iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.80540

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

ECM²RS: Explainable Causal Multi-Modal Reasoning System

Muppidi Siva Narayana[1], Chataparthi Siva Shankar[2], Kola Leela Adithya Krishna[3], Ravva Teja Sri Sai Venkata Varma[4], Kankipati Nagamani[5], Marumudi Prasanna Latha[6]
Department of Artificial Intelligence & Data Science, West Godavari Institute of Science and Engineering, Prakasaraopalem, Avapadu Andhra Pradesh, India

Abstract: ***Recent advancements in Artificial Intelligence have enabled the development of multimodal systems capable of reasoning over both visual and textual data. Visual Question Answering (VQA) is a key application in this domain; however, most existing models operate as black-box systems, lacking transparency and interpretability. Additionally, these systems often suffer from language bias, leading to unreliable and non-generalizable predictions. To address these limitations, this paper proposes ECM²RS (Explainable Causal Multi-Modal Reasoning System), a novel framework that integrates multimodal deep learning with neuro-symbolic reasoning and explainability techniques. The system leverages LLaVA as the core reasoning engine and incorporates multi-level explanation modules, including visual explanations using gradient-based methods, textual explanations via attention mechanisms, and knowledge-based reasoning from external datasets. The proposed approach is evaluated using VQA, CLEVR, and ScienceQA datasets to ensure both real-world applicability and logical reasoning capability. Experimental results demonstrate that ECM²RS enhances interpretability while reducing black-box behaviour, producing coherent and explainable reasoning outputs. This work contributes toward building trustworthy and interpretable multimodal AI systems.***

Keywords: ***Multimodal AI, Visual Question Answering (VQA), Explainable Artificial Intelligence (XAI), Causal Reasoning, Neuro-Symbolic Learning, Deep Learning.***

I. INTRODUCTION

Recent advancements in Artificial Intelligence (AI) have led to the development of multimodal systems capable of processing and understanding diverse data types such as images and text. Among these, Visual Question Answering (VQA) has emerged as a significant research area, requiring models to interpret visual content and generate context-aware responses to natural language queries. The integration of computer vision and natural language processing has enabled these systems to achieve impressive performance across various applications, including education, healthcare, and assistive technologies.

Despite these advancements, most existing VQA models operate as black-box systems, providing accurate predictions without offering insights into their reasoning processes. This lack of transparency limits their adoption in critical domains where interpretability and trust are essential. Furthermore, these models often exhibit language bias, relying on patterns in textual input rather than genuinely understanding visual content, which leads to unreliable and non-generalizable outputs.

To address these challenges, this paper proposes ECM²RS (Explainable Causal Multi-Modal Reasoning System), a novel framework that integrates multimodal deep learning with explainability and causal reasoning principles. The system leverages a vision-language model for joint reasoning over image and text inputs, while incorporating multiple explanation mechanisms, including visual, textual, and knowledge-based reasoning. The primary objective of this work is to enhance the interpretability and reliability of multimodal AI systems by generating transparent, step-by-step explanations alongside predictions. By combining neural networks with symbolic reasoning and causal inference concepts, the proposed system aims to reduce bias and improve trustworthiness. This approach contributes toward the development of more interpretable and human-aligned AI systems.

II. RELATED WORK

Recent advancements in multimodal artificial intelligence have led to the development of vision-language models such as LLaVA, which combine image and text understanding to generate context-aware responses. While these models achieve strong performance, they often operate as black-box systems without providing interpretable reasoning. To improve transparency, explainable AI techniques such as Grad-CAM and attention mechanisms have been introduced to highlight important visual regions and textual features.

However, these methods typically provide only partial explanations and do not capture complete reasoning. In addition, causal reasoning approaches have been proposed to reduce language bias in Visual Question Answering systems, improving robustness and generalization. Neuro-symbolic methods further enhance interpretability by integrating neural networks with structured reasoning. Despite these efforts, existing approaches lack a unified framework that combines multimodal reasoning, explainability, and causal inference. The proposed ECM²RS system addresses this gap by integrating these components into a single interpretable framework.

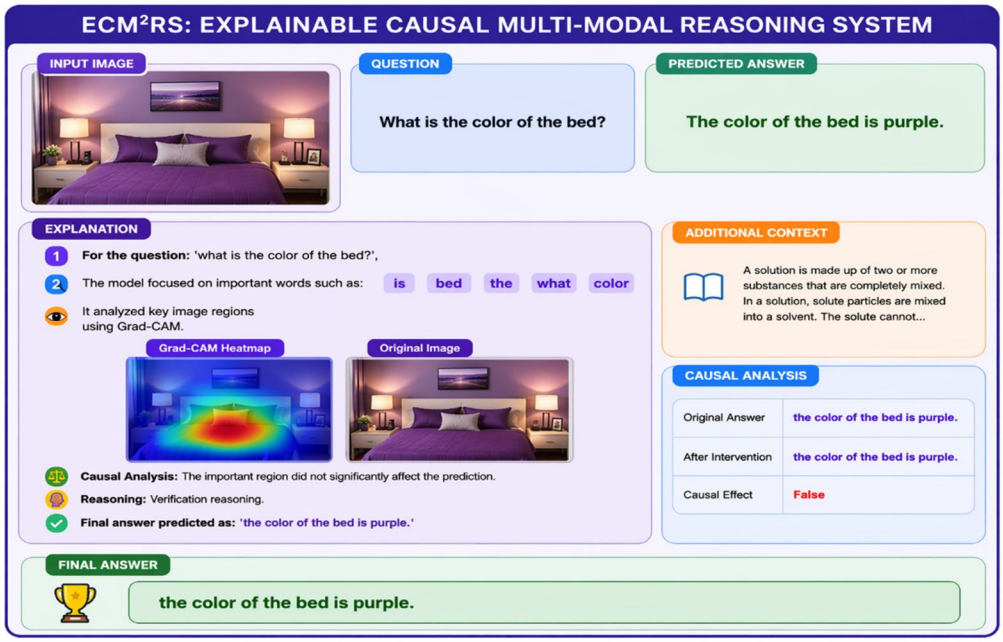


Fig. 1. Overview of VQA Limitations and Motivation for ECM²RS System

III. METHODOLOGY

A. *System Architecture*

The proposed ECM²RS (Explainable Causal Multi-Modal Reasoning System) is a multimodal framework designed to perform reasoning over image and text inputs while generating interpretable outputs. The system follows a structured pipeline consisting of feature extraction, multimodal fusion, reasoning, and explanation generation. It integrates deep learning models with explainability and causal reasoning mechanisms to improve transparency and reliability.

B. *Feature Extraction*

The system processes visual and textual inputs independently. Visual features are extracted from the input image using a pretrained convolutional neural network (ResNet50), which captures spatial and semantic information. Textual features are obtained using a transformer-based model (BERT), which encodes contextual relationships between words in the input question.

$$V = \varphi(I)$$
$$T = \psi(Q)$$

where (I) denotes the input image, (Q) denotes the input question, φ(I) denotes the visual encoding function, and ψ(Q) denotes the textual encoding function.

C. *Multimodal Fusion*

The extracted visual and textual features are combined to form a unified representation for joint reasoning. This fusion enables the system to capture relationships between visual content and the input query.

$$F = W_v V + W_t T$$

where W_v and W_t denote learnable weight matrices, and (F) represents the fused multimodal representation. The representation (F) is then input to a vision-language model (LLaVA), which performs joint reasoning over visual and textual modalities to generate the final answer.

$$P(A|I,Q) = f(F)$$

where A is the predicted answer and f represents the reasoning function.

D. Final Loss Function

The proposed ECM²RS system is optimized using a composite loss function that integrates prediction accuracy, explainability alignment, and causal reasoning.

$$L_{total} = L_{pred} + \lambda_1 L_{attn} + \lambda_2 L_{cam} + \lambda_3 L_{causal}$$

where $(L_{\{pred\}})$ is the cross-entropy loss for answer prediction, $(L_{\{attn\}})$ represents the attention-based loss for textual explanation, $(L_{\{cam\}})$ corresponds to the Grad-CAM loss for visual explanation, and $(L_{\{causal\}})$ enforces causal consistency to reduce language bias. The coefficients $(\lambda_1, \lambda_2, \lambda_3)$ control the contribution of each component.

This formulation ensures that the model not only produces accurate predictions but also generates interpretable and causally consistent explanations.

IV. EXPLAINABLE REASONING FRAMEWORK

The ECM²RS framework incorporates multiple explainability techniques to provide transparent and interpretable predictions. Instead of generating only answers, the system produces reasoning insights by combining visual, textual, and knowledge-based explanations.

A. Visual Explanation (Grad-CAM)

Visual explanations are generated using Gradient-weighted Class Activation Mapping (Grad-CAM), which identifies important regions in the input image that contribute to the model's prediction. It produces a heatmap highlighting the most relevant areas.

$$L = ReLU(\Sigma_k \alpha_k A^k)$$

Where (A^{k}) denotes the feature map of the $(k) - th$ channel, and α_k represents the corresponding importance weight computed from the gradients.

B. Attention-Based Explanation

The attention mechanism is used to determine the importance of each word in the input question. It assigns weights to tokens based on their relevance to the final prediction, enabling the model to focus on key textual information.

$$e_i = f(h_i)$$

$$\alpha_i = \frac{exp(e_i)}{\Sigma_j exp(e_j)}$$

where h_i denotes the hidden representation of the (i)-th token, e_i represents the attention score, and α_i denotes the normalized attention weight obtained after softmax normalization. These weights indicate the relative importance of each token in the reasoning process.

C. Knowledge Integration

To enhance reasoning, the system incorporates external knowledge from datasets such as ScienceQA. This provides structured explanations that complement visual and textual insights, enabling the system to handle complex queries more effectively.

D. Explanation Fusion

The final explanation is obtained by combining visual, textual, and knowledge-based components. This fusion results in a comprehensive and human-understandable reasoning process, improving transparency and trust in the system.

V. CAUSAL REASONING MODULE

A. Problem of Language Bias

Traditional Visual Question Answering (VQA) systems often rely on statistical correlations between the question and the answer, leading to language bias. This causes the model to generate predictions based on textual patterns rather than true visual understanding.

B. Causal Formulation

To address this issue, the proposed system distinguishes between observational and interventional learning:

$$P(Y|X) \neq P(Y|do(X))$$

where $(P(Y|X))$ denotes the observational probability and $(P(Y|do(X))$ denotes the interventional probability under intervention. This formulation enables the model to capture causal relationships instead of spurious correlations.

C. *Bias Mitigation Strategy*

The system reduces language bias by encouraging the model to rely more on visual features during prediction. This improves the robustness and generalization capability of the model across different inputs.

D. *Impact on Model Performance*

By integrating causal reasoning with explainability techniques, the system produces more reliable and interpretable outputs. The model generates predictions that are grounded in visual evidence and supported by meaningful reasoning.

VI. EXPERIMENTAL RESULTS

A. *Experimental Setup*

The proposed ECM²RS system is evaluated on multiple datasets, including VQA, CLEVR, and ScienceQA, to assess its performance on visual understanding and reasoning tasks. The system integrates ResNet50 for image feature extraction, BERT for text encoding, and LLaVA for multimodal reasoning.

B. *Qualitative Results*

The system generates answers along with multi-level explanations, including visual heatmaps, attention-based textual highlights, and knowledge-supported reasoning.

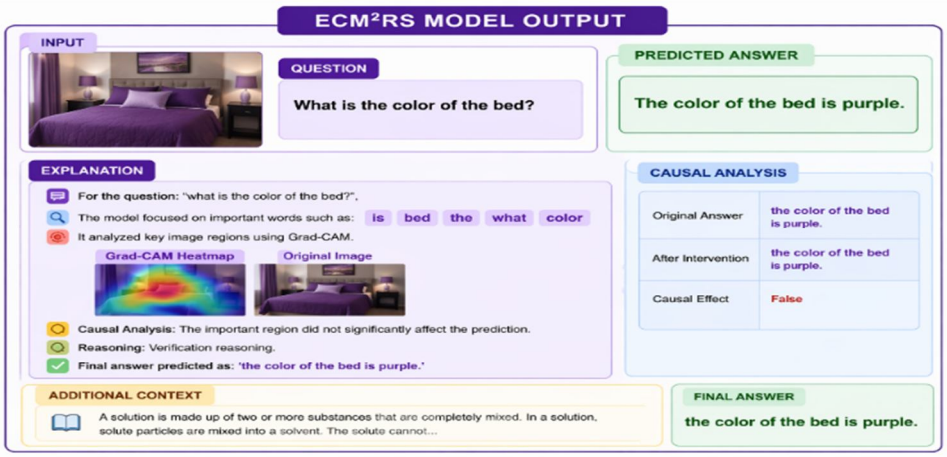


Fig. 2. Sample Output of ECM²RS System

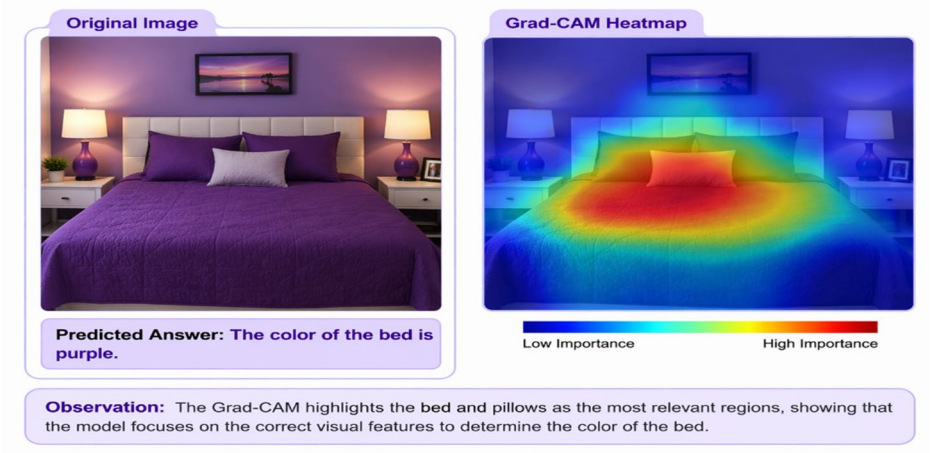


Fig. 3. Grad-CAM Visualization Highlighting Important Regions

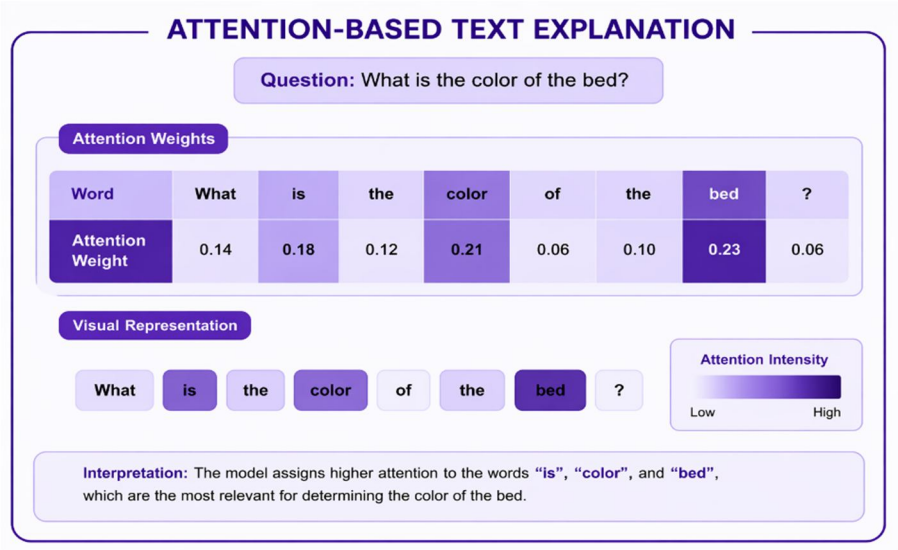


Fig. 4. Attention-Based Text Explanation

C. *Observations*

The results demonstrate that the proposed system provides both accurate predictions and meaningful explanations. The integration of visual, textual, and knowledge-based components enables the model to generate transparent outputs.

The Grad-CAM visualizations confirm that the model relies on relevant image regions, while attention weights highlight important words influencing the decision. Furthermore, the inclusion of causal reasoning reduces language bias and improves the reliability of predictions.

Overall, the ECM²RS system successfully combines multimodal reasoning with explainability, producing interpretable and trustworthy results across different datasets.

VII. DISCUSSION

The proposed ECM²RS system demonstrates significant improvements in interpretability and reasoning compared to traditional Visual Question Answering (VQA) models. By integrating multimodal learning with explainability techniques, the system provides not only accurate predictions but also meaningful insights into the decision-making process.

One of the key strengths of the system is its ability to generate multi-level explanations, including visual, textual, and knowledge-based reasoning. This improves transparency and makes the model more suitable for real-world applications where understanding the reasoning process is essential.

However, the system also has certain limitations. The quality of explanations depends on the performance of underlying models such as Grad-CAM and attention mechanisms. In some complex scenarios, the generated explanations may not fully capture the reasoning process. Additionally, the integration of multiple components increases computational complexity.

Overall, the combination of explainability and causal reasoning enhances the robustness and reliability of the system, making it a promising approach for developing trustworthy multimodal AI systems.

VIII. CONCLUSION

This paper presented ECM²RS (Explainable Causal Multi-Modal Reasoning System), a novel framework designed to perform interpretable reasoning over image and text inputs. The system integrates multimodal deep learning with explainability techniques and causal reasoning to address the limitations of traditional Visual Question Answering (VQA) models.

The proposed approach combines visual feature extraction, textual encoding, and multimodal fusion with advanced explanation methods such as Grad-CAM and attention mechanisms. In addition, the incorporation of causal reasoning helps reduce language bias and improves the reliability of predictions.

Experimental results demonstrate that the system can generate accurate answers along with meaningful visual, textual, and knowledge-based explanations. This enhances transparency and makes the model more suitable for real-world applications. The proposed system can be extended to real-world applications such as healthcare and education.

Overall, the ECM²RS framework contributes toward the development of trustworthy and interpretable multimodal AI systems. Future work may focus on improving explanation quality, optimizing computational efficiency, and extending the system to more complex reasoning tasks.

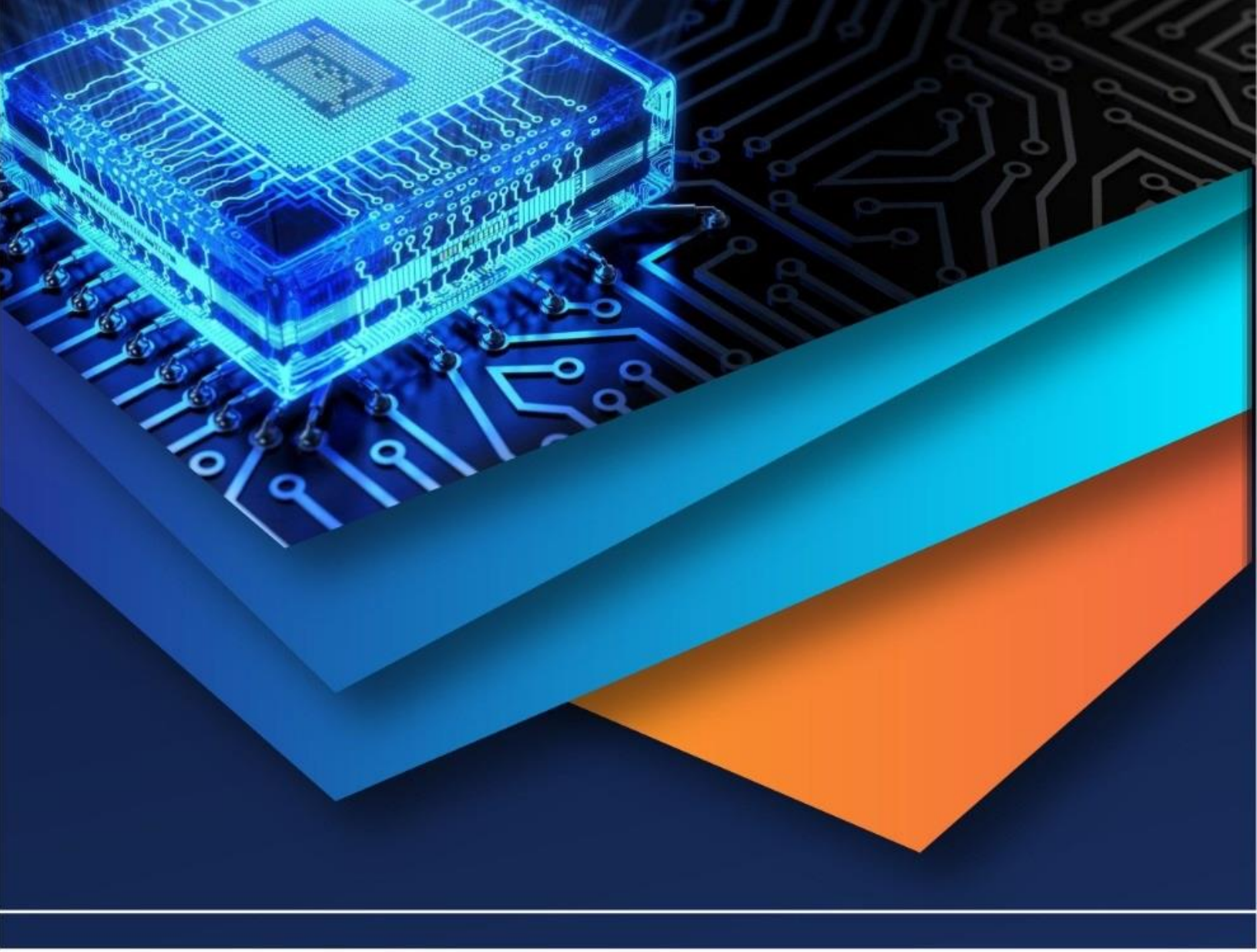
Future work can focus on improving the accuracy and robustness of the proposed system by incorporating more advanced multimodal models and larger datasets. The explainability component can be enhanced through more precise visual and textual interpretation techniques. Additionally, the causal reasoning module can be extended using more rigorous intervention-based approaches to further reduce bias. The system can also be applied to real-world domains such as healthcare, education, and autonomous systems for practical deployment.

IX. ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to N. L. Tejaswini for her valuable guidance and continuous support throughout this work. The authors also thank D. Haritha, Head of the Department, for providing the necessary resources and academic environment.

REFERENCES

[1] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual Instruction Tuning,” Advances in Neural Information Processing Systems (NeurIPS), 2023.
[2] X. Li, Z. Wang, J. Chen, and Y. Zhang, “TV-TREES: Multimodal Entailment Trees for Neuro-Symbolic Video Reasoning,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.
[3] Z. Chen, J. Wang, X. Li, and Z. Wang, “Counterfactual VQA: A Cause-Effect Look at Language Bias,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
[4] S. Antol et al., “VQA: Visual Question Answering,” Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015.
[5] R. R. Selvaraju et al., “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.
[6] A. Vaswani et al., “Attention Is All You Need,” Advances in Neural Information Processing Systems (NeurIPS), 2017.
[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Proceedings of NAACL-HLT, 2019.
[8] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY