Proceedings of CONF-SEML 2026 Symposium: Importance of Machine Learning Methods and Analysis in Engineering
DOI: 10.54254/2755-2721/2026.AS31461

Research on Multimodal Large Language Models for Visual Question Answering: Advances and Challenges

Liangyu Mei

Leeds Joint School, Southwest Jiaotong University, Chengdu, China
el23lm@leeds.ac.uk

Abstract. Multimodal understanding, which requires models to jointly reason over visual and linguistic information, has become a core challenge in artificial intelligence (AI). Visual Question Answering (VQA) stands as a paradigmatic task for investigating these multimodal reasoning capabilities. While early VQA systems relied on task-specific architectures, recent breakthroughs in Multimodal Large Language Models (MLLMs) have significantly reshaped the field by proposing unified, instruction-driven multimodal reasoning frameworks. By conducting a systematic literature review, this paper scrutinizes the evolution of VQA from traditional CNN–LSTM-based models to modern MLLM-based approaches. The review centers on representative architectures and training paradigms, including BLIP-2 and LLaVA, to analyze how large language models and pretrained vision encoders are integrated for flexible and open-ended visual reasoning. In addition, this paper identifies and deliberates on critical challenges confronting contemporary MLLMs, encompassing modality imbalance, insufficient cross-modal alignment, and hallucinations. This paper concludes that while MLLMs have substantially expanded the application scope and functional capabilities of VQA systems, they still grapple with reliable visual grounding and balanced multimodal fusion. Addressing these limitations is paramount for constructing trustworthy and robust VQA systems, and future research should prioritize improving alignment mechanisms and mitigating hallucinations in multimodal reasoning.

Keywords: Visual Question Answering, Multimodal Large Language Models, Multimodal Reasoning, Vision–Language Alignment

1. Introduction

Multimodal understanding, which entails jointly reasoning over visual and linguistic information, has become an important research topic in artificial intelligence [1]. Visual Question Answering (VQA) serves as a representative setting to study this capability, as it explicitly demands grounded reasoning between images and natural language queries. However, conventional VQA systems were predominantly constructed based on task-specific pipelines, wherein visual and textual features were processed in isolation and fused at a downstream stage [2-4]. Such designs constrain the models' capacity to perform flexible reasoning and generalize beyond predefined tasks, underscoring the imperative for more unified multimodal frameworks.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.AS31461

Recent advances in large-scale pretraining have spurred the development of Multimodal Large Language Models (MLLMs), which integrate powerful visual encoders with large language models to support instruction-driven multimodal reasoning [5]. Representative studies, such as BLIP-2 [6], have shown that freezing pre-trained vision and language backbones while incorporating lightweight adapter modules enables efficient achievement of state-of-the-art performance. Likewise, LLaVA [7] has demonstrated that instruction tuning can be extended to the vision–language domain, enabling models to respond to diverse user queries in a conversational manner. These works illustrate that VQA is no longer regarded as an isolated task, but rather as one aspect of a broader multimodal reasoning capability enabled by foundation models. Concurrently, recent studies also reveal unresolved issues, including insufficient visual grounding and over-reliance on language priors.

Against this background, this paper scrutinizes how MLLMs are reshaping VQA by reviewing representative architectures and training paradigms. It centers on critical challenges such as modality imbalance, cross-modal alignment, and hallucination. By synthesizing current advancements and inherent limitations, this study seeks to foster a more comprehensive understanding of MLLM-based VQA systems, thereby offering valuable insights for future research endeavors and practical implementations in the field of multimodal reasoning.

2. Comparison of tradition VQA and MLLMs

2.1. Limitations of traditional VQA

Traditional Visual Question Answering (VQA) typically entails an image paired with a natural language query, where the model is required to predict an appropriate answer. Early VQA approaches commonly adopt a dual-stream architecture, in which Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs) are utilized to process visual and textual modalities separately. CNN-based backbones such as VGGNet [8] and ResNet [9] extract discriminative visual features from images, while LSTMs [10] encode the sequential semantics of questions. The resulting representations are subsequently fused and mapped to a predefined answer space through a classification module [2].

Building upon this paradigm, a suite of attention-based models have been proposed to enhance visual grounding. For instance, the Stacked Attention Network (SAN) [11] implements multi-step attention over image regions to progressively refine relevant visual evidence. Likewise, the Bilinear Attention Network (BAN) [12] introduces bilinear pooling mechanisms to model fine-grained interactions between visual regions and linguistic tokens. Although these models enhance performance by reinforcing cross-modal correlations, they predominantly depend on hand-engineered fusion strategies and task-specific architectural configurations.

Despite these advances, traditional VQA models are plagued by several inherent limitations. First, visual and linguistic modalities are often processed in isolation and only integrated at later stages, leading to inadequate cross-modal alignment and shallow multimodal reasoning [13]. Second, most VQA systems formulate answer prediction as a classification problem over a fixed label space, which constrains their generalizability to scenarios beyond dataset-specific distributions. Moreover, such models grapple with compositional reasoning, multi-step inference, and instruction-following behavior, making them poorly suited for open-ended and interactive multimodal tasks.

These limitations underscore a fundamental bottleneck in conventional VQA frameworks: they are engineered to address narrowly defined tasks rather than facilitate general-purpose multimodal reasoning. This has catalyzed a paradigm shift toward Multimodal Large Language Models

DOI: 10.54254/2755-2721/2026.AS31461

(MLLMs), which seek to unify vision and language understanding within a flexible, scalable, and instruction-driven framework.

2.2. Overview of MLLMs

Over the recent years, the field has witnessed a significant transition from task-specific VQA architectures to foundation models capable of handling multimodal reasoning. These burgeoning MLLMs not only surmount the alignment and generalization bottlenecks of conventional counterparts but also revolutionize the formulation and solution of visual question answering.

Instead of confining answer prediction to a fixed label space, MLLMs support open-ended, instruction-driven reasoning, empowering them to address diverse question types, generate explanations, and engage in multi-turn interactions anchored in visual content. This shift transforms VQA from a narrowly defined recognition task into a more general multimodal reasoning problem.

Multimodal Large Language Models (MLLMs) are large-scale foundational frameworks constructed on large language models that integrate pretrained vision encoders with language models, enabling unified perception and reasoning across visual and textual modalities, and generating outputs anchored in multimodal context [14]. Rather than treating vision and language as discrete processing pipelines, MLLMs seek to co-model cross-modal semantic representations within a unified reasoning framework.

In general, an MLLM comprises a visual encoding component responsible for extracting semantically rich representations from input images. These visual encoders are predominantly built upon large-scale pretrained models such as CLIP [15], ViLT [16], and BLIP [17], which acquire transferable visual–linguistic representations via contrastive image–text learning and cross-modal alignment. Through extensive pretraining on large-scale image-text corpora, these encoders provide high-level visual features that capture both object-level semantics and contextual information.

To address the modality gap between visual feature representations and the LLM, MLLMs incorporate an adapter module that projects visual embeddings into a format compatible with the token-based input space of the language model. Representative designs include the Querying Transformer (Q-Former) used in BLIP-2 [6] and the lightweight linear projection layers utilized in LLaVA [7]. These modules convert visual embeddings into a compact set of visual tokens, enabling efficient cross-modal interaction while circumventing the prohibitive computational cost incurred by end-to-end fine-tuning of large vision encoders within LLMs.

At the core of an MLLM resides the large language model (LLM), functioning as the primary reasoning engine. Endowed with robust in-context learning capacities, comprehensive world knowledge, and instruction-compliant capabilities, the LLM conducts conditional reasoning on both textual inputs and visual tokens. Leveraging this unified representation space, the model is able to generate coherent and contextually grounded natural language responses, supporting flexible multimodal reasoning across a wide range of vision–language tasks.

3. Representative case studies

3.1. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models

BLIP-2 is a representative Multimodal Large Language Model (MLLM) that proposes a lightweight Querying Transformer (Q-Former) to bridge the frozen vision encoder and large language model (LLM). In this architecture, both the vision encoder and LLM remain frozen during training—

DOI: 10.54254/2755-2721/2026.AS31461

meaning their parameters are not updated—while the Q-Former serves as the only trainable component. Its core function is to distill semantically salient visual information and convert it into a compact set of visual tokens compatible with the LLM's input format. These tokens are then fed into the LLM, which jointly processes them with textual inputs, enabling unified multimodal reasoning [6].

Unlike earlier multimodal systems that depended on late fusion of visual and textual features, BLIP-2 enables integrated reasoning across both modalities. This enables the model to not only generate accurate answers but also produce interpretable reasoning rationales to justify its outputs—in contrast to traditional VQA models that merely output a probability distribution over predefined candidate answers. Despite having significantly fewer trainable parameters, BLIP-2 achieves state-of-the-art performance across diverse vision-language benchmarks, while remaining computationally efficient. Furthermore, the model demonstrates strong zero-shot generalization to diverse tasks such as image captioning and image-text retrieval, highlighting its versatility as a universal multimodal foundation model.

3.2. LLaVA: large language and vision assistant

In recent years, there has been a burgeoning trend in academia toward developing language-augmented foundation vision models [7]. These models demonstrate proficiency in specific vision-language tasks such as classification, segmentation, and image captioning. Nevertheless, each model is generally confined to a single predefined task and lacks the capacity to parse diverse instructional prompts, with its behavior restricted to the implicit task objectives encoded during model development. Furthermore, language in these models is primarily utilized to describe image content, rather than being fully leveraged to support more sophisticated interactions such as free-form user queries, commands, or follow-up questions. In other words, they lack interactivity and adaptability.

To mitigate these drawbacks, LLaVA (Large Language and Vision Assistant) was proposed to achieve multimodal conversational capabilities, enabling it to handle various vision-language tasks and interact with users in a dialogue-like manner. The architecture of LLaVA consists of a CLIP [15] vision encoder, a Vicuna LLM [18], and a trainable projection matrix that projects visual features into linguistic embedding tokens with consistent dimensionality to the language model's word embedding space.

Another key contribution of LLaVA is that it applies instruction tuning to the vision-language domain for the first time. Specifically, it reframes conventional image-text datasets (e.g., COCO) into instruction-following paradigms, enabling the model to acquire natural interaction capabilities with users grounded in visual inputs. LLaVA exhibits strong zero-shot performance and human-like conversational abilities across various multimodal benchmarks, representing a notable advancement in unified vision–language reasoning.

4. Discussion

The advent of MLLMs has fundamentally reshaped the landscape of Visual Question Answering (VQA). By fusing large language models (LLMs) with high-performance visual encoders, paradigmatic systems such as BLIP-2 and LLaVA exhibit unified multimodal reasoning capabilities and attain notable zero-shot generalization. However, despite these breakthroughs, several key challenges remain that prevent MLLM-based VQA from reaching human-level understanding.

DOI: 10.54254/2755-2721/2026.AS31461

4.1. Modality imbalance

Contemporary research has corroborated that numerous MLLMs exhibit a pronounced modality imbalance, wherein models over-rely on linguistic priors ingrained in their linguistic backbones while underutilizing visual representations furnished by vision encoders [19]. As a result, the visual modality contributes less effectively to the final prediction, causing the model to behave more like a text-only LLM in the face of ambiguous inputs. This imbalance often manifests as incorrect visual grounding, such as failing to identify critical objects or misinterpreting spatial relations, and can lead to hallucinated responses wherein models confidently assert the existence of non-existent visual content.

4.2. Proper alignment

Achieving proper alignment between multimodal inputs persists as a pivotal challengefor current MLLMs [7]. Visual features and textual representations differ substantially in both structure and semantic granularity, which makes it difficult for models to establish precise correspondences between image regions and the words or concepts expressed in a query. In many cases, visual information is aggregated into high-level embeddings before being passed to the language model, leading to the loss of fine-grained spatial or relational nuances. When alignment is suboptimal, the model may fail to attend to critical visual evidence, misinterpret object attributes or relationships, or rely excessively on linguistic priors learned during pretraining. As such, responses may be partially visually grounded, factually inconsistent with visual inputs, or outright hallucinatory. Enhancing cross-modal alignment is therefore paramount to facilitating accurate visual grounding and reliable multimodal reasoning in MLLMs.

4.3. Hallucinations

The problem of hallucinations arises because MLLMs inherently inherit the risks and failure modes of their underlying LLMs. In particular, hallucinations denote cases where the model generates text responses that are inconsistent with—or entirely unrelated to—the visual input [20]. Given the linguistic backbone's overwhelming dominance during the inference phase, MLLMs may produce confidently articulated yet visually ungrounded answers—particularly in open-ended VQA settings. As a result, hallucinations pose a serious challenge to developing trustworthy, safe, and reliable multimodal systems [21]. They not only degrade VQA performance but also constrain the deployment of MLLMs in high-stakes, real-world applications where factual congruence with visual evidence is critical.

5. Conclusion

This paper systematically delineates the evolutionary trajectory of Visual Question Answering (VQA), spanning from early task-specific multimodal architectures to the recent advent of Multimodal Large Language Models (MLLMs). Early CNN–LSTM architectures enabled basic visual–textual reasoning but were hindered by inadequate semantic alignment, limited generalization, and task-specific design. With the integration of stronger visual encoders and large language models, systems such as BLIP-2, LLaVA, and GPT-4V now provide unified multimodal pipelines capable of instruction-based reasoning across a diverse spectrum of visual tasks.

This research contributes to the existing body of knowledge by providing a structured and conceptually rigorous analysis of how Multimodal Large Language Models (MLLMs) are reshaping

DOI: 10.54254/2755-2721/2026.AS31461

the Visual Question Answering (VQA) paradigm. While prior studies have either focused on task-specific VQA architectures or examined MLLMs from a general multimodal perspective, this work bridges the gap between these two research strands by explicitly situating VQA within the broader MLLM framework. In doing so, it elucidates the paradigmatic shift from classification-centric VQA systems toward instruction-driven, open-ended multimodal reasoning.

This study is limited by its qualitative and literature-based methodology, which focuses on conceptual analysis rather than empirical validation. Consequently, the conclusions are derived from existing studies and representative models, and may not fully capture the performance variability of MLLMs across different datasets or application scenarios. In addition, the analysis focuses predominantly on widely adopted architectures such as BLIP-2 and LLaVA, which may bias the discussion toward current mainstream design choices. An additional limitation pertains to the lack of quantitative benchmarking or controlled experiments, which may be mitigated in future inquiries via systematic empirical comparisons and large-scale validation studies.

Future research may prioritize mitigating the identified constraints through the implementation of systematic empirical evaluations of MLLMs across diverse VQA benchmarks and real-world scenarios. In particular, future research may explore improved cross-modal alignment mechanisms, stronger visual grounding strategies, and training objectives that explicitly mitigate hallucinatory outputs. In addition, the author will consider extending the analysis to emerging architectures and multimodal datasets as the field continues to evolve. Overall, this study provides new insights into how MLLMs are reshaping VQA and highlights the importance of balanced modality fusion and reliable visual grounding. By illuminating prevailing challenges and emerging opportunities, this work lays the foundation for the development of more trustworthy and robust multimodal reasoning systems.

References

[1] Kim, B.S., Kim, J., Lee, D., Jang, B. (2023) Visual question answering: A survey of methods, datasets, evaluation, and challenges. ACM Computing Surveys, 56: 1–39.
[2] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D. (2015) VQA: Visual question answering. International Journal of Computer Vision, 123: 4–31.
[3] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L. (2018) Bottom-up and top-down attention for image captioning and visual question answering. Computer Vision and Image Understanding, 173: 20–31.
[4] Teney, D., Anderson, P., He, X., van den Hengel, A. (2018) Tips and tricks for visual question answering: Learnings from the 2017 challenge. Computer Vision and Image Understanding, 173: 72–82.
[5] Caffagni, D., Cocchi, F., Barsellotti, L., Moratelli, N., Sarto, S., Baraldi, L., Cornia, M., Cucchiara, R.（2024）The revolution of multimodal large language models: a survey. arXiv preprint, arXiv: 2402.12451.
[6] Li, J., Li, D., Savarese, S., Hoi, S.C.H. (2023) BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning (ICML). Honolulu, USA. pp.19730–19742.
[7] Liu, H., Li, C., Wu, Q., Lee, Y.J.（2023）Visual instruction tuning. arXiv preprint, arXiv: 2304.08485.
[8] Simonyan, K., Zisserman, A. (2015) Very deep convolutional networks for large-scale image recognition. arXiv preprint, arXiv: 1409.1556.
[9] He, K., Zhang, X., Ren, S., Sun, J. (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.
[10] Hochreiter, S., Schmidhuber, J. (1997) Long short-term memory. Neural Computation, 9: 1735–1780.
[11] Yang, Z., He, X., Gao, J., Deng, L., Smola, A. (2016) Stacked attention networks for image question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas. pp. 21–29.
[12] Kim, J.-H., Jun, J., Zhang, B.-T. (2018) Bilinear attention networks. In: Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS). Montréal. pp. 1564–1574.

DOI: 10.54254/2755-2721/2026.AS31461

[13] Lu, J., Yang, J., Batra, D., Parikh, D. (2016) Hierarchical question-image co-attention for visual question answering. In: Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS). Barcelona. pp. 289–297.
[14] Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E. (2024) A survey on multimodal large language models. IEEE Transactions on Pattern Analysis and Machine Intelligence.
[15] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I. (2021) Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML). Virtual Conference. pp. 8748–8763.
[16] Kim, W., Son, B., Kim, I. (2021) ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML). Virtual Conference. pp. 5583–5594.
[17] Li, J., Li, D., Xiong, C., Hoi, S.C.H. (2022) BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In: Proceedings of the 39th International Conference on Machine Learning (ICML). Baltimore, USA. pp. 12888–12900.
[18] Chiang, W.-L., Li, Z., Sheng, Y., Zhang, Z., Wu, J., Wang, Y., Zhuang, Y., Zhou, Y., Zheng, X., & Stoica, I. (2023) Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90% ChatGPT Quality. arXiv preprint, arXiv: 2303.18223.
[19] Liu, C., Xiong, T., Chen, Y., Chen, R., Wu, Y., Guo, J., Zhou, T., & Huang, H. (2024) Modality-Balancing Preference Optimization of Large Multimodal Models by Adversarial Negative Mining. arXiv preprint, arXiv: 2402.12020.
[20] Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., & Shou, M. Z. (2024) Hallucination of Multimodal Large Language Models: A Survey. arXiv preprint, arXiv: 2404.18930.
[21] Huang, W., Liu, H., Guo, M., & Gong, N. Z. (2023) Visual Hallucinations of Multi-modal Large Language Models. arXiv preprint, arXiv: 2309.07819.