Who is in the Spotlight: The Hidden Bias Undermining Multimodal Retrieval-Augmented Generation

Jiayu Yao[1*] **Shenghua Liu**[1* †] **Yiwei Wang**[2] **Lingrui Mei**[1*]
Baolong Bi[1*] **Yuyao Ge**[1*] **Zhecheng Li**[3] **Xueqi Cheng**[1]
[1]Institute of Computing Technology, Chinese Academy of Sciences
[2]University of California, Merced
[3]University of California, San Diego
yaojiayu25@mails.ucas.ac.cn, {liushenghua, meilingrui22, geyuyao24z}@ict.ac.cn

Abstract

Multimodal Retrieval-Augmented Generation (RAG) systems have become essential in knowledge-intensive and open-domain tasks. As retrieval complexity increases, ensuring the robustness of these systems is critical. However, current RAG models are highly sensitive to the order in which evidence is presented, often resulting in unstable performance and biased reasoning, particularly as the number of retrieved items or modality diversity grows. This raises a central question: *How does the position of retrieved evidence affect multimodal RAG performance?* To answer this, we present the first comprehensive study of position bias in multimodal RAG systems. Through controlled experiments across text-only, image-only, and mixed-modality tasks, we observe a consistent U-shaped accuracy curve with respect to evidence position. To quantify this bias, we introduce the Position Sensitivity Index (PSI_p) and develop a visualization framework to trace attention allocation patterns across decoder layers. Our results reveal that multimodal interactions intensify position bias compared to unimodal settings, and that this bias increases logarithmically with retrieval range. These findings offer both theoretical and empirical foundations for position-aware analysis in RAG, highlighting the need for evidence reordering or debiasing strategies to build more reliable and equitable generation systems. Our code and experimental resources are available at https://github.com/Theodyy/Multimodal-Rag-Position-Bias.

1 Introduction

The growing demand for multimodal interaction has driven the development of multimodal Retrieval-Augmented Generation (RAG) systems, which integrate heterogeneous data sources (text, images, audio) to achieve comprehensive information understanding (Abootorabi et al., 2025). This technological advancement has enabled breakthroughs across diverse domains: academic research leverages frameworks like Taichu-mRAG, OmniSearch, VARAG (Faysse et al., 2024) and GraphRAG (Edge et al., 2025) for knowledge discovery, while industrial applications employ DocPrompting (Zhou et al., 2023), UniFashion (Zhao et al., 2024), RAG-Driver (Yuan et al., 2024) and Img2Loc (Zhou et al., 2024b) for real-world problem solving. In professional fields, systems like RULE (Xia et al., 2024b) and MMed (Xia et al., 2024a) enhance medical diagnostics. In other scenarios, SoccerRAG (Strand et al., 2024) and MMRA (Wu et al., 2024) research demonstrates the social computing and entertainment applications. These successes highlight multimodal RAG's potential in open-domain question answering and knowledge-intensive scenarios.

However, the reliability of multimodal RAG systems faces critical challenges as applications expand. Current systems exhibit vulnerability when faced with complex retrieval problems: excessive or insufficient retrieved content often induces hallucination. Ensuring that generated outputs faithfully adhere to the provided context is crucial for overall system robustness (Bi et al., 2024), yet even with optimal corpus selection, the positional arrangement of retrieved results significantly impacts answer reliability when fed to generation models. This instability aligns with emerging research on systematic position bias in contemporary Large Language Models (LLMs) and Vision-Language Models (VLMs) (Tan et al., 2024; Zhang et al., 2024). These models disproportionately focus on the start and end positions of input sequences while neglecting middle content, a phenomenon we term "middle-loss". However, the existing re-

*Authors from affiliation [1] are also affiliated with: Key Laboratory of Network Data Science and Technology, ICT, CAS; State Key Laboratory of AI Safety; University of Chinese Academy of Sciences.
†Corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 15183–15193
November 4-9, 2025

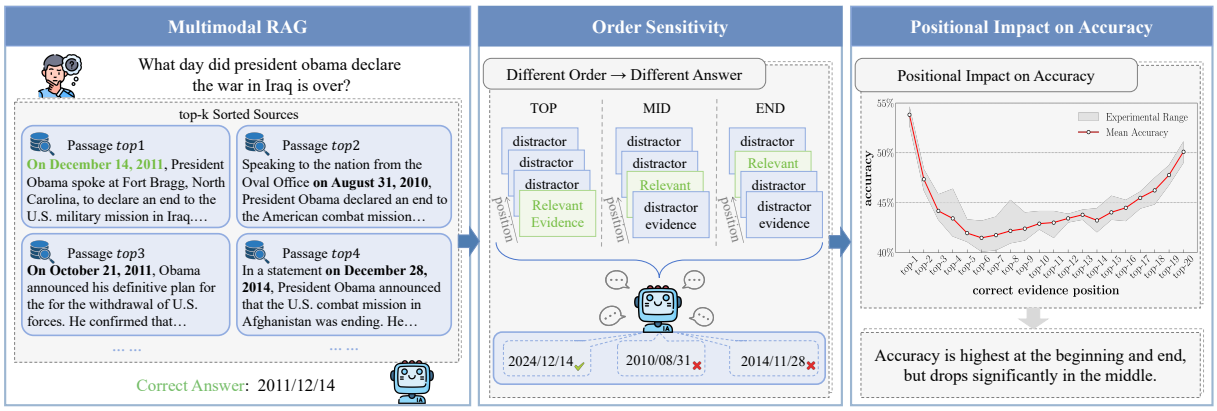

Figure 1: Illustration of position bias in multimodal RAG. **Left** An example of RAG for open question answering, where the prompt contains the question and k retrieved documents sorted by relevance. **Mid** Different positions of the search results may lead to different answers. **Right** Accuracy analysis when correct document ranks at position i ($i \in k, k = 20$) - gray blocks display 10 random experiments, red line indicates mean accuracy.

search presents two critical gaps. First, current studies primarily address single-modal RAG systems (e.g. text-based), lacking systematic investigation of multimodal scenarios. Second, conventional evaluation metrics (e.g. NDCG, MRR) fail to quantify positional sensitivity, with no interpretable framework established for bias analysis in multimodal contexts.

To address these limitations, we obtained the following research results through controlled experiments and interpretability analysis:

- We show that the multimodal RAG system has position bias (Figure 1), and in text, image and mixed-modality scenarios, the generative model assigns different levels of attention to the evidence at different positions, forming a U-shaped accuracy curve.
- We propose a position sensitivity index PSI_p to quantify the amplitude of position bias, and show that cross-modal interactions and larger retrieval scales logarithmically intensify this effect, offering guidance for robust system deployment.
- Through systematic attention visualization analysis, we experimentally validate the above conclusions and uncover layer-specific sparsity patterns in cross-modal attention across decoder hierarchies.

2 Position Bias

To better understand how the position of retrieved evidences affect reasoning in multimodal RAG systems, we begin with a set of controlled experiments designed to probe position bias across diverse input modalities. This section presents both the experimental setup and key empirical findings, revealing a consistent pattern of position-induced performance fluctuation. By systematically perturbing the position of gold evidence while holding content constant, we are able to isolate and quantify the impact of sequence order on model behavior across text-only, image-only, and mixed-modality contexts. To simulate real-world open-domain question answering and knowledge-intensive scenarios, we conducted controlled experiments across three modality configurations: text-only, image-only, and image-text mixed. As depicted in Figure 2, our framework emulates practical RAG workflows where systems process multiple evidence documents (text passages, charts, or multimodal pairs) alongside user queries to generate answers. Through systematic dataset perturbation, we preserve semantic content while manipulating gold evidence positions relative to distractors, with 10 randomized experiments per configuration to ensure statistical robustness. Through this design, we isolate positional effects from content biases, enabling precise characterization of context ordering sensitivity.

2.1 Benchmark

Data For text-based QA simulation, we leverage the MS-MARCO passage ranking benchmark (Nguyen et al., 2016), constructing triplets comprising a query, one gold passage containing the answer, and two topically relevant but non-answer distractor passages. Gold passage permutation across three positions (Top/Mid/End) mimics real-world

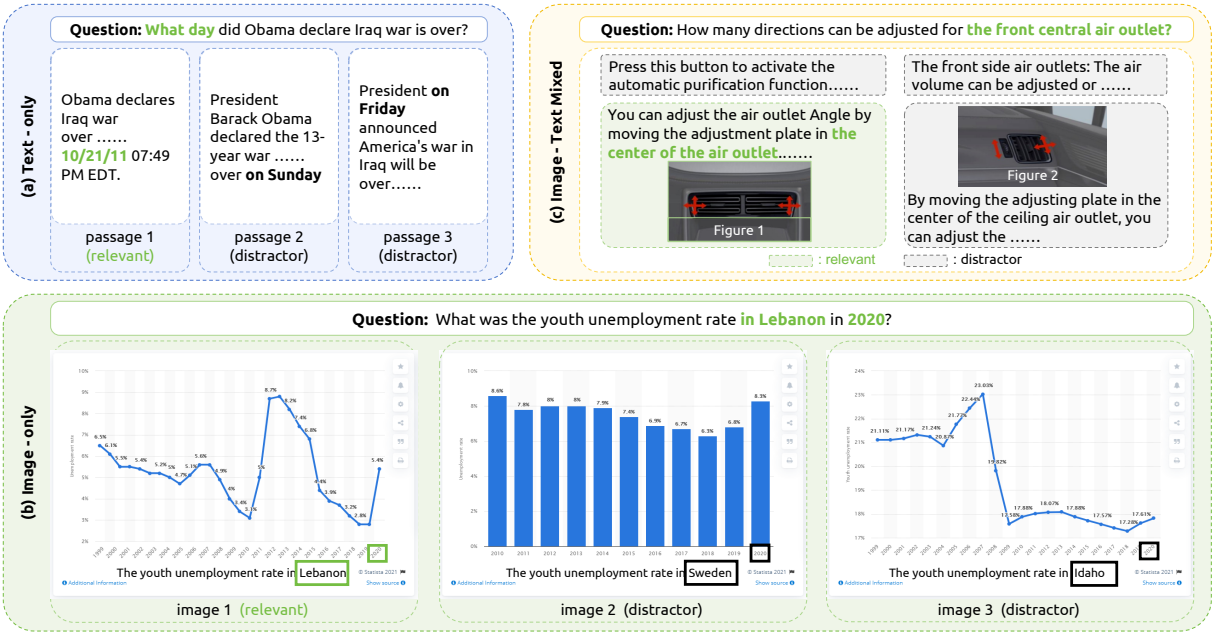

Figure 2: Order sensitivity evaluation across modalities. Relevant evidence is permuted across top, middle, and end positions in (a) Text-only, (b) Image-only, and (c) Image-text mixed settings. Controlled experiments isolate positional effects, showing consistent fluctuations in reasoning accuracy across all modalities.

multi-document retrieval scenarios.

In VQA tasks, we use ChartQA (Masry et al., 2022) to create chart triplets (one answer-containing gold chart and two distractors) with identical positional variations to evaluate pure image processing. For image-text mixed-modality evaluation, we use the VEGA dataset (Zhou et al., 2024a). Each sample in this dataset includes a question, several image-text pairs, and a golden answer based on a specific pair. In our experiment, for each query, we selected the correct image-text pairs and randomly chose two pairs of unrelated images and paragraphs as interference pairs. Then, we set the same position transformation to evaluate whether the multimodal RAG model shows sequence sensitivity when reasoning about interlaced visual and textual evidence.

Models Our evaluation protocol preserves validity through frozen model parameters and deactivated fine-tuning functions, ensuring comparisons reflect inherent architectural biases rather than training artifacts. For the retrieval stage, we implement cross-modal retrieval using the VisRAG-Ret (Yu et al., 2025). In the generation stage, we systematically evaluate three cutting-edge open-source instruction-tuned models (Qwen2-VL-7B-Instruct (Wang et al., 2024a), Llama-3.2-11B-Vision-Instruct (AI@Meta, 2024), MiniCPM-v2.6 (OpenBMB, 2024)) and the closed-source GPT-4o (OpenAI, 2024), which support multi-image input integration and cross-modal attention interactions.

2.2 Empirical Patterns of Positional Bias

Our experiments are conducted on a high-performance computing system with 8×NVIDIA A100 GPUs (80GB memory each), an Intel Xeon Platinum 8336C processor (128 threads @3.5GHz), and 2TB RAM. The software environment runs on Ubuntu 20.04 with CUDA 12.4 for GPU acceleration.

Our experiments reveal the systematically existing position bias patterns in multimodal scenarios through controlled simulation of RAG scenarios. As shown in the Table 1, the average results based on ten randomized experiments indicate: When the key evidence documents are located in the middle of the input sequence, the generative model's ability generally decreases, while the top and end positions maintain a relatively high accuracy rate, forming a typical U-shaped performance curve. This phenomenon stably exists in text (MS-MARCO), image (ChartQA), and mixed-modality (VEGA) scenarios. Among them, GPT-4o shows a significant advantage in the recovery ability at the tail position (the accuracy at the tail of the image task is 11% higher than that in the middle). It can be seen from the results that the performance of MiniCPM-

Modality	Location	MiniCPM-v2.6	Qwen2-VL-7B	Llama-3.2-11B-Vision-Instruct	GPT-4o
Text-only	Top	0.5321(±0.007)	0.4182(±0.009)	0.3199(±0.012)	0.5083(±0.010)
	Mid	0.4963(±0.011)	0.3810(±0.010)	0.2992(±0.009)	0.4513(±0.008)
	End	0.5075(±0.010)	0.3882(±0.008)	0.3095(±0.011)	0.4679(±0.014)
Image-only	Top	0.5911(±0.010)	0.4859(±0.011)	0.2256(±0.009)	0.7333(±0.007)
	Mid	0.5506(±0.009)	0.4790(±0.008)	0.2103(±0.013)	0.7059(±0.012)
	End	0.5593(±0.007)	0.5114(±0.010)	0.2019(±0.010)	0.8125(±0.008)
Image-Text Mixeds	Top	0.5057(±0.009)	0.3891(±0.009)	0.2496(±0.011)	0.7557(±0.010)
	Mid	0.4599(±0.010)	0.3544(±0.008)	0.2272(±0.010)	0.7021(±0.008)
	End	0.4671(±0.013)	0.3923(±0.013)	0.2202(±0.009)	0.7462(±0.009)

Table 1: Accuracy of generative models across different modalities and evidence positions. Results are reported with standard deviations (from ten randomized experiments) shown in parentheses. The consistent U-shaped trend indicates that middle positions yield lower accuracy than top and end positions.

v2.6 and Qwen2-VL-7B-Instruct is relatively weak in multiple tasks, which is mainly attributed to their deficiencies in following instructions.

To further validate the generalizability of these findings, we evaluated five additional state-of-the-art vision-language models (mPLUG-Owl2, Fuyu-8B, Kosmos-2, Gemini 1.5, and Claude 3 Opus) representing diverse training pipelines and modality fusion strategies. We conducted controlled experiments on the ChartQA dataset with a retrieval depth of $k = 3$. The results consistently revealed a U-shaped accuracy pattern across all new models, where performance significantly declined when the correct evidence was placed in the middle of the input sequence. These results confirm that position sensitivity is not an artifact of a specific architecture but a systemic and generalizable phenomenon across a wide range of multimodal RAG systems.

In summary, our controlled experiments across multiple datasets and models consistently confirm that position bias is a systemic phenomenon in multimodal RAG systems. The U-shaped accuracy pattern is robust across modalities, architectures, and retrieval depths, with evidence at the middle positions persistently underestimated. These findings underscore the order sensitivity intrinsic to multimodal RAG and motivate the need for quantification and mechanistic analysis, which we pursue in the following sections.

3 Evaluation and Analysis

3.1 Quantification of Bias Amplitude

To quantify the intensity of the position bias problem of large models under three modalities, we define the position sensitivity as:

$$PSI_p = \frac{1}{p}\sum_{i\in\mathcal{T}_p} A_i - \frac{1}{p}\sum_{j\in\mathcal{B}_p} A_j. \quad (1)$$

Among them, $A_i$represents the accuracy rate when the evidence is placed at the ith position; $\mathcal{T}_p$ and $\mathcal{B}_p$ respectively represent the set of p position indexes with the highest and lowest accuracy rates. When p=1 and the highest/lowest positions are respectively at the beginning, the end and the middle, this indicator is denoted as PSI.

Based on the accuracy data in the above Table 1, we calculated the PSI values for MS-MARCO (text), ChartQA (image), and VEGA (image-text mixed) scenarios, respectively. The results are summarized in Table 2.

Modality / Model	MiniCPM	Qwen2-VL	Llama-3.2	GPT-4o
Text-only	0.0358	0.0372	0.0207	0.0570
Image-only	0.0405	0.0324	0.0237	0.1066
Image-Text Mixed	0.0458	0.0379	0.0294	0.0536

Table 2: Comparison of Position sensitivity PSI of each model in the three modalities.

It can be seen from the Table 2 that there are systematic differences in the position sensitivity of the model under different modal scenarios. In the text-only scenario, the PSI values of all models are between 0.020 and 0.050, indicating that the influence of the positions of text evidence on the generation performance is relatively balanced and moderate. Secondly, the image-only task amplified the sensitivity of most models. The PSI of GPT-4o soared to 0.1066, which was 86.9% higher

than the text-only baseline. Paradoxically, Qwen2-VL exhibits the opposite behavior with reduced sensitivity, possibly due to its vision-centered architecture enhancing robustness to disturbances in image sequences. In the mixed-modality scenario, the sensitivity of most models is further amplified. This result indicates that the cross-modal attention mechanism in the multimodal interaction process will further magnify the existing positional bias, resulting in the generation performance being more sensitive to the order of evidence arrangement. In terms of the performance differences among models, Llama-3.2 shows a relatively higher sensitivity ($PSI \approx 0.045$) under the mixed-modality conditions, suggesting that its cross-modal fusion strategy may be more dependent on the sequence order; The sensitivity of GPT-4o is relatively balanced, indicating that it has a stronger adaptability to positional disturbances. Based on the above findings, in the next section, we will further explore the influence mechanism of the retrieval scale on the position bias amplitude through the increment experiment of the retrieval quantity analysis.

To further validate our controlled "gold + distractor" setup in comparison with more realistic noisy retrieval scenarios, we conducted additional experiments on ChartQA using Qwen2-VL-7B with $k = 5$, comparing a clean setting where all retrieved evidence was relevant with noisy settings where distractors were progressively replaced by unrelated documents from TextVQA. The results indicate that as noise increases, overall accuracy gradually decreases, reaching up to a 5% drop at 80% noise, while the position sensitivity index (PSI_p) correspondingly diminishes. Compared with the fully relevant setting, introducing retrieval noise leads the model to distribute its attention more diffusely, interpret the context more conservatively, and respond less sensitively to the placement of evidence. This behavior is expected, as PSI_p fundamentally measures a model's sensitivity to the position of useful evidence, which depends on the model's ability to identify and prioritize relevant content and manifests as a preference for certain positions, such as the top or end of the evidence sequence. Under substantial noise, however, the model struggles to distinguish gold evidence from distractors, resulting in blurred attention distributions and a weakened positional signal, and consequently, the position bias becomes less pronounced, which is reflected in lower PSI_p scores.

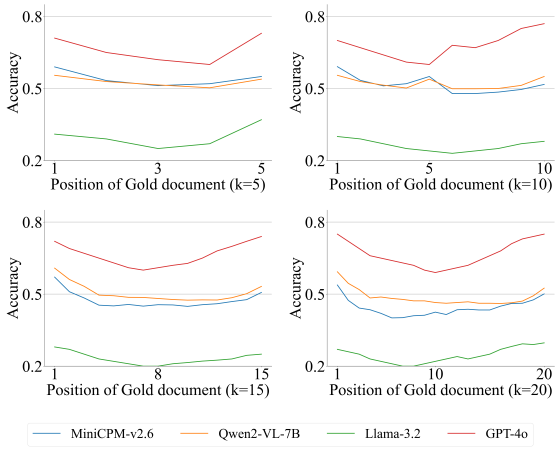

Figure 3: Accuracy under varying retrieval sizes. On the ChartQA dataset, four models are evaluated with retrieval sizes $k \in \{5, 10, 15, 20\}$. As k increases, the U-shaped accuracy curves become more pronounced.

3.2 Bias under Varying Retrieval Sizes

To further explore the robustness and causes of the positional bias phenomenon, in this section, multiple random experiments were conducted on the above four models on the ChartQA dataset. The four subgraphs in Figure 3 respectively show the corresponding expanded U-shaped curves at $k \in \{5, 10, 15, 20\}$. For each k, we successively place the correct evidence at position i ($i \in k$), record and plot the average accuracy curve of each model. As the number of retrieval results increases, the accuracy rate at the middle position continuously decreases, while the changes at the top and end positions are relatively small, thereby further magnification the position bias.

Subsequently, we calculated PSI_p based on the average accuracy rate of each position and the sample variance. The results are shown in the Table 3. As shown in the table, the PSI_p of all models shows a continuous growth trend with the increase of the retrieval quantity k : When k increased from 3 to 20, the sensitivity of MiniCPM-v2.6 rose from approximately 0.018 to 0.060, and Qwen2-VL increased from 0.020 to 0.080. Llama-3.2 and GPT-4o also increased from 0.030 and 0.028 to 0.085 and 0.082 respectively. The slight increase in the variance term indicates that the results of multiple random experiments are stable and reliable. This phenomenon intuitively reflects that as the number of retrieval results increases, the model neglects the evidence in the middle more seriously. However, the top and end position can still maintain

a relatively high level of attention, resulting in a further widening of the performance gap between the different positions, and the model's bias from the sequence order becomes more severe.

k	MiniCPM-v2.6	Qwen2-VL-7B	Llama-3.2-11B	GPT-4o
3	$\mathbf{0.040}_{(\pm 0.012)}$	$\mathbf{0.032}_{(\pm 0.010)}$	$\mathbf{0.024}_{(\pm 0.011)}$	$\mathbf{0.107}_{(\pm 0.009)}$
5	$\mathbf{0.078}_{(\pm 0.009)}$	$\mathbf{0.052}_{(\pm 0.013)}$	$\mathbf{0.065}_{(\pm 0.010)}$	$\mathbf{0.113}_{(\pm 0.010)}$
10	$\mathbf{0.094}_{(\pm 0.010)}$	$\mathbf{0.114}_{(\pm 0.009)}$	$\mathbf{0.097}_{(\pm 0.011)}$	$\mathbf{0.119}_{(\pm 0.011)}$
15	$\mathbf{0.108}_{(\pm 0.012)}$	$\mathbf{0.134}_{(\pm 0.011)}$	$\mathbf{0.105}_{(\pm 0.009)}$	$\mathbf{0.137}_{(\pm 0.012)}$
20	$\mathbf{0.119}_{(\pm 0.013)}$	$\mathbf{0.137}_{(\pm 0.012)}$	$\mathbf{0.118}_{(\pm 0.014)}$	$\mathbf{0.152}_{(\pm 0.013)}$

Table 3: Position sensitivity PSI_p (mean ± variance) of each model under varying retrieval sizes. Results on ChartQA with retrieval sizes $k \in \{3, 5, 10, 15, 20\}$.

Finally, to characterize the rate at which the degree of bias increases with k, we fit the relationship of each model PSI_p with respect to $ln(k)$ to a linear model,

$$PSI_p \;=\; \alpha \,\ln(k) \;+\; \beta. \tag{2}$$

With Llama-3.2, for example, the least squares fitting get $\alpha = 0.035, \beta = 0.010$, goodness-of-fit $R^2 = 0.986$. Similarly, the slopes α of the four models are all between 0.030 and 0.040, verifying the law that the position bias linearly amplifies with the logarithmic growth of the retrieval scale. This result suggests that when designing a multimodal RAG system, the excessive number of retrieval items should be carefully controlled or a position reweighting mechanism should be introduced to alleviate the significant sequence bias caused by the increase of k.

3.3 Preliminary Mitigation Attempts

To provide a preliminary assessment of current solutions' effectiveness, we empirically evaluated two heuristic mitigation methods on the ChartQA dataset using the Qwen2-VL-7B model with a retrieval size of $k = 5$: (1) symmetric reordering of evidence and (2) position-equalizing prompts. Our findings indicate that symmetric reordering improved average accuracy from 49.21% to 55.35% (a 6.14% increase) while reducing the Position Sensitivity Index (PSI_p) from 0.0324 to 0.0279. Similarly, using a positional prompt increased accuracy to 53.62% (a 4.41% increase) and lowered the PSI_p to 0.0247. While these approaches moderately reduce bias and improve accuracy, they fail to eliminate the issue, confirming the persistence of position bias and reinforcing the need for more principled mitigation techniques.

4 In-depth Exploration of Hidden Bias

4.1 Visualization Methodology

Our study investigates the causes of position bias in multimodal RAG systems via a multi-level visual analysis framework based on the Qwen2-VL-7B and Llama-3.2-11B-Vision-Instruct models. The framework integrates cross-modal attention heatmap visualization and quantitative bias metrics to systematically analyze attention behaviors. The methodology is organized as follows.

Under a fixed retrieval setting $k = 3$, we construct controlled sequences by placing the correct evidence at three positions within the retrieved inputs: top (position 1), middle (position 2), and end (position 3). All decoder parameters remain frozen to ensure consistent decoding across configurations. Cross-modal attention weights are extracted from the 14th decoder layer—empirically selected as optimal (justified in the next section)—and rendered as interpretable visualizations.

All input images are preprocessed using bilinear interpolation to a unified resolution of 616×644 to ensure spatial consistency. Text tokens are generated from templated prompts and tokenized with the model's native tokenizer. For visual inputs, a Vision Transformer backbone is employed to divide images into spatial patches:

$$T_v = \text{PatchEmbed}(I) \in R^{(H/28)\times(W/28)\times d}, \tag{3}$$

where I denotes the input image, PatchEmbed represents the patch embedding operation, and H, W, d correspond to the image height, width, and embedding dimension, respectively. The resulting tensor T_v represents the visual tokens arranged in a 2D spatial grid.

Once the visual and textual tokens are obtained, we extract cross-modal attention maps from key text tokens (e.g., entities or numerals) to the spatial grid of image tokens. To enhance interpretability, the raw attention matrix is normalized within a defined Region of Interest (ROI):

$$A_{\text{norm}}(i,j) = \frac{A(i,j) - \min(A_{\text{ROI}})}{\max(A_{\text{ROI}}) - \min(A_{\text{ROI}})}, \tag{4}$$

where $A(i,j)$ is the attention weight at coordinate (i,j), and A_{ROI} denotes the subregion associated with projected keyword relevance. The normalized attention maps are then converted into visual overlays. Specifically, we compute an overlay $\mathcal{O}^{(i)}$ via

a visualization mapping function $\mathcal{M}$ applied to the attention distributions:

$$\mathcal{O}^{(i)} = \mathcal{M}(\alpha_{\text{src}}^{(i)}, \alpha_{\text{tgt}}^{(i)}), \tag{5}$$

where $\mathcal{M}$ maps normalized attention scores from white (low intensity) to red (high intensity). The heatmap is blended with the original image using transparency-based overlaying with a fixed alpha value ($\alpha = 0.5$):

$$\text{Overlay} = \alpha I + (1 - \alpha) \cdot \text{Heatmap}, \tag{6}$$

preserving both semantic content and attention saliency.

To support quantitative analysis of positional bias, we further generate "attention difference heatmaps." These visualizations highlight contrastive patterns in attention allocation. Additionally, a position bias matrix is computed:

$$\Delta A_{\text{pos}} \in R^{3\times 3}, \tag{7}$$

along with region-specific attention scores:

$$S_{\text{ROI}} = \frac{1}{N_{\text{patch}}} \sum_{(x,y)\in \text{ROI}} A_{\text{norm}}(x, y), \tag{8}$$

and global attention scores:

$$S_{\text{global}} = \frac{1}{HW} \sum_{x=1}^{H} \sum_{y=1}^{W} A_{\text{norm}}(x, y), \tag{9}$$

to quantify the model's sensitivity to evidence position. Final visualizations are rendered via Matplotlib's subplot interface for comparative, multi-view presentation.

4.2 Layer-Specific Sparsity Analysis

To quantify the sparsity degree of cross-modal attention distribution in different decoding layers, we adopt the 'Normalized 2D Entropy' (N2E) index. The indicator definition is as follows: Given a cross-modal attention weight matrix of a certain layer $A \in R^{H\times W}$, first normalize it:

$$p_{ij} \;=\; \frac{A_{ij}}{\sum_{m=1}^{H} \sum_{n=1}^{W} A_{mn}}. \tag{10}$$

Next, we compute the two-dimensional Shannon entropy:

$$H_{2D}(A) \;=\; -\sum_{i=1}^{H} \sum_{j=1}^{W} p_{ij} \,\ln p_{ij}. \tag{11}$$

This entropy is then normalized to the range $[0, 1]$:

$$\text{N2E}(A) \;=\; \frac{H_{2D}(A)}{\ln(H\,W)}\,. \tag{12}$$

Finally, the two-dimensional sparsity index is derived as:

$$S_{2D}(A) \;=\; 1 - \text{N2E}(A). \tag{13}$$

Among them, $S_{2D} = 1$ indicates extreme sparsity (attention is highly concentrated on certain patches), $S_{2D} = 0$ indicates a completely uniform distribution. For Qwen2-VL-7B-instruct and Llama-3.2-11B-Vision-Instruct, Respectively to extract cross modal cross attention (text and visual) matrix in each layer $\ell = 0, 1, \ldots, 27$, and the calculation results of three typical layers (shallow layer $\ell = 3$, middle layer $\ell = 14$, and deep layer $\ell = 24$) are taken as shown in the Table 4.

Model / Hierarchy	$\ell = 3$ (shallow layer)	$\ell = 14$ (middle layer)	$\ell = 24$ (deep layer)
Qwen2-VL-7B	0.42	**0.72**	0.68
Llama-3.2-11B	0.45	**0.75**	0.70

Table 4: Layer-wise sparsity index $S_{!2D}$ of cross-modal attention.

From the table, we observe that both models exhibit lowest sparsity in shallow layers, suggesting more uniformly distributed attention. In contrast, middle layers exhibit the highest sparsity, indicating that attention becomes highly concentrated on semantically critical regions, which is a hallmark of effective cross-modal feature alignment. The sparsity remains relatively high in deeper layers, likely reflecting the model's focus on reasoning-critical visual regions in support of text generation. Based on this quantitative trend, we hypothesize that in the 28-layer decoder of Qwen2-VL, the shallow layers ($\ell = 0 \sim 11$) mainly focus on unimodal processing (e.g., text self-attention or local visual features), the middle layers ($\ell = 12 \sim 19$) progressively enhance cross-modal attention and are critical for cross-modal integration, and the deep layers ($\ell = 20 \sim 27$) primarily engage in text-centric reasoning. Therefore, in the following sections, we select the 14th layer for attention heatmap visualization and position bias analysis, as it represents the point of strongest cross-modal interaction.

4.3 Attention Discrepancy and Visualization

To further investigate the position bias in cross-modal attention, we visualize the attention maps

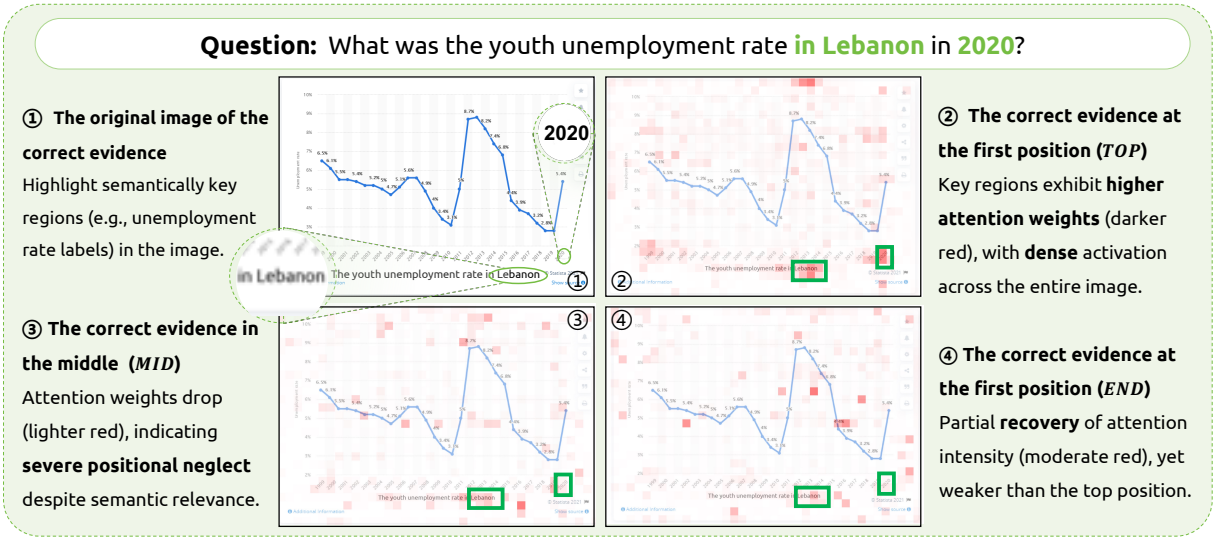

Figure 4: Cross-modal attention heatmaps illustrate the impact of evidence position. For the query on Lebanon's 2020 youth unemployment, attention is strongest at the top, weakens in the middle, and partially recovers at the end, confirming position-sensitive attention allocation.

at the 14th decoder layer under three document position configurations: *Top*, *Mid*, and *End*. Figure 4 shows the semi-transparent overlays of the model's attention heatmaps overlaid on the corresponding input images, where only one image contains the correct evidence, and the other two serve as distractors. A concrete example is illustrated in Figure 4: When analyzing the youth unemployment rate in Lebanon (2020), the model focuses on semantically critical regions such as the numeric labels and axis titles in the bar chart (highlighted by black boxes). Across all cases, we observe that the model consistently attends to semantically relevant regions in the correct image, confirming that the 14th layer captures meaningful cross-modal interactions. However, the magnitude and focus of attention vary noticeably depending on the position of the correct evidence in the retrieval sequence.

To quantitatively examine how attention varies with position, we compute patch-wise differences between attention maps across configurations and visualize them in Figure 5. Each heatmap illustrates the difference between two positional settings (Top-Mid, Top-End, and Mid-End), where red indicates increased attention and blue indicates reduced attention for the first position relative to the second.

It can be observed from the difference heatmaps that when the correct document is located at the beginning position, the overall attention received is higher than when it is located in the middle and the end, and the attention received by the local key areas is significantly higher than that at the other two positions. This emphasizes the strong bias of the model towards the main evidence. Moreover, when the correct evidence is located at the end, the overall attention score is also higher than when it is located in the middle position, which confirms that the model tends to underestimate the middle evidence in its cross-modal reasoning.

In summary, the attention heatmaps and their pairwise differences jointly confirm a position-sensitive bias in cross-modal attention allocation. The *Top* and *End* positions induce stronger focus on relevant evidence, while the *Mid* position leads to diluted attention. These findings provide both visual and quantitative support for incorporating position-aware strategies such as evidence reweighting or positional prompts to mitigate attention imbalance.

5 Related Work

Single-mode RAG position bias Position bias is a known challenge in single-mode Retrieval-Augmented Generation (RAG) (Wang et al., 2024b). LLMs often exhibit sensitivity to information position in long contexts, with higher attention at sequence ends and neglected middles, a phenomenon linked to pre-training preferences in Transformer models (Coelho et al., 2024). Similarly, graph data serialization can alter LLM perception of topological structures (Ge et al., 2024). Prompt-based attention direction has been proposed to mitigate this bias (Zhang et al., 2024).

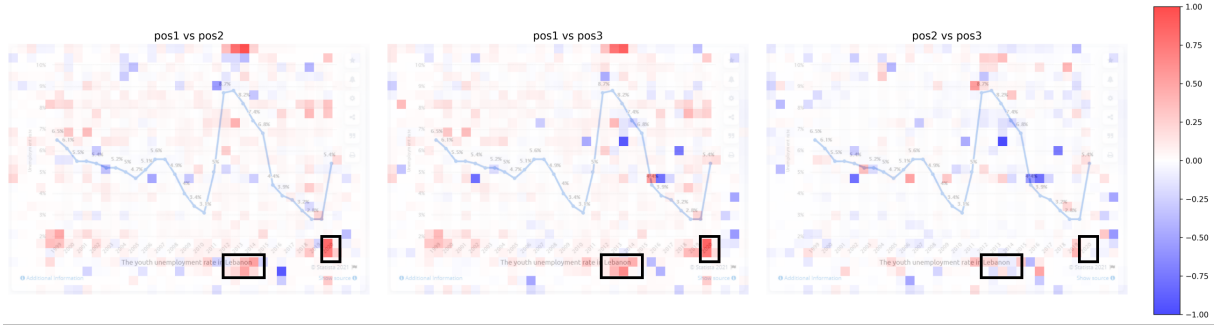

Figure 5: Visualization of attention differences at layer 14 under different gold document position configurations. Each subfigure shows patch-wise attention differences between two settings. The red-blue colormap indicates relative attention (red = higher in the first configuration; blue = higher in the second). Black boxes highlight semantically relevant regions with notable shifts.

Multimodal RAG systems Advancements in multimodal RAG include Google's ColPali framework for end-to-end document understanding using delayed interaction encoders (Faysse et al., 2024). Other approaches involve multi-agent cooperative retrieval planning for enhanced reasoning robustness and cross-modal allocation-causal reasoning architectures targeting semantic gaps with joint embeddings and hierarchical retrieval, notably in medical image QA.

Position Bias in Multimodal RAG While efforts to quantify position bias include metrics like ListT5 for text retrieval (Yoon et al., 2024) and Landmark Embedding with location-aware objectives (Luo et al., 2024), current quantitative methods are often limited to text-only tasks or do not readily extend to multimodal scenarios. In such multimodal contexts, position bias also significantly affects Multimodal LLMs (MLLMs). Work by (Tan et al., 2024) demonstrated that MLLMs exhibit this bias, favoring visual features when placed at sequence beginnings or ends. Although this study confirmed position bias in RAG settings and suggested prompting for mitigation, its experimental design involved a limited number of candidates (e.g., four items). This setup may not capture potential attention dilution effects present in real-world scenarios involving larger-scale retrieval results (e.g., more than five candidates), a limitation our work addresses.

6 Conclusion

Our study establishes foundational insights into position bias in multimodal retrieval-augmented generation (RAG) systems, uncovering systematic performance instability rooted in the ordering of retrieved evidence. Our experiments reveal that multimodal reasoning exhibits heightened sensitivity to positional arrangements compared to unimodal settings, with accuracy following a distinct U-shaped trajectory as retrieval scope expands. To quantify the bias, we introduce a position sensitivity metric PSI_p and an interpretable diagnostic framework, demonstrating that cross-modal attention mechanisms disproportionately prioritize sequence extremities while neglecting middle content. Our work provides a fairer evaluation framework and theoretical support for the design of multimodal RAG systems.

Limitations

While our study establishes foundational insights into position bias in multimodal RAG systems, several directions warrant further exploration. First, the experimental scope is constrained to retrieval scales $k \leq 20$ and models under 11B parameters due to computational resource limitations(requiring more than 16 GPUs), leaving larger-scale scenarios and frontier architectures (e.g. Claude 3) for future investigation. Second, while we empirically observe model-specific sensitivity patterns (e.g. Llama-3.2's higher bias), the analysis does not systematically correlate these differences with architectural designs like attention mechanisms and cross-modal fusion strategies, which could deepen theoretical understanding. Third, although attention visualization reveals critical bias propagation patterns, a unified theoretical framework explaining how positional encoding interacts with multimodal fusion remains to be developed. Fourth, our analysis focuses on static attention distributions, whereas temporal dynamics in multi-step reasoning tasks, where positional effects may evolve, re-

quire dedicated study. Finally, operationalizing real-time debiasing mechanisms (e.g. dynamic position reweighting) in practical systems presents an open engineering challenge. These limitations collectively outline promising avenues for advancing both the theory and application of robust multimodal RAG systems.

Ethical Statement

This research does not involve human subjects, personally identifiable information, or sensitive data. All experiments are conducted on publicly available benchmark datasets and models (e.g., Qwen2-VL-7B-instruct, Llama-3.2-11B-Vision-Instruct). We strictly follow the terms of use and licensing agreements of these resources. Our analysis aims to improve the robustness and fairness of multimodal retrieval-augmented generation systems by identifying and mitigating potential positional biases. We believe that our findings can contribute to building more interpretable and responsible AI systems.

Acknowledgements

This work is supported partially by the Strategic Priority Research Program of the CAS under Grants No. XDB0680102 and the National Natural Science Foundation of China under Grant Nos. 62472408, 62372431, 62441229, and 62377043.

References

Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. 2025. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. *Preprint*, arXiv:2502.08826.

AI@Meta. 2024. Llama 3 model card.

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, and 1 others. 2024. Context-dpo: Aligning language models for context-faithfulness. *arXiv preprint arXiv:2412.15280*.

João Coelho, Bruno Martins, Joao Magalhaes, Jamie Callan, and Chenyan Xiong. 2024. Dwell in the beginning: How language models embed long documents for dense retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 370–377, Bangkok, Thailand. Association for Computational Linguistics.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2025. From local to global: A graph rag approach to query-focused summarization. *Preprint*, arXiv:2404.16130.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Colpali: Efficient document retrieval with vision language models. *Preprint*, arXiv:2407.01449.

Yuyao Ge, Shenghua Liu, Baolong Bi, Yiwei Wang, Lingrui Mei, Wenjie Feng, Lizhe Chen, and Xueqi Cheng. 2024. Can graph descriptive order affect solving graph problems with llms? *Preprint*, arXiv:2402.07140.

Kun Luo, Zheng Liu, Shitao Xiao, Tong Zhou, Yubo Chen, Jun Zhao, and Kang Liu. 2024. Landmark embedding: A chunking-free embedding method for retrieval augmented long-context large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3268–3281, Bangkok, Thailand. Association for Computational Linguistics.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268.

OpenAI. 2024. Hello, gpt-4o | openai.

OpenBMB. 2024. openbmb/minicpm-v-2_6.

Aleksander Theo Strand, Sushant Gautam, Cise Midoglu, and Pål Halvorsen. 2024. Soccerrag: Multimodal soccer information retrieval via natural queries. *Preprint*, arXiv:2406.01273.

Zhijie Tan, Xu Chu, Weiping Li, and Tong Mo. 2024. Order matters: Exploring order sensitivity in multimodal large language models. *Preprint*, arXiv:2410.16983.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yiwei Wang, Yujun Cai, Muhao Chen, Yuxuan Liang, and Bryan Hooi. 2024b. Primacy effect of chatgpt. *Preprint*, arXiv:2310.13206.

Siwei Wu, Kang Zhu, Yu Bai, Yiming Liang, Yizhi Li, Haoning Wu, Jiaheng Liu, Ruibo Liu, Xingwei Qu, Xuxin Cheng, and 1 others. 2024. Mmra: A benchmark for multi-granularity multi-image relational association. *arXiv preprint arXiv:2407.17379*.

Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. 2024a. Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085*.

Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. 2024b. RULE: Reliable multimodal RAG for factuality in medical vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1081–1093, Miami, Florida, USA. Association for Computational Linguistics.

Soyoung Yoon, Eunbi Choi, Jiyeon Kim, Hyeongu Yun, Yireun Kim, and Seung-won Hwang. 2024. ListT5: Listwise reranking with fusion-in-decoder improves zero-shot retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2287–2308, Bangkok, Thailand. Association for Computational Linguistics.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *Preprint*, arXiv:2410.10594.

Jianhao Yuan, Shuyang Sun, Daniel Omeiza, Bo Zhao, Paul Newman, Lars Kunze, and Matthew Gadd. 2024. Rag-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model. *arXiv preprint arXiv:2402.10828*.

Meiru Zhang, Zaiqiao Meng, and Nigel Collier. 2024. Can we instruct LLMs to compensate for position bias? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12545–12556, Miami, Florida, USA. Association for Computational Linguistics.

Xiangyu Zhao, Yuehan Zhang, Wenlong Zhang, and Xiao-Ming Wu. 2024. Unifashion: A unified vision-language model for multimodal fashion retrieval and generation. In *Conference on Empirical Methods in Natural Language Processing*.

Chenyu Zhou, Mengdan Zhang, Peixian Chen, Chaoyou Fu, Yunhang Shen, Xiawu Zheng, Xing Sun, and Rongrong Ji. 2024a. Vega: Learning interleaved image-text comprehension in vision-language large models. *Preprint*, arXiv:2406.10228.

Shuyan Zhou, Uri Alon, Frank F. Xu, Zhiruo Wang, Zhengbao Jiang, and Graham Neubig. 2023. Docprompting: Generating code by retrieving the docs. In *International Conference on Learning Representations (ICLR)*, Kigali, Rwanda.

Zhongliang Zhou, Jielu Zhang, Zihan Guan, Mengxuan Hu, Ni Lao, Lan Mu, Sheng Li, and Gengchen Mai. 2024b. Img2loc: Revisiting image geolocalization using multi-modality foundation models and image-based retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2749–2754.