# Multimodal Adaptive Retrieval Augmented Generation through Internal Representation Learning

Ruoshuang Du*, Xin Sun†, Qiang Liu†, Bowen Song‡, Zhongqi Chen‡, Weiqiang Wang‡, Liang Wang†
*School of Information Science and Technology, Shanghaitech University
†Institute of Automation, Chinese Academy of Sciences
‡Ant Group

***Abstract*—Visual Question Answering systems face reliability issues due to hallucinations, where models generate answers misaligned with visual input or factual knowledge. While Retrieval Augmented Generation frameworks mitigate this issue by incorporating external knowledge, static retrieval often introduces irrelevant or conflicting content, particularly in visual RAG settings where visually similar but semantically incorrect evidence may be retrieved. To address this, we propose Multimodal Adaptive RAG (MMA-RAG), which dynamically assesses the confidence in the internal knowledge of the model to decide whether to incorporate the retrieved external information into the generation process. Central to MMA-RAG is a decision classifier trained through a layer-wise analysis, which leverages joint internal visual and textual representations to guide the use of reverse image retrieval. Experiments demonstrated that the model achieves a significant improvement in response performance in three VQA datasets. Meanwhile, ablation studies highlighted the importance of internal representations in adaptive retrieval decisions. In general, the experimental results demonstrated that MMA-RAG effectively balances external knowledge utilization and inference robustness in diverse multimodal scenarios.**

## I. Introduction

Large language models (LLMs) have achieved remarkable success in a wide range of natural language understanding and generation tasks [1]. However, despite their impressive capabilities, LLMs are known to suffer from a critical limitation: hallucination [2]–[4]. This phenomenon refers to the generation of outputs that are factually inaccurate, unverifiable, or inconsistent with the provided input.

To address this limitation, Retrieval-Augmented Generation (RAG) has emerged as a promising solution [5], [6]. RAG improves language model performance by incorporating external knowledge retrieved from large-scale textual corpora to complement the model's internal parameterized representations [7]–[12]. This approach has been shown to improve response accuracy and factual consistency, particularly in knowledge-intensive tasks where reliance on static, pre-trained parameters alone may lead to outdated or incorrect information [13], [14].

Although early implementations were confined to purely textual settings, recent work has extended this idea to multimodal contexts, leading to multimodal RAG frameworks [15]–[18]. In multimodal RAG, models generate image-question conditioned retrieval queries, sourcing both text and images from external databases to support reasoning. These systems have shown promise on complex tasks such as Visual Question Answering (VQA) [19]. For example, REVIVE uses regional visual representations combined with the knowledge retrieved to improve the accuracy of the answers [20]. Similarly, MuRAG improves response generation by incorporating information from external knowledge bases, enabling more effective joint reasoning over images and texts [21]. The multimodal alignment model introduced a multimodal large language model-based reclassification step, selecting the most relevant knowledge from the top candidates retrieved to improve performance [22].

Reverse Image Retrieval (RIR) can be viewed as a specialized variant of multimodal RAG, in which additional multimodal context is provided by retrieving visually similar images from web-based sources [23]. In this approach, screenshots or related images retrieved based on the query image are integrated with the original input, thereby enriching the model's input space with complementary visual information. Importantly, the advantage of RIR does not typically lie in directly supplying the correct answer, but rather in facilitating more accurate grounding and interpretation of the query through contextually relevant visual cues.

Despite the aforementioned advantages, visual retrieval augmented generation introduces a more challenging failure mode than its text-only counterpart. Images returned by visual retrieval are often highly similar in appearance, yet semantically inconsistent with the query. As illustrated in Fig. 1, queries concerning plants from the *Lamiaceae* family may retrieve visually similar species such as Horehound, resulting in evidence that appears highly convincing but is in fact incorrect. This phenomenon, referred to as *visual similarity with semantic mismatch*, is more difficult to defend against than interference in text-based RAG systems. Consequently, effective mitigation requires joint reasoning over visual and textual features, enabling the model to simultaneously assess the visual similarity of retrieved content and its semantic consistency with the query.

Similar to RIR, many existing multimodal RAG methods implicitly assume that external information is always beneficial, often introducing irrelevant or misleading evidence—particularly in cases where the model already pos-

Corresponding author: qiang.liu@nlpr.ia.ac.cn

sesses sufficient internal knowledge—thereby causing retrieval redundancy that is likely to degrade overall model performance [24], [25].

We propose MMA-RAG[1], a multimodal adaptive RAG framework that mitigates harmful retrieval from visually similar but semantically incorrect evidence. It aligns textual and visual hidden states into a unified representation and trains a four-class classifier to predict the impact of retrieval on the correctness of the answer. Crucially, we argue that simple last-layer representations may not fully capture the nuanced misalignment between visual and textual modalities. Through a comprehensive layer-wise analysis of the model's internal states, we observe that the semantic alignment between vision and text evolves differently across network depths. Text-only features show limited discriminative capability in shallow layers and become effective only in deeper layers, whereas multimodal features achieve high detection accuracy even in the early layers. This observation indicates that multimodal fusion is crucial for the early identification of erroneous or misleading evidence. Motivated by this finding, MMA-RAG strategically selects and fuses the most informative intermediate representations to train a multimodal joint-feature classifier. Finally, guided by the classifier's predictions, the model adaptively employs the RIR mechanism, avoiding the introduction of harmful external information while ensuring the generation of correct answers whenever possible. The core contributions of this paper are summarized as follows:

- We propose MMA-RAG, a multimodal adaptive retrieval augmented generation framework that predicts the utility of RIR from internal multimodal representations to mitigate harmful retrieval in visual question answering tasks.
- We perform a layer-wise analysis of multimodal large language models, revealing how visual and textual confidence signals evolve and informing the selection of internal features for hallucination detection.
- We design an internal-representation-based retrieval utility classifier that integrates multimodal features to assess whether external retrieval improves response correctness.
- Extensive experiments on three knowledge-intensive VQA benchmarks with multiple vision–language backbone models demonstrate that MMA-RAG outperforms standard retrieval-based methods and existing baselines.

## II. Methodology

Although RIR can provide valuable external context by retrieving visually similar images, it can also introduce harmful samples that mislead the model, especially when the retrieved images contain semantically irrelevant or contradictory content. In such cases, relying on external retrieval may cause the model to generate incorrect answers, even when the original image alone would have sufficed for correct reasoning. As illustrated in the example depicted in Fig. 1, the use of RIR yields the answer "The plant is in the horehound family", which is in fact incorrect. The accurate response corresponds to the case without RIR, namely, "The plant family of this image is the Lamiaceae family, also known as the mint family".

To address such issues in VQA tasks, we propose a **Multimodal Adaptive Retrieval-Augmented Generation (MMA-RAG)** framework to address the challenge of harmful retrievals in VQA tasks. The core idea is to adaptively determine whether externally retrieved images should be incorporated into the generation process to minimize the negative impact of irrelevant or misleading visual information. If the retrieved images are helpful to improve the accuracy of the answers, the MMA-RAG framework adopts the RAG approach, incorporating all images and the corresponding question as input to the generation model. If the retrieved images introduce noise and degrade answer quality, the framework relies solely on the original image and the question for answer generation.

MMA-RAG is designed to make adaptive use of retrieved visual content in a multimodal VQA setting. It consists of three key components:

**1) Reverse Image Retrieval**: For each VQA instance, the input consists of a query $Q$ and an image $I_1$. We perform RIR using $I_1$ querying visually similar images on Google and capturing screenshots of the retrieved results. These screenshots serve as an additional input image $I_2$, which may subsequently be fed into the large model along with the original question $Q$ and image $I_1$.

**2) Abstract feature**: In purely text-based large language models, the hidden states of the exact answer token can serve as key tokens for error detection [26]. We extend this observation to multimodal large language models and perform a layer-wise analysis, showing that multimodal fusion leads to more accurate error detection in multimodal settings. Specifically, we evaluate error-detection capability across different layers using the Idefics2-8B backbone on the OK-VQA dataset. Hidden states are extracted from every even-numbered Transformer layer as well as the final layer. We compare multiple feature configurations: textual hidden states at key token positions alone, and textual states combined with vision features using different pooling methods. As shown in Fig. 2, the results reveal several key insights into the internal decision-making process of the model.

**Primacy of Multimodal Fusion.** Error-detection methods that integrate visual features consistently outperform those based solely on textual hidden states across all layers. This observation indicates that visual representations play a critical role in assessing the internal certainty of the model and in deciding whether external retrieval is required.

**Layer-wise Evolution of Information.** Text-only features exhibit limited predictive capability in shallow layers ranging from layer 0 to layer 10, but their performance improves substantially in deeper layers. In contrast, multimodal features reach high accuracy much earlier, particularly within the intermediate layers from layer 2 to layer 16. This trend suggests that the alignment between visual and textual cues becomes sufficiently established at mid-network depths to effectively support retrieval gating.

[1]https://github.com/DearYoyoa/Multimodal-Adaptive-RAG

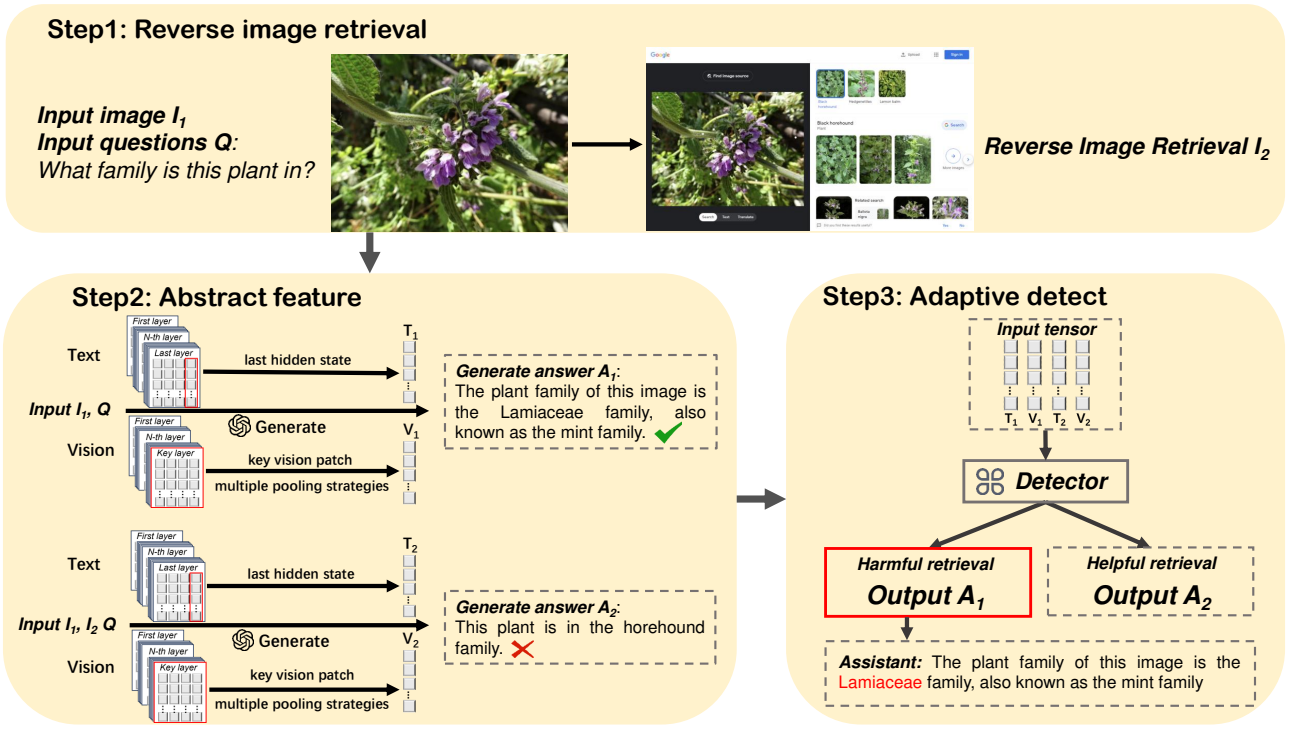


Fig. 1. Overview of the multimodal adaptive RAG framework

**Limited Sensitivity to Pooling Strategies.** Both average and maximum pooling over visual features result in similar high-accuracy regions for error detection. This suggests that, when visual information is globally aggregated, the detector mainly relies on image-level semantics, while the specific pooling operator has only a minor impact. Therefore, we

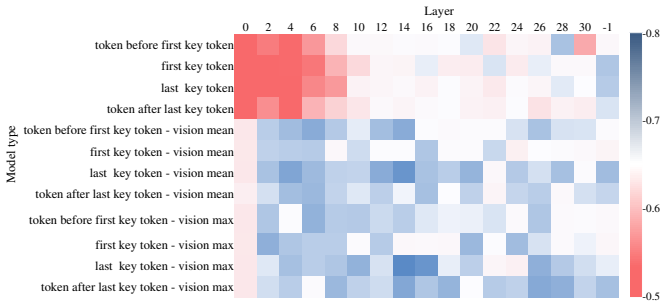


Fig. 2. Heatmap of the classifier accuracy across different key tokens and layers on the OK-VQA dataset using the Idefics2-8B model

leverage the model's internal textual and visual hidden states to train the adaptive classifier, as these representations preserve rich semantic and contextual information. Although key tokens serve as effective diagnostic signals that represent posterior knowledge in retrospective analysis, they are inaccessible during the inference phase of an adaptive system. Consequently, for the textual feature $T_1$, we utilize the hidden state from the final decoding step. This state synthesizes information from the question, the image, and the generated prefix, thereby naturally reflecting the model's current belief state. For the image feature, after feeding the input image into the model, multiple layers of visual representations are obtained. We select a specific intermediate layer and compute the average pooling over all patch embeddings within that layer. The resulting vector serves as the compact and informative vision feature $V_1$ that represents the input image. We jointly input $Q$, $I_1$, and $I_2$ into the model, following a similar procedure as described above, to extract the corresponding textual characteristic $T_2$ and vision feature $V_2$.

To effectively capture both semantic and visual cues, we concatenate the generated feature $T_1$, $V_1$, $T_2$, $V_2$ to form a unified representation $H_c$ for classification.

$$H_c = \text{Concat}(T_1, V_1, T_2, V_2)$$

**3) Adaptive detect**: The extracted data $H_c$ are utilized to train a four-class classifier. Considering both performance and efficiency, we adopt a multilayer perceptron architecture for the classifier, which is employed to evaluate the retrieval utility under different conditions as a retrieval gating mechanism.

$$\hat{y} = \arg\max_{y} \text{Softmax}(f(H_c))$$

where $\hat{y}$ represents the classification result. Specifically, there are four possible scenarios: 1) $S_1$: Both using external retrieval and not using external retrieval result in an incorrect answer. 2) $S_2$: Using external retrieval leads to a correct answer, while not using it results in an incorrect answer. 3) $S_3$: Using external

retrieval results in an incorrect answer, while not using it leads to a correct answer. 4) $S_4$: Both using external retrieval and not using external retrieval result in the correct answer.

At inference time, we introduce two alternative retrieval trigger strategies based on the classifier's four-way prediction. These strategies represent two opposing preferences regarding the reliance on external retrieval.

**RIR-Pessimistic Strategy.** This pessimistic-oriented strategy adopts a cautious stance toward external retrieval. Retrieval is triggered only when it is predicted to be essential, i.e. when using retrieved images leads to a correct answer, and not using them would result in an incorrect one. In all other scenarios, the model discards the retrieved content and relies solely on the original image and the question. This strategy minimizes the risk of introducing harmful noise, favoring the default path unless a clear benefit is identified.

$$R = \begin{cases} 1 & \text{if } \hat{y} = S_2 \\ 0 & \text{if } \hat{y} \neq S_2 \end{cases}$$

where R decides whether to use RIR.

**RIR-Optimistic Strategy.** This optimistic oriented strategy takes a more liberal approach to retrieval usage, reflecting a retrieval-favoring bias. It triggers external retrieval in all cases except when it is predicted to degrade performance, that is, when the retrieved images introduce noise and hurt the response quality. In this view, an external visual context is generally helpful and should be included unless explicitly harmful.

$$R = \begin{cases} 1 & \text{if } \hat{y} \neq S_3 \\ 0 & \text{if } \hat{y} = S_3 \end{cases}$$

When R is 1, choose to use RIR to generate the answer, that is, input the original image, screenshot image and question together into the model to generate the answer; otherwise, do not use the screenshot image.

The two strategies embody different stances toward external retrieval. Our subsequent experiments investigate their hierarchical impact across different datasets. This flexible decision layer enables the system to trade off robustness and completeness while adapting to the specific characteristics of the data or application domain.

## III. Experiments

In this section, we evaluate the effectiveness of the multimodal adaptive classifier across multiple datasets. We use model-judged accuracy as metrics for answer correctness. We systematically compare the performance of four configurations: zero-shot prompt, few-shot prompt, few-shot prompt with RIR, and our proposed method MMA-RAG, which adaptively determines whether to apply RIR based on a classifier under the few-shot setting.

### A. Datasets and Knowledge Bases

Knowledge-intensive VQA requires the integration of external information beyond images. Benchmarks such as OK-VQA [27] highlight this need, while previous work improves performance through knowledge extraction [28], cross-modal retrieval frameworks [21], and hybrid graph representations [29]. Encyclopedic-VQA [30] is a large-scale benchmark with 221K QA pairs and multiple images per query, focusing on fine-grained, instance-level understanding that requires encyclopedic knowledge, and has become a standard testbed for knowledge-intensive vision tasks [31], [32]. InfoSeek [33], [34] further expands this setting with 1.3M questions over 11K visual entities from various domains, combining human-annotated and automatically generated QA pairs to support large-scale visual information retrieval.

TABLE I
Training and evaluation sample sizes of the three datasets.

| Dataset | Infoseek | OK-VQA | E-VQA |
|---|---|---|---|
| Training | 3740 | 1000 | 1000 |
| Evaluation | 1646 | 4989 | 3345 |

Our dataset is randomly sampled from these public benchmarks and converted into screenshot-based inputs for training and evaluation, with statistics summarized in Table 1.

### B. Backbone Models and Metrics

We use several strong vision–language backbones. Idefics2-8B [35] and Idefics3-8B [36] support flexible image–text inputs and achieve competitive VQA performance. The Qwen2-VL series [37] introduces several key advancements in vision language models to enhance the model's ability to process images of varying resolutions, while Qwen2.5-VL further introduces unified image–video processing, improved cross-modal alignment, and adaptive visual tokenization.

Qwen2.5-Instruct is used as an automatic evaluator to assess answer correctness and compute accuracy.

### C. Baselines

**RIR** generates responses by incorporating reverse image retrieval [23]. Building upon RIR, we further consider several baselines that make adaptive decisions about whether to use RIR. **Chain-of-Thought (CoT)** encourages the model to produce explicit intermediate reasoning steps [38]. **P(true)** serves as a standard confidence-based baseline, which estimates correctness by calculating the probability that the model affirms its own answer [39]. **CLIP** measures image–text semantic alignment using joint embeddings [40].

## IV. Results

### A. Main Results

We trained the classifier for Multimodal Adaptive Retrieval Augmented Generation using visual features and textual hidden states and conducted experiments on three datasets: InfoSeek, OK-VQA, and Encyclopedic-VQA. Given the varying class distributions across datasets, we apply dataset-specific class weights to balance the contribution of different classes during training, thereby ensuring the reliability and validity of the experiments. The retrieval process involved obtaining similar images through Google search, capturing screenshots

TABLE II
ACCURACY SCORES ON THE INFOSEEK, E-VQA, AND OK-VQA TEST DATASETS, USING IDEFICS2-8B, IDEFICS3-8B, AND QWEN2VL-7B AS BACKBONES, WITH QWEN25_INSTRUCT USED TO JUDGE ANSWER CORRECTNESS. **BOLD** INDICATES STATE-OF-THE-ART PERFORMANCE

| Model | Method | InfoSeek | OK-VQA | E-VQA |
|---|---|---|---|---|
| Qwen2VL | Zero shot | 19.9 | 58.7 | 10.0 |
| | Few shot | 15.9 | 58.5 | 14.1 |
| | RIR | 23.3 | 62.2 | 19.8 |
| | CoT | 20.4 | 46.7 | 14.4 |
| | P(true) | 22.6 | 53.3 | 19.1 |
| | CLIP | 20.9 | 56.0 | 18.5 |
| | **MMA-RAG** | **23.9** | **62.4** | **20.0** |
| Idefics2 | Zero shot | 15.1 | 53.8 | 13.0 |
| | Few shot | 14.2 | 58.7 | 12.7 |
| | RIR | 17.2 | 56.7 | 14.1 |
| | CoT | 15.5 | 49.0 | 14.1 |
| | P(true) | 14.1 | 49.1 | 13.8 |
| | CLIP | 16.4 | 54.8 | 14.3 |
| | **MMA-RAG** | **20.3** | **60.1** | **14.6** |
| Idefics3 | Zero shot | 12.5 | 43.2 | 11.9 |
| | Few shot | 8.5 | 46.0 | 6.3 |
| | RIR | 17.5 | 56.6 | 12.5 |
| | CoT | 14.3 | 43.7 | 10.9 |
| | P(true) | 12.9 | 43.7 | 11.0 |
| | CLIP | 12.6 | 50.0 | 10.0 |
| | **MMA-RAG** | **18.1** | **58.3** | **12.6** |

and using these screenshots alongside the original images and questions as input to generate final answers. The classifier's role was to assess the utility of such retrievals; if beneficial for generating correct answers, external retrieval was employed; otherwise, it was omitted. During the testing process, Qwen2.5-instruct was used to evaluate whether the answers were correct and the final evaluation metric was accuracy. Tab. II shows the main experimental results.

Across different models and datasets, the performance of zero-shot and few-shot prompts is inconsistent, primarily due to variations in dataset types and task characteristics. To ensure experimental validity, we adopted the few-shot prompt in both the RIR and MMA-RAG experiments. RIR systems that use Google to retrieve similar images have significantly improved the accuracy of the generated answers. However, in certain cases, the use of retrieved external images leads to a degradation in response quality, producing incorrect responses that would not have occurred if the model had relied solely on the original image and question. We refer to such instances as "harmful samples", where the introduction of additional visual information inadvertently misguides the model. Compared with the original RIR model, the multimodal adaptive RAG model achieves varying degrees of improvement in three datasets. Compared with the confidence-based P(true) baseline, the auxiliary-model-based CLIP approach, and reasoning-based CoT method, MMA-RAG consistently delivers substantial and statistically significant performance gains across all evaluated datasets. This improvement is primarily attributed to the model's ability to predict and suppress the influence of harmful samples, thus avoiding unnecessary or detrimental retrievals and preserving the model's ability to generate correct answers.

For certain model configurations on OK-VQA and E-VQA, MMA-RAG yields only marginal improvements, likely because the proportion of harmful samples introduced by RIR is low, leaving limited room for denoising. This also indicates that MMA-RAG is relatively safe: when retrieval is beneficial, the gating mechanism tends to keep retrieval enabled and is unlikely to introduce adverse effects.

### B. Feature Robustness

We further conducted exploratory experiments on the extraction of visual features, including selecting patch embeddings from different Transformer layers and applying various pooling strategies, such as average pooling and max pooling to obtain global image representations.

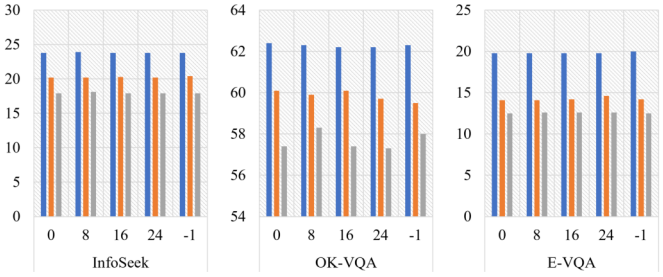


Fig. 3. Impact of layer selection on visual feature representation

As illustrated in Fig. 3, the vertical axis represents the response accuracy and the horizontal axis corresponds to the layer index. Despite differences in feature extraction strategies, overall performance variation remained relatively small. This limited variance can be attributed to the robustness of the downstream classifier to feature extraction strategies. Since the middle and later layers of vision transformers already produce stable semantic representations, and pooling differences are largely smoothed out by the subsequent MLP, the overall performance is less sensitive to such variations.

### C. Ablation Study

The consistent improvements observed across the datasets can be attributed to the effective exploitation of internal hidden states by the multimodal adaptive RAG model. In particular, both the textual hidden states, which encapsulate the semantic representation and contextual reasoning derived from the question and the generated response, and the visual features, which encode salient spatial and perceptual information from the input image, play a pivotal role in informing adaptive retrieval decisions. To better understand the individual and combined contributions of these features, we conducted a series of ablation studies, analyzing their respective impacts on the classifier's performance within the MMA-RAG framework.

In this experiment, we trained the classifier on three datasets using only hidden textual states or only visual features, while keeping all other settings unchanged.

As shown in the results of Tab. III , given the limited accuracy of large models in generating answers, classifiers

TABLE III
PERFORMANCE OF ABLATION STUDY ON MULTIMODAL ADAPTIVE RAG WITH IDEFICS2-8B. THE PERFORMANCE OF THE MODEL IS JOINTLY INFLUENCED BY TEXTUAL AND VISUAL FEATURES.

| Method | Infoseek | OK-VQA | E-VQA |
|---|---|---|---|
| RIR | 17.2 | 56.7 | 14.1 |
| wo-text | 17.9 | 58.6 | 14.2 |
| wo-vision | 19.3 | 59.9 | 14.2 |
| **MM Adaptive RAG** | **20.3** | **60.1** | **14.6** |

that incorporate visual features outperform those relying solely on hidden text states. This suggests that in VQA tasks, extracting and utilizing visual features may enhance the ability to determine the effectiveness of external retrievals. Similarly, classifiers that lack textual hidden states have also exhibited a decline in performance. This has demonstrated that hidden textual states inherently contain implicit cues regarding the accuracy of the generated answers. Therefore, by effectively integrating hidden text states with visual features, we were able to extract maximally valuable information, thus enhancing our understanding of the model decision-making process.

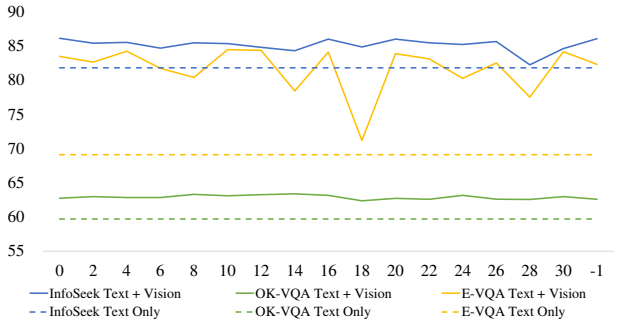


Fig. 4. Layer-wise comparison of classifier accuracy using textual and multimodal features on the InfoSeek, OK-VQA, and E-VQA datasets

Additionally, we conduct a layer-wise comparative study using the Idefics2-8B model to compare classifiers that rely solely on textual features or visual features with those that jointly incorporate textual and visual features. As shown in Fig. 4 and Fig. 5, the vertical axis represents the classifier accuracy, while the horizontal axis corresponds to different network layers. The combined use of textual and visual features consistently leads to improved accuracy across datasets, and both modalities are indispensable across all layers. These results further validate the effectiveness of our approach.

### *D. RIR Strategy Comparison*

To further examine the impact of retrieval trigger policies, we conduct a comparative study between *RIR-Optimistic* and *RIR-Pessimistic* strategies. Experiments are performed with Idefics2-8B and Idefics3-8B on three datasets: InfoSeek, OK-VQA, and E-VQA. For each dataset, we report the response accuracy under both strategies while keeping all other settings fixed.

As shown in Fig. 6, a clear dataset-dependent preference emerges between the two retrieval strategies, where the vertical

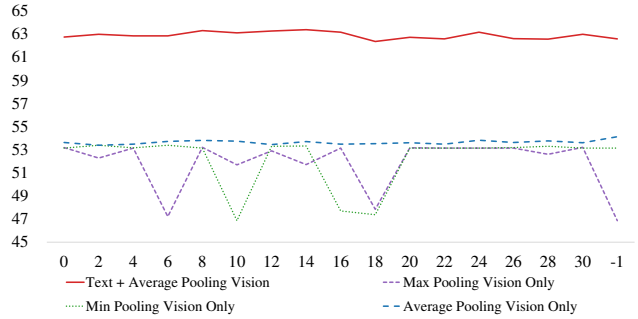


Fig. 5. Layer-wise comparison of classifier accuracy on OK-VQA dataset

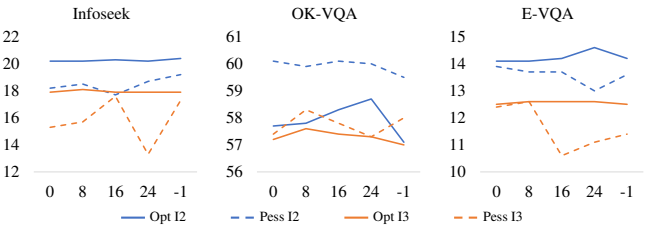


Fig. 6. Performance comparison between different RIR strategies

axis denotes the response accuracy and the horizontal axis corresponds to the layer index. On OK-VQA, the RIR-Pessimistic Strategy consistently outperforms the RIR-Optimistic Strategy, whereas on InfoSeek and E-VQA the opposite trend is observed.

This divergence stems from the different roles of external visual information across datasets. In OK-VQA, which emphasizes common-sense reasoning, retrieved images often introduce irrelevant noise, making the RIR-Pessimistic Strategy more robust. In contrast, InfoSeek and E-VQA rely on fine-grained, instance-level knowledge, where external visual evidence provides useful complementary cues and improves accuracy. Consistent trends across the backbone indicate that strategy preference is driven by dataset characteristics, highlighting the need for adaptive retrieval policies.

## V. CONCLUSION

In this paper, we propose MMA-RAG, a multimodal adaptive RAG framework that leverages internal visual and textual representations to predict retrieval utility and selectively incorporate external evidence. By filtering harmful retrieval, it improves answer correctness and robustness in knowledge-intensive VQA. Experiments across benchmarks and backbones confirm consistent gains over standard RAG and existing baselines.

## ACKNOWLEDGMENT

This research work is supported by the National Key Research and Development Program under Grant 2023YFC3305203, and the National Natural Science Foundation of China under Grant 62576339.

## REFERENCES


[1] M. Shao, A. Basit, R. Karri, and M. Shafique, "Survey of different large language model architectures: Trends, benchmarks, and challenges," *IEEE Access*, 2024.

[2] S. Li, Y. Yu, H. Yang, Y. Zhou, and C. Ji, "Him-rag: A heuristic framework for iterative multi-source retrieval-augmented generation," in *2025 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2025, pp. 1–8.

[3] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1–55, 2025.

[4] D. Park, Z. Qian, G. Han, and S.-N. Lim, "Mitigating dialogue hallucination for large vision language models via adversarial instruction tuning," *arXiv preprint arXiv:2403.10492*, 2024.

[5] Y. Huang and J. Huang, "A survey on retrieval-augmented text generation for large language models," *arXiv preprint arXiv:2404.10981*, 2024.

[6] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, vol. 2, p. 1, 2023.

[7] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[8] S. Wu, Y. Xiong, Y. Cui, H. Wu, C. Chen, Y. Yuan, L. Huang, X. Liu, T.-W. Kuo, N. Guan *et al.*, "Retrieval-augmented generation for natural language processing: A survey," *arXiv preprint arXiv:2407.13193*, 2024.

[9] S. Peng, W. Zhang, and D. Qu, "Think before retrieving: Query-response relevance retrieval augmented generation," in *2025 International Joint Conference on Neural Networks (IJCNN)*, 2025, pp. 1–8.

[10] X. Sun, Z. Chen, X. Zheng, Q. Liu, S. Wu, B. Song, Z. Wang, W. Wang, and L. Wang, "Kbqa-r1: Reinforcing large language models for knowledge base question answering," *arXiv preprint arXiv:2512.10999*, 2025.

[11] L. Chen, X. Yang, Y. Lu, J. Zhang, X. Sun, Q. Liu, S. Wu, J. Dong, and L. Wang, "Poisonarena: Uncovering competing poisoning attacks in retrieval-augmented generation," *arXiv preprint arXiv:2505.12574*, 2024.

[12] X. Sun, Z. Chen, Q. Liu, S. Wu, B. Song, W. Wang, Z. Wang, and L. Wang, "Predict the retrieval! test time adaptation for retrieval augmented generation," *arXiv preprint arXiv:2601.11443*, 2024.

[13] P. Zhao, H. Zhang, Q. Yu, Z. Wang, Y. Geng, F. Fu, L. Yang, W. Zhang, J. Jiang, and B. Cui, "Retrieval-augmented generation for ai-generated content: A survey," *arXiv preprint arXiv:2402.19473*, 2024.

[14] X. Sun, J. Xie, Z. Chen, Q. Liu, S. Wu, Y. Chen, B. Song, W. Wang, Z. Wang, and L. Wang, "Divide-then-align: Honest alignment based on the knowledge boundary of rag," *arXiv preprint arXiv:2505.20871*, 2024.

[15] R. Zhao, H. Chen, W. Wang, F. Jiao, X. L. Do, C. Qin, B. Ding, X. Guo, M. Li, X. Li *et al.*, "Retrieving multimodal information for augmented generation: A survey," *arXiv preprint arXiv:2303.10868*, 2023.

[16] M. Yasunaga, A. Aghajanyan, W. Shi, R. James, J. Leskovec, P. Liang, M. Lewis, L. Zettlemoyer, and W.-t. Yih, "Retrieval-augmented multimodal language modeling," *arXiv preprint arXiv:2211.12561*, 2022.

[17] Q. Yu, Z. Xiao, B. Li, Z. Wang, C. Chen, and W. Zhang, "Mramg-bench: A beyondtext benchmark for multimodal retrieval-augmented multimodal generation," *arXiv preprint arXiv:2502.04176*, 2025.

[18] W. Zhai, "Sam-rag: An self-adaptive framework for multimodal retrieval-augmented generation," in *2025 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2025, pp. 1–8.

[19] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[20] Y. Lin, Y. Xie, D. Chen, Y. Xu, C. Zhu, and L. Yuan, "Revive: Regional visual representation matters in knowledge-based visual question answering," *Advances in neural information processing systems*, vol. 35, pp. 10 560–10 571, 2022.

[21] W. Chen, H. Hu, X. Chen, P. Verga, and W. W. Cohen, "Murag: Multimodal retrieval-augmented generator for open question answering over images and text," *arXiv preprint arXiv:2210.02928*, 2022.

[22] B. Chen, A. Khare, G. Kumar, A. Akula, and P. Narayana, "Seeing beyond: Enhancing visual question answering with multi-modal retrieval," in *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, 2025, pp. 410–421.

[23] J. Xu, M. Moor, and J. Leskovec, "Reverse image retrieval cues parametric memory in multimodal llms," *arXiv preprint arXiv:2405.18740*, 2024.

[24] Y. Li, Y. Li, X. Wang, Y. Jiang, Z. Zhang, X. Zheng, H. Wang, H.-T. Zheng, F. Huang, J. Zhou *et al.*, "Benchmarking multimodal retrieval augmented generation with dynamic vqa dataset and self-adaptive planning agent," *arXiv preprint arXiv:2411.02937*, 2024.

[25] L. Mei, S. Mo, Z. Yang, and C. Chen, "A survey of multimodal retrieval-augmented generation," *arXiv preprint arXiv:2504.08748*, 2025.

[26] H. Orgad, M. Toker, Z. Gekhman, R. Reichart, I. Szpektor, H. Kotek, and Y. Belinkov, "Llms know more than they show: On the intrinsic representation of llm hallucinations," *arXiv preprint arXiv:2410.02707*, 2024.

[27] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "Ok-vqa: A visual question answering benchmark requiring external knowledge," in *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 2019, pp. 3195–3204.

[28] J. Wu, J. Lu, A. Sabharwal, and R. Mottaghi, "Multi-modal answer validation for knowledge-based vqa," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 3, 2022, pp. 2712–2721.

[29] Z. Zhu, J. Yu, Y. Wang, Y. Sun, Y. Hu, and Q. Wu, "Mucko: Multi-layer cross-modal knowledge reasoning for fact-based visual question answering," *arXiv preprint arXiv:2006.09073*, 2020.

[30] T. Mensink, J. Uijlings, L. Castrejon, A. Goel, F. Cadar, H. Zhou, F. Sha, A. Araujo, and V. Ferrari, "Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3113–3124.

[31] G. Van Horn, E. Cole, S. Beery, K. Wilber, S. Belongie, and O. Mac Aodha, "Benchmarking representation learning for natural world image collections," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 884–12 893.

[32] T. Weyand, A. Araujo, B. Cao, and J. Sim, "Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2575–2584.

[33] Y. Chen, H. Hu, Y. Luan, H. Sun, S. Changpinyo, A. Ritter, and M.-W. Chang, "Can pre-trained vision and language models answer visual information-seeking questions?" *arXiv preprint arXiv:2302.11713*, 2023.

[34] H. Hu, Y. Luan, Y. Chen, U. Khandelwal, M. Joshi, K. Lee, K. Toutanova, and M.-W. Chang, "Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12 065–12 075.

[35] H. Laurençon, L. Tronchon, M. Cord, and V. Sanh, "What matters when building vision-language models?" *Advances in Neural Information Processing Systems*, vol. 37, pp. 87 874–87 907, 2024.

[36] H. Laurençon, A. Marafioti, V. Sanh, and L. Tronchon, "Building and better understanding vision-language models: insights and future directions," in *Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models*, 2024.

[37] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[38] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[39] S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson *et al.*, "Language models (mostly) know what they know," *arXiv preprint arXiv:2207.05221*, 2022.

[40] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.