A Review on VQA: Methods, Tools and Datasets

Mayank Agrawal
Department of Computer Engineering and Applications,
GLA University, Mathura, India
mayank.agrawal@gla.ac.in

Anand Singh Jalal
Department of Computer Engineering and Applications,
GLA University, Mathura, India
asjalal@gla.ac.in

Himanshu Sharma
Department of Computer Engineering and Applications,
GLA University, Mathura, India
himanshu.sharma@gla.ac.in

2023 International Conference on Computer Science and Emerging Technologies (CSET) | 979-8-3503-4173-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/CSET58993.2023.10346815

***Abstract*— An new area called "visual question answering" (VQA) seeks to integrate CV with NLP. In order to get correct results, it entails creating models that can comprehend both textual questions and visual input, represented by videos or images. Applications for VQA systems include content-based image retrieval, medical image analysis, autonomous vehicles, and human-computer interaction. However, there are a number of difficulties in developing efficient VQA models, such as ambiguity in questions, sophisticated reasoning, processing multi-modal data, and data bias. To solve these issues and enhance the functionality and interpretability of VQA models, researchers are consistently investigating novel techniques, such as attention mechanisms, fusion tactics, and transformer-based architectures. In addition to a review of the existing problems and potential future developments in the area of visual question answering, this paper offers an overview of VQA methods, datasets, and tools.Finally, we go through potential routes for VQA and image comprehension research in the future.**

Keywords— VQA, Multimodal Learning, Visual Attention

I. INTRODUCTION

The difficult and multidisciplinary job of VQA combines CV and NLP. It tries to create models that can understand and respond to queries about visual content, such as images or videos. With the help of the VQA paradigm, robots will be able to comprehend visual data and provide meaningful replies by bridging the gap between visual perception and verbal comprehension. In order to provide reliable responses, the VQA job requires processing both the visual input and the text-based questions. To generate the right answers, the model must recognize the subtleties and semantics of the questions, as well as the content and context of the visual image. The classification of images [1, 2], object identification [3, 4], and activity recognition [5, 6, 7] are only a few of the CV tasks that have greatly benefited from recent advances in DL and CV research. Deep CNNs can do image classification [2] on par with humans when given enough data. Visual encoders and question encoders are the two fundamental parts of VQA models, respectively. In order to extract useful visual information, the visual encoder examines the visual input (such as image or video frames). CNNs that have already been trained on complex image classification [2] tasks are often used to do this. On the other hand, the question encoder transforms the textual questions into a meaningful representation by using strategies like Recurrent Neural Networks (RNNs) or Transformers. After being collected, the visual and question features are integrated and put into a fusion module. The model can successfully represent their interactions by fusing visual and textual data by to the fusion module. The two modalities may be combined using a variety of fusion techniques, including concatenation, element-wise multiplication, or attention methods. For predicting the answer to the question, the fused representation is then run through a classification layer. To create a probability distribution across a predetermined set of answer possibilities, the classification layer may be a fully connected layer followed by a softmax function.

II. VQA METHODS

Various techniques and strategies are used in Visual Question Answering (VQA) models. Here are a few methods that are often used:

A. Fusion-based Methods

Fusion-based approaches [8] try to integrate textual andvisual characteristics to provide answers.

There are several fusion techniques, including:

- Late fusion is the merging of visual and textual elements at a late stage of the model, usually by concatenating them or via element-wise processes.
- In early fusion, visual and textual information are combined at an early stage of the model, often by feeding them into a common network architecture.

B. Multimodal Learning

Approaches to multimodal learning [9] seek to teach combined representations of textual and visual elements. Typical strategies include:

- Multimodal Embeddings: Visual and textual information are projected into a single embedding space in multimodal embeddings, where their semantic similarity is maximized.

979-8-3503-0082-6/23/$31.00 ©2023 IEEE

Authorized licensed use limited to: Cambridge Institute of Technology - Bengaluru. Downloaded on March 19,2025 at 05:52:40 UTC from IEEE Xplore. Restrictions apply.

- Multimodal Attention: The most relevant areas of the visual input and the words in the question are alignedand focused on using attention processes.
- Multimodal Transformers: Transformer designs have been modified for VQA tasks after being developed for natural language processing.

C. *Memory Networks*

Memory networks [10] use external memory to store and access pertinent data.

D. *Reinforcement Learning*

The creation of responses in VQA models may be improvedby using reinforcement learning approaches [11].

E. *Ensemble Methods*

Ensemble methods [12] involve combining the predictionsof multiple VQA models.

F. *Visual Attention*

The model is able to concentrate on certain areas or objectsin the visual input thanks to visual attention techniques [13]. Table I., which is explained below, compares the specificsof several VQA methods.

TABLE I. Comparison of various VQA methods

Method	**Advantages**	**Limitations**
Fusion-based Methods [8]	• Capture interactions between visual and textual information.	• Inability to accurately depict complicated interactions between textual and visual modalities.
Multimodal Learning [9]	• Identify the semantic relationships between text and image. .	• Dependency on the quality and availability of aligned visual and textual data.
Memory Networks [10]	• Enable explicit memory access and reasoning capabilities. • Suitable for complex reasoning tasks.	• Complexity in designing and training memory mechanisms. • Computational overhead due to memory access.
Reinforcement Learning [11]	• Effective for optimizing task-specific rewards. • Enable iterative refinement of predictions.	• Requires additional training and may be sensitive to reward design. • Can be computationally expensive.
Ensemble Methods [12]	• Improved performance through model averaging or combination. • Reduced overfitting and increased robustness.	• Increased model complexity and resource requirements. • Additional computational overhead during inference.
Visual Attention [13]	• Capture fine-grained information by attending to relevant regions. • Interpretability through visual attention visualization.	• Difficulty in modeling long-range dependencies. • Sensitivity to noise or inaccuracies in visual attention mechanisms.

III. TOOLS

Using CV and NLP approaches together is called VQA. Here are a few VQA tools and libraries that are often used:

A. *PyTorch*

A well-liked DL structure called PyTorch [14] offers a versatile and effective platform for creating VQA models.

B. *TensorFlow*

Another popular deep learning framework, TensorFlow [15], provides a complete set of VQA tools.

C. *Keras*

On top of TensorFlow, Keras [16] is a user-friendly deep learning

D. *NLTK (Natural Language Toolkit)*

A Python package created expressly for tasks involving natural language processing is called NLTK [17]. It offers capabilities for text processing tasks like as tokenization, stemming, part-of-speech tagging, and others.

E. *Hugging Face Transformers*

Hugging Face Transformers [18] is a library that provides cutting-edge pre-trained models for applications involving natural language comprehension.

F. *TorchVision*

The PyTorch ecosystem's TorchVision [19] module offers pre-trained models, datasets, and image modifications especially for computer vision workloads.

The comparison information for several VQA tools is included in Table II, which is detailed below.

TABLE II. Comparison of various VQA tools

Tool	**Description**	**Advantages**	**Limitations**
PyTorch [14]	A deep learning	• A user-friendly	• Steeper learning curve

Authorized licensed use limited to: Cambridge Institute of Technology - Bengaluru. Downloaded on March 19,2025 at 05:52:40 UTC from IEEE Xplore. Restrictions apply.

	framework that offers a productive and adaptable foundation for creating VQA models.	interface written in Python. •. Graph of dynamic computation	than with other frameworks. • Need knowledge of deep learning principles and Python.
TensorFlow [15]	A well-liked deep learning framework with many features and tools for creating VQA models.	Wide variety of pre-built models and tools, scalable and effective computing, and suitability for production deployment.	• For novices, there may be a steeper learning curve. • Requires a working knowledge of Python and deep learning principles.
Keras [16]	An easy-to-use interface is provided by a user-friendly deep learning framework that operates on top of TensorFlow and allows for the creation and training of VQA models.	• A user-friendly and simple API. • Models with training are available.	• Limited adaptability in comparison to simpler frameworks like PyTorch or TensorFlow. • May not work with sophisticated modifications.
NLTK (Natural Language Toolkit) [17]	A Python package that offers capabilities for text processing, tokenization, stemming, and other activities targeted towards natural language processing.	• A vast selection of text-processing tools. • Multiple languages are supported. • Solid academic support.	• Mainly focused on activities related to natural language processing; not specifically designed for VQA-specific processes. • Limited support for computer vision
Hugging Face Transformers [18]	A collection of modern pre-trained models for NLU tasks, such as BERT, GPT, and RoBERTa, which may be used in VQA models for text encoding and semantic interprettation.	• Easy access to potent language models that have already been trained. • A big set of utilities and models. • A sizable and vibrant community.	• Mainly focused on challenges involving language comprehension; not specifically designed for computer vision or VQA-specific activities. • Limited support for computer vision.
TorchVision [19]	A component of the PyTorch ecosystem that offers pre-trained models, datasets, and image modifycations especially for computer vision applications, helpful for the extraction of visual features for VQA models.	• Seamless PyTorch integration. • A variety of vision models that have been trained. • Effective image conversions.	• Mainly focused on computer vision tasks; not specifically designed for operations including natural language processing or VQA. • Limited support for NLP.

IV. DATASETS

Datasets for Visual Question Answering (VQA) are necessary for developing and testing VQA models. They are made up of images that are linked with text questions and the answers to those questions. These datasets make it possible to design and test VQA models, verifying their effectiveness and generalizability across a range of visual and textual inputs. Here are a few well-known VQA datasets:

A. VQA

One of the first and most popular VQA datasets [20] is the original VQA dataset. It includes more than 200,000 images from the MS COCO dataset, each of which is linked to three open-ended questions and 10 potential solutions.

Authorized licensed use limited to: Cambridge Institute of Technology - Bengaluru. Downloaded on March 19,2025 at 05:52:40 UTC from IEEE Xplore. Restrictions apply.

B. VQA 2.0

In order to overcome some of the biases and restrictions in the first release, VQA 2.0 [21] is an expanded version of the VQA dataset.

C. Visual Genome

More than 100,000 images with detailed annotations, including object instances, properties, and connections between objects, may be found in the Visual Genome [22] collection.

D. CLEVR

To assess compositional reasoning skills in VQA models, researchers created the CLEVR dataset [23].

E. GQA

An addition to the Visual Genome dataset that seeks to solve some of its shortcomings is the GQA (Visual Genome Question Answering) dataset [24].

F. OK-VQA

The object-based reasoning is the main emphasis of the OK-VQA (Object Knowledge for Visual Question Answering) dataset [25].

On the basis of their characteristics, sizes and question types, Table III compares several Visual Question Answering (VQA) datasets:

TABLE III. Comparison of various VQA datasets

Dataset	**Characteristics**	**Size**	**Question Types**
VQA [20]	One of the first and most well-known VQA datasets.	204K images	Open-ended questions, Yes/No questions
VQA 2.0 [21]	A more comprehensive variation of the first VQA dataset.	204K images	Open-ended questions, Yes/No questions
Visual Genome [22]	Has thorough annotations for sophisticated reasoning.	108K images	Open-ended questions
CLEVR [23]	Intended for evaluating compositional reasoning.	700K images	Compositional questions
GQA [24]	An expansion of Visual Genome that includes a variety of questions.	113K images	Open-ended questions, Yes/No questions
OK-VQA [25]	Emphasizes object-based reasoning.	150K images	Object-based questions

V. APPLICATIONS

Visual Question Answering (VQA) has many uses in many different domains. Some of the main uses of VQA include:

A. Human-Computer Interaction

By allowing people to ask computers questions in normal language regarding visual content, VQA may improve human-computer interaction. It has uses in chat-bots, voice-activated systems, and virtual assistants.

B. Content-based Image Retrieval

Using text-based searches, VQA may be used to find videos or images. Users may use natural language to describe what they are seeking for while searching for particular visual material.

C. Educational Tools

VQA may be used into teaching resources to help students comprehend visual content. It enables people to query diagrams or images and get the right answers.

D. Visual Assistance in Navigation

VQA may help navigation systems by deciphering visual cues from the environment. Users may ask questions about their surroundings, and the system can respond with pertinent advice.

E. Social Media Applications

By enabling users to comment on or ask questions about other people's published videos or images, VQA may improve social media networks.

VI. LITERATURE SURVEY

This phase will describe many research publications on the VQA framework.

By creating counterfactual images, the authors of this study [26] present an interpretability strategy for VQA models. The created image is intended to deviate from the original image as little as possible, which causes the VQA model to provide a different result. Additionally, their method guarantees the realism of the image that is created.

The authors of this study [27] have put out a comprehensive, end-to-end retriever-reader architecture for knowledge-based VQA. Their study differs from previous research in two ways. They thoroughly investigate the advantages of multi-modal implicit knowledge derived from trained vision-language models first. The second step is the deliberate construction of a new technique to handle the error propagation issue resulting from the explicit knowledge use.

Authorized licensed use limited to: Cambridge Institute of Technology - Bengaluru. Downloaded on March 19,2025 at 05:52:40 UTC from IEEE Xplore. Restrictions apply.

The questioner, the oracle, and the answerer are the three parts of the conversation-based VQA (Co-VQA) architecture that the authors of this study [28] have developed. Using an extended HRED model, the questioner poses the sub-questions, and Oracle responds to each one individually. The question-answer pair is used to update the visual representation progressively in ACVRM for Answerer.

The issue of VQA in low labelled data regime, which has received little attention in the literature, is discussed in this study [29] by the authors. To add appropriate inductive biases to the VQA model, they use a data augmentation strategy to extend the original little labelled data. According to their findings, accuracy has increased by up to 34% compared to baselines trained simply on the first labelled data.

A The authors of this publication [30] have presented a knowledge description framework (KDF) that unifies their knowledge-based VQA (Kb-VQA) research findings, including KHT, the knowledge pyramid, the knowledge theory model, and KIA, and identifies the direction of knowledge flow. Furthermore, they make the point that more in-depth information may help Kb-VQA to enhance its outcomes on challenging and unresolved problems.

The authors of this study [31] have suggested MedVInT, a generative model designed to progress this important medical endeavour. By matching visual information from a previously learned vision encoder with language models, MedVInT is taught. They also provide a scalable method for building PMC-VQA, a substantial MedVQA dataset that includes 227k VQA pairings over 149k pictures and spans a variety of modalities and disorders.

Through their devised mechanism to correlate the answering embeddings with the fused image-question features, the authors of this paper [32] have presented a semi-open framework for medical VQA that successfully incorporates answer semantic information into the answer class prediction process, which significantly improves accuracy.

Prophet is a conceptually simple framework developed by the authors of this research [33] to provide the GPT-3 response heuristics for knowledge-based VQA. To be more specific, they first train a VQA model that has never been trained on a specific knowledge-based VQA dataset. Theynext remove the model's two distinct categories of complementing response heuristics—answer candidates andanswer-aware instances.

In this study [34], two network designs—LSTM and TD—have been looked at potential explanation generators. Their technique generates textual explanations that are accessible by humans while maintaining SOTA VQA accuracy on the GQA-REX (77.59%) and VQA-E (71.48%) datasets.

The trade-off between VQA accuracy and the grounding capabilities of the presently used SOTA transformer-based approaches is shown by the authors in this study [35]. They suggest using transformer encoder layers in conjunction with text-guided capsule representation.

VII. Issues and challenges

To create reliable and accurate VQA systems, researchers and developers must address a number of issues and challenges that get related to visual question answering (VQA). Among the main challenges are:

A. Ambiguity and Variability in Questions

In VQA, questions may be very complex and ambiguous, which can result in several legitimate answers. For an answer to be truthful, it is essential to comprehend the many ways that a question may be posed and to address each one.

B. Complex Reasoning and Understanding

Complex reasoning skills including spatial thinking, counting, comparing, and logical processes are often needed for VQA. Creating models that are capable of such reasoning is a difficult task.

C. Handling Multi-Modal Data

Processing visual and textual data is a part of VQA. It is difficult to successfully combine different modalities and comprehend how they interact, particularly when one modality is more helpful than the other.

D. Data Bias and Imbalance

When specific question types or answer distributions are overrepresented in VQA datasets, it might cause differences in model performance across various question types.

E. Lack of Explicit Visual References

Certain questions may not make clear use of certain visual components, making it difficult for models to comprehend the situation and pinpoint the pertinent areas to address.

VIII. Conclusion

We provide a thorough review of the most recent techniques, datasets, tools, and problems in this topic, which is quickly developing, in our research. The survey emphasized the value of VQA in bridging the gap between NLP and computer vision, allowing computers to comprehend and respond to questions concerning visual content. Throughout the survey, we have discussed a range of VQA methodologies, including fusion-based approaches, multimodal learning, memory networks, and attention processes. These methods show the range of tactics that may be used to effectively blend visual and textual information. Several well-known VQA datasets, each of which tackles a specific assessment issue, are examined, including VQA, VQA 2.0, Visual Genome, and CLEVR. High-quality and diverse datasets are crucial for both training and evaluating VQA models. Additionally, we go through the challenges and issues that VQA faces, such as confusing questions, advanced reasoning, handling multi-modal data, and dataset biases. These problems need to be addressed in order to provide more accurate, reliable, and intelligible VQA systems. In order to aid in the development and testing of VQA models, we have also discussed a variety of tools and libraries, including Hugging Face Transformers, PyTorch, TensorFlow, and others.

Authorized licensed use limited to: Cambridge Institute of Technology - Bengaluru. Downloaded on March 19,2025 at 05:52:40 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[2] 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).

[3] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

[4] Singh, Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28.

[5] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2625-2634).

[6] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 1725-1732).

[7] Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems, 27.

[8] Zhang, D., Cao, R., & Wu, S. (2019). Information fusion in visual question answering: A survey. Information Fusion, 52, 268-280.

[9] Ilievski, I., & Feng, J. (2017). Multimodal learning and reasoning for visual question answering. Advances in neural information processing systems, 30.

[10] Su, Z., Zhu, C., Dong, Y., Cai, D., Chen, Y., & Li, J. (2018). Learning visual knowledge memory networks for visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7736-7745).

[11] Fountoukidou, T., & Sznitman, R. (2023). A reinforcement learning approach for VQA validation: An application to diabetic macular edema grading. Medical image analysis, 87, 102822.

[12] Han, X., Wang, S., Su, C., Huang, Q., & Tian, Q. (2021). Greedy gradient ensemble for robust visual question answering. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1584-1593).

[13] S. M. G, S. R. D, M. T. R, T. Rajan, V. K and S. H. K, "Intelligent Systems for Medical Diagnostics with the Detection of Diabetic Retinopathy at Reduced Entropy," 2023 International Conference on Network, Multimedia and Information Technology (NMITCON), Bengaluru, India, 2023, pp. 1-8, doi: 10.1109/NMITCON58196.2023.10276174.

[14] Sharma, H., & Jalal, A. S. (2022). Improving visual question answering by combining scene-text information. Multimedia Tools and Applications, 81(9), 12177-12208.

[15] Brugués i Pujolràs, J., Gómez i Bigordà, L., & Karatzas, D. (2022, May). A Multilingual Approach to Scene Text Visual Question Answering. In International Workshop on Document Analysis Systems (pp. 65-79). Cham: Springer International Publishing.

[16] Bansal, M., Gadgil, T., Shah, R., & Verma, P. (2019). Medical Visual Question Answering at Image CLEF 2019-VQA Med. In CLEF (Working Notes).

[17] Yusuf, A. A., Chong, F., & Xianling, M. (2022). Evaluation of graph convolutional networks performance for visual question answering on reasoning datasets. Multimedia Tools and Applications, 81(28), 40361-40370.

[18] Biten, A. F., Litman, R., Xie, Y., Appalaraju, S., & Manmatha, R. LaTr: Layout-Aware Transformer for Scene-Text VQA Supplementary Material.

[19] Lin, J., Ren, X., Zhang, Y., Liu, G., Wang, P., Yang, A., & Zhou, C. (2022). Transferring General Multimodal Pretrained Models to Text Recognition. arXiv preprint arXiv:2212.09297.

[20] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision (pp. 2425-2433).

[21] Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2017). Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6904-6913).

[22] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., ... & Fei-Fei, L. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision, 123, 32-73.

[23] Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., & Girshick, R. (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2901-2910).

[24] Hudson, D. A., & Manning, C. D. (2019). Gqa: A new dataset for real-world visual reasoning and compositional question answering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 6700-6709).

[25] Marino, K., Rastegari, M., Farhadi, A., & Mottaghi, R. (2019). Ok-vqa: A visual question answering benchmark requiring external knowledge. In Proceedings of the IEEE/cvf conference on computer vision and pattern recognition (pp. 3195-3204).

[26] Boukhers, Z., Hartmann, T., & Jürjens, J. (2022). Coin: Counterfactual image generation for vqa interpretation. arXiv preprint arXiv:2201.03342.

[27] Guo, Y., Nie, L., Wong, Y., Liu, Y., Cheng, Z., & Kankanhalli, M. (2022, October). A unified end-to-end retriever-reader framework for knowledge-based VQA. In Proceedings of the 30th ACM International Conference on Multimedia (pp. 2061-2069).

[28] Wang, R., Qian, Y., Feng, F., Wang, X., & Jiang, H. (2022). Co-VQA: Answering by interactive sub question sequence. arXiv preprint arXiv:2204.00879.

[29] Askarian, N., Abbasnejad, E., Zukerman, I., Buntine, W., & Haffari, G. (2022). Inductive biases for low data vqa: a data augmentation approach. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 231-240).

[30] S. Madhuri G, K. Chokkanathan, M. T R, M. M. Musthafa, V. K and V. V, "MLPDR: High Performance ML Algorithms for the Prediction of Diabetes Retinopathy," 2023 International Conference on Network, Multimedia and Information Technology (NMITCON), Bengaluru, India, 2023, pp. 1-7, doi: 10.1109/NMITCON58196.2023.10276078.

[31] Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., & Xie, W. (2023). Pmc-vqa: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415.

[32] Liu, Y., Wang, Z., Xu, D., & Zhou, L. (2023, June). Q2atransformer: Improving medical vqa via an answer querying decoder. In International Conference on Information Processing in Medical Imaging (pp. 445-456). Cham: Springer Nature Switzerland.

[33] Shao, Z., Yu, Z., Wang, M., & Yu, J. (2023). Prompting large language models with answer heuristics for knowledge-based visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14974-14983).

[34] Vaideeswaran, R., Gao, F., Mathur, A., & Thattai, G. (2022). Towards reasoning-aware explainable vqa. arXiv preprint arXiv:2211.05190.

[35] Khan, A. U., Kuehne, H., Gan, C., Lobo, N. D. V., & Shah, M. (2022, October). Weakly Supervised Grounding for VQA in Vision-Language Transformers. In European Conference on Computer Vision (pp. 652-670). Cham: Springer Nature Switzerland.

Authorized licensed use limited to: Cambridge Institute of Technology - Bengaluru. Downloaded on March 19,2025 at 05:52:40 UTC from IEEE Xplore. Restrictions apply.