Self-Aware-MRAG: Training-Free Uncertainty-Guided Evidence Control for Multimodal RAG

Anonymous ACL submission

Abstract

Multimodal retrieval-augmented generation (MRAG) improves factuality by grounding generation in external evidence, yet it is often brittle because evidence is *used* in a largely static and indiscriminate manner. In practice, MRAG systems tend to always retrieve and then fuse evidence with fixed ordering/weighting, which can inject noise for easy queries and exacerbate position bias for hard ones. We propose **Self-Aware-MRAG**, a training-free framework that uses **cross-modal uncertainty** as an explicit control signal for evidence usage. Self-Aware-MRAG estimates uncertainty from complementary textual, visual, and cross-modal signals, and uses it to (i) *route retrieval* (skip / text / image / both) and (ii) modulate *position-aware fusion* via relevance-guided reordering and adaptive decay reweighting. Across OK-VQA and four additional MRAG benchmarks, Self-Aware-MRAG improves attribution precision by **+17.1 pp** over the strongest competitor in our setting and reduces position bias by **49.6%**, while maintaining competitive accuracy at a matched retrieval rate/budget (see tables 1 and 2).

1 Introduction

Multimodal large language models (MLLMs) have made rapid progress in perception and reasoning, yet they still hallucinate on knowledge-intensive queries and struggle to update static parametric knowledge. Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) mitigates this by grounding generation in external evidence at inference time. In multimodal settings, however, MRAG pipelines are often brittle: a common recipe is to *always* retrieve and concatenate top-k evidence. This not only incurs unnecessary latency for easy queries, but can also inject noisy or conflicting context that actively harms answer quality.

Code: https://github.com/X/Aware-MRAG

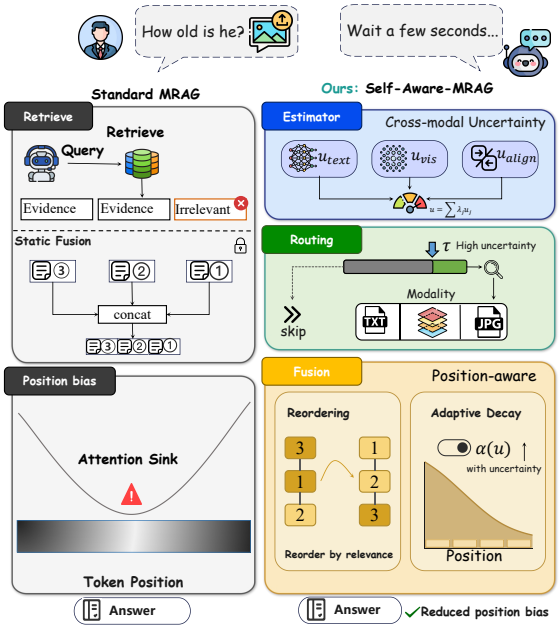

Figure 1: **Overview of Self-Aware-MRAG.** Standard MRAG often retrieves evidence indiscriminately and fuses it with static ordering/weighting. Self-Aware-MRAG estimates internal cross-modal uncertainty (textual, visual, and alignment) and uses it as a control signal to (i) trigger *modality-specific* retrieval only when needed, and (ii) modulate position-aware fusion to reduce position bias.

We argue that the core missing capability is **evidence-use control**—the system should decide *when* to seek external evidence, *what modality* to seek, and *how strongly* to rely on the resulting evidence. Current MRAG pipelines typically lack a principled estimate of the model's knowledge gap, so retrieval is triggered indiscriminately and modality choice (text vs. image) is often fixed. The problem is compounded at fusion time: long-context LMs exhibit position bias (Liu et al., 2024), over-attending to early/late evidence while underutilizing middle content, and existing MRAG systems commonly apply reranking or fixed ordering heuristics (Yu et al., 2024; Chen et al., 2024). Cru-

cially, a key dynamic is overlooked: *susceptibility to position bias depends on the model's epistemic state*. When the model is uncertain, it is more vulnerable to spurious or misplaced evidence; static fusion policies treat all queries the same, leading to unstable evidence use.

We present **Self-Aware-MRAG**, a training-free framework that turns **cross-modal uncertainty** into an explicit control signal for MRAG. Self-Aware-MRAG estimates uncertainty from three internal sources—textual dispersion, visual attention scatter, and text–vision alignment divergence—to decide *whether* to retrieve and *which modality* to retrieve. It then performs **uncertainty-modulated position-aware fusion** by combining relevance-guided reordering with adaptive decay weighting, reducing position bias while preserving useful context. Empirically, Self-Aware-MRAG improves robustness across benchmarks, yielding substantial gains in attribution precision and large reductions in position bias under comparable retrieval overhead.

We summarize our contributions as **one framework, one key mechanism, and one set of evidence**:

- **A unified framework (evidence-use control).** We propose a training-free MRAG framework that uses **cross-modal uncertainty** as a unified control signal to decide *when* to retrieve, *which modality* to retrieve, and how to use retrieved evidence.
- **A key mechanism (MRAG evidence fusion policy).** We propose an uncertainty-conditioned *evidence fusion* policy for MRAG that combines relevance-guided reordering with adaptive position decay, reducing position-driven misuse of retrieved evidence in long-context generation.
- **Evidence.** Experiments on OK-VQA and four additional MRAG benchmarks show substantial gains in attribution precision and large reductions in position bias under comparable retrieval overhead.

2 Related Work

Conventional retrieval-augmented generation (RAG) augments language models with non-parametric memory to improve factuality and currency (Lewis et al., 2020; Guu et al., 2020; Izacard and Grave, 2021). A growing line of work further studies *adaptive retrieval*, where the system triggers retrieval only when needed during generation. Representative approaches include reflection-style control that interleaves retrieve–generate–critique (Asai et al., 2024) and confidence-based triggering at token level (Jiang et al., 2023). Recent large-scale analyses suggest that uncertainty-based triggering often yields strong efficiency–accuracy trade-offs compared with more complex multi-stage pipelines (Moskvoretskii et al., 2025). Our work follows this general direction but targets a harder multimodal setting, where the system must decide not only *whether/when* to retrieve but also *which modality* (text vs. image) to retrieve, and how to robustly *use* retrieved evidence under long-context biases.

Multimodal RAG (MRAG) integrates visual and textual evidence for QA and reasoning (Abootorabi et al., 2025; Chen et al., 2022; Yu et al., 2024; Hu et al., 2023; Chen et al., 2024; Adjali et al., 2024; Zhang et al., 2024; Wang et al., 2025; Zhai, 2024; Cheng et al., 2024). Most MRAG pipelines retrieve a fixed number of candidates and concatenate them before generation; stronger variants incorporate reranking, structured fusion, or iterative retrieval. For example, MuRAG adopts FiD-style fusion (Chen et al., 2022); VisRAG and RagVL rely on reranking heuristics or MLLM-based rerankers (Yu et al., 2024; Chen et al., 2024); mR^{2}AG explores multi-round retrieval (Zhang et al., 2024). Agent-style test-time control has also been explored (e.g., SAM-RAG, ViDoRAG) (Zhai, 2024; Wang et al., 2025). In contrast to these largely static pipelines, we focus on a *training-free* evidence-use control mechanism driven by internal cross-modal signals.

Uncertainty estimation has been studied extensively, ranging from Bayesian approximations and deep ensembles (Gal and Ghahramani, 2016; Lakshminarayanan et al., 2017) to linguistic and information-theoretic proxies for generators (Kuhn et al., 2023). Prior work also derives uncertainty from internal representations to guide retrieval decisions (Yao et al., 2025). We extend this principle to MRAG by combining textual dispersion, visual attention scatter, and text–vision alignment divergence into a unified cross-modal uncertainty signal. Crucially, we treat uncertainty as a control knob not only for retrieval routing but also for evidence fusion in long contexts.

Finally, long-context LMs exhibit position bias, over-attending to early/late evidence while underutilizing the middle (Liu et al., 2024). Existing

MRAG systems typically apply static fusion policies (concatenation, reranking, or fixed ordering) that do not adapt to query difficulty (Yu et al., 2024; Chen et al., 2024). Our key difference is *uncertainty-conditioned fusion control*: when the model is uncertain, it is more vulnerable to spurious or misplaced evidence, and thus benefits from stronger emphasis on highly relevant evidence; when uncertainty is low, aggressive reweighting is less necessary. Accordingly, Self-Aware-MRAG modulates both evidence ordering and position weighting based on cross-modal uncertainty, yielding more stable evidence use.

Some dataset-leading systems improve performance via reinforcement fine-tuning, tools/notes, or heavier retrieval resources; these settings are not directly comparable to training-free MRAG with fixed backbones and budgets, but are complementary. Self-Aware-MRAG is an inference-time controller and can, in principle, be layered on top to improve evidence-usage stability.

3 Method

We consider multimodal QA with a question q, an image I, and a generator M_θ. Self-Aware-MRAG uses *cross-modal uncertainty* as a control signal to decide (i) *when/what* evidence to retrieve and (ii) *how* retrieved evidence is placed and weighted in the context for generation. We assume access to the generator's internal states (representations/attentions), which is available for the open-weight MLLMs used in our experiments.

3.1 Cross-Modal Uncertainty Estimation

In MRAG, ambiguity can stem from language, vision, or their interaction. We therefore estimate three complementary uncertainty signals and aggregate them into a scalar gate $u \in [0,1]$, while retaining component-wise signals for retrieval routing.

Textual uncertainty (u_{text}). We use internal representations after encoding (q, I) to quantify semantic dispersion. Let $H \in \mathbb{R}^{n\times d}$ be the final-layer token representations and $G = \frac{1}{d} HH^\top$ the Gram matrix. We define

$$u_{\text{text}} = 1 - \text{Norm}\left(\frac{\lambda_{\max}(G)}{\text{tr}(G)}\right), \qquad (1)$$

where $\lambda_{\max}(\cdot)$ is the largest eigenvalue and $\text{tr}(\cdot)$ is the trace; a less concentrated spectrum indicates higher uncertainty. This proxy follows representation-based uncertainty signals used in prior work (Yao et al., 2025) and requires only a single forward pass without draft generation.

Visual uncertainty (u_{vis}). Visual ambiguity often arises when the model cannot localize relevant regions and attends diffusely to the image. Let $A \in \mathbb{R}^{n\times m}$ be the cross-attention map from n text tokens to m image patches and $a_j = \frac{1}{n}\sum_{i=1}^{n} A_{ij}$ the average attention on patch j. We measure attention diffuseness via the variance of $\{a_j\}_{j=1}^{m}$ and map it to uncertainty:

$$u_{\text{vis}} = 1 - \text{Norm}\left(\frac{1}{m}\sum_{j=1}^{m}(a_j - \bar{a})^2\right), \qquad (2)$$

where $\bar{a}$ is the mean attention and $\text{Norm}(\cdot)$ rescales values to $[0,1]$ using robust statistics on a held-out set. Diffuse attention (low variance) yields high uncertainty, while focused attention yields low uncertainty.

Alignment uncertainty (u_{align}). To capture cross-modal inconsistency, we embed q and I with CLIP encoders, obtaining $e_{\text{text}}, e_{\text{vis}} \in \mathbb{R}^d$. We form normalized feature-activation profiles $P_{\text{text}} = \text{softmax}(e_{\text{text}}/T)$ and $P_{\text{vis}} = \text{softmax}(e_{\text{vis}}/T)$ (with T selected on validation), and measure their divergence:

$$u_{\text{align}} = \text{Norm}\big(\text{JSD}(P_{\text{text}} \parallel P_{\text{vis}})\big), \qquad (3)$$

where larger JSD indicates stronger cross-modal mismatch and thus higher uncertainty.

Unified gate. We aggregate the three components into a calibrated uncertainty gate:

$$u = \lambda_1 u_{\text{text}} + \lambda_2 u_{\text{vis}} + \lambda_3 u_{\text{align}}, \quad \text{s.t.} \sum_i \lambda_i = 1, \qquad (4)$$

with λ_i set to $1/3$ by default (or tuned on validation). All components are standardized to $[0,1]$ using robust statistics from a held-out set and fixed at test time.

3.2 Uncertainty-Guided Retrieval Routing

Self-Aware-MRAG triggers retrieval only when the model is uncertain and routes retrieval to the most informative modality. We first apply a selective trigger:

$$\text{RETRIEVE} \iff u > \tau, \qquad (5)$$

where τ is tuned on validation (or fixed across tasks). When retrieval is triggered, we optionally scale the retrieval budget with uncertainty,

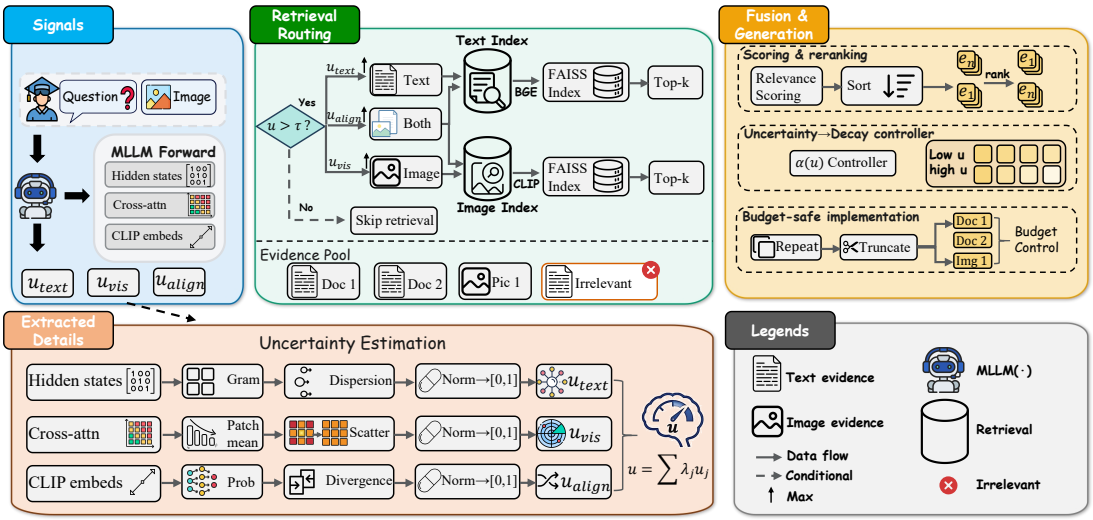

Figure 2: Overview of Self-Aware-MRAG. Self-Aware-MRAG treats MRAG as *evidence-use control*: it (1) estimates cross-modal uncertainty u from internal signals (text/vision/alignment), (2) skips retrieval when confident ($u \leq \tau$) and otherwise triggers and routes retrieval to TEXT/IMAGE/BOTH, and (3) fuses retrieved evidence via relevance-guided reordering and uncertainty-modulated position-aware decay under a fixed context budget to mitigate position-driven misuse of evidence.

$k(u) = k_{\min} + \lfloor (k_{\max} - k_{\min}) \cdot u \rfloor$; in our main setting we keep the maximum budget fixed and report the resulting retrieval rate.

We then select retrieval modality using the component-wise profile: if $u_{\text{align}} > \tau_{\text{align}}$, we retrieve *both* modalities to resolve potential cross-modal inconsistency; otherwise, we retrieve the modality with dominant uncertainty between u_{text} and u_{vis}. Equivalently,

$$\text{MODALITY} = \begin{cases} \text{BOTH}, & u_{\text{align}} > \tau_{\text{align}}, \\ \text{IMAGE}, & u_{\text{vis}} > u_{\text{text}}, \\ \text{TEXT}, & \text{otherwise}. \end{cases} \quad (6)$$

3.3 Uncertainty-Modulated Position-Aware Fusion

Retrieval alone is insufficient: MRAG must *use* retrieved evidence reliably. Since retrieved evidence is a controllable part of the context, how it is placed and weighted can amplify or mitigate long-context position bias (Liu et al., 2024). We therefore design an MRAG *evidence fusion policy* that controls (i) evidence placement via reordering and (ii) evidence weighting via uncertainty-modulated position decay.

Relevance-guided reordering (placement control). Given retrieved evidence chunks $\{d_j\}_{j=1}^{k}$, we compute a multimodal relevance score

$$r_j = \text{sim}_{\text{text}}(q, d_j) + \beta \cdot \text{sim}_{\text{vis}}(I, d_j), \quad (7)$$

and sort evidence by r_j in descending order, placing the most relevant chunks in positions where the generator tends to attend more strongly.

Uncertainty-modulated position decay (weighting control). We assign a position-dependent weight to the j-th evidence chunk,

$$w_j = \exp(-\alpha(u)\, j), \quad \alpha(u) = \alpha_0(1 + \kappa u), \quad (8)$$

where $\alpha(u)$ increases with uncertainty. When uncertainty is high, the model is more vulnerable to noisy or spurious retrieved evidence and benefits from stronger concentration on top-ranked evidence; when uncertainty is low, aggressive concentration is less necessary and a flatter weighting preserves useful complementary context.

Training-free implementation under fixed budget. To remain training-free and avoid modifying attention operators, we implement w_j via token-level repetition under a fixed total token budget: chunk d_j is repeated $\lceil w_j \cdot R \rceil$ times and the concatenated context is truncated to match the baseline context length. We enforce the same maximum context length as all baselines; repetition is compensated by truncation to keep the total token budget identical.We set the repetition cap to $R = 3$ in all experiments; the hyperparameter ranges are summarized in Appendix/Table 5.

Efficiency. Uncertainty estimation requires a single forward pass and adds minor overhead, while

selective triggering skips retrieval for confident queries, reducing average retrieval calls; we report retrieval rate/overhead in table 2. Additional end-to-end efficiency accounting (latency decomposition and effective input tokens) is provided in Appendix E.

4 Experiments

4.1 Settings

Benchmarks. We evaluate Self-Aware-MRAG on five public multimodal QA benchmarks: OK-VQA (Marino et al., 2019), A-OKVQA (Schwenk et al., 2022), WebQA (Chang et al., 2022), MultiModalQA (MMQA) (Talmor et al., 2021), and MRAG-Bench (Hu et al., 2025). We report each benchmark's official QA metric (Acc/F1/EM as applicable) and summarize overall performance in table 1. Following prior MRAG evaluations, we additionally use OK-VQA to report retrieval robustness and evidence-usage metrics (Recall@5, Faithfulness, Attribution Precision, and Position Bias; table 2).

Metrics. For overall QA performance, we report the official accuracy/F1/EM metrics of each dataset and summarize them in table 1. On OK-VQA, we additionally report: **Recall@5** (retrieval coverage under a fixed budget), **Faithfulness** (attribution consistency), **Attribution Precision** (precision of cited evidence), and **Position Bias** measured as Jensen–Shannon divergence from a relevance-aligned target distribution (lower is better). We also report **Retrieval Rate** (%) as an efficiency indicator.

Implementation Details. We employ Qwen3-VL-8B-Instruct as the backbone generator (Bai et al., 2025). Retrieval relies on BGE-large-en-v1.5 for text embeddings (Xiao et al., 2023) and CLIP ViT-L/14 for image embeddings (Radford et al., 2021). We build the dense index and perform nearest-neighbor search with FAISS (Johnson et al., 2017). Uncertainty aggregation uses equal weights ($\omega_i = 1/3$) by default, with the retrieval threshold τ tuned on the validation set.We additionally report a τ / retrieval-rate trade-off curve in Appendix B (Figure 3) to demonstrate that uncertainty acts as a controllable knob: retrieval can be reduced substantially while maintaining or improving evidence-use quality.

We standardize the maximum context length across baselines by using the same evidence budget and truncation rule. Statistical significance is assessed via paired bootstrap tests on test instances ($p < 0.05$, 10,000 resamples).

4.2 Baselines

We compare Self-Aware-MRAG against representative multimodal RAG systems: RagVL (Chen et al., 2024), MuRAG (Chen et al., 2022), VisRAG (Yu et al., 2024), mR^{2}AG (Zhang et al., 2024), ViDoRAG (Wang et al., 2025), and SAM-RAG (Zhai, 2024). All baselines are evaluated under the same backbone and context budget for a controlled comparison.

4.3 Main Results

Observations. Compared with the strongest baseline (SAM-RAG), Self-Aware-MRAG improves average performance by **+2.19pp** with consistent gains across all five datasets (table 1). Beyond QA accuracy, we further evaluate retrieval robustness and evidence usage on OK-VQA (next subsection).

4.4 Robustness and Efficiency on OK-VQA

Recall parity, better evidence use. Notably, Recall@5 remains comparable across all methods (54.42–54.95%), aligning with our goal of improving *evidence-use control* rather than raw retrieval strength. With recall parity, Self-Aware-MRAG achieves two key improvements: (i) *reduced cost* by skipping retrieval for confident queries (retrieval rate: 78.53% vs 82–100% for baselines), and (ii) *improved precision* through uncertainty-aware fusion that filters noise and focuses on high-relevance evidence.

4.5 Ablation and Analysis

We ablate the uncertainty components and the position-aware fusion module under a unified evaluation protocol. table 3 reports stepwise gains together with evidence-use and robustness metrics.

Key findings. (i) Adding uncertainty components steadily improves accuracy and faithfulness, with cross-modal alignment yielding the largest single gain (+1.84pp). (ii) The "+ Visual Uncertainty" variant slightly increases position bias (0.382→0.392), suggesting that incorporating additional multimodal context without position-aware fusion can exacerbate position sensitivity. (iii) Position-aware fusion sharply mitigates the lost-in-the-middle effect (0.365→0.148), and the full model further improves to 0.135. (iv) Under the

Table 1: Overall results across datasets under a controlled setting (same backbone, retriever, and context budget). All values are in % unless noted; Avg is the mean over datasets. ΔAvg (pp) denotes the absolute improvement over the strongest baseline in this controlled setting (SAM-RAG). This comparison is not intended as a leaderboard SOTA claim across different backbones or fine-tuning/tool-augmented systems.

	Datasets (higher is better)						Summary
Method	OK-VQA	A-OKVQA	WebQA	MultiModalQA	MRAG-Bench	Avg (%)	ΔAvg (pp)
Self-Aware-MRAG (Ours)	**67.14**	**59.46**	**70.81**	**53.42**	**76.87**	**65.54**	+2.19
SAM-RAG (Zhai, 2024)	65.53	57.31	68.54	51.12	74.24	63.35	0.00
RagVL (Chen et al., 2024)	63.82	55.23	67.54	50.21	73.12	61.98	−1.37
mR^{2}AG (Zhang et al., 2024)	64.22	55.93	66.82	49.54	72.33	61.77	−1.58
ViDoRAG (Wang et al., 2025)	62.11	53.14	69.52	49.83	72.41	61.40	−1.95
VisRAG (Yu et al., 2024)	62.54	53.81	65.23	48.17	71.55	60.26	−3.09
MuRAG (Chen et al., 2022)	60.15	50.42	62.13	45.34	58.41	55.29	−8.06

Table 2: Robustness and evidence-use evaluation on OK-VQA under a controlled MRAG setting. **Acc.** is answer accuracy; **Faith.** measures evidence faithfulness; **AttrPrec.** is attribution precision of cited evidence; **PosBias** quantifies position-induced brittleness (lower is better); **Recall@5** is retrieval recall at top-5; **Retrieval Rate** is the fraction of samples that trigger retrieval. ΔPosBias is computed relative to SAM-RAG as $\Delta\text{PosBias} = \text{PosBias} - \text{PosBias}_{\text{SAM-RAG}}$ (negative indicates reduced position bias).

Method	Recall@5 (%)	Faith. (%)	Attr. Prec. (%)	Pos. Bias ↓	Retr. Rate (%)	ΔPosBias(↓)
MuRAG	54.42	44.82	38.53	0.395	100.00	+0.127
RagVL	54.63	50.51	45.82	0.327	100.00	+0.059
VisRAG	54.58	48.24	42.31	0.296	100.00	+0.028
ViDoRAG	54.81	49.53	44.12	0.282	100.00	+0.014
mR^{2}AG	54.67	52.23	48.64	0.312	85.42	+0.044
SAM-RAG	54.78	56.41	52.13	0.268	82.14	0.000
Self-Aware-MRAG (Ours)	**54.95**	**60.50**	**69.20**	**0.135**	**78.53**	**-0.133**

unified citation protocol, attribution precision improves monotonically with added components and increases from the penultimate variant to the full model (63.45%→69.20%, +5.75pp), consistent with uncertainty-modulated filtering reducing noisy evidence when uncertainty is high. (v) Retrieval Rate (Table 3) helps contextualize the gains: improvements should be interpreted jointly with how often retrieval is triggered.

4.6 Further Discussions

Effects of backbone models. To assess whether Self-Aware-MRAG generalizes beyond a single MLLM, we evaluate under two backbones: Qwen3-VL-8B and Llama-3.2-11B, while keeping the retrieval budget (top-$k = 5$) and evaluation protocol unchanged. As shown in table 4, Self-Aware-MRAG consistently improves performance over the same MRAG baseline (MuRAG) across backbones: it improves accuracy by **+6.99pp** on Qwen3-VL-8B and **+5.33pp** on Llama-3.2-11B. Notably, Recall@5 remains nearly identical (differences $\leq$0.55pp), indicating that the gains are not driven by stronger retrieval coverage but by better evidence-use control. In addition, Self-Aware-MRAG yields large robustness improvements, including higher faithfulness and attribution precision and substantially lower position bias, suggesting that our control mechanism transfers across backbones.

Uncertainty quality: risk and correlation. We further examine whether the estimated uncertainty behaves as a meaningful control signal. We report AURC (area under the risk–coverage curve; lower is better) as a selective prediction metric, and Spearman ρ between uncertainty and answer error as a correlation measure. Across both backbones, Self-Aware-MRAG substantially reduces AURC (45.21→18.45 on Qwen3-VL-8B; 48.65→24.12 on Llama-3.2-11B) and increases Spearman correlation (0.18→0.56; 0.15→0.49), indicating that higher uncertainty aligns with higher risk. These results support our use of cross-modal uncertainty as a control knob for retrieval triggering and fusion aggressiveness.

Position-aware fusion mitigates position-driven evidence misuse. Finally, we connect uncertainty-conditioned control to long-context robustness. Despite similar retrieval recall,

Table 3: Incremental ablation variants on OK-VQA. Each step cumulatively adds one component to the previous variant. ΔAcc denotes the absolute accuracy gain (percentage points) over the previous row. **All variants are evaluated with an identical citation-output protocol**, so **Attr.** (attribution precision) is reported for every row. **Retrieval Rate** is the fraction of samples that trigger retrieval. **Pos. Bias** measures position-induced brittleness (lower is better)

Variant	Acc (%)	ΔAcc (pp)	F1 (%)	Recall@5 (%)	Faith. (%)	Attr. (%)	Retrieval Rate (%)	Pos. Bias ↓
Baseline (MuRAG)	60.15	0.00	48.50	54.42	44.80	55.00	100.00	0.395
+ Text Uncertainty	61.28	+1.13	49.95	53.68	46.20	56.50	92.10	0.382
+ Visual Uncertainty	62.90	+1.62	51.80	54.95	47.90	58.50	95.00	0.392
+ Cross-Modal Align.	64.74	+1.84	55.20	54.92	57.40	62.80	88.30	0.365
+ Position-Aware	65.86	+1.12	56.55	54.96	58.10	63.45	86.10	0.148
Full (Self-Aware-MRAG)	**67.14**	+1.28	**58.10**	54.95	**60.50**	**69.20**	82.40	**0.135**

Table 4: Backbone generalization and uncertainty analysis on OK-VQA. All values are in % except Pos.Bias and Spearman ρ. Lower is better for Pos.Bias and AURC (risk), higher is better otherwise.

Backbone	**Method**	**Acc.**	**ΔAcc**	**Recall@5**	**Faith.**	**Attr.Prec.**	**Pos.Bias↓**	**AURC↓**	**Spearman ρ ↑**
Qwen3-VL-8B	Baseline (MuRAG)	60.15	–	54.42	44.82	38.53	0.395	45.21	0.18
Qwen3-VL-8B	Self-Aware-MRAG (Ours)	67.14	+6.99	54.95	60.50	69.20	0.135	18.45	0.56
Llama-3.2-11B	Baseline (MuRAG)	59.12	–	54.28	43.65	37.14	0.412	48.65	0.15
Llama-3.2-11B	Self-Aware-MRAG (Ours)	64.45	+5.33	54.83	58.92	67.45	0.153	24.12	0.49

Self-Aware-MRAG consistently reduces position bias (0.395→0.135 and 0.412→0.153) while improving attribution precision by large margins. This suggests that controlling evidence placement/weighting conditioned on uncertainty can stabilize evidence usage, especially under long-context positional effects.

Efficiency trade-offs and practical deployment. Self-Aware-MRAG improves efficiency by skipping retrieval on confident queries while keeping the per-call retrieval budget fixed. Combined with the recall parity observed above, the improvements in faithfulness and attribution precision indicate a favorable accuracy–robustness–efficiency trade-off: we retrieve no more evidence than baselines, but use evidence more precisely and reduce position-driven failures. We provide full hyperparameter settings (thresholds and normalization statistics) and additional sensitivity results in the Appendix for reproducibility.

5 Conclusion

In this paper, we propose Self-Aware-MRAG to improve evidence usage in multimodal retrieval-augmented generation. The main idea of Self-Aware-MRAG is to treat MRAG as *evidence-use control* driven by cross-modal uncertainty. Specifically, Self-Aware-MRAG estimates fine-grained uncertainty from internal textual, visual, and alignment signals, uses it to selectively trigger and route modality-specific retrieval, and modulates evidence fusion to reduce position-driven misuse of retrieved context. We conduct extensive experiments on five multimodal QA benchmarks and show that Self-Aware-MRAG achieves consistent gains in overall accuracy, while substantially improving robustness on OK-VQA—notably increasing attribution precision and reducing position bias under comparable retrieval recall and lower retrieval rate. These results suggest that uncertainty is a practical control knob for building more reliable and efficient multimodal RAG systems.We plan to release the code and evaluation scripts to facilitate reproducibility.

Limitations

Access assumptions. Self-Aware-MRAG estimates uncertainty using signals that require access to model internals (e.g., hidden representations and/or attention-related statistics). This assumption may not hold for strictly black-box APIs that only return final text outputs, limiting direct applicability in such settings.

Proxy design choices. Some components rely on practical proxies rather than fully explicit mechanisms. In particular, our position-aware weighting is implemented via token repetition under a fixed context budget, which only approximates fine-grained attention reweighting and may interact with truncation or context management in ways that differ from an explicit reweighting operator.

additional training-free positional debiasing baselines. Recent work proposes inference-time

positional debiasing by modifying internal positional signals, e.g., scaling a positional hidden-state channel in causal LMs, or scaling positional embeddings layer-wise, as well as training-free attention interpolation for multi-image LVLMs. While promising, these approaches require model-internal intervention (hidden-state/positional embedding/attention mask) that is not directly exposed in our current black-box VLM evaluation pipeline and may not transfer one-to-one to evidence-chunk packing in RAG. As a lightweight training-free baseline within our setting, we include order randomization controls (Shuffle/Reverse) and will further incorporate model-internal debiasing baselines in the final version.

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. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16776–16809, Vienna, Austria. Association for Computational Linguistics.

Omar Adjali, Olivier Ferret, Sahar Ghannay, and Herv'e Le Borgne. 2024. Multi-level information retrieval augmented generation for knowledge-based visual question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16499–16513, Miami, Florida, USA. Association for Computational Linguistics.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *arXiv preprint*. ArXiv:2511.21631.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2022. WebQA: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16495–16504.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. 2022. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhanpeng Chen, Chengjin Xu, Yiyan Qi, and Jian Guo. 2024. Mllm is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training. *CoRR*, abs/2407.21439.

Xin Cheng, Xun Wang, Xingxing Zhang, Tao Ge, Si-Qing Chen, Furu Wei, Huishuai Zhang, and Dongyan Zhao. 2024. xrag: Extreme context compression for retrieval-augmented generation with one token. In *Advances in Neural Information Processing Systems*, volume 37.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Wenbo Hu, Jia-Chen Gu, Zi-Yi Dou, Mohsen Fayyaz, Pan Lu, Kai-Wei Chang, and Nanyun Peng. 2025. MRAG-Bench: Vision-centric evaluation for retrieval-augmented multimodal models. In *International Conference on Learning Representations (ICLR)*.

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, David A. Ross, Cordelia Schmid, and Alireza Fathi. 2023. REVEAL: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 23369–23379. IEEE.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023*

Conference on Empirical Methods in Natural Language Processing, pages 7969–7992, Singapore. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *arXiv preprint*. ArXiv:1702.08734.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*. ArXiv:2302.09664.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, pages 6402–6413. Curran Associates, Inc.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*.

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

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3195–3204.

Viktor Moskvoretskii, Maria Marina, Mikhail Salnikov, Nikolay Ivanov, Sergey Pletenev, Daria Galimzianova, Nikita Krayko, Vasily Konovalov, Irina Nikishina, and Alexander Panchenko. 2025. Adaptive retrieval without self-knowledge? bringing uncertainty back home. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6355–6384, Vienna, Austria. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A benchmark for visual question answering using world knowledge. In *Computer Vision – ECCV 2022*, Lecture Notes in Computer Science, pages 146–162. Springer.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. MultiModalQA: Complex question answering over text, tables and images. In *International Conference on Learning Representations (ICLR)*.

Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. 2025. ViDoRAG: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9113–9134, Suzhou, China. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2023. C-pack: Packed resources for general chinese embeddings. *arXiv preprint*. ArXiv:2309.07597.

Zijun Yao, Weijian Qi, Liangming Pan, Shulin Cao, Linmei Hu, Weichuan Liu, Lei Hou, and Juanzi Li. 2025. SeaKR: Self-aware knowledge retrieval for adaptive retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27022–27043, Vienna, Austria. 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. 2024. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. *CoRR*, abs/2410.10594.

Wenjia Zhai. 2024. Sam-rag: Self-adaptive multimodal retrieval-augmented generation. *CoRR*, abs/2410.11321.

Tao Zhang, Ziqi Zhang, Zongyang Ma, Yuxin Chen, Zhongang Qi, Chunfeng Yuan, Bing Li, Junfu Pu, Yuxuan Zhao, Zehua Xie, Jin Ma, Ying Shan, and Weiming Hu. 2024. mr^{2}ag: Multimodal retrieval-reflection-augmented generation for knowledge-based vqa. *CoRR*, abs/2411.15041.

A Experimental Details

Backbones and retrievers. We use Qwen3-VL-8B-Instruct as the primary vision-language backbone for generation, and additionally evaluate on Llama-3.2-11B-Vision-Instruct for generalization. For retrieval, we employ BGE-large-en-v1.5 for text embeddings and CLIP-ViT-L/14 for image embeddings. All models are used as-is without additional fine-tuning.

Retrieval corpora across datasets. We follow a unified retrieval setup per dataset: (i) OK-VQA / A-OKVQA retrieve from Wikipedia-3M (text + associated images); (ii) MRAG-Bench uses the benchmark-provided image corpus with captions as documents; (iii)MultiModalQA retrieves from the dataset-provided text/table/image documents; (iv) WebQA directly uses the provided text/image facts as the evidence pool. For indexing, we use FAISS Flat for exact nearest-neighbor search. Text documents are chunked at the paragraph level with a maximum of 512 tokens per chunk when applicable.

Fairness across baselines. To ensure controlled comparisons, all methods use the same retriever configuration, the same candidate pool, and the same retrieval budget. Unless explicitly stated otherwise, we retrieve top-$k = 5$ evidence items for all methods, and keep the maximum number of images in the input context bounded (up to 20 images). We also standardize the maximum context length and apply the same truncation policy across all methods.

Hyperparameters. table 5 summarizes key hyperparameters for Self-Aware-MRAG We use equal uncertainty weights ($\omega_1 = \omega_2 = \omega_3 = \frac{1}{3}$) by default. Thresholds are tuned on a held-out validation set (10% of training data) and then fixed for test evaluation. Specifically, τ is selected from $\{0.25, 0.30, 0.35, 0.40, 0.45\}$ with a preference for maintaining retrieval rate below 85%; τ_{align} is selected from $\{0.4, 0.5, 0.6\}$; the alignment temperature T is selected from $\{0.05, 0.1, 0.2, 0.5\}$; κ is selected from $\{0.25, 0.5, 0.75, 1.0\}$; and β is selected from $\{0.1, 0.2, 0.3, 0.5\}$. Unless noted, we keep these values fixed across datasets.

B Retrieval-Rate Trade-off by Varying τ

To characterize uncertainty as a controllable knob, we vary the retrieval trigger threshold τ while keeping *all other settings fixed* (same backbone, retriever, top-k, context budget, and citation-output protocol as in Tables 2 and 3). For each τ, we evaluate on the OK-VQA test set and record the resulting retrieval trigger rate together with Acc/AttrPrec/PosBias. We plot metrics against the **retrieval rate** (more interpretable than τ), and mark the operating point used in the main paper (chosen on validation).

C Uncertainty Estimation Details

Normalization. We standardize each raw uncertainty score to $[0, 1]$ using robust statistics computed on a calibration set sampled from the validation split. For a raw score s, we compute:

$$\text{Norm}(s) = \text{clip}\left(\frac{s - \text{median}(S)}{1.5 \cdot \text{IQR}(S)} + 0.5, 0, 1\right), \tag{9}$$

where S denotes the raw scores computed on 500 validation instances, and IQR is the interquartile range. The resulting normalization parameters are fixed and reused for test-time evaluation.

Textual uncertainty (spectral dispersion). We extract final-layer hidden states $H \in \mathbb{R}^{n \times d}$ for the prompt tokens. We form the Gram matrix $G = \frac{1}{d} H H^\top$ and compute the ratio between the leading eigenvalue and the trace:

$$u_{\text{text}} = 1 - \text{Norm}\left(\frac{\lambda_{\max}(G)}{\text{tr}(G)}\right), \tag{10}$$

where a smaller leading-eigen ratio indicates more dispersed representations (higher uncertainty), following the spectral-dispersion intuition.

Visual uncertainty (attention scatter). Let $A \in \mathbb{R}^{n \times m}$ be the text-to-image cross-attention from n text tokens to m image patches. We average attention across heads and layers to obtain $\bar{A}$, and compute patch-wise mean attention $a_j = \frac{1}{n} \sum_{i=1}^{n} \bar{A}_{ij}$. Visual uncertainty is the complement of the normalized patch variance:

$$u_{\text{vis}} = 1 - \text{Norm}\left(\frac{1}{m} \sum_{j=1}^{m} (a_j - \bar{a})^2\right), \tag{11}$$

where $\bar{a}$ is the mean of $\{a_j\}$. Intuitively, diffuse attention (low variance) indicates the model cannot localize relevant regions, leading to higher uncertainty.

Table 5: Hyperparameter settings for Self-Aware-MRAG. "Tuned" indicates selected on validation; others are fixed by default.

Component	Symbol	Value	Range / Note
Uncertainty weights	$\omega_1, \omega_2, \omega_3$	$\frac{1}{3}$	fixed
Retrieval trigger	τ	0.35	$\{0.25, 0.30, 0.35, 0.40, 0.45\}$ (tuned)
Alignment conflict	τ_{align}	0.50	$\{0.4, 0.5, 0.6\}$ (tuned)
Fusion relevance weight	β	0.30	$\{0.1, 0.2, 0.3, 0.5\}$ (tuned)
Base decay rate	α_0	0.10	$\{0.05, 0.10, 0.15\}$ (tuned)
Decay modulator	κ	0.50	$\{0.25, 0.5, 0.75, 1.0\}$ (tuned)
Repetition factor	R	3	$\{2, 3, 4, 5\}$ (fixed)
Alignment temperature	T	0.10	$\{0.05, 0.1, 0.2, 0.5\}$ (tuned)
Retrieval budget	k	5	fixed across methods

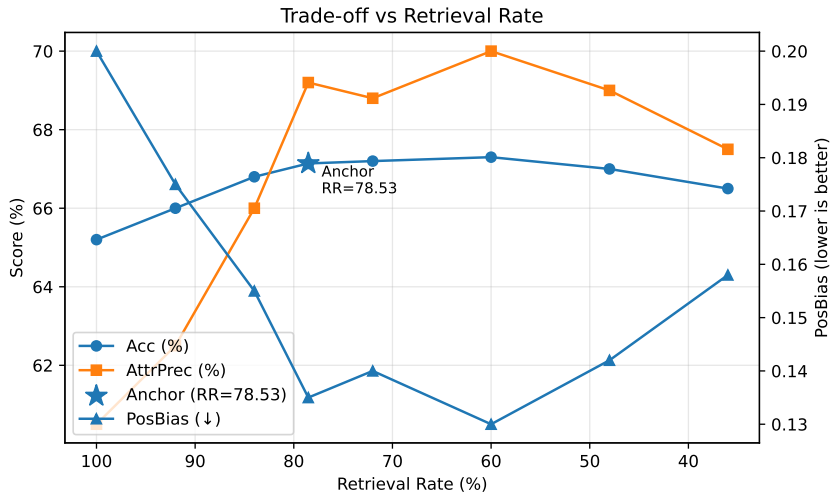

Figure 3: Trade-off curve induced by varying the retrieval threshold τ on OK-VQA. The x-axis shows the retrieval trigger rate; the left y-axis reports Acc and AttrPrec, and the right y-axis reports PosBias (lower is better). The starred point denotes the operating point used for the main results.

Alignment uncertainty (cross-modal discrepancy). We compute CLIP embeddings for the question text and image, $e_{\text{text}}, e_{\text{vis}} \in \mathbb{R}^d$, and convert them to normalized feature-importance profiles: $P_{\text{text}} = \text{softmax}(e_{\text{text}}/T)$ and $P_{\text{vis}} = \text{softmax}(e_{\text{vis}}/T)$. We then compute:

$$u_{\text{align}} = \text{Norm}(\text{JSD}(P_{\text{text}} \parallel P_{\text{vis}})), \qquad (12)$$

where a larger divergence indicates greater cross-modal inconsistency.

D Evaluation Protocols

QA metrics. We report dataset-official QA metrics (e.g., Accuracy/F1/EM) for OK-VQA, A-OKVQA, WebQA, MM-VQA, and MRAG-Bench.

Retrieval and evidence-use metrics on OK-VQA. We follow the OK- protocol for Recall@5, Faithfulness, and Attribution Precision. Faithfulness checks whether a generated answer span can be supported by at least one retrieved evidence segment after normalization; Attribution Precision measures the precision of cited evidence among retrieved items under the same protocol.

Unified citation-output protocol (enforced for all methods). To ensure fair cross-method comparisons for Faithfulness and Attribution Precision, we evaluate *all* methods (ours and baselines) with the same citation constraints and the same answer format.

Prompt requirement. We append the following instruction to the user prompt for every method: *"Answer the question. Every factual claim MUST be supported by at least one citation to the retrieved evidence. Cite evidence using [E#], where # indexes the retrieved evidence chunks in the provided context. If the question cannot be answered using the retrieved evidence, output 'Insufficient evi-*

dence'."

Failure handling. If a model outputs an answer with (i) no citations, or (ii) citations with invalid indices (e.g., [E99] when only top-k chunks exist), the instance receives zero supported citations for citation-dependent metrics. This prevents unfairly benefiting methods that do not comply with citation constraints.

Position bias. We quantify position bias as Jensen–Shannon divergence (JSD) between (i) the model's empirical attention mass over evidence positions and (ii) a relevance-aligned target distribution. Let the retrieved evidence be ordered into k chunks, and let $p \in \Delta^k$ be the normalized attention mass over the k evidence chunks, computed by aggregating the generator's attention from answer tokens to tokens belonging to each evidence chunk. Let r_j be the relevance score of chunk j (the same relevance used for reordering), and define the target distribution $q_j = \frac{\exp(r_j)}{\sum_{t=1}^{k} \exp(r_t)}$. We compute:

$$\text{PosBias} = \text{JSD}(p \parallel q), \tag{13}$$

where lower values indicate closer alignment between attention allocation and relevance (i.e., reduced position-driven misuse).

Clarification (PosBias). Since PosBias uses a relevance-aligned target distribution constructed from the same relevance scores used for reordering, it should be viewed as an *alignment-to-relevance* metric (not fully model-agnostic positional brittleness).

Model-agnostic order sensitivity. We therefore additionally define order-sensitivity tests that keep the retrieved evidence set fixed but change only the order: random-shuffle variance over N permutations and best-vs-worst (relevance-sorted vs. reverse-sorted) performance gaps; smaller variance/gaps indicate reduced ordering sensitivity.

Uncertainty quality. We report AURC (area under the risk–coverage curve; lower is better) as a selective prediction metric, and Spearman ρ between uncertainty and answer error to measure monotonic correlation between predicted uncertainty and risk.

Statistical testing and reproducibility. We fix random seeds for evaluation and assess statistical significance via paired bootstrap tests on test instances (10,000 resamples, $p < 0.05$). We will release evaluation scripts and configuration files after the review period.

Table 6: Results on Llama-3.2-11B-Vision-Instruct on OK-VQA.

Method	Acc	Faith	Attr	Bias ↓	AURC ↓
Baseline (MuRAG)	59.12	43.65	37.14	0.412	48.65
Self-Aware-MRAG	**64.45**	**58.92**	**67.45**	**0.153**	**24.12**

E Efficiency Accounting

Why Retrieval Rate alone is insufficient. Retrieval Rate captures how often retrieval is invoked, but our pipeline also incurs (i) an uncertainty-estimation forward pass and (ii) possible token repetition for evidence weighting. We therefore explicitly define the full efficiency accounting below.

End-to-end latency decomposition. For each query, we decompose wall-clock latency into:

$$t_{\text{total}} = t_{\text{unc}} + \mathbb{I}[u > \tau] \cdot (t_{\text{retr}} + t_{\text{prep}}) + t_{\text{gen}},$$

where t_{unc} is the uncertainty estimation cost, t_{retr} is retrieval time, t_{prep} covers evidence formatting/packing (including repetition or truncation), and t_{gen} is generation time.

Token budget and repetition. All methods are evaluated under the same fixed context budget. When repetition is used to approximate evidence weights, we apply truncation after packing to keep the effective input tokens within the same budget; we additionally report the *effective input tokens* after packing/truncation to make the cost transparent.

Measurement protocol (reproducibility). Latency is measured on the same hardware with identical batch size, decoding settings, and caching/warmup procedure for all methods. We report mean (and optionally std) over the evaluation set or a fixed-size subset for stable measurement.

F Additional Results: Backbone Generalization

Discussion. table 6 shows that Self-Aware-MRAG consistently improves both accuracy and robustness metrics under the Llama backbone, supporting that uncertainty-driven evidence-use control transfers across different MLLM architectures.