Journal of Science Engineering Technology and Management Science
Volume 02, Issue 08, August 2025 *ISSN: 3049-0952*
www.jsetms.com *DOI:10.63590/jsetms.2025.v02.i08.437-443*

A MULTIMODEL GPT BASED APPLICATION TO HELP BLIND USERS VISUAL PICTURE EXTENSIONS

[1] N. JOSHIKA,[2] S. SRIKANTH,[3] S. SURAJ,[4] S. PRADEEP,[5] Mrs. A. NANDINI SREE
[1234] Students, [5] Assistant Professor
Department Of Computer Science and Design
Teegala Krishna Reddy Engineering College, Meerpet, Balapur, Hyderabad-500097

To Cite this Article
N. Joshika, S. Srikanth, S. Suraj, S. Pradeep, Mrs. A. Nandini Sree, "A Multimodel Gpt Based Application To Help Blind Users Visual Picture Extensions", Journal of Science Engineering Technology and Management Science, Vol. 02, Issue 08, August 2025,pp: 437-443, DOI: http://doi.org/10.63590/jsetms.2025.v02.i08.pp437-443
Submitted: 12-07-2025 *Accepted: 18-08-2025* *Published: 25-08-2025*

ABSTRACT
The quick headway in Fake Insights (AI) has opened modern conceivable outcomes for upgrading availability for outwardly impeded clients. This venture presents a multimodal GPT-based application that consistently coordinating Normal Dialect Preparing (NLP) and picture acknowledgment innovations to supply comprehensive talked depictions of visual substance. The framework empowers real-time interaction by translating objects, scenes, and feelings in pictures and changing over them into important stories, in this way bridging the crevice between visual and non-visual mediums. The application offers transformative benefits in different spaces. For outwardly disabled clients, it moves forward every day availability by deciphering complex visuals into expressive sound. In instruction, it helps by clarifying classroom visuals, making learning comprehensive. In e-commerce, it encourages nitty gritty item depictions, improving the shopping involvement. Whereas the framework illustrates solid execution with clear visual substance, challenges continue in precisely deciphering complex scenes, enthusiastic prompts, and vague settings, as well as in accomplishing real-time preparing on low- resource gadgets. This extend emphasizes a adaptable and user-friendly plan, utilizing progressed profound learning models such as Convolutional Neural Systems (CNNs) for picture handling and strong NLP strategies for conversational coherence. The design is outlined to handle multimodal inputs successfully, with future cycles arranged to address current restrictions, counting way better scene translation and compliance with information protection directions. By progressing the capabilities of AI-driven conversational interfacing, this extend lays the foundation for more comprehensive innovation arrangements. The continuous refinement and extension of these frameworks have the potential to revolutionize availability, instruction, and client encounters over assorted divisions.

I. INTRODUCTION

In a world where visual media are part of daily life, accessibility and inclusion become essential to assist visually impaired individuals. This is one of the goals of Visual Question Answering (VQA). VQA is a challenging research area joining advancements in Computer Vision (CV) and Natural Language Processing (NLP). VQA applications entail comparing the semantic information in a visual media against the semantic elements embedded in a question in natural language. This paper proposes a VQA application to support people with different visual abilities in exploring their surrounding environment through information about images framed by their smartphones. The aim is to help to

mentally visualize or create a mental representation of what they cannot or can hardly see.

1.1 MOTIVATION

The Multimodal GPT-based Application to Help Blind Users Visualize Picture Extensions represents a groundbreaking endeavor aimed at leveraging advanced technologies to enhance accessibility and inclusivity. In an era where technology bridges countless barriers, visually impaired individuals often face challenges in interpreting visual content—a domain where this project aspires to make a transformative difference.

By combining Natural Language Processing (NLP) with state-of-the-art image recognition, this system transcends traditional limitations, offering real-time spoken descriptions of images. Its potential to empower visually impaired users extends beyond accessibility, providing tools that could reshape education, e-commerce, and day-to-day experiences.

Documenting this project ensures that its core concepts, methodologies, and achievements are preserved, shared, and built upon. This documentation will serve as a guide for future innovations, fostering a deeper understanding of the system's architecture and its real-world applications. It also stands as a testament to the dedication and creativity of the team behind this impactful solution.

With further research and refinement, the project's aspirations of tackling complex visual interpretations and emotional contexts could revolutionize human-computer interactions, underscoring the importance of this documentation as a stepping stone for the next generation of technological breakthroughs.

1.2 PROBLEM STATEMENT

The failure of current chatbot frameworks to comprehend multimodal inputssuch as content and imagescohesively postures genuine challenges to their comprehension of complicated visuals passionate circumstances and vague demands real-time preparing on low-resource gadgets protecting conversational coherence and giving outwardly impeded clients pertinent portrayals are all made more challenging by these issues the encourage utilize of ai-driven chatbots in down to earth settings such as availability instruction and retail is compelled by these imperatives creating a solid context-aware chatbot that combines cutting-edge picture acknowledgment and characteristic dialect preparing nlp innovation points to address these issues by encouraging smooth characteristic intuitive and improving openness for individuals with visual disabilities.

1.3 SCOPE AND OBJECTIVE

The objective of this venture is to create a multimodal GPT-based application that coordinating progressed Characteristic Dialect Preparing (NLP) and picture acknowledgment innovations to improve availability for outwardly impeded people. The framework points to supply real-time talked portrayals of visual substance, empowering clients to pick up significant experiences into their environment. By tending to challenges such as translating complex visuals, keeping up conversational coherence, and working on low-resource devices, the extend endeavors to make a vigorous, user-friendly, and context- aware chatbot. Eventually, this development looks for to bridge the crevice between visual and non-visual mediums, contributing to inclusivity and empowering broader applications in availability, instruction, and e-commerce.

Objectives:

- Improve Openness:

Create a framework that gives outwardly disabled clients with precise and important talked portrayals of pictures, empowering superior interaction with visual substance.

- Consistent Multimodal Integration:

Combine content and picture preparing capabilities to guarantee cohesive and common intuitive in real-time.

- Move forward Relevant Understanding:

Empower the chatbot to decipher complex scenes, passionate signals, and vague inquiries for

improved client involvement.

- Guarantee Proficiency:

Optimize the framework to function successfully on low-resource gadgets without compromising execution.

- Grow Convenience:

Plan a versatile, measured arrangement that can be adjusted for assorted applications in openness, instruction, retail, and past.

- Promote Information Protection:

Guarantee compliance with information protection directions, such as GDPR, to construct client believe and defend touchy data.

II. LITERATURE SURVEY

Conversational Image Recognition Chatbots

Researchers have combined Natural Language Processing (NLP) and image recognition technologies to create chatbots capable of real-time discussions about visual content. These systems can identify objects, explain scenes, and interpret emotions. While effective for simple visuals, challenges remain in handling complex scenes, maintaining conversational coherence, and ensuring accessibility for visually impaired users. Advanced models like CNNs and Transformer architectures have been employed to address these challenges, though limitations persist in ambiguous queries and real- time performance.

Accessibility for Visually Impaired Users

Chatbot systems leveraging image-to-speech technologies have shown promise in aiding visually impaired individuals by converting images into spoken descriptions. Despite this, these systems often struggle with providing detailed contextual understanding, especially in visually complex or emotionally nuanced scenarios. Enhancements in multimodal inputs integration and context-aware dialogue tracking have been proposed to bridge this gap.

Improving Real-Time Processing on Low-Resource Devices

The dependency of AI systems on high-quality and diverse datasets has been identified as a major barrier to their performance. Researchers have focused on optimizing resource usage by implementing scalable backend architectures and modular updates. This has led to the exploration of lightweight AI models and frameworks tailored for low-resource environments, enabling more accessible and efficient chatbot applications.

Multimodal Integration for Enhanced Coherence

Studies have demonstrated the potential of combining NLP and image recognition technologies to create a cohesive user experience. Techniques like context-aware dialogue tracking improve conversational coherence and enhance user satisfaction. These approaches also facilitate the seamless integration of multimodal inputs, making interactions more intuitive and effective.

Ethical Considerations and Privacy Compliance

Compliance with data privacy regulations, such as GDPR, is a recurring theme in the literature. Ensuring data security and privacy in multimodal systems has been highlighted as a crucial aspect of developing user-friendly and trustworthy chatbot solutions.

Challenges in Complex Visual Interpretation

Addressing the interpretation of complex visual content, including emotional cues, remains a significant focus area. Current research efforts are exploring the use of advanced deep learning models capable of handling intricate visual and contextual scenarios, paving the way for systems with enhanced interpretative abilities.

EXISTING SYSTEM

• Description:

Conversational chatbots are integrated with Natural Language Processing (NLP) and image

recognition to process real-time user inputs, including text and images. They facilitate intuitive user interaction through conversational interfaces.

• Disadvantages:

1. Limited ability to interpret complex visual scenes or emotional expressions accurately.
2. Struggles with maintaining smooth, coherent multi-turn conversations.
3. Inadequate handling of ambiguous or unclear questions.
4. Dependency on high-quality, diverse training data for effective performance.
5. Challenges with real-time processing on low-resource devices.

PROPOSED SYSTEM

An advanced AI-powered chatbot system that integrates deep learning techniques like CNNs for visual data and NLP models for linguistic data, enabling more natural and efficient interaction.

• Advantages:

1. Enhanced ability to interpret complex visual scenes and emotional cues.
2. Seamless integration of multi-modal inputs (text and images) for a cohesive user experience.
3. Improved conversational coherence through context-aware dialogue tracking.
4. Scalable backend architecture supporting modular updates and new features.
5. Ensured data privacy and compliance with regulations like GDPR.

III. MODULE DESCRIPTION

1. Client Interface (UI) Module

- **Portrayal:**

This module is the point of interaction for clients. It permits clients to input information through content, voice, or picture, and shows the yield in an open arrange, either outwardly or through sound. For outwardly impeded clients, the interface incorporates voice acknowledgment for input and text-to-speech (TTS) for yield. The UI is outlined to be straightforward and available, with compatibility for screen perusers and other assistive innovations.

- **Key Capacities:**
  - Acknowledge client input by means of content, voice, or picture.
  - Show comes about or play sound yields.
  - Give a responsive, available interface for outwardly disabled clients.

2. Input Preparing Module

- **Portrayal:**

This module is capable for taking care of distinctive sorts of input given by the client, counting content, voice, and pictures. It oversees voice-to-text change, forms printed questions, and advances pictures to the picture acknowledgment unit.

- **Key Capacities:**
  - Content Input:

Acknowledges client questions in content frame, forms them, and passes them to the NLP show.

  - Voice Input:

Changes over talked words into content utilizing speech-to-text innovation for encourage handling.

  - Picture Input:

Captures pictures utilizing the camera module and plans them for investigation by the picture acknowledgment framework.

3. Characteristic Dialect Preparing (NLP) Module

- **Portrayal:**

The NLP module is capable for understanding and handling the content input from the client. It employments progressed NLP models (e.g., Transformer-based models like GPT) to translate client inquiries, identify aim, and create context-aware reactions. The module empowers the framework to

lock in in significant, coherent discussions with clients.

- **Key Capacities:**
  - Handle and get it client questions.
  - Identify aim and give important reactions.
  - Produce text-based reactions to client inputs.

4. Picture Acknowledgment Module

- **Portrayal:**

This module employments progressed picture acknowledgment procedures (such as Convolutional Neural Systems (CNNs)) to analyze and translate visual substance. When the client gives an picture, this module recognizes objects, scenes, and conceivable passionate prompts.

- **Key Capacities:**
  - Analyze pictures to recognize objects, scenes, or content.
  - Identify visual highlights like passionate prompts or complex scenes.
  - Give a point by point portrayal of the picture to the client, either through content or voice.

5. Text-to-Speech (TTS) Module

- **Description:**

The TTS module is responsible for converting text-based responses (from the NLP and Image Recognition modules) into audible speech. This module plays a crucial role in making the system accessible for visually impaired users, as it enables the application to "speak" responses and descriptions aloud.

- **Key Functions:**
- Convert text into clear, natural-sounding speech.
- Provide real-time spoken descriptions for users.
- Support multiple languages and accents for diverse user needs.

IV. SYSTEM DESIGN

SYSTEM ARCHITECTURE

Fig. System Architecture

V. OUTPUT SCREENS

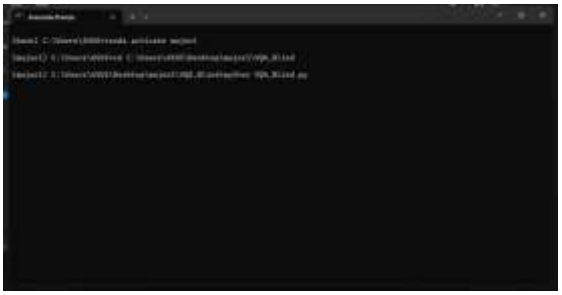

Fig : 1

Fig : 2

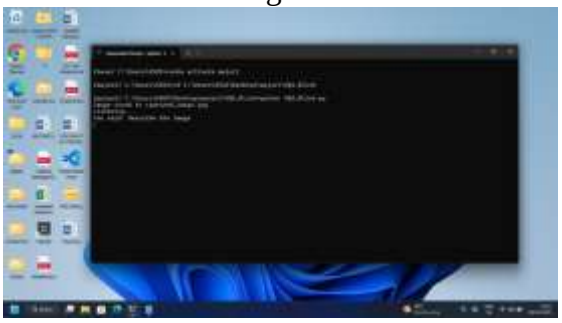

Fig: 3

Fig: 4

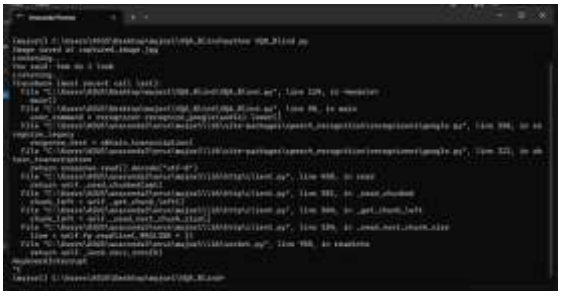

Fig: 5

VI. CONCLUSION

The development of a multimodal GPT-based application for assisting visually impaired users represents a significant stride toward improving accessibility and inclusivity. By integrating advanced Natural Language Processing (NLP) and image recognition technologies, the system bridges the gap between visual and non-visual mediums, providing meaningful and context-aware descriptions of visual content. This innovative solution empowers visually impaired users to better understand their surroundings, enhancing their daily interactions and independence. Beyond accessibility, this application has the potential to transform various industries, including education and e-commerce, by offering dynamic, user-friendly interfaces and insightful visual interpretations. Despite its promising capabilities, current limitations in handling complex scenes, emotional cues, and ambiguous inputs highlight the need for ongoing research and enhancement. Future efforts will focus on overcoming these challenges, enabling the system to achieve greater accuracy, conversational coherence, and efficiency. This project underscores the transformative power of combining multimodal AI

technologies to create intuitive and impactful user experiences. With continued development, such applications can play a crucial role in shaping an inclusive digital future, benefiting not only visually impaired users but also broader audiences across diverse sectors.

REFERENCE

[1] B Kundan, Sangaralingam P. Combining Machine Learning and Deep Learning in the Retinopathy Diagnostic Algorithm for Enhanced Detection of DR and DME. J Neonatal Surg [Internet]. 2025Apr.2 [cited 2025Apr.9];14(5):128-40. Available from: https://www.jneonatalsurg.com/index.php/jns/article/view/2914.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition. 6077–6086.

[3] Silvio Barra, Carmen Bisogni, Maria De Marsico, and Stefano Ricciardi. 2021. Visual question answering: Which investigated applications? Pattern Recognition Letters 151 (2021), 325–331.

[4] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. 2010. Vizwiz: nearly real-time answers to visual questions. In Proceedings of the 23nd annual ACM symposium on User interface software and technology. 333– 342.

[5] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See https://vicuna. lmsys. org (accessed 09 January 2024) (2023).

[6] Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P Bigham. 2019. Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 939–948.

[7] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In Proceedings of the IEEE conference on computer vision and pattern recognition. 3608–3617.

[8] Md Farhan Ishmam, Md Sakib Hossain Shovon, MF Mridha, and Nilanjan Dey. 2024. From image to language: A critical analysis of visual question answering

(vqa) approaches, challenges, and opportunities. Information Fusion (2024), 102270.

[9] Walter S Lasecki, Phyo Thiha, Yu Zhong, Erin Brady, and Jeffrey P Bigham. 2013. Answering visual questions with conversational crowd assistants. In Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility. 1–8.

[10] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023).

[11] S. Manmadhan and B. C. Kovoor. 2020. Visual question answering: a state-of-the- art review. In Artificial Intelligence Review (2020).