ConfRAG: Confidence-Calibrated RAG for Hallucination-Free Visual Question Answering

Zhangchi Feng
CCSE, Beihang University
Beijing, China
zcmuller@buaa.edu.cn

Jingyuan Wang
CCSE, Beihang University
Beijing, China
wangjingyuan03@buaa.edu.cn

Yangyifei Luo
CCSE, Beihang University
Beijing, China
luoyangyifei@buaa.edu.cn

Dongdong Kuang
CCSE, Beihang University
Beijing, China
kuangdd@buaa.edu.cn

Zhongyuan Wang
CCSE, Beihang University
Beijing, China
wangzy23@buaa.edu.cn

Yaowei Zheng
CCSE, Beihang University
Beijing, China
hiyouga@buaa.edu.cn

Richong Zhang*
CCSE, Beihang University
Beijing, China
zhangrc@act.buaa.edu.cn

Abstract

Vision Large Language Models (VLLMs) have achieved remarkable success in visual question answering, but suffer from critical hallucination problems, generating confident-sounding but factually incorrect responses. While Retrieval-Augmented Generation (RAG) offers promising solutions, existing multi-modal RAG approaches face three key limitations: retrieval strategies that ignore model confidence, lack of effective hallucination detection, and models not trained to express uncertainty. We propose ConfRAG, a confidence-calibrated retrieval-augmented generation framework that systematically addresses these limitations through three core innovations. First, our confidence-aware retrieval mechanism employs several confidence thresholds to filter high-quality evidence during both image-based and web-based retrieval. Second, our hybrid hallucination detection module uses practical rules—generation termination analysis and average token probability assessment—to identify unreliable content. Third, our IDK-aware training strategy independently optimizes three specialized pipelines (direct QA, image RAG, web RAG) using quality-based sampling to teach appropriate uncertainty expression. Comprehensive experiments on the Meta CRAG-MM challenge demonstrate ConfRAG's effectiveness, achieving 7th place overall with consistent performance across all three tasks. Notably, our IDK training transforms severely negative baselines into positive performance, demonstrating dramatic hallucination reduction while maintaining competitive accuracy. Our code and data are available at https://github.com/BUAADreamer/ConfRAG.

*Corresbonding author: zhangrc@act.buaa.edu.cn.

KDD Cup '25, Toronto, Canada

CCS Concepts

• **Computing methodologies** → **Natural language generation**;
• **Information systems** → **Language models**.

Keywords

Multi-modal Retrieval, Visual Question Answering

ACM Reference Format:
Zhangchi Feng, Jingyuan Wang, Yangyifei Luo, Dongdong Kuang, Zhongyuan Wang, Yaowei Zheng, and Richong Zhang. 2025. ConfRAG: Confidence-Calibrated RAG for Hallucination-Free Visual Question Answering. In *KDDCup Workshop of SIGKDD '25: KDDCup Workshop of the 31th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 7 pages.

1 Introduction

Vision Large Language Models (VLLMs) have achieved remarkable success in visual question answering (VQA) [8, 14], demonstrating sophisticated multi-modal understanding capabilities. However, they suffer from a critical hallucination problem: generating confident-sounding but factually incorrect responses. This issue is particularly severe when handling long-tail entities, complex multi-hop reasoning, or queries requiring integration of recognition, OCR, and knowledge retrieval capabilities. The fundamental challenge lies in the models' tendency to generate plausible responses even when lacking sufficient information, leading to a trade-off between informativeness and reliability.

While Retrieval-Augmented Generation (RAG) offers a promising solution by grounding responses in external knowledge [3, 5, 13], existing multi-modal RAG approaches face three key limitations. First, traditional retrieval strategies rely solely on similarity metrics without considering model confidence, potentially incorporating misleading information. Second, current systems lack effective hallucination detection mechanisms to assess response reliability. Third, models are not explicitly trained to express uncertainty, resulting in overconfident responses even in knowledge-insufficient scenarios.

We propose ConfRAG, a confidence-calibrated retrieval-augmented generation framework that systematically addresses these limitations through three core innovations. Our confidence-aware retrieval mechanism employs dual-stage confidence thresholds to filter high-quality evidence during both image-based and web-based retrieval. The hallucination detection module uses two practical rules—generation termination analysis and average token probability assessment—to identify unreliable content. Most importantly, our IDK-aware training strategy independently optimizes three specialized pipelines (direct QA, image RAG, web RAG) using quality-based sampling to teach appropriate uncertainty expression.

ConfRAG employs a cascaded architecture that progresses through these pipelines based on confidence assessment, preferring simpler approaches when they demonstrate adequate reliability. This design reflects our empirical finding that direct QA achieves 20% accuracy, making it worthwhile to attempt before engaging complex retrieval, while web-based retrieval shows higher information recall than image-based approaches.

Comprehensive experiments on the CRAG-MM benchmark validate our approach's effectiveness. ConfRAG achieves 7th place overall with consistent performance across all three tasks: 8th place in Task 1 (single-source augmentation), 9th place in Task 2 (multi-source augmentation), and 7th place in Task 3 (multi-turn QA). Notably, our IDK training transforms severely negative baselines (scores ranging from -96 to -141) into positive performance, demonstrating dramatic hallucination reduction while maintaining competitive accuracy.

The key contributions of this work are: (1) A confidence-calibrated RAG framework with dual-threshold retrieval filtering and hybrid hallucination detection, (2) An IDK-aware training strategy with pipeline-specific optimization and quality-based sampling, (3) A cascaded architecture design informed by empirical analysis of different information sources, and (4) Comprehensive experimental validation achieving top-10 rankings across all CRAG-MM tasks with consistent hallucination suppression.

2 ConfRAG Framework

The ConfRAG framework consists of several interconnected components designed to work synergistically for hallucination-free visual question answering. We present the overall architecture and detail each component's functionality and design rationale.

2.1 Overall Architecture

ConfRAG employs a modular architecture that seamlessly integrates visual and textual processing pipelines with confidence-calibrated retrieval mechanisms. The framework processes queries through multiple stages:

Given an input query q and an associated image I, the system first extracts relevant features and constructs retrieval queries. The visual pipeline processes the image through a vision encoder to extract semantic features, while simultaneously performing entity detection and grounding. The textual pipeline analyzes the query to identify key phrases and construct effective search queries.

The retrieval stage operates on two parallel tracks: image-based retrieval and web-based retrieval. For image retrieval, the system searches a knowledge graph indexed by images, retrieving similar images with structured information. For web retrieval, the system queries a text corpus to obtain relevant passages.

The retrieved information is then processed through a reranking module that uses confidence scores to prioritize the most relevant and reliable evidence. Finally, the generation module synthesizes the retrieved information with the original query and image to produce a grounded answer, with confidence calibration applied throughout to detect and suppress potential hallucinations.

2.2 Confidence-Aware Retrieval

The confidence-aware retrieval mechanism is central to ConfRAG's ability to suppress hallucinations. Traditional retrieval methods typically rely on similarity rank alone, which can lead to the inclusion of superficially related but ultimately misleading information. Our approach incorporates model confidence at multiple levels to improve retrieval quality.

For image-based retrieval, we employ a two-stage grounding process. First, we use a multimodal large language model to generate descriptive phrases from the input image. The generated phrases are then processed by a visual grounding model to localize entities within the image. The grounding process utilizes two confidence thresholds: a text confidence threshold $\tau_{text} = 0.3$ and a bounding box confidence threshold $\tau_{box} = 0.2$. High-confidence detections are defined as:

$$\mathcal{E}_{high} = \{e_i | c_{text}(e_i) > \tau_{text} \text{ and } c_{box}(e_i) > \tau_{box}\}$$

The retrieval process follows a coarse-to-fine approach, first performing initial retrieval to obtain candidate images, then applying reranking to refine the results. The reranking scores are used as confidence measures, and we filter out irrelevant images based on a similarity threshold to ensure only highly relevant visual content is retained.

For web-based retrieval, we first generate descriptive phrases from the input query using the same multimodal model. These phrases are then used to search for similar web page chunks through text-based retrieval. The web retrieval also follows the coarse-to-fine paradigm: initial text similarity matching followed by reranking. Similar to image retrieval, we apply similarity thresholds on the reranked results to filter out irrelevant page chunks, ensuring only high-quality textual information is included in the final retrieved set.

We note that while confidence thresholds could theoretically be applied at both the coarse retrieval stage and the reranking stage, our empirical evaluation revealed that applying thresholds during the initial coarse retrieval phase leads to suboptimal performance. Through extensive experimentation, we found that filtering candidates too aggressively in the early stages can exclude potentially relevant content that might be correctly identified through the more sophisticated reranking process. Therefore, our final design applies confidence-based filtering only at the reranking stage, allowing the initial retrieval to maintain sufficient recall while relying on the refined confidence scores from reranking to ensure precision.

2.3 Hybrid Hallucination Detection Module

The hybrid hallucination detection module operates on generated responses to identify potentially unreliable content. Rather than

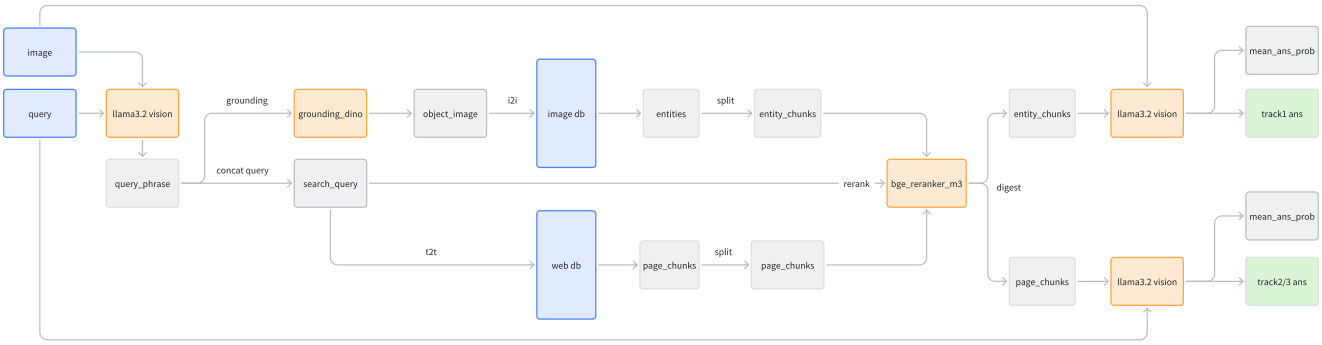

Figure 1: ConfRAG Inference Pipeline

employing complex multi-faceted approaches, we implement a simple yet effective two-rule confidence assessment strategy.

Our first rule examines the generation termination condition. When the model stops generation due to reaching the maximum length limit rather than naturally concluding with an end-of-sequence token, we interpret this as an indicator of low confidence. Overly verbose responses often suggest that the model lacks certainty and is attempting to compensate through verbosity. In such cases, we automatically classify the response as unreliable and replace it with an "I don't know" (IDK) response. Specifically, when the generated sequence length exceeds 75 tokens, we treat this verbosity as a low-confidence signal and trigger the same IDK fallback.

The second rule involves computing the average token probability across the entire generated sequence. For a generated sequence of tokens $t_1, t_2, ..., t_n$, we calculate:

$$\bar{p} = \frac{1}{n} \sum_{i=1}^{n} p(t_i | t_{<i})$$

where $p(t_i|t_{<i})$ represents the probability of token t_i given the preceding context. This average probability is then compared against a predefined threshold τ. If $\bar{p} < \tau$, indicating insufficient confidence, the response is replaced with an IDK response.

In practice, there is a controllable coverage–accuracy trade-off. Increasing the maximum generation length caps and/or lowering the mean token probability threshold τ raises the proportion of answered responses (reducing IDK rate), whereas decreasing the caps and/or raising τ yields more conservative behavior with fewer hallucinations but more IDK. We select operating points by sweeping these knobs on a held-out set to satisfy a target balance between coverage and reliability.

Based on this confidence assessment mechanism, we establish a cascaded question-answering framework consisting of three sequential stages: direct QA, image-guided RAG QA, and web-guided RAG QA. The system progresses to the next stage only when the current stage produces a response with insufficient confidence, ensuring that simpler approaches are preferred when they demonstrate adequate reliability. The rationale behind this cascaded design is twofold: first, we observe that the model's direct question-answering capability achieves approximately 20% accuracy, making it worthwhile to attempt direct responses before engaging more complex retrieval mechanisms. Second, our empirical analysis reveals that web-based retrieval demonstrates higher effective information recall rates compared to image-based retrieval, justifying the progression from image-guided to web-guided RAG when enhanced retrieval becomes necessary. Practically, this ordering lets the stronger pipeline (web-RAG) serve as a safety backstop, while lighter pipelines (direct QA, image-RAG) handle easy cases first at lower cost. Depending on application needs, the order can be rearranged (e.g., web before image), or all three pipelines can be executed in parallel with a final merge module to produce the final answer.

2.4 IDK-Aware Training Strategy

A critical innovation in ConfRAG is the pipeline-specific training strategy, which recognizes the distinct characteristics of different information retrieval approaches. Our analysis reveals that the three core pipelines—direct QA, image RAG, and web RAG—operate with fundamentally different mechanisms and information sources, leading us to adopt an independent training approach for each pipeline.

The training process consists of four key steps. First, we observe that the direct QA, image RAG, and web RAG pipelines have minimal interdependence, prompting us to decouple their training procedures. Second, for each question in the training set, we obtain complete input-output pairs from all three pipelines and employ a large language model to score each output against the ground truth answer, classifying responses as either correct or incorrect.

Third, we construct supervised fine-tuning (SFT) datasets using a quality-based sampling strategy. Outputs evaluated as correct by the scoring model are directly incorporated as positive training examples, while those deemed incorrect are modified to "I don't know" responses, creating a balanced dataset of both confident answers and uncertainty expressions:

$$\mathcal{D}_{pipeline} = \mathcal{D}_{correct} \cup \mathcal{D}_{idk}$$

where $\mathcal{D}_{correct}$ contains high-quality question-answer pairs and $\mathcal{D}_{idk}$ includes uncertainty-expressing examples.

Finally, we perform pipeline-specific fine-tuning using each pipeline's curated dataset to train specialized multimodal models. During inference, we deploy the corresponding specialized model

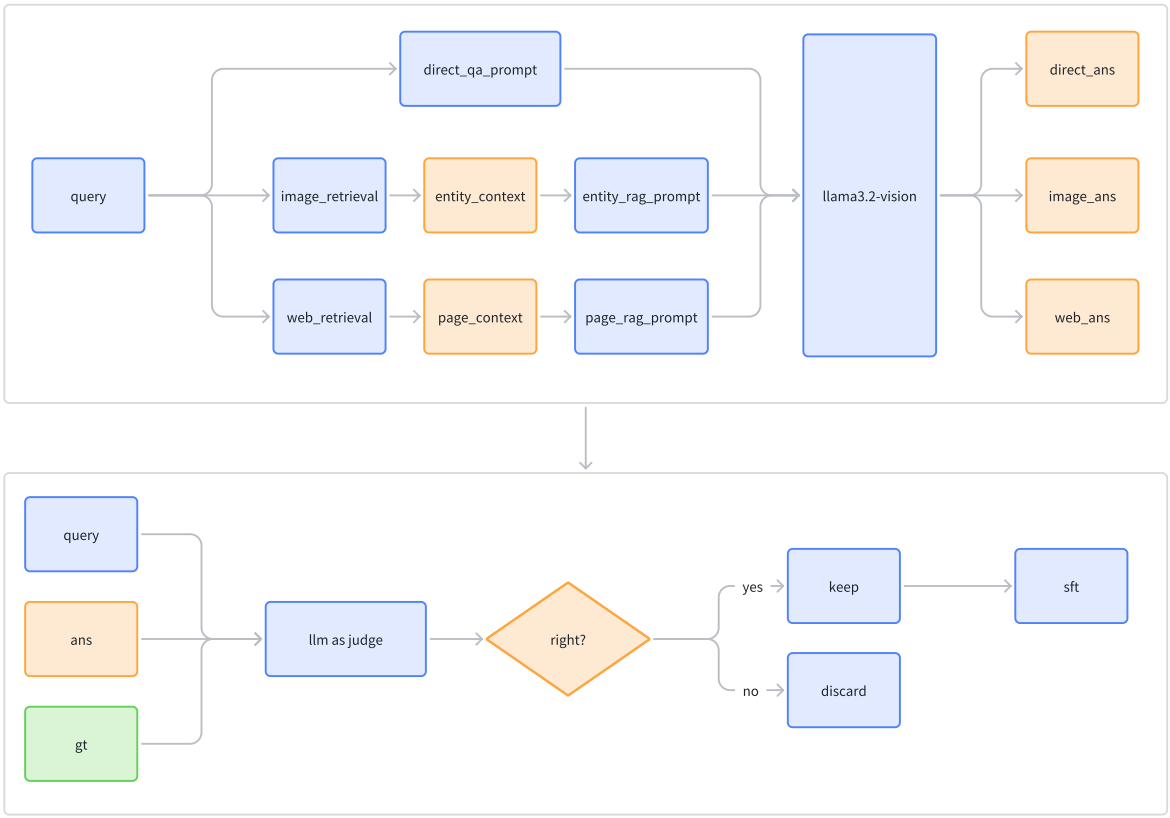

Figure 2: Hallucination Suppression Training Pipeline

for each pipeline, ensuring optimal performance tailored to the specific characteristics and requirements of each information retrieval approach.

2.5 Additional Experimental Explorations

In the course of developing ConfRAG, we conducted additional experiments on two promising approaches that, while showing some merit, ultimately provided limited improvements to the overall pipeline performance.

2.5.1 Off-the-shelf IDK Classification. Rather than having multimodal large language models directly determine "I don't know" responses, we explored using dedicated off-the-shelf models as specialized IDK classifiers. This approach aimed to separate the classification task from the generation task, potentially improving the reliability of uncertainty detection.

Given the CRAG-MM competition's model size constraint of 1.5B parameters, we experimented with two compact yet capable models: Qwen3-0.6B [10] and InternVL3-1B [2]. The IDK classifier was designed to take the original query and image (image only for internvl) as input and output a binary decision indicating whether the model should respond with "I don't know" instead of attempting to generate an answer.

We use LlamaFactory [15] to train the classifier. However, both models demonstrated suboptimal performance in this specialized classification task. The dedicated IDK classifiers failed to achieve the desired balance between appropriate uncertainty expression and maintaining reasonable answer coverage. The models either exhibited excessive conservatism, classifying too many answerable questions as requiring IDK responses, or insufficient caution, missing cases where uncertainty expression would be appropriate. This poor calibration ultimately led to performance degradation compared to our integrated approach where the same model handles both answer generation and uncertainty assessment.

2.5.2 CLIP-based Composed Image Retrieval. We investigated enhancing our image retrieval capabilities through composed image retrieval [4] using CLIP, which combines textual and visual information to construct more sophisticated query representations for image-to-image search.

Our approach leveraged CLIP's ability to encode both images and text into a shared embedding space. For a given query containing both textual description and visual content, we computed separate CLIP embeddings for the text $\mathbf{t}$ and the image $\mathbf{i}$, then combined them through summation followed by L2 normalization to create a unified query representation $\mathbf{q}_u$:

$$\mathbf{q}_u = \frac{\text{CLIP}_{text}(\mathbf{t}) + \text{CLIP}_{image}(\mathbf{i})}{||\text{CLIP}_{text}(\mathbf{t}) + \text{CLIP}_{image}(\mathbf{i})||_2}$$

To train this compositional image retrieval model, we performed semantic search between each query question and all entity items using Qwen3-Embedding-8B. For each query, we retrieved its top-5 most semantically similar entities based on similarity scores. We used these top-5 items as pseudo-labels to provide supervisory signals for training. These retrieved entities were then mapped to their associated target images, resulting in query-target image pairs for compositional alignment training.

To facilitate smoother alignment during the compositional retrieval training process, we froze the parameters of the CLIP image encoder, which is consistent with the encoder used to produce the provided image embeddings, and trained only the text encoder.

The training process optimized the alignment between the compositional query representations and the target image embeddings using contrastive learning objectives. We progressively investigated the impact of various training configurations and hyperparameters on retrieval performance.

Experimental results demonstrated incremental improvements over baseline image retrieval:

Table 1: CLIP-based Composed Image Retrieval Results

Method	Recall@5
Pure image-to-image retrieval	0.22
Zero-shot composed image retrieval	0.26
Training on pseudo-labeled dataset	0.31
w increased batch size	0.336

The final configuration achieved a notable 53% relative improvement over the baseline image retrieval approach.

However, when integrated into the complete ConfRAG pipeline, the compositional image retrieval approach achieved slightly lower overall performance compared to our main grounding-based method. While the improved recall in isolated retrieval evaluation was promising, the computational overhead and the marginal gains in end-to-end question answering accuracy did not justify the added complexity. This finding reinforced our decision to prioritize the more efficient and equally effective grounding-based approach in our final framework design.

3 Experiments

We conduct comprehensive experiments to evaluate the effectiveness of ConfRAG on the CRAG-MM benchmark, which provides a challenging testbed for multi-modal question answering systems. Our experiments are designed to assess both the accuracy of generated answers and the reduction in hallucination rates.

3.1 Experimental Setup

We implement ConfRAG using a comprehensive set of state-of-the-art models optimized for different components of our framework. For the core multimodal large language model, we employ Meta's Llama 3.2 Vision 11B model [7] as the base architecture, which provides strong visual-language understanding capabilities. We use unsloth[1] to fine-tune Llama 3.2 Vision with Lora.

For visual processing components, we utilize OpenAI's clip-vit-large-patch14-336 model [9] for image retrieval, which excels at cross-modal similarity matching between images and text queries. The visual grounding module employs IDEA's (International Digital Economy Academy) grounding-dino-base model [6], enabling precise entity localization with the confidence thresholds described in our approach.

The text retrieval pipeline incorporates BAAI's bge-large-en-v1.5 model [12] for initial text similarity matching during the coarse retrieval stage, followed by BAAI's bge-reranker-v2-m3 model [1] for refined reranking to improve retrieval precision. This two-stage approach ensures both high recall and precision in text-based evidence gathering.

For evaluation and quality assessment during training, we employ Alibaba's Qwen-Plus API as the scoring model to automatically evaluate response quality and guide our IDK-aware training strategy.

All experiments are conducted on NVIDIA L20s GPUs with 48GB memory. The the CRAG-MM challenge benchmark[2] consists of three tasks: single-source augmentation (Task1), multi-source augmentation (Task2), and multi-turn conversational QA (Task3) [11].

For evaluation metrics, we use the standard CRAG-MM scoring system: Perfect (1.0), Acceptable (0.5), Missing (0.0), and Incorrect (-1.0). We report accuracy, hallucination count, missing count, and overall score.

3.2 Main Results

3.2.1 Local Validation Results. We first present results from local experiments using 20% sampled data (388 questions) from the public test set to validate our approach.

The local validation results demonstrate consistent effectiveness of our IDK-aware training strategy across all three pipelines.

For direct QA, the IDK training successfully transforms a severely negative baseline (score -96) into positive performance, with epoch2 achieving the first positive score (4) through dramatic hallucination reduction from 176 to 31.

For image RAG, the baseline shows the most severe hallucination problem (score -141) among all three approaches. However, our IDK training demonstrates remarkable improvement: epoch1 immediately reduces hallucinations from 214 to only 10, though with increased missing responses (370). The training progression shows steady improvement, with epoch4 achieving the best performance (score 9) by optimally balancing accuracy (25), minimal hallucinations (16), and reasonable missing responses (347).

For web RAG, despite starting with the highest baseline accuracy (127), the severe hallucination problem (245) results in a negative score (-118). The IDK training with simple recognition ground truth consistently maintains positive scores across all epochs, with epoch2 achieving the overall best performance.

The results reveal that while web RAG demonstrates higher accuracy potential, image RAG shows the most dramatic improvement in hallucination control through IDK training, transforming from

[1] https://github.com/unslothai/unsloth
[2] https://www.aicrowd.com/challenges/meta-crag-mm-challenge-2025

Table 2: Official Leaderboard Results across All Tasks (Overall Ranking: 7th)

Task	Method	Subset	Rank	Acc	Hal	Miss	Score
Task 1	direct_idk+query_phrase	All	8th	98	42	2448	0.022
		Ego		77	33	1812	-
		Web		21	9	636	-
	pure_image_rag	All	-	111	85	2392	0.01
		Ego		73	57	1792	-
		Web		38	28	600	-
Task 2	pure_web_rag	All	9th	263	150	2175	0.044
		Ego		219	120	1583	-
		Web		44	30	592	-
Task 3	pure_web_rag	All	7th	499	171	3243	0.084

Table 3: Hallucination Suppression Training Results (20% Public Test Sample)

Method	Acc	Hal	Miss	Score
Base direct_qa				
Base	80	176	132	-96
direct_idk_data_train validation				
epoch1	41	45	302	-4
epoch2	35	31	322	4
epoch3	23	23	342	0
Base pure_image_rag				
Base	73	214	101	-141
pure_image_rag_idk validation				
epoch1	8	10	370	-2
epoch2	21	20	347	1
epoch3	26	24	338	2
epoch4	25	16	347	9
epoch5	26	20	342	6
Base pure_web_rag				
Base	127	245	16	-118
pure_web_rag_idk validation + train				
epoch1	82	75	231	7
epoch2	109	100	179	9
epoch3	77	68	243	9
epoch4	93	91	204	2
epoch5	88	86	214	2

the worst baseline performance to competitive results. This validates our approach's effectiveness across different modalities and information sources.

3.2.2 Official Leaderboard Results. We subsequently evaluated our best performing configurations on the official CRAG-MM leaderboard across all three tasks. Table 2 presents the comprehensive results, achieving an overall 7th place ranking.

The official leaderboard results validate our approach across all three tasks, achieving competitive rankings and demonstrating consistent performance improvements with effective hallucination control. Our final overall ranking of 7th place reflects the robustness and effectiveness of our ConfRAG framework.

In Task 1 (Single-source Augmentation), our direct QA with IDK training achieved 8th place with a score of 0.022, demonstrating superior hallucination control despite moderate accuracy. The pure image RAG approach, while not submitted for official ranking, shows complementary performance with higher accuracy (111 vs 98) but increased hallucinations.

Task 2 (Multi-source Augmentation) secured 9th place with pure web RAG achieving the highest task-specific accuracy (263) and maintaining effective hallucination suppression (score 0.044). This ranking demonstrates the effectiveness of web-based retrieval augmentation in complex multi-source scenarios.

Task 3 (Multi-turn QA) achieved our best individual task ranking of 7th place with a score of 0.084. This performance highlights our framework's particular strength in handling complex conversational contexts while maintaining consistent hallucination control across multiple interaction turns.

Key findings across all tasks include: (1) IDK-aware training consistently reduces hallucinations while maintaining competitive accuracy, contributing to stable rankings across tasks, (2) confidence-based threshold settings prove crucial for optimal performance, and (3) the framework's scalability from single-source to multi-turn scenarios is validated by achieving top-10 rankings across all tasks, culminating in an overall 7th place finish.

3.3 Confidence Calibration Analysis

To validate the effectiveness of our IDK-aware training strategy, we conducted a comprehensive analysis of sequence-level confidence patterns before and after training. This experiment specifically examines how well our training approach enhances the discriminative power between correct and incorrect responses.

As shown in Table 4, our IDK-aware training significantly improves the model's ability to distinguish between correct and incorrect responses through confidence scores. Before training, the average token probability for correct sequences (0.88) was only marginally higher than that for incorrect sequences (0.85), with

Table 4: Average Token Probability Analysis Before and After IDK Training

Training Stage	Correct	Incorrect	Difference
Before Training	0.88	0.85	0.03
After Training	0.85	0.75	0.10

a minimal difference of 0.03. This narrow margin makes it challenging for confidence-based hallucination detection to effectively identify unreliable responses.

After applying our IDK-aware training strategy, the model demonstrates substantially better confidence calibration. While the average probability for correct sequences decreases slightly to 0.85, the probability for incorrect sequences drops significantly to 0.75, resulting in a much larger difference of 0.10. This 3.3x improvement in confidence discrimination provides a more reliable signal for our hallucination detection module to identify and suppress incorrect responses. This enhanced confidence differentiation directly validates two key aspects of our approach: (1) the IDK-aware training successfully teaches the model to express lower confidence when generating incorrect responses, and (2) the improved confidence calibration enables more effective operation of our hallucination detection rules, particularly the average token probability threshold mechanism.

4 Limitations

While ConfRAG effectively reduces hallucinations, several limitations remain. First, the method shows a relatively high missing rate across pipelines, which we find is strongly correlated with the base model's accuracy—stronger base RAG performance leads to lower missing rates after IDK-aware training. This indicates that retrieval effectiveness fundamentally constrains the model's ability to answer confidently, suggesting future work should improve information recall to balance uncertainty control and answer completeness. Second, the framework treats image-based and web-based sources independently in a cascaded manner, failing to exploit their complementary nature. Future research could explore agent-based RAG approaches that jointly integrate visual and textual evidence through more advanced multi-source fusion mechanisms to further enhance overall performance.

5 Conclusion

We presented ConfRAG, a confidence-calibrated retrieval-augmented generation framework that addresses hallucination problems in visual question answering. Through confidence-aware retrieval, hybrid hallucination detection, and IDK-aware training strategies, our approach transforms severely negative baselines into positive performance while maintaining competitive accuracy. ConfRAG achieves 7th place overall on the CRAG-MM challenge with consistent hallucination reduction across all three tasks. Future work includes extending confidence calibration to other multi-modal tasks and developing adaptive threshold mechanisms for different question complexities.

References

[1] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. arXiv:2402.03216 [cs.CL]

[2] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24185–24198.

[3] Zhangchi Feng, Dongdong Kuang, Zhongyuan Wang, Zhijie Nie, Yaowei Zheng, and Richong Zhang. 2024. Easyrag: Efficient retrieval-augmented generation framework for automated network operations. *arXiv preprint arXiv:2410.10315* (2024).

[4] Zhangchi Feng, Richong Zhang, and Zhijie Nie. 2024. Improving composed image retrieval via contrastive learning with scaling positives and negatives. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 1632–1641.

[5] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997* 2, 1 (2023).

[6] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2023. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. arXiv:2303.05499 [cs.CV]

[7] Meta AI. 2024. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.

[8] Jielin Qiu, Andrea Madotto, Zhaojiang Lin, Paul A. Crook, Yifan Ethan Xu, Babak Damavandi, Xin Luna Dong, Christos Faloutsos, Lei Li, and Seungwhan Moon. 2024. SnapNTell: Enhancing Entity-Centric Visual Question Answering with Retrieval Augmented Multimodal LLM. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 247–266. doi:10.18653/v1/2024.findings-emnlp.14

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[10] Qwen Team. 2025. Qwen3 Technical Report. arXiv:2505.09388 [cs.CL] https://arxiv.org/abs/2505.09388

[11] Jiaqi Wang, Xiao Yang, Kai Sun, Parth Suresh, Sanat Sharma, Adam Czyzewski, Derek Andersen, Surya Appini, Arkav Banerjee, Sajal Choudhary, Shervin Ghasemlou, Ziqiang Guan, Akil Iyer, Haidar Khan, Lingkun Kong, Roy Luo, Tiffany Ma, Zhen Qiao, David Tran, Wenfang Xu, Skyler Yeatman, Chen Zhou, Gunveer Gujral, Yinglong Xia, Shane Moon, Nicolas Scheffer, Nirav Shah, Eun Chang, Yue Liu, Florian Metze, Tammy Stark, Zhaleh Feizollahi, Andrea Jessee, Mangesh Pujari, Ahmed Aly, Babak Damavandi, Rakesh Wanga, Anuj Kumar, Rohit Patel, Wen tau Yih, and Xin Luna Dong. 2025. CRAG-MM: Multi-modal Multi-turn Comprehensive RAG Benchmark. arXiv:2510.26160 [cs.CV] https://arxiv.org/abs/2510.26160

[12] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-Pack: Packaged Resources To Advance General Chinese Embedding. arXiv:2309.07597 [cs.CL]

[13] Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, Lei Chen, Nicolas Scheffer, Yue Liu, Nirav Shah, Rakesh Wanga, Anuj Kumar, Wen-tau Yih, and Xin Luna Dong. 2024. CRAG - Comprehensive RAG Benchmark. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 10470–10490. https://proceedings.neurips.cc/paper_files/paper/2024/file/1435d2d0fca85a84d83ddcb754f58c29-Paper-Datasets_and_Benchmarks_Track.pdf

[14] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*, Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (Eds.). PMLR, 57730–57754. https://proceedings.mlr.press/v235/yu24o.html

[15] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372* (2024).