lee et al., 2025). All documents are first segmented into chunks of 2,000 characters. For each query, we retrieve the top-5 most relevant passages based on vector similarity. These retrieved passages are then concatenated and provided as the context to the language model for answer generation. We compare the performance of DLMs against LLMs on question answering benchmarks (detailed experimental setups are provided in Section 5.1).

The results are shown in Table 1 and Figure 1. On the one hand, DLMs achieve factual accuracy scores comparable to or even surpassing their autoregressive counterparts on metrics such as F1. This appears to stem from their strong inclination to copy content verbatim from the provided context. However, compared to high recall, the absolute precision of DLM-generated responses often emerges as a distinct performance bottleneck. These results suggest that while directly adopting RAGs can empower DLMs to copy factually grounded text from retrieved knowledge, it may also include considerably redundant and irrelevant information. Consequently, the answer's relevance to the query's semantics is not fully optimized.

Overall, the results suggest that combining DLMs with RAGs provides promising potentials, but directly adopting the off-the-shelf RAG does not achieve optimal results. In the following subsections, we provide in-depth analyses for this phenomenon and motivate our proposed method.

3.2 Response Semantic Drift

To gain further insights, we take a closer look at the iterative denoising process in discrete DLMs and observe a critical phenomenon: the generated content, while potentially factually related to the context, is not fully faithful to the query. Concretely, the outputs tend to be redundant and gradually diverge from the core semantics of the query. We term this problem affecting the generation quality as Response Semantic Drift (RSD).

Formally, given a generated answer text a, we segment it into a consecutive sequence of sentences $S = \{s_1, s_2, ..., s_{|S|}\}$. We measure the average semantic incoherence between adjacent sentences

within the generated answer, calculated as follows:

$$\mathrm{RSD}(a) = \frac{1}{|S|-1} \sum_{i=1}^{|S|-1} \mathrm{Dist}(s_i, s_{i+1}), \quad (1)$$

where $\mathrm{Dist}(s_i, s_{i+1})$ denotes the semantic distance between two adjacent sentences s_i and s_{i+1}. A larger value indicates more pronounced semantic drift between consecutive sentences, corresponding to poor overall coherence.

We employ cosine distance based on a high-performance pre-trained semantic encoder to quantify the semantic difference between sentences:

$$\mathrm{Dist}(s_i, s_{i+1}) = 1 - \mathrm{cosine}(\mathrm{Enc}(s_i), \mathrm{Enc}(s_{i+1})). \quad (2)$$

We provide an illustrative example in Figure 2. Consider a simple query such as "What year was the Eiffel Tower built?" along with its relevant context. A straight-forward DLM-RAG baseline model might generate an answer as shown in Figure 2(a). Although each sentence is independently correct and derived from the retrieved context, the response as a whole exhibits significant semantic drift. For instance, the content may jump from the tower's structural features to its construction timeline, thereby failing to precisely focus on the query's semantics regarding the construction date. In contrast, an ideal response would directly address the core of the query as shown in Figure 2(b). Alternatively, it could expand on the core topic while maintaining high thematic coherence, e.g., "The Eiffel Tower was built from 1887 to 1889. Its construction took two years." In this response, the RSD between consecutive sentences is minimal.

3.3 Semantic Drift in the Denoising Process

The problem of RSD stems from a fundamental conflict between the local optimization of the DLM's denoising process and the global semantic requirements of question answering with RAG tasks. The standard diffusion denoising process relies on strategies based on confidence or entropy. These strategies operate on a step-wise basis, which are inherently local and greedy. At each iteration, its primary objective is to select the most probable token based on the retrieved context and the current generative state. However, this process of locally optimal denoising lacks a global alignment for semantics. In question answering tasks, the objective is to maintain global alignment of the entire sequence with the query's semantics. The

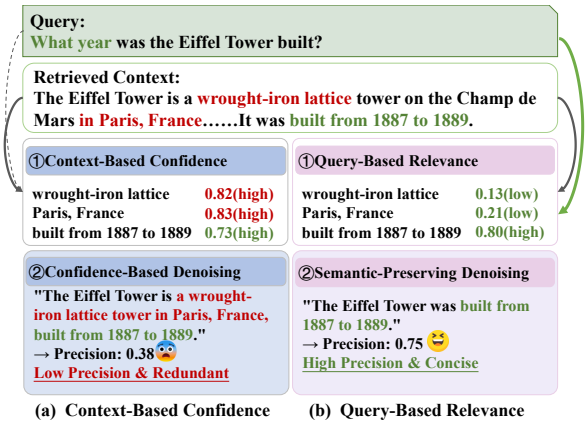


Figure 2: An illustration comparing different denoising strategies in DLMs. (a) Context-Based Confidence. The model relies on its internal confidence score, which is potentially biased by the global retrieved context, causing the model to generate irrelevant information. (b) Query-Based Relevance. An ideal approach maintains a persistent focus on the query's semantics, which allows the model to prioritize relevant facts and generate a concise and relevant response.

final generated answer must remain strictly aligned with the semantics of the initial query. Since standard denoising strategies overlook this overall goal, the influence of the original query is reduced over successive iterations. Consequently, the generation process falls into a series of local decisions, and minor deviations at each step gradually accumulate and amplify throughout the iterative process. Finally, the model is likely to forget the query and generate technically correct but irrelevant answers. Therefore, resolving this issue requires a strategy that can guide the denoising process and enforce a global semantic constraint, especially the alignment with the query's semantics, which is the key motivation for our proposed method.

4 Method

In this section, we introduce our method in detail. Our central idea is to introduce a query-relevance-guided denoising strategy at each iterative denoising step, which dynamically adjusts denoising priorities by evaluating the semantic relevance between candidate tokens at masked positions and the query, thereby guiding the generation process to consistently anchor to the query semantics.

4.1 Overall Framework

The framework of SPREADis illustrated in Figure 3. We introduce a Query-Token Semantic Relevance Assessment module at each denoising step,

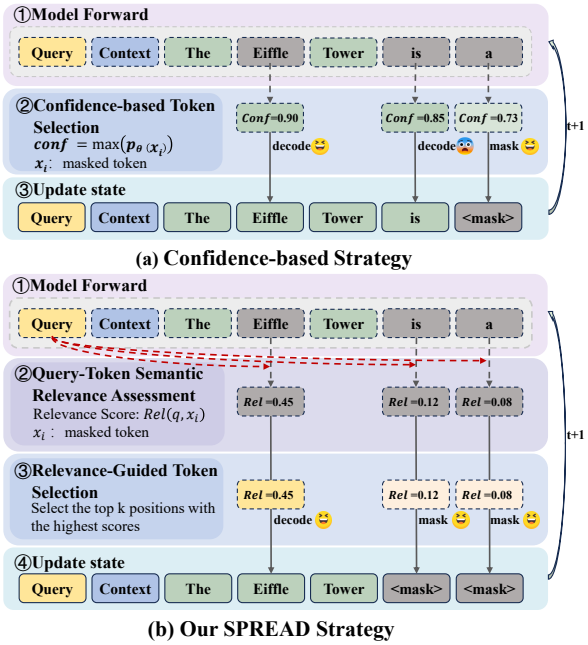


Figure 3: An overview and illustrative comparison of our proposed method: (a) Confidence-based Strategy: The existing methods rely on local model confidence for token selection, lacking guidance from the global query semantics and leading to semantic drift during generation. (b) Our SPREAD strategy: We assess the relevance of each candidate token to the query and directly guides the token selection, actively steering the generation process to maintain continuous alignment with the query semantics.

following the standard forward pass of the DLM. Without requiring additional encoding overhead, this module uses the model's internal representations in the forward pass to compute a semantic relevance score between each masked position and the query. A Relevance-Guided Token Selection strategy then utilizes these scores to determine which token predictions should be decoded in the current step and which should remain masked for refinement in subsequent steps. This process is repeated in every iteration, ensuring that continuous semantic alignment is maintained throughout the entire denoising process. Next, we elaborate on these two modules.

4.2 Query-Token Semantic Similarity Assessment

We realize the semantic alignment by measuring semantic similarity. Our method is based on the principle that a token's similarity to the query in the semantic space directly measures its consistency with the query's overall semantics. By computing and utilizing this semantic relevance, we can effectively steer the generation process to remain anchored to the query's semantics. To precisely measure the consistency of these local generation decisions with the global query semantics, we design a semantic relevance assessment module based on the model's internal representations.

Contextualized Embedding Extraction Given the sequence state x_t at the denoising step t and the query q, we utilize the model's own forward pass to obtain contextualized hidden layer representations. Let the model output hidden states for all layers be $\mathbf{H} \in \mathbb{R}^{L \times N \times d}$, where L is the number of layers, N is the sequence length, and d is the hidden dimensionality. We select the hidden state of the last layer as the final contextualized representation where the representation of the i-th position is denoted as $\mathbf{h}_i \in \mathbb{R}^d$. Compared to static token embeddings, $\mathbf{h}_i$ incorporates deep semantic information from the sequence context, allowing it to more accurately reflect the semantic content. This approach requires only a single model forward pass and introduces no additional computation for token embeddings.

Semantic Similarity Computation To maintain consistency in evaluation, the query q also obtains its complete semantic representation through a forward pass of the same model. We compute the cosine similarity between the hidden state $\mathbf{h}_i$ and the query representation $\mathbf{h}_q$ as the measure of their semantic relevance:

$$\mathrm{Sim}(i, q) = \frac{\mathbf{h}_i \cdot \mathbf{h}_q}{\|\mathbf{h}_i\| \|\mathbf{h}_q\|}. \tag{3}$$

This metric is aligned in computational principle with the RSD metric defined in section 3.

Then we map the similarity to a normalized relevance score. To smoothly amplify high-relevance signals, we apply the Sigmoid function $\sigma(\cdot)$:

$$\mathrm{Rel}(i, q) = \sigma(\mathrm{Sim}(i, q)), \tag{4}$$

where $\mathrm{Rel}(i, q) \in (0, 1)$ directly represents the estimated probability that the content at position i is semantically aligned with the query q.

4.3 Relevance-Guided Token Selection

Based on the computed $\mathrm{Rel}(i, q)$, SPREAD employs a straightforward and efficient strategy to guide the denoising process.

At denoising step t, the model produces an initial prediction distribution for all masked positions

$i \in M_t$. We first calculate the relevance scores $\{\text{Rel}(i, q)\}_{i \in M_t}$ for these positions. Subsequently, we select the top k positions with the highest scores, where k is a predefined or dynamically determined hyperparameter, and decode the predicted tokens at these positions, replacing the original masks. The remaining masked positions retain their masked status, waiting for further evaluation and prediction in subsequent iterative steps.

By prioritizing the retention of the token most relevant to the query, our strategy directly suppresses the accumulation of unrelated information and enforces alignment with the query's semantics, which reduces RSD, and thereby enhances the precision of the generated answers. The pseudo-code of SPREAD is shown in Appendix A.

5 Experiments

This section presents a systematic experimental evaluation to assess the effectiveness of our proposed SPREAD method. Our experiments are designed to answer three key research questions:

- **Q1:** Can SPREAD consistently improve answer precision and suppress RSD across various datasets and model baselines?
- **Q2:** Does SPREAD show superior performance compared to existing DLM denoising strategies?
- **Q3:** How does SPREAD affect different aspects of DLMs, such as copy rate and efficiency?

5.1 Experimental Setup

Datasets To comprehensively evaluate model performance across diverse knowledge retrieval and integration scenarios, we conduct experiments on six widely-used open-domain question-answering (QA) datasets: Natural Questions (NQ) (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017), which focus on factual QA, the multi-hop QA datasets HotpotQA (Yang et al., 2018), MuSiQue (Trivedi et al., 2022), and MultiHop-RAG (Tang and Yang, 2024), which require reasoning by synthesizing information from multiple documents, and UltraDomain (Qian et al., 2025), which challenges the model's capability for in-depth understanding within a vertical domain. These datasets encompass varying difficulties and requirements. The statistics and more details are provided in Appendix B.

Evaluation Metrics Guided by the problem analysis in Section 3, we employ Precision (denoted as P) and Response Semantic Drift (RSD) as our core evaluation metrics, directly measuring the factual accuracy and internal semantic coherence of the the answers, respectively. Additionally, we use the Copy Rate (denoted as CR) and Redundancy (denoted as R) as auxiliary analytical metrics to gain deeper insights into the model's generation behavior, with both metrics calculated at the word level. Specifically, Copy Rate quantifies the extent of verbatim reproduction from the source context by computing the proportion of overlapping words between the generated answer and the provided context. Redundancy measures the repetitiveness within the generated answer itself, indicating an inefficient or circular information presentation.

Baseline Models We select two representative open-source DLMs as our base models: LLaDA-8B-Instruct (Nie et al., 2025) and Dream-7B-Instruct (Ye et al., 2025). For LLaDA, we compare three denoising strategies, including the native random strategy and confidence-based strategy in their original article, and our SPREAD. For Dream, we compare its entropy-based strategy and maskgit-plus strategy, and SPREAD.

Implementation Details We use a dense vector retriever based on the NV-Embed-v2 encoder, with document chunks of 2,000 characters, retrieving the top-5 relevant passages as context for each query. To ensure a fair comparison, all experiments, including baselines and our method, share the same generation hyper-parameters: the number of diffusion steps is 128, the maximum new token length is 512, and the sampling temperature is 0.1. All reported results are averaged across all queries within that dataset.

5.2 Main Results

To answer **Q1** and **Q2**, Table 2 provides a comprehensive performance comparison between SPREAD and the baselines, evaluated on both LLaDA and Dream base models with all denoising strategies across all datasets.

The results demonstrate that SPREAD consistently and significantly improves the precision of generated answers in all experimental settings, achieving an average absolute increase of 15.34% on LLaDA and 30.90% on Dream. This directly shows that guiding the generation process with query semantics effectively filters out irrelevant

	Model	NQ		Trivia		HotpotQA		MuSiQue		Multihop		UltraDomain	
		P	RSD	P	RSD	P	RSD	P	RSD	P	RSD	P	RSD
LLM	Qwen2.5-7B	21.02	32.88	37.81	16.47	45.40	26.41	19.61	40.50	25.53	49.95	13.42	72.42
	Qwen2.5-32B	15.25	47.80	28.42	28.42	33.77	43.72	16.47	60.95	22.62	54.66	13.47	71.86
	LLaMA3-8B	14.47	59.21	32.55	29.23	35.22	38.45	13.19	55.78	12.99	61.79	18.97	71.37
DLM	LLaDA (random)	20.63	29.35	30.38	16.09	35.60	27.19	19.28	28.16	28.07	37.37	21.58	63.60
	LLaDA (low-conf)	26.05	23.01	36.04	11.12	41.38	24.40	23.91	20.92	44.91	24.28	23.89	65.57
	LLaDA (SPREAD)	**30.37**	**21.79**	**42.46**	**8.91**	**48.71**	**19.26**	**26.67**	**17.82**	**48.73**	**19.79**	**29.33**	**61.48**
	Dream (maskgit-plus)	37.11	12.60	47.08	3.48	52.42	6.04	31.35	4.51	37.05	32.13	36.24	62.62
	Dream (entropy)	35.25	12.61	44.30	3.44	50.16	5.54	29.33	4.02	35.85	32.12	41.69	59.27
	Dream (SPREAD)	**52.79**	**4.83**	**55.71**	**1.29**	**59.37**	**4.05**	**37.78**	**1.13**	**59.09**	**9.46**	**58.19**	**24.66**

Table 2: The results of different methods for the question answering task. P indicates Precision (↑) and RSD indicates Response Semantic Drift (↓). The best results are in **bold** and the second-best results are underlined.

information, leading to more focused and accurate responses. Corresponding to the gain in precision, SPREAD also significantly reduces Response Semantic Drift, with an average relative reduction exceeding 10.92% on LLaDA and 61.18% on Dream. This validates our core hypothesis that aligning the generation trajectory with the query semantics fundamentally mitigates RSD between sentences, thereby enhancing overall coherence.

5.3 Further Analysis

To address **Q3** and gain further insights into our method, we conduct a series of analyses.

5.3.1 Impact on Copy Rate

In this subsection, we analyze the model's behavior in copying content from retrieved context. As shown in Figure 4, SPREAD does not significantly alter the overall Copy Rate of the answers. Meanwhile, Redundancy is decreased significantly. The results support our argument that in RAG, how many retrieved contents are copied alone are insufficient measurements, but what content is copied is far more critical. By guiding the model to prioritize copying content strongly related to the query semantics, SPREAD significantly enhances the quality and integration of the retrieved information for enhanced question answering ability.

5.3.2 Multi-dimensional Analysis

To provide an integrated view of our method's performance across multiple evaluation dimensions, we use a radar chart to summarize five key metrics: F1, Recall, and Precision, together with ROUGE-1 and ROUGE-L which measures unigram and longest-sequence overlap against reference texts,

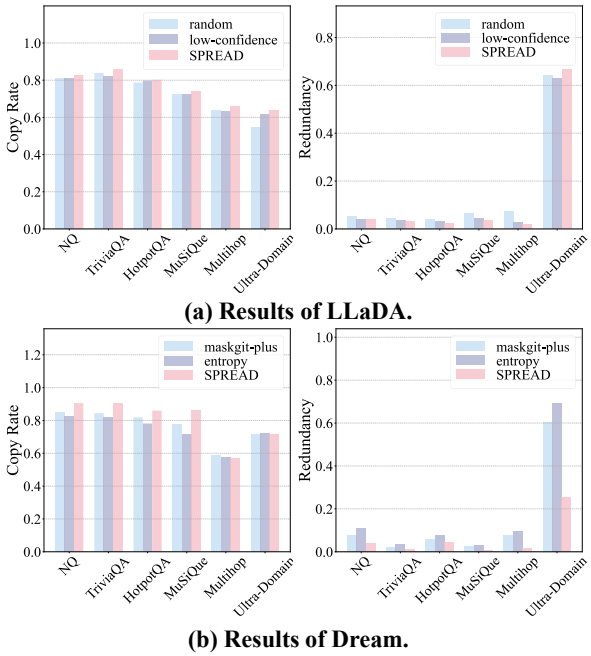


Figure 4: A comparative analysis of Copy Rate and Redundancy across different denoising strategies.

respectively. As shown in Figure 5, SPREAD covers a noticeably larger area than all baseline strategies, demonstrating its overall superiority. The chart highlights that our method achieves dominant performance in four of the five metrics, including F1, Precision, ROUGE-L, and ROUGE-1. This reflects an enhanced ability to generate answers that are factually accurate. Meanwhile, its Recall score remains on par with the strongest baselines, which indicates that our method boosts precision and coherence without sacrificing its capacity to cover relevant information from the retrieved context. This holistic evaluation confirms that SPREAD delivers a more balanced and effective generation quality,

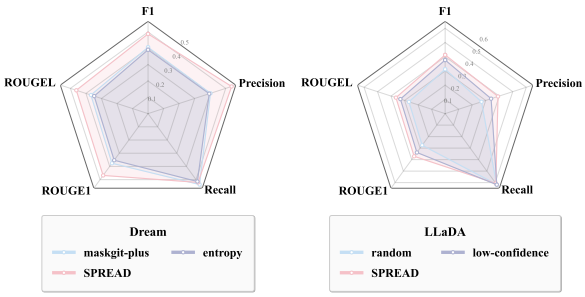


Figure 5: The results of multi-dimensional comparison of different masking strategies.

Model	Strategy	Avg Time (s)	Tokens/sec
Dream	maskgit-plus	19.64	30.72
	entropy	**18.34**	**32.41**
	SPREAD	19.02	31.49
LLaDA	random	**27.33**	**21.08**
	low-confidence	27.39	21.03
	SPREAD	27.56	20.90

Table 3: Efficiency comparison of different denoising strategies. The best results are in **bold**.

excelling in the aspects most crucial for reliable question answering tasks.

5.3.3 Efficiency Analysis

To comprehensively assess the practical efficiency of our method, we conduct a detailed efficiency analysis comparing the computational cost of SPREAD against various baseline denoising strategies. We adopt two metrics, including the Average Generation Time, measured for generating full answers on the validation set, which reflects the end-to-end cost of the iterative denoising process, and Tokens/second, which directly compares the decoding speed across different models. All experiments are conducted on a single NVIDIA H200 GPU with 141 GB of memory.

As shown in Table 3, the computational overhead introduced by SPREAD is minimal. Compared to the baseline, SPREAD incurs an increase of only 0.23s on LLaDA and 0.68s on Dream. This slight overhead is attributed to the computation of query-token semantic relevance scores, which is extremely efficient as it reuses the model's existing hidden states without requiring extra encoding.

In summary, our SPREAD strategy achieves significant improvements in answer precision and coherence by introducing only negligible computational time overhead, which ensures its practicality and efficiency for real-world applications.

Metric Pair	LLaDA	Dream
RSD-Precision	-0.36	-0.24
RSD-Redundancy	0.46	0.55
RSD-Copy Rate	-0.30	-0.16

Table 4: The Pearson correlation of RSD with Precision, Copy Rate and Redundancy.

5.3.4 Relationship with Other Metrics

To empirically analyze the relationship between RSD and classical evaluation metrics in text generation, we calculate the Pearson correlation coefficients between RSD and Precision, Redundancy, and Copy Rate, with results presented in Table 4.

The results reveal a moderate negative correlation between RSD and precision across both DLM backbones, statistically confirming that semantic incoherence within generated text is a plausible factor that impairs generation precision.

Furthermore, the results show a strong positive correlation between RSD and redundancy, affirming their intrinsic connection in terms of redundant information. In other words, our metric can potentially identify generated content that is novel in wording but suffers from topic jumps or logical fractures between sentences, a blind spot for existing redundancy metrics.

Lastly, RSD shows a weak negative correlation with copy rate. The result suggests that a high copy rate per se is not related to semantic drift, i.e., a model can perform "high-quality copying" by coherently integrating valuable information from the retrieved context and generate consistent answers.

In sum, the empirical results further validate that RSD can systematically quantifies the internal semantic coherence of generated answers.

6 Conclusion

This work investigates the integration of discrete DLMs within the RAG paradigm. We first identify and formalize a key challenge: Response Semantic Drift (RSD), which quantifies the progressive deviation from query semantics in generated answers. To address this limitation, we propose SPREAD, a novel framework that steers the generation trajectory using a query-relevance-guided denoising strategy. Comprehensive experiments demonstrate that SPREAD delivers consistent and significant improvements by effectively enhancing answer precision and substantially reducing RSD.

7 Limitations

While SPREAD demonstrates significant efficacy, its performance is contingent on the relevance of retrieved context, while co-designing the retriever and generator can create a feedback loop for enhanced performance. Additionally, we current only validate our method on benchmark QA tasks and we aim to extend it for more open-ended generation tasks such as story writing.

References

Anonymous. 2025. Retrieval-augmented diffusion language model for generative commonsense reasoning. In *Submitted to ACL Rolling Review - February 2025*.

Marianne Arriola, Aaron Gokaslan, Justin T Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sekhar Sahoo, and Volodymyr Kuleshov. 2025. Block diffusion: Interpolating between autoregressive and diffusion language models. In *The Thirteenth International Conference on Learning Representations*.

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

Si-An Chen, Lesly Miculicich, Julian Eisenschlos, Zifeng Wang, Zilong Wang, Yanfei Chen, Yasuhisa Fujii, Hsuan-Tien Lin, Chen-Yu Lee, and Tomas Pfister. 2024. Tablerag: Million-token table understanding with language models. *Advances in Neural Information Processing Systems*, 37:74899–74921.

Tianqi Chen, Shujian Zhang, and Mingyuan Zhou. 2025. Dlm-one: Diffusion language models for one-step sequence generation. *arXiv preprint arXiv:2506.00290*.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6491–6501.

Guangze Gao, Zixuan Li, Chunfeng Yuan, Jiawei Li, Wu Jianzhuo, Yuehao Zhang, Xiaolong Jin, Bing Li, and Weiming Hu. 2025. D-rag: Differentiable retrieval-augmented generation for knowledge graph question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35386–35405.

Ishaan Gulrajani and Tatsunori B Hashimoto. 2023. Likelihood-based diffusion language models. *Advances in Neural Information Processing Systems*, 36:16693–16715.

Haoyu Huang, Yongfeng Huang, Junjie Yang, Zhenyu Pan, Yongqiang Chen, Kaili Ma, Hongzhi Chen, and James Cheng. 2025. Retrieval-augmented generation with hierarchical knowledge. *arXiv preprint arXiv:2503.10150*.

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.

Rehan Khan and Amaan Syed. 2025. Measuring semantic drift across generational corpora: A framework using pretrained embeddings. *Preprints*.

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*, 7:452–466.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. Nv-embed: Improved techniques for training llms as generalist embedding models. In *The Thirteenth International Conference on Learning Representations*.

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, Yefan Zhou, Dilxat Muhtar, Lu Yin, Shilin Yan, Li Shen, Yi Liang, Soroush Vosoughi, and Shiwei Liu. 2025a. Diffusion language models know the answer before decoding. In *NeurIPS 2025 Workshop on Efficient Reasoning*.

Tianyi Li, Mingda Chen, Bowei Guo, and Zhiqiang Shen. 2025b. A survey on diffusion language models. *arXiv preprint arXiv:2508.10875*.

Xinze Li, Sen Mei, Zhenghao Liu, Yukun Yan, Shuo Wang, Shi Yu, Zheni Zeng, Hao Chen, Ge Yu, Zhiyuan Liu, and 1 others. 2025c. Rag-ddr: Optimizing retrieval-augmented generation using differentiable data rewards. In *The Thirteenth International Conference on Learning Representations*.

Aaron Lou, Chenlin Meng, and Stefano Ermon. 2024. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Proceedings of the 41st International Conference on Machine Learning*, pages 32819–32848.

Sabbir Mollah, Rohit Gupta, Sirnam Swetha, Qingyang Liu, Ahnaf Munir, and Mubarak Shah. 2025. The telephone game: Evaluating semantic drift in unified models. *arXiv preprint arXiv:2509.04438*.

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. In *ICLR 2025 Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy*.

Hongjin Qian, Zheng Liu, Peitian Zhang, Kelong Mao, Defu Lian, Zhicheng Dou, and Tiejun Huang. 2025. Memorag: Boosting long context processing with global memory-enhanced retrieval augmentation. In *Proceedings of the ACM on Web Conference 2025*, pages 2366–2377.

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

Ava Spataru. 2024. Know when to stop: A study of semantic drift in text generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3656–3671.

Yixuan Tang and Yi Yang. 2024. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. In *First Conference on Language Modeling*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Renzhi Wang, Jing Li, and Piji Li. 2023. Infodiffusion: Information entropy aware diffusion process for non-autoregressive text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13757–13770.

Zheng Wang, Shu Teo, Jieer Ouyang, Yongjun Xu, and Wei Shi. 2024a. M-rag: Reinforcing large language model performance through retrieval-augmented generation with multiple partitions. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1966–1978.

Zheng Wang, Shu Teo, Jieer Ouyang, Yongjun Xu, and Wei Shi. 2024b. M-rag: Reinforcing large language model performance through retrieval-augmented generation with multiple partitions. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1966–1978.

Qingyan Wei, Yaojie Zhang, Zhiyuan Liu, Dongrui Liu, and Linfeng Zhang. 2025. Accelerating diffusion large language models with slowfast: The three golden principles. *arXiv preprint arXiv:2506.10848*.

Minkai Xu, Tomas Geffner, Karsten Kreis, Weili Nie, Yilun Xu, Jure Leskovec, Stefano Ermon, and Arash Vahdat. 2025. Energy-based diffusion language models for text generation. In *The Thirteenth International Conference on Learning Representations*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380.

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*.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. *Advances in Neural Information Processing Systems*, 37:121156–121184.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. 2024. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*.

Jiawei Zhou and Lei Chen. 2025. Openrag: Optimizing rag end-to-end via in-context retrieval learning. In *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*.

A Algorithm Details

In this section, we provide the pseudo-code of our method in Algorithm 1.

B More Details about Datasets

Detailed dataset descriptions are as follows.

- **Natural Questions** (Kwiatkowski et al., 2019): A large-scale question answering benchmark dataset consisting of real user queries paired with Wikipedia pages.

- **TriviaQA** (Joshi et al., 2017): A large-scale reading comprehension dataset with complex, compositional questions and independently collected evidence documents, designed to challenge cross-sentence and multi-document reasoning.

- **HotpotQA** (Yang et al., 2018): A challenging question answering dataset built on Wikipedia, containing question-answer pairs that require

Dataset	NQ	Trivia	HotpotQA	MuSiQue	MultiHop	UltraDomain
Num of Queries	1,000	1,000	1,000	1,000	1,000	958
Num of Passages	7,830	1,432	66,581	29,898	609	161
Average Length of Question	9.19	13.21	16.17	13.96	45.99	12.74
Average Length of Answer	2.61	2.02	2.44	2.73	1.30	39.31

Table 5: Dataset Statistics

finding and reasoning over multiple supporting documents, with sentence-level annotated supporting facts.

- **MuSiQue** (Trivedi et al., 2022): A systematically constructed multi-hop question answering dataset consisting of 2-4 hop questions generated via a bottom-up composition of single-hop queries.
- **MultiHop-RAG** (Tang and Yang, 2024): A benchmark dataset for retrieval-augmented generation that evaluates models on multi-hop queries requiring iterative retrieval and reasoning across multiple documents.
- **UltraDomain** (Qian et al., 2025): A benchmark dataset spanning diverse specialized domains, designed to evaluate models on long-context, high-level queries that require holistic understanding and multi-information synthesis.

Specifically, the MultiHop-RAG dataset is licensed under ODC-BY 1.0, MuSiQue under CC BY 4.0, and HotpotQA under CC BY-SA 4.0. The use of other established benchmarks follows standard scholarly citation and fair use conventions. We utilize all datasets strictly as benchmarks for experimental evaluation, which aligns with their intended research use. The statistics of these datasets are shown in Table 5.

Algorithm 1 Semantic-Preserving Retrieval-Augmented Diffusion (SPREAD)

Require: Query q, retrieved context c, discrete diffusion model M, the total denoising steps T, tokens to retain per step k

Ensure: Generated answer $\mathbf{x}_0$

$\mathbf{x}_T \leftarrow [\text{MASK}] * L$
$\mathbf{h}_q \leftarrow M.\text{encode_query}(q)$
Prefix q and c to the model input.
for $t = T$ **down to** 1 **do**
  // Step 1: Obtain Contextualized Representations
  $\mathbf{h} \leftarrow M.\text{forward}(\mathbf{x}_t)$
  $M_t \leftarrow \text{GetMaskedPositions}(\mathbf{x}_t)$
  // Step 2: Compute Relevance Scores
  $\mathbf{Rel} \leftarrow \text{zeros}(|M_t|)$
  for $idx, i \in \text{enumerate}(M_t)$ **do**
    $\mathbf{Sim}(\mathbf{i}, \mathbf{q}) \leftarrow \text{cosine}(\mathbf{h}_i, \mathbf{h}_q)$
    $\mathbf{Rel}(\mathbf{i}, \mathbf{q}) \leftarrow \sigma(\mathbf{Sim}(\mathbf{i}, \mathbf{q}))$
  end for
  // Step 3: Relevance-Guided Top-k Selection & Update
  $\text{TopKIndices} \leftarrow \text{Argsort}(\mathbf{Rel}(\mathbf{i}, \mathbf{q}))[:k]$
  for $j \in M_t[\text{TopKIndices}]$ **do**
    $\mathbf{x}_t[j] \leftarrow M.\text{predict_token}(\mathbf{x}_t, j)$
  end for
  $\mathbf{x}_{t-1} \leftarrow \mathbf{x}_t$
end for
return $\mathbf{x}_0$

C Ethical Considerations

While our proposed method aims to enhance general language model performance and is verified in benchmark question answering tasks without any specific ethical designs, potential risks include its misuse for generating persuasive misinformation and the perpetuation of biases from the underlying semantic encoder. Integrating source verification and fairness testing could reduce these risks.

D AI Usage Statement

During the preparation of this paper, we utilize AI writing assistants to enhance the quality of our writing. Its primary role was to assist in checking for grammatical errors, correcting spelling, improving sentence structure and polishing phrasing.We hereby declare that all core scientific ideas, experimental designs, data analysis, and final conclusions are the original work of the authors. The authors take full responsibility for the accuracy and originality of the manuscript's content.