Revealing the Attention Floating Mechanism in Masked Diffusion Models

Xin Dai[1], **Pengcheng Huang**[1], **Zhenghao Liu**[1†], **Shuo Wang**[2†],
Yukun Yan[2], **Chaojun Xiao**[2], **Yu Gu**[1], **Ge Yu**[1], **Maosong Sun**[2]
[1]School of Computer Science and Engineering, Northeastern University, Shenyang, China
[2]Department of Computer Science and Technology, Tsinghua University, Beijing, China

Abstract

Masked diffusion models (MDMs), which leverage bidirectional attention and a denoising process, are narrowing the performance gap with autoregressive models (ARMs). However, their internal attention mechanisms remain under-explored. This paper investigates the attention behaviors in MDMs, revealing the phenomenon of *Attention Floating*. Unlike ARMs, where attention converges to a fixed sink, MDMs exhibit dynamic, dispersed attention anchors that shift across denoising steps and layers. Further analysis reveals its Shallow Structure-Aware, Deep Content-Focused attention mechanism: shallow layers utilize floating tokens to build a global structural framework, while deeper layers allocate more capability toward capturing semantic content. Empirically, this distinctive attention pattern provides a mechanistic explanation for the strong in-context learning capabilities of MDMs, allowing them to double the performance compared to ARMs in knowledge-intensive tasks. All codes and datasets are available at https://github.com/NEUIR/Attention-Floating.

1 Introduction

Large language models (LLMs) (Touvron et al., 2023; Qwen et al., 2024) have achieved remarkable success across a wide range of generation and reasoning tasks (Wei et al., 2022). The prevailing approach for LLMs has been autoregressive models (ARMs), where a Transformer is trained to predict tokens from left to right. Recent research on ARMs has uncovered a notable *attention sink* (Gu et al., 2024) phenomenon: a significant portion of the attention mass is systematically absorbed by a few initial tokens at the start of the sequence, which function as prominent static anchors for attention allocation. This rigid attention pattern biases the information flow toward early positions, which can lead to the "lost-in-the-middle" issue (Liu et al., 2024; Yao et al., 2025).

Diffusion language models (DLMs) (Li et al., 2022; Liu et al., 2023; Shabalin et al., 2025b,a; Tae et al., 2025) have recently emerged as a promising alternative to the autoregressive paradigm, generating text through a multi-step denoising process that relaxes the rigid left-to-right generation constraint in ARMs. As the most prominent instantiation of this paradigm, masked diffusion models (MDMs) begin with a fully masked sequence, then iteratively predict and fill a subset of masked positions across a series of denoising steps. These models employ bidirectional attention, enabling each position to attend to all others. As the sequence evolves from a fully masked state to a fully visible one, its visibility changes across denoising stages. However, the impact of the gradual denoising process on the attention mechanism remains unclear. Existing work has primarily analysed attention phenomena in ARMs, such as induction heads (Olsson et al., 2022) and attention sinks (Gu et al., 2024), yet the internal workings of MDMs are still largely unexplored.

In this paper, we investigate the attention mechanism of MDMs and systematically analyze their attention dynamics. Similar to ARMs, MDMs also exhibit a subset of token positions that receive disproportionately large attention mass. However, unlike ARMs where this behavior often concentrates at a fixed sink at the start of the sequence (typically <BOS>), we identify the phenomenon of *attention floating*: in MDMs, these dominant-attention anchors are dispersed across multiple positions and can shift across denoising steps and layers (Figure 2). We refer to tokens at these positions as *floating tokens*. Furthermore, we uncover a distinctive mechanism underlying this behavior: *Shallow Structure-Aware, Deep Content-Focused*. Through a geometric decomposition of the pre-softmax attention logits (QK) and an analysis of the special-

†Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 8266–8286

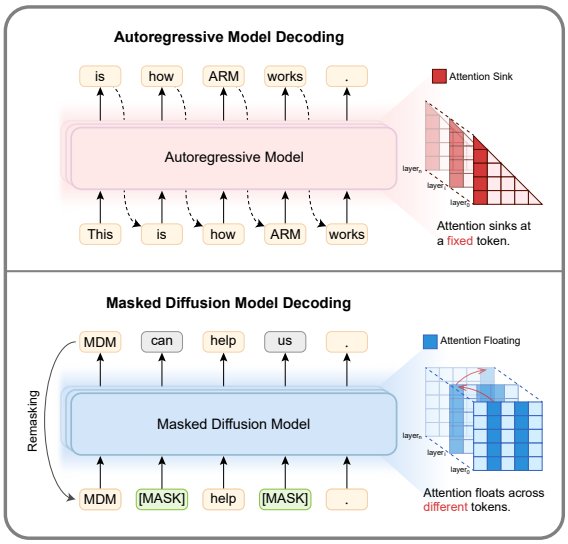


Figure 1: Comparison of ARMs and MDMs.

ization of retrieval heads, we demonstrate that the shallow layers rely on structurally floating tokens to form a global framework, while the deeper layers gradually shift attention toward tokens that carry semantic information.

Empirically, we demonstrate that the attention floating mechanism enhances knowledge utilization from in-context inputs. MDMs nearly double the performance gains of ARMs on knowledge-intensive tasks. To further dissect the underlying drivers of this performance gap, we conduct a series of stress tests to evaluate the models' resilience against contextual noise, positional biases, and complex evidence layouts. These tests reveal that MDMs maintain greater stability in diverse configurations than ARMs (Yao et al., 2025). Finally, we employ region-level attention flow analysis, a coarse-grained visualization that tracks how attention mass moves between regions of the input (e.g., <BOS>, query, and evidence), to uncover the underlying mechanism: unlike ARMs that remain rigidly anchored at the sequence start, MDMs dynamically reorganize their internal information pathways to track relevant context actively. These findings provide a mechanistic explanation for the superior robustness of MDMs under in-context learning.

2 Background: Generative Paradigms and Attention Mechanisms

In this section, we provide the background for our analysis. We first clarify the architectural differences between autoregressive models (ARMs) and masked diffusion models (MDMs), which motivate our investigation into their attention mechanisms. Subsequently, we introduce the attention sink definition and phenomenon in ARMs, which serves as a comparative baseline for characterizing the "attention floating" phenomenon in MDMs.

2.1 Generative Paradigms: From Autoregressive Decoding to MDMs

To understand the attention mechanisms in LLMs, we begin by contrasting the architectural differences between autoregressive models (ARMs) and masked diffusion models (MDMs) in their decoding paradigms.

As shown in Figure 1, the dominant paradigm, ARMs, generate text through a strict left-to-right sequential prediction process. This relies on causal self-attention, where each token is restricted to attending only to its predecessors. In contrast, MDMs employ bidirectional attention, enabling each token to attend to all other positions in the sequence, regardless of their order, at every denoising step. Starting from a fully masked sequence, MDMs conduct denoising steps to recover the target response during inference. Specifically, they progressively reconstruct the text over diffusion time steps, updating multiple tokens in a single step, which facilitates parallel generation. The transition from causal to bidirectional attention fundamentally alters the flow of information within LLMs. Unlike ARMs, which face limitations such as restricted receptive fields, MDMs provide a global context view. However, it remains unclear how this bidirectional visibility, coupled with time-dependent denoising, impacts the model's attention patterns. To explore this, we first provide background on existing analyses of attention mechanisms in Section 2.2.

2.2 Attention Analysis in Language Models

A large body of prior work has investigated how to analyze attention mechanisms in language models in order to identify the most influential parts of the input (Clark et al., 2019). With the emergence of LLMs, research has predominantly focused on ARMs. Recent studies show that ARMs exhibit a distinctive *attention sink* phenomenon (Gu et al., 2024), which has been widely analysed and exploited in practical settings such as streaming inference and long-context generation.

Specifically, in Transformer-based ARMs, left-to-right dependencies are enforced via self-attention with a causal mask. When a token se-

quence $X = \{x_1, \ldots, x_n\}$ is fed into the model, ARMs encode it using the self-attention mechanism. At the l-th layer, the attention weight from the i-th token to the j-th token is obtained by averaging the attention scores over the m attention heads of ARMs:

$$A^{\ell}_{i \to j} = \frac{1}{m} \sum_{h=1}^{m} \left[\text{Softmax}\left(\frac{Q_i^{(\ell,h)} K^{(\ell,h)\top}}{\sqrt{d_h}} \right) \right]_j, \tag{1}$$

where $Q^{(\ell,h)}$ and $K^{(\ell,h)}$ denote the query and key matrices of the h-th attention head at layer ℓ, and d_h is the head dimension. Then, the average attention received by position j at layer ℓ is computed by taking the mean of the attention weights over all n input tokens X:

$$A^{\ell}_{j} = \frac{1}{n} \sum_{i=1}^{n} A^{\ell}_{i \to j}, \tag{2}$$

Finally, the j-th token is regarded as the sink token if it satisfies the following criteria:

$$A^{\ell}_{j} > \frac{1}{n-1} \sum_{\substack{k=1 \\ k \neq j}}^{n} A^{\ell}_{k} + \epsilon, \tag{3}$$

where k ranges over all positions in the input sequence X and ϵ is a predefined threshold. In all experiments, we set $\epsilon = 3$. In ARMs, early tokens are visible to all subsequent positions under the causal mask, making them natural global sinks, so the <BOS> token frequently dominates in practice. In contrast, MDMs adopt bidirectional attention mechanisms. It remains unclear whether the attention sink phenomenon persists under such attention patterns, which motivates our further investigation.

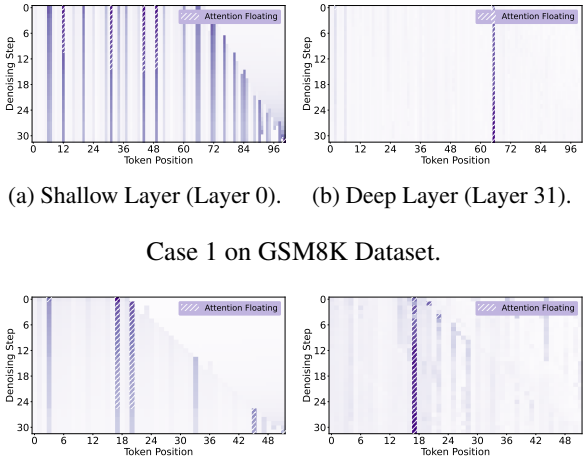
3 Attention: Floating Rather than Sinking in MDMs

Following the attention sink phenomenon observed in ARMs, this section further investigates the attention mechanism of MDMs and reveals the presence of *attention floating*. We then analyse the category of tokens that receive a larger attention weight (floating tokens) in MDMs.

3.1 Attention Floating Phenomenon in MDMs

To characterize how attention evolves across different layer depths and denoising steps, we analyze the attention floating mechanism of MDMs through attention visualization.

Attention Floating Visualization. To illustrate the attention behavior in MDMs, we visualize the per-token attention weights over the input sequence X from the shallow layer ($A_j^{\ell=0}$) to the deep layer ($A_j^{\ell=31}$), computed using Eq. 2.



(a) Shallow Layer (Layer 0). (b) Deep Layer (Layer 31).

Case 1 on GSM8K Dataset.

(c) Shallow Layer (Layer 0). (d) Deep Layer (Layer 31).

Case 2 on 2wiki Dataset.

Figure 2: Positional Drift of Attention Floating across Different Layers and Denoising Steps in MDM.

In MDMs, we observe an *attention floating* phenomenon, where positions receiving large attention mass shift across layers and further drift as denoising progresses. Moreover, this pattern exhibits task-dependent variations. Figure 2 compares two representative cases on GSM8K and 2WikiMQA and shows that the *attention floating* phenomenon is consistently present across both layer and task. Across layers, the floating tokens concentrate at clearly different token positions: in the shallow layer (Figure 2a and Figure 2c), high-attention columns spread across multiple positions and gradually shift toward later token positions as the denoising step increases (from around position 17 before step 14 to position 34 after step 14, and reaching position 46 after step 26 in Figure 2c), whereas in the deep layer (Figure 2b and Figure 2d) they become much sparser and concentrate on different positions. Moreover, in the same layer, the floating positions differ across tasks: Case 1 and Case 2 exhibit high-attention columns at different token positions, suggesting that floating adapts to task-specific input structure rather than remaining fixed at a single position. We provide additional visualizations across all layers for MDMs in Figure 10 and Figure 11 (Appendix A.2).

These observations motivate two complementary analyses along layers and task settings. Section 4 takes a layer-wise view to decompose attention

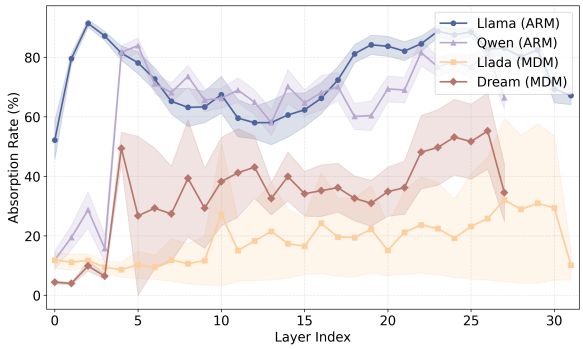


Figure 3: Layer-Wise Attention Absorption Rate in ARMs and MDMs. ARMs concentrate a disproportionately large fraction of attention on a single sink token (typically <BOS>) across layers, whereas MDMs exhibit consistently lower and more distributed absorption, consistent with attention floating.

behavior in MDMs. Section 5 takes a task-wise view to further examine the link between attention floating and robustness in learning from context.

Attention Absorption Rate. We then identify the positions $\mathcal{S}$ of the tokens receiving dominant attention weights using Eq. 3, and use them to quantify how sink tokens in ARMs or floating tokens in MDMs collectively absorb attention.

Specifically, we define the attention absorption rate of the ℓ-th layer as:

$$\text{Absorb}(\mathcal{S}, \ell) = \sum_{j \in \mathcal{S}} A_j^{\ell} \times 100\% \tag{4}$$

where A_j^{ℓ} denotes the head-averaged attention received by position j at layer ℓ, and $\mathcal{S}$ represents the position set of sink or floating tokens. Figure 3 presents the absorption rates of typical ARMs (Llama and Qwen) and MDMs (Llada and Dream) across different layers. Both ARMs use a single <BOS> token as the sink token and exhibit an extremely strong attention sink phenomenon. The sink token absorbs a disproportionately large fraction of the total attention mass across layers. In contrast, MDMs yield lower absorption values. This clear discrepancy indicates that ARMs induce a rigid concentration of attention around the sink token, whereas MDMs display a weaker and more distributed absorption pattern, which aligns precisely with our notion of *attention floating*.

3.2 Which Tokens Become Floating Tokens

Beyond positional drift, we further investigate which types of tokens are prone to becoming floating tokens in MDMs.

For each identified floating token, we classify its underlying vocabulary item into two categories: (i) structural tokens, including special control tokens (e.g., <BOS>, <|mdm_mask|>) as well as conventional formatting tokens such as punctuation and other layout symbols; and (ii) lexical tokens, covering content words and subwords. Unlike ARMs, where the attention sink phenomenon is largely dominated by a single special token <BOS>, as shown in Table 7, floating tokens in MDMs are predominantly composed of high-frequency structural tokens. Within this structural category, approximately 2% of all detected floating tokens correspond to the model-specific denoising mask token <|mdm_mask|>. These tokens function less as carriers of semantic content and more as "structural controllers" that signal text boundaries and layout organization, thereby anchoring local context and stabilizing the overall sequence structure. The frequency breakdown of floating tokens is summarized in Appendix A.2.

4 Shallow Structure–Aware, Deep Content–Focused Attention in MDMs

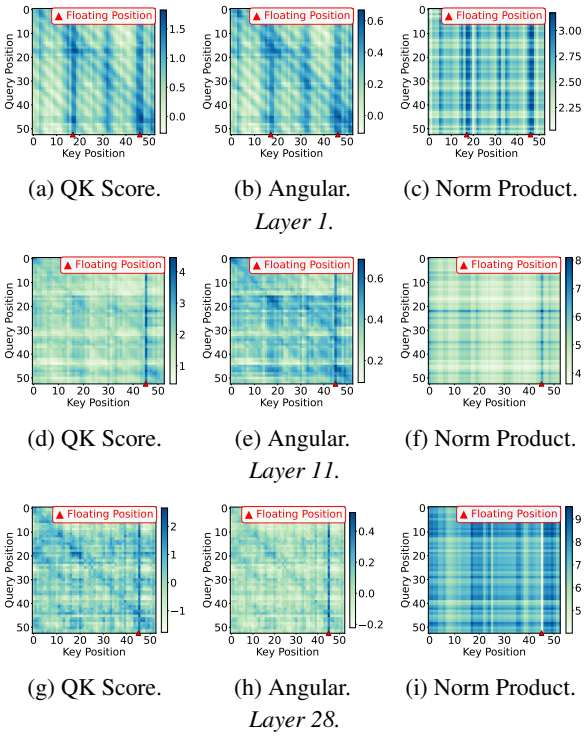
In this section, we take a layer-wise, model-level view of attention in MDMs. We analyze how geometric factors of attention vary across layers, leading to the hypothesis that attention floating shifts from structural bias in shallow layers to semantic content bias in deeper layers, and we then verify this hypothesis by locating retrieval-specialized heads across depth and measuring their contribution to context-following behavior.

Self-Attention Decomposition. To systematically analyze attention floating in MDMs, we start from the QK scoring mechanism analysis.

Existing work (Gu et al., 2024) on ARMs has revealed that the salience of sink positions is primarily manifested as a systematic advantage in the directional (angular) term, while column-wise differences in the scale (norm product) term are relatively weak, and this phenomenon can be summarized as a form of *key bias*. To disentangle the effect of vector magnitude from directional alignment in the representation space of MDMs, we explicitly decompose the QK score as follows:

$$\mathbf{Q}\mathbf{K}^{\top} = \|\mathbf{Q}\| \, \|\mathbf{K}\| \cos\theta, \tag{5}$$

where $\mathbf{Q}$ and $\mathbf{K}$ denote the query and key vectors and θ is the angle between them. This decomposition separates the contribution of vector norms $\|\mathbf{Q}\| \, \|\mathbf{K}\|$ from that of angular alignment $\cos\theta$.



(a) QK Score. (b) Angular. (c) Norm Product.

Layer 1.

(d) QK Score. (e) Angular. (f) Norm Product.

Layer 11.

(g) QK Score. (h) Angular. (i) Norm Product.

Layer 28.

Figure 4: QK Geometric Decomposition across Different Layers in MDM. Floating positions become most prominent in the middle layers, where their QK advantage is jointly supported by angular alignment and norm magnitude; in deeper layers, the norm advantage weakens and attention becomes less exclusively tied to floating positions, indicating a shift toward content-focused attention.

Figure 4 illustrates the QK decomposition, and heatmaps covering all layers are provided in Appendix A.2. The horizontal axis corresponds to key positions, while the vertical axis corresponds to query positions. These floating tokens, identified from attention statistics, are marked along the horizontal axis. At each depth, we jointly present the pre-softmax score map (QK Score) together with its decomposed directional alignment component (Angular) and scale component (Norm Product). This visualization enables a direct in-depth comparison to assess whether the QK advantage of floating key columns over other key columns is mainly attributable to stronger angular alignment, scale amplification, or their combined effect.

The visualization results for MDMs indicate that, in shallow layers, the QK distribution (Figure 4a) is relatively dispersed, with no clear contrast formed between floating and non-floating positions. This suggests that shallow-layer attention is still in an exploratory phase, not yet having established a stable structural anchoring pattern. As depth increases, the QK contrast between floating and non-floating positions reaches its peak: in Figure 4d, the vertical stripes at floating positions become particularly prominent, while the QK scores in non-floating regions remain relatively low. Combined with our empirical finding in Section 3.2 that the floating region is almost entirely occupied by structural tokens, this pattern suggests that attention at this stage preferentially latches onto structural anchors to stabilize denoising and information aggregation. At this stage, the QK score advantage of floating tokens is jointly driven by direction (Figure 4e) and scale (Figure 4f). In deeper layers (roughly after layer 20), the scale term in Figure 4i exhibits an opposite trend: floating positions show substantially smaller scale magnitudes than non-floating positions. Consequently, the scale term no longer contributes positively to the elevated QK scores of floating columns, and the remaining advantage of the QK score can be attributed more consistently to the directional alignment term (Figure 4h) itself. Meanwhile, the QK scores at non-floating positions also rise noticeably, causing the overall distribution between floating and non-floating positions to become more balanced. This indicates that deep-layer attention, while maintaining focus on structural anchors, also begins to allocate more weight to tokens carrying semantic content.

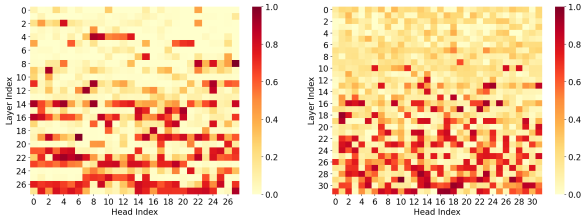
Synthesizing these layer-wise patterns, we propose the following hypothesis: shallow layers are still exploring global information; as depth increases, attention relies more heavily on structural anchors and establishes a stable structural framework; deep layers then build upon this framework by progressively reallocating the attention toward tokens that carry semantic content. We refer to this transition from shallow exploration, through increasingly strong structural anchoring, to deep-layer content-driven behavior as the *Shallow Structure-Aware, Deep Content-Focused* attention mechanism. As a quantitative proxy, we compute the layer-wise structural-to-lexical attention ratio and find that it generally decreases from shallow to deep layers, further supporting our hypothesis (Appendix Table 5).

Retrieval Head Analyses. To further verify the above hypothesis, we conduct an analysis following Wu et al. (2024) on retrieval-specialized heads.

Specifically, we assign each attention head a retrieval score that quantifies how frequently the head allocates top-k ($k = 5$) attention to the ground-truth needle tokens during final answer genera-

Model	Open-Domain QA						Multi-hop QA		Slot Filling		Avg.
	NQ		TQA		Marco QA		HotpotQA		T-REx		
	Acc	*F1*	*Acc*	*F1*	*Acc*	*F1*	*Acc*	*F1*	*Acc*	*F1*	
Autoregressive Models											
Llama	15.69	32.07	61.52	70.78	0.30	16.78	10.61	23.83	23.28	34.01	28.89
Llama w/ RAG	35.92	50.63	74.84	81.15	0.90	15.15	20.93	29.80	24.60	28.61	36.25 (+7.37)
Qwen	14.95	27.20	52.58	59.14	0.20	11.20	19.05	27.75	31.20	37.65	28.09
Qwen w/ RAG	34.16	50.07	72.61	80.35	0.50	16.11	24.50	34.05	30.02	34.97	37.74 (+9.64)
Masked Diffusion Models											
Dream	17.48	27.13	47.81	53.37	0.53	13.36	16.96	24.30	27.54	33.22	26.16
Dream w/ RAG	38.66	53.22	76.39	82.32	0.90	19.25	24.23	34.17	29.46	35.42	39.39 (+13.23)
Llada	11.77	21.47	37.11	42.88	0.70	11.48	15.18	22.55	28.92	33.26	22.52
Llada w/ RAG	**55.30**	**59.78**	**84.90**	**85.89**	**2.33**	**30.52**	**35.79**	**39.40**	**43.02**	**45.09**	**48.20** (+25.68)

Table 1: Overall Performance of ARMs and MDMs. The **best** results are highlighted.



(a) Dream. (b) Llada.

Figure 5: Retrieval Head Analysis in MDMs. High-scoring retrieval-specialized heads are concentrated in the middle and deeper layers, supporting the transition from shallow structural anchoring to deeper content-focused retrieval.

Method	NQ	TQA	Macro QA	HotpotQA	T-REx
Dream	38.66	76.39	0.90	24.23	29.46
w/o Random Heads	37.59	73.74	0.78	23.04	27.58
w/o Retrieval Heads	**18.01**	**37.64**	**0.34**	**10.32**	**12.00**

Table 2: Performance of Masking Random Heads vs. Retrieval Heads in Dream-7B-Instruct.

Method	NQ	TQA	Macro QA	HotpotQA	T-REx
Llada	55.30	84.90	2.33	35.79	43.02
w/o Random Heads	54.88	84.38	2.17	35.69	42.34
w/o Retrieval Heads	**43.53**	**74.36**	**1.77**	**27.75**	**41.26**

Table 3: Performance of Masking Random Heads vs. Retrieval Heads in Llada-8B-Instruct.

tion. A higher score indicates that the head effectively routes information from relevant context tokens into the output, whereas a lower score suggests that the head rarely contributes to context-following behavior. The resulting retrieval-score heatmap (Figure 5) reveals that, in MDMs, high-scoring heads are predominantly concentrated in the middle and deeper layers. This pattern suggests that, with increasing depth, MDMs progressively allocate greater attention capacity to content-sensitive retrieval heads that track context-bearing tokens. Such a depth-wise transition from structurally oriented floating behavior to content-centric retrieval behavior is exactly in line with the *Shallow Structure-Aware, Deep Content-Focused* attention mechanism hypothesis.

Beyond this correlational pattern, we further conduct a causal head-ablation study. Specifically, we mask the top-50 retrieval heads identified by the retrieval score, and compare the result against masking the same number of random heads. As summarized in Table 2 and Table 3, masking retrieval heads causes a substantially larger performance drop than masking random heads in both Dream and Llada. This suggests that retrieval-specialized heads are not merely correlated with context-following behavior, but make non-trivial functional contributions to model performance.

5 Attention Floating Improves Robustness in Learning from Context

In this section, we transition from analyzing internal mechanisms to empirically evaluating the model's capability to learn from context. We begin in Section 5.1 by evaluating overall performance on a range of knowledge-intensive tasks, where we find that MDMs benefit more substantially from retrieved context than ARMs. Then, to better understand the sources of this performance disparity, Section 5.2 shows that *attention floating* contributes to robust learning from context through a set of stress tests. Finally, Section 5.3 investigates the information aggregation process through region-level attention flow, offering a mechanistic explanation for the empirically observed gains.

5.1 Performance of ARMs and MDMs in Learning Knowledge from Contexts

This section examines the overall performance of autoregressive models (ARMs) and masked diffusion models (MDMs) across knowledge-intensive tasks, with and without retrieved context. For the controlled comparison, we use two representative ARM backbones, Qwen2.5-7B-Instruct and Llama3-8B-Instruct, and two representative MDM backbones, Dream-7B-Instruct and Llada-8B-Instruct. These models are chosen to be broadly comparable in scale and efficiency, so that the observed differences are less likely to be explained by model size alone. Detailed dataset statistics and model scaling effectiveness are provided in Appendix A.2.

As shown in Table 1, autoregressive baselines achieve slightly higher average scores than MDMs in the close-book QA setting, which is consistent with their stronger parametric capacity. However, once we incorporate query-retrieved passages as contextual input, MDMs not only close the performance gap with ARMs, but also outperform all ARM w/ RAG baselines across all QA scenarios. MDMs w/ RAG achieve over 19.5% average improvement compared to their corresponding baseline models, which is more than twice the gain observed for ARMs when augmented with retrieval (ARMs w/ RAG obtain 8.5% improvements). These results indicate that, although MDMs start from a slightly weaker parametric baseline, they are substantially more effective at transforming retrieved evidence into end-task performance gains. In the following experiments, we further investigate how *attention floating* enables a more retrieval-sensitive and context-driven utilization of external knowledge.

5.2 Effectiveness of Attention Floating with Contextual Stress Testing

To further examine how the *attention floating* supports robustness under context variations, we conduct a systematic evaluation along three key dimensions: (i) contextual noise interference, (ii) position perturbation, and (iii) evidence integration.

Contextual Noise Interference. To evaluate model robustness under different signal-to-noise conditions, we adopt an experimental setup following prior work (Hsieh et al., 2024): we fix the context to contain exactly one gold document, gradually increase the number of unrelated distractor documents, and observe how model performance changes. As shown in Figure 6, as the number of distractor documents increases, the performance of the ARMs exhibits a gradual degradation, which may be attributed to the autoregressive generation architecture. In contrast, the MDM demonstrates substantially stronger noise resilience: even when the number of distractor documents is significantly increased, its accuracy degrades more mildly. This further indicates that MDMs are better at alleviating the impact of irrelevant documents, supporting the view that the bidirectional attention mechanism of MDMs provides a form of global denoising.

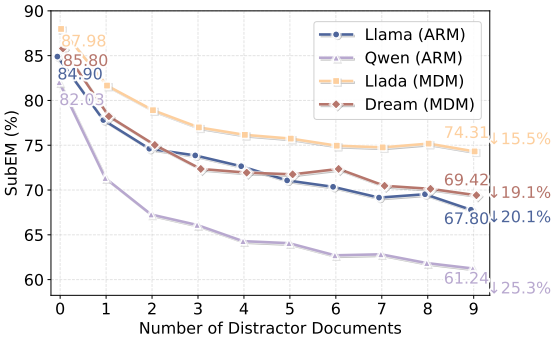
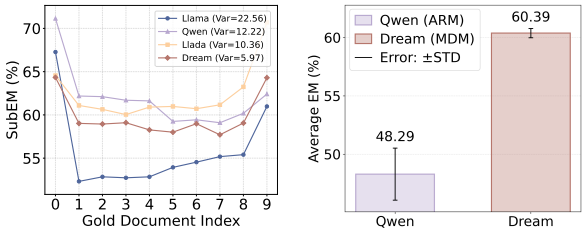


Figure 6: Performance with Varying Numbers of Distractor Documents. As the number of distractor documents grows, ARMs degrade more steadily, whereas MDMs remain more stable, indicating stronger robustness to irrelevant context.

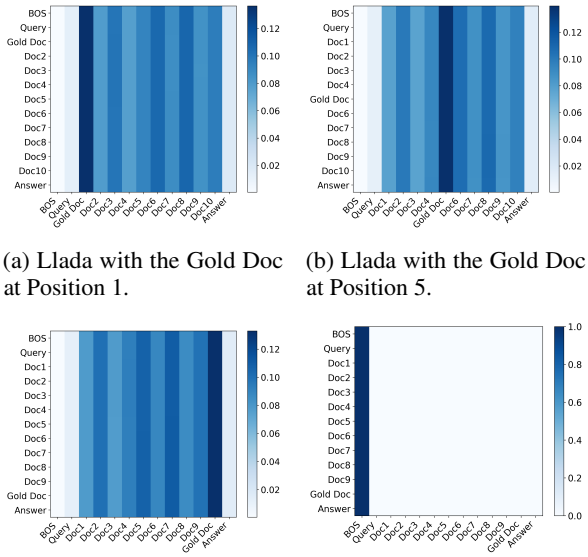
Position Perturbation. Beyond the impact of contextual noise, the position of key information within a long context also affects model performance (Liu et al., 2024). Specifically, we evaluate a multi-document QA task by systematically varying the position of the gold document among distractors to examine the model's sensitivity to evidence location. As illustrated in Figure 7a, ARMs exhibit a characteristic U-shaped performance curve, with accuracy peaking when the gold evidence is located near the boundaries and deteriorating when it appears in the middle. In contrast, MDMs show substantially smaller performance variance across different positions, suggesting that they are less sensitive to the location of key information. This positional robustness can be attributed to the *attention floating* mechanism: Unlike ARMs, which exhibit rigid sinking around <BOS> and a strong recency bias, MDMs can actively reorganize their attention distribution.

Evidence Integration. We further investigate whether the architecture of MDMs leads to more ro-



(a) Position Shift. (b) Evidence Integration.

Figure 7: Performance under (a) Position Perturbation and (b) Evidence Integration Scenarios. (a) ARMs exhibit a U-shaped sensitivity to the position of the gold document, while MDMs maintain lower variance across positions. (b) MDMs are also more stable under different evidence distributions, suggesting stronger integration of scattered supporting evidence.



(a) Llada with the Gold Doc at Position 1. (b) Llada with the Gold Doc at Position 5.

(c) Llada with the Gold Doc at Position 10. (d) Llama with Gold Doc at Positions 1, 5, 10.

Figure 8: Region-Level Attention Flow of Llada (a–c) and Llama (d) with Gold Doc at Different Positions.

bust behavior in complex scenarios, particularly in multi-hop reasoning tasks where the reasoning outcome requires the integration of information from scattered evidence. Specifically, while keeping the content unchanged, we systematically perturb the distribution of these evidence documents within the input context. As shown in Figure 7b, the ARM exhibits pronounced sensitivity to different evidence distribution, as evidenced by achieving a higher variance score than MDMs. In contrast, the MDM achieves superior average performance while maintaining remarkable stability. The minimal variance indicates that the MDM is largely insensitive to the distribution of gold documents, demonstrating its robust capability to effectively integrate evidence from retrieved documents.

5.3 Attention Floating Analysis via Region-Level Attention Flow

To better understand the underlying mechanisms that lead to the superior performance of MDMs across different scenarios, we analyze their internal information flow through the lens of attention mechanisms. This perspective enables a fine-grained examination of how models dynamically allocate attention across different input regions.

To characterize the information flow across all layers, we adopt an attention-flow-style region-level influence matrix (Abnar and Zuidema, 2020). Concretely, we aggregate head-averaged attention across layers into a position-level influence matrix, and subsequently group contiguous positions into coarse-grained regions corresponding to <BOS>, Query, Doc1–Doc10, and Answer. The formal definition of the attention flow procedure is provided in Appendix A.3. As illustrated in Figure 8 and Figure 9, we observe a clear qualitative distinction between ARMs and MDMs. For ARMs, the dominant peak of the region-level flow remains tightly concentrated around <BOS>, exhibiting minimal sensitivity to changes in evidence location. In contrast, for MDMs, relocating the gold document results in a corresponding shift of the high-intensity band in the region-level flow toward the ground-truth document segments. These results indicate that the dominant attention flow in ARMs behaves as a rigid sink in <BOS> and is insensitive to the position of the gold document. In comparison, MDMs actively reorganize attention distributions to capture semantically relevant information, thus mitigating the influence of distracting documents.

6 Related Work

Autoregressive models (ARMs) remain the dominant paradigm but suffer from high latency and positional bias due to their sequential, causal mechanism. To alleviate these limitations, a growing body of studies (Hersche et al., 2025; Liu et al., 2025a; Ni et al., 2025; Shao et al., 2025) on diffusion language models (DLMs) have explored replacing sequential autoregressive sampling with parallel generation via denoising, aiming to retain strong generation quality while improving parallelism and efficiency. Within the landscape of DLMs, recent research has increasingly converged on masked diffusion models (MDMs) (Nie et al., 2025; Ye et al.,

2025), which operate natively over the discrete token vocabulary. Unlike approaches that require mapping to continuous latent spaces, MDMs typically initialize from sequences filled with special mask symbols. They define a forward process that injects masking noise and a reverse process that predicts tokens at masked positions in parallel to reconstruct the text. While early MDMs (He et al., 2023; Zheng et al., 2023) were relatively small in parameter and training scale, leaving a noticeable performance gap compared to ARMs, recent scaled-up MDMs (Wu et al., 2025; Liu et al., 2025b) have successfully narrowed this gap on general language modeling and downstream benchmarks, demonstrating their potential as a competitive alternative.

In recent years, research on the mechanistic understanding and interpretability of large models has progressed rapidly. In ARMs, prior work has developed circuit-style analysis frameworks to elucidate decomposable attention heads and compositional mechanisms, and has demonstrated that induction heads can mechanistically account for in-context learning (Elhage et al., 2021; Olsson et al., 2022). Studies on long-context behavior have further established clear empirical baselines for positional sensitivity (Liu et al., 2024). In parallel, the phenomenon of attention disproportionately concentrating on early sequence positions, known as the attention sink, has been systematically analysed (Xiao et al., 2024; Gu et al., 2024). Similar mechanistic analyses have been extended to DLMs, including investigations that dissect internal features and sources of bias, as well as work that traces the evolution of interpretable concepts along the denoising trajectory (Shi et al., 2025; Tinaz et al., 2025). More recently, attention-sink analyses have also been applied to DLMs, examining how sink behavior interacts with the denoising process (Rulli et al., 2025). Collectively, these lines of work provide both the methodological foundation and comparative reference points for our attention-mechanism analysis in MDMs.

7 Conclusion

This paper investigates MDMs from the perspective of their attention behavior. We first identify and formalize a phenomenon termed *attention floating*, demonstrating that MDMs exhibit weaker, more mobile, and structurally less concentrated attention patterns. Our analyses reveal a Shallow Structure–Aware, Deep Content–Focused attention mechanism. We then empirically evaluate the MDM's ability to learn from context and conduct a series of experiments to show the crucial role of the attention floating mechanism. These findings deepen the understanding of the internal working mechanisms of MDMs.

8 Limitations

This work has limitations that also point to promising future directions. First, our analysis mainly focuses on one category of Diffusion Language Models (DLLMs)–Masked Diffusion Models (MDMs). While this focus allows for a controlled and in-depth investigation of attention floating under masked denoising dynamics, we do not explicitly evaluate other DLLM variants with different diffusion or denoising mechanisms. Future work could extend this analysis framework to a broader family of DLLMs and systematically compare how different architectures and denoising strategies affect floating patterns and retrieval gains. Second, our contributions are primarily centered on mechanism analysis, and we have not yet proposed training or inference methods that can directly improve performance based on these findings; future work may explore turning floating signals into learnable regularizers or developing denoising-style context modeling for multi-document evidence integration, thereby closing the loop from mechanism to method.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62576082). This work is also supported by the AI9Stars community.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of ACL*, pages 4190–4197.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, and 1 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, (1):12.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. 2024. When attention sink emerges in language models: An empirical view. *ArXiv preprint*.

Zhengfu He, Tianxiang Sun, Qiong Tang, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. 2023. DiffusionBERT: Improving generative masked language models with diffusion models. In *Proceedings of ACL*, pages 4521–4534.

Michael Hersche, Samuel Moor-Smith, Thomas Hofmann, and Abbas Rahimi. 2025. Soft-masked diffusion language models. *ArXiv preprint*.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? *ArXiv preprint*.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. 2022. Diffusion-lm improves controllable text generation. In *Proceedings of NeurIPS*.

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

Guangyi Liu, Zeyu Feng, Yuan Gao, Zichao Yang, Xiaodan Liang, Junwei Bao, Xiaodong He, Shuguang Cui, Zhen Li, and Zhiting Hu. 2023. Composable text controls in latent space with ODEs. In *Proceedings of EMNLP*, pages 16543–16570.

Jingyu Liu, Xin Dong, Zhifan Ye, Rishabh Mehta, Yonggan Fu, Vartika Singh, Jan Kautz, Ce Zhang, and Pavlo Molchanov. 2025a. Tidar: Think in diffusion, talk in autoregression. *ArXiv preprint*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, pages 157–173.

Xiaoran Liu, Yuerong Song, Zhigeng Liu, Zengfeng Huang, Qipeng Guo, Ziwei He, and Xipeng Qiu. 2025b. Longllada: Unlocking long context capabilities in diffusion llms. *ArXiv preprint*.

Jinjie Ni, Qian Liu, Chao Du, Longxu Dou, Hang Yan, Zili Wang, Tianyu Pang, and Michael Qizhe Shieh. 2025. Training optimal large diffusion language models. *ArXiv preprint*.

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*.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, and 1 others. 2022. In-context learning and induction heads. *ArXiv preprint*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2024. Qwen2.5 technical report. *ArXiv preprint*.

Maximo Eduardo Rulli, Simone Petruzzi, Edoardo Michielon, Fabrizio Silvestri, Simone Scardapane, and Alessio Devoto. 2025. Attention sinks in diffusion language models. *ArXiv preprint*.

Alexander Shabalin, Viacheslav Meshchaninov, Egor Chimbulatov, Vladislav Lapikov, Roman Kim, Grigory Bartosh, Dmitry Molchanov, Sergey Markov, and Dmitry Vetrov. 2025a. Tencdm: Understanding the properties of the diffusion model in the space of language model encodings. In *Proceedings of AAAI*, 23, pages 25110–25118.

Alexander Shabalin, Viacheslav Meshchaninov, and Dmitry Vetrov. 2025b. Smoothie: Smoothing diffusion on token embeddings for text generation. *ArXiv preprint*.

Chenyang Shao, Sijian Ren, Fengli Xu, and Yong Li. 2025. Diffuse thinking: Exploring diffusion language models as efficient thought proposers for reasoning. *ArXiv preprint*.

Yingdong Shi, Changming Li, Yifan Wang, Yongxiang Zhao, Anqi Pang, Sibei Yang, Jingyi Yu, and Kan Ren. 2025. Dissecting and mitigating diffusion bias via mechanistic interpretability. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8192–8202.

Jaesung Tae, Hamish Ivison, Sachin Kumar, and Arman Cohan. 2025. Tess 2: A large-scale generalist diffusion language model. *ArXiv preprint*.

Berk Tinaz, Zalan Fabian, and Mahdi Soltanolkotabi. 2025. Emergence and evolution of interpretable concepts in diffusion models. *ArXiv preprint*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *ArXiv preprint*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of NeurIPS*.

Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, Ligeng Zhu, Ping Luo, Song Han, and

Enze Xie. 2025. Fast-dllm: Training-free acceleration of diffusion llm by enabling kv cache and parallel decoding. *ArXiv preprint*.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2024. Retrieval head mechanistically explains long-context factuality. *ArXiv preprint*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *Proceedings of ICLR*.

Jiayu Yao, Shenghua Liu, Yiwei Wang, Lingrui Mei, Baolong Bi, Yuyao Ge, Zhecheng Li, and Xueqi Cheng. 2025. Who is in the spotlight: The hidden bias undermining multimodal retrieval-augmented generation. *ArXiv preprint*.

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*.

Lin Zheng, Jianbo Yuan, Lei Yu, and Lingpeng Kong. 2023. A reparameterized discrete diffusion model for text generation. *ArXiv preprint*.

A Appendix

A.1 License

This section summarizes the licenses (or usage terms) of the datasets used in our experiments.

All datasets are used under their respective licenses and agreements, which permit academic research use: Natural Questions (CC BY-SA 3.0 License); TriviaQA and 2WikiMQA (Apache 2.0 License); HotpotQA (CC BY-SA 4.0 License); T-REx (CC BY-SA 4.0 License); GSM8K (MIT License), and MS MARCO QA, which is provided under the MS MARCO Terms and Conditions for non-commercial research purposes.

A.2 Additional Experimental Details

Dataset Statistics for Knowledge-Intensive Tasks. We evaluate all models on a suite of retrieval-augmented knowledge-intensive benchmarks that follow the evaluation configuration of Li et al. (2025). The datasets span open-domain QA, multi-hop reasoning, and slot-filling style questions. For each dataset, we use the same retrieval corpus and context construction strategy as in Li et al. (2025), and report results on evaluation-only splits. Table 6 summarizes the basic statistics of the datasets used in our experiments.

Robustness across ARM Scales. To further examine whether the weaker RAG gains of ARMs could be explained by the relatively small ARM backbones used in the main comparison, we conduct an additional scaling study on larger ARM models under the same setup as Table 4. As shown in the table, larger ARMs achieve stronger absolute performance, but the gains brought by RAG do not increase monotonically with model size and become very limited at larger scales.

Additional Visualizations of Positional Drift. We include additional visualizations of positional drift in attention floating for Llada and Dream across denoising steps. Figure 10 and Figure 11 provide qualitative support for the layer-dependent drift patterns discussed in the main text.

High-Frequency Floating Token Statistics. Table 7 reports the most frequent floating tokens ranked by their overall proportion. The statistics show that floating tokens are dominated by high-frequency structural symbols (e.g., newline, end-of-text, mask tokens, and punctuation), while lexical content tokens appear much less frequently, supporting our structural-controller interpretation.

Layer-wise Structural-to-Lexical Attention Ratio. To provide a direct quantitative proxy for the proposed Shallow Structure-Aware, Deep Content-Focused hypothesis, we compute, for each layer, the ratio of attention mass allocated to structural tokens over lexical tokens. Structural and lexical token categories follow the definition in Section 3.2. As shown in Table 5, the ratio is highest in shallow layers and generally decreases toward middle and deep layers, indicating that attention is initially more concentrated on structural tokens and gradually shifts toward content-bearing lexical tokens with increasing depth. We also observe a rebound at the final layer, which we attribute to the output-interface behavior of the last layer.

Layer-wise QK Decomposition Heatmaps. We provide visualizations of the layer-wise QK decomposition across all layers, including the QK scores, angular and norm products in Figure 12–Figure 17, confirming that the proposed Shallow Structure-Aware, Deep Content-Focused attention mechanism is not an artifact of cropping or local windows.

A.3 Attention Flow for Multi-Document RAG

This section provides the formal definition of the attention flow procedure following (Abnar and Zuidema, 2020) used in our analysis.

Token-Level Influence Matrix. To account for residual connections in Transformers, we first augment the attention weights with a residual term:

$$\bar{A}^{\ell}_{i\to j} = \alpha A^{\ell}_{i\to j} + (1-\alpha)\delta_{i\to j}, \tag{6}$$

where α is a hyperparameter controlling the balance between attention and residual connections, and $\delta_{i\to j}$ is the Kronecker delta. Following Abnar and Zuidema (2020), we set $\alpha = 0.5$. Since adding the residual term changes the row sums, we re-normalize using the average attention received by each position:

$$\tilde{A}^{\ell}_{i\to j} = \frac{\bar{A}^{\ell}_{i\to j}}{\sum_{i=1}^{n} \bar{A}^{\ell}_{i\to j}}, \tag{7}$$

where n denotes the length of sequence. We then accumulate attention across layers via matrix multiplication. Let $\tilde{\mathbf{A}}^{\ell} \in \mathbb{R}^{n\times n}$ denote adjusted attention matrix at layer ℓ. The token-level influence

Method	NQ	TQA	Marco QA	HotpotQA	T-REx	Avg.
Qwen2.5-14B-Instruct	32.86	70.09	2.67	26.25	44.60	35.29
Qwen2.5-14B-Instruct w/ RAG	48.06	84.11	4.33	31.43	42.20	42.03 (+6.74)
Qwen2.5-32B-Instruct	38.16	73.08	3.00	28.75	49.40	38.48
Qwen2.5-32B-Instruct w/ RAG	49.82	85.23	5.00	31.07	44.00	43.03 (+4.55)
Qwen2.5-72B-Instruct	45.94	80.56	3.67	32.32	52.40	42.98
Qwen2.5-72B-Instruct w/ RAG	49.47	85.79	5.00	32.14	42.60	43.00 (+0.02)

Table 4: Scaling Behavior of RAG-Induced Improvements in ARMs.

Layer	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Structural/Lexical Ratio	0.91	0.82	0.74	0.68	0.69	0.70	0.56	0.44	0.41	0.43	0.26	0.46	0.39	0.30	0.37	0.35
Layer	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Structural/Lexical Ratio	0.24	0.34	0.32	0.34	0.39	0.31	0.30	0.37	0.42	0.27	0.28	0.22	0.24	0.22	0.24	0.67

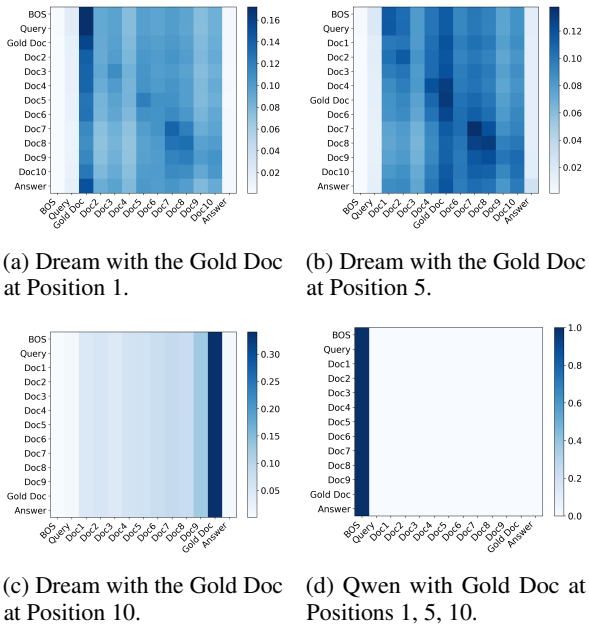
Table 5: Layer-wise Structural-to-Lexical Attention Mass Ratio.

Task	Dataset	Total
Open-domain QA	Natural Questions (2019)	2,837
	TriviaQA (2017)	5,359
	MARCO QA (2016)	3,000
Multi-hop QA	HotpotQA (2018)	5,600
Slot Filling	T-REx (2018)	5,000

Table 6: Dataset statistics for Knowledge-Intensive Tasks.

Rank	Token	Prop.	Type
1	\n	61.09%	Structural
2	<|endoftext|>	28.70%	Structural
3	␣	3.34%	Structural
4	<|mdm_mask|>	2.13%	Structural
5	,	1.23%	Structural
6	.	0.87%	Structural
7	)	0.53%	Structural
8	?	0.38%	Structural

Table 7: High-Frequency Floating Token Statistics.



(a) Dream with the Gold Doc at Position 1.

(b) Dream with the Gold Doc at Position 5.

(c) Dream with the Gold Doc at Position 10.

(d) Qwen with Gold Doc at Positions 1, 5, 10.

Figure 9: Region-Level Attention Flow for Dream (a–c) and Qwen (d) under Different Evidence Positions.

matrix is:

$$\mathbf{R} = \prod_{\ell=1}^{L} \tilde{\mathbf{A}}^{\ell}, \tag{8}$$

where $R_{i \to j}$ represents the cumulative flow of information from position i to j across all layers.

Region-Level Influence Matrix. In the multi-document RAG setting, we partition the sequence into regions corresponding to <BOS>, Query, Doc1–Doc10, and Answer. Let $\mathcal{I}_p$ be the set of token indices in region p and $\mathcal{I}_q$ the set for region q. We aggregate token-level influences into a region-level matrix $R^{\text{region}} \in \mathbb{R}^{P \times P}$ via:

$$R^{\text{region}}_{p \to q} = \frac{1}{|\mathcal{I}_p|\,|\mathcal{I}_q|} \sum_{i \in \mathcal{I}_p} \sum_{j \in \mathcal{I}_q} R_{i \to j}. \tag{9}$$

where P is the number of regions. For visualization, we row-normalize R^{region} so that each row represents the relative distribution of outgoing influence from a source region to all target regions.

To verify that the observed positional behavior is consistent across model families, we additionally visualize the region-level attention flow for Dream and Qwen under different gold-document positions. As shown in Figure 9, Dream exhibits a clear shift of high-intensity flow toward the relocated evidence regions, while Qwen remains rigidly sunk in <BOS> region.

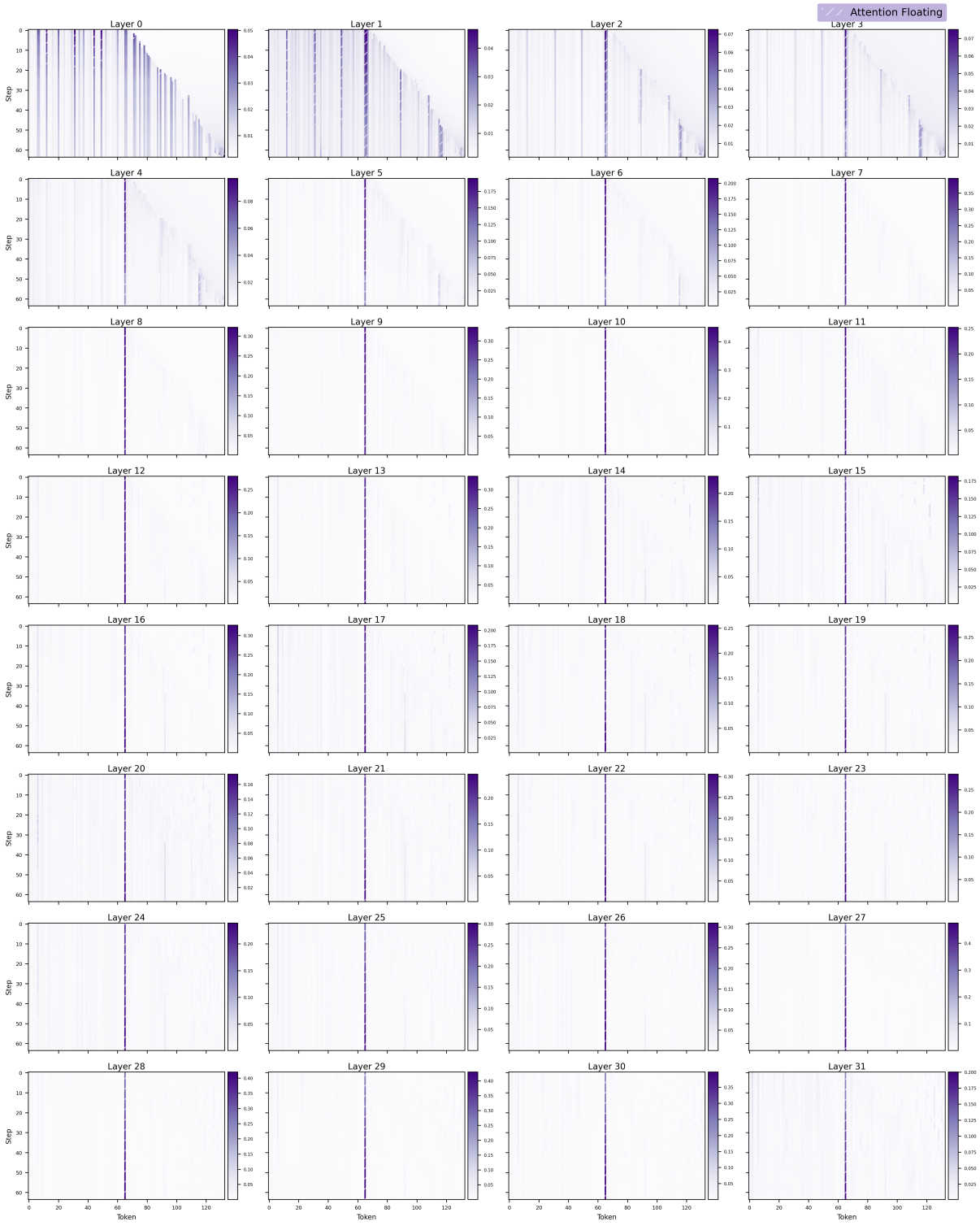


Figure 10: Positional Drift of Attention Floating in Llada.

Figure 11: Positional Drift of Attention Floating in Dream.

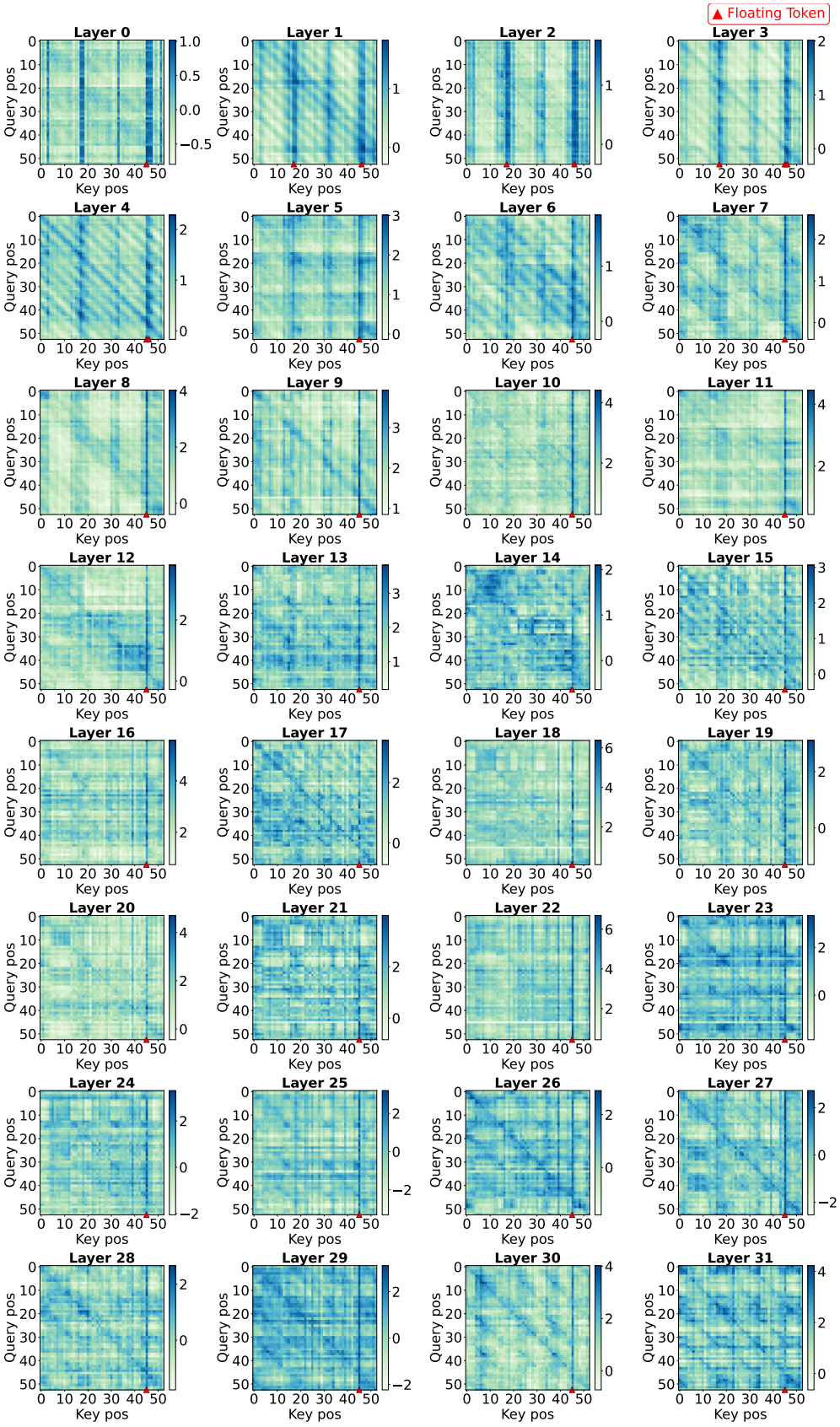


Figure 12: QK Score in Llada.

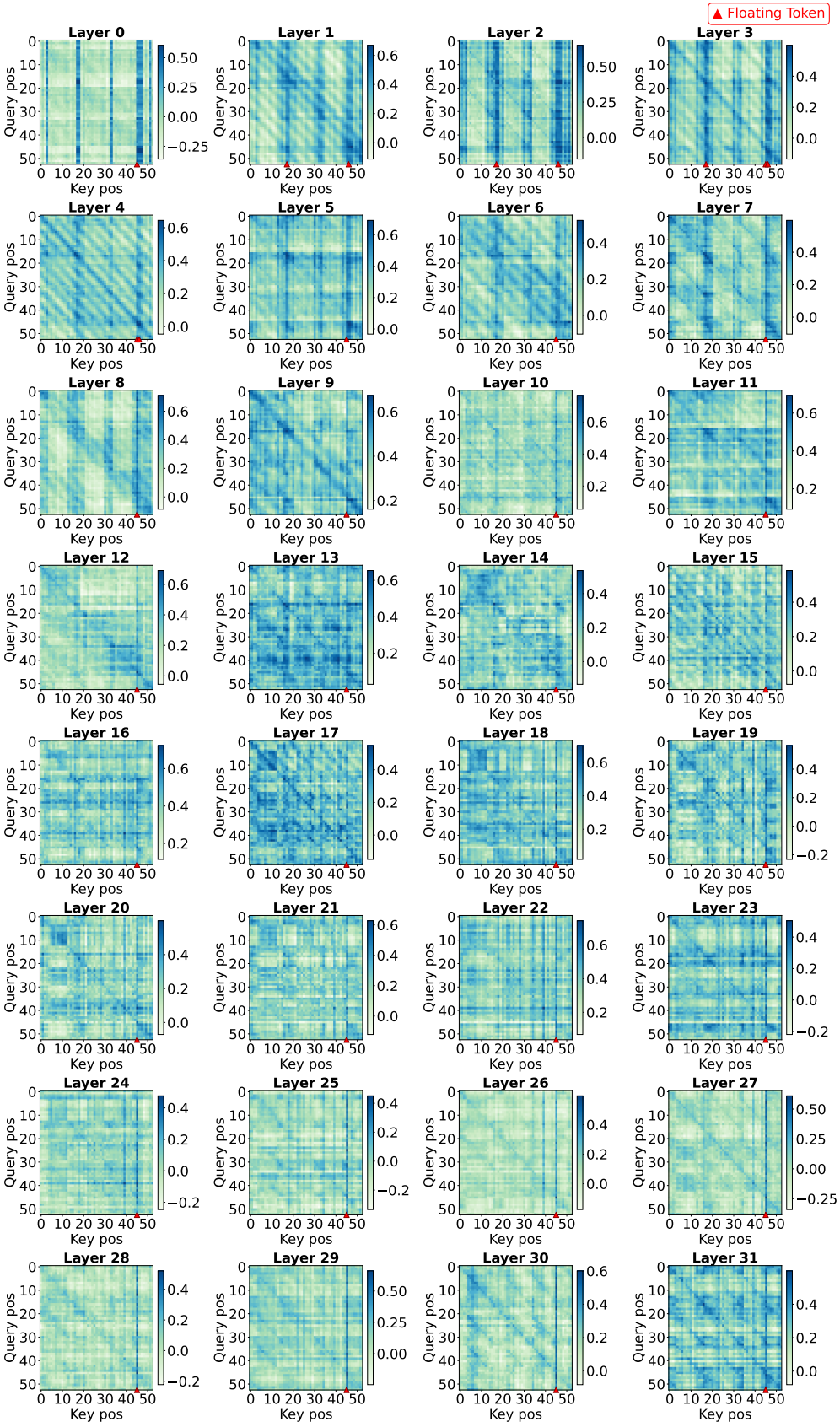


Figure 13: Angular in Llada.

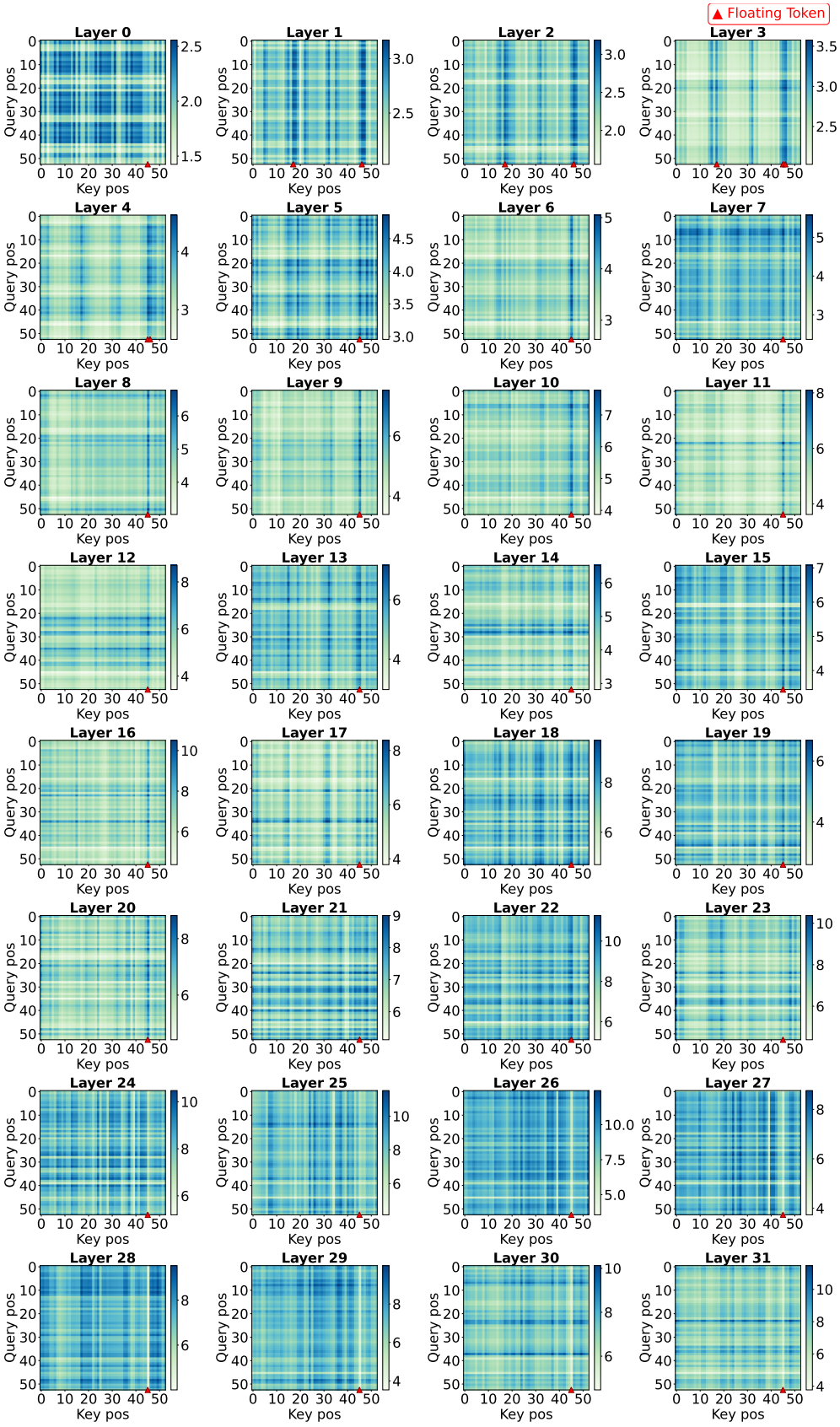


Figure 14: Norm Product in Llada.

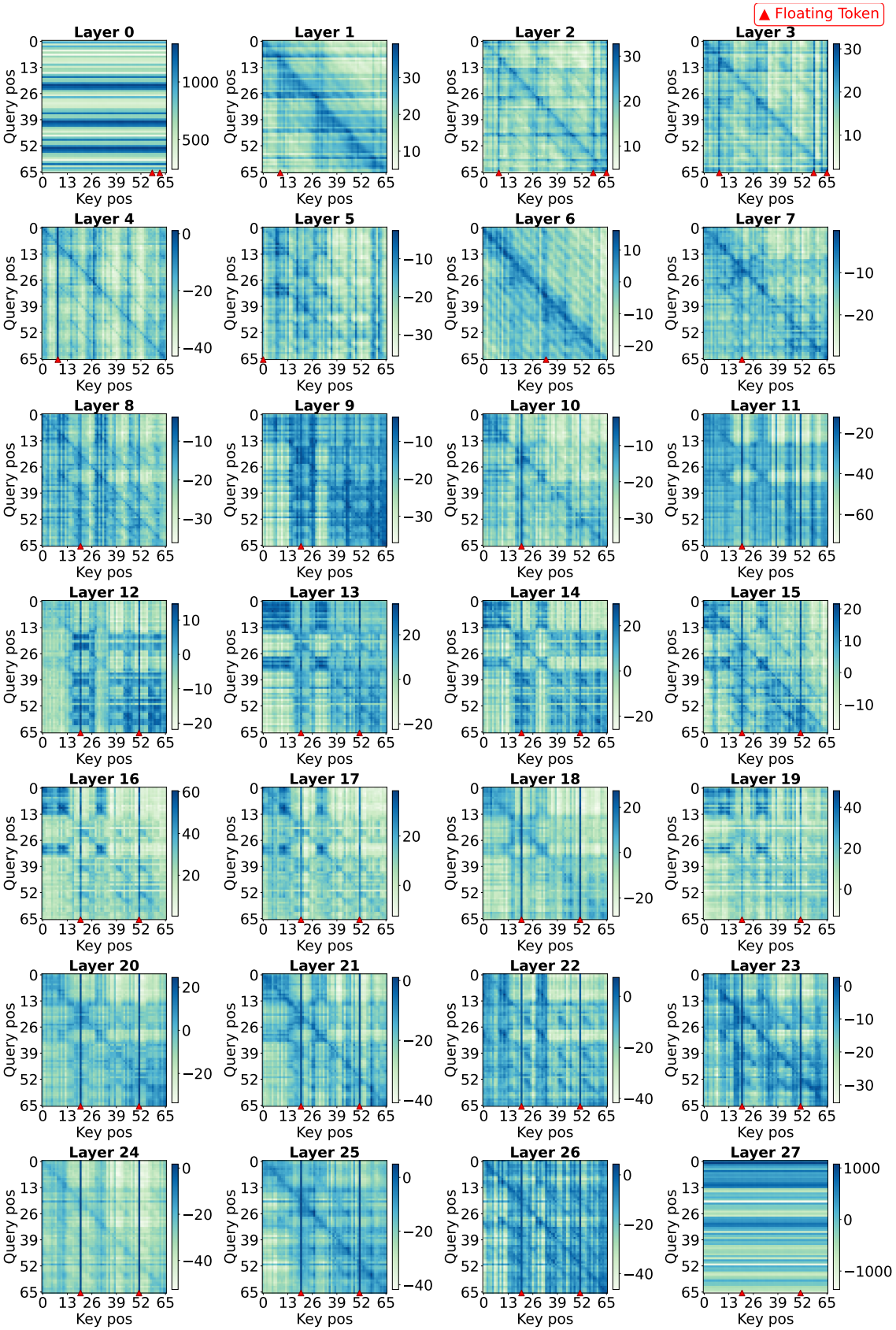


Figure 15: QK Score in Dream.

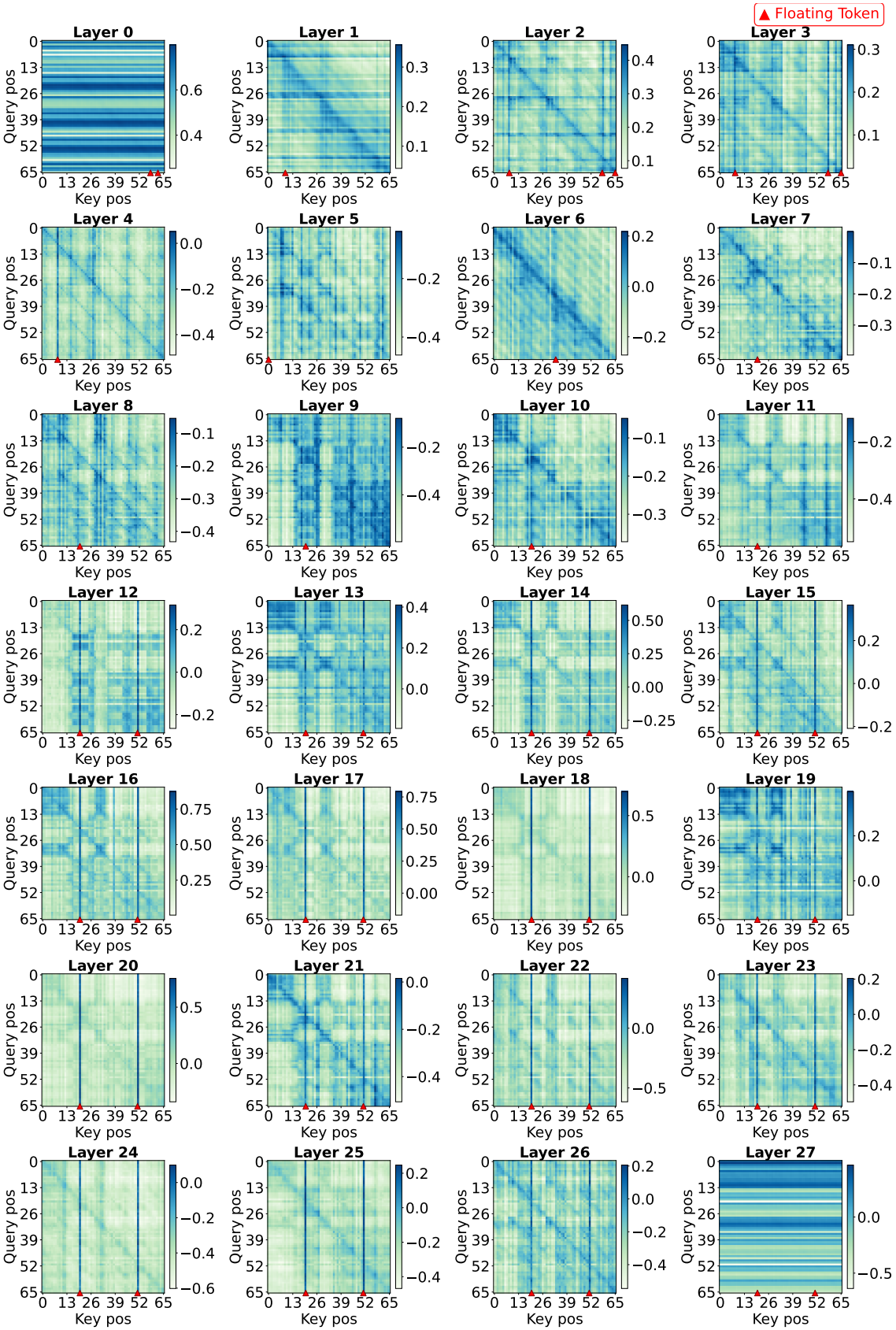


Figure 16: Angular in Dream.

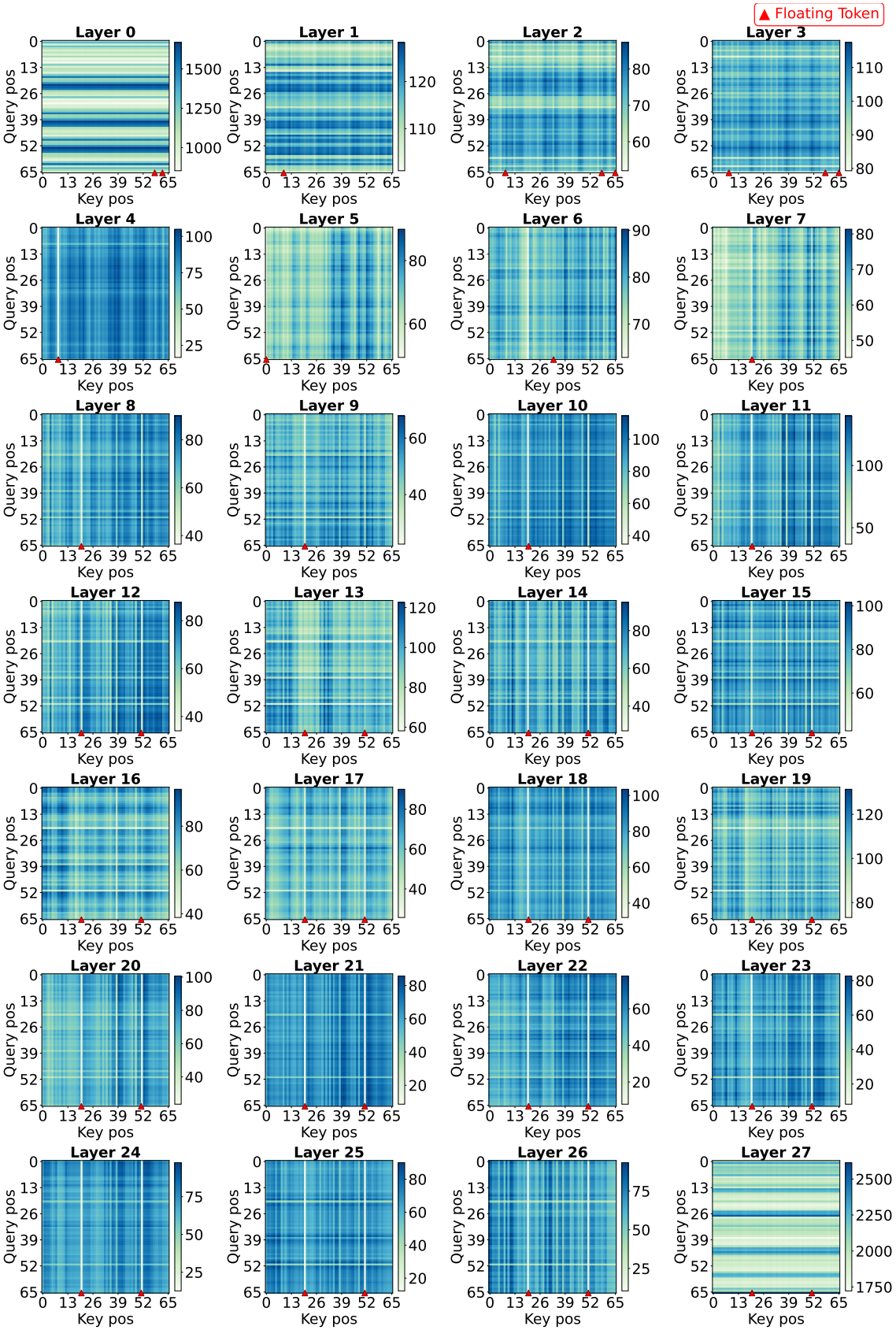


Figure 17: Norm Product in Dream.