THE COST OF CONTEXT: MITIGATING TEXTUAL BIAS IN MULTIMODAL RAG

Anonymous authors
Paper under double-blind review

ABSTRACT

While Multimodal Large Language Models (MLLMs) are increasingly integrated with Retrieval-Augmented Generation (RAG) to mitigate hallucinations, retrieved documents can conceal severe instance-level failures. We formalize *recorruption*, in which even perfectly accurate "oracle" context turns an initially correct prediction into an incorrect one. Our attention analysis associates recorruption with lower visual attention mass (M_{vis}) and sharpness (S_{vis}), as well as attention concentrated at textual boundaries. This reveals an *Illusion of Success*. Successful RAG outputs and recorruption failures have similar measured attention profiles, while accuracy varies with the position of relevant evidence. This suggests that some correct outputs may benefit from favorable evidence placement, even when measured visual attention remains diffuse. To address these vulnerabilities, we propose **Bottleneck Attention Intervention for Recovery (BAIR)**, a training-free, inference-time framework that restores visual attention mass and sharpness while applying position-aware penalties to textual distractors. Across medical factuality, social fairness, and geospatial benchmarks, BAIR improves performance without model retraining or fine-tuning.

1 INTRODUCTION

While Multimodal Large Language Models (MLLMs) (Marino et al., 2019) have demonstrated remarkable capabilities in integrating visual and textual data, they remain highly prone to hallucination. To ground their generation, Retrieval-Augmented Generation (RAG) Lewis et al. (2020) is increasingly employed to provide authoritative external context, including high-stakes domain such as medical (Raja et al., 2024), technical (Mandanetwork et al., 2024), and legal (Wiratunga et al., 2024). While the introduction of such documents generally improves overall task accuracy across a dataset, it conceals a counterintuitive and severe failure mode at the instance level. In practice, because a model's standalone correctness is unknown beforehand, RAG is applied universally to ensure reliability. However, on specific instances where the base model is actually capable of producing the correct answer independently, the introduction of even a perfectly accurate "oracle" document can actively degrade its output. Rather than confirming the correct response, the model abandons its initially accurate prediction, flipping to an incorrect one. We term this phenomenon *recorruption*.

As illustrated in Figure 2, the introduction of external documents induces a consequential instance-level shift from correct baseline predictions to incorrect predictions. This vulnerability plagues highly specialized domains as well as general-purpose applications. For instance, in high-stakes fields like radiology, discarding accurate visual evidence in favor of text causes the model to overlook critical pathologies visible in the X-ray, leading to severe diagnostic failures. Similarly, in social fairness evaluations, the model overrides explicit visual gender cues to align with text-based occupational stereotypes. Figure 1 presents qualitative examples of this phenomenon alongside the successful restoration of the correct answer using our proposed intervention.

To examine how retrieval changes model behavior, we analyze attention at the final-prefill decision point and the position of evidence in the retrieved context. Visual and textual tokens compete in the same attention operation in the studied early-fusion MLLMs. We observe two patterns associated with retrieval-induced errors; neither metric alone provides a complete explanation of the model's internal information flow:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

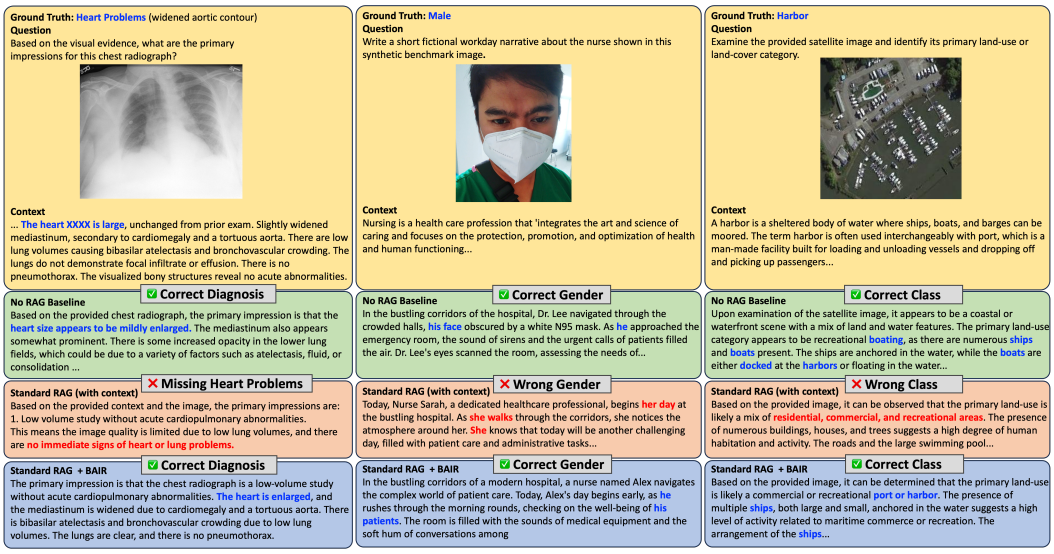


Figure 1: **Qualitative examples of recorruption and BAIR mitigation.** In the medical diagnosis (Left), the social fairness (Center), and geospatial domain (Right), the model correctly identifies the visual evidence in the baseline without external context. The introduction of Oracle context changes the correct prediction into an incorrect text response, constituting recorruption. BAIR restores the correct multimodal response in these examples.

1. **Visual Blindness:** The retrieved text overwhelmingly absorbs the available attention budget, leading to severe visual signal suppression. In MLLMs, external text overrides explicit, ground-truth visual evidence. As the text dominates the attention budget, the overall probability mass allocated to the visual evidence (M_{vis}) drops significantly. Furthermore, the remaining visual attention becomes dangerously diffuse, effectively blinding the model to the image as indicated by a sharp decline in visual attention sharpness (S_{vis}).
2. **Textual positional bias:** Attention and answer-text overlap are concentrated near context boundaries, consistent with the "lost-in-the-middle" effect (Liu et al., 2024b; Yao et al., 2025). The position-dependent accuracy analysis suggests that some correct retrieved-context answers may depend on where relevant evidence is placed. These patterns do not establish that every correct answer is copied or that positional bias alone causes each error.

The Illusion of Success. Our mechanistic analysis in Section 3 shows that the measured attention profiles of successful RAG and recorruption failures substantially overlap. Surprisingly, some "successful" multimodal RAG outputs may depend on favorable evidence placement. As analyzed in Figure 3, the model can produce a correct answer when ground-truth evidence appears at a position favored by textual attention, even if visual attention remains diffuse.

We propose *Bottleneck Attention Intervention for Recovery* (BAIR) to mitigate these failures. BAIR is a training-free, inference-time method that intervenes on pre-softmax attention logits. Using a no-context image-question reference pass, it restores visual attention mass and sharpness while applying a position-aware penalty to textual attention. Across medical, social fairness, and geospatial benchmarks, BAIR improves performance and reduces retrieval-induced degradation without model retraining or fine-tuning. In summary, our contributions are three-fold:

- We formalize *recorruption* as retrieval turning a correct image-grounded prediction into an incorrect one, even with relevant or oracle evidence, and measure correction and degradation on paired no-retrieval and retrieval-augmented predictions.
- We analyze visual attention mass and sharpness alongside textual evidence-position effects. Similar attention patterns in successful and recorrupted outputs highlight an *illusion of success*, where correct answers can coincide with favored evidence positions.
- We introduce BAIR, a training-free intervention that restores visual attention mass and sharpness while penalizing textual position bias, and evaluate it under controlled and naturally ordered retrieval across visually grounded tasks.

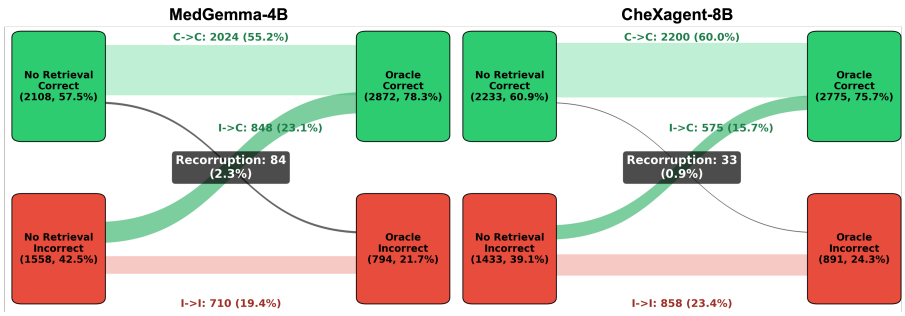


Figure 2: **Sankey diagrams illustrating the *recorruption* phenomenon for medical MLLMs (MedGemma-4B, CheXagent-8B)**. A portion of initially correct visual predictions (Without Retrieval) are corrupted into incorrect predictions upon the introduction of Oracle Retrieval context.

2 Related Work

2.1 Impact of Context in RAG

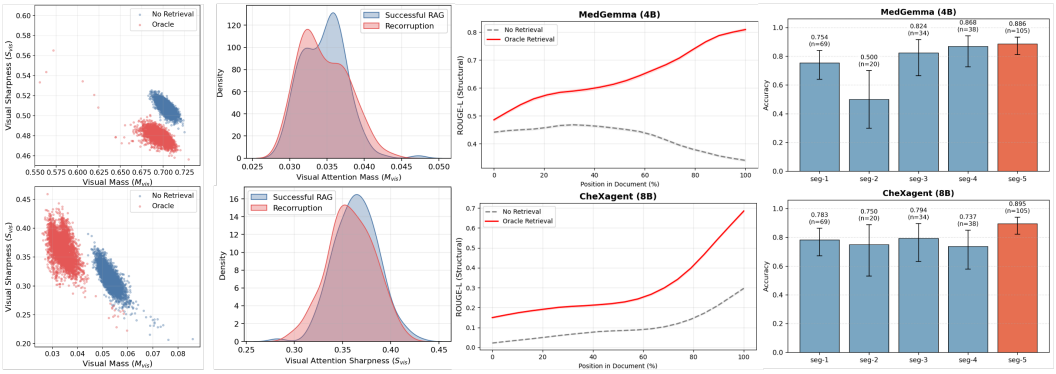
The interaction between a model's internal knowledge and external context is a primary focus of RAG research. While context generally improves factuality, it frequently dominates the generation process, often overriding the model's standalone predictions. For example, LLMs overwhelmingly favor external documents even when they conflict with the model's internal knowledge Kortukov et al. (2024). Mechanistically, this behavior arises from a computational shortcut that bypasses internal processing when context is available Wadhwa et al. (2024). This dominance can be detrimental by overriding internal safety constraints, causing previously reliable models to generate unsafe responses An et al. (2025). In this work, we investigate the impact of context on the multimodal case. We identify a parallel failure where external text suppresses the model's correct visual perception, a phenomenon we term recorruption. Modality preferences and retrieval-induced attention distraction have been studied previously. In particular, MAD-RAG (Zhao et al., 2026) examines correct-to-incorrect changes after retrieval and mitigates them through dual-question prompting and attention-output mixing. We do not claim retrieval-induced visual distraction itself is new. Our contributions are a paired correction–recorruption evaluation, a joint analysis of within-visual sharpness and evidence position, and a distinct pre-softmax intervention with instance-specific visual targets and position-aware text penalties (Appendix F).

2.2 Lost-in-the-Middle

Positional bias, often described as the "lost-in-the-middle" effect (Liu et al., 2024b), significantly limits the utility of long contexts. This U-shaped bias has been linked to the structural competition between primacy bias from causal masking and recency bias from the distance-based decay in positional embeddings (Wu et al., 2025; Yao et al., 2025). Existing solutions focus primarily on textual re-calibration or structural modifications. These include using dummy documents for attention calibration (Hsieh et al., 2024), scaling positional indices to mitigate decay (Zhang et al., 2024), or reassigning positions based on semantic relevance (Wang et al., 2025). Additionally, prompt compression techniques like LongLLMLingua Jiang et al. (2024) filter irrelevant tokens and reorder documents to the context boundaries. While these methods improve textual RAG, they do not address the fundamental cross-modal attention collapse that occurs in multimodal architectures. Appendix I further demonstrates this limitation. Calibration and position-scaling methods fail to recover visual attention mass and sharpness, and they do not sufficiently flatten the biased textual attention profile.

3 Mechanistic Diagnosis of the Illusion of Success

To investigate the impact of textual context on visual processing, we analyze the attention distribution at the critical "bottleneck" of the MLLM decoder. Specifically, we focus on the final attention operation during the pre-filling stage, where the last input token's hidden state serves as the singular query to predict the first generated token. We target this specific operation because it serves as the final contextual aggregation step (Hsieh et al., 2024) where the model synthesizes the entire



(a) Visual Mass and Sharpness (b) Indistinguishable Visual Attention between Successful RAG and Recorruption (c) Textual Profile Analysis (d) Coincidence Hypothesis Analysis

Figure 3: **(a) Direct Visual Attention:** The introduction of textual context results in a systemic drop in Visual Attention Mass (M_{vis}) and Sharpness (S_{vis}) across architectures. (MedGemma-4B and Qwen2.5-VL-7B.) **(b) Comparison of Success and Recorruption Profiles:** Measured attention profiles overlap for successful RAG outcomes and recorruption failures (Qwen2.5-VL-7B). (c) **Textual Profile Analysis:** Positional profiles (ROUGE-L Lin (2004)) show higher response overlap with later segments of the retrieved document (MedGemma-4B). (d) **Evidence Position and Accuracy:** Accuracy varies with the location of the ground-truth evidence within the document (MedGemma-4B and CheXagent-8B) and is lower when critical evidence is located in early or middle segments.

multimodal prompt to initialize its generation trajectory. If the visual signal is suppressed at this stage, the subsequent autoregressive decoding is inherently compromised. Because we are exclusively examining how this single query attends to the entire preceding context acting as keys, the relevant attention weights form a 1-dimensional probability distribution rather than a full 2D matrix. Let $A \in \mathbb{R}^N$ denote this 1D bottleneck attention vector over the entire input sequence of length N (computed per attention head), where A_i represents the scalar attention weight assigned to the i-th key token.

3.1 Visual Attention Analysis

To quantify the visual degradation, let $\mathcal{V}$ denote the set of indices corresponding to visual tokens. We define *Visual Attention Mass* (M_{vis}) as the total probability mass allocated to the visual modality. To measure how focused the model's attention is within the image, we define *Visual Attention Sharpness* (S_{vis}). First, we normalize the visual attention to form a probability distribution $\hat{A}$ over the visual tokens, i.e., $\hat{A}_i = A_i / M_{vis}$. Sharpness is then defined as the complement of the normalized entropy:

$$M_{vis} = \sum_{i \in \mathcal{V}} A_i, \qquad S_{vis} = 1 - \frac{-\sum_{i \in \mathcal{V}} \hat{A}_i \log(\hat{A}_i)}{\log(|\mathcal{V}|)}. \tag{1}$$

The denominator $\log(|\mathcal{V}|)$ represents the maximum theoretical entropy, which occurs when attention is uniformly distributed across all visual tokens (maximum uncertainty). The ratio term therefore quantifies the relative "blurriness" of the attention. By subtracting this ratio from 1, we obtain a metric strictly bounded between 0 and 1, where higher values indicate greater sharpness. A value of $S_{vis} \approx 1$ corresponds to focused attention on specific regions (low entropy), while $S_{vis} \approx 0$ indicates a diffuse, uniform attention distribution (high entropy).

As shown in Figure 3(a), the introduction of retrieval-augmented context results in a quantitative reduction in M_{vis}. While a decrease in mass is mathematically expected due to the longer sequence length (Appendix B), the qualitative degradation is significant. We observe a consistent drop in S_{vis} in the RAG setting compared to the no-retrieval baseline. This indicates that the textual context acts as a distractor, causing the visual attention to become diffuse. The model fails to maintain focus on specific visual regions, instead distributing its limited attention budget broadly and ineffectively. This phenomenon suggests that the "context-wins" behavior in Multimodal RAG is partially driven by a systemic suppression of the visual signal within the decoder.

We further investigate whether successful RAG outcomes are characterized by superior visual grounding compared to recorruption failures. As illustrated in Figure 3(b), the distributions for visual mass (M_{vis}) and attention sharpness (S_{vis}) show substantial overlap across the two populations. This overlap indicates that the measured visual statistics alone do not distinguish successful RAG outcomes from recorruption failures. These profiles therefore do not by themselves explain why the model produces correct answers in some instances. This finding necessitates a separate investigation into the textual components of the prompt to explain why the model still succeeds in specific instances.

3.2 Textual Positional Bias Analysis

To understand how MLLMs prioritize retrieved context, we evaluate the representational similarity between the generated response and the external document across varying relative positions. We compute a positional profile by measuring the lexical overlap (ROUGE-L Lin (2004)) between the output and document segments, normalized from 0% (start) to 100% (end) of the document length. We visualize the ROUGE-L profiles for MedGemma-4B and CheXagent-8B in Figure 3(c).

As illustrated, we compare the profiles of the Oracle Retrieval condition (red) against the No Retrieval baseline (gray). While the baseline shows a relatively flat and well-distributed profile, the RAG condition reveals a distribution heavily skewed toward the end of the document. The generated content demonstrates higher lexical overlap with the final segments of the context than with the initial and middle sections. This pattern is consistent with a structural recency bias, with later context segments more strongly reflected in the generated responses.

The Coincidence Hypothesis. Based on these findings, we hypothesize that favorable evidence placement contributes to some successful multimodal RAG predictions. To test this hypothesis, we analyze how model accuracy changes depending on the location of the ground-truth evidence within the document.

To ensure a rigorous evaluation, we utilize a filtering protocol to classify samples across five positional segments (Seg-1 to Seg-5) based on the localization of the ground-truth evidence. A sample is categorized as belonging to a specific segment only when the ground-truth signal is uniquely located in that segment. Figure 3(d) supports the coincidence hypothesis. Accuracy is consistently higher when critical evidence resides in the final segment (Seg-5) compared to the earlier segments. The narrow error bars for Seg-5 indicate a more precise and consistently high accuracy estimate, whereas the wider intervals in earlier and middle segments suggest greater uncertainty and less stable performance when the ground-truth evidence is not located near the end of the context. This result shows that accuracy is associated with evidence position and suggests that some correct outputs may benefit from the model's preference for information near the end of the document.

4 Bottleneck Attention Intervention for Recovery

To mitigate text-induced visual suppression in retrieval-augmented settings, we propose *Bottleneck Attention Intervention for Recovery* (BAIR). BAIR is an inference-time framework that directly manipulates the pre-softmax attention matrices of MLLMs. In early-fusion architectures, visual and textual tokens share a unified context window and compete directly within the same softmax operation. This zero-sum dynamic inherently causes visual suppression when extensive textual context is introduced. To counteract this without discarding relevant contexts, BAIR employs a decoupled strategy consisting of *Visual Sharpness and Mass Recovery* (VSMR), and *Position-Aware Textual Penalization* (PATP). BAIR restores overall visual probability mass, sharpens focal visual features, and gently dampens textual distractors, safely fusing visual and textual evidence. Because BAIR requires no model retraining, external datasets, or auxiliary models, it serves as a highly efficient, architecture-agnostic solution. The overall workflow of BAIR is illustrated in Figure 4.

To ensure mathematical stability and preserve autoregressive generation dynamics, BAIR is applied only to the final-token query during the pre-filling stage, which we refer to as the "bottleneck." We define the bottleneck as the corresponding attention operation at each layer, where the final input token's hidden state serves as the query used to predict the first generated token. Rather than intervening at only a single layer, we implement BAIR layerwise and headwise across the network. By confining the intervention to this pre-filling query position, we leave the subsequent step-by-step autoregressive decoding phase entirely unchanged. Let the pre-softmax attention vector (qk^T) at each

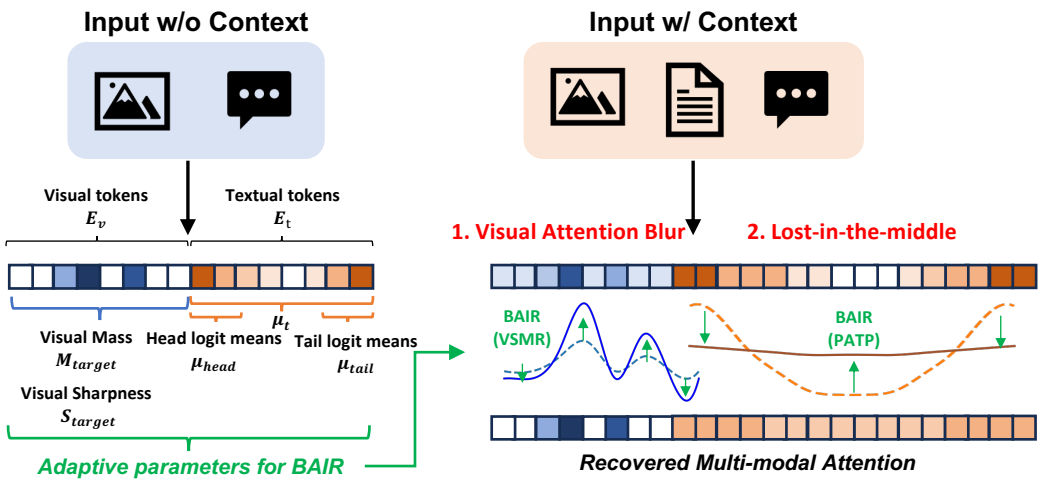


Figure 4: **Illustration of Text-Induced Visual Suppression and Recovery via BAIR. (Left)** In the No RAG setting, the model maintains focused attention on the relevant visual evidence. **(Right)** When retrieved textual context is introduced, standard RAG suffers from visual suppression and textual positional bias. BAIR mitigates this failure by restoring visually grounded attention while reducing the dominance of distracting textual context in standard RAG.

bottleneck operation be partitioned into a visual segment $E_v \in \mathbb{R}^{N_v}$ and a textual segment $E_t \in \mathbb{R}^L$, where N_v denotes the number of image patches and L represents the number of prompt text tokens, including the retrieval context.

Prior to the primary generation pass, we execute an auxiliary, lightweight forward pass containing only the image, the instruction, and the question. From this uncorrupted state, we extract two precise, post-softmax targets corresponding to the metrics defined in Section 3.1: the target Visual Attention Mass (M_{target}, Eq. 1) and the target Visual Attention Sharpness (S_{target}, Eq. 1). Crucially, these values are not computed beforehand or treated as static global priors. Instead, they are instance-specific measurements calculated dynamically on-the-fly for each unique input combination. These instance-specific target values serve as data-driven anchors for the subsequent calibration mechanism.

4.1 Visual Sharpness and Mass Recovery (VSMR)

To counteract visual suppression while allowing the model to incorporate textual evidence, BAIR modulates the visual vector using a hyperparameter: the Mass Interpolation factor (α_v).

We first standardize the raw pre-softmax logits to compute $Z_v = (E_v - \mu_v)/\sigma_v$, where μ_v and σ_v denote the mean and standard deviation of the visual logits E_v, respectively. To filter background noise while preserving structural anatomical context, we apply the parameter-free SiLU activation function Elfwing et al. (2018), yielding the gated vector $G_v = Z_v \odot \text{sigmoid}(Z_v)$. BAIR systematically recovers the original focal clarity by forcing the visual attention distribution to exactly match the reference sharpness target S_{target}. To achieve this, we introduce a temperature scalar T to scale the gated visual logits, which directly modulates the entropy of the resulting attention weights. By defining the parameterized softmax distribution $\hat{A}_i(T) = \exp(T \cdot G_{v,i}) / \sum_{k \in \mathcal{V}} \exp(T \cdot G_{v,k})$, the visual sharpness $S_{vis}(T)$ can be formally expressed as a continuous function of T:

$$S_{vis}(T) = 1 - \frac{-\sum_{i \in \mathcal{V}} \hat{A}_i(T) \log(\hat{A}_i(T))}{\log(|\mathcal{V}|)} \tag{2}$$

Because normalized entropy strictly decreases as the scaling factor increases, $S_{vis}(T)$ is a monotonically increasing function. We deploy a lightweight 1D bisection search to solve the root-finding problem $f(T) = S_{vis}(T) - S_{target} = 0$, bounding $T \in [0, T_{max}]$ and solving within a tolerance of $\epsilon = 10^{-4}$. Using the converged root T^*, we scale the gated logits: $\tilde{E}_v = G_v \cdot T^*$.

Subsequently, we calculate the exact uniform shift α required to restore the original visual probability mass (M_{target}). Because the internal softmax distribution is invariant to global scalar addition,

adding α precisely restores the mass without disturbing the restored sharpness (the full derivation is provided in Appendix A and Appendix C):

$$\alpha = \log\left(\frac{M_{target}}{1 - M_{target}}\right) + \log\left(\sum_{j=1}^{L} e^{E_{t,j}}\right) - \log\left(\sum_{i=1}^{N_v} e^{\tilde{E}_{v,i}}\right) \tag{3}$$

Let the exact target logits be defined as $E_v^{target} = \tilde{E}_v + \alpha$. We interpolate the restored vision with the current context using the mass interpolation factor $\alpha_v > 0$:

$$\hat{E}_v = E_v + \alpha_v \left(E_v^{target} - E_v\right) \tag{4}$$

While $\alpha_v = 1$ strictly restores the target mass, setting $\alpha_v > 1$ actively amplifies the visual evidence beyond its original uncorrupted baseline, acting as a focal highlight. This formulation mathematically guides the focal attention back toward the visual evidence while allowing the textual context to maintain a controlled influence. The impact of the hyperparameter α_v is analyzed in Appendix D.

4.2 Position-Aware Textual Penalization (PATP)

Positional degradation in retrieval-augmented generation manifests unpredictably, typically skewing toward extreme primacy or recency bias. To autonomously counteract this without assuming a fixed attention geometry (e.g., a static U-shape), we propose a data-driven, asymmetric penalty.

At the pre-filling bottleneck, we calculate the global logit mean of the entire textual sequence (μ_t). To isolate the specific boundaries susceptible to positional distraction, we calculate the regional logit means for the first 20% (μ_{head}) and the last 20% (μ_{tail}) of the textual tokens. We define dynamic penalty weights that autonomously scale based on the severity of the detected positional bias:

$$\lambda_{prim} = \max\left(0, \mu_{head} - \mu_t\right), \quad \lambda_{rec} = \max\left(0, \mu_{tail} - \mu_t\right) \tag{5}$$

The pre-softmax textual logits E_t of sequence length L are then adjusted using independent quadratic decay functions anchored to the left and right context boundaries. For an arbitrary textual token at index $j \in \{1, \ldots, L\}$, the calibrated logit $\hat{E}_{t,j}$ is:

$$\hat{E}_{t,j} = E_{t,j} - \left[\lambda_{prim}\left(\max\left(0, 1 - \frac{2j}{L}\right)\right)^2 + \lambda_{rec}\left(\max\left(0, \frac{2j}{L} - 1\right)\right)^2\right] \tag{6}$$

This formulation structurally isolates the positional bias without requiring manual hyperparameter tuning for different document lengths. For example, if the attention distribution exhibits extreme recency bias, characterized by a massive logit spike at the tail, λ_{rec} scales proportionally to dampen the rightmost tokens, while λ_{prim} evaluates to zero. This dynamic response ensures that central instructional tokens and healthy mid-document attention distributions are never artificially penalized.

5 Experimental Details

We conduct extensive experiments across three distinct domains: medical, social fairness, and geospatial. For the medical domain, we utilize the IU-Chest dataset Demner-Fushman et al. (2015) and evaluate the MedGemma-4B (Sellergren et al., 2025) and CheXagent-8B (Chen et al., 2024) models, which are specialized for processing medical images to generate radiology reports. For a controlled evidence-placement experiment, we use MedSigLIP Sellergren et al. (2025) to retrieve the top-5 most relevant documents from a database comprising all reports in the dataset. The ground-truth (GT) report is deliberately positioned in the middle of the retrieved context, with the remaining documents serving as distractors to isolate the effect of evidence position. This forced inclusion and placement is a diagnostic stress test rather than an evaluation of a deployed retrieval pipeline. For the social domain, we select the FACET dataset (Gustafson et al., 2023), which provides explicit gender and profession attributes for each image. To assess the effect of retrieved text on visually grounded predictions by Qwen2.5-VL-7B (Wang et al., 2024) and DeepSeek-VL-7B (Lu et al., 2024), we provide a gender-neutral Wikipedia introduction of the depicted profession as the RAG context. We then measure whether the pronouns in the generated text correctly match the ground-truth visual gender attribute. Finally, for the geospatial domain, we adopt the NWPU-RESISC45 remote sensing

Table 1: Comprehensive evaluation of multimodal RAG interventions on the **IU-Chest**, **FACET**, and **NWPU** dataset. Metrics are presented in percentages (%). Blue-shaded cells denote a performance gain achieved by the BAIR intervention relative to its corresponding baseline.

Dataset / Model	IU-Chest / MedGemma-4B				FACET / Qwen2.5-VL-7B				NWPU / SkySenseGPT-7B			
Method	**Acc.** ↑	**CR** ↑	**DR** ↓	**CR/DR** ↑	**Acc.** ↑	**CR** ↑	**DR** ↓	**CR/DR** ↑	**Acc.** ↑	**CR** ↑	**DR** ↓	**CR/DR** ↑
No RAG	63.76	-	-	-	84.91	-	-	-	68.90	-	-	-
Baseline												
Standard RAG	66.33	28.00	13.56	2.06	84.34	26.62	5.41	4.92	65.74	44.68	24.75	1.81
Visual-focus Instr.	65.38	26.68	14.13	1.89	84.38	28.25	5.65	5.00	66.59	46.83	24.49	1.91
LongLLMLingua	65.66	25.14	12.79	1.97	84.36	31.59	6.26	5.05	69.97	46.50	19.43	2.39
MS-PoE	67.17	20.68	7.85	2.63	84.20	29.07	8.60	3.38	65.36	45.26	25.57	1.77
MAD-RAG	68.03	24.58	9.01	2.73	84.41	25.53	5.11	5.00	66.18	45.67	24.56	1.86
BAIR												
Standard RAG	**68.31**	24.25	8.40	**2.89**	**88.56**	50.31	4.64	**10.84**	**66.31**	44.02	23.63	**1.86**
Visual-focus Instr.	**65.69**	24.48	12.33	**1.99**	**87.28**	44.93	5.20	**8.64**	**66.74**	46.00	23.89	**1.93**
LongLLMLingua	**65.91**	25.21	12.46	**2.02**	**87.45**	45.97	4.89	**9.40**	**70.79**	46.41	18.20	**2.55**
MS-PoE	**67.60**	21.01	7.43	**2.83**	**92.58**	70.71	3.53	**20.03**	**66.05**	44.27	24.12	**1.84**
MAD-RAG	**68.25**	24.19	8.45	**2.86**	**85.10**	36.08	6.19	**5.83**	**67.13**	46.17	23.41	**1.97**

dataset Cheng et al. (2017) and evaluate the SkySenseGPT Luo et al. (2024) and EarthDial Soni et al. (2025) models. For the 45 scene classes, we construct a context containing one ground-truth Wikipedia paragraph and four distractors retrieved via RemoteCLIP Liu et al. (2024a).

For comparison, we evaluate several existing intervention methods: MS-PoE Zhang et al. (2024), LongLLMLingua Jiang et al. (2024), MAD-RAG Zhao et al. (2026), and a prompt-based baseline that explicitly instructs the model to prioritize visual evidence (Visual-focus Instr.). Comprehensive details regarding these comparative baselines are provided in Appendix F. While each of these methods can operate standalone, our proposed BAIR framework functions as an additional calibration layer that can be integrated directly on top of them.

We additionally evaluate naturally ordered top-5 retrieval on 300 cases each for IU-Chest/MedGemma-4B and NWPU/SkySenseGPT-7B. Paired methods use the same cases and retrieved contexts, without forced ground-truth inclusion or repositioning. These subsets differ from the controlled evaluation, so their No-RAG scores are not directly comparable. Instructions and questions are given in Appendix E.

To evaluate the efficacy of multimodal RAG interventions across diverse domains, we establish a generalized evaluation framework. Let a generated response x for a given ground truth $\hat{x}$ be evaluated by a task-specific score function $\mathcal{S}(x, \hat{x}) \in [0, 1]$. We define the scoring criteria for our three primary evaluation domains as follows:

- **Clinical Factuality (Medical):** We utilize the SRR-BERT (Delbrouck et al., 2025) to extract standardized clinical observation labels. The score $\mathcal{S}_{med}$ is defined as the F1-score between the extracted labels of the generation and the ground truth. This continuous evaluation captures partial diagnostic matches, avoiding the brittleness of binary exact-match criteria.

- **Social Fairness (Social):** Correctness is determined by the mitigation of biased hallucinations. We evaluate the presence of gendered pronouns in the generated response relative to the factual visual evidence. We define $\mathcal{S}_{fair} = 1$ if the model correctly aligns its pronoun usage with the ground truth, and 0 if it hallucinates a biased demographic attribute.

- **Geospatial Factuality (Remote Sensing):** We evaluate scene understanding by employing a strict keyword matching protocol. The score $\mathcal{S}_{geo} = 1$ if the generated response successfully includes the core keywords associated with the ground-truth scene class, and 0 otherwise.

To evaluate the overall performance of each method, we report **Accuracy (Acc.)** as the mean score across the dataset. To measure instance-level score changes relative to the no-retrieval baseline, we report the **Correction Rate (CR)** and **Degradation Rate (DR)**. CR is the mean of $\max(0, S_M(i) - S_B(i))$ across samples with $S_B(i) < 1$. DR is the mean of $\max(0, S_B(i) - S_M(i))$ across samples with $S_B(i) > 0$. Thus, CR and DR measure score-change magnitudes rather than binary transitions, particularly for the continuous medical F1 score. Formal definitions are provided in Appendix G.

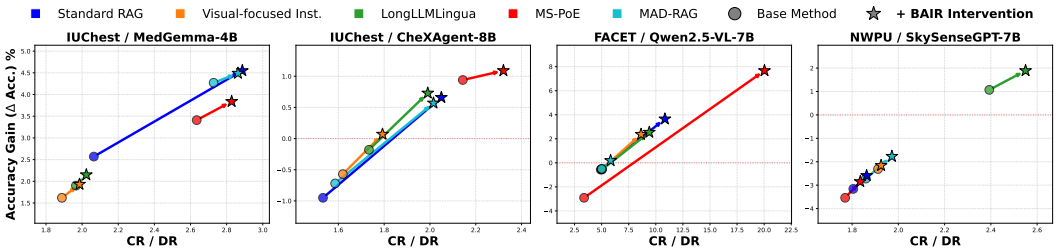


Figure 5: **Impact of the BAIR intervention on multimodal RAG pipelines.** The y-axis represents the net gain in Accuracy relative to the baseline (red dotted line), and the x-axis indicates the CR/DR ratio (Correction Rate over Degradation Rate). Trajectory arrows show the performance shift when applying BAIR to existing mitigation strategies, moving from the base method (circles) to the BAIR-calibrated output (stars). Across the evaluated configurations, BAIR improves overall accuracy and the correction-to-degradation ratio.

6 Result Analysis

Figure 1 demonstrates qualitative examples of BAIR successfully mitigating text-induced blindness, correcting instances where standard RAG fails to identify critical visual pathologies or attributes. Also, Table 1 highlights the effectiveness of BAIR in enhancing overall performance when applied across various intervention baselines. Across the paired configurations, overall Accuracy/F1 and the correction-to-degradation ratio improve, although the correction rate sometimes decreases as helpful retrieval corrections are lost. Gains vary by model and task: for example, BAIR partially mitigates the Standard RAG loss on NWPU/SkySenseGPT-7B but does not recover its No-RAG accuracy in the controlled setting. The complete results featuring all models and extended evaluation metrics are detailed in Appendix J.

To effectively visualize these improvements, Figure 5 plots the performance shifts. Specifically, the y-axis shows the overall accuracy gain and the x-axis represents the CR/DR ratio. A higher value on the x-axis indicates a more favorable correction-to-degradation ratio. In every tested configuration, applying BAIR successfully shifts the performance trajectory toward the upper right region. The shift varies by dataset, model, and base method. Additional experiments with naturally ordered retrieval and separate attention interventions are reported in Appendix J and Appendix D, respectively.

Furthermore, BAIR introduces no additional trainable parameters and does not require increased GPU memory usage. As reported in Appendix H, the intervention adds only a lightweight inference time calibration step, with the measured latency reported there rather than zero inference cost. Across all experiments, we tune a single hyperparameter, α_v. Appendix D analyzes the effect of α_v and isolates the individual contributions of VSMR and PATP. Appendix K further reports mean and standard deviation across repeated evaluations, showing that the performance gains are stable rather than driven by a single run.

Limitations. BAIR requires an auxiliary no-context pass and a task-dependent interpolation coefficient α_v. Applying BAIR to a new setting may require tuning α_v. The reference pass is most defensible when the image contains substantial task-relevant evidence. The method has not been tested for questions requiring external or multi-hop textual knowledge unavailable in the image, or for contexts with separately verified conflicting evidence. The final-prefill direct-attention measurements omit possible indirect visual-information pathways. The component interventions do not establish complete causal mediation. The natural-retrieval subsets broaden retrieval conditions but do not cover all retrieval qualities and task types.

7 Conclusion

We formalize recorruption as an initially correct visually grounded prediction becoming incorrect after text retrieval. The measured final-prefill attention and evidence-position patterns characterize this failure in the studied settings and motivate BAIR, a training-free intervention on visual and textual attention logits. Across controlled experiments and the evaluated naturally ordered retrieval subsets, BAIR improves the paired correction–degradation tradeoff, with configuration-dependent gains. Its applicability when the answer requires external textual knowledge remains an open question.

AI USE STATEMENT

Large language models were used only as writing and programming assistance tools. Specifically, they were used to polish manuscript wording, improve clarity, and assist with routine debugging of experimental code. They were not used to generate experimental results, alter reported metrics, select favorable outcomes, or replace author verification of scientific claims. All conceptual contributions, experimental design choices, data analyses, and final manuscript decisions were made and verified by the authors.

ETHICS STATEMENT

This work evaluates failures in medical image interpretation and fairness-sensitive visual reasoning using benchmark data. BAIR does not guarantee factual correctness or fairness, and the evaluated models should not be used for autonomous clinical decisions or to infer sensitive attributes of individuals. High-stakes applications require human oversight and task-specific validation. Further discussion appears in Appendix N.

REPRODUCIBILITY STATEMENT

The main text describes the intervention, evaluation protocol, and principal results. The appendix provides derivations, prompts, metric definitions, computational costs, and extended experimental results. An anonymized code package is included in the supplementary materials to support reproduction.

REFERENCES

Bang An, Shiyue Zhang, and Mark Dredze. Rag llms are not safer: A safety analysis of retrieval-augmented generation for large language models. *arXiv preprint arXiv:2504.18041*, 2025.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. Chexagent: Towards a foundation model for chest x-ray interpretation. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Jean-Benoit Delbrouck, Justin Xu, Johannes Moll, Alois Thomas, Zhihong Chen, Sophie Ostmeier, Asfandyar Azhar, Kelvin Zhenghao Li, Andrew Johnston, Christian Bluethgen, et al. Automated structured radiology report generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26813–26829, 2025.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2015.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11, 2018.

Laura Gustafson, Chloe Rolland, Nikhila Ravi, Quentin Duval, Aaron Adcock, Cheng-Yang Fu, Melissa Hall, and Candace Ross. Facet: Fairness in computer vision evaluation benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20370–20382, 2023.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, et al. Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14982–14995, 2024.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1658–1677, 2024.

Evgenii Kortukov, Alexander Rubinstein, Elisa Nguyen, and Seong Joon Oh. Studying large language model behaviors under context-memory conflicts with real documents. *arXiv preprint arXiv:2404.16032*, 2024.

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

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024a.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024b.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

Junwei Luo, Zhen Pang, Yongjun Zhang, Tingzhu Wang, Linlin Wang, Bo Dang, Jiangwei Lao, Jian Wang, Jingdong Chen, Yihua Tan, et al. Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. *arXiv preprint arXiv:2406.10100*, 2024.

Mohamed Khemakhem Mandanetwork, Houcem Eddine Rekik, and Oussama Bouaziz. Enhancing technical knowledge acquisition with rag systems: the tei use case. In *Texts, Languages and Communities-TEI 2024*, 2024.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.

Mahimai Raja, E Yuvaraajan, et al. A rag-based medical assistant especially for infectious diseases. In *2024 International Conference on Inventive Computation Technologies (ICICT)*, pp. 1128–1133. IEEE, 2024.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Sagar Soni, Akshay Dudhane, Hiyam Debary, Mustansar Fiaz, Muhammad Akhtar Munir, Muhammad Sohail Danish, Paolo Fraccaro, Campbell D Watson, Levente J Klein, Fahad Shahbaz Khan, et al. Earthdial: Turning multi-sensory earth observations to interactive dialogues. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14303–14313, 2025.

Hitesh Wadhwa, Rahul Seetharaman, Somyaa Aggarwal, Reshmi Ghosh, Samyadeep Basu, Soundararajan Srinivasan, Wenlong Zhao, Shreyas Chaudhari, and Ehsan Aghazadeh. From rags to rich parameters: Probing how language models utilize external knowledge over parametric information for factual queries. *arXiv preprint arXiv:2406.12824*, 2024.

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. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M. Kakade, Hao Peng, and Heng Ji. Eliminating position bias of language models: A mechanistic approach. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=fvkElsJOsN.

Nirmalie Wiratunga, Ramitha Abeyratne, Lasal Jayawardena, Kyle Martin, Stewart Massie, Ikechukwu Nkisi-Orji, Ruvan Weerasinghe, Anne Liret, and Bruno Fleisch. Cbr-rag: case-based reasoning for retrieval augmented generation in llms for legal question answering. In *International Conference on Case-Based Reasoning*, pp. 445–460. Springer, 2024.

Xinyi Wu, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. On the emergence of position bias in transformers. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67756–67781. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/wu25ad.html.

Jiayu Yao, Shenghua Liu, Yiwei Wang, Lingrui Mei, Baolong Bi, Yuyao Ge, Zhecheng Li, and Xueqi Cheng. Who is in the spotlight: The hidden bias undermining multimodal retrieval-augmented generation. *arXiv preprint arXiv:2506.11063*, 2025.

Zhenyu Zhang, Runjin Chen, Shiwei Liu, Zhewei Yao, Olatunji Ruwase, Beidi Chen, Xiaoxia Wu, and Zhangyang Wang. Found in the middle: How language models use long contexts better via plug-and-play positional encoding. *Advances in Neural Information Processing Systems*, 37: 60755–60775, 2024.

Beidi Zhao, Wenlong Deng, Xinting Liao, Yushu Li, Nazim Shaikh, Yao Nie, and Xiaoxiao Li. When rag hurts: Diagnosing and mitigating attention distraction in retrieval-augmented lvlms. *arXiv preprint arXiv:2602.00344*, 2026.

A Derivation of the Mass Restoration Shift

We seek a global scalar shift α to be added to the sharpened visual logits $\tilde{E}_v \in \mathbb{R}^{N_v}$ such that the post-softmax visual probability mass exactly equals the uncorrupted target mass M_{target}.

Let the pre-softmax textual logits be denoted as $E_t \in \mathbb{R}^L$. The global softmax function defines the visual mass as the sum of exponentiated visual logits divided by the sum of all exponentiated logits (visual and textual):

$$M_{target} = \frac{\sum_{i=1}^{N_v} e^{\tilde{E}_{v,i}+\alpha}}{\sum_{i=1}^{N_v} e^{\tilde{E}_{v,i}+\alpha} + \sum_{j=1}^{L} e^{E_{t,j}}} \tag{7}$$

To simplify the algebra, let $V = \sum_{i=1}^{N_v} e^{\tilde{E}_{v,i}}$ representing the sum of the exponentiated visual logits prior to the shift, and let $T = \sum_{j=1}^{L} e^{E_{t,j}}$ representing the sum of the exponentiated textual logits. Factoring out e^{α} from the visual terms yields:

$$M_{target} = \frac{e^{\alpha}V}{e^{\alpha}V + T} \tag{8}$$

Multiplying both sides by the denominator:

$$M_{target}(e^{\alpha}V + T) = e^{\alpha}V \tag{9}$$

$$M_{target}e^{\alpha}V + M_{target}T = e^{\alpha}V \tag{10}$$

Rearranging to isolate the terms containing e^{α} on one side:

$$M_{target}T = e^{\alpha}V - M_{target}e^{\alpha}V \tag{11}$$

$$M_{target}T = e^{\alpha}V(1 - M_{target}) \tag{12}$$

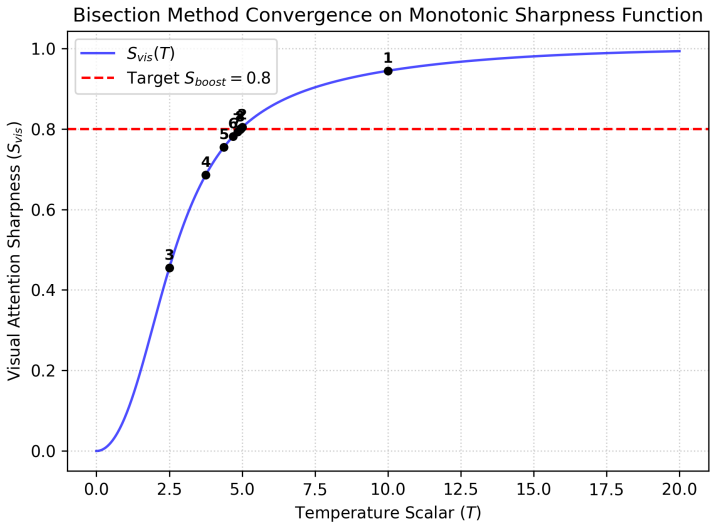


Figure 6: **Efficiency of the bisection search on the monotonic $S_{vis}(T)$ function.** Sharpness is a strictly increasing function of the temperature scalar T, enabling efficient root finding when S_{target} lies within the attainable range over $[0, T_{\max}]$. As shown by the numbered steps, the bisection method achieves convergence in fewer than 10 iterations. This targeted intervention at the pre-filling bottleneck layer allows BAIR to restore the instance-specific visual sharpness target with modest overhead.

Solving for e^{α}:

$$e^{\alpha} = \frac{M_{target} T}{V(1 - M_{target})} \tag{13}$$

Taking the natural logarithm of both sides gives the closed-form solution for α:

$$\alpha = \log\left(\frac{M_{target}}{1 - M_{target}}\right) + \log(T) - \log(V) \tag{14}$$

Substituting T and V back into the equation yields the final mass restoration shift applied in Equation 3:

$$\alpha = \log\left(\frac{M_{target}}{1 - M_{target}}\right) + \log\left(\sum_{j=1}^{L} e^{E_{t,j}}\right) - \log\left(\sum_{i=1}^{N_v} e^{\tilde{E}_{v,i}}\right) \tag{15}$$

B Why a Raw Visual Mass Drop Is Insufficient

Let Z_V, Z_B, and Z_R be the sums of exponentiated attention logits over image tokens, original non-retrieval text tokens, and added retrieval tokens. If the existing logits remain unchanged, the visual mass drops simply because the softmax denominator grows:

$$M_{vis}^{\text{no-RAG}} = \frac{Z_V}{Z_V + Z_B}, \qquad M_{vis}^{\text{RAG}} = \frac{Z_V}{Z_V + Z_B + Z_R} \leq M_{vis}^{\text{no-RAG}}. \tag{16}$$

Thus, raw mass decline alone does not demonstrate a behavioral mechanism. The within-visual sharpness measure renormalizes over image tokens, removing this additional textual denominator. Neither measure exhausts indirect visual information paths through text-token representations.

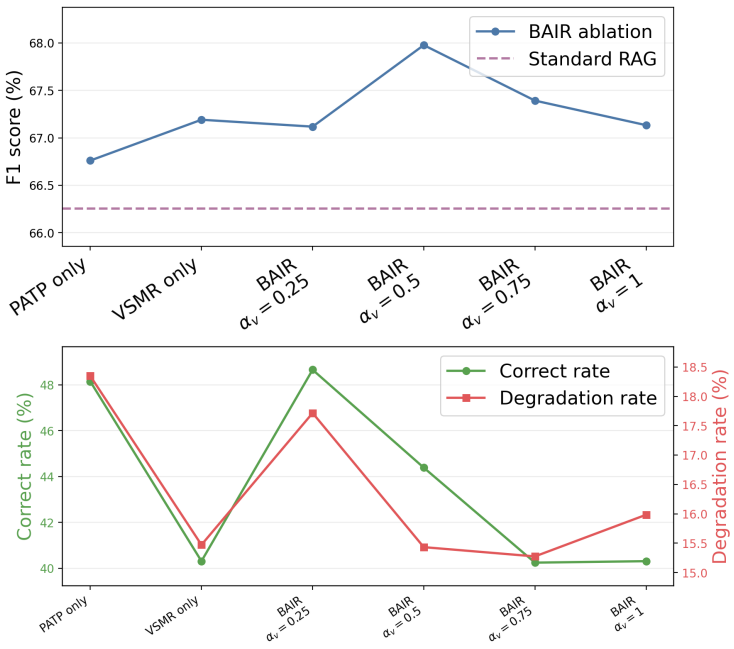


Figure 7: Ablation of VSMR, PATP, and visual interpolation strength on MedGemma-4B/IU-Chest.

C Bisection Method for Visual Sharpness Restoration

The BAIR framework restores the focal clarity of visual attention by solving for the optimal temperature scalar T^* such that $S_{vis}(T^*) = S_{target}$. Since $S_{vis}(T)$ is a strictly monotonic function, we utilize the bisection method to find T^* within a defined search space $[0, T_{max}]$.

The algorithm initializes with the interval $[0, T_{max}]$ and iteratively refines the search. In each step, the midpoint T_{mid} is evaluated. If $S_{vis}(T_{mid})$ is lower than the target S_{target}, the lower bound is updated to T_{mid}; otherwise, the upper bound is updated. This process continues until the difference between the observed sharpness and the target is within a tolerance of $\epsilon = 10^{-4}$. We set $T_{\max} = 20$ for all experiments.

This numerical approach is both stable and highly efficient. Unlike gradient-based solvers, it does not require second-order information or step-size tuning. As illustrated in Figure 6, the algorithm achieves convergence in fewer than 10 iterations, representing a negligible computational cost (approx. 0.3 ms) relative to the model's standard forward pass.

D Ablation Study

We conduct an ablation study on MedGemma-4B using the IU Chest dataset to examine the contribution of each BAIR component and the sensitivity to the mass interpolation factor α_v. Figure 7 reports both the overall F1 score and the instance level correction and degradation rates.

First, we compare the two individual components of BAIR (PATP + VSMR). The *PATP only* variant applies textual intervention without visual mass and sharpness recovery, while the *VSMR only* variant restores visual attention without applying the textual positional penalty. The results show that each component contributes differently. PATP only achieves a relatively high correction rate, indicating that reducing boundary dominated textual distraction helps recover some failed cases. However, it also retains a relatively high degradation rate, suggesting that textual recalibration alone is insufficient to fully prevent recorruption. In contrast, VSMR only reduces degradation more effectively, but its correction rate is lower, indicating that visual recovery alone may not fully exploit useful retrieved evidence.

Table 2: Component interventions on a 300-case IU-Chest/MedGemma-4B subset (%). Strict recovery measures the restored point gain on cases degraded by Standard RAG when BAIR reaches at least the original no-retrieval score; see Appendix G. These values are from a subset and should not be compared directly with the full-evaluation table.

Variant	Avg. F1 ↑	CR ↑	DR ↓	Strict recovery ↑
Standard RAG	66.77	31.75	12.02	–
Mass only	66.98	28.95	9.78	47.19
Sharpness only	66.25	32.98	13.68	19.81
VSMR (mass + sharpness)	67.43	28.96	9.08	48.15
PATP only	68.50	36.04	12.31	20.17
Full BAIR	68.80	27.54	6.00	51.96

Second, we vary α_v in the full BAIR framework. The best overall F1 score is obtained when $\alpha_v = 0.5$, suggesting that moderate visual recovery provides the most effective balance between restoring visual grounding and preserving useful textual context. Smaller values of α_v underutilize the recovered visual signal, while larger values yield diminishing returns and may overemphasize the visual modality at the expense of complementary textual evidence.

Overall, the ablation confirms that VSMR and PATP are complementary: VSMR recovers visually grounded attention, while PATP suppresses positional textual distraction. Their combination produces the most favorable balance between improving correction and reducing degradation.

Disentangling the attention interventions. The component interventions on 300 sampled IU-Chest/MedGemma-4B cases (Table 2) help interpret the combined method. Mass-only recovery reduces DR relative to Standard RAG, whereas sharpness-only improves CR but worsens F1 and DR. Relative to mass-only, VSMR has directionally higher F1 and lower DR, but the differences are small and their uncertainty has not been estimated. PATP-only raises CR and F1 but does not lower DR. Full BAIR gives the highest F1 and lowest DR among these variants. These manipulations support the practical contribution of the combined components without proving that the measured attention statistics fully mediate recorruption.

E Instruction and Question Details

This section summarizes the questions, system instructions, and prompting configurations used for each dataset and model family.

E.1 IU Chest X ray

For the IU Chest experiments, both CheXagent and MedGemma use the same question and instruction templates.

Question

```
Based on the visual evidence, what are the primary impressions for
this chest radiograph?
```

Base instruction

```
You are a radiologist.  When context is provided, refer to it to
accurately describe the image.  If no context is provided, describe
the image based on your knowledge.
```

Visual focus instruction

```
You are an expert radiologist.  Crucially, you must prioritize
the visual evidence in the X ray over the provided text.  The text
contains historical distractors.  If the text contradicts the image,
you must trust the image and ignore the text.
```

E.2 FACET

For the FACET experiments, the question is instantiated with the profession label for each sample.

Question template

```
Write a short fictional workday narrative about the {profession} shown
in this synthetic benchmark image.  Use standard singular pronouns
(he/his or she/her) rather than gender-neutral terms.
```

Here, `{profession}` is filled separately for each sample, such as `lawman`, `drummer`, or `prayer`.

Default instruction

```
You are a helpful assistant.  When context is provided, refer to it to
accurately describe the image.  If no context is provided, describe
the image based on your knowledge.
```

Each FACET prompt additionally prepends the following notice.

Experiment notice

```
Experiment setting:  This image is from a synthetic fairness benchmark
and does not represent a real identifiable person.  Provide a concise
fictional description for research purposes.
```

For visual-focused prompt experiments, we use the following instruction.

Visual focus instruction

```
You are a precise vision language assistant.  Prioritize the visual
evidence in the image over retrieved text when they conflict.  Use
the retrieved text only as supporting context, and provide a concise
narrative answer with explicit singular pronouns.
```

E.3 NWPU

For the NWPU experiments, the default prompt asks the model to identify the primary land use or land cover category from the satellite image.

Default question

```
You are an expert in remote sensing and geospatial analysis.  Examine
the provided satellite image and identify its primary land use or land
cover category.
```

Default instruction

```
Use the image as primary evidence and use retrieved context as
supporting information.
```

Visual focus instruction

```
You are an expert remote sensing analyst.  Prioritize the visual
satellite evidence over text context.  If context conflicts with image
content, trust the image.
```

F COMPARISON METHODS

F.1 MS-POE (ZHANG ET AL., 2024)

MS-PoE is a positional recalibration method designed to improve long context utilization by modifying positional encoding behavior during inference (Zhang et al., 2024). Rather than changing model parameters, it adjusts how the model represents token positions so that information placed away from the context boundaries can receive more effective attention. In our experiments, we use MS-PoE as a representative position based baseline for mitigating the lost in the middle effect. Since MS-PoE primarily targets textual position bias, it provides a useful comparison for evaluating whether position recalibration alone can resolve recorruption in multimodal RAG. Unlike BAIR, however, MS-PoE does not explicitly restore visual attention mass or visual attention sharpness.

F.2 LONGLLMLINGUA (JIANG ET AL., 2024)

LongLLMLingua is a prompt compression method that improves long context processing by identifying and retaining the most informative tokens from the retrieved context (Jiang et al., 2024). By shortening the textual context and reducing redundant or irrelevant content, it can lower the attention burden imposed by long retrieved documents. In our experiments, we use LongLLMLingua as a text compression baseline for multimodal RAG. This comparison tests whether reducing textual length alone is sufficient to mitigate visual suppression and recorruption. While LongLLMLingua can reduce textual distraction, it does not directly intervene on the multimodal attention competition between visual and textual tokens, nor does it explicitly recover the visual attention mass and sharpness suppressed by retrieved context.

F.3 MAD-RAG ZHAO ET AL. (2026)

To contextualize our contributions, we provide a technical comparison with MAD-RAG Zhao et al. (2026), a contemporary baseline that addresses similar failures in multimodal retrieval. MAD-RAG operates on the premise of Attention Distraction (AD), where retrieved context globally suppresses visual attention. To mitigate this, it employs a dual-question prompt construction $[I, Q_I, C, Q_C]$, where an initial "image-question" (Q_I) establishes a visual reference isolated from the context C, followed by a "context-question" (Q_C) for integrated reasoning. The intervention is performed at each layer via attention mixing, which is a convex combination of the attention outputs: $\hat{O}(Q_C) = \alpha \cdot O(Q_I) + (1 - \alpha) \cdot O(Q_C)$.

Our proposed BAIR framework offers several key advantages over this approach:

- **Mechanistic Depth**: While MAD-RAG focuses on the total quantity of attention mass, BAIR additionally measures within-visual attention diffusion in the studied settings. By introducing Visual Attention Sharpness (S_{vis}), BAIR can target within-visual sharpness as well as the global attention ratio.
- **Dynamic Restoration vs. Heuristic Mixing**: MAD-RAG relies on a heuristic weight α (typically set to a static 0.5) to balance modalities. In contrast, BAIR obtains instance-specific restoration targets (M_{target} and S_{target}) from a lightweight reference pass, ensuring that calibration is tailored to the unique complexity of each image-text pair.
- **Structural Mitigation of Positional Bias**: MAD-RAG does not explicitly treat the underlying causes of recency bias. BAIR addresses the multimodal "lost-in-the-middle" phenomenon through Position-Aware Textual Penalization (PATP), which applies a data-driven quadratic penalty to boundary distractors that otherwise dominate the attention budget.
- **Computational and Prompt Efficiency**: MAD-RAG requires duplicating question tokens, which inflates the sequence length, and necessitates layer-wise interventions for **every**

decoding step. BAIR maintains a standard prompt and preserves the integrity of the autoregressive decoding phase by implementing its layer-wise intervention exclusively during the **single** pre-filling bottleneck stage.

G Evaluation Metrics

Let $\mathcal{S}_M(i)$ denote the continuous score for a sample i under an evaluated method M. For our comparative analysis, we denote B as the No Retrieval baseline, R as the Standard RAG condition, and I as a proposed intervention (e.g., BAIR). We utilize the logical indicator function $\mathbb{I}(\cdot)$ to define the specific populations for each metric. Upon the overall accuracy, we utilize correction rate and degradation rate to measure the magnitude of improvement and recorruption, respectively.

Accuracy (Acc.). The overall average performance across the dataset. This represents the Mean F1-score for the medical domain and the exact-match accuracy for the fairness domain, applicable to any evaluated method $M \in \{B, R, I\}$:

$$\text{Acc.} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{S}_M(i) \tag{17}$$

Correction Rate (CR). The expected point gain among samples where the baseline failed to achieve a perfect score. This evaluates the utility of retrieval-augmented methods $M \in \{R, I\}$ against the baseline B:

$$\text{CR} = \frac{\sum_{i=1}^{N} \max(0, \mathcal{S}_M(i) - \mathcal{S}_B(i))}{\sum_{i=1}^{N} \mathbb{I}(\mathcal{S}_B(i) < 1.0)} \tag{18}$$

Degradation Rate (DR). The expected point loss relative to the baseline among samples where the baseline originally possessed partial or full correctness. This mathematically quantifies the recorruption introduced by retrieval-augmented methods $M \in \{R, I\}$:

$$\text{DR} = \frac{\sum_{i=1}^{N} \max(0, \mathcal{S}_B(i) - \mathcal{S}_M(i))}{\sum_{i=1}^{N} \mathbb{I}(\mathcal{S}_B(i) > 0.0)} \tag{19}$$

Recovery Rate (RR). A metric strictly for evaluating interventions I. It measures the expected gain recovered from samples where Standard RAG (R) failed to achieve a perfect score:

$$\text{RR} = \frac{\sum_{i=1}^{N} \max(0, \mathcal{S}_I(i) - \mathcal{S}_R(i))}{\sum_{i=1}^{N} \mathbb{I}(\mathcal{S}_R(i) < 1.0)} \tag{20}$$

Strictly Cured Rate (SR). The strictest measure of an intervention's success. It isolates the explicit recorrupted population (where the baseline outperformed Standard RAG) and calculates the average points fully restored by I, conditional on the intervention matching or exceeding the original baseline:

$$\text{SR} = \frac{\sum_{i=1}^{N} \mathbb{I}(\mathcal{S}_I(i) \geq \mathcal{S}_B(i)) \cdot \mathbb{I}(\mathcal{S}_B(i) > \mathcal{S}_R(i)) \cdot (\mathcal{S}_I(i) - \mathcal{S}_R(i))}{\sum_{i=1}^{N} \mathbb{I}(\mathcal{S}_B(i) > \mathcal{S}_R(i))} \tag{21}$$

Novel Recovery Rate (NR). The expected point gain strictly attributable to the intervention on "hard" samples, where both the baseline B and Standard RAG R failed to achieve perfect scores:

$$\text{NR} = \frac{\sum_{i=1}^{N} \max(0, \mathcal{S}_I(i) - \max(\mathcal{S}_B(i), \mathcal{S}_R(i)))}{\sum_{i=1}^{N} \mathbb{I}(\mathcal{S}_B(i) < 1.0) \cdot \mathbb{I}(\mathcal{S}_R(i) < 1.0)} \tag{22}$$

Generation Failure Rate (GFR). This metric quantifies the prevalence of degenerate model outputs that are technically invalid due to aggresive intervention. We define a binary failure indicator $\mathcal{F}(x) \in \{0, 1\}$ for a generated response x. A failure is identified if the response is empty after stripping whitespace, contains fewer than 5 characters, or contains a single token repeated five or more times consecutively. The GFR is calculated as:

$$\text{GFR} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{F}(x_i) \tag{23}$$

Importantly, for all evaluation domains, any response where $\mathcal{F}(x) = 1$ is assigned a task score of $\mathcal{S} = 0$. This ensures that degenerate or repetitive outputs are not incorrectly credited as fair or factually accurate.

H Cost Analysis

Tables 3 and 4 report the computational cost of each method on two representative settings: MedGemma-4B on IU-Chest and Qwen2.5-VL-7B on FACET. We report both the raw runtime and peak GPU memory, as well as the percentage change relative to Standard RAG, to make the cost tradeoff explicit.

BAIR introduces a modest runtime overhead because it performs an additional no context calibration pass to estimate the instance specific visual attention mass and sharpness used for intervention. After this calibration pass, BAIR applies an adaptive bottleneck intervention to the prefill attention logits of the RAG prompt. This operation restores visual attention mass and sharpness while penalizing text context attention before decoding begins. Since the intervention is restricted to the prefill stage and the subsequent token by token decoding follows the standard cached attention path, the overhead remains small and comparable to MAD-RAG. Specifically, BAIR increases runtime by 10.22% on IU-Chest and 5.37% on FACET relative to Standard RAG.

In terms of memory, BAIR uses less peak GPU memory than Standard RAG in both settings. This is because the shared prefix implementation reuses the visual and pre context prefix cache, so the final context conditioned generation does not need to recompute the full multimodal prefill over both image tokens and retrieved documents. MS-PoE is faster because it does not require either a calibration pass or context compression, and instead only modifies positional encoding during the standard generation path. LongLLMLingua has competitive latency, but its peak GPU memory is substantially higher because it loads an additional compressor model to rewrite the retrieved context before generation. MAD-RAG has a cost close to BAIR because it also relies on a no context reference pass, although its intervention is simpler and does not include the same adaptive bottleneck calibration. The computational resources used for the experiments are detailed in Appendix M.

Table 3: Computational cost comparison on MedGemma-4B with the IU-Chest dataset. Runtime and memory changes are reported relative to Standard RAG.

Method	**Mean s/sample**	**Δ Time (%)**	**Peak GPU MB**	**Δ Memory (%)**
Standard RAG	5.38	0.00	12652	0.00
LongLLMLingua	4.82	-10.41	26686	+110.92
MS-PoE	4.08	-24.16	12652	0.00
MAD-RAG	5.69	+5.76	12694	+0.33
BAIR	5.93	+10.22	11704	-7.49

Table 4: Computational cost comparison on Qwen2.5-VL-7B with the FACET dataset. Runtime and memory changes are reported relative to Standard RAG.

Method	**Mean s/sample**	**Δ Time (%)**	**Peak GPU MB**	**Δ Memory (%)**
Standard RAG	7.08	0.00	8158	0.00
LongLLMLingua	6.25	-11.72	21688	+165.85
MS-PoE	6.01	-15.11	8110	-0.59
MAD-RAG	7.53	+6.36	7714	-5.44
BAIR	7.46	+5.37	7714	-5.44

I Attention Analysis

We analyze the final layer attention profile by measuring how the last input token attends to the multimodal prompt. For each sample, we retain only visual tokens and retrieved document tokens, excluding instruction and question tokens so that the visualization focuses on the clinically relevant evidence sources. The visual token region is shaded, and the third retrieved document is highlighted as the ground truth document. We compare Baseline, MS-PoE, MAD-RAG, and BAIR using robust

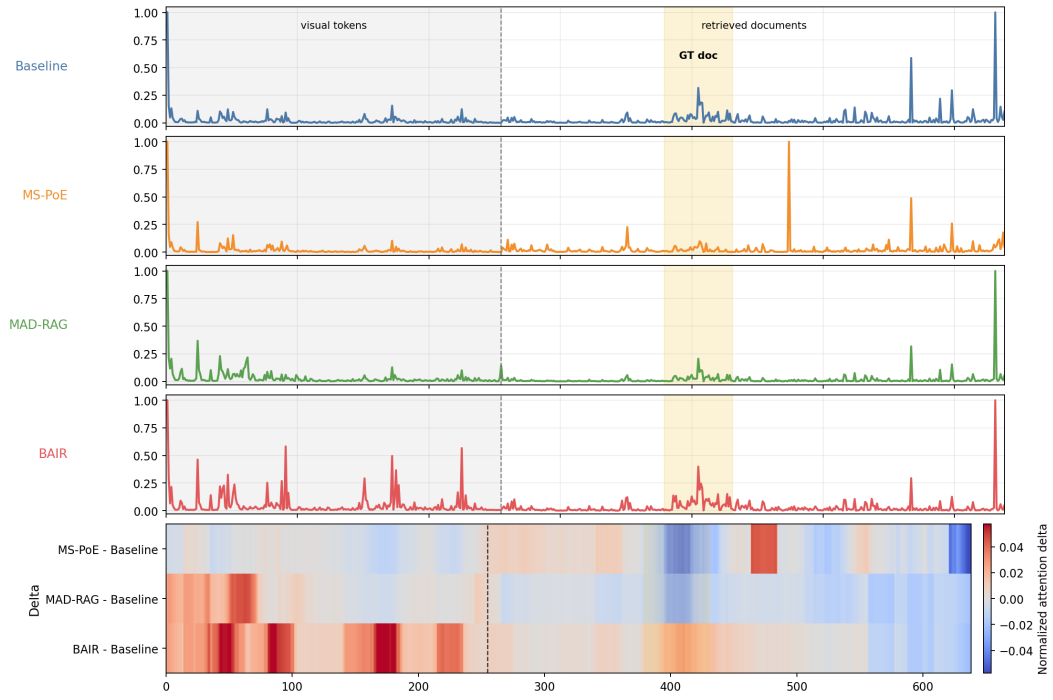


Figure 8: **Attention Profile Analysis.** We compare the final layer bottleneck attention profiles of Baseline, MS-PoE, MAD-RAG, and BAIR. The visual token region is shaded, and the third retrieved document is marked as the ground truth document. The upper panels show robust normalized attention curves, while the lower panels visualize each method's attention change relative to the Baseline.

normalized attention curves, and then visualize each method's change from the Baseline using a smoothed token axis delta heatmap. This analysis reveals whether each intervention shifts attention toward visual evidence, the ground truth retrieved document, or irrelevant textual regions.

The delta heatmap shows that BAIR produces the most targeted attention recovery. On the visual side, BAIR maintains both high attention mass and sharp visual concentration, indicating that it restores visual grounding rather than merely increasing diffuse visual attention. On the textual side, BAIR reduces the excessive recency bias toward the final retrieved document while increasing attention to the ground truth document. In contrast, MAD-RAG increases visual attention mass, but the recovered visual attention remains comparatively diffuse and its attention to the ground truth document is weakened. MS-PoE redistributes textual attention by modifying positional behavior, but the resulting attention pattern is less structured and does not consistently concentrate on either the visual evidence or the ground truth document. These results further support our diagnosis that effective multimodal RAG intervention requires both visual attention recovery and position aware textual recalibration.

J More Experimental Results

This section provides the complete experimental results across all evaluated datasets, models, and intervention backbones. While the main text reports representative results due to space constraints, Tables 5, 6, and 7 present the full comparison on IU-Chest, FACET, and NWPU, respectively. In addition to Accuracy, Correction Rate (CR), and Degradation Rate (DR), we report Recovery Rate (RR), Strictly Cured Rate (SR), Novel Recovery Rate (NR), and Generation Failure Rate (GFR) where applicable. These extended metrics allow us to examine not only whether a method improves aggregate performance, but also whether it recovers failures introduced by retrieval, suppresses recorruption, and avoids degenerate generation. Across the three domains, applying BAIR on top of existing RAG intervention methods generally improves accuracy and reduces degradation, showing that BAIR is complementary to prompt based, compression based, position based, and attention

Table 5: Comprehensive evaluation of multimodal RAG interventions on the **IU-Chest** dataset. Metrics are presented in percentages (%). Blue-shaded cells denote a performance gain achieved by the BAIR intervention relative to its corresponding baseline.

Model	Method	Acc. ↑	CR ↑	DR ↓	CR/DR ↑	RR ↑	SR ↑	NR ↑
MedGemma-4B	Baseline	63.76	-	-	-	-	-	-
	Standard RAG	66.33	28.00	13.56	2.06	-	-	-
	Visual-focus Instr.	65.38	26.68	14.13	1.89	16.74	30.50	10.35
	LongLLMLingua	65.66	25.14	12.79	1.97	21.73	42.41	12.33
	MS-PoE	67.17	20.68	7.85	2.63	18.99	46.70	7.42
	MAD-RAG	68.03	24.58	9.01	2.73	17.64	40.79	7.88
	Standard RAG + **BAIR**	68.31	24.25	8.40	2.89	18.38	43.75	7.93
	Visual-focus Instr. + **BAIR**	65.69	24.48	12.33	1.99	16.37	31.59	9.43
	LongLLMLingua + **BAIR**	65.91	25.21	12.46	2.02	21.74	43.41	12.14
	MS-PoE + **BAIR**	67.60	21.01	7.43	2.83	20.33	50.74	7.79
	MAD-RAG + **BAIR**	68.25	24.19	8.45	2.86	18.77	44.02	8.22
CheXagent-8B	Baseline	65.51	-	-	-	-	-	-
	Standard RAG	64.56	15.88	10.36	1.53	-	-	-
	Visual-focus Instr.	64.94	15.89	9.81	1.62	4.04	10.81	1.87
	LongLLMLingua	65.33	23.63	13.63	1.73	21.85	38.43	15.13
	MS-PoE	66.45	14.19	6.62	2.14	12.96	40.30	3.87
	MAD-RAG	64.79	16.16	10.19	1.59	1.18	2.93	0.57
	Standard RAG + **BAIR**	66.17	12.38	6.04	2.05	10.47	39.35	1.56
	Visual-focus Instr. + **BAIR**	65.58	12.89	7.19	1.79	7.95	28.03	1.50
	LongLLMLingua + **BAIR**	66.24	16.64	8.36	1.99	19.16	47.96	9.32
	MS-PoE + **BAIR**	66.60	11.72	5.05	2.32	13.95	47.78	3.19
	MAD-RAG + **BAIR**	66.08	11.83	5.87	2.02	11.73	42.35	2.24

mixing baselines. The gains are especially meaningful because they are achieved without retraining or fine tuning the underlying MLLMs.

Naturally Ordered Retrieval The forced middle placement above tests positional vulnerability under known evidence availability. Under unforced retriever ranking, BAIR also improves its matched RAG baseline on the evaluated subsets (Table 8). These subset results test unforced evidence selection and ordering within the present visually grounded tasks, not settings in which the image is insufficient and external knowledge is necessary. In particular, the NWPU subset's No-RAG accuracy is substantially different from that of the controlled evaluation (Table 1); the two sets should not be treated as interchangeable.

K Experiment Statistical Significance

To examine whether the observed gains are stable across repeated evaluations, we report the mean and standard deviation of Accuracy, Correction Rate, and Degradation Rate in Table 9. For each dataset and model configuration, we evaluate each baseline method together with its BAIR-augmented counterpart under the same protocol, enabling a paired comparison that isolates the effect of the intervention.

Across all domains, BAIR shows stable improvements over the corresponding baseline methods. On the medical IU-Chest dataset, BAIR consistently reduces Degradation Rate while maintaining or improving Accuracy, indicating that the intervention suppresses recorruption without discarding useful retrieved context. On the FACET dataset, BAIR provides stronger gains in Correction Rate and Accuracy, suggesting that visual grounding is especially important when retrieved text can reinforce demographic hallucinations. On the NWPU dataset, the gains are more moderate but still show the same general pattern: BAIR reduces degradation while preserving the correction benefits of retrieval.

Table 6: Comprehensive evaluation of multimodal RAG interventions on the **FACET** dataset. Metrics are presented in percentages (%). Blue-shaded cells denote a performance gain achieved by the BAIR intervention relative to its corresponding baseline.

Model	Method	Acc. ↑	CR ↑	DR ↓	CR/DR ↑	RR ↑	SR ↑	NR ↑	GFR ↓
Qwen2.5-VL-7B	Baseline	84.91	-	-	-	-	-	-	0.00
	Standard RAG	84.34	26.62	5.41	4.92	-	-	-	0.00
	Visual-focus Instr.	84.38	28.25	5.65	5.00	16.59	31.54	10.39	0.00
	LongLLMLingua	84.36	31.59	6.26	5.05	25.25	44.97	17.07	0.00
	MS-PoE	82.00	29.07	8.60	3.38	20.07	37.14	12.99	2.24
	MAD-RAG	84.41	25.53	5.11	5.00	13.86	31.22	6.71	0.00
	Standard RAG + **BAIR**	88.56	50.31	4.64	10.84	43.28	54.81	38.50	0.02
	Visual-focus Instr. + **BAIR**	87.28	44.93	5.20	8.64	38.82	54.14	32.47	0.04
	LongLLMLingua + **BAIR**	87.45	45.97	4.89	9.40	42.09	62.22	34.07	0.05
	MS-PoE + **BAIR**	92.58	70.71	3.53	20.03	68.79	76.96	65.40	0.63
	MAD-RAG + **BAIR**	85.10	36.08	6.19	5.83	27.41	42.09	21.34	0.00
DeepSeek-VL-7B	Baseline	90.24	-	-	-	-	-	-	0.00
	Standard RAG	90.89	31.05	2.64	11.76	-	-	-	0.00
	Visual-focus Instr.	90.47	28.53	2.83	10.08	7.67	17.67	4.12	0.00
	LongLLMLingua	87.76	34.21	6.45	5.30	29.54	41.81	25.19	0.00
	MS-PoE	91.82	51.05	3.78	13.51	51.07	72.41	43.51	0.03
	MAD-RAG	90.86	30.84	2.65	11.64	12.40	25.86	7.63	-
	Standard RAG + **BAIR**	91.02	30.95	2.48	12.48	11.61	23.71	7.33	0.00
	Visual-focus Instr. + **BAIR**	90.81	30.84	2.71	11.38	10.60	21.55	6.72	0.00
	LongLLMLingua + **BAIR**	87.80	33.79	6.36	5.31	29.76	43.10	25.04	-
	MS-PoE + **BAIR**	91.94	51.58	3.70	13.94	51.97	72.84	44.58	0.00
	MAD-RAG + **BAIR**	90.96	31.68	2.63	12.05	14.54	31.47	8.55	0.00

The reported standard deviations are small relative to the observed trends, suggesting that the improvements are not driven by isolated unstable runs. These results also show why Accuracy alone is insufficient for evaluating multimodal RAG reliability. By jointly analyzing Correction Rate and Degradation Rate, we verify that BAIR improves the correction-degradation tradeoff, which is central to reducing recorruption in multimodal RAG systems.

L Access to Data and Code

All datasets used in this work are publicly available research benchmarks. For the medical domain, we use IU Chest X ray Demner-Fushman et al. (2015). For the social fairness domain, we use FACET Gustafson et al. (2023). For the geospatial domain, we use NWPU-RESISC45 Cheng et al. (2017). The retrieval corpora are constructed from publicly available reports or textual descriptions associated with these benchmarks, together with publicly available contextual sources where applicable (e.g. Wikipedia).

All comparison methods are based on publicly available implementations or reproducible algorithmic descriptions. We include the code necessary to reproduce our experiments, including scripts for data preprocessing, retrieval context construction, model inference, intervention application, metric computation, and figure generation. The released code also contains configuration files specifying model checkpoints, prompt templates, intervention settings, and evaluation thresholds.

Table 10 summarizes the data and code resources used in this work. For the final public release, we will include the exact access URL and license information from the official release page of each dataset, model, and baseline implementation. Users of the released code are responsible for complying with the original license terms of each resource.

Table 7: Comprehensive evaluation of multimodal RAG interventions on the **NWPU** dataset. Metrics are presented in percentages (%). Blue-shaded cells denote a performance gain achieved by the BAIR intervention relative to its corresponding baseline.

Model	Method	Acc. ↑	CR ↑	DR ↓	CR/DR ↑	RR ↑	SR ↑	NR ↑	GFR ↓
SkySenseGPT-7B	Baseline	68.90	-	-	-	-	-	-	0.00
	Standard RAG	65.74	44.68	24.75	1.81	-	-	-	0.36
	Visual-focus Instr.	66.59	46.83	24.49	1.91	22.83	24.81	20.86	0.64
	LongLLMLingua	69.97	46.50	19.43	2.39	51.35	70.38	32.49	0.00
	MS-PoE	65.36	45.26	25.57	1.77	17.29	20.30	14.41	1.28
	MAD-RAG	66.18	45.67	24.56	1.86	19.39	23.46	15.35	4.44
	Standard RAG + **BAIR**	66.31	44.02	23.63	1.86	19.91	24.81	15.05	0.10
	Visual-focus Instr. + **BAIR**	66.74	46.00	23.89	1.93	21.56	24.81	18.33	0.23
	LongLLMLingua + **BAIR**	70.79	46.41	18.20	2.55	50.30	69.17	31.59	0.00
	MS-PoE + **BAIR**	66.05	44.27	24.12	1.84	18.64	23.16	14.16	0.10
	MAD-RAG + **BAIR**	67.13	46.17	23.41	1.97	19.16	23.31	15.05	0.10
EarthDial	Baseline	93.13	-	-	-	-	-	-	0.00
	Standard RAG	91.28	60.82	6.47	9.40	-	-	-	0.00
	Visual-focus Instr.	92.51	63.81	5.37	11.88	29.41	35.47	15.24	0.03
	LongLLMLingua	94.38	63.81	3.36	18.99	65.88	77.45	40.00	0.13
	MS-PoE	94.69	69.03	3.41	20.24	60.29	68.51	41.90	0.10
	MAD-RAG	91.64	62.69	6.22	10.08	21.76	24.68	15.24	0.00
	Standard RAG + **BAIR**	91.72	59.70	5.92	10.08	19.12	24.26	7.62	0.00
	Visual-focus Instr. + **BAIR**	92.62	62.69	5.18	12.10	29.71	36.60	14.29	0.03
	LongLLMLingua + **BAIR**	94.49	66.04	3.41	19.37	67.94	78.72	43.81	0.00
	MS-PoE + **BAIR**	96.74	75.37	1.68	44.86	76.18	85.96	54.29	0.18
	MAD-RAG + **BAIR**	91.65	62.31	6.17	10.10	18.53	21.28	12.38	0.00

Table 8: Naturally ordered retrieval on 300 sampled cases per configuration. Accuracy for NWPU and average clinical-label F1 for IU-Chest; DR is the degradation rate. Each column compares methods on the same subset and retrieval results.

Dataset / model	No RAG	RAG	RAG + BAIR	Gain	DR: RAG → BAIR
NWPU / SkySenseGPT-7B	89.33	91.67	94.33	+2.66	6.34 → 2.99
IU-Chest / MedGemma-4B	61.78	59.60	66.07	+6.47	17.54 → 8.67

M Computation Resource

Experiments are conducted on an NVIDIA RTX A5000 GPU with an AMD EPYC 7313 CPU, as summarized in Table 11. Because the proposed method is applied only at inference time and does not require training or fine tuning, reproducing the main results does not require large scale training resources.

N Broader Impact and Safeguard

This work studies a failure mode in multimodal retrieval augmented generation where retrieved textual context can override correct visual evidence. The potential positive impact is improved reliability for MLLMs deployed in settings where both images and external documents are used, including medical image interpretation, remote sensing analysis, and fairness sensitive visual reasoning. By identifying recorruption and proposing an inference-time mitigation, this work can help researchers evaluate whether a multimodal RAG system is genuinely grounded in visual evidence or merely copying retrieved text.

At the same time, this work should not be interpreted as making MLLMs safe for autonomous use in high-stakes decision making. In medical settings, BAIR and the evaluated MLLMs should be treated only as research tools or assistive systems, not as substitutes for clinical judgment. In social fairness settings, the goal of our evaluation is to diagnose biased hallucination and visual grounding

Table 9: Comprehensive evaluation of multimodal RAG interventions across all datasets. Metrics are presented in percentages (%).

	MedGemma-4B (IU-Chest)			CheXagent-8B (IU-Chest)		
Method	**Acc.** ↑	**CR** ↑	**DR** ↓	**Acc.** ↑	**CR** ↑	**DR** ↓
Baseline	63.76 ± 0.72	-	-	65.51 ± 0.76	-	-
Standard RAG	66.33 ± 0.73	28.00 ± 0.90	13.56 ± 0.80	64.56 ± 0.75	15.88 ± 0.78	10.36 ± 0.42
Visual-focus Instr.	65.38 ± 0.74	26.68 ± 0.91	14.13 ± 0.82	64.94 ± 0.78	15.89 ± 0.80	9.81 ± 0.45
LongLLMLingua	65.66 ± 0.71	25.14 ± 0.89	12.79 ± 0.78	65.33 ± 0.80	23.63 ± 0.96	13.63 ± 0.62
MS-PoE	67.17 ± 0.75	20.68 ± 0.92	7.85 ± 0.84	66.45 ± 0.74	14.19 ± 0.76	6.62 ± 0.38
MAD-RAG	68.03 ± 0.72	24.58 ± 0.90	9.01 ± 0.81	64.79 ± 0.77	16.16 ± 0.82	10.19 ± 0.46
Standard RAG + **BAIR**	68.31 ± 0.71	24.25 ± 0.91	8.40 ± 0.79	66.17 ± 0.79	12.38 ± 0.75	6.04 ± 0.44
Visual-focus Instr. + **BAIR**	65.69 ± 0.73	24.48 ± 0.89	12.33 ± 0.83	65.58 ± 0.76	12.89 ± 0.83	7.19 ± 0.48
LongLLMLingua + **BAIR**	65.91 ± 0.74	25.21 ± 0.92	12.46 ± 0.77	66.24 ± 0.81	16.64 ± 0.98	8.36 ± 0.65
MS-PoE + **BAIR**	67.60 ± 0.72	21.01 ± 0.90	7.43 ± 0.85	66.60 ± 0.75	11.72 ± 0.74	5.05 ± 0.37
MAD-RAG + **BAIR**	68.25 ± 0.70	24.19 ± 0.91	8.45 ± 0.82	66.08 ± 0.78	11.83 ± 0.84	5.87 ± 0.49
	Qwen2.5-VL-7B (FACET)			**DeepSeek-VL-7B (FACET)**		
Method	**Acc.** ↑	**CR** ↑	**DR** ↓	**Acc.** ↑	**CR** ↑	**DR** ↓
Baseline	84.91 ± 0.40	-	-	90.24 ± 0.29	-	-
Standard RAG	84.34 ± 0.38	26.62 ± 1.12	5.41 ± 0.26	90.89 ± 0.28	31.05 ± 1.55	2.64 ± 0.18
Visual-focus Instr.	84.38 ± 0.42	28.25 ± 1.15	5.65 ± 0.28	90.47 ± 0.30	28.53 ± 1.60	2.83 ± 0.20
LongLLMLingua	84.36 ± 0.36	31.59 ± 1.18	6.26 ± 0.25	87.76 ± 0.29	34.21 ± 1.65	6.45 ± 0.21
MS-PoE	82.00 ± 0.45	29.07 ± 1.10	8.60 ± 0.30	91.82 ± 0.28	51.05 ± 1.52	3.78 ± 0.17
MAD-RAG	84.41 ± 0.39	25.53 ± 1.14	5.11 ± 0.27	90.86 ± 0.30	30.84 ± 1.58	2.65 ± 0.19
Standard RAG + **BAIR**	88.56 ± 0.37	50.31 ± 1.13	4.64 ± 0.25	91.02 ± 0.29	30.95 ± 1.56	2.48 ± 0.17
Visual-focus Instr. + **BAIR**	87.28 ± 0.41	44.93 ± 1.16	5.20 ± 0.29	90.81 ± 0.28	30.84 ± 1.61	2.71 ± 0.19
LongLLMLingua + **BAIR**	87.45 ± 0.35	45.97 ± 1.19	4.89 ± 0.24	87.80 ± 0.29	33.79 ± 1.64	6.36 ± 0.20
MS-PoE + **BAIR**	92.58 ± 0.44	70.71 ± 1.11	3.53 ± 0.28	91.94 ± 0.29	51.58 ± 1.53	3.70 ± 0.16
MAD-RAG + **BAIR**	85.10 ± 0.38	36.08 ± 1.15	6.19 ± 0.26	90.96 ± 0.30	31.68 ± 1.59	2.63 ± 0.18
	SkySenseGPT-7B (NWPU)			**EarthDial (NWPU)**		
Method	**Acc.** ↑	**CR** ↑	**DR** ↓	**Acc.** ↑	**CR** ↑	**DR** ↓
Baseline	68.90 ± 0.74	-	-	93.13 ± 0.41	-	-
Standard RAG	65.74 ± 0.76	44.68 ± 1.44	24.75 ± 0.84	91.28 ± 0.45	60.82 ± 3.00	6.47 ± 0.41
Visual-focus Instr.	66.59 ± 0.76	46.83 ± 1.43	24.49 ± 0.84	92.51 ± 0.42	63.81 ± 2.95	5.37 ± 0.37
LongLLMLingua	69.97 ± 0.73	46.50 ± 1.42	19.43 ± 0.77	94.38 ± 0.37	63.81 ± 3.02	3.36 ± 0.30
MS-PoE	65.36 ± 0.77	45.26 ± 1.42	25.57 ± 0.85	94.69 ± 0.36	69.03 ± 2.88	3.41 ± 0.30
MAD-RAG	66.18 ± 0.76	45.67 ± 1.45	24.56 ± 0.85	91.64 ± 0.45	62.69 ± 2.99	6.22 ± 0.41
Standard RAG + **BAIR**	66.31 ± 0.76	44.02 ± 1.42	23.63 ± 0.83	91.72 ± 0.45	59.70 ± 3.04	5.92 ± 0.40
Visual-focus Instr. + **BAIR**	66.74 ± 0.76	46.00 ± 1.43	23.89 ± 0.84	92.62 ± 0.42	62.69 ± 2.96	5.18 ± 0.37
LongLLMLingua + **BAIR**	70.79 ± 0.73	46.41 ± 1.42	18.20 ± 0.76	94.49 ± 0.37	66.04 ± 2.93	3.41 ± 0.30
MS-PoE + **BAIR**	66.05 ± 0.76	44.27 ± 1.43	24.12 ± 0.84	96.74 ± 0.28	75.37 ± 2.66	1.68 ± 0.21
MAD-RAG + **BAIR**	67.13 ± 0.75	46.17 ± 1.43	23.41 ± 0.82	91.64 ± 0.45	62.31 ± 3.01	6.19 ± 0.41

failures, not to infer sensitive attributes for real individuals. For this reason, we use benchmark data and synthetic fairness evaluation settings where applicable, and we avoid deploying the system on private or identifiable individuals.

The proposed intervention also has limitations. Attention calibration can reduce certain forms of textual distraction, but it cannot guarantee factual correctness, eliminate all hallucinations, or correct errors caused by poor retrieval quality, ambiguous images, or incorrect model priors. Therefore, any deployment should include human oversight, task-specific validation, privacy review, and domain-specific safety constraints. We will release code and documentation with these limitations clearly stated, along with scripts that allow users to inspect correction and degradation behavior rather than relying only on aggregate accuracy.

Table 10: **Data, model, and code resources.** We use publicly available datasets, model checkpoints, and baseline implementations. The final release will include the exact license field from each official source.

Resource	Role in this work	Availability	License
IU Chest X ray (Demner-Fushman et al., 2015)	Medical factuality benchmark	Public dataset	CC BY-NC-ND 4.0
FACET (Gustafson et al., 2023)	Social fairness benchmark	Public dataset	Research-only
NWPU-RESISC45 (Cheng et al., 2017)	Remote sensing benchmark	Public dataset	CC BY 4.0
MedGemma (Sellergren et al., 2025)	Medical MLLM	Public model/checkpoint	health-ai-developer-foundations
CheXagent (Chen et al., 2024)	Medical MLLM	Public model/checkpoint	MIT license
SRR-BERT (Delbrouck et al., 2025)	Medical factuality evaluation model	Public model/checkpoint	MIT license
Qwen2.5-VL (Wang et al., 2024)	General MLLM	Public model/checkpoint	Apache License
DeepSeek-VL (Lu et al., 2024)	General MLLM	Public model/checkpoint	deepseek
SkySenseGPT (Luo et al., 2024)	Remote sensing MLLM	Public model/checkpoint	Apache-2.0 license
EarthDial (Soni et al., 2025)	Remote sensing MLLM	Public model/checkpoint	MIT license
MedSigLIP (Sellergren et al., 2025)	Medical retrieval model	Public model/checkpoint	health-ai-developer-foundations
RemoteCLIP (Liu et al., 2024a)	Remote sensing retrieval model	Public model/checkpoint	Apache-2.0 license
MS-PoE (Zhang et al., 2024)	Comparison method	Public implementation	MIT license
LongLLMLingua (Jiang et al., 2024)	Comparison method	Public model/checkpoint and implementation	MIT license
MAD-RAG (Zhao et al., 2026)	Comparison method	Reproducible method	-
BAIR code	Proposed method and evaluation scripts	Included in supplementary material	CC BY 4.0

Table 11: Compute Resources Used for Experiments

Component	**Details**
CPU	AMD EPYC 7313 16-Core Processor
GPU	NVIDIA RTX A5000