Team Metamorphosis: Meta CRAGMM- KDD CUP

Harshavardhan Etcherla*
AI Garage, Mastercard
Gurugram, India
harshavardhan.etcherla@mastercard.com

Aditi Rai*
AI Garage, Mastercard
Gurugram, India
aditi.rai@mastercard.com

Rakshit Rao
AI Garage, Mastercard
Gurugram, India
rakshit.rao@mastercard.com

Alekhya Bhatraju
AI Garage, Mastercard
Gurugram, India
alekhya.bhatraju@mastercard.com

Diksha Shrivastava
AI Garage, Mastercard
Gurugram, India
diksha.shrivastava@mastercard.com

Shivam Arora
AI Garage, Mastercard
Gurugram, India
Shivam.Arora@mastercard.com

Abstract

Vision-Language Models (VLMs) show remarkable capabilities in multi-modal reasoning but struggle with hallucinations, long tail recognition, and grounding under real-world conditions such as those found in wearable devices. The Meta CRAG-MM Challenge 2025 presents a benchmark to evaluate these issues across three tasks: Single-source Augmentation, Multi-source Augmentation, and Multi-turn QA. We present a modular MM-RAG pipeline that explicitly targets factuality, interpretability, and robustness in such scenarios. Our system combines object-aware image cropping, domain-specific identification via a LoRA-finetuned LLaMA 3.2 Vision Instruct model [4], CLIP, and BGE based dual retrieval pipelines, and structured Chain-of-Thought (CoT) reasoning followed by a hallucination sensitive summarizer. We observe that visual preprocessing and identification significantly improve retrieval quality, while CoT prompting enhances answer consistency. Early results show reduced hallucination and improved truthfulness over a baseline LLM setup. We outline remaining challenges and future directions, including contrastive answer alignment, domain generalization, and inference optimization, toward developing real-time, grounded VLLMs for egocentric vision tasks.

CCS Concepts

• **VLMs → multi-modal, multi-turn QA**.

Keywords

VLMs, MM-RAG QA system

ACM Reference Format:
Harshavardhan Etcherla, Aditi Rai, Rakshit Rao, Alekhya Bhatraju, Diksha Shrivastava, and Shivam Arora. 2025. Team Metamorphosis: Meta CRAGMM- KDD CUP. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25), August 3–7, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*Both authors contributed equally to this research.

KDD '25, Toronto, ON, Canada

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The growing capabilities of Vision-Language Models (VLMs) have opened new frontiers in multi-modal AI, from visual question answering (VQA) to interactive perception in augmented reality systems. Among these, Vision Language Large Models (VLLMs) such as GPT-4V and LLaVA have demonstrated remarkable fluency in combining textual and visual input. However, these models often hallucinate facts, perform poorly in long tailed or fine-grained visual categories and lack the structured reasoning required in complex, real-world environments, particularly those involving egocentric vision from smart glasses or mobile agents. The Meta CRAG-MM Challenge 2025 provides a rigorous benchmark for these shortcomings across three tasks: (1) Single source Augmentation: answering factual questions using structured image-based knowledge graphs (KGs), (2) Multi source Augmentation: integrating open-domain web search results for grounding and (3) Multi turn QA: conducting coherent multi turn conversations grounded in egocentric visual context. The benchmark includes over 5,000 annotated examples across 14 real-world domains, emphasizing truthfulness, visual specificity, and multi-hop reasoning. Evaluation metrics span truthfulness score, hallucination rate, completeness of responses, and latency of queries. We address these challenges with a structured MM-RAG pipeline comprising six modules: visual cropping using Grounding DINO, fine-grained object identification via a LoRA-adapted LLaMA 3.2 model [4], dual source retrieval using CLIP (visual) and BGE (textual) embeddings, a query-aware reranking stage, and CoT based answer reasoning followed by an abstention-aware summarizer. Our early experiments demonstrate substantial improvements in factual grounding, particularly for visually complex and domain-specific queries. Compared to an LLM only baseline, our system reduces hallucinations and improves the truthfulness of responses. Our contributions are as follows: We design a modular MM-RAG pipeline optimized for egocentric, multi-domain, multi-turn visual QA. We finetune a LLaMA-3.2 Vision-Instruct model using LoRA [4] for domain-specific recognition (e.g., car makes/models). We introduce an image-cropping and object-ID step to enrich query generation and enhance retrieval precision. We combine CoT prompting and hallucination-aware summarization to improve answer traceability and factual reliability. We validate our design through ablations and error analysis with early results showing significant reduction in hallucination over baseline methods.

2 Related Work

2.1 Vision-Language Models and Visual Question Answering

Vision-Language Models (VLMs) aim to jointly understand visual and textual data. Early models such as ViLBERT[13] and VisualBERT[11] employed dual-stream architectures for image-text alignment, while later models like CLIP[16] and ALIGN [7] learned cross-modal embeddings at scale via contrastive learning. More recent systems such as BLIP-2 [10], Flamingo [1], and GPT-4V have demonstrated strong generalization in zero-shot and few-shot visual question answering (VQA). Despite this progress, these models often rely entirely on their internal parametric knowledge and are susceptible to *hallucinations* and factual errors, especially when dealing with rare or long tail concepts.

2.2 Hallucination in VLMs

Hallucination refers to the phenomenon where models generate fluent but unfaithful outputs that are unsupported or contradicted by the input. Studies have shown this to be a persistent issue in both language-only and vision-language models [6]. Mitigation strategies include prompt engineering, fine-tuning with curated datasets and most notably, *retrieval augmentation*.

2.3 Retrieval-Augmented Generation (RAG)

The RAG framework [9] augments language models with non-parametric memory by retrieving relevant external knowledge and conditioning generation on it. Originally developed for text-only QA tasks, RAG has been extended to the multi-modal setting by integrating vision features into the retrieval and generation pipeline. Approaches like REVEAL[5] and MM-ReAct[24] combine recognition, web search and reasoning, showing promise in improving factuality and interpretability.

2.4 Multi-modal RAG and Egocentric QA

Multi Modal Retrieval-Augmented Generation (MM-RAG) is recent paradigm that synthesizes queries from both image and textual inputs, retrieves information from structured (e.g., knowledge graphs) and unstructured (e.g., web) sources and grounds the answer in the retrieved context. This is especially relevant for *egocentric QA*, a domain where the input images are personal, diverse and context rich. Systems like ViperGPT[18] and MM-ReAct[24] exemplify attempts to modularize perception and reasoning, but they often fall short in grounding answers across long tailed domains and multi-step reasoning.

Our work builds on these insights by integrating image-text grounding, OCR, domain-specific retrieval and knowledge-aware answer synthesis into a unified MM-RAG pipeline tailored for the CRAG-MM benchmark.

3 Tasks and Dataset Overview

The Meta CRAG-MM Challenge 2025 introduces a comprehensive benchmark for evaluating the capabilities of multi-modal systems in grounded visual question answering. The dataset and tasks are designed to test the limits of current Vision-Language Models (VLMs), particularly in settings involving egocentric vision, factual grounding, and complex reasoning. The benchmark consists of over 5,000 images and question-answer pairs, collected from diverse domains and contexts, including a substantial portion sourced from smart glasses (Ray-Ban Meta).

3.1 Dataset Composition

The CRAG-MM contains three parts of the data: the image set, the QA set and the contents for retrieval.

3.1.1 Image set. CRAG-MM contains two types of images: egocentric images and normal images. The egocentric images were collected using Ray-Ban Meta Smart Glasses 4 from a first-person perspective. Normal images were collected from publicly available images on the Web.

3.1.2 Question Answer Pairs. The CRAG-MM covers 14 domains: Book, Food, General object recognition, Math and science, Nature, Pets, Plants and Gardening, Shopping, Sightseeing, Sports and games, Style and fashion, Text understanding, Vehicles and Others, representing popular use cases that wearable device users would like to engage with. It also includes four types of question, ranging from simple questions that can be answered based on the image to complex questions that require retrieving multiple sources and synthesizing an answer.

The four types of questions in the benchmark are as follows:

- *Simple questions:* Questions asking for simple facts.
  - *Simple recognition:* This can be directly answered from the image (e.g., "What brand is the milk?" or "Who wrote this book?" where the brand name and the book author are shown on the image).
  - *Simple knowledge:* Requires external knowledge for the answers (e.g., "What is the price of this sofa on Amazon?").
- *Multi-hop questions:* Questions that require chaining multiple pieces of information to compose the answer (e.g., "What other movies have the director of this movie directed in the past?").
- *Comparison and Aggregation questions:* Questions requiring aggregating or comparing multiple pieces of information (e.g., "Which drinks do not contain added sugar among these?" or "Is this cheaper on Amazon?").
- *Reasoning questions:* Questions about an entity that cannot be directly looked up and require reasoning to answer (e.g., "Can the dryer be used in Europe?" where the image shows a dryer).

3.1.3 Retrieval Contents. The data set includes a mock image search API and a mock web search API to simulate real-world knowledge sources from which RAG solutions retrieve.

- *Image search:* A mock image search API takes an image as input and returns similar images with structured metadata from a mock KG.

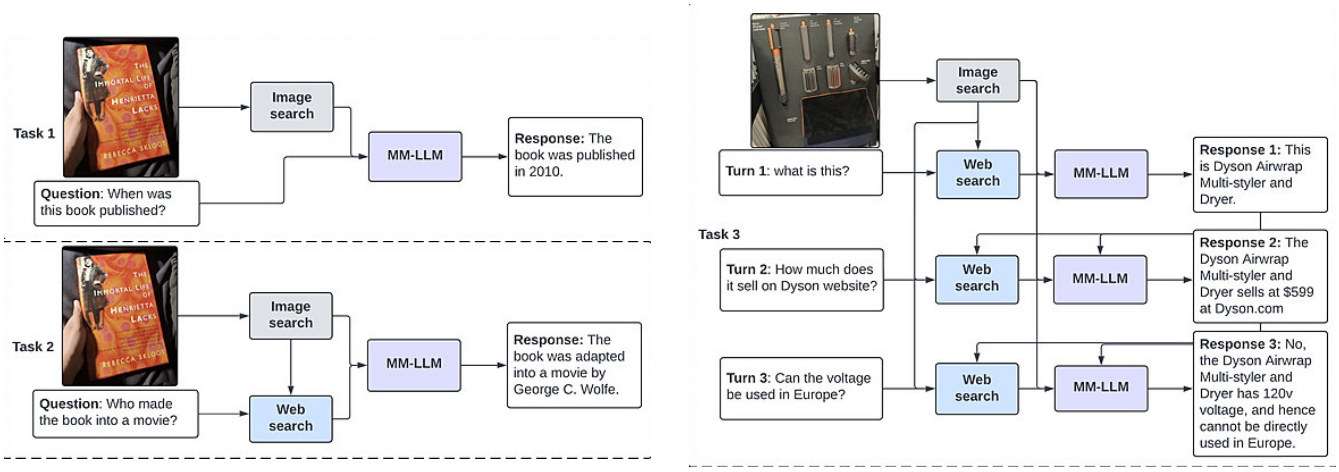

Figure 1: Overview of Tasks 1, 2, and 3. Left column shows Task 1 and Task 2 . Right column shows Task 3.

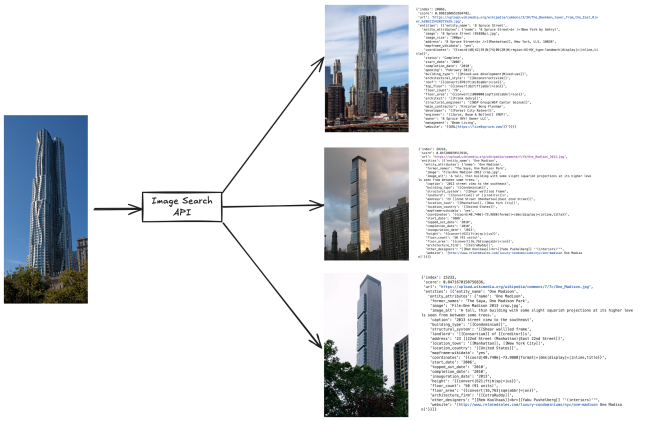

Figure 2: Working of Image search API

- *Text-Based Web Search:* A text search API takes a text query as input and returns relevant web pages (URLs, page titles, snippets, last updated time).

3.2 Task Definitions

The challenge consists of three tasks, each progressively increasing in difficulty and designed to test different capabilities of MM-RAG systems:

3.2.1 Task 1: Single-Source Augmentation. Given an image and a question, the system retrieves supporting evidence from a structured internal knowledge base (mock-KG) and generates a grounded answer. This task evaluates retrieval and grounding with domain-specific structured data.

- **Goal:** To test the basic answer generation capability of MM-RAG systems.
- Provides an image mock API to access information from an underlying image-based mock KG. The mock KG is indexed by the image and stores structured data associated with the image. Answers to the questions may or may not exist in the mock KG. The mock API takes an image as input and returns similar images from the mock KG, along with structured data associated with each image to support answer generation.

3.2.2 Task 2: Multi-Source Augmentation. Extends Task 1 by allowing the system to additionally retrieve from external web sources. The model must synthesize relevant information from both internal and open-domain data to provide grounded answers.

- **Goal:** To test how well the MM-RAG system synthesizes information from different sources.
- In addition to Task 1, this task provides a web search mock API as a second retrieval source. The Web pages may provide useful information to answer the question, but they may also contain noise.

3.2.3 Task 3: Multi-Turn QA. Introduces a dialogue-style setup in which the model answers a sequence of related questions based on the same image. This task evaluates memory, temporal coherence and conversational reasoning.

- **Goal:** To test context understanding for smooth multiturn conversations.
- This task tests the system's ability to conduct multi-turn conversations. Each conversation contains 2–6 turns. Except for the first turn, questions in later turns may or may not require the image for answering.

3.3 Evaluation Metrics

3.3.1 SINGLE-TURN QA.

- For each question in the evaluation set, the answer is scored as:
- Perfect (fully correct) → Score: 1.0
- Acceptable (useful but with minor non-harmful errors) → Score: 0.5
- Missing (e.g., "I don't know", "I'm sorry I can't find …") → Score: 0.0
- Incorrect (wrong or irrelevant answer) → Score: -1.0
- **Truthfulness Score:** The average score across all examples in the evaluation set for a given MM-RAG system.

3.3.2 MULTI-TURN QA. We adapt the method in [2], which is closest to the information-seeking flavor of conversations. In particular, we stop a conversation when the answers in two consecutive turns are wrong and consider answers to all remaining questions in the same conversation as missing mimicking the behaviour of real users when they lose trust or feel frustrated after repeated failures. We then take the average score of all multi-turn conversations.

4 Team Metamorphosis: Meta CRAGMM: Methodology

We present a structured pipeline that reduces hallucination by combining visual cropping, object recognition, multi-modal retrieval and stepwise reasoning. Our architecture is grounded in modular retrieval-augmented generation (RAG) with emphasis on accurate visual grounding, targeted image/text retrieval and reasoning via chain-of-thought [20] prompting. Our system effectively balances vision and language grounding to provide interpretable and factual responses. This architecture composed of six sequential modules, each tailored to reduce noise and improve factual grounding.

4.1 Base Model

For our RAG multi-modal reasoning pipeline, we adopted the LLaMA 3.2 11B Vision-Instruct [14] model developed by Meta as the base large language model. This model, designed to handle both visual and textual modalities, served as the backbone for our identification and reasoning components. However, during early-stage evaluations, we observed that the pretrained model exhibited limited reliability when it came to fine-grained recognition tasks. In particular, its performance was suboptimal in identifying visually specific categories such as car models, animal species and brand logos, domains where subtle texture, shape and compositional cues are critical.

To address these limitations, we conducted parameter-efficient finetuning using the Low-Rank Adaptation (LoRA) [4] approach. Specifically, we applied LoRA [4] to all attention modules within the transformer architecture, setting the rank parameter to $r = 32$ and the scaling factor to $\alpha = 16$. This configuration allowed for efficient adaptation without fully updating the large base model, making it computationally tractable while preserving generalization. The model was fine-tuned for one epoch using the Stanford Cars dataset [8], publicly available on Kaggle. It contains 16,185 images across 196 car classes. The data is split into 8,144 training images and 8,041 testing images. Each class typically represents a specific make, model, and year (e.g., 2012 Tesla Model S, 2012 BMW M3 Coupe). We used a learning rate of 10^{-6}. This fine-tuning phase led to significantly improved performance in object identification.

4.2 Image Cropping Block

The first stage of our pipeline focuses on localizing the visual region most relevant to the user query. For this, we employ the `IDEA-Research/grounding-dino-base` [12] model, which takes a natural image, and a text query as input and predicts bounding boxes for regions that semantically align with the query. This model acts as a visual grounding module, essential for reducing irrelevant context and focusing downstream modules on the query-specific visual information.

The inputs to this block are the original image and a user-provided natural language query. The model outputs bounding box coordinates that mark where the object or region of interest lies in the image. We used these coordinates to extract cropped patches from the image. These patches are then passed on to subsequent modules for classification, retrieval, or reasoning.

To ensure that the cropping process is both accurate and semantically meaningful, we apply two confidence thresholds. A **box threshold of 0.5** is used to discard bounding boxes with weak visual grounding, while a **text threshold of 0.3** ensures that the selected boxes are meaningfully aligned with the language query. Only bounding boxes that satisfy both thresholds are retained for cropping otherwise, we preserve the image as is.

This step plays a crucial role in reducing hallucinations later in the pipeline. By limiting the visual input to only those regions that are relevant to the query, we prevent downstream models from attending to irrelevant or misleading parts of the image. It also improves the performance of vision-language retrieval models, which benefit from tightly focused, noise-free visual representations.

4.3 Object Identification Block

Following the cropping step, the next objective is to interpret and classify the content within the query-relevant region of the image. For this, we use our base model finetuned on the Stanford Cars dataset. This block receives the cropped image region as input and produces a textual identification of the object present in that region, for example, the car's make, model, or type.

The motivation for this block lies in the fact that many multi-modal queries are lacking sufficient information unless the object of interest is clearly identified. Simply providing the cropped image to a retrieval model may not sufficiently inform the search about the semantics of the object. Instead, producing a textual label (e.g., "Honda Civic" or "Audi A4") serves two purposes: it enriches the retrieval query and grounds the reasoning steps that follow.

The input to this block is a single cropped image, and query and the output is a concise natural language phrase that identifies the object. In our implementation, the model generates a class label in free form text, allowing flexibility to express uncertain or partial predictions when the object is ambiguous or partially visible.

This classification output is later concatenated with the user's original question to create a semantically rich retrieval prompt. For example, the question "What is the horsepower?" becomes "What is the horsepower of the 2012 Honda Civic?" after identification. This modification helps both image-based and text-based retrieval systems to find more contextually appropriate results.

Fine-tuning the LLaMA 3.2-11B Vision Instruct model on the Stanford Cars dataset was critical for high-quality identification, as the dataset provides detailed annotations for real-world vehicle types. This specialized training helps the model generalize better across unseen examples while remaining grounded in the domain of vehicle identification.

Ultimately, this block enhances both the retrieval relevance and downstream reasoning quality. It serves as a bridge between visual

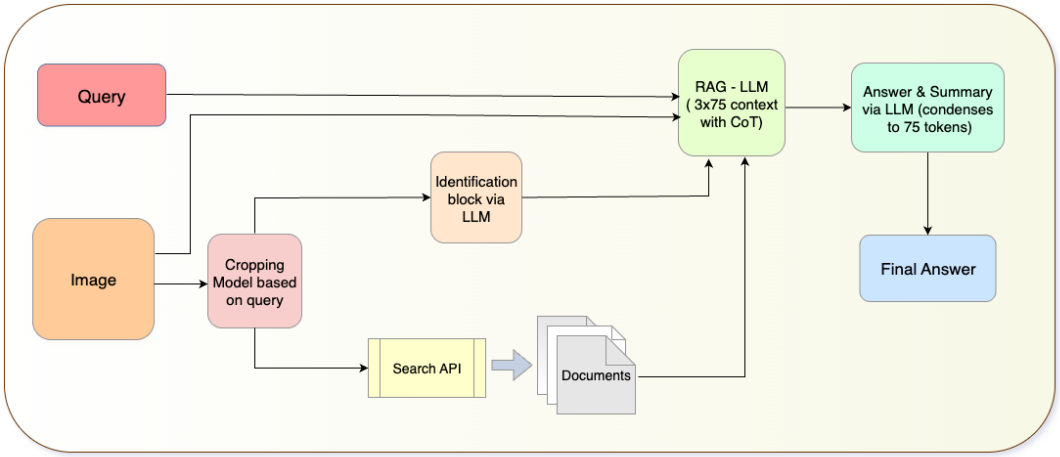

Figure 3: Model Architecture for Task 1: Single-Source Augmentation

recognition and language-based understanding—contributing to reduced ambiguity, better contextual alignment, and stronger factual grounding in the final answer.

4.4 Image Search via CRAG-MM API

After identifying the object in the cropped image, we aim to gather supporting context using visual similarity-based retrieval. This block uses the CRAG-MM image search API to retrieve relevant metadata and knowledge snippets from a multi-modal Knowledge Graph (KG). The API leverages a vision transformer model, specifically `openai/clip-vit-large-patch14-336` [16], which encodes the cropped image into a high-dimensional embedding space. This embedding is then compared against pre-indexed image representations in the KG using cosine similarity.

The input to this block is the cropped image generated from the previous stage. The output is a set of top-K visually similar entries from the knowledge base, each consisting of structured metadata—such as captions, object descriptors, positional information, and snippet context. We empirically set K = 5 after filtering results based on similarity thresholds to ensure precision.

This step is crucial as it retrieves images visually similar to the query image, allowing us to extract and utilize their associated metadata. This metadata not only aids in filtering and ranking relevant images but also provides interpretability by offering insight into the reasoning behind their retrieval.

4.5 Textual Search via CRAG-MM API

To complement visual retrieval, we perform semantic search over text using the CRAG-MM text API. The input to this block is a concatenation of the object identification result and the user query, forming a more context-aware prompt. This enriched text input is embedded using the `BAAI/bge-large-en-v1.5` [22] model—a powerful sentence-level embedding model optimized for dense retrieval tasks. The model encodes the combined query and searches for semantically similar entries in the textual portion of the Knowledge Graph.

This block returns the top 15 textual snippets, each selected based on similarity scores and filtered via thresholding to avoid noisy or unrelated matches. These snippets may contain factual details, product specifications, or common knowledge about the object of interest. Examples of Identification + Query include "Audi A4. What is the price of this car?", "Tesla Model 3. What are the features of this model?" etc

The key purpose of this block is to use web search directly to answer the query which is available on web. While the image search can match based on appearance, text search allows the system to retrieve domain knowledge and descriptive content that is essential for question answering. It is especially useful when answering queries that are inherently textual, such as "What is the fuel efficiency of this model?" or "When was this car manufactured?"

By combining both visual and textual retrieval, we ensure that the system benefits from both grounded perceptual evidence and explicit language-based information. This fusion strengthens the foundation for the final reasoning step, leading to answers that are both accurate and well-supported by external knowledge.

4.6 Chain-of-Thought (CoT) LLM Reasoning

To generate accurate and interpretable answers, we employed a Chain-of-Thought (CoT) [20] LLM that leverages the RAG-enhanced context constructed from the query, the image and the retrieval results from both image and text search pipelines. This module is designed to mimic a structured human reasoning process by breaking the response generation into multiple coherent thought steps. Specifically, the model is allowed to think and respond over three sequential reasoning stages, each constrained to a maximum of 75 tokens, making the total output capped at 225 tokens. This capping choice is due to the time constraints imposed by the hackathon. This design encourages the model to elaborate step-by-step rather than jumping directly to a final answer. The reasoning flow involves identifying relevant facts, correlating visual and textual information and progressively refining its inference. By capturing intermediate thoughts before finalizing the answer, this approach not only improves interpretability but also reduces the likelihood of

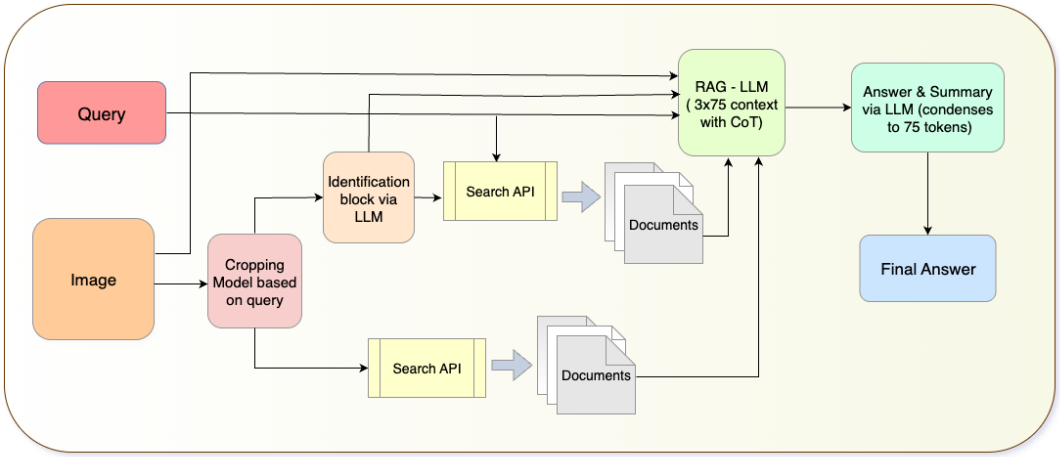

Figure 4: Model Architecture for Task 2 and 3: Multi-Source Augmentation and Multi-Turn QA

hallucinated or unsupported conclusions. The full reasoning trace, including all intermediate steps and the final answer, is preserved as the output of this block, which is then passed to the summarization model for final polishing.

4.7 Answer & Summary LLM

Multi-modal models often face a structural bottleneck when dealing with large images and long textual contexts, especially in transformer-based architectures where image tokens tend to dominate the early layers of attention. This leads to a phenomenon where essential text snippets or reasoning chains are diluted or ignored. To overcome this, we introduced a separate summarization stage using a text-only LLM that operates independently of the visual modality. This model receives the entire output from the CoT [20] reasoning module, including both the thought process and conclusion and generates a concise, coherent, and grounded response. Importantly, we explicitly instructed the model to respond with "I don't know" when the evidence or reasoning was insufficient to generate a factual answer. This safety mechanism acted as a guardrail against hallucinations, helping us maintain a higher factual precision. This final LLM stage was instrumental in improving our overall leaderboard score by ensuring clarity, brevity, and correctness in the final output. For multi-turn conversations, we maintain context by appending answers from turns 1 to (n-1), and we introduce fresh RAG information when processing the nth query. This ensures continuity while integrating new evidence.

5 Experiments

Conducted a series of experiments to explore the performance of different MM-RAG configurations for grounded question answering on the CRAG-MM dataset. The experimental setup evolved iteratively, beginning with a lightweight LLM baseline and progressively integrating retrieval, query synthesis and re-ranking components.

5.1 Phase 1 Improvements

5.1.1 Initial Baseline: LLM-Only Answering. Our first experiment involved a simple baseline where answers were generated directly using a large language model (LLM) without any retrieval. The input to the LLM consisted of the question concatenated with an image caption or summary produced by LLama 3.2 Vision Instruct model. While this setup could handle basic recognition-style questions, it often hallucinated facts or failed to retrieve long-tail knowledge not encoded in the model's parameters.

5.1.2 Baseline RAG Agent. Next, we implemented a retrieval-augmented baseline that included the following pipeline:

(1) **Input:** Image + question
(2) **Image summarization + Query generation:** The model generates a descriptive summary and some search queries. The queries and summary are merged to form search queries.
(3) **Retrieval:** The query is passed to a web or KG-based retriever which is search API in our case
(4) **Answering:** The retrieved text passages are provided as context to the LLM, which generates a final answer.

Although this approach improved factual grounding in some cases, the overall performance was limited by the quality of the search queries and the retrieval results.

5.1.3 API Observations and Query Generation Challenges. Manual inspection of the API behaviour revealed that poor search queries frequently led to irrelevant or low-quality retrieval results. This had a significant downstream impact on answer quality. We identified several recurring problems:

- Lack of named entities or keywords or mislabelled entities in generated queries
- Overly generic queries (e.g., "what is this" or "describe the object")
- Redundancy or hallucinated attributes in queries

To address this, we experimented with prompt engineering and fine-tuning of the query generation module to make the queries more specific, grounded and information dense.

Experiment	Truthful	Missing	Hallucination	Accuracy
Initial Baseline	-0.332	0.231	0.550	0.218
Query RAG Agent	-0.519	0.026	0.746	0.227
Reranking	-0.308	0.214	0.547	0.239

(a) Phase 1 Experiments

Experiment	Truthful	Missing	Hallucination	Accuracy
Cropping	-0.478	0.088	0.695	0.217
Final Model (CoT + FT)	-0.279	0.159	0.560	0.281

(b) Phase 2 Experiments

Table 1: Quantitative results from Phase 1 (left) and Phase 2 (right) experiments.

5.1.4 Reranking: Slow vs Fast Pipelines. To improve retrieval relevance, we incorporated reranking mechanisms:

- **Fast Reranker:** A lightweight cross-encoder sentence-transformers/all-MiniLM-L6-v2 [17] model applied to the retrieved documents to select the top-20.
- **Slow Reranker:** A larger, more accurate reranker (`BAAI/bge-reranker-large` [21]) to filter from top 20 to top 3.

We observed that reranking consistently improved answer grounding and fluency. However, the slow reranker introduced significant latency, making it more suitable for offline or reranking in evaluation-time settings.

5.2 Phase 2 Enhancements

In Phase 2, organizers had changed the reranker for text with accurate one so we removed the reranking step and focused on blocks and model improvement. We implemented a series of enhancements to overcome the limitations identified in Phase 1 and to target more domain-specific and compositional queries. The focus was on improving visual input fidelity, enabling step-by-step reasoning and incorporating task-specific adaptation.

5.2.1 Image Cropping and Preprocessing. We introduced image cropping heuristics to improve the relevance of visual features provided to the captioning and recognition modules. Many egocentric images included background clutter or peripheral noise by cropping around salient regions (e.g., based on object detection), we were able to improve the focus and quality of downstream image summaries.

5.2.2 Chain-of-Thought (CoT) [20] Prompting. To better handle complex reasoning, we adopted *Chain-of-Thought* (CoT) prompting strategies within the LLM. This enabled more interpretable, step-by-step answer generation, especially for multi-hop or comparative questions. We observed that CoT prompting often led to more faithful reasoning paths and reduced hallucination compared to direct answer generation.

5.2.3 Domain Adaptation: Stanford Cars Finetuning [8]. Recognizing performance bottlenecks on fine-grained domains such as vehicle classification, we finetuned a vision encoder using the Stanford Cars dataset. This improved the system's ability to recognize specific car makes and models, leading to better query generation and answer grounding for automotive-related queries in the CRAG-MM dataset.

6 Conclusions and Future work

While our current system showed strong performance across the benchmark tasks, several key enhancements could further improve its accuracy, efficiency and generalization. One of our primary directions is to incorporate contrastive loss mechanisms during fine-tuning, which would explicitly encourage alignment between the generated answers and ground truth annotations. We conducted preliminary experiments with contrastive learning to align image-summary-text representations more tightly. While the initial implementation was incomplete and not fully tuned, it showed early promise in helping the model distinguish semantically close but visually distinct concepts. Future iterations may explore stronger contrastive supervision and joint training strategies This would help reduce subtle factual drift and guide the model toward producing more grounded outputs. Additionally, we aim to improve the model's ability to front-load answers delivering direct and concise information within the 75-token generation limit while still preserving stepwise reasoning for traceability. On the engineering front, we plan to optimize inference pipelines to reduce latency, improve batching strategies and minimize GPU memory overhead. Lastly, extending the identification module to handle additional domains such as species classification, bicycles, or machinery will expand the system's applicability and robustness to diverse visual contexts.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Paul H Luc, Antoine Miech, Malcolm Reynolds, Wellington Hsu, et al. 2022. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198* (2022).

[2] Yuntao Bai, Siyan Chen, Di Xiao, Yuxuan Zhang, Zhengyuan Zhang, Yajing Liu, Zihan Li, Yangfeng Ji, Kai Han, Wenpeng Yin, et al. 2024. MT-Bench-101: A Fine-Grained Benchmark for Evaluating Large Language Models in Multi-Turn Dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*. https://aclanthology.org/2024.acl-long.401/

[3] Tian Gao, Yuming Tang, Canwen Zhou, Yizhong Zhang, Jing Liu, Shuohang Wang, Ruoxi Xie, Xiang Lin, Zhiyuan Liu, Maosong Sun, et al. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997* (2023). https://arxiv.org/abs/2312.10997

[4] Edward J. Hu, Yelong Shen, Phil Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Wei Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685* (2021). https://arxiv.org/abs/2106.09685

[5] Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. 2023. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 23369–23379.

[6] Zhe Ji, Tian Gao, Zhiyuan Liu, Maosong Sun, et al. 2023. Survey of Hallucination in Natural Language Generation. *arXiv preprint arXiv:2303.12756* (2023).

[7] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*. PMLR, 4904–4916.

[8] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3D Object Representations for Fine-Grained Categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*. IEEE, 554–561.

[9] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[10] Junnan Li, Dongxu Li, Chunyuan Xiong, and Steven C. H. Hoi. 2023. BLIP-2: Bootstrapped Language-Image Pretraining. *arXiv preprint arXiv:2301.12597* (2023).

[11] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).

[12] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2024. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. arXiv:2303.05499 [cs.CV] https://arxiv.org/abs/2303.05499

[13] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* 32 (2019).

[14] Meta AI. 2024. Meta LLaMA3.2 11B Vision-Instruct. https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct. Multimodal instruction-tuned model released September 2024; access via Hugging Face; knowledge cutoff December 2023.

[15] Xuefei Qiu, Qian Yao, Yi Tang, Nan Duan, Hua Wu, and Haifeng Wang. 2024. SnapNTell: Enhancing Entity-Centric Visual Question Answering with Retrieval Augmented Multimodal LLM. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. https://aclanthology.org/2024.findings-emnlp.14/

[16] Alec Radford, Jong Wook Kim, Luke Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pam 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 (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

[17] Iryna Reimers, Nils and Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[18] Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11888–11898.

[19] 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

[20] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, Vol. 35. Curran Associates, Inc., 24824–24837.

[21] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. BAAI/bge-reranker-large: A cross-encoder reranking model (English & Chinese, 560M parameters). https://huggingface.co/BAAI/bge-reranker-large. Accessed: 2025-07-25.

[22] Canwen Xu, Yuning Gong, Yiming Liu, Bowen Zhou, Yijin Hou, Yuxuan Cui, and Qian Yang. 2023. BAAI General Embedding (BGE): A Series of General Text Embedding Models. https://huggingface.co/BAAI/bge-large-en-v1.5. Accessed: 2025-07-24.

[23] Weijia Yang, Xiaotao Zhang, Xinyu Xie, Yuxuan Hu, Yiheng Cui, Dongxu Zhang, Long Zhou, Wenhui Wang, Yuxuan Wu, Zeqi Yuan, et al. 2024. CRAG: A Comprehensive Benchmark for Retrieval-Augmented Generation. In *NeurIPS 2024 Datasets and Benchmarks Track*. https://proceedings.neurips.cc/paper_files/paper/2024/hash/1435d2d0fca85a84d83ddcb754f58c29-Abstract-Datasets_and_Benchmarks_Track.html

[24] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381* (2023).

[25] Yujia Yu, Jiahui Zhu, Yujie Shen, Rujun Luo, Jian Liu, Xinyang Wang, Ziwei Liu, and Shiyu Chang. 2023. MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities. *arXiv preprint arXiv:2308.02490* (2023). https://arxiv.org/abs/2308.02490

A System Prompts Used in the Pipeline

A.1 Identification Module

```
SYSTEM_PROMPT = ("""You are a helpful assistant that truthfully
      identifies the
object of interest in the image, demonstrative pronoun is about in
      the query
given. The object can be an thing, an abstraction or a phenomenon
      etc that is
being denoted based. You need to identify it as specific as
      possible.""")
```

A.2 Retrieval-Augmented Generation (RAG) Module

```
SYSTEM_PROMPT = """
You are a helpful assistant that answers truthfully for the
      question.
You need to think in the following manner.

1. Question Understanding: Understand what the [Question] is asking
      and identify
what are the things needed to answer the query about the image.

2. Analysis: Analyse the information given in RAG Context and find
      the answers
for above step either in the image or in this RAG Context or your
      own knowledge.

3. Draft Initial Answer (if answerable): Synthesize 'Key Facts'
      into a
comprehensive draft. Ensure every claim is directly supported.

4. Final Answer: If confident, output only the 'Final Answer'
      derived from your
draft. Else, output ONLY "I don't know."

Rules:
Please give only answer to the question using proper nouns about
      the image and
context below and mention the key elements supporting the answer.
If there are comparisons in the query, report differences in
      numerical or
non-numerical properties.
If you are unsure of the answer or context is insufficient, say only
"I don't know".
If context is contradicting or noisy, consider the most reliable or
      frequent
information, otherwise say "I don't know".
"""
```

A.3 Summarization Module

```
SYSTEM_PROMPT = ("""You are a ultra-precise assistant. Your task is
      to provide
the answer and 1--2 supporting sentences as concise, factually
      accurate, and
explicitly grounded answer (approx. 70--75 tokens) to the
      'Question' from the
previous responses and context generated.

The judge penalizes hallucination even if the answer is concise, if
      it is wrong,
unrelated, misidentified, or does not identify the demonstrative
      pronoun
correctly in the query. They value grounded answers with evidence
      supporting the
subject of interest.

ABSOLUTE RULES:
1. Every claim must be verifiable from Consolidated_Textual_Context
      or
Initial_Answer.
2. Focus on core validated facts and include brief grounding cues
      (e.g.,
"seen in image", "reported by context").
3. Target length is 70--75 tokens; accuracy overrides length.
4. Correctly identify the subject of interest; otherwise say "I
      don't know".
5. If Initial_Answer is "I don't know", or context is unreliable,
      output only
"I don't know".

Output must be ONLY the Final Answer or "I don't know".
""")
```