Conversational Image Recognition Chatbot Using Vision-Language Models

[1]Dr. P. Harini, [2]M. Triveni, [3]M. Hema Srivalli, [4]N. Sudheer Vali

[1]Professor & Head, Department of Computer Science and Engineering, St. Ann's College of Engineering and Technology, Chirala-523187, Andhra Pradesh, India.

[2,3,4]U. G Students, Department of Computer Science and Engineering, St. Ann's College of Engineering and Technology, Chirala-523187, Andhra Pradesh, India.

ABSTRACT

This project presents a Conversational Image Recognition Chatbot that integrates Artificial Intelligence, Computer Vision, and Vision-Language Models to provide intelligent image analysis through natural language conversations. The system enables users to upload an image and interact with it by asking questions about its contents, such as identifying objects, describing scenes, extracting text, and understanding contextual relationships. It ensures accurate and real-time image understanding by leveraging advanced deep learning techniques and multimodal AI models. Unlike traditional image recognition systems that only classify objects or generate static captions, the proposed system supports interactive, context-aware conversations, allowing users to ask follow-up questions and receive detailed responses. Conventional image recognition applications often lack conversational capabilities, contextual reasoning, and personalised interaction, making them less effective for complex visual understanding tasks. There is a need for an intelligent, user-friendly, and cost-effective image recognition system that combines computer vision with conversational artificial intelligence to deliver accurate, explainable, and human-like responses. The proposed chatbot addresses these limitations by providing an efficient and scalable solution suitable for applications in education, healthcare, tourism, accessibility, customer support, and smart digital assistants.

Keywords: ***Artificial Intelligence (AI), Computer Vision, Image Recognition, Vision-Language Model (VLM), Natural Language Processing (NLP), Deep Learning, Conversational AI, Gemini AI, FastAPI, Next.js.***

INTRODUCTION

Image recognition has become one of the most significant applications of Artificial Intelligence and Computer Vision due to recent advancements in deep learning and Vision-Language Models. Traditional image recognition systems are capable of identifying objects or generating simple image

descriptions but often lack the ability to interact with users through natural conversations.

This project combines advanced image recognition with conversational artificial intelligence to provide an intelligent chatbot capable of analysing uploaded images and answering user queries in real time. The proposed system not only improves image understanding but also enhances user interaction by enabling context-aware conversations, making image analysis more accurate, accessible, and user-friendly.

LITERATURE SURVEY

Conversational image recognition systems have gained significant attention in recent years due to rapid advancements in Artificial Intelligence, Computer Vision, and Natural Language Processing technologies. These systems combine image understanding with conversational capabilities to provide users with intelligent and interactive responses based on visual content.

The core concept of conversational image recognition revolves around analyzing uploaded images using deep learning and Vision-Language Models, enabling accurate object detection, scene understanding, text recognition, and contextual reasoning through natural language interactions.in this many there.

RELATED WORK

Several researchers have contributed to the development of image recognition and visual question answering systems with the aim of improving image analysis, accessibility, and human-computer interaction. Early research focused on image classification and object detection using convolutional neural networks, where systems could identify objects but offered limited user interaction. Recent developments have introduced Vision-Language Models that integrate computer vision with natural language processing, allowing users to ask questions about images and receive context-aware responses. These intelligent systems provide improved accuracy, flexibility, and a more natural conversational experience compared to traditional image recognition approaches.

EXISTING SYSTEM

The existing image recognition systems mainly depend on object classification, image captioning, or keyword-based search to identify visual content. Users can only obtain predefined outputs such as object names or short descriptions, with very limited interaction. These systems are unable to understand complex user queries or maintain conversational context while analyzing images. Most traditional image recognition applications lack intelligent

reasoning, contextual understanding, and real-time conversational capabilities. They cannot effectively answer follow-up questions or explain the relationships between objects within an image. Overall, the existing systems provide limited user interaction, reduced flexibility, and insufficient visual understanding for real-world applications.

PROPOSED SYSTEM

The proposed system implements an AI-driven Conversational Image Recognition Chatbot that combines Computer Vision, Vision-Language Models, and Natural Language Processing to provide intelligent image understanding through natural conversations. It allows users to upload an image and ask questions about its contents in real time. The chatbot identifies objects, describes scenes, extracts text, recognizes visual relationships, and generates context-aware responses based on user queries. Unlike conventional image recognition systems, the proposed solution supports multi-turn conversations, enabling users to ask follow-up questions while maintaining conversational context. By leveraging advanced multimodal AI models, the system provides accurate, interactive, and human-like responses, making image analysis more efficient, accessible, and suitable for various real-world applications.

SYSTEM ARCHITECTURE

The architecture of the Conversational Image Recognition Chatbot consists of a user interface, image upload module, FastAPI backend, Vision-Language Model (Gemini API), image processing module, and response generation module. The user uploads an image through the web interface, which is sent to the backend for processing. The Vision-Language Model analyzes the uploaded image by identifying objects, scenes, text, and visual relationships.The generated response is then displayed through the chat interface, allowing users to continue the conversation with follow-up questions. This architecture ensures intelligent, real-time, and interactive image understanding through conversational AI.

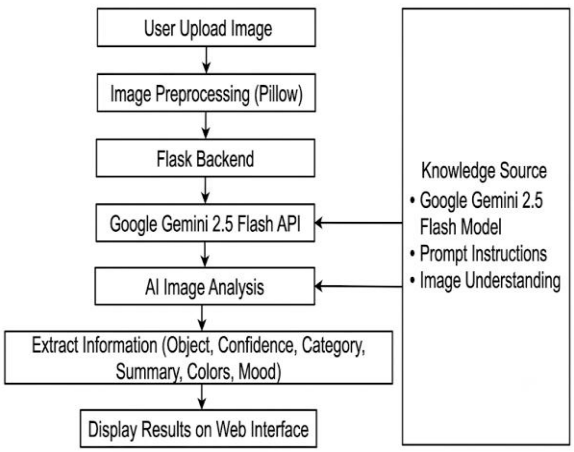

Fig 1: System Architecture

METHODOLOGY DESCRIPTION

The methodology of the Conversational Image Recognition Chatbot is designed to provide intelligent image understanding and natural language interaction using modern Artificial Intelligence and Computer Vision technologies. The system follows a structured approach that includes image acquisition, image preprocessing, feature extraction, multimodal analysis using a Vision-Language Model, natural language response generation, and conversational interaction. This workflow enables the chatbot to analyze uploaded images accurately, understand user queries, and generate meaningful, context-aware responses in real time.

RESULTS AND DISCUSSION

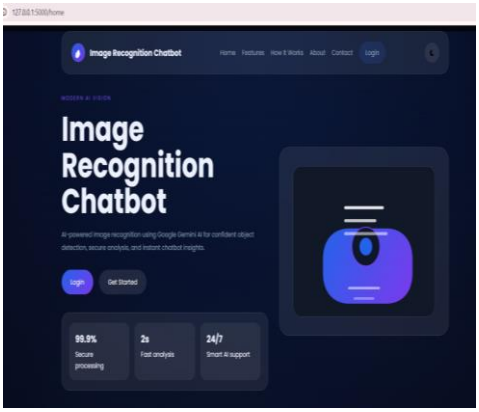

Fig 2: Home Page

The home page serves as the main interface of the Conversational Image Recognition Chatbot. It provides users with an attractive and user-friendly interface for accessing the image recognition system. Users can navigate through the dashboard, upload images, and view AI-generated image analysis results. The interface is designed using HTML, CSS, JavaScript, and Flask to provide a smooth user experience.

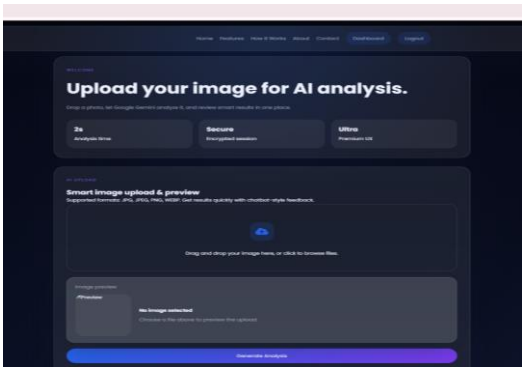

Fig 3: Dashboard Page

The dashboard is the central module of the application where users upload images for analysis. It includes drag-and-drop image upload functionality, image preview, and an option to generate AI-based image recognition results. The dashboard communicates with the Flask backend, which processes the uploaded image using the Google Gemini API**.**

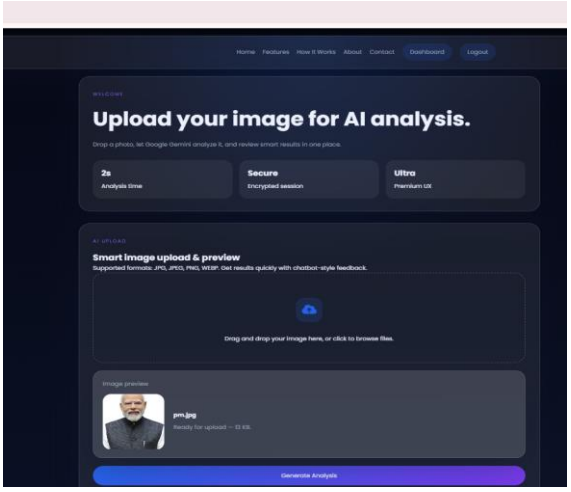

Fig 4: Image Upload Page

The Image Upload module allows users to upload images in JPG, JPEG, PNG, or WEBP formats. After uploading, the image is securely stored on the server and prepared for AI analysis. Validation checks ensure that only supported image formats are accepted.

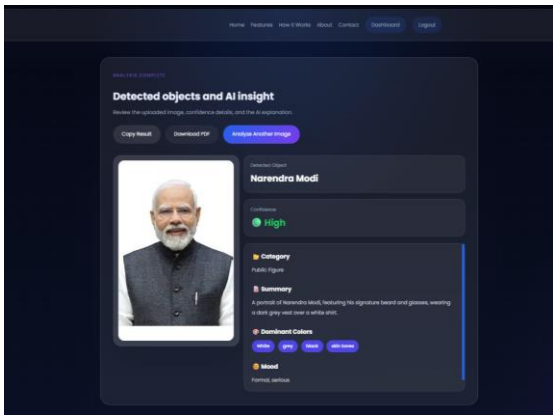

Fig 5: AI Image Analysis Result

After an image is uploaded, it is processed by the Flask backend and sent to the **Google Gemini 2.5 Flash** model for analysis. The AI model examines the visual content and generates structured information, including the detected object, confidence level, category, summary, dominant colours, and overall mood of the image. The generated results are then displayed in a well-organised dashboard, allowing users to quickly understand the contents of the uploaded image. This demonstrates the effectiveness of the proposed Conversational Image Recognition Chatbot in providing accurate, meaningful, and user-friendly image analysis through advanced multimodal artificial intelligence.

CONCLUSION

The proposed **Conversational Image Recognition Chatbot** provides an intelligent and user-friendly solution for analysing images using Google's **Gemini 2.5 Flash** multimodal AI model. The system enables users to upload images and receive meaningful information, including the detected object, confidence level, category, summary, dominant colours, and overall mood. By integrating artificial intelligence with a Flask-based web application, the system delivers accurate image understanding through a simple and interactive interface. The chatbot enhances the user experience by presenting image analysis in a conversational and organized format, making it suitable for educational, business,and general-purpose applications. Overall, the proposed system demonstrates the effectiveness of multimodal AI in providing fast, reliable, and informative image recognition without requiring users to have technical expertise.

FUTURE SCOPE

The proposed Conversational Image Recognition Chatbot can be further enhanced by integrating advanced computer vision and deep learning technologies to improve image recognition accuracy and support more complex visual analysis tasks. Face recognition capabilities can be incorporated to identify registered users or

recognize well-known public figures using secure and authorized databases. Optical Character Recognition (OCR) can be integrated to extract and summarize text from documents, receipts, invoices, and handwritten notes. The system can also be extended to support real-time webcam image recognition, multilingual image descriptions, voice-based interaction, and automatic image caption generation. Future improvements may include cloud deployment, mobile application support, user authentication with image history, image similarity search, and integration with Internet of Things (IoT) devices for smart applications. These enhancements will make the chatbot more intelligent, scalable, and suitable for various real-world applications across education, healthcare, security, business, and smart automation.

REFERENCES

[1] P. Harini, "Sms based wireless e-notice board," *International Journal of Emerging Technology and Advanced Engineering,* vol. 3, no. 3, pp. 181-185, 2013.

[2] P. Harini, "Enhanced packet delivery techniques using crypto-logic riddle on jamming attacks for wireless communication medium," *Int. J. Latest Trends Eng. Technol,* vol. 2, no. 4, pp. 469-478, 2013.

[3] P. Harini, "[Retracted] Machine Learning Algorithms Are Applied in Nanomaterial Properties for Nanosecurity," *Journal of Nanomaterials,* vol. 2022, no. 1, p. 5450826, 2022.

[4] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). *Attention Is All You Need*. Advances in Neural Information Processing Systems (NeurIPS).

[5] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). *An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale*. International Conference on Learning Representations (ICLR).

[6] Radford, A., Kim, J. W., Hallacy, C., et al. (2021). *Learning Transferable Visual Models From Natural Language Supervision*. International Conference on Machine Learning (ICML).

[7] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

[8] Simonyan, K., & Zisserman, A. (2015). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. International Conference on Learning Representations (ICLR).

[9] He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep Residual Learning for Image Recognition*. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[10] OpenCV Documentation. (2025). Available: https://docs.opencv.org/

[11] Mozilla Developer Network (MDN). (2025). *HTML Documentation*. Available: https://developer.mozilla.org/docs/Web/HTML

[12] Mozilla Developer Network (MDN). (2025). *CSS Documentation*. Available: https://developer.mozilla.org/docs/Web/CSS

[13] Mozilla Developer Network (MDN). (2025). *JavaScript Guide*. Available: https://developer.mozilla.org/docs/Web/JavaScript

[14] Jinja Documentation. (2025). *Jinja Template Designer Documentation*. Available: https://jinja.palletsprojects.com/

[15] World Wide Web Consortium (W3C). (2025).*Web Standards*. Available: https://www.w3.org/

[16] Brownlee, J. (2019). *Deep Learning for Computer Vision*. Machine Learning Mastery.