IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, April 2026
Impact Factor: 8.2

Multimodal Conversational Chatbot Using Image Recognition and NLP

S. Fowjiya, A. Abinaya, V. Keerthana, C. Kiruthika
Department of Computer Science and Technology
Vivekanandha College of Engineering for Women (Autonomous), Tiruchengode, Namakkal, Tamil Nadu, India
fowjiya@vcew.ac.in, abinayaabi6677@gmail.com,
keerthanavelusamy2004@gmail.com, kiruthikachandran11@gmail.com

Abstract: *The Pluggable Visual Conversational Intelligence Engine proposed in this study allows users to engage in multi-turn conversations based on a single uploaded image. The system can understand visual inputs and preserve context throughout a session, in contrast to standard text-only chatbots. It combines session-based memory with pretrained vision-language models and offers REST APIs for simple application integration. Instead of developing new AI models, the emphasis is on creating a modular, deployable, and controlled framework. All things considered, the system shows how conversational AI and picture interpretation may be integrated into a structured, expandable pipeline appropriate for academic and future domain-specific applications.*

Keywords: Visual Conversational Intelligence, Image-grounded Dialogue, Multi-turn Conversation, Vision–Language Models

I. INTRODUCTION

Conversational AI is widely used in digital platforms and academic systems, but most chatbots use only text input . In real-world academic settings, many queries are related to images such as lab equipment, diagrams, and environments, which traditional chatbots cannot effectively handle, limiting their usefulness. Although recent vision–language models can process both images and text, they are typically offered as closed, centralized systems that lack transparency, customization, and ease of deployment in institutional environments. This creates a gap between advanced AI capabilities and their practical use in academia. This work aims to bridge that gap by developing a modular and deployable visual conversational system that enhances existing chatbots with image understanding capabilities.

II. EXISTING SYSTEM

The existing system in conversational artificial intelligence primarily relies on text-based chatbot solutions that use natural language processing techniques to respond to user queries. These systems are widely implemented in educational platforms, customer service portals, and institutional applications. However, they are limited to processing only textual inputs and cannot interpret or analyse visual data. Most traditional chatbot function using rule-based approaches or predefined responses, along with basic machine learning models, and they often have limited capability in maintaining context during multi-turn conversations.

III. PROPOSED SYSTEM

The proposed Pluggable Visual Conversational Intelligence Engine (PVCI) is an advanced system designed to enable intelligent interaction using both images and text. It enhances traditional chatbot systems by allowing users to upload an image and ask multiple questions related to it, enabling more meaningful and interactive communication. The system uses a pretrained vision–language model such as BLIP to process images and extract important visual features. At the same time, transformer-based models are used to understand user queries in natural language, ensuring accurate interpretation of questions. A multimodal fusion mechanism combines both visual and textual information to generate

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, April 2026
Impact Factor: 8.2

context-aware and relevant responses. Additionally, a session memory module stores the uploaded image and previous interactions, allowing users to have continuous multi-turn conversations based on the same image.

IV. METHODOLOGY

4.1 Image Understanding Module:

The Image Understanding Module is responsible for processing user-uploaded images and extracting meaningful visual information. Its main functions include accepting images from users, resizing and normalizing them for consistent input, and converting them into embeddings using a Vision Transformer.

4.2 Text Understanding Module:

The Text Understanding Module is responsible for processing natural language questions and converting them into semantic representations. Its main functions include tokenizing user input, generating contextual embeddings using a Transformer model, and capturing the semantic meaning of the text. This module supports multiple question formats and ensures that the textual information can be effectively aligned with visual features.

4.3 Multimodal Fusion Module:

The Multimodal Fusion Module is responsible for integrating image and text features to enable cross-modal reasoning. Its main functions include aligning image embeddings with text embeddings, performing cross-attention to capture interactions between modalities, and generating a fused multimodal representation. This allows the system to provide image-grounded responses that consider both visual and textual context.

4.4 Session Memory and Control Module:

The Session Memory and Control Module is responsible for maintaining conversational context across multiple interactions. Its main functions include storing uploaded images, recording previous questions and answers, and preserving session-specific information. The module also implements validation rules and guardrails to ensure that responses remain consistent, controlled, and contextually relevant during multi-turn conversations.

4.5 Deployment and Integration Module:

The Deployment and Integration Module is responsible for enabling the practical usability of the system. Its main functions include providing a Gradio-based user interface for easy interaction, exposing REST APIs for communication between modules, supporting integration with external applications, and ensuring modular and scalable deployment. This ensures that the multimodal system can be efficiently utilized in real-world scenarios.

V. SYSTEM ARCHITECTURE

5.1 Image Processing & Feature Extraction Module:

This module handles all visual inputs and converts them into meaningful representations. Its main functions include:

- Image Upload & Validation: Accept images from the user and ensure proper format and size.
- Preprocessing: Resize, normalize, and optionally apply augmentation for consistency.
- Feature Extraction: Use Vision Transformers (ViT) or CNNs to extract object-level and scene-level embeddings.
- Output: Structured embeddings that represent visual content for the fusion module.

5.2 Natural Language Understanding (NLU) Module:

This module processes textual queries and converts them into semantic embeddings. Its main functions include:

- Text Preprocessing: Tokenization, normalization, and handling of different question formats.
- Contextual Embedding Generation: Use Transformer-based models (like BERT or GPT variants) to capture meaning.
- Semantic Representation: Encode text so that it can be aligned with image embeddings.
- Output: Contextual text embeddings ready for multimodal fusion.

5.3 Multimodal Fusion & Reasoning Module:

This module merges visual and textual information for meaningful responses. Its main functions include:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, April 2026
Impact Factor: 8.2

• Cross-Modal Alignment: Align image embeddings with text embeddings for coherent reasoning.
• Cross-Attention Mechanism: Capture interactions between the modalities.
• Fused Representation Generation: Produce a unified embedding containing both visual and textual context.
• Decision Layer: Determine what information is relevant for answering the user query.

5.4 Conversation Management & Memory Module:

This module maintains context and supports multi-turn interactions. Its main functions include:
• Session Memory: Store uploaded images, past queries, and responses.
• Context Tracking: Keep track of conversation state to enable coherent multi-turn dialogue.
• Validation & Safety: Ensure responses are consistent, safe, and contextually accurate.
• Output: Context-aware embeddings and memory cues for response generation.

5.5 Deployment & Integration Module:

This module ensures that the chatbot can be accessed and integrated efficiently. Its main functions include:
• User Interface: Provide a frontend interface using tools like Gradio or Streamlit.
• API Services: Expose REST or GraphQL APIs for communication between modules or integration with external apps.
• Scalability & Monitoring: Deploy on cloud platforms with scalable infrastructure; monitor usage and performance.
• Modular Architecture: Support independent updates for text, image, or fusion modules without downtime.

Result Display & Report Generation:

The Result Display & Report Generation Module shows the outputs of the multimodal chatbot in a clear and easy-to-understand way and also creates reports for later analysis. It displays the chatbot's text answers and gives visual feedback by highlighting objects or areas in uploaded images, like using boxes or heatmaps. The module can combine text explanations with annotated images, so users can interact with the visuals to understand them better. It keeps a record of the conversation, including images, questions, and answers, and can create summaries of detected objects, scene descriptions, and common queries

5.6 FLOWCHART

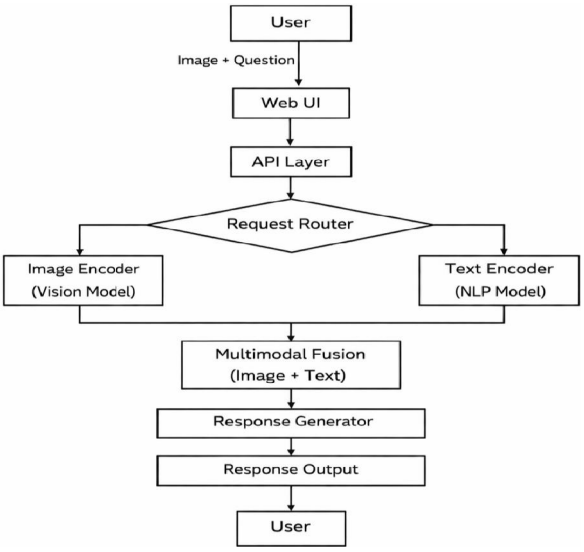

Fig: 5.1 Flowchart

DOI: 10.48175/IJARSCT-32408

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, April 2026
Impact Factor: 8.2

VI. RESULTS AND DISCUSSION

6.1 OUTPUT

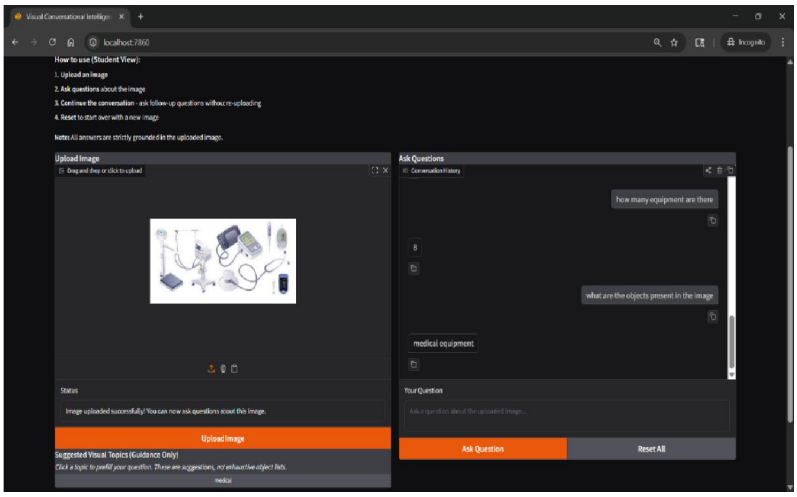

Fig 6.1.1 Image Understanding

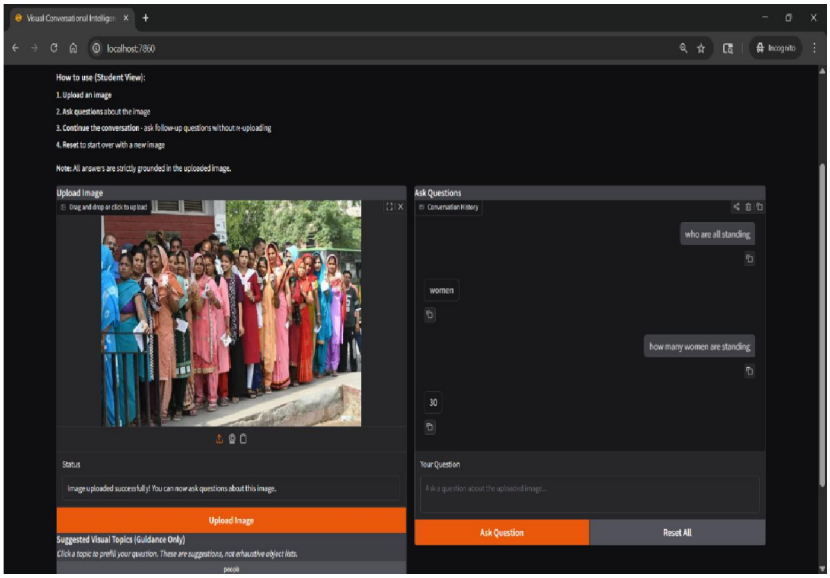

Fig 6.1.2 Text Understanding

VII. CONCLUSION

The Multimodal Conversational Chatbot successfully integrates image recognition and natural language processing to provide intelligent, context-aware responses. By combining visual and textual understanding, the system can interpret user queries more effectively, answer questions about images, and support interactive multi-turn conversations. The use of advanced modules for image processing, text understanding, and multimodal fusion ensures accurate reasoning and meaningful responses. Additionally, session memory and report generation enhance usability by keeping track of interactions and providing structured outputs for analysis. Overall, this project demonstrates the potential of multimodal AI systems to improve human-computer interaction, making digital assistants more intuitive, informative, and versatile for research, educational, and real-world applications.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 2, April 2026
Impact Factor: 8.2

VIII. ACKNOWLEDGMENT

We sincerely thank our mentors and faculty advisors for their guidance and support throughout this project. We also acknowledge the developers and open-source communities of machine learning, image recognition, and NLP tools that made this work possible. Finally, we are grateful to our friends and colleagues for their encouragement and valuable feedback.

REFERENCES

[1] Radford, A., et al., "Learning Transferable Visual Models From Natural Language Supervision," ICML, 2021. https://arxiv.org/abs/2103.00020
[2] Li,J., et al., "BLIP: Bootstrapping Language–Image Pre-training," arXiv, 2022. https://arxiv.org/abs/2201.12086
[3] Antol, S., et al., "VQA: Visual Question Answering," ICCV, 2015. https://arxiv.org/abs/1505.00468
[4] Das, A., et al., "Visual Dialog," CVPR, 2017. https://arxiv.org/abs/1611.08669
[5] HuggingFace – https://huggingface.co
[6] Gradio – https://www.gradio.app