Adaptive Knowledge-Enriched Visual Question Answering via Retrieval Reflection Mechanisms

Aswathi P
dept. of Computer Science and Engineering
NSS College of Engineering
Palakkad, India
aswathipcmd@gmail.com

***Abstract*—Visual Question Answering (VQA) has emerged as a critical task in multimodal artificial intelligence, requiring systems to understand both visual content and natural language queries. While recent approaches leverage Retrieval-Augmented Generation (RAG) to incorporate external knowledge, they often retrieve information indiscriminately, leading to increased computational overhead and reduced answer relevance due to noisy context. This paper proposes an adaptive and lightweight framework for knowledge-based VQA that integrates retrieval-reflection and relevance-reflection mechanisms. The system intelligently determines when external knowledge is required and selectively filters only the most relevant information before answer generation. Image understanding is performed using a vision-language model, while knowledge retrieval is implemented using TF-IDF and cosine similarity over a curated local knowledge base. Experimental observations demonstrate that the proposed approach reduces unnecessary retrieval operations, improves contextual grounding, and enhances answer reliability. The framework achieves a balanced performance on both purely visual and knowledge-intensive queries while maintaining computational efficiency.**

***Index Terms*—Visual Question Answering, Retrieval-Augmented Generation, Multimodal AI, Knowledge Retrieval, LLM, TF-IDF**

I. Introduction

Visual Question Answering (VQA) has emerged as a prominent research area in multimodal artificial intelligence, where the objective is to enable machines to understand visual content and respond to natural language questions. By combining techniques from computer vision and natural language processing, VQA systems aim to bridge the gap between visual perception and language understanding. Such systems have found increasing relevance in real-world applications, including intelligent virtual assistants, educational tools, and assistive technologies for visually impaired individuals, where accurate interpretation of visual scenes is essential.

Conventional VQA approaches primarily depend on visual features extracted directly from images to generate answers. While these methods perform reasonably well for questions related to observable attributes such as object recognition, counting, or color identification, they face significant limitations when the question requires knowledge beyond the visible content. In many practical scenarios, answering a question involves reasoning with external or background knowledge. For instance, identifying a monument in an image may be achievable through visual cues, but answering a question like "When was the Taj Mahal built?" requires historical information that is not inherently present in the image itself. This highlights a fundamental gap in traditional VQA systems, where reliance on visual data alone is insufficient for comprehensive understanding.

To address this challenge, recent advancements have introduced Retrieval-Augmented Generation (RAG) frameworks, which incorporate external knowledge sources during the answer generation process. In these systems, relevant information is retrieved from large text corpora, such as Wikipedia, and combined with visual and textual inputs to produce more informed responses. While this approach improves the ability to handle knowledge-intensive queries, it also introduces new challenges. Existing RAG-based VQA systems often perform retrieval for every query, regardless of whether external knowledge is actually needed. This leads to unnecessary computational overhead and increased response latency. Moreover, these systems typically utilize all retrieved content without adequate filtering, which can introduce irrelevant or noisy information into the reasoning process, ultimately degrading answer quality.

Motivated by these limitations, this work proposes an adaptive framework termed Retrieval-Reflection Augmented VQA, which aims to make the retrieval process more intelligent and efficient. Instead of blindly retrieving external knowledge for every query, the proposed system first determines whether retrieval is necessary. If required, it further refines the retrieved content by selecting only the most relevant information before passing it to the answer generation module. This selective approach not only reduces computational cost but also improves the relevance and reliability of generated answers.

The key contributions of this work lie in the design of two novel mechanisms: a retrieval-reflection module that dynamically decides the necessity of knowledge retrieval, and a relevance-reflection module that filters out noisy or irrelevant information from retrieved data. In addition, the system integrates vision understanding, knowledge retrieval, and large language models into a unified and lightweight framework. To enhance trustworthiness, a confidence-based ranking strategy is also incorporated, enabling the system to provide more reliable and interpretable responses. Together, these contributions present a more efficient and robust approach to knowledge-

based visual question answering.

II. RELATED WORKS

Visual Question Answering (VQA) has evolved as a challenging multimodal task that requires joint reasoning over visual and textual information. Early VQA systems were primarily based on deep learning architectures trained on large-scale annotated image–question pairs. While these models achieved strong performance on benchmark datasets, they were largely limited to questions that could be answered using visual cues alone. Their inability to incorporate external or background knowledge made them less effective in handling real-world, knowledge-intensive queries.

With the emergence of Multimodal Large Language Models (MLLMs), recent research has shifted toward enhancing VQA systems by integrating external knowledge sources. These models leverage both visual representations and powerful language reasoning capabilities, enabling more flexible and generalizable solutions. In this context, zero-shot approaches such as ZPVQA attempt to reduce dependency on labeled datasets by converting visual inputs into textual descriptions and applying prompt-based reasoning. Although such methods improve scalability and reduce training cost, they still rely heavily on the implicit knowledge stored within the model, which limits their effectiveness for queries requiring explicit external information[10].

To address these limitations, Retrieval-Augmented Generation (RAG) has been widely adopted in VQA systems. RAG-based frameworks dynamically retrieve relevant knowledge from external sources, such as Wikipedia or large document collections, and incorporate it into the answer generation process. Several studies have explored different retrieval strategies to improve performance. For example, hierarchical retrieval approaches such as Wiki-LLaVA organize the retrieval process into multiple stages, enabling more structured and effective access to large knowledge repositories. Similarly, MMKB-RAG introduces knowledge-boundary-aware mechanisms that determine whether retrieval is necessary, thereby reducing irrelevant context and improving answer accuracy. Multi-RAG further extends this idea by incorporating multiple modalities, including text, images, and videos, to enrich contextual understanding[4],[3],[7].

Despite these advancements, traditional RAG-based approaches often rely on unstructured textual data, which can introduce significant noise into the reasoning process. This not only increases the computational burden but also negatively impacts answer quality. To mitigate this issue, reflection-based frameworks such as mR²AG introduce adaptive strategies that include retrieval-reflection and relevance-reflection mechanisms. These methods aim to selectively retrieve and filter information, thereby improving reasoning efficiency and reducing noise in the generated responses[1].

In addition to text-based retrieval, recent works have explored structured knowledge representations to enhance VQA performance. The mKG-RAG framework incorporates multimodal knowledge graphs into the retrieval process, enabling the model to capture relationships between entities and perform more precise reasoning. While this approach improves knowledge grounding, it also introduces additional complexity in graph construction and maintenance[2].

Memory-augmented approaches represent another direction in this domain. Frameworks such as REVEAL utilize large-scale multimodal knowledge memory systems that store diverse information sources, including image–text pairs and knowledge graph triples. By jointly training retrieval and generation components, these models improve both retrieval quality and reasoning capability. However, they require substantial computational resources and large-scale training data, which may limit their practical applicability[6].

Overall, existing approaches have made significant progress in integrating external knowledge into VQA systems. However, several challenges remain unresolved. Many RAG-based methods still retrieve unnecessary or irrelevant information, leading to noise and inefficiency. Multimodal alignment between visual and textual data continues to be a complex issue, and structured or memory-based approaches often introduce additional computational overhead. These limitations highlight the need for a more efficient and adaptive framework that can intelligently control the retrieval process and ensure the use of only relevant knowledge.

III. PROPOSED METHODOLOGY

The proposed framework is designed as an adaptive multimodal pipeline that integrates image understanding, knowledge retrieval, and language generation for Visual Question Answering (VQA). Unlike conventional Retrieval-Augmented Generation (RAG) systems that perform retrieval for every query, the proposed method introduces reflection-based mechanisms to selectively control retrieval and filter relevant knowledge. This results in a more efficient and reliable system capable of handling both purely visual and knowledge-intensive queries.

The overall workflow of the system is illustrated conceptually as a sequential pipeline, where each module contributes to progressively refining the input information before generating the final answer.

A. Dataset and Knowledge Base Construction

To support knowledge-based reasoning, a curated local knowledge base is constructed using textual data derived from publicly available sources such as Wikipedia. The dataset is designed to include diverse categories of general knowledge, ensuring coverage across multiple domains. The knowledge base primarily consists of information related to:

- Historical landmarks and monuments
- Scientific concepts and technological topics
- General knowledge such as geography and environment

All documents are stored locally to reduce dependency on external APIs during runtime and to ensure faster retrieval. Each document is organized in a structured textual format, enabling efficient indexing and search operations. This lightweight knowledge base is sufficient for validating the effectiveness of

the proposed adaptive retrieval framework while maintaining low computational overhead.

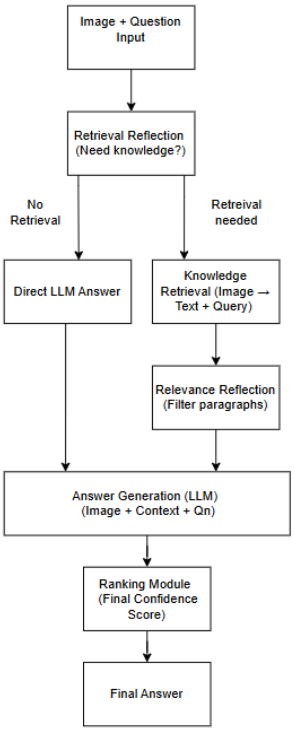

Fig. 1. overall workflow of the proposed system.

B. Data Preprocessing and Representation

Before retrieval, all textual data undergoes preprocessing to convert raw text into a structured and searchable format. This step plays a crucial role in improving retrieval accuracy and efficiency. Initially, documents are cleaned by removing unnecessary symbols, stopwords, and redundant information. The processed text is then transformed into numerical representations using the Term Frequency–Inverse Document Frequency (TF-IDF) technique. TF-IDF helps in identifying the importance of words within each document relative to the entire corpus. Each document is represented as a vector in a high-dimensional space, capturing its semantic relevance. These vectors are stored and later used for similarity comparison during query processing. For efficient retrieval, cosine similarity is employed to measure the closeness between the query vector and document vectors.

C. Input Processing

The system accepts a multimodal input consisting of an image and a natural language question. The image provides visual context, while the question specifies the required information. To enable integration with the retrieval system, the image is first processed using a vision-language model, which converts visual content into a descriptive textual representation. This textual description serves as an intermediate representation that bridges the gap between visual and textual modalities.

D. Retrieval-Reflection Mechanism

The retrieval-reflection module is a key component of the proposed system, responsible for determining whether external knowledge retrieval is necessary for a given query. Instead of performing retrieval for every input, the system analyzes the question using a keyword-based scoring mechanism. Queries that involve visual attributes such as color, count, or object identification are classified as visual questions, and retrieval is skipped. In contrast, queries requiring factual or background knowledge are classified as knowledge-based questions, triggering the retrieval process. This adaptive decision-making significantly reduces unnecessary computation and prevents the inclusion of irrelevant knowledge, thereby improving system efficiency.

E. Knowledge Retrieval

When retrieval is required, the system generates a query by combining the image description with the user's question. This combined query captures both visual context and semantic intent. The query is transformed into a TF-IDF vector and compared with document vectors stored in the knowledge base. Cosine similarity is used to identify the most relevant documents, and the top-K documents with the highest similarity scores are retrieved. This approach ensures that only contextually relevant information is selected for further processing.

F. Relevance-Reflection Mechanism

To further refine the retrieved information, a relevance-reflection module is introduced. Retrieved documents are divided into smaller textual units, such as paragraphs, to enable fine-grained filtering. Each paragraph is evaluated independently by computing its similarity with the query. Only those segments that exceed a predefined relevance threshold are selected, while irrelevant or noisy content is discarded. This step plays a crucial role in improving answer quality by ensuring that only meaningful and contextually aligned information is passed to the answer generation module.

G. Answer Generation

The filtered knowledge, along with the image description and the original question, is provided as input to a Large Language Model (LLM) based on the LLaMA architecture. The model generates a natural language response by reasoning over both visual and textual inputs. In addition to the answer, the model also produces a confidence score indicating the reliability of the generated response.

H. Ranking Mechanism

To ensure reliability, a final ranking mechanism is applied to compute the overall confidence of the generated answer. The final score is calculated by combining multiple factors, including retrieval relevance, filtering effectiveness, and model confidence.

Final Score=Retrieval Score×Relevance Score×Answer Confidence

This composite scoring mechanism provides a more robust estimate of answer reliability and enhances the interpretability of the system.

IV. EXPERIMENTAL SETUP

The experimental study is implemented as a modular framework integrating multiple components for image understanding, knowledge retrieval, and answer generation. The overall implementation is carried out using Python, with a Flask-based backend to manage system operations and handle user interactions. The framework is designed to support efficient processing of multimodal inputs while maintaining scalability and modularity.

For visual understanding, the system utilizes a vision-enabled large language model accessed through the Groq API. This model is responsible for converting input images into descriptive textual representations and also contributes to the final answer generation process. The integration of vision and language capabilities enables the system to bridge the gap between visual content and textual reasoning.

Knowledge retrieval is performed using a classical information retrieval approach based on Term Frequency–Inverse Document Frequency (TF-IDF). The TF-IDF implementation provided by the Scikit-learn library is used to transform textual data into vector representations. These vectors are then compared using cosine similarity to identify the most relevant documents for a given query. To further improve retrieval efficiency, especially for larger datasets, Facebook AI Similarity Search (FAISS) is incorporated as an optional indexing mechanism. FAISS enables faster similarity search by organizing vector representations into optimized data structures, thereby reducing retrieval time.

Numerical computations and vector operations throughout the system are handled using the NumPy library, ensuring efficient processing of high-dimensional data. The entire pipeline is structured into distinct functional modules, each responsible for a specific task within the VQA process.

The system architecture consists of five primary modules. The Retrieval-Reflection module determines whether external knowledge retrieval is required based on the nature of the query, thereby avoiding unnecessary computation. The Knowledge Retrieval module generates a query by combining the image description and the user's question, and retrieves relevant documents from the knowledge base using TF-IDF and similarity measures. The Relevance-Reflection module further refines the retrieved information by filtering out irrelevant content at a finer granularity. The Answer Generator module leverages a LLaMA-based language model to produce the final response based on the filtered knowledge and visual context. Finally, the Ranking module computes a confidence score by combining multiple evaluation factors, ensuring the reliability and interpretability of the generated answers.

This modular design allows each component to operate independently while contributing to a cohesive and efficient pipeline, making the system suitable for real-world multimodal question answering applications.

V. EVALUATION METRICS

To evaluate the performance of the proposed system, multiple metrics are considered to capture both answer quality and retrieval effectiveness. Answer accuracy is used to measure the correctness of generated responses with respect to expected answers. Retrieval precision evaluates the relevance of documents retrieved from the knowledge base, while filtering efficiency measures the ability of the relevance-reflection module to remove irrelevant information. In addition, a retrieval reduction rate is introduced to quantify how effectively the system avoids unnecessary retrieval operations. The final response confidence is computed using a combined scoring mechanism, reflecting the reliability of the generated answer. These metrics together provide a comprehensive evaluation of both efficiency and answer quality.

VI. RESULTS AND DISCUSSIONS

The proposed system is evaluated through a series of qualitative experiments using real-world image–question pairs. The evaluation focuses on assessing the system's ability to handle both purely visual queries and knowledge-based queries by leveraging its adaptive retrieval and reflection mechanisms. The results are obtained from the deployed Flask-based system interface, where users provide an image and a corresponding question, and the system generates an answer along with confidence and retrieval information .

A. Performance on Knowledge-Based Queries

The system demonstrates strong performance on knowledge-intensive questions that require external information beyond the visual content. For instance, when provided with an image of the Taj Mahal and the question "When was it officially opened?", the system correctly identifies the need for external knowledge and activates the retrieval mechanism. The retrieved contextual information is then filtered using the relevance-reflection module before being passed to the language model.

The generated response correctly states that the Taj Mahal was completed in 1653, while also providing additional contextual information about its construction timeline. This indicates that the system not only retrieves relevant knowledge but also integrates it effectively during answer generation.

The retrieval decision is supported by a confidence score (e.g., 30%), reflecting the system's internal assessment of the necessity of external knowledge. This behavior aligns with the design of the retrieval-reflection module, which uses keyword-based heuristics to classify queries .

B. Performance on Identity and Contextual Queries

The system also performs effectively on queries that involve identifying contextual attributes such as nationality. In the case of an image depicting a well-known personality (e.g.,

Mahatma Gandhi), the system correctly identifies the individual and provides the appropriate answer, such as "The person depicted in the image is Indian."

In such cases, the system may still activate retrieval due to the presence of knowledge-related keywords in the query (e.g., "nationality"). However, the final answer remains concise and accurate, indicating that the retrieval and filtering mechanisms do not introduce unnecessary noise. This demonstrates the robustness of the relevance-reflection module in maintaining answer quality.

C. Performance on Visual-Only Queries

For purely visual questions, the system effectively avoids unnecessary retrieval operations. For example, when asked "What is the color of the bird?", the system correctly classifies the query as a visual question and bypasses the retrieval stage. The answer is generated directly using visual understanding, producing a response such as "The colors of the bird are yellow, orange, green, and blue."

This behavior highlights the effectiveness of the retrieval-reflection mechanism in reducing computational overhead. By skipping retrieval when it is not required, the system minimizes latency and avoids introducing irrelevant knowledge into the reasoning process.

D. Effectiveness of Adaptive Retrieval

A key observation from the experimental results is the system's ability to dynamically switch between retrieval-based and non-retrieval-based processing. The retrieval-reflection module plays a crucial role in this decision-making process by analyzing the query using predefined patterns and scoring mechanisms.

As implemented in the system, the retrieval decision is based on keyword matching and scoring thresholds, allowing the system to distinguish between visual and knowledge-based queries efficiently . This adaptive behavior significantly improves efficiency compared to traditional RAG systems, which perform retrieval for every query regardless of necessity.

E. Impact of Relevance-Reflection Filtering

The relevance-reflection module contributes significantly to improving answer quality by filtering retrieved documents at a finer granularity. Instead of passing entire documents to the language model, the system evaluates individual paragraphs using similarity scores and selects only those that are relevant to the query.

This filtering mechanism reduces noise and ensures that the language model receives focused and meaningful context. As a result, the generated answers are more precise and less prone to hallucination. The implementation uses cosine similarity between embeddings to determine relevance, ensuring semantic alignment between the query and retrieved content .

F. Confidence and Ranking Analysis

The system provides multiple confidence indicators, including retrieval confidence, answer confidence, and a final composite score. The final score is computed using a multiplicative ranking function that combines retrieval relevance, filtering effectiveness, and answer confidence .

In the observed results, the system consistently produces high answer confidence values (close to 1.0), indicating strong confidence in generated responses. The inclusion of a composite score enhances interpretability and provides users with an indication of answer reliability.

G. Discussion

The experimental results demonstrate that the proposed mR2AG framework effectively addresses key limitations of traditional VQA and RAG systems. By introducing adaptive retrieval and relevance filtering, the system achieves a balance between efficiency and answer quality. The ability to selectively perform retrieval reduces unnecessary computation and improves response time, while the relevance-reflection mechanism ensures that only useful information contributes to answer generation. Additionally, the integration of a vision-language model with a lightweight knowledge base enables the system to handle a wide range of queries without requiring large-scale datasets. However, the current evaluation is primarily qualitative and based on a curated knowledge base, which limits the ability to perform standardized benchmarking. Future work should include evaluation on established datasets such as OK-VQA to further validate performance.

VII. CONCLUSION

This paper presented an adaptive and lightweight framework for knowledge-based Visual Question Answering, referred to as mR2AG (Multimodal Retrieval-Reflection-Augmented Generation). The proposed system integrates image understanding, selective knowledge retrieval, and language generation into a unified pipeline, with a primary focus on improving efficiency and answer reliability. Unlike conventional Retrieval-Augmented Generation approaches that perform retrieval for every query, the proposed method introduces a retrieval-reflection mechanism to determine whether external knowledge is required. This significantly reduces unnecessary computation for purely visual questions. In addition, a relevance-reflection mechanism is employed to filter retrieved information at a finer granularity, ensuring that only contextually relevant content is used during answer generation. Experimental observations demonstrate that the system effectively handles both visual and knowledge-based queries. The adaptive retrieval strategy minimizes noise and latency, while the filtering process improves the precision of generated answers. Furthermore, the inclusion of a confidence-based ranking mechanism enhances the interpretability and trustworthiness of the system outputs. Overall, the proposed framework provides a practical solution to the limitations of traditional VQA and RAG systems by balancing accuracy, efficiency, and scalability. The modular design also allows easy extension to larger knowledge bases and more advanced retrieval techniques. However, the current work is evaluated on a curated knowledge base and qualitative test scenarios, which limits direct comparison with benchmark datasets. Future work

will focus on extending the system to large-scale datasets such as OK-VQA, incorporating advanced dense retrieval methods, and improving reasoning capabilities of the language model to further enhance performance.

VIII. FUTURE WORK

The proposed mR2AG framework can be further improved by expanding the knowledge base to support a wider range of real-world queries. Incorporating advanced retrieval techniques, such as dense embedding-based methods, can enhance semantic matching and retrieval accuracy. The retrieval-reflection mechanism may also be extended using learning-based approaches to improve decision-making in complex queries. In addition, future work will focus on evaluating the system on standard benchmark datasets such as OK-VQA and improving multimodal reasoning capabilities using more advanced vision-language models.

REFERENCES

[1] Zhang, T., Zhang, Z., Ma, Z., Chen, Y., Qi, Z., Yuan, C., Li, B., Pu, J., Zhao, Y., Xie, Z., Ma, J., Shan, Y., Hu, W. (2024). mR²AG: Multimodal retrieval-reflection-augmented generation for knowledge-based VQA. arXiv preprint arXiv:2411.15041. https://arxiv.org/abs/ 2411.15041

[2] Yuan, X., Ning, L., Fan, W., Li, Q. (2025, August). mKG-RAG: Multimodal knowledge graph-enhanced RAG for visual question answering. arXiv preprint arXiv:2508.05318. https://doi.org/10.48550/arXiv.2508.05318

[3] Ling, Z., Guo, Z., Huang, Y., An, Y., Xiao, S., Lan, J., Zhu, X., Zheng, B. (2025). MMKB-RAG: A multi-modal knowledge-based retrieval-augmented generation framework. arXiv preprint arXiv:2504.10074. https://arxiv.org/abs/ 2504.10074

[4] Caffagni, D., Cocchi, F., Moratelli, N., Sarto, S., Cornia, M., Baraldi, L., Cucchiara, R. (2024). Wiki-LLaVA: Hierarchical retrieval-augmented generation for multimodal LLMs. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (pp. 1818–1826). IEEE. https://doi.org/10.1109/CVPRW63382.2024.00188

[5] Yan, Y., Xie, W. (2024, November). EchoSight: Advancing visual-language models with Wiki knowledge. In Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 1538–1551). Association for Computational Linguistics. https://aclanthology.org/2024.findings-emnlp.83

[6] Hu, Z., Iscen, A., Sun, C., Wang, Z., Chang, K.-W., Sun, Y., Schmid, C., Ross, D. A., Fathi, A. (2023). Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 23369–23379). IEEE. https://doi.org/10.1109/ CVPR52729.2023.02238

[7] Mao, M., Perez-Cabarcas, M. M., Kallakuri, U., Waytowich, N. R., Lin, X., Mohsenin, T. (2025). Multi-RAG: A multimodal retrieval-augmented generation system for adaptive video understanding. arXiv preprint arXiv:2505.23990. https://arxiv.org/abs/ 2505.23990

[8] Mensink, T., Uijlings, J., Castrejon, L., Goel, A., Cadar, F., Zhou, H., Sha, F., Araujo, A., Ferrari, V. (2023). Encyclopedic VQA: Visual questions about detailed properties of fine-grained categories. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 3090–3101). IEEE. https://doi.org/10.1109/ICCV51070.2023.00289

[9] Chen, Y., Hu, H., Luan, Y., Sun, H., Changpinyo, S., Ritter, A., Chang, M.-W. (2023, December). Can pre-trained vision and language models answer visual information-seeking questions? In H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 14948–14968). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.925

[10] Hu, N., Zhang, X., Zhang, Q., Huo, W., You, S. (2025). ZPVQA: Visual question answering of images based on zero-shot prompt learning. IEEE Access, 13, 50849–50859. https://doi.org/10.1109/ACCESS.2025.3550942

[11] Liu, H., Li, C., Li, Y., Lee, Y. J. (2024). Improved baselines with visual instruction tuning. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 26286–26296). IEEE. https://doi.org/10.1109/CVPR52733.2024.02484