Adaptive Reasoning Engine for Knowledge-Based Visual Question Answering (ARE-VQA)

Adarsh Gupta and Abhishek Kumar

Indian Institute of Technology Guwahati

Abstract. Knowledge-Based Visual Question Answering (KB-VQA) requires an AI system to reason about an image using external, real-world knowledge. Recent advancements have shifted from complex, trained architectures to training-free pipelines that leverage the implicit knowledge of Large Language Models (LLMs). However, treating the LLM as a monolithic black box is inefficient and struggles to adapt to questions of varying complexity. We propose the Adaptive Reasoning Engine for KB-VQA (ARE-VQA), a modular, training-free pipeline that mimics human analytical processes. Our system first triages questions by complexity and required knowledge source. It then builds a focused context by gathering relevant visual details and conditionally retrieving external knowledge. For complex questions, a planner module decomposes them into simpler, solvable sub-questions. A dynamic tool selector then chooses the optimal prompting strategy to answer each atomic question. Finally, a synthesizer composes the intermediate steps into a coherent final answer. By orchestrating these specialized modules, ARE-VQA aims to provide a more accurate, interpretable, and efficient solution to the KB-VQA challenge.

Keywords: Visual Question Answering, Knowledge-Based VQA, Large Language Models, Modular Reasoning.

1 Introduction

Visual Question Answering (VQA) is a fundamental task in artificial intelligence, aiming to build systems that can answer natural language questions about an image. The task serves as a benchmark for an AI's ability to combine visual perception with linguistic understanding. However, a more challenging frontier in this domain is Knowledge-Based VQA (KB-VQA), which is defined by questions that require external, real-world knowledge not explicitly present in the image's pixels [7]. For example, given an image of a specific brand's soda can, a KB-VQA system might need to identify its parent company.

Historically, KB-VQA systems relied on complex pipelines that linked visual entities to structured knowledge bases like DBPedia or Wikidata, requiring intricate retrieval and reasoning modules. The field has recently undergone a paradigm shift with the advent of Large Language Models (LLMs). Researchers

found that these models implicitly store a vast amount of world knowledge, reframing the problem from structured data retrieval to effectively prompting an LLM with visual context to elicit the correct reasoning.

While this modern, training-free approach is powerful, it often applies a one-size-fits-all strategy to all questions. This can be inefficient for simple queries and ineffective for complex, multi-step ones. Our project addresses this gap by proposing the Adaptive Reasoning Engine for KB-VQA (ARE-VQA). Instead of a single model, we are designing a multi-stage agent that first analyzes a question's structure, gathers the necessary evidence (both visual and external), decomposes the problem if needed, and selects the right tool for each step. This system-first approach promises a more robust, efficient, and interpretable method for solving KB-VQA tasks.

2 Related Works

The KB-VQA landscape has been reshaped since 2022 by the dominance of the training-free paradigm, which leverages large, frozen LLMs as general-purpose reasoners. Progress in this area is now primarily driven by innovations in prompt engineering and intelligent context generation rather than novel model architectures.

This trend is clearly illustrated by the evolution of state-of-the-art models. An early and effective demonstration was a "Simple Baseline" approach [2], which achieved impressive results by simply generating a set of image captions and feeding the most semantically relevant ones to a frozen LLaMA model. This highlighted the power of providing even basic visual context to an LLM. Subsequent works have focused on making this context generation more task-aware. For instance, PromptCap [3] generates captions that are dynamically guided by the question itself, ensuring the textual description is rich with details specifically relevant to the query. Taking this a step further, the Prophet model [1] employs a multi-stage prompting method where a weaker VQA model first generates "answer heuristics," which are then used to construct a more potent and focused prompt for a powerful LLM like GPT-3. These methods showcase a clear progression towards crafting increasingly sophisticated prompts to better unlock the LLM's latent knowledge.

Parallel to this, the Retrieval-Augmented VQA (RA-VQA) paradigm remains crucial for questions requiring highly specific or dynamic knowledge not stored in an LLM's weights. Here too, the trend is towards deeper integration. Models like FLMR (2023) [4] have moved beyond using off-the-shelf retrievers by jointly training the retriever and answer generator, ensuring the retrieved knowledge is optimized for the VQA task. The most recent evolution, seen in models like ReAuSE (2025) [5], seamlessly integrates the retriever into the LLM itself, creating a built-in, generative search engine that can decide when and what to retrieve as part of its reasoning process. Our proposed work draws inspiration from this trend of intelligent context generation but organizes it within a more explicit, modular reasoning framework.

3 Proposed Pipeline: ARE-VQA

We propose a five-stage pipeline designed to reason about visual questions in an adaptive and human-like manner. The system intelligently routes queries through different modules based on their intrinsic properties. We plan to develop and evaluate this pipeline on the A-OKVQA dataset [6], as its focus on commonsense reasoning is an excellent match for our LLM-based approach. Below is a brief description of the ARE-VQA pipeline.

1. **Module 1: Triage Router.** This is the entry point. An LLM call classifies the input question along two axes: its complexity (ATOMIC vs. COMPOSITIONAL) and its required knowledge source (VISUAL vs. KNOWLEDGE-BASED). This initial decision dictates the entire subsequent path.
2. **Module 2: Context Builder.** This module gathers the evidence. First, it extracts key visual entities from the image using pre-trained models for captioning, object detection, and OCR. Then, if the Triage Router flagged the question as KNOWLEDGE-BASED, this module performs an additional step: it retrieves a snippet of external knowledge about the key visual entities from a source like Wikipedia or via a direct LLM query. The output is a compact, highly relevant context containing both visual and external facts.
3. **Module 3: Query Planner.** This "divide-and-conquer" module is only activated if the question is classified as COMPOSITIONAL. An LLM is prompted to act as a planner, breaking the complex question down into an ordered list of simple, ATOMIC sub-questions.
4. **Module 4: Tool Selector & Executor.** This is the core answering unit. It receives one ATOMIC question at a time (either the original or a sub-question). An LLM-based router selects the best "tool" for the job from a predefined set of prompt templates, such as a standard VQA prompt, an OCR-focused prompt, or a knowledge-integration prompt that uses the external facts gathered in Module 2. The executor then uses the selected prompt and the built context to generate an answer.
5. **Module 5: Synthesizer.** If the original question was ATOMIC, the answer from the executor is the final answer. If it was COMPOSITIONAL, this final module uses an LLM to synthesize the sequence of answers to the sub-questions into a single, coherent, natural language response.

This modular architecture allows our system to apply complex reasoning and knowledge retrieval only when necessary, making it more efficient and robust than a monolithic approach.

4 Results and Analysis

4.1 Quantitative Evaluation on A-OKVQA

We evaluated ARE-VQA on the A-OKVQA dataset [6], which comprises approximately 19,000 question–answer pairs. To analyze performance across different

reasoning dimensions, we triaged the dataset into four subsets based on question type: Visual–Atomic (V–A), Visual–Compositional (V–C), Knowledge–Atomic (K–A), and Knowledge–Compositional (K–C). Each subset isolates distinct reasoning challenges—ranging from direct perception to multi-hop reasoning with external knowledge.

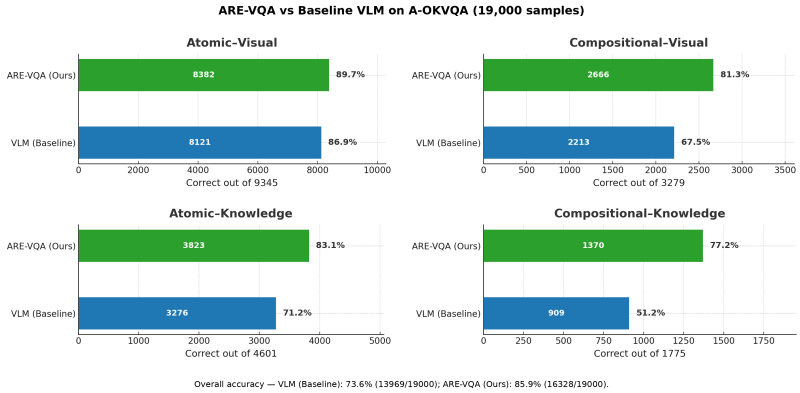

Fig. 1. Performance comparison between the baseline Visual Language Model (VLM) and our proposed ARE-VQA across the four reasoning categories. Each bar indicates both the absolute correct count and the corresponding accuracy percentage. Overall, ARE-VQA achieves 85.9% accuracy compared to the baseline's 73.6%.

As shown in Figure 1, ARE-VQA consistently outperforms the baseline across all categories. The largest gains are observed for compositional and knowledge-based queries, demonstrating the effectiveness of our modular reasoning design. Specifically:

- **Visual–Atomic (V–A):** ARE-VQA improves baseline accuracy from 86.9% to 89.7%, showing that adaptive tool selection benefits even simple perception-based queries.
- **Visual–Compositional (V–C):** A substantial gain from 67.5% to 81.3% validates the Query Planner's ability to handle multi-step visual reasoning.
- **Knowledge–Atomic (K–A):** Accuracy increases from 71.2% to 83.1%, attributed to the Context Builder's improved retrieval and evidence synthesis.
- **Knowledge–Compositional (K–C):** The most significant improvement—from 51.2% to 77.2%—illustrates how both the Context Builder and Query Planner modules synergistically enhance complex reasoning.

Overall, ARE-VQA achieves an aggregate accuracy of 85.9% (16,328 correct answers out of 19,000), outperforming the baseline by over 12 percentage points.

4.2 Ablation Study 1: Impact of Query Planner

To quantify the contribution of the Query Planner module, we removed it and directly prompted the LLM with the original compositional questions. As shown in Figure 2, performance dropped notably for both compositional subsets.

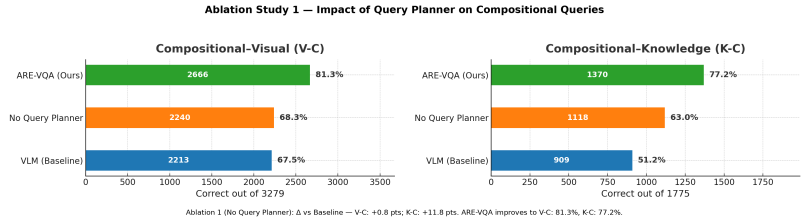

Fig. 2. Ablation Study 1 — Effect of removing the Query Planner. Without question decomposition, ARE-VQA's accuracy decreases from 81.3% to 68.3% on Visual–Compositional (V–C) and from 77.2% to 63.0% on Knowledge–Compositional (K–C).

This demonstrates that the Query Planner is essential for structured reasoning. Even though the baseline performs modestly on compositional tasks (67.5% on V–C and 51.2% on K–C), our planner-equipped model significantly improves performance by enabling the LLM to answer sub-questions sequentially and aggregate their results coherently.

4.3 Ablation Study 2: Impact of Context Builder

We next examined the influence of the Context Builder, which aggregates image-based evidence and retrieves external knowledge. When this module was disabled, the model relied only on the raw question and minimal visual context.

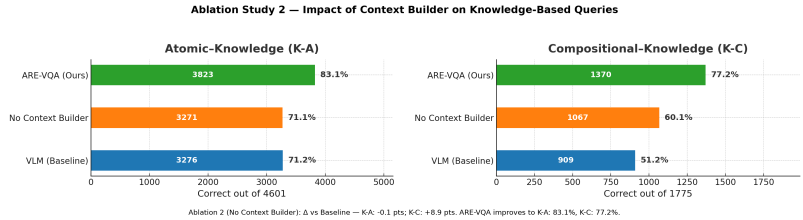

Fig. 3. Ablation Study 2 — Effect of removing the Context Builder. Without context aggregation, accuracy drops from 83.1% to 71.1% for Knowledge–Atomic (K–A) and from 77.2% to 60.1% for Knowledge–Compositional (K–C).

The absence of contextual enrichment particularly hinders knowledge-based reasoning. The results show a ∼12-point drop for atomic knowledge questions (K–A) and a ∼17-point drop for compositional knowledge questions (K–C), confirming that rich, targeted context is crucial for reliable LLM reasoning.

4.4 Qualitative Insights

Qualitatively, we observe that the Query Planner helps isolate specific visual entities and logically order intermediate steps, producing explanations closer to human reasoning. The Context Builder, in contrast, enhances factual grounding and reduces hallucination by explicitly introducing retrieved evidence. Together, these modules allow ARE-VQA to balance structured reasoning with factual precision.

References

1. Shao, Z., et al.: Prompting Large Language Models With Answer Heuristics for Knowledge-Based VQA. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023).
2. Tiong, A., et al.: A Simple Baseline for Knowledge-Based Visual Question Answering. In: Findings of the Association for Computational Linguistics: EMNLP (2022).
3. Hu, Y., et al.: PromptCap: Prompt-guided Image Captioning for VQA. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023).
4. Li, C., et al.: Fine-grained Late-interaction Multi-modal Retrieval for RA-VQA. In: Advances in Neural Information Processing Systems (NeurIPS) (2023).
5. Luo, S., et al.: RE-AU-SE: Retrieval-Augmented Visual Question Answering via Built-in Autoregressive Search Engines. In: Proceedings of the AAAI Conference on Artificial Intelligence (2024).
6. Schwenk, D., et al.: A-OKVQA: A Benchmark for Visual Question Answering using World Knowledge. In: European Conference on Computer Vision (ECCV) (2022).
7. Marino, K., et al.: OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019).